Causal Inference with Measurement Error

by

Di Shu

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

Waterloo, Ontario, Canada, 2018

Examining Committee Membership

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

External Examiner:	Dr. Erica Moodie Associate Professor, McGill University
Supervisor:	Dr. Grace Yi Professor
Internal Member:	Dr. Cecilia Cotton Associate Professor
Internal Member:	Dr. Yeying Zhu Assistant Professor
Internal-External Member:	Dr. Ning Jiang Assistant Professor, Department of Systems Design Engineering

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

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

Abstract

Causal inference methods have been widely used in biomedical sciences and social sciences, among many others. With different assumptions, various methods have been proposed to conduct causal inference with interpretable results. The validity of most existing methods, if not all, relies on a crucial condition: all the variables need to be precisely measured. This condition, however, is commonly violated. In many applications, the collected data are not precisely measured and are subject to measurement error. Ignoring measurement error effects can lead to severely biased results and misleading conclusions. In order to obtain reliable inference results, measurement error effects should be carefully addressed.

Outside the context of causal inference, research on measurement error problems has been extensive and a large body of methods have been developed. In the paradigm of causal inference, however, there is limited research on measurement error problems, although an increasing, but still scarce, literature has emerged. Certainly, this is an area that deserves in-depth investigation. Motivated by this, this thesis focuses on causal inference with measurement error. We investigate the impact of measurement error and propose methods which correct for measurement error effects for several useful settings. This thesis consists of nine chapters.

As a preliminary, Chapter 1 gives an introduction to causal inference, measurement error and other features such as missing data, as well as an overview of existing methods on causal inference with measurement error. In this chapter we also describe the problems of our interest that will be investigated in depth in subsequent chapters.

Chapter 2 considers estimation of the causal odds ratio, the causal risk ratio and the causal risk difference in the presence of measurement error in confounders, possibly time-varying. By adapting two correction methods for measurement error effects applicable for the noncausal context, we propose valid methods which consistently estimate the causal effect measures for settings with error-prone confounders. Furthermore, we develop a linear combination based method to construct estimators with improved asymptotic efficiency.

Chapter 3 focuses on the inverse-probability-of-treatment weighted (IPTW) estimation of causal parameters under marginal structural models with error-contaminated and

time-varying confounders. To account for bias due to imprecise measurements, we develop several correction methods. Both the so-called stabilized and unstabilized weighting strategies are covered in the development.

In Chapter 4, measurement error in outcomes is of concern. For settings of inverse probability weighting (IPW) estimation, we study the impact of measurement error for both continuous and binary outcome variables and reveal interesting consequences of the naive analysis which ignores measurement error. When a continuous outcome variable is mismeasured under an additive measurement error model, the naive analysis may still yield a consistent estimator; when the outcome is binary, we derive the asymptotic bias in a closed-form. Furthermore, we develop consistent estimation procedures for practical scenarios where either validation data or replicates are available. With validation data, we propose an efficient method. To provide protection against model misspecification, we further develop a doubly robust estimator which is consistent even when one of the treatment model and the outcome model is misspecified.

In Chapter 5, the research problem of interest is to deal with measurement error generated from more than one sources. We study the IPW estimation for settings with mismeasured covariates and misclassified outcomes. To correct for measurement error and misclassification effects simultaneously, we develop two estimation methods to facilitate different forms of the treatment model. Our discussion covers a broad scope of treatment models including typically assumed logistic regression models as well as general treatment assignment mechanisms.

Chapters 2-5 emphasize addressing measurement error effects on causal inference. In applications, we may be further challenged by additional data features. For instance, missing values frequently occur in the data collection process in addition to measurement error. In Chapter 6, we investigate the problem for which both missingness and misclassification may be present in the binary outcome variable. We particularly consider the IPW estimation and derive the asymptotic biases of three types of naive analysis which ignore either missingness or misclassification or both. We develop valid estimation methods to correct for missingness and misclassification effects simultaneously. To provide protection against misspecification, we further propose a doubly robust correction method.

Doubly robust estimators developed in Chapter 6 offer us a viable way to address issues of model misspecification and they can be easily applied for practical problems. However, such an appealing property does not say that doubly robust estimators have no weakness. When both the treatment model and the outcome model are misspecified, such estimators will not necessarily be consistent. Driven by this consideration, in Chapter 7, we propose new estimation methods to correct for effects of misclassification and/or missingness in outcomes. Differing from the doubly robust estimators which are constructed based on a *single* treatment model and a *single* outcome model, the new methods are developed by considering a *set* of treatment models and a *set* of outcome models. Such enlargements of the associated models enable us to construct consistent estimators which will enjoy the so-called multiple robustness, a property that has been discussed in the literature of missing data.

To expedite the application of our developed methods, we implement the proposed methods in Chapter 4 and develop an R package for general users. The details are included in Chapter 8. The thesis concludes with a discussion in Chapter 9.

Acknowledgements

I have many thanks to offer. First and foremost, I want to express my deepest gratitude to my supervisor Dr. Grace Y. Yi, who has always offered her guidance and support throughout my Ph.D. studies. She helped to broaden my horizons, encouraged me to deepen the work and explore the unknowns. Without her advice and insights, it would be impossible for me to complete this thesis. Her advice, not limited to research, will continue to benefit me in the future. I am so lucky to have Dr. Yi as my supervisor.

I also thank Dr. Cecilia Cotton, Dr. Yeying Zhu, Dr. Erica Moodie and Dr. Ning Jiang for serving as my committee members and providing thoughtful and invaluable comments.

In addition, I am grateful to the department faculty and staff, for the knowledge I have learnt and all the help and support I have received.

Many thanks go to my friends for providing support, encouragement and fun, especially Nathalie Moon, who has always been there to encourage me, help me and listen to me.

Last but certainly not least, I would like to thank my family for their love and support. You are always my source of motivation and power.

To my family

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Causal inference and measurement error are two important aspects of statistics. Many authors have studied them separately, and there is limited work on combining the two features although this area is attracting increasing attention. In this thesis, we consider several important problems concerning causal inference with measurement error. This chapter presents a brief introduction of some key aspects, which serves as a preliminary to later chapters. Starting from Chapter 2, we consider causal inference and measurement error simultaneously. In addition, this chapter also gives a brief introduction to missing data problems, which is of relevance to the development of Chapters 6 and 7. Research problems of our interest are introduced, and they will be investigated in depth in later chapters.

1.1 Basics of Causal Inference

This section introduces some key ideas and concepts in causal inference, which are closely related to the later chapters.

1.1.1 Causality and Neyman-Rubin Causal Model

In scientific research, the causal effect is often of ultimate interest. For example, does smoking cause lung cancer? If it does, what is the magnitude of this causal effect? To answer these questions, we should first define "causality". However, the meaning of causality is often vague and depends on the context.

Consider the following statement: smoking causes lung cancer. In strict sense, this statement means that individuals will get lung cancer for sure if they smoke, and individuals will never get lung cancer if they never smoke. Obviously, such a simple and strict explanation usually does not hold. In reality, an individual who smokes may or may not develop lung cancer, and an individual who never smokes may or may not develop lung cancer as well. Therefore, to establish the existence of a causal relationship, the language of probability and statistics should be introduced. The difficulty is how to use statistical methods to describe the causality.

In 1965, Hill (1965) proposed his criteria for causality. Hill's criteria list the following guidelines for establishing causality: *strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment and analogy.* The criteria have been widely used in epidemiological studies. However, they do not provide a mathematical or statistical framework to define and estimate the causal effects. According to Pearl (2009a), "Causality is not mystical or metaphysical. It can be understood in terms of simple processes, and it can be expressed in a friendly mathematical language, ready for computer analysis."

To explicitly describe the causal framework, the Neyman-Rubin causal model was proposed (Neyman, 1923; Rubin, 1974; Holland, 1986). Specifically, we let i indicate the individual index. We let $Y_{i,r}$ be the potential outcome that would have been observed had the subject i been given treatment r, and let $Y_{i,s}$ be the potential outcome that would have been observed had the subject i been given treatment s. The causal effect of treatment r versus s can be assessed by comparing $Y_{i,r}$ and $Y_{i,s}$. In reality, however, we can never observe both $Y_{i,r}$ and $Y_{i,s}$ simultaneously for $r \neq s$. When the individual is assigned treatment s, $Y_{i,r}$ is unobserved (thus counterfactual) and vice versa. Thus $Y_{i,r}$ and $Y_{i,s}$ can never be compared directly for subject i. Although the *individual-based* causal effect cannot be obtained due to unavailability of all potential outcomes, the *population-based* causal effect

can be estimated if suitable conditions are imposed. In particular, we are interested in comparing $E(Y_r)$ and $E(Y_s)$, where the expectations are taken with respect to the same population: our target population; and Y_r and Y_s represent the potential outcomes for an individual in the population had the individual been given treatments r and s, respectively. The difference $E(Y_r) - E(Y_s)$ represents the average treatment effect of treatment r versus s. In most cases, by letting $r = 1$ and $s = 0$, we deal with the binary treatment case and are interested in estimation of $E(Y_1) - E(Y_0)$. Extensions are available; see Imbens (2000) for multi-valued treatments and Hirano and Imbens (2004) for continuous treatments.

Throughout this thesis, we will stick to the Neyman-Rubin causal model framework, which is also called the *potential outcome* or *counterfactual* framework.

1.1.2 Standard Assumptions for Causal Inference

The Neyman-Rubin causal model framework defines the potential outcomes which cannot be observed simultaneously. This can be regarded as a missing data problem. Like in statistical inference with missing data, some assumptions need to be made to enable us to conduct valid causal estimation. Let T and X denote the observed binary treatment indicator and a vector of pre-treatment covariates, respectively. We list the fundamental assumptions in causal inference (e.g., Rosenbaum and Rubin, 1983; Lunceford and Davidian, 2004; Cole and Frangakis, 2009; Hernán and Robins, 2016) as follows:

Assumption 1 (No Interference): the treatment taken by one subject has no effect on the potential outcomes of another subject.

Assumption 2 (Consistency): the potential outcome under the observed treatment is equal to the observed outcome Y, i.e., $Y = TY_1 + (1 - T)Y_0$.

Assumption 3 (No Unmeasured Confounding): the treatment received is independent of the potential outcomes, given confounders X, i.e., $(Y_1, Y_0) \perp\!\!\!\perp T \mid X$.

Assumption 4 (Positivity): given the confounders X, all subjects have a positive probability to receive treatment or not, i.e., $0 < P(T = 1|X) < 1$.

The no unmeasured confounders assumption implies that controlling for X is sufficient to eliminate confounding bias. If this assumption is violated, the confounding bias caused by unmeasured confounders generally cannot be eliminated, and the resulting causal results could be misleading due to the residual confounding bias.

The positivity assumption is also required, because no comparison between the treated and untreated groups can be made if no data exist for the treated or untreated group.

Unfortunately, like missing data analysis, causal inference usually involves untestable assumptions.

1.1.3 Propensity Score Methods and Beyond

The propensity score e is defined as the probability of receiving treatment given observed covariates or confounders; i.e., $e(X) = P(T = 1|X)$. Rosenbaum and Rubin (1983) showed that $T \perp\!\!\!\perp X \mid e(X)$. This equality implies that given $e(X)$, X offers no more information on treatment assignment. So $e(X)$ is a good scalar summary of X without losing information on treatment assignment. Moreover, given $e(X)$, the distribution of X should be well balanced in the treated and untreated groups, indicating that the confounding bias caused by X can be eliminated by controlling for a scalar $e(X)$, rather than controlling for a vector X, which can be of high dimension.

Given causal inference assumptions, Rosenbaum and Rubin (1983) showed that $(Y_1, Y_0) \perp\!\!\!\perp T \mid e(X)$. Therefore, given $e(X)$, the distribution of potential outcomes are independent of the treatment assignment.

Above theoretical results justify the key role that propensity scores play in causal estimation: controlling for propensity score eliminates confounding bias. Based on propensity scores, causal inference methods have been proposed using matching (Rosenbaum and Rubin, 1983, 1985; Heckman et al., 1998; Ho et al., 2007), stratification (Rosenbaum and Rubin, 1983, 1984), covariance adjustment (Rosenbaum and Rubin, 1983) and inverse probability weighting (IPW)(Horvitz and Thompson, 1952; Rosenbaum, 1987, 1998; Robins et al., 2000; Lunceford and Davidian, 2004). The propensity score methods are easy to implement and have been increasingly used.

Propensity score methods are commonly employed to analyze data arising from observational studies where confounding is pervasive. In some randomized trials where the randomization is not properly implemented, i.e., distributions of the baseline variables are imbalanced between the treated and untreated groups even after randomization, adjusting for the baseline variables helps reduce the residual confounding (e.g., Kahan et al., 2014). Williamson et al. (2014) demonstrated the IPW adjustment can account for chance imbalance of the baseline variables and increase the precision of the estimated treatment effect in randomized trials. Therefore, propensity score methods are applicable to both observational studies and randomized trials.

All propensity score methods share the same first step: estimate the propensity score. In many cases, we fit a regression model (often, a logistic regression model) to relate treatment assignment T and covariates X and estimate the propensity score by calculating the fitted probability of the regression model. Machine learning methods have also been used (McCaffrey et al., 2004; Westreich et al., 2010; Lee et al., 2010). Kang and Schafer (2007) found that the causal effects estimation is vulnerable to model misspecification and slight misspecification of the propensity score model could lead to severely biased causal effects estimates. To improve the robustness, doubly robust estimators have been proposed (e.g., Robins et al., 1994; Scharfstein et al., 1999; Bang and Robins, 2005; Qin et al., 2008; Cao et al., 2009; Tan, 2010), where by doubly robust we mean the estimator is consistent even when either the treatment model or the outcome model is misspecified. Imai and Ratkovic (2014) proposed the covariate balancing propensity score (CBPS) which incorporates the covariate balancing property of propensity score into the estimation procedure. Fan et al. (2016) proposed a doubly robust estimator by generalizing CBPS.

1.1.4 Extensions to Time-Dependent Treatment Studies

The concept of propensity scores is initially designed for time-independent treatment studies (or point-treatment studies). However, it is common that many studies may involve time-dependent treatments and time-dependent confounders which are risk factors for the outcomes, predict subsequent treatment, and are predicted by the past treatment. In such a setting, standard methods for adjusting for confounding can lead to biased causal

estimates (Robins et al., 2000). To consistently estimate the causal effects, Robins and his colleagues have developed several methods, including inverse-probability-of-treatment weighted (IPTW) estimation of marginal structural models, parametric g-computation formula, g-estimation of structural nested models and the iterative conditional expectations estimation; see Robins (1989), Robins et al. (1992), Robins (1997), Robins (1998), Robins (1999), Robins et al. (2000) and Hernán et al. (2000) for details. In the presence of time-dependent confounding, these methods can be employed to produce consistent causal estimates under certain assumptions. For example, Polis et al. (2013) suggested using them to control for time-dependent confounders for the assessment of causal effect of hormonal contraception on HIV acquisition.

Among those methods, the IPTW estimation of marginal structural models has been widely used, partially due to its straightforward implementation and close relation to the standard analysis. However, its popularity is not a guarantee that investigators can utilize it blindly, because its validity hinges on a set of assumptions that may not hold for every situation.

1.1.5 Other Methods

The aforementioned methods, including the propensity score methods and the IPTW estimation of marginal structural models in particular, have been widely used. However, they cannot solve all the problems in causal inference. For example, propensity score methods require that all confounders which need to be adjusted for are fully observed and violation of this would lead to biased causal estimates. Instrumental variable estimation (e.g., Baiocchi et al., 2014; Hernán and Robins, 2016) is a method which can deal with such violation. It has its own set of assumptions without requiring observing all the confounders which need to be adjusted for. Principal stratification (Frangakis and Rubin, 2002) is an approach that yields causal effects within principal strata. Targeted learning (van der Laan and Rose, 2011) is a method which incorporates various machine learning tools into estimation procedure, with the aim of conducting efficient causal inference in a more data-driven way.

We did not name all the causal inference methods here. Our aim is to introduce the basic idea and concept in causal inference. For example, we did not discuss causal diagrams

(Greenland et al., 1999; Hernán and Robins, 2016) in the previous subsections, although it is a very important tool for the illustration of relationships between variables. It displays the causal mechanism and tells us which variables should be adjusted for and therefore guides the model specification in causal inference.

Comprehensive treatments of causal inference can be found in Rothman et al. (2008), Pearl (2009a), van der Laan and Rose (2011), Imbens and Rubin (2015), Hernán and Robins (2016) and Rosenbaum (2017).

1.2 Measurement Error Problems

Unlike mathematical theories, all statistical analysis procedures involve using data, which need to be measured and collected. To guarantee the validity of a certain statistical method, in addition to its own set of assumptions, another crucial assumption is implicitly required: all the measurements of variables are precise. When this assumption is violated, measurement error problems arise.

In this section, basic concepts of measurement error models are presented.

1.2.1 Sources of Measurement Error

Measurement error can arise for both continuous and discrete variables, and the measurement error in discrete variables is usually called misclassification. In practice, many data are subject to measurement error due to various reasons (e.g., Fuller, 1987; Gustafson, 2003; Carroll et al., 2006; Buonaccorsi, 2010; Yi, 2017). We list some reasons here. First, the true value of the data is a long-term average, which is impossible to measure at a certain visit. For example, the long-term average blood pressure can never be directly measured by a clinic visit. Second, the data set includes some self-reported items in a questionnaire, such as smoking status and dietary intake, which are subject to recall bias. Third, each diagnostic test has its own test sensitivity and test specificity. With a positive probability, a diagnostic test can wrongly identify a subject. For example, in medical diagnosis, it is possible for an undiseased patient to be diagnosed with a disease. Unfortunately, there

is a trade-off between the false positive rate and false negative rate. Decreasing the false positive rate may inflate the false negative rate, and vice versa. Fourth, the measurement equipment may only report measurement in a pre-specified range, and the variable with value outside the range will be mismeasured as the limit of the pre-specified range. Fifth, the measurement error can be caused by mis-recording. For example, suppose the true height of an individual is 170cm, it can be misrecorded as 178cm into the spreadsheet, because of human error. Even the birth date in a birth certificate can be wrong, although it seems not likely to happen. Sixth, precise measurements may be too expensive to afford, with a given budget.

A world without measurement error would be perfect. Can we really avoid measurement error? The answer is "No". Development of technology may make better measurement equipment and lower the cost of precise measurement. But, no matter how careful we are, recall bias and human error can never be eliminated. Although measurement error can not be eliminated for many studies, the efforts we have made in study design and data collection to try to reduce measurement error can improve the data quality. For example, although the true long-term average of a quantity can never be observed, we can take the average of multiple measurements to reduce some random variation.

1.2.2 Measurement Error Models

Measurement error models are statistical models describing the underlying mechanism of measurement error. They present the relationship between the observed error-prone measurements and other variables. Various measurement error models have been proposed in the literature, and our intent is not to make an exhaustive list of them. Instead, we selectively present the most commonly used models for continuous variables and for discrete variables. More complicated measurement error models can be found in Carroll et al. (2006) and Yi (2017). Let W denote the observed measurements of unobserved true covariates X. Both W and X are assumed to be scalar variables for ease of exposition.

For continuous variables, the most widely used model is the classical additive model (Carroll et al., 2006):

$$W = X + U, \tag{1.1}$$

where $E(U|X) = 0$, and often U is assumed to be independent of X and follow a normal distribution.

Another model is complement to model (1.1) and called the Berkson model (Berkson, 1950; Carroll et al., 2006):

$$X = W + U, \tag{1.2}$$

where $E(U|W) = 0$, and often U is assumed to be independent of W and follow a normal distribution.

Note that compared to model (1.1), in model (1.2), the positions of X and W are exchanged. Although two models look similar, they have different implications. In model (1.1), the variation of W is larger than X, since $Var(W) = Var(X) + Var(U)$. On the contrary, the variation of W is smaller than X in model (1.2), since $Var(X) = Var(W) + Var(U)$.

For discrete variables, the measurement error mechanism can be characterized by the misclassification probability $P(W = r|X = s)$, where r and s are possible realizations of W and X. In the situation where X and W are binary, the probabilities $P(W = 1|X = 1)$ and $P(W = 0|X = 0)$ are called the sensitivity and specificity, respectively. Like in the continuous variable case, sometimes, it is preferred to exchange the positions of X and W, and to model $P(X = s|W = r)$. Adding extra variables to the misclassification probability can help characterize more complicated measurement error mechanisms.

1.2.3 Data Requirements for Measurement Error Models

Although Section 1.2.2 outlines some useful measurement error models, it does not tell us how to handle the associated parameters such as $Var(U)$, or the misclassification probabilities $P(W = r|X = s)$ and $P(X = s|W = r)$ which are often unknown. Here we briefly outline the strategy or requirement for dealing with this issue.

Case 1: No Extra Information:

In this case, we specify reasonable values for $Var(U)$, $P(W = r|X = s)$ and $P(X = s|W = r)$. Such user-specified reasonable values may be borrowed from similar studies. We

can further specify a series of specified values, to see if the estimation results are sensitive to the specification. This process is the sensitivity analysis.

Case 2: Replication Data or Validation Data Available:

In this case, we have some extra information. By replication data we mean that the measurements of X are taken multiple times. By using the replication data, $Var(U)$ can be estimated (Carroll et al., 2006), when the measurement error is additive, and is independent and identically distributed across individuals. By validation data, we mean that the data include observed values of X. The validation data is called *internal* when it is a subset of the original data. The validation data can also be *external*, when it is independent of the original data. By using validation data, we can directly model the relationship between X and W, because they are both observed. When external validation data are used, it is necessary to carefully assess the appropriateness of transferring the measurement error mechanism for the external validation data to the original data.

1.2.4 Measurement Error Effects

In the absence of measurement error, standard analysis methods are valid if their own assumptions hold. It has been extensively studied and well documented that, in the presence of measurement error, naively substituting mismeasured variables into the standard regression analysis procedures can lead to severely biased estimated regression parameters (e.g., Fuller, 1987; Gustafson, 2003; Carroll et al., 2006; Buonaccorsi, 2010; Yi, 2017). We take a simple linear regression for example and write the regression model (Yi, 2017, Section 2.2):

$$Y = \beta_0 + \beta X + \epsilon, \tag{1.3}$$

where ϵ is independent of X, β_0 and β are the regression parameters of interest, and ϵ is distributed with mean 0.

However, X is unobserved and mismeasured as W, according to measurement error model (1.1). Model (1.3) can be re-expressed as

$$Y = \beta_0 + \beta W - \beta U + \epsilon. \tag{1.4}$$

If error term $-\beta U + \epsilon$ were independent of W and has mean 0, model (1.4) will be a standard linear regression model and the ordinary least square estimator of (1.4) will be a consistent estimator of β. However, $-\beta U + \epsilon$ is not independent of W, due to the correlation between U and W. As a result, model (1.4) is not a standard regression method relating Y and W. Furthermore, it can been shown analytically that naively conducting regression of Y on W leads to an asymptotically inconsistent estimator (Fuller, 1987).

1.2.5 Correcting for Measurement Error Effects

To conduct a valid estimation procedure, the measurement error effects should be well adjusted for. If there is a way to recover X based on W, the problem is solved. However, it is impossible. By model (1.1), we can never get X from W.

A fully consistent estimation method adjusting for measurement error usually depends on the standard analysis procedure. For example, consistent correction methods for linear models and logistic models may be different.

It is difficult to develop approaches which are easy to implement and applicable for all settings. However, two commonly-used approximately consistent approaches are available: the regression calibration method (Prentice, 1982; Rosner et al., 1989, 1990; Carroll et al., 2006) and the simulation-extrapolation (SIMEX) method (Cook and Stefanski, 1994). In addition to regression calibration and SIMEX, other general approaches have been developed, including the moment reconstruction (MR) method (Freedman et al., 2004) and multiple imputation (Cole et al., 2006; Freedman et al., 2008; Blackwell et al., 2017).

For comprehensive treatments of measurement error, see Fuller (1987), Gustafson (2003), Carroll et al. (2006), Buonaccorsi (2010), and Yi (2017).

1.3 Missing Data

Statistical analyses make inferences about the target population from sample data. Imperfect data present a challenge for the validity of standard statistical analyses which are

often developed under stringent assumptions and data. Mismeasured data discussed in Section 1.2.2 fall into the category of imperfect data. In this section, we describe another commonly-seen type of imperfect data: missing data.

1.3.1 Missingness Mechanisms

Missing data frequently arise in practice due to various reasons. In surveys, non-response occurs when respondents cannot be reached by phone call or mail, or refuse to answer sensitive questions. Sometimes, the respondents just do not know the answer. In longitudinal studies, dropout happens when patients stop showing up at follow-ups. In some situations, missingness is caused by administrative error. For example, the interviewer may forget to record the responses.

It is essential to understand the reasons why missing data occur, in order to conduct valid inferences. Let D be an $n \times p$ matrix of data, where n and p are the sample size and the number of variables, respectively. Define M to be the matrix of missing indicator with $M_{ij} = 1$ if the jth variable of the ith subject D_{ij} is missing and $M_{ij} = 0$ otherwise for $i = 1, \ldots, n$, $j = 1, \ldots, p$. Let D_{obs} and D_{miss} denote the collections of the observed and missing components of D, respectively. Three types of missingness mechanisms may be defined for $f(M|D)$, the conditional distribution of M, given D (e.g., Little and Rubin, 2002):

Missing Completely At Random (MCAR): $f(M|D) = f(M)$;

Missing At Random (MAR): $f(M|D) = f(M|D_{obs})$;

Missing Not At Random (MNAR): $f(M|D) \neq f(M|D_{obs})$ and the distribution of M depends on D_{miss};

where $f(M)$ stands for the marginal distribution of M and $f(M|D_{obs})$ represents the conditional distribution of M, given the observed data D_{obs}.

The MCAR assumption assumes that the missingness is independent of both observed and unobserved data, and is missing completely at random. If the MCAR assumption holds, then we are safe to regard the observed data as being representative for the whole

data we would have seen in the absence of missingness. It is reasonable to assume MCAR, for example, when the scanner stops working all of a sudden to scan students' answers to certain questions for grading.

The MAR assumption assumes that the missingness is independent of the unobserved data conditioning on the observed data. With MAR, we may be able to model the missingness process using the observed data and further to adjust for the missingness effect.

The MNAR assumption assumes that the missingness depends on the unobserved data, even after conditioning on the observed data. When the underlying missingness mechanism is MNAR, it is usually a difficult situation.

1.3.2 Handling Missing Data

Missing data commonly arise in practice and has attracted extensive research attention. Many methods have been developed to handle missing data, including maximum likelihood approaches (e.g., Rubin, 1976) such as the EM algorithm (e.g., Dempster et al., 1977), multiple imputation (e.g., Rubin, 1987, 1996), inverse probability weighting (e.g., Horvitz and Thompson, 1952), semiparametric methods (e.g., Robins et al., 1994, 1995), and empirical likelihood (Qin and Lawless, 1994; Owen, 2001) based approaches (e.g., Qin and Zhang, 2007; Han and Wang, 2013; Han, 2014b), among many others. For comprehensive treatments of missing data problems, see Little and Rubin (2002) and Tsiatis (2007).

Many methods developed for missing data in noncausal contexts can be applied to estimate treatment effects in causal inference. A unique feature in causal inference is that we can never observe all the potential outcomes for a subject at the same time. Thus, to some degree, this feature can be regarded as a missing data problem, and methods developed for missing data problems can be used to estimate treatment effects in causal inference.

It is interesting to note that multiple imputation serves as a correction method in both measurement error and missing data problems (Cole et al., 2006; Freedman et al., 2008; Blackwell et al., 2017). The idea comes from the following observation: missing data can be regarded as extreme measurement error, suggesting that measurement error can

be can be dealt with using missing data methods. Other common methods for handling measurement error and missing data problems include the observed likelihood approach and the EM algorithm (e.g., Yi, 2017).

1.4 Overview of Methods on Causal Inference with Measurement Error

Causal inference and measurement error have been extensively but separately studied. Recently, there has been increasing but still relatively little attention to causal inference with measurement error. Outside the context of causal inference methods, it has been well documented that ignoring measurement error effects can lead to biased results and misleading conclusions (e.g., Fuller, 1987; Gustafson, 2003; Carroll et al., 2006; Buonaccorsi, 2010; Yi, 2017). Similarly, the presence of error-prone data presents a challenge to the validity of causal inference methods. This section provides an overview of existing methods on dealing with measurement error in causal inference. Discussion on this topic can also be found in Yi (2017, Section 9.2).

Using causal diagrams, Hernán and Cole (2009) qualitatively depicted four types of measurement errors of exposure and outcome: independent nondifferential, dependent nondifferential, independent differential, and dependent differential. In graph-based causal inference, Pearl (2009b) proposed a matrix adjustment for measurement error effects and introduced correction methods for binary and linear models. Kuroki and Pearl (2014) discussed the correction for measurement errors in causal inference using graphical techniques. Edwards et al. (2015a) demonstrated that bias due to mismeasurements can be incorporated into the potential outcomes framework and considered together with other types of bias.

In mediation analysis, Ogburn and VanderWeele (2012) examined bias caused from nondifferential misclassification of a binary mediator. Blakely et al. (2013) investigated the effects of mediator misclassification on the estimation of direct and indirect effects.

Under misclassified exposure, Lewbel (2007) discussed identification and estimation is-

sues on the estimation of conditional average effect. Babanezhad et al. (2010) compared the performance of inverse probability of treatment weighted estimators, doubly robust estimators, G-estimators, propensity score adjusted estimators and ordinary least squares estimators. Braun et al. (2016, 2017) proposed estimation methods to correct for exposure misclassification in propensity score methods using validation data. Imai and Yamamoto (2010) discussed identification issues and sensitivity analysis for causal inference with differential measurement error.

In the presence of error-prone covariates, Regier et al. (2014) conducted a simulation study to understand the impact of measurement error on the IPTW estimation for marginal structural models. The simulation results not only reveal well-understood effects of attenuation and augmentation but also uncover two unanticipated effects: null effects and sign reversals. McCaffrey et al. (2013) proposed consistent inverse probability-weighted estimators with time-independent covariates that are subject to measurement error. Lockwood and McCaffrey (2016) studied matching and weighting estimators for settings where mismeasured covariates are time-independent. Kyle et al. (2016) applied the simulation-extrapolation method to deal with measurement error in time-varying covariates under marginal structural models. With a single covariate subject to mean reverting measurement error, Lenis et al. (2016) adapted the simulation-extrapolation method to a doubly robust estimator of the average treatment effect. For the adjustment for measurement error in covariates in propensity score methods, also see Raykov (2012) for a latent variable approach and Hong et al. (2017) for a Bayesian approach.

When the outcome variable is subject to misclassification, Gravel and Platt (2018) developed a weighting approach to deal with misclassification effects on the estimation of marginal causal odds ratio.

To adjust for unmeasured confounding, Stürmer et al. (2005) proposed the so-called propensity score calibration method using validation data, which apply the regression calibration method with treating propensity scores in the main data as error-prone data. The execution of this method can be viewed as applying the regression calibration method to correct for measurement error in confounders of the main data. Therefore, the propensity score calibration method can handle measurement error. In other words, measurement error can be treated as a special form of unmeasured confounding.

As described in Section 1.3.2, multiple imputation is a generally applicable tool for measurement error problems, like the regression calibration and simulation-extrapolation method. In the context of causal inference, multiple imputation has been applied to reduce bias due to error-prone covariates in propensity score analysis (Webb-Vargas et al., 2017) and misclassified exposure in marginal structural models (Edwards et al., 2015b).

1.5 Problems of Our Interest

In this thesis, we investigate six interesting and important problems.

The first problem concerns estimation of three widely used causal effect measures when the outcome variable is binary: the causal odds ratio, the causal risk ratio and the causal risk difference. In many applications, both confounding bias and measurement error occur. They should be adjusted for in the estimation of these causal effect measures. It is of interest to develop consistent estimation methods for the causal effect measures with error-prone and possibly time-varying confounders.

In the second problem, we move to the framework of marginal structural models. In the presence of time-dependent confounders which are also affected by the previous treatment, Robins et al. (2000) proposed the inverse-probability-of-treatment weighted (IPTW) estimation method to consistently estimate causal parameters in marginal structural models. This method is, however, vulnerable to the quality of data, because these methods were developed under the assumption that measurements were precisely collected. Kyle et al. (2016) applied the simulation extrapolation method to adjust for measurement error in time-varying covariates. It is of interest to explore more methods to correct for measurement error effects.

The third problem focuses on measurement error in outcomes, which can be a serious concern in causal inference but receives rather limited attention. Among existing causal inference methods, the inverse probability weighting (IPW) estimation methods enjoy easy implementation and transparent interpretations. It adjusts for the confounder effects by re-weighting the data so that the weighted data may be treated as if being collected from randomized controlled trials. It is interesting to investigate the consequence of ignoring

measurement error in outcomes in IPW estimation and develop methods to correct for measurement error effects.

Two types of error are considered simultaneously in the fourth problem. In the presence of measurement error in both covariates and outcomes, to our best knowledge, there has been no research on addressing measurement error effects on causal inference. We focus on the IPW estimation of average treatment effects for settings with mismeasured covariates and misclassified outcomes. It is of interest to develop estimation methods to correct for measurement error and misclassification effects simultaneously.

In addition to measurement error, our fifth problem is connected to missing data. Outside the context of causal inference, methods have been proposed to handle measurement error in covariates and missingness in responses simultaneously (Yi, 2008; Yi et al., 2011, 2012, 2015b). We consider both missingness and misclassification in outcomes. That is, the outcome data for some subjects are missing. For the rest of subjects, the outcome data are available but misclassified. In the presence of both missingness and misclassification in outcomes, to our best knowledge, there is no available work to address missingness and misclassification effects on causal inference. We are interested in the investigation of bias caused from ignoring missingness and/or misclassification in IPW estimation as well as the development of methods to eliminate missingness and misclassification effects.

Our six problem is driven by a desire to achieve robustness. Recently, to further pursue more protection against model misspecification than doubly robust methods, multiple robust estimation methods have been developed (Han and Wang, 2013; Chan and Yam, 2014; Han, 2014a,b, 2016; Chen and Haziza, 2017). An estimator is said to be multiply robust if it is consistent when at least one of the multiple postulated models is correctly specified. Multiply robust estimation methods may lose their multiple robustness if ignoring measurement error and/or missingness. To our best knowledge, none of the existing methods on causal inference with measurement error were developed for the settings of multiply robust estimation. We are interested in the development of multiply robust estimation of causal effects with missing and/or misclassified outcomes.

In short, we consider the following six research problems in this thesis:

- Problem 1: Simultaneous estimation of the causal odds ratio, the causal risk ratio and the causal risk difference with mismeasured and possibly time-varying confounders
- Problem 2: Estimation of causal parameters in marginal structural models with time-dependent confounders subject to measurement error
- Problem 3: Bias analysis and correction methods (including doubly robust estimation) for IPW estimation with measurement error in outcomes
- Problem 4: Weighted estimation of causal effects correcting for measurement error in covariates and misclassification in outcomes simultaneously
- Problem 5: Bias analysis and correction methods (including doubly robust estimation) for IPW estimation with misclassification and/or missingness in outcomes
- Problem 6: Multiply robust causal inference methods with misclassification and/or missingness in outcomes taken into account

1.6 Outline of the Thesis

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

In Chapter 2, we consider estimation of causal effect measures with error-prone confounders, possibly time-varying. By adapting two correction methods for measurement error effects developed outside the context of causal inference methods, we develop valid methods to consistently estimate the causal effect measures with measurement error taken into account. We further develop a linear combination method to improve asymptotic efficiency.

In Chapter 3, we are interested in IPTW estimation of causal parameters of marginal structural models with error-contaminated and time-varying confounders. To correct for measurement error effects, we develop several correction methods. The proposed methods have both the so-called stabilized and unstabilized weighting versions.

In Chapter 4, we focus on causal inference with measurement error in outcomes. When a continuous outcome variable is mismeasured under an additive measurement error model, we find that ignoring measurement error in IPW estimation may still yield a consistent estimator. When the outcome is binary, we derive the asymptotic bias of the IPW estimator and develop several consistent estimation methods with and without extra data sources. With validation data, our proposed estimator is valid and efficient. To provide protection against model misspecification, we further develop a doubly robust estimator which is consistent even when one of the treatment model and the outcome model is misspecified.

In Chapter 5, we study IPW estimation with mismeasured covariates and misclassified outcomes. We develop two estimation methods to correct for measurement error and misclassification effects simultaneously. Our discussion covers a broad scope of treatment models including commonly-used logistic regression models as well as general treatment assignment mechanisms.

In Chapter 6, we investigate causal inference with outcomes subject to both missingness and misclassification. We investigate the impact of ignoring missingness and/or misclassification, and propose consistent methods to correct for missingness and misclassification effects simultaneously. We further propose a doubly robust correction method to provide protection against misspecification.

In Chapter 7, we propose multiply robust methods to (1) pursue more protection against model misspecification than doubly robust estimation, (2) eliminate misclassification effect in outcomes and (3) eliminate missingness effect in outcomes. The proposed estimators are guaranteed to be consistent when either the set of multiple postulated treatment models or the set of multiple postulated outcome models contains a correctly specified model.

To expedite the application of the proposed methods in Chapter 4, we develop an R package for general users. Chapter 8 describes this package in detail.

Finally, the thesis concludes with a discussion in Chapter 9.

Chapter 2

Estimation of Causal Effect Measures

This chapter deals with Problem 1 discussed in Section 1.5. Section 2.1 describes the notation and framework in the absence of measurement error. In Section 2.2 we propose the adaptive conditional score method and the adaptive unbiased estimating equation method to correct for measurement error effects. In addition, we develop a linear combination method with improved efficiency. Section 2.3 includes sensitivity analyses of the data arising from the Framingham Heart Study. Section 2.4 extends the development to accommodating settings with time-dependent treatment.

2.1 Notation and Framework

2.1.1 Setup and Assumptions

For any individual, let T denote the observed binary treatment indicator, taking value 1 if the individual receives the treatment and 0 otherwise. Let Y be the observed binary outcome variable, and Z and X be confounders, where the Z are precisely measured, and the X are error-prone.

We specify the treatment model:

$$\text{logit}\{P(T=1|Z,X)\} = \gamma_0 + \boldsymbol{\gamma}_{\text{Z}}^{\text{T}} Z + \boldsymbol{\gamma}_{\text{X}}^{\text{T}} X, \tag{2.1}$$

where $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}_{\mathrm{Z}}^{\mathrm{T}}, \boldsymbol{\gamma}_{\mathrm{X}}^{\mathrm{T}})^{\mathrm{T}}$ are regression parameters.

Let Y_1 and Y_0 denote the potential outcomes that would have been observed had a subject been treated and untreated, respectively. Thus, $E(Y_1) = P(Y_1 = 1)$ represents the probability of experiencing the outcome event had the entire population received the treatment, and $E(Y_0) = P(Y_0 = 1)$ stands for the probability of experiencing the outcome event had the entire population been untreated.

For the following development, we assume the fundamental causal inference assumptions described in Section 1.1.2 with X replaced by $(X^{\mathrm{T}}, Z^{\mathrm{T}})^{\mathrm{T}}$.

2.1.2 Causal Effect Measures

Three quantities of central interest are the *causal* odds ratio, *causal* risk ratio and the *causal* risk difference, expressed respectively as follows:

$$\psi_{\mathrm{OR}} = \frac{E(Y_1)/\{1 - E(Y_1)\}}{E(Y_0)/\{1 - E(Y_0)\}}, \tag{2.2}$$

$$\psi_{\mathrm{RR}} = \frac{E(Y_1)}{E(Y_0)}, \tag{2.3}$$

and

$$\psi_{\mathrm{RD}} = E(Y_1) - E(Y_0). \tag{2.4}$$

These measures have causal interpretations since they compare the treated and untreated individuals in the same population based on counterfactual outcomes. In contrast, the *associational* odds ratio, *associational* risk ratio and the *associational* risk difference can be formulated from the viewpoint of association measures for observational studies:

$$\phi_{\mathrm{OR}} = \frac{E(Y|T = 1)/\{1 - E(Y|T = 1)\}}{E(Y|T = 0)/\{1 - E\{Y|T = 0)\}},$$

$$\phi_{\mathrm{RR}} = \frac{E(Y|T = 1)}{E(Y|T = 0)},$$

and

$$\phi_{\text{RD}} = E(Y|T=1) - E(Y|T=0).$$

In general, $\phi_{\text{OR}} \neq \psi_{\text{OR}}$, $\phi_{\text{RR}} \neq \psi_{\text{RR}}$ and $\phi_{\text{RD}} \neq \psi_{\text{RD}}$, because the associational odds ratio, associational risk ratio and the associational risk difference compare the differences of the two subpopulations ($T = 1$ versus $T = 0$, or treated versus untreated) of the original population. The imbalance in covariates between the two subpopulations may distort the true relationship between treatment and outcome. Therefore, the associational quantities fail to provide a causal interpretation.

2.1.3 Estimation in the Absence of Measurement Error

To obtain consistent estimators of ψ_{OR}, ψ_{RR} and ψ_{RD}, it suffices to estimate $E(Y_1)$ and $E(Y_0)$ consistently in (2.2), (2.3) and (2.4). To highlight the idea, our discussion is addressed to ψ_{OR} only; the development for ψ_{RR} and ψ_{RD} is the same except that (2.3) and (2.4) will replace (2.2).

Let e be the propensity score which is defined as the conditional probability of treatment assignment given confounders, i.e., by (2.1),

$$e = P(T=1|Z,X) = \frac{1}{1+\exp(-\gamma_0 - \boldsymbol{\gamma}_{\text{Z}}^{\text{T}} Z - \boldsymbol{\gamma}_{\text{X}}^{\text{T}} X)}.$$

Suppose we have a sample of size n. For subject $i = 1, \ldots, n$, let T_i be the observed binary treatment indicator and Y_i be the observed binary outcome variable; let Z_i and X_i be confounders where the Z_i are precisely measured and the X_i are error-prone; and let $Y_{i,1}$ and $Y_{i,0}$ be the potential outcomes that would have been observed had subject i been treated and untreated, respectively. Let $(\hat{\gamma}_0, \hat{\boldsymbol{\gamma}}_{\text{Z}}^{\text{T}}, \hat{\boldsymbol{\gamma}}_{\text{X}}^{\text{T}})^{\text{T}}$ be the maximum likelihood estimates for logistic regression parameters $(\gamma_0, \boldsymbol{\gamma}_{\text{Z}}^{\text{T}}, \boldsymbol{\gamma}_{\text{X}}^{\text{T}})^{\text{T}}$ in (2.1). Then e_i is estimated by

$$\hat{e}_i = \hat{P}(T_i=1|Z_i,X_i) = \frac{1}{1+\exp(-\hat{\gamma}_0 - \hat{\boldsymbol{\gamma}}_{\text{Z}}^{\text{T}} Z_i - \hat{\boldsymbol{\gamma}}_{\text{X}}^{\text{T}} X_i)}, \tag{2.5}$$

Given $\hat{e}_i$, we estimate $E(Y_1)$ and $E(Y_0)$ using the method described by Lunceford and

Davidian (2004) and Hernán and Robins (2016):

$$\hat{E}(Y_1) = \left(\sum_{i=1}^{n} \frac{T_i}{\hat{e}_i}\right)^{-1} \sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}_i} \tag{2.6}$$

and

$$\hat{E}(Y_0) = \left(\sum_{i=1}^{n} \frac{1-T_i}{1-\hat{e}_i}\right)^{-1} \sum_{i=1}^{n} \frac{(1-T_i) Y_i}{1-\hat{e}_i} \tag{2.7}$$

respectively.

Consequently, ψ_{OR} is consistently estimated by using (2.6) and (2.7) in combination with (2.2) and (2.5). Let $\hat{\psi}_{\text{OR}}$ denote the resulting estimator.

To characterize the associated variability, we use the bootstrap resampling technique (Efron, 1982) to obtain variance estimator for $\hat{\psi}_{\text{OR}}$. We resample the data at the individual level with replacement for B times where B is a user-specified number. Each resampled data have sample size equals to the original data. Let $\hat{\psi}_{\text{OR},b}$ denote the resulting estimate for ψ_{OR} obtained from the bth resample, where $b = 1, \ldots, B$. Then the bootstrap-based variance estimate for estimator $\hat{\psi}_{\text{OR}}$ is

$$\widehat{Var}(\hat{\psi}_{\text{OR}}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\psi}_{\text{OR},b} - \frac{\sum_{b=1}^{B} \hat{\psi}_{\text{OR},b}}{B}\right)^2.$$

Let $\hat{\psi}_{\text{OR}}(\alpha)$ be the $(1-\alpha)100\%$ quantile of $\{\hat{\psi}_{\text{OR},b} : b = 1, \ldots, B\}$ where α is a constant between 0 and 1. Then the resulting $(1-\alpha)100\%$ confidence interval for the corresponding parameter is given by $(\hat{\psi}_{\text{OR}}(\alpha/2), \hat{\psi}_{\text{OR}}(1-\alpha/2))$.

2.2 Estimation in the Presence of Measurement Error

2.2.1 Measurement Error Model

The validity of estimators described in Section 2.1 requires the precisely measured X_i. However, when the X_i are subject to measurement error, ignoring this feature and naively

using the procedure of Section 2.1 can lead to seriously biased estimates. To address this issue, suppose the confounders X_i are error-prone, and X_i^* is an observed measurement of X_i. Assume that

$$X_i^* = X_i + \epsilon_i \tag{2.8}$$

for $i = 1, \ldots, n$, where the ϵ_i are independent of each other and of $\{T_i, X_i, Z_i, Y_{i,1}, Y_{i,0}\}$. Assume that the error terms ϵ_i follow $N(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$ with covariance matrix $\mathbf{\Sigma}_\epsilon$. To highlight the key idea, we assume that $\mathbf{\Sigma}_\epsilon$ is known for now. Measurement error model (2.8) characterizes settings where the observed value X_i^* fluctuates around the true confounder value X_i with an error term.

2.2.2 Accommodating Measurement Error Effects

By (2.5), the $\hat{e}_i$ are vulnerable to mismeasurement in X_i, and the estimation method in Section 2.1 needs to be modified to accommodate the measurement error effects. Adapting the development of Stefanski and Carroll (1987) who discussed correcting measurement error effects for parameter estimation under logistic regression models, here we introduce a modified estimator to correct for measurement errors on estimation of the conditional probability e. Specifically, we propose to estimate e_i with

$$\hat{e}_i = \hat{P}(T_i = 1 | Z_i, \hat{\Delta}_i) = \frac{1}{1 + \exp(-\hat{\gamma}_0 - \hat{\boldsymbol{\gamma}}_{\mathrm{Z}}^{\mathrm{T}} Z_i - \hat{\boldsymbol{\gamma}}_{\mathrm{X}}^{\mathrm{T}} \hat{\Delta}_i)}, \tag{2.9}$$

where $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0, \hat{\boldsymbol{\gamma}}_{\mathrm{Z}}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}_{\mathrm{X}}^{\mathrm{T}})^{\mathrm{T}}$ is a consistent estimator of $\boldsymbol{\gamma}$, and $\hat{\Delta}_i = X_i^* + (T_i - 1/2)\mathbf{\Sigma}_\epsilon \hat{\boldsymbol{\gamma}}_{\mathrm{X}}$. Compared to (2.5), $\hat{\Delta}_i$ shows how measurement error is addressed; the naive analysis which disregards the difference between X_i^* and X_i simply replaces X_i with X_i^* in (2.5); the correction method based on (2.9) incorporates the degree of measurement error which pertains to $\mathbf{\Sigma}_\epsilon$. The form of $\hat{\Delta}_i$ also suggests that correction of measurement error effects depends on the treatment status as well as the true covariate effect $\boldsymbol{\gamma}_{\mathrm{X}}$. An equivalent formula to (2.9) was considered by McCaffrey et al. (2013).

Using (2.9) to estimate e_i requires a consistent estimator of $\boldsymbol{\gamma}$. While there is no unique way to obtain a consistent estimator of $\boldsymbol{\gamma}$, here we particularly employ two methods for

this purpose.

Adaptive Conditional Score Method

The first method modifies the conditional score method proposed by Stefanski and Carroll (1987). A consistent estimator of $\boldsymbol{\gamma}$, denoted as $\hat{\boldsymbol{\gamma}}_{\text{SC}}$, is derived by solving

$$\sum_{i=1}^{n} \left\{ T_i - \frac{1}{1+\exp(-\gamma_0 - \boldsymbol{\gamma}_{\text{Z}}^{\text{T}} Z_i - \boldsymbol{\gamma}_{\text{X}}^{\text{T}} \Delta_i)} \right\} \begin{pmatrix} 1 \\ Z_i \\ \Delta_i \end{pmatrix} = \mathbf{0}$$

for γ_0, $\boldsymbol{\gamma}_{\text{Z}}$ and $\boldsymbol{\gamma}_{\text{X}}$, where $\Delta_i = X_i^* + (T_i - 1/2)\boldsymbol{\Sigma}_\epsilon \boldsymbol{\gamma}_{\text{X}}$.

As a result, $\hat{e}_i$ can be calculated by replacing $\hat{\boldsymbol{\gamma}}$ in (2.9) with $\hat{\boldsymbol{\gamma}}_{\text{SC}}$ and substituting the resulting $\hat{e}_i$ into the procedures in Section 2.1 gives the estimator of ψ_{OR}, denoted as $\hat{\psi}_{\text{OR}}^{\text{SC}}$. Its variance estimate can be obtained using the bootstrap resampling method as described in Section 2.1.

Adaptive Unbiased Estimating Equation Method

The second approach modifies the method proposed by Huang and Wang (2001). Their method is to directly construct unbiased estimating functions using the observed surrogate measurement X_i^*. That is, we estimate $\boldsymbol{\gamma}$ by solving

$$(T_i - 1) \begin{pmatrix} 1 \\ Z_i \\ X_i^* \end{pmatrix} + T_i \exp(-\gamma_0 - \boldsymbol{\gamma}_{\text{Z}}^{\text{T}} Z_i - \boldsymbol{\gamma}_{\text{X}}^{\text{T}} X_i^* - \boldsymbol{\gamma}_{\text{X}}^{\text{T}} \boldsymbol{\Sigma}_\epsilon \boldsymbol{\gamma}_{\text{X}}/2) \begin{pmatrix} 1 \\ Z_i \\ X_i^* + \boldsymbol{\Sigma}_\epsilon \boldsymbol{\gamma}_{\text{X}} \end{pmatrix} = \mathbf{0}$$

for γ_0, $\boldsymbol{\gamma}_{\text{Z}}$ and $\boldsymbol{\gamma}_{\text{X}}$. Let $\hat{\boldsymbol{\gamma}}_{\text{HW}}$ denote the resulting estimator of $\boldsymbol{\gamma}$, which is consistent under regularity conditions.

Similarly, replacing $\hat{\boldsymbol{\gamma}}$ in (2.9) with $\hat{\boldsymbol{\gamma}}_{\text{HW}}$ and substituting the resulting $\hat{e}_i$ into the procedures in Section 2.1 gives the estimator of ψ_{OR}, denoted as $\hat{\psi}_{\text{OR}}^{\text{HW}}$. Its variance estimate

of the estimator can be obtained using the bootstrap resampling method as described in Section 2.1.

Linear Combination of Consistent Estimators

Although $\hat{\psi}_{\rm OR}^{\rm SC}$ and $\hat{\psi}_{\rm OR}^{\rm HW}$ are consistent estimators of $\psi_{\rm OR}$, their efficiency may differ. We want to construct a new estimator with higher efficiency than both $\hat{\psi}_{\rm OR}^{\rm SC}$ and $\hat{\psi}_{\rm OR}^{\rm HW}$, in addition to being consistent. To do this, we consider linear combinations of $\hat{\psi}_{\rm OR}^{\rm SC}$ and $\hat{\psi}_{\rm OR}^{\rm HW}$ and then identify an estimator with the smallest variance (e.g., Yi and He, 2006).

For $0 \leq c \leq 1$, consider the linear combination:

$$\hat{\psi}_{\rm OR}^{\rm comb}(c) = c\hat{\psi}_{\rm OR}^{\rm SC} + (1-c)\hat{\psi}_{\rm OR}^{\rm HW}. \tag{2.10}$$

Noting that

$$\begin{aligned} Var\{\hat{\psi}_{\rm OR}^{\rm comb}(c)\} = &\{Var(\hat{\psi}_{\rm OR}^{\rm SC}) + Var(\hat{\psi}_{\rm OR}^{\rm HW}) - 2Cov(\hat{\psi}_{\rm OR}^{\rm SC}, \hat{\psi}_{\rm OR}^{\rm HW})\}c^2 \\ &-\{2Var(\hat{\psi}_{\rm OR}^{\rm HW}) - 2Cov(\hat{\psi}_{\rm OR}^{\rm SC}, \hat{\psi}_{\rm OR}^{\rm HW})\}c + Var(\hat{\psi}_{\rm OR}^{\rm HW}), \end{aligned}$$

we minimize $Var\{\hat{\psi}_{\rm OR}^{\rm comb}(c)\}$ with respect to c, and let $c_{\rm opt}$ be the corresponding value of c, given by

$$c_{\rm opt} = \frac{Var(\hat{\psi}_{\rm OR}^{\rm HW}) - Cov(\hat{\psi}_{\rm OR}^{\rm SC}, \hat{\psi}_{\rm OR}^{\rm HW})}{Var(\hat{\psi}_{\rm OR}^{\rm SC}) + Var(\hat{\psi}_{\rm OR}^{\rm HW}) - 2Cov(\hat{\psi}_{\rm OR}^{\rm SC}, \hat{\psi}_{\rm OR}^{\rm HW})}.$$

Therefore, the resultant estimator $\hat{\psi}_{\rm OR}^{\rm opt} = c_{\rm opt}\hat{\psi}_{\rm OR}^{\rm SC} + (1-c_{\rm opt})\hat{\psi}_{\rm OR}^{\rm HW}$ is the optimal estimator among the class of estimators of form (2.10).

To calculate the estimator $\hat{\psi}_{\rm OR}^{\rm opt}$, we need to determine the coefficient $c_{\rm opt}$. For this purpose, we design a bootstrap-based procedure. Specifically, we resample the data with replacement for B times where B is a user-specified number. Let $\hat{\psi}_{{\rm OR},b}^{\rm SC}$ and $\hat{\psi}_{{\rm OR},b}^{\rm HW}$ denote the resulting estimates of estimators $\hat{\psi}_{\rm OR}^{\rm SC}$ and $\hat{\psi}_{\rm OR}^{\rm HW}$ using the bth resample for $b = 1, \ldots, B$. Calculate

$$\widehat{Var}(\hat{\psi}_{\rm OR}^{\rm SC}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\psi}_{{\rm OR},b}^{\rm SC} - \frac{\sum_{b=1}^{B}\hat{\psi}_{{\rm OR},b}^{\rm SC}}{B}\right)^2,$$

$$\widehat{Var}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\psi}_{\mathrm{OR},b}^{\mathrm{HW}} - \frac{\sum_{b=1}^{B}\hat{\psi}_{\mathrm{OR},b}^{\mathrm{HW}}}{B}\right)^2,$$

and

$$\widehat{Cov}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}, \hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\psi}_{\mathrm{OR},b}^{\mathrm{SC}} - \frac{\sum_{b=1}^{B}\hat{\psi}_{\mathrm{OR},b}^{\mathrm{SC}}}{B}\right)\left(\hat{\psi}_{\mathrm{OR},b}^{\mathrm{HW}} - \frac{\sum_{b=1}^{B}\hat{\psi}_{\mathrm{OR},b}^{\mathrm{HW}}}{B}\right).$$

Then we approximate c_{opt} by

$$\hat{c}_{\mathrm{opt}} = \frac{\widehat{Var}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}) - \widehat{Cov}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}, \hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}})}{\widehat{Var}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}) + \widehat{Var}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}) - 2\widehat{Cov}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}, \hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}})}.$$

In implementing this procedure, we need to ensure the resulting $\hat{c}_{\mathrm{opt}}$ is reasonable, which means two conditions must be satisfied. Firstly, $\hat{c}_{\mathrm{opt}}$ must lie between 0 and 1. Secondly, $\widehat{Var}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}) + \widehat{Var}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}) - 2\widehat{Cov}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}, \hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}) \geq 0$ since

$$\begin{aligned}
& Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}) + Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}) - 2Cov(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}, \hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}) \\
= \;& Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}) + Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}) - 2\sqrt{Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}})Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}})}Cor(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}, \hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}) \\
\geq \;& Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}) + Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}) - 2\sqrt{Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}})Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}})} \\
\geq \;& 2\sqrt{Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}})Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}})} - 2\sqrt{Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}})Var(\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}})} \\
= \;& 0.
\end{aligned}$$

If one of these two conditions is not satisfied, we set $\hat{c}_{\mathrm{opt}}$ to be either 0 or 1. That is, we take $\hat{\psi}_{\mathrm{OR}}^{\mathrm{opt}}$ to be $\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}$ if $\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}$ has a smaller variance than $\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}$, or take $\hat{\psi}_{\mathrm{OR}}^{\mathrm{opt}}$ to be $\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}$ if $\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}$ has a smaller variance than $\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}$.

The variance estimate for $\hat{\psi}_{\mathrm{OR}}^{\mathrm{opt}}$ is obtained by

$$\widehat{Var}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{opt}}) = \hat{c}_{\mathrm{opt}}^2\widehat{Var}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}) + (1-\hat{c}_{\mathrm{opt}})^2\widehat{Var}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}) + 2\hat{c}_{\mathrm{opt}}(1-\hat{c}_{\mathrm{opt}})\widehat{Cov}(\hat{\psi}_{\mathrm{OR}}^{\mathrm{SC}}, \hat{\psi}_{\mathrm{OR}}^{\mathrm{HW}}).$$

To obtain a confidence interval for ψ_{OR}, we consider bootstrap-based estimates $\mathcal{E} = \{\hat{c}_{\mathrm{opt}}\hat{\psi}_{\mathrm{OR},b}^{\mathrm{SC}} + (1-\hat{c}_{\mathrm{opt}})\hat{\psi}_{\mathrm{OR},b}^{\mathrm{HW}} : b = 1, \ldots, B\}$. Let $\hat{\psi}_{\mathrm{OR}}^{\mathrm{opt}}(\alpha)$ be the $(1-\alpha)100\%$ quantile of $\mathcal{E}$, where α is a

constant between 0 and 1. Then an approximate $(1-\alpha)100\%$ confidence interval of ψ_{OR} is given by $(\hat{\psi}^{\mathrm{opt}}_{\mathrm{OR}}(\alpha/2), \hat{\psi}^{\mathrm{opt}}_{\mathrm{OR}}(1-\alpha/2))$.

2.3 Empirical Example: Framingham Heart Study

To illustrate the proposed methods, we conduct sensitivity analyses of the data arising from the Framingham Heart Study (e.g., Carroll et al., 2006). The data set consists of records of 1615 men aged 31 to 65. The observed outcome Y is the occurrence of coronary heart disease within an eight-year period following Exam 3. Our objective is to estimate the causal effects of T, the smoking status at Exam 1, on the occurrence of coronary heart disease. Confounders considered here are age, serum cholesterol (CHOL) and systolic blood pressure (SBP), where SBP and CHOL are error-prone since the true values of long-term average measurements are unobserved. Let Z denote age. According to Carroll et al. (2006), let $X = (X_1, X_2)^{\mathrm{T}}$, where X_1 denote transformed long-term average SBP, $X_1 = \log(\mathrm{SBP} - 50)$, and let X_1^* be a transformed observed measurement of SBP. Let X_2 denote transformed long-term average CHOL, $X_2 = \log(\mathrm{CHOL})$, and let X_2^* be a transformed observed measurement of CHOL. Such transformation strategies were originally discussed by Cornfield (1962) and then applied by Carroll et al. (1984); the purpose was to make the transformed observations distributed reasonably close to normal distributions. We define $X^* = (X_1^*, X_2^*)^{\mathrm{T}}$ and assume that X^* and X are modeled by (2.8). We apply the proposed methods and naive analysis to the data with $\mathbf{\Sigma}_\epsilon$ set as $V = \begin{pmatrix} 0.0126 & 0.000673 \\ 0.000673 & 0.00846 \end{pmatrix}$, an estimate obtained by (Carroll et al., 2006, page 118). In using the bootstrap algorithm for variance estimates, we take $B = 1000$.

Table 2.1 summarizes the results. All the methods indicate statistically significant causal effects of smoking on the occurrence of coronary heart disease at the nominal level 0.05. Estimates obtained from the proposed methods are slightly larger than the naive estimates.

We further perform sensitivity analyses to evaluate how sensitive estimation of causal effect measures is to different values of $\mathbf{\Sigma}_\epsilon$. We consider four scenarios which assume the

Table 2.1: Analysis results for the estimation of causal effects of smoking on the occurrence of coronary heart disease in the presence of measurement error in systolic blood pressure and serum cholesterol: estimate (EST), bootstrap standard error (SE) and 95% confidence interval (CI)

Method	Measure	EST	SE	95% CI
naive	ψ_{OR}	1.754	0.468	(1.138, 2.949)
	ψ_{RR}	1.688	0.425	(1.127, 2.746)
	ψ_{RD}	0.036	0.013	(0.009, 0.060)
ACS	ψ_{OR}	1.763	0.473	(1.141, 2.997)
	ψ_{RR}	1.696	0.429	(1.129, 2.814)
	ψ_{RD}	0.036	0.013	(0.009, 0.060)
AUEE	ψ_{OR}	1.769	0.473	(1.143, 3.001)
	ψ_{RR}	1.702	0.429	(1.132, 2.811)
	ψ_{RD}	0.036	0.013	(0.009, 0.060)
LC	ψ_{OR}	1.763	0.473	(1.141, 2.997)
	ψ_{RR}	1.696	0.428	(1.129, 2.814)
	ψ_{RD}	0.036	0.013	(0.009, 0.060)

ACS: adaptive conditional score method; AUEE: adaptive unbiased estimating function method; LC: linear combination estimator based on ACS and AUEE.

same form (2.8) for the measurement error model but the covariance matrix $\mathbf{\Sigma}_\epsilon$ is specified differently. We take $\mathbf{\Sigma}_\epsilon = \delta M$ where δ is set as a value in $[0, 3]$ and M is a 2×2 matrix. In Scenario 1, we set $M = V$; this scenario retains the correlation structure of V but alters the magnitude. In Scenario 2, we take $M = \begin{pmatrix} 0.0126 & 0 \\ 0 & 0 \end{pmatrix}$; this scenario explores situations where only measurement error in SBP is incorporated in the analysis. In Scenario 3, we take $M = \begin{pmatrix} 0 & 0 \\ 0 & 0.00846 \end{pmatrix}$; this scenario facilitates situations where only CHOL is treated as error-prone. In Scenario 4, we take $M = \begin{pmatrix} 0.0126 & 0 \\ 0 & 0.00846 \end{pmatrix}$; this scenario describes that both SBP and CHOL as error-prone but assumes $\mathbf{\Sigma}_\epsilon$ to be diagonal, i.e., measurement error in SBP and CHOL are treated independent.

The specification of these four scenarios is driven by the following consideration. By (2.8), $Var(X^*) = Var(X) + \mathbf{\Sigma}_\epsilon$ and hence $Var(X_i^*) \geq \mathbf{\Sigma}_\epsilon^{ii}$ where $i = 1, 2$ and $\mathbf{\Sigma}_\epsilon^{ii}$ is the (i, i) element in $\mathbf{\Sigma}_\epsilon$. The range of δ (i.e., from 0 to 3) is reasonably broad to reflect plausible scenarios in sensitivity analyses; estimates of the reliability of X_i, defined as $Var(X_i)/Var(X_i^*) = (Var(X_i^*) - \mathbf{\Sigma}_\epsilon^{ii})/Var(X_i^*)$, change from 0.28 to 1 for X_1 and from 0.20 to 1 for X_2, where the empirical estimate of $Var(X^*)$ is $\widehat{Var}(X^*) = \begin{pmatrix} 0.0525 & 0.00406 \\ 0.00406 & 0.0316 \end{pmatrix}$.

We display the results in Figures 2.1 and 2.2 for the estimated odds ratio, risk ratio and risk difference obtained from the proposed methods. Figure 2.1 displays the results for Scenarios 1 and 4, where both SBP and CHOL are treated as error-prone. Figure 2.2 shows the results for Scenarios 2 and 3, where only one of SBP and CHOL is regarded as error-prone. Figure 2.1 shows that as δ increases, the estimates of causal effects first become larger and then smaller. Figure 2.2 shows that the estimates of causal effects become larger as the measurement error in SBP increases, assuming that CHOL is error-free. The estimates of causal effects become smaller as the measurement error in CHOL increases, assuming that SBP is error-free. Interestingly, the measurement error in SBP and CHOL has opposite effects on estimation of causal measures. As a result, when both SBP and CHOL are assumed to be subject to error, the effects of measurement error in

SBP and CHOL may interplay and are difficult to be visualized. Although Figures 2.1 and 2.2 present different patterns that are related to different scenarios of measurement error, the magnitudes of the estimates are fairly similar, which suggests that our conclusion on the causal effects of smoking on the occurrence of coronary heart disease is reasonably robust to different kinds of measurement error.

2.4 Extensions to Time-Dependent Treatment

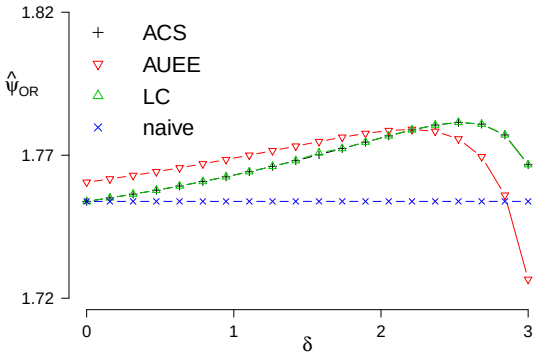
In this section we generalize the proposed methods in the previous sections to the setting with a time-dependent treatment.

2.4.1 Notation and Setting

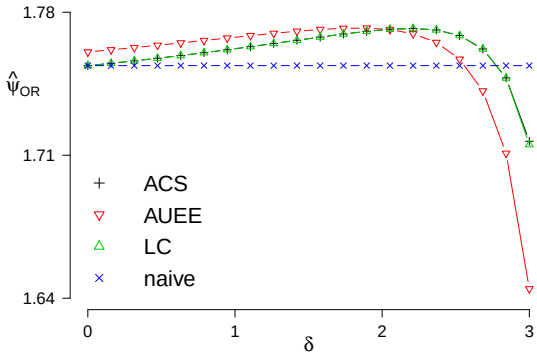
Consider K follow up visits. Let $a(k)$ be the potential binary treatment indicator at visit k, where $k = 0, 1, \ldots, K$; and $\bar{a}(k) = \{a(u) : 0 \leq u \leq k, u \text{ is an integer}\}$ is the potential treatment history up to and including visit k. Let $\bar{a} = \bar{a}(K)$, and $Y_{\bar{a}}$ be the potential outcome that would have been observed had the subject experienced treatment history $\bar{a}$.

For subject i, let $Y_{i,\bar{a}}$ denote the potential outcome that would have been observed had this subject experienced the treatment history $\bar{a}$. Let $A_i(k)$ and $\bar{A}_i(k) = \{A_i(u) : 0 \leq u \leq k, u \text{ is an integer}\}$ represent the actually observed treatment at visit k and the actually observed treatment history up to and including visit k, respectively; and we write $\bar{A}_i = \bar{A}_i(K)$. Define $A_i(-1) = 0$. Let $(Z_i^{\mathrm{T}}(k), X_i^{\mathrm{T}}(k))^{\mathrm{T}}$ be the vector of time-dependent confounders at visit k, where the $Z_i(k)$ are precisely measured, and the $X_i(k)$ are subject to measurement error. Let $\bar{Z}_i(k) = \{Z_i(u) : 0 \leq u \leq k, u \text{ is an integer}\}$ and $\bar{X}_i(k) = \{X_i(u) : 0 \leq u \leq k, u \text{ is an integer}\}$ be the confounder histories up to and including visit k; $\bar{Z}_i = \bar{Z}_i(K)$ and $\bar{X}_i = \bar{X}_i(K)$. Let Y_i be the binary outcome for subject i measured after visit K. Confounders are assumed to be measured before treatment.

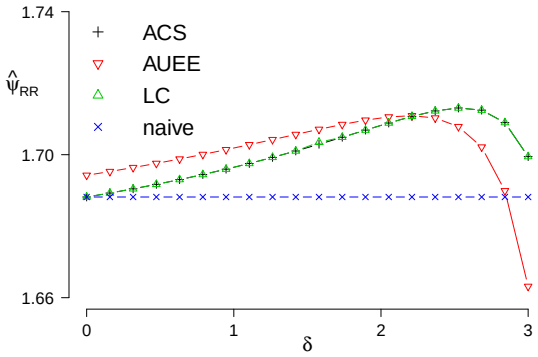
Analogous to the assumptions in time-independent settings, we make the following assumptions for all k.



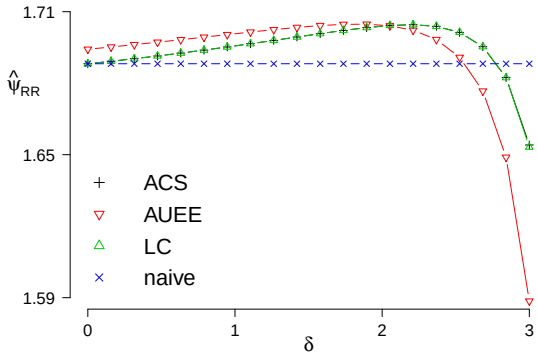
$\hat{\psi}_{\mathrm{OR}}$ resulted from various $\boldsymbol{\Sigma}_\epsilon$ for Scenario 1



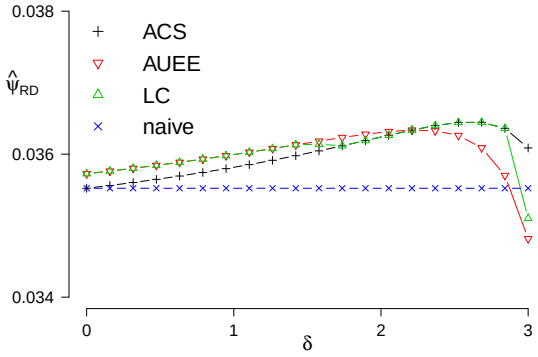
$\hat{\psi}_{\mathrm{OR}}$ resulted from various $\boldsymbol{\Sigma}_\epsilon$ for Scenario 4



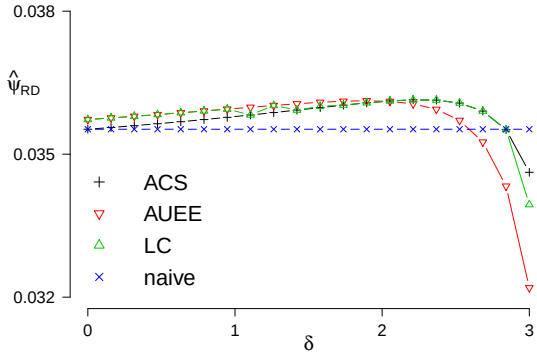
$\hat{\psi}_{\mathrm{RR}}$ resulted from various $\boldsymbol{\Sigma}_\epsilon$ for Scenario 1



$\hat{\psi}_{\mathrm{RR}}$ resulted from various $\boldsymbol{\Sigma}_\epsilon$ for Scenario 4

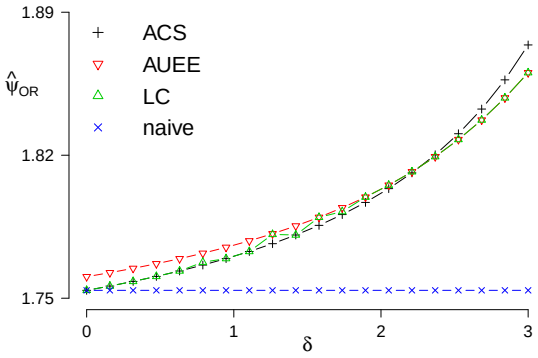


$\hat{\psi}_{\mathrm{RD}}$ resulted from various $\boldsymbol{\Sigma}_\epsilon$ for Scenario 1

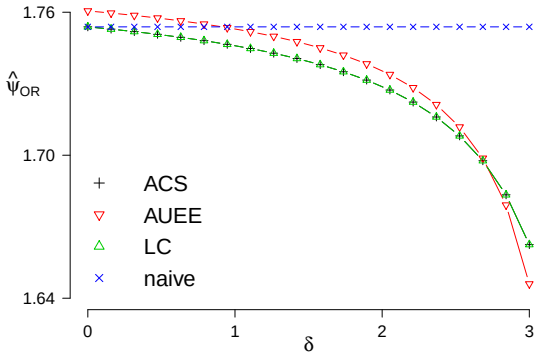


$\hat{\psi}_{\mathrm{RD}}$ resulted from various $\boldsymbol{\Sigma}_\epsilon$ for Scenario 4

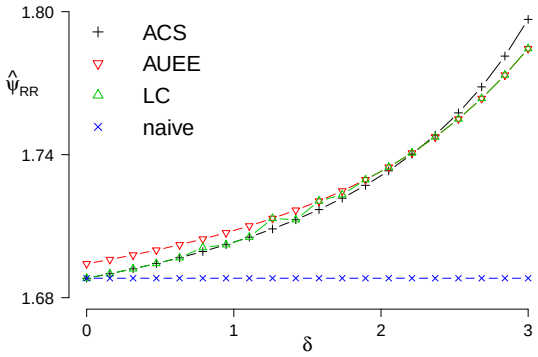
Figure 2.1: Estimated causal odds ratio, causal risk ratio and causal risk difference assuming measurement error in both SBP and CHOL: Scenario 1 and Scenario 4



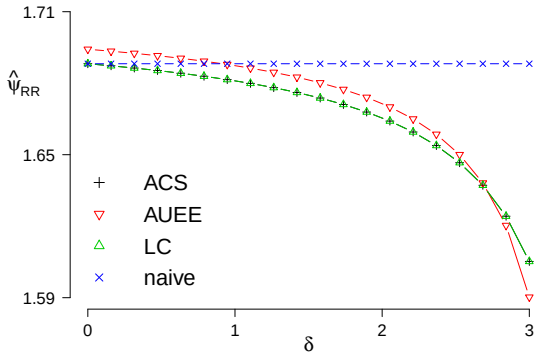
$\hat{\psi}_{\mathrm{OR}}$ resulted from various $\boldsymbol{\Sigma}_\epsilon$ for Scenario 2



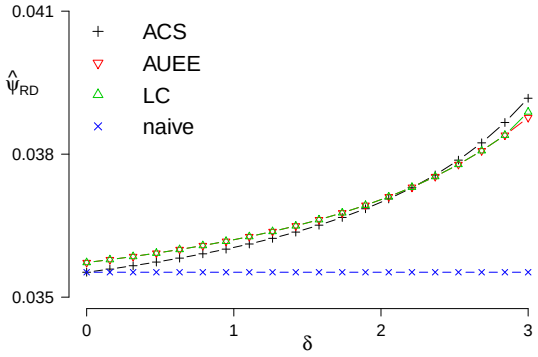
$\hat{\psi}_{\mathrm{OR}}$ resulted from various $\boldsymbol{\Sigma}_\epsilon$ for Scenario 3



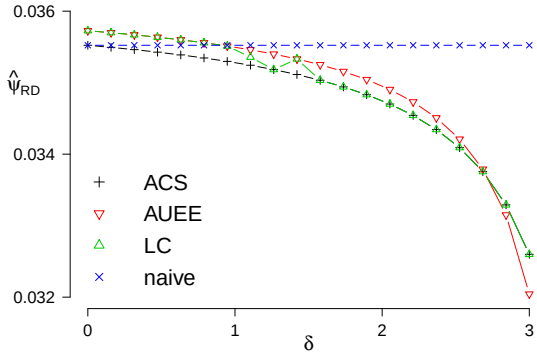
$\hat{\psi}_{\mathrm{RR}}$ resulted from various $\boldsymbol{\Sigma}_\epsilon$ for Scenario 2



$\hat{\psi}_{\mathrm{RR}}$ resulted from various $\boldsymbol{\Sigma}_\epsilon$ for Scenario 3



$\hat{\psi}_{\mathrm{RD}}$ resulted from various $\boldsymbol{\Sigma}_\epsilon$ for Scenario 2



$\hat{\psi}_{\mathrm{RD}}$ resulted from various $\boldsymbol{\Sigma}_\epsilon$ for Scenario 3

Figure 2.2: Estimated causal odds ratio, causal risk ratio and causal risk difference assuming only one of SBP and CHOL is error-prone: Scenario 2 and Scenario 3

Assumption 1 (No Interference): the treatment taken by subject j has no effect on the potential outcomes of subject i for all $i \neq j$.

Assumption 2 (Consistency): $Y_{\bar{A}} = Y$.

Assumption 3 (No Unmeasured Confounding): $P(\bar{A}|\bar{Z}, \bar{X}, Y_{\bar{a}}) = P(\bar{A}|\bar{Z}, \bar{X})$.

Assumption 4 (Positivity): $0 < P\{A(k) = 1|\bar{A}(k-1), \bar{Z}(k), \bar{X}(k)\} < 1$.

In addition, we assume

Assumption 5 (Markov Assumption): $P(\bar{A}|\bar{Z}, \bar{X}) = \prod_{k=0}^{K} P\{A(k)|\bar{A}(k-1), Z(k), X(k)\}$.

Assumption 5 is reasonable when the previous confounders have no effect on the treatment assignment, given the previous treatments and current confounders.

With any two potential treatment histories $\bar{a}_1$ and $\bar{a}_0$, the causal odds ratio, causal risk ratio and the causal risk difference are given by

$$\psi_{\text{OR}} = \frac{E(Y_{\bar{a}_1})/\{1 - E(Y_{\bar{a}_1})\}}{E(Y_{\bar{a}_0})/\{1 - E(Y_{\bar{a}_0})\}},$$

$$\psi_{\text{RR}} = \frac{E(Y_{\bar{a}_1})}{E(Y_{\bar{a}_0})},$$

and

$$\psi_{\text{RD}} = E(Y_{\bar{a}_1}) - E(Y_{\bar{a}_0}),$$

respectively.

2.4.2 Estimation with Measurement Error Effects Accommodated

Suppose the confounders $X_i(k)$ are subject to measurement error, and X_{ik}^* is the observed version of $X_i(k)$. Assume that

$$X_{ik}^* = X_i(k) + \epsilon_{ik} \tag{2.11}$$

for $i = 1, \ldots, n$ and $k = 0, \ldots, K$, where the ϵ_{ik} are independent of each other, and of $\{A_i(k), X_i(k), Z_i(k) : k = 0, 1, \ldots, K\}$ and $Y_{i,\bar{a}}$. Assume that the error terms ϵ_{ik} follow

$N(\mathbf{0}, \mathbf{\Sigma}_{\epsilon k})$, with covariance matrix $\mathbf{\Sigma}_{\epsilon k}$. Again, to highlight the key idea, we assume that $\mathbf{\Sigma}_{\epsilon k}$ is known.

Suppose that the treatment indicators are modeled by:

$$\text{logit}[P\{A(k) = 1|\bar{A}(k-1), Z(k), X(k)\}] = \gamma_{0k} + \boldsymbol{\gamma}_{\text{A}k}^{\text{T}} \bar{A}(k-1) + \boldsymbol{\gamma}_{\text{Z}k}^{\text{T}} Z(k) + \boldsymbol{\gamma}_{\text{X}k}^{\text{T}} X(k), \quad (2.12)$$

where $\boldsymbol{\gamma}_k = (\gamma_{0k}, \boldsymbol{\gamma}_{\text{A}k}^{\text{T}}, \boldsymbol{\gamma}_{\text{Z}k}^{\text{T}}, \boldsymbol{\gamma}_{\text{X}k}^{\text{T}})^{\text{T}}$ is a vector of regression parameters for $k = 0, 1, \ldots, K$.

To consistently estimate ψ_{OR}, ψ_{RR} and ψ_{RD}, it suffices to consistently estimate $E(Y_{\bar{a}_1})$ and $E(Y_{\bar{a}_0})$. The following theorem establishes the consistency of the proposed estimators, whose proof is included in Appendix A.

Theorem 2.1. *Suppose Assumptions 1 to 5, (2.11) and (2.12) hold. Then the causal mean $E(Y_{\bar{a}})$ under treatment history $\bar{a}$ can be consistently estimated by*

$$\hat{E}(Y_{\bar{a}}) = \frac{\sum_{i=1}^{n} \hat{w}_i Y_i I(\bar{A}_i = \bar{a})}{\sum_{i=1}^{n} \hat{w}_i I(\bar{A}_i = \bar{a})},$$

where $I(\cdot)$ is the indicator function,

$$\hat{w}_i = \prod_{k=0}^{K} \left(1 + \exp[\{-\hat{\gamma}_{0k} - \hat{\boldsymbol{\gamma}}_{\text{A}k}^{\text{T}} \bar{A}_i(k-1) - \hat{\boldsymbol{\gamma}}_{\text{Z}k}^{\text{T}} Z_i(k) - \hat{\boldsymbol{\gamma}}_{\text{X}k}^{\text{T}} \hat{\Delta}_i(k)\} \cdot \{2A_i(k) - 1\}]\right),$$

and $\hat{\Delta}_i(k) = X_{ik}^* + \{A_i(k) - 1/2\}\mathbf{\Sigma}_{\epsilon k}\hat{\boldsymbol{\gamma}}_{\text{X}k}$ with $(\hat{\gamma}_{0k}, \hat{\boldsymbol{\gamma}}_{\text{A}k}^{\text{T}}, \hat{\boldsymbol{\gamma}}_{\text{Z}k}^{\text{T}}, \hat{\boldsymbol{\gamma}}_{\text{X}k}^{\text{T}})^{\text{T}}$ being a consistent estimator of $(\gamma_{0k}, \boldsymbol{\gamma}_{\text{A}k}^{\text{T}}, \boldsymbol{\gamma}_{\text{Z}k}^{\text{T}}, \boldsymbol{\gamma}_{\text{X}k}^{\text{T}})^{\text{T}}$.

We note that setting $K = 0$ shows that the resulting estimators $\hat{E}(Y_1)$ and $\hat{E}(Y_0)$ can be expressed by (2.6) and (2.7), respectively, with $\hat{e}_i$ given by (2.9). Therefore, the proposed estimator based on (2.9) is a special case of Theorem 2.1. Finally, estimation of γ_k $(k = 0, \ldots, K)$ may be performed using the conditional score method and the unbiased estimating equation method described in Section 2.2.2. Estimators of ψ_{OR}, ψ_{RR} and ψ_{RD} can be obtained in an analogous manner of the previous sections.

2.4.3 Simulation Studies

We conduct simulation studies to assess the performance of the proposed methods described in Section 2.2.2: the adaptive conditional score (ACS) method, the adaptive unbiased estimating equations (AUEE) method, and the linear combination (LC) method. For comparison purposes, we also consider the naive analysis which ignores the measurement error effects, and the benchmark method which uses the generated measurements of X_i rather than surrogate values X_i^*. These five methods are displayed under the headings "ACS", "AUEE", "LC", "naive" and "benchmark" in the following tables.

In Setting 1, consider two time points. For the ith subject, at visit 0, the confounder $X_i(0)$ is a scalar and is generated from $N(0,1)$ and the treatment indicator $A_i(0)$ is drawn from a Bernoulli distribution with probability $1/[1+\exp\{-\gamma_{\text{X0}}X_i(0)\}]$. At visit 1, the confounder $X_i(1)$ is generated from $N(X_i(0)+A_i(0)-0.5,1)$, and the treatment indicator $A_i(1)$ is drawn from a Bernoulli distribution with probability $1/[1+\exp\{-\gamma_{\text{A1}}A_i(0)-\gamma_{\text{X1}}X_i(1)\}]$. Finally, the outcome Y_i is generated from a Bernoulli distribution with probability $1/[1+\exp\{0.5+X_i(1)-A_i(0)-A_i(1)\}]$.

Set $\gamma_{\text{X0}}=0.3$, $\gamma_{\text{A1}}=0.2$ and $\gamma_{\text{X1}}=0.3$. The measurement error model is taken as (2.11) where $\boldsymbol{\Sigma}_{\epsilon k}$ becomes a scalar, now denoted $\sigma^2_{\epsilon k}$; we set $\sigma_{\epsilon k}$ to be 0.5 or 1 for all k to feature different degrees of measurement error. Sample sizes $n=1000$ and $n=5000$ are considered, and 5000 simulations are run for each parameter configuration. The number of bootstrap replicates is set to $B=1000$.

We are interested in estimating the causal odds ratio, causal risk ratio and the causal risk difference regarding two potential treatment plans $\bar{a}_1=\{1,1\}$ and $\bar{a}_0=\{0,0\}$.

The average relative bias in percent (ReBias%), average bootstrap-based standard errors (ASE), empirical standard error (ESE) and the coverage percentage (CP) of 95% confidence intervals are reported, where the relative bias is calculated as the bias divided by the true value.

Table 2.2 reports the results. The naive analysis produces severe bias and its performance becomes worse as the magnitude of measurement error increases. The proposed methods all present satisfactory results in terms of bias, although the estimated causal odds

ratio displays somewhat noticeable finite sample bias when the sample size is 1000. Unsurprisingly, variance estimates yielded from the proposed methods are larger than those produced from the benchmark method. The AUEE method seems to produce slightly larger variances than the ACS method does. The linear combination method shows improved variance estimates although not substantial. Compared to the ACS method, the linear combination method has smaller ASEs and sometimes larger ESEs.

In Setting 2, we consider two mismeasured confounders, i.e., vector $X_i(k)$ has two elements. For the ith subject, at visit 0, the confounders $(Z_i(0), X_i^{\mathrm{T}}(0))^{\mathrm{T}}$ are generated from a multivariate normal distribution with mean $(0,0,0)^{\mathrm{T}}$ and covariance matrix Σ_{c}, where

$$\Sigma_{\mathrm{c}} = \begin{pmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.5 \\ 0.2 & 0.5 & 1 \end{pmatrix},$$

and the treatment indicator $A_i(0)$ is generated from a Bernoulli distribution with probability $1/[1+\exp\{-\gamma_{\mathrm{z0}} Z_i(0) - \gamma_{\mathrm{x0}}^{\mathrm{T}} X_i(0)\}]$. At visit 1, the confounders $(Z_i(1), X_i^{\mathrm{T}}(1))^{\mathrm{T}}$ are generated from a multivariate normal distribution with mean $(Z_i(0), X_i^{\mathrm{T}}(0))^{\mathrm{T}} + (A_i(0) - 0.5, A_i(0) - 0.5, A_i(0) - 0.5)^{\mathrm{T}}$ and covariance matrix Σ_{c}, and the treatment indicator $A_i(1)$ is drawn from a Bernoulli distribution with probability $1/[1+\exp\{-\gamma_{\mathrm{a1}} A_i(0) - \gamma_{\mathrm{z1}} Z_i(1) - \gamma_{\mathrm{x1}}^{\mathrm{T}} X_i(1)\}]$. The outcome Y_i is generated from a Bernoulli distribution with probability $1/[1+\exp\{0.5 + (1,1,1)(Z_i(1), X_i^{\mathrm{T}}(1))^{\mathrm{T}} - A_i(0) - A_i(1)\}]$.

Set $\gamma_{\mathrm{z0}} = 0.1$, $\gamma_{\mathrm{x0}} = (0.2, 0.3)^{\mathrm{T}}$, $\gamma_{\mathrm{a1}} = 0.2$, $\gamma_{\mathrm{z1}} = 0.1$, and $\gamma_{\mathrm{x1}} = (0.3, 0.2)^{\mathrm{T}}$. The measurement error model (2.11) is assumed with $\mathbf{\Sigma}_{\epsilon 0} = V_0$, $\mathbf{\Sigma}_{\epsilon 1} = V_1$, or with $\mathbf{\Sigma}_{\epsilon 0} = 4V_0$, $\mathbf{\Sigma}_{\epsilon 1} = 4V_1$, where

$$V_0 = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix} \quad \text{and} \quad V_1 = \begin{pmatrix} 0.15 & 0.05 \\ 0.05 & 0.15 \end{pmatrix}.$$

Unlike Setting 1, Setting 2 includes both precisely measured confounders $(Z_i(0), Z_i(1))$ as well as error-prone confounders $(X_i^{\mathrm{T}}(0), X_i^{\mathrm{T}}(1))$ which are allowed to be correlated with each other. Sample sizes $n = 2000$ and $n = 5000$ are respectively considered, and 5000 simulations are run for each parameter configuration. Set the number of bootstrap repli-

Table 2.2: Simulation results for Setting 1 with the correctly-specified measurement error model: average relative bias in percent (ReBias%), average standard error (ASE), empirical standard error (ESE) and coverage percentage (CP)

			$\sigma_{\epsilon k} = 0.5$				$\sigma_{\epsilon k} = 1$			
n	Method	Measure	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
1000	naive	ψ_{OR}	-6.936	0.321	0.318	91.4	-19.728	0.282	0.273	71.1
		ψ_{RR}	-3.266	0.090	0.090	91.3	-8.983	0.085	0.083	71.0
		ψ_{RD}	-11.778	0.038	0.038	91.2	-31.779	0.039	0.039	70.4
	benchmark	ψ_{OR}	1.371	0.346	0.342	95.2	1.161	0.344	0.332	95.1
		ψ_{RR}	0.241	0.093	0.093	95.3	0.200	0.093	0.091	95.2
		ψ_{RD}	-0.245	0.038	0.038	95.3	-0.424	0.038	0.037	95.2
	ACS	ψ_{OR}	1.466	0.356	0.352	94.9	1.658	0.391	0.373	95.6
		ψ_{RR}	0.265	0.096	0.095	95.2	0.352	0.105	0.102	95.6
		ψ_{RD}	-0.220	0.039	0.039	95.1	-0.194	0.042	0.041	95.6
	AUEE	ψ_{OR}	1.943	0.371	0.365	94.7	2.916	0.462	0.402	95.5
		ψ_{RR}	0.459	0.099	0.098	95.1	0.873	0.122	0.109	95.3
		ψ_{RD}	0.303	0.040	0.040	95.1	1.233	0.045	0.043	95.6
	LC	ψ_{OR}	0.885	0.355	0.352	94.7	0.943	0.388	0.373	95.7
		ψ_{RR}	0.113	0.096	0.095	95.1	0.137	0.104	0.102	95.5
		ψ_{RD}	-0.276	0.039	0.039	95.0	-0.525	0.042	0.041	95.6
5000	naive	ψ_{OR}	-7.950	0.138	0.138	77.8	-20.492	0.122	0.121	11.1
		ψ_{RR}	-3.434	0.040	0.040	78.2	-9.117	0.038	0.037	12.4
		ψ_{RD}	-11.561	0.017	0.017	77.7	-31.457	0.018	0.017	11.3
	benchmark	ψ_{OR}	0.264	0.149	0.149	94.8	0.232	0.149	0.147	94.8
		ψ_{RR}	0.058	0.041	0.041	94.5	0.041	0.041	0.041	95.0
		ψ_{RD}	-0.036	0.017	0.017	94.8	-0.076	0.017	0.017	95.0
	ACS	ψ_{OR}	0.285	0.153	0.153	94.8	0.311	0.165	0.163	94.8
		ψ_{RR}	0.065	0.042	0.042	94.7	0.061	0.045	0.045	95.0
		ψ_{RD}	-0.026	0.017	0.017	94.9	-0.054	0.019	0.018	94.9
	AUEE	ψ_{OR}	0.364	0.157	0.157	94.6	0.505	0.172	0.169	95.0
		ψ_{RR}	0.098	0.043	0.043	94.5	0.143	0.047	0.047	94.7
		ψ_{RD}	0.062	0.018	0.018	94.3	0.179	0.019	0.019	94.6
	LC	ψ_{OR}	0.140	0.153	0.153	94.8	0.033	0.165	0.164	94.8
		ψ_{RR}	0.040	0.042	0.042	94.7	-0.005	0.045	0.045	94.9
		ψ_{RD}	-0.030	0.017	0.017	94.9	-0.115	0.018	0.018	94.9

ACS: adaptive conditional score method; AUEE: adaptive unbiased estimating function method; LC: linear combination estimator based on ACS and AUEE.

cates be $B = 1000$. The goal is to estimate the causal odds ratio, causal risk ratio and the causal risk difference regarding two potential treatment plans $\bar{a}_1 = \{1, 1\}$ and $\bar{a}_0 = \{0, 0\}$.

Table 2.3 summarizes the results. As expected, the proposed methods all present satisfactory results. The AUEE method shows larger variances than the ACS method does. The linear combination method yields similar performance to the ACS method.

It is noted that the validity of the proposed methods requires correct specification of the parameters for the measurement error model. We now investigate the impact of misspecifying $\mathbf{\Sigma}_{\epsilon k}$. For Setting 1, we set $\sigma_{\epsilon k} = 0.5$ to generate data but misspecify it as $\sigma^*_{\epsilon k} = 0.4$ or $\sigma^*_{\epsilon k} = 0.6$ when conducting estimation. For Setting 2, we let the true covariance matrices be $\mathbf{\Sigma}_{\epsilon 0} = 2V_0$ and $\mathbf{\Sigma}_{\epsilon 1} = 2V_1$ for data generation but misspecify them as $\mathbf{\Sigma}^*_{\epsilon 0} = V_0$, $\mathbf{\Sigma}^*_{\epsilon 1} = V_1$, or $\mathbf{\Sigma}^*_{\epsilon 0} = 3V_0$, $\mathbf{\Sigma}^*_{\epsilon 1} = 3V_1$ in estimation procedures. Sample size $n = 5000$ is considered and 5000 simulations are run for each parameter configuration. The results are summarized in Tables 2.4 and 2.5. As expected, misspecifying the parameters for the measurement error models can lead to seriously biased estimates and coverage percentages that are significantly below 95% when using the proposed methods. In application, if there is no good knowledge about the parameter values for the measurement error models, it is sensible to conduct sensitivity analyses by applying the proposed estimators to examine how sensitive the results would be under a series of possible values of the parameters for the measurement error model.

Supplementary Material: An Alternative Approach

An alternative approach estimates the causal mean $E(Y_{\bar{a}})$ under treatment $\bar{a}$ by

$$\tilde{E}(Y_{\bar{a}}) = \frac{\sum_{i=1}^{n} \hat{w}_i Y_i I(\bar{A}_i = \bar{a})}{n}$$

where $I(\cdot)$ and $\hat{w}_i$ are the same as described in Theorem 2.1. Simulation results using $\tilde{E}(Y_{\bar{a}})$ with correctly-specified measurement error model are reported in Tables 2.6 and 2.7. Tables 2.8 and 2.9 summarize the simulation results using $\tilde{E}(Y_{\bar{a}})$ with misspecified covariance matrices of measurement error.

Table 2.3: Simulation results for Setting 2 with the correctly-specified measurement error model: average relative bias in percent (ReBias%), average standard error (ASE), empirical standard error (ESE) and coverage percentage (CP)

			$\boldsymbol{\Sigma}_{\epsilon 0} = V_0$, $\boldsymbol{\Sigma}_{\epsilon 1} = V_1$				$\boldsymbol{\Sigma}_{\epsilon 0} = 4V_0$, $\boldsymbol{\Sigma}_{\epsilon 1} = 4V_1$			
n	Method	Measure	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
2000	naive	ψ_{OR}	-6.237	0.065	0.065	90.0	-19.188	0.057	0.057	47.1
		ψ_{RR}	-2.915	0.038	0.038	89.9	-9.071	0.036	0.036	47.0
		ψ_{RD}	15.200	0.026	0.026	90.2	47.631	0.026	0.026	46.8
	benchmark	ψ_{OR}	0.480	0.069	0.069	95.0	0.514	0.069	0.070	94.3
		ψ_{RR}	0.092	0.039	0.039	94.8	0.093	0.039	0.039	94.4
		ψ_{RD}	-0.002	0.026	0.026	95.0	-0.041	0.026	0.026	94.3
	ACS	ψ_{OR}	0.488	0.071	0.071	94.9	0.769	0.077	0.076	94.7
		ψ_{RR}	0.091	0.040	0.040	95.0	0.189	0.043	0.042	94.8
		ψ_{RD}	0.034	0.027	0.027	94.9	-0.395	0.029	0.029	94.6
	AUEE	ψ_{OR}	1.550	0.082	0.081	94.5	2.400	0.093	0.089	94.1
		ψ_{RR}	0.508	0.045	0.045	94.5	0.836	0.050	0.049	94.2
		ψ_{RD}	-1.931	0.030	0.030	94.5	-3.460	0.033	0.033	94.0
	LC	ψ_{OR}	0.311	0.071	0.071	94.8	0.500	0.077	0.076	94.5
		ψ_{RR}	0.057	0.040	0.040	95.0	0.122	0.043	0.043	94.7
		ψ_{RD}	0.084	0.027	0.027	94.9	-0.275	0.029	0.029	94.6
5000	naive	ψ_{OR}	-6.496	0.041	0.041	82.3	-19.537	0.036	0.036	11.4
		ψ_{RR}	-2.962	0.024	0.024	82.6	-9.161	0.023	0.023	11.0
		ψ_{RD}	15.177	0.016	0.017	82.5	47.938	0.017	0.017	11.2
	benchmark	ψ_{OR}	0.150	0.043	0.043	94.9	0.138	0.043	0.044	94.4
		ψ_{RR}	0.022	0.024	0.024	94.8	0.012	0.024	0.025	94.2
		ψ_{RD}	0.083	0.016	0.016	94.9	0.131	0.016	0.017	94.4
	ACS	ψ_{OR}	0.216	0.044	0.045	94.9	0.241	0.047	0.048	94.4
		ψ_{RR}	0.047	0.025	0.025	94.7	0.049	0.027	0.027	94.2
		ψ_{RD}	-0.036	0.017	0.017	94.9	-0.009	0.018	0.018	94.3
	AUEE	ψ_{OR}	0.671	0.050	0.051	94.5	0.836	0.055	0.055	94.3
		ψ_{RR}	0.229	0.028	0.028	94.5	0.287	0.030	0.031	94.4
		ψ_{RD}	-0.896	0.019	0.019	94.5	-1.143	0.020	0.021	94.3
	LC	ψ_{OR}	0.179	0.044	0.045	94.9	0.140	0.047	0.048	94.3
		ψ_{RR}	0.042	0.025	0.025	94.7	0.025	0.027	0.027	94.2
		ψ_{RD}	-0.032	0.017	0.017	94.9	0.024	0.018	0.018	94.3

$V_0 = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$, $V_1 = \begin{pmatrix} 0.15 & 0.05 \\ 0.05 & 0.15 \end{pmatrix}$; ACS: adaptive conditional score method; AUEE: adaptive unbiased estimating function method; LC: linear combination estimator based on ACS and AUEE.

Table 2.4: Simulation results for Setting 1 with misspecified variance for the measurement error model: average relative bias in percent (ReBias%), average standard error (ASE), empirical standard error (ESE) and coverage percentage (CP)

		$\sigma^*_{\epsilon k} = 0.4$				$\sigma^*_{\epsilon k} = 0.6$			
Method	Measure	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
naive	ψ_{OR}	-7.972	0.138	0.138	77.6	-7.867	0.139	0.138	78.9
	ψ_{RR}	-3.449	0.040	0.040	77.6	-3.408	0.040	0.040	78.7
	ψ_{RD}	-11.600	0.017	0.017	77.0	-11.450	0.017	0.017	78.5
benchmark	ψ_{OR}	0.208	0.149	0.149	94.8	0.326	0.149	0.148	94.7
	ψ_{RR}	0.031	0.041	0.041	94.5	0.073	0.041	0.041	94.9
	ψ_{RD}	-0.118	0.017	0.017	94.7	0.040	0.017	0.017	94.9
ACS	ψ_{OR}	-3.136	0.147	0.147	91.9	5.472	0.163	0.163	89.0
	ψ_{RR}	-1.374	0.041	0.041	92.2	2.174	0.044	0.044	89.8
	ψ_{RD}	-4.696	0.017	0.017	91.9	6.711	0.017	0.017	89.5
AUEE	ψ_{OR}	-3.071	0.150	0.150	92.1	5.579	0.168	0.168	89.2
	ψ_{RR}	-1.347	0.042	0.042	92.2	2.218	0.045	0.045	89.6
	ψ_{RD}	-4.619	0.017	0.018	91.8	6.824	0.018	0.018	89.4
LC	ψ_{OR}	-3.260	0.147	0.147	91.8	5.299	0.163	0.163	89.3
	ψ_{RR}	-1.395	0.041	0.041	92.0	2.144	0.044	0.044	89.8
	ψ_{RD}	-4.697	0.017	0.017	91.9	6.702	0.017	0.017	89.6

ACS: adaptive conditional score method; AUEE: adaptive unbiased estimating function method; LC: linear combination estimator based on ACS and AUEE.

Table 2.5: Simulation results for Setting 2 with misspecified working covariance matrices: average relative bias in percent (ReBias%), average standard error (ASE), empirical standard error (ESE) and coverage percentage (CP)

		$\boldsymbol{\Sigma}^*_{\epsilon 0} = V_0$, $\boldsymbol{\Sigma}^*_{\epsilon 1} = V_1$				$\boldsymbol{\Sigma}^*_{\epsilon 0} = 3V_0$, $\boldsymbol{\Sigma}^*_{\epsilon 1} = 3V_1$			
Method	Measure	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
naive	ψ_{OR}	-11.824	0.039	0.039	51.8	-11.836	0.039	0.039	52.8
	ψ_{RR}	-5.438	0.023	0.024	52.5	-5.444	0.023	0.024	52.4
	ψ_{RD}	28.018	0.017	0.017	51.8	28.049	0.017	0.017	52.8
benchmark	ψ_{OR}	0.077	0.043	0.043	95.0	0.088	0.043	0.044	94.6
	ψ_{RR}	-0.013	0.024	0.024	94.7	-0.012	0.024	0.025	94.6
	ψ_{RD}	0.246	0.016	0.016	95.0	0.230	0.016	0.017	94.7
ACS	ψ_{OR}	-6.559	0.042	0.042	82.1	8.995	0.051	0.051	78.7
	ψ_{RR}	-2.996	0.024	0.024	82.2	3.819	0.027	0.028	78.3
	ψ_{RD}	15.345	0.017	0.017	82.2	-18.505	0.018	0.018	78.6
AUEE	ψ_{OR}	-6.175	0.046	0.046	86.4	9.656	0.060	0.059	79.6
	ψ_{RR}	-2.838	0.027	0.027	86.2	4.073	0.032	0.032	79.0
	ψ_{RD}	14.561	0.019	0.019	86.3	-19.656	0.021	0.021	79.5
LC	ψ_{OR}	-6.595	0.042	0.042	82.0	8.873	0.051	0.051	79.0
	ψ_{RR}	-3.002	0.024	0.024	82.1	3.790	0.027	0.028	78.5
	ψ_{RD}	15.348	0.017	0.017	82.2	-18.458	0.018	0.018	78.6

$V_0 = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$ and $V_1 = \begin{pmatrix} 0.15 & 0.05 \\ 0.05 & 0.15 \end{pmatrix}$

ACS: adaptive conditional score method; AUEE: adaptive unbiased estimating function method; LC: linear combination estimator based on ACS and AUEE.

Table 2.6: Supplementary Material: Simulation results using $\tilde{E}(Y_{\bar{a}})$ for Setting 1 with correctly-specified measurement error model

			$\boldsymbol{\Sigma}_{\epsilon k} = 0.5^2$				$\boldsymbol{\Sigma}_{\epsilon k} = 1^2$			
n	Method	Measure	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
1000	naive	ψ_{OR}	-6.871	0.332	0.325	92.0	-19.740	0.286	0.276	71.2
		ψ_{RR}	-3.228	0.091	0.090	91.8	-8.937	0.086	0.084	71.6
		ψ_{RD}	-11.737	0.039	0.039	91.8	-31.774	0.040	0.039	70.7
	benchmark	ψ_{OR}	1.630	0.364	0.355	94.9	1.434	0.362	0.345	95.3
		ψ_{RR}	0.265	0.094	0.093	95.0	0.230	0.094	0.091	95.4
		ψ_{RD}	-0.085	0.039	0.039	94.9	-0.242	0.039	0.038	95.3
	ACS	ψ_{OR}	1.773	0.379	0.367	95.0	2.129	0.501	0.404	95.6
		ψ_{RR}	0.295	0.097	0.096	95.0	0.372	0.108	0.104	95.6
		ψ_{RD}	-0.037	0.040	0.040	95.1	-0.002	0.044	0.043	95.6
	AUEE	ψ_{OR}	1.961	0.381	0.372	95.1	2.651	0.831	0.410	95.3
		ψ_{RR}	0.390	0.099	0.098	94.9	0.579	0.112	0.106	95.4
		ψ_{RD}	0.199	0.040	0.040	94.9	0.631	0.046	0.044	95.5
	LC	ψ_{OR}	0.775	0.372	0.363	94.9	0.739	0.425	0.393	95.4
		ψ_{RR}	0.035	0.097	0.096	94.9	-0.037	0.106	0.103	95.6
		ψ_{RD}	-0.661	0.040	0.040	95.0	-1.245	0.044	0.043	95.5
5000	naive	ψ_{OR}	-8.071	0.141	0.141	77.8	-20.579	0.123	0.122	12.0
		ψ_{RR}	-3.432	0.040	0.040	78.1	-9.079	0.038	0.038	13.3
		ψ_{RD}	-11.695	0.017	0.017	77.1	-31.524	0.018	0.018	12.1
	benchmark	ψ_{OR}	0.252	0.154	0.154	94.4	0.312	0.154	0.153	94.8
		ψ_{RR}	0.038	0.041	0.042	94.5	0.058	0.041	0.041	95.0
		ψ_{RD}	-0.090	0.017	0.017	94.5	-0.007	0.017	0.017	94.9
	ACS	ψ_{OR}	0.281	0.159	0.159	95.0	0.418	0.175	0.174	95.0
		ψ_{RR}	0.047	0.042	0.043	94.7	0.076	0.046	0.046	94.7
		ψ_{RD}	-0.074	0.018	0.018	94.8	0.015	0.019	0.019	95.1
	AUEE	ψ_{OR}	0.317	0.161	0.161	94.5	0.493	0.177	0.175	94.9
		ψ_{RR}	0.061	0.043	0.044	94.5	0.112	0.047	0.047	94.6
		ψ_{RD}	-0.032	0.018	0.018	94.6	0.114	0.020	0.020	95.0
	LC	ψ_{OR}	-0.090	0.158	0.158	94.8	-0.175	0.173	0.172	95.0
		ψ_{RR}	-0.006	0.042	0.043	94.6	-0.061	0.046	0.046	94.8
		ψ_{RD}	-0.200	0.018	0.018	94.6	-0.378	0.019	0.019	95.0

ACS: adaptive conditional score method; AUEE: adaptive unbiased estimating function method; LC: linear combination estimator based on ACS and AUEE.

Table 2.7: Supplementary Material: Simulation results using $\tilde{E}(Y_{\bar{a}})$ for Setting 2 with correctly-specified measurement error model

			$\boldsymbol{\Sigma}_{\epsilon 0} = V_0$, $\boldsymbol{\Sigma}_{\epsilon 1} = V_1$				$\boldsymbol{\Sigma}_{\epsilon 0} = 4V_0$, $\boldsymbol{\Sigma}_{\epsilon 1} = 4V_1$			
n	Method	Measure	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
2000	naive	ψ_{OR}	-5.843	0.069	0.068	91.0	-18.669	0.059	0.059	51.6
		ψ_{RR}	-2.842	0.041	0.041	90.9	-9.015	0.039	0.039	52.4
		ψ_{RD}	14.488	0.028	0.027	90.9	46.513	0.027	0.027	51.8
	benchmark	ψ_{OR}	0.727	0.074	0.073	95.3	0.647	0.074	0.073	94.8
		ψ_{RR}	0.143	0.043	0.042	95.1	0.096	0.043	0.043	95.0
		ψ_{RD}	-0.366	0.028	0.027	95.2	-0.149	0.028	0.028	94.8
	ACS	ψ_{OR}	0.761	0.076	0.075	95.3	0.947	0.083	0.081	95.1
		ψ_{RR}	0.149	0.044	0.043	95.1	0.200	0.047	0.047	95.2
		ψ_{RD}	-0.368	0.029	0.028	95.3	-0.552	0.031	0.030	95.1
	AUEE	ψ_{OR}	1.247	0.081	0.080	94.7	1.590	0.088	0.086	94.6
		ψ_{RR}	0.356	0.046	0.046	94.6	0.489	0.050	0.049	94.6
		ψ_{RD}	-1.229	0.030	0.030	94.6	-1.760	0.033	0.032	94.7
	LC	ψ_{OR}	0.185	0.075	0.074	95.2	0.124	0.082	0.080	94.8
		ψ_{RR}	-0.044	0.043	0.043	95.0	-0.092	0.047	0.046	94.9
		ψ_{RD}	0.094	0.028	0.028	95.2	0.296	0.031	0.030	94.9
5000	naive	ψ_{OR}	-6.300	0.043	0.043	83.7	-19.030	0.037	0.037	14.3
		ψ_{RR}	-2.948	0.026	0.026	84.4	-9.099	0.024	0.025	15.3
		ψ_{RD}	14.836	0.017	0.017	83.9	46.801	0.017	0.017	14.5
	benchmark	ψ_{OR}	0.210	0.046	0.046	94.5	0.200	0.046	0.046	94.4
		ψ_{RR}	0.026	0.027	0.027	94.7	0.017	0.027	0.027	94.5
		ψ_{RD}	0.028	0.017	0.018	94.5	0.061	0.017	0.018	94.3
	ACS	ψ_{OR}	0.251	0.047	0.047	94.7	0.370	0.051	0.051	94.4
		ψ_{RR}	0.040	0.027	0.028	94.6	0.079	0.029	0.030	94.4
		ψ_{RD}	-0.036	0.018	0.018	94.6	-0.199	0.019	0.019	94.4
	AUEE	ψ_{OR}	0.465	0.050	0.050	94.4	0.597	0.054	0.054	94.8
		ψ_{RR}	0.135	0.029	0.029	94.8	0.182	0.031	0.031	94.5
		ψ_{RD}	-0.433	0.019	0.019	94.5	-0.626	0.020	0.021	94.7
	LC	ψ_{OR}	0.028	0.047	0.047	94.7	-0.033	0.050	0.051	94.7
		ψ_{RR}	-0.042	0.027	0.027	94.5	-0.069	0.029	0.030	94.4
		ψ_{RD}	0.111	0.018	0.018	94.6	0.191	0.019	0.019	94.5

$V_0 = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$, $V_1 = \begin{pmatrix} 0.15 & 0.05 \\ 0.05 & 0.15 \end{pmatrix}$; ACS: adaptive conditional score method; AUEE: adaptive unbiased estimating function method; LC: linear combination estimator based on ACS and AUEE.

Table 2.8: Supplementary Material: Simulation results using $\tilde{E}(Y_{\bar{a}})$ for Setting 1 with misspecified working covariance matrix $\mathbf{\Sigma}^*_{\epsilon k} \neq \mathbf{\Sigma}_{\epsilon k} = 0.5^2$

		$\mathbf{\Sigma}^*_{\epsilon k} = 0.4^2$				$\mathbf{\Sigma}^*_{\epsilon k} = 0.6^2$			
Method	Measure	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
naive	ψ_{OR}	-8.058	0.141	0.140	78.2	-7.909	0.142	0.141	78.9
	ψ_{RR}	-3.431	0.040	0.040	78.4	-3.373	0.040	0.040	79.4
	ψ_{RD}	-11.677	0.017	0.017	78.1	-11.468	0.017	0.017	78.8
benchmark	ψ_{OR}	0.245	0.154	0.154	95.0	0.417	0.155	0.153	94.8
	ψ_{RR}	0.031	0.041	0.041	94.9	0.095	0.041	0.041	95.1
	ψ_{RD}	-0.099	0.017	0.017	95.0	0.128	0.017	0.017	95.0
ACS	ψ_{OR}	-3.161	0.152	0.150	92.4	5.734	0.173	0.172	89.4
	ψ_{RR}	-1.366	0.041	0.041	92.3	2.189	0.044	0.044	89.6
	ψ_{RD}	-4.725	0.018	0.017	92.2	6.921	0.018	0.018	89.6
AUEE	ψ_{OR}	-3.130	0.153	0.152	92.4	5.770	0.174	0.174	89.1
	ψ_{RR}	-1.351	0.042	0.042	92.4	2.208	0.045	0.045	89.5
	ψ_{RD}	-4.685	0.018	0.018	92.4	6.964	0.018	0.018	89.3
LC	ψ_{OR}	-3.457	0.151	0.150	91.8	5.237	0.171	0.171	90.0
	ψ_{RR}	-1.402	0.041	0.041	92.1	2.109	0.044	0.044	89.9
	ψ_{RD}	-4.797	0.018	0.017	92.2	6.701	0.018	0.018	89.6

ACS: adaptive conditional score method; AUEE: adaptive unbiased estimating function method; LC: linear combination estimator based on ACS and AUEE.

Table 2.9: Supplementary Material: Simulation results using $\tilde{E}(Y_{\bar{a}})$ for Setting 2 with misspecified working covariance matrices $\boldsymbol{\Sigma}^*_{\epsilon 0} \neq \boldsymbol{\Sigma}_{\epsilon 0} = 2V_0$, $\boldsymbol{\Sigma}^*_{\epsilon 1} \neq \boldsymbol{\Sigma}_{\epsilon 1} = 2V_1$

		$\boldsymbol{\Sigma}^*_{\epsilon 0} = V_0$, $\boldsymbol{\Sigma}^*_{\epsilon 1} = V_1$				$\boldsymbol{\Sigma}^*_{\epsilon 0} = 3V_0$, $\boldsymbol{\Sigma}^*_{\epsilon 1} = 3V_1$			
Method	Measure	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
naive	ψ_{OR}	-11.458	0.040	0.040	57.4	-11.463	0.040	0.040	57.7
	ψ_{RR}	-5.390	0.025	0.025	58.5	-5.395	0.025	0.025	58.5
	ψ_{RD}	27.282	0.017	0.017	57.9	27.294	0.017	0.017	57.8
benchmark	ψ_{OR}	0.172	0.046	0.046	94.9	0.187	0.046	0.046	94.9
	ψ_{RR}	0.007	0.027	0.027	94.8	0.006	0.027	0.027	94.7
	ψ_{RD}	0.109	0.017	0.017	94.9	0.093	0.017	0.018	94.8
ACS	ψ_{OR}	-6.308	0.043	0.044	84.4	8.914	0.055	0.056	81.0
	ψ_{RR}	-2.960	0.026	0.027	84.5	3.814	0.031	0.031	81.7
	ψ_{RD}	14.885	0.018	0.018	84.5	-18.278	0.019	0.020	80.9
AUEE	ψ_{OR}	-6.107	0.046	0.046	86.3	9.128	0.059	0.059	81.7
	ψ_{RR}	-2.871	0.028	0.028	86.5	3.909	0.032	0.032	81.0
	ψ_{RD}	14.484	0.019	0.019	86.4	-18.645	0.020	0.020	81.4
LC	ψ_{OR}	-6.464	0.043	0.047	83.7	8.361	0.055	0.055	82.5
	ψ_{RR}	-3.024	0.026	0.027	83.7	3.632	0.030	0.031	82.5
	ψ_{RD}	14.987	0.018	0.018	84.4	-17.740	0.019	0.019	81.7

$V_0 = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$ and $V_1 = \begin{pmatrix} 0.15 & 0.05 \\ 0.05 & 0.15 \end{pmatrix}$

ACS: adaptive conditional score method; AUEE: adaptive unbiased estimating function method; LC: linear combination estimator based on ACS and AUEE.

Chapter 3

Inverse-Probability-of-Treatment Weighted Estimation of Causal Parameters with Error-Prone Data

This chapter deals with Problem 2 discussed in Section 1.5. Section 3.1 describes the notation and framework with error free data. Section 3.2 presents the measurement error model we consider. Section 3.3 describes methods to account for the measurement error effects on IPTW estimation with error-prone and possibly time-dependent confounders. In section 3.4, we perform simulation studies to assess the performance of the explored methods in finite samples, and apply our methods to conduct sensitivity analyses for NHEFS data.

3.1 Notation and Framework

Suppose subjects in the study are assessed at discrete time points $0, 1, \ldots, K$. For $k = 0, 1, \ldots, K$, let $A(k)$ denote the binary treatment indicator at visit k and $\bar{A}(k) = \{A(u) : u = 0, 1, \ldots, k\}$ the treatment history up to and including visit k. Write $\bar{A} = \bar{A}(K)$. Let $(Z^{\mathrm{T}}(k), X^{\mathrm{T}}(k))^{\mathrm{T}}$ be the vector of possibly time-dependent confounders that are measured at visit k, where the $Z(k)$ are precisely measured and the $X(k)$ are subject to measurement

error. Both $X(k)$ and $Z(k)$ can be univariate or multivariate. Depending on the application context, $Z(k)$ and $X(k)$ may represent the true measurements precisely collected at time point k, or the average measurements over the time period between visit $k-1$ and visit k; sometimes, $Z(k)$ and $X(k)$ may even refer to the cumulative average measurements up to and including time point k. Let $\bar{Z}(k) = \{Z(u) : u = 0, 1, \ldots, k\}$ and $\bar{X}(k) = \{X(u) : u = 0, 1, \ldots, k\}$ be the confounder histories up to and including visit k, $\bar{Z} = \bar{Z}(K)$, and $\bar{X} = \bar{X}(K)$. Let Y denote the outcome that is observed at the end of the study which can be either continuous or discrete.

For $k = 0, 1, \ldots, K$, let $a(k)$ be a realization of the treatment indicator at visit k and $\bar{a}(k) = \{a(u) : u = 0, 1, \ldots, k\}$ be the corresponding treatment history up to and including visit k. Write $\bar{a} = \bar{a}(K)$. Let $Y_{\bar{a}}$ denote the potential outcome that would have been observed had a subject experienced treatment history $\bar{a}$, and let $Y_{i,\bar{a}}$ denote the potential outcome for subject i that would have been observed had this subject experienced treatment history $\bar{a}$. Suppose we have a sample of observations from n subjects. We add subscript i to symbols from time to time to indicate information for subject i in the sample. For instance, $A_i(k)$ represents the observed treatment at visit k for subject i.

We make the following assumptions for all k.

Assumption 1 (No Interference): the treatment taken by subject j has no effect on the potential outcomes of subject i for all $i \neq j$.

Assumption 2 (Consistency): $Y_{\bar{A}} = Y$.

Assumption 3 (No Unmeasured Confounding): $P(\bar{A}|\bar{Z}, \bar{X}, Y_{\bar{a}}) = P(\bar{A}|\bar{Z}, \bar{X})$.

Assumption 4 (Positivity): $0 < P\{A(k) = 1|\bar{A}(k-1), \bar{Z}(k), \bar{X}(k)\} < 1$.

3.1.1 Model Setup

Assume a marginal structural model for the mean of the potential outcome with treatment history $\bar{a}$:

$$E(Y_{\bar{a}}) = h(\bar{a}; \boldsymbol{\beta}), \tag{3.1}$$

where $\boldsymbol{\beta}$ is the vector of causal parameters of interest, and $h(\cdot)$ is a known function. Model (3.1) basically facilitates the marginal feature (i.e., the mean) of potential outcomes by postulating its relationship with the treatment history (Robins et al., 2000; Daniel et al., 2013).

In reality, we can only observe the treatment history $\bar{A}$ and associated outcome Y rather than $\bar{a}$ and all the potential outcomes $Y_{\bar{a}}$. In association studies, one may employ a model form as in (3.1) to describe the conditional mean of the observed outcome, given the observed treatment history $\bar{A}$:

$$E(Y|\bar{A}) = h(\bar{A}; \boldsymbol{\alpha}), \tag{3.2}$$

where $\boldsymbol{\alpha}$ is the vector of associational regression parameters. However, in the presence of time-dependent confounders, which may also be predicted by the previous treatment, fitting model (3.2) to the observed data generally yields biased estimation for the causal effect $\boldsymbol{\beta}$ (Robins et al., 2000).

3.1.2 IPTW Estimation of Causal Effects

To produce a consistent estimator for $\boldsymbol{\beta}$, Robins et al. (2000) proposed the IPTW estimation method which includes the following two steps.

Step 1 (Weight Estimation):

For each subject i, determine the weight

$$w_i = \prod_{k=0}^{K} \frac{1}{P\{A_i(k)|\bar{A}_i(k-1), \bar{Z}_i(k), \bar{X}_i(k)\}}, \tag{3.3}$$

which requires the determination of conditional probabilities $P\{A_i(k)|\bar{A}_i(k-1), \bar{Z}_i(k), \bar{X}_i(k)\}$ for $k = 0, 1, \ldots, K$. Since the $A_i(k)$ are binary variables, we invoke conventional modeling

techniques to characterize these conditional probabilities:

$$\text{logit}[P\{A_i(k)=1|\bar{A}_i(k-1),\bar{Z}_i(k),\bar{X}_i(k)\}]=\gamma_{0k}+\boldsymbol{\gamma}_{Ak}^{\mathrm{T}}\bar{A}_i(k-1)+\boldsymbol{\gamma}_{Zk}^{\mathrm{T}}\bar{Z}_i(k)+\boldsymbol{\gamma}_{Xk}^{\mathrm{T}}\bar{X}_i(k), \tag{3.4}$$

where $\boldsymbol{\gamma}_k=(\gamma_{0k},\boldsymbol{\gamma}_{Ak}^{\mathrm{T}},\boldsymbol{\gamma}_{Zk}^{\mathrm{T}},\boldsymbol{\gamma}_{Xk}^{\mathrm{T}})^{\mathrm{T}}$ is the vector of regression parameters for $k=0,1,\ldots,K$. We fit (3.4) to the observed data and obtain an estimator for $\boldsymbol{\gamma}_k$ for each k, thus yielding the estimator $\hat{w}_i$ of w_i from (3.3).

Weights defined by (3.3) can be quite large when some probabilities in the denominator are close to 0, resulting in unstable numerical results; these weights are thereby referred to as *unstabilized* weights. To produce more stable results, Robins et al. (2000) and Daniel et al. (2013) suggested to alternatively take the *stabilized* weights:

$$sw_i=\prod_{k=0}^{K}\frac{P\{A_i(k)|\bar{A}_i(k-1)\}}{P\{A_i(k)|\bar{A}_i(k-1),\bar{Z}_i(k),\bar{X}_i(k)\}}, \tag{3.5}$$

whose denominators are the same as those in (3.3). The numerator of sw_i requires specification of models for the $A(k)$ conditioning on $\bar{A}(k-1)$:

$$P\{A_i(k)=1|\bar{A}_i(k-1)\}=g(\bar{A}_i(k-1);\theta_{0k},\boldsymbol{\theta}_{Ak}), \tag{3.6}$$

where $g(\cdot)$ is a link function, and θ_{0k} and $\boldsymbol{\theta}_{Ak}$ are regression parameters. Often, the logistic model is used (Robins et al., 2000):

$$\text{logit}[P\{A_i(k)=1|\bar{A}_i(k-1)\}]=\theta_{0k}+\boldsymbol{\theta}_{Ak}^{\mathrm{T}}\bar{A}_i(k-1). \tag{3.7}$$

Let $\widehat{sw}_i$ denote the resulting estimate of sw_i.

Step 2 (Fitting the Weighted Outcome Model):

For $i=1,\ldots,n$, assign weights $\hat{w}_i$ or $\widehat{sw}_i$ to subject i and fit model (3.2) accordingly. Let $\hat{\boldsymbol{\beta}}$ denote the resulting estimator of parameter $\boldsymbol{\alpha}$ in model (3.2); and $\hat{\boldsymbol{\beta}}$ is the so-called the IPTW estimator for the causal effect $\boldsymbol{\beta}$ of model (3.1).

3.2 Measurement Error Model

The validity of IPTW estimation discussed in Section 3.1.2 requires the weights to be correctly specified. However, when the $X_i(k)$ are mismeasured, disregarding the difference between $X_i(k)$ and the observed version and naively using the procedure of Section 3.1.2 would produce biased estimates. In fact, (3.3), (3.4) and (3.5) show that the denominators of w_i or sw_i are vulnerable to mismeasurements in $X_i(k)$.

We consider the situation where the $X_i(k)$ are mismeasured and other variables are precisely measured. Let X_{ik}^* be an observed measurement of $X_i(k)$. Assume that conditional on $X_i(k)$,

$$X_{ik}^* = X_i(k) + \epsilon_{ik} \tag{3.8}$$

for $i = 1, \ldots, n$ and $k = 0, \ldots, K$, where the ϵ_{ik} and the $X_i(k)$ are independent, and the ϵ_{ik} are independent across different i and k. Assume that the error terms ϵ_{ik} follow $N(\mathbf{0}, \mathbf{\Sigma}_{\epsilon k})$, with covariance matrix $\mathbf{\Sigma}_{\epsilon k}$. To highlight the key idea, we assume that $\mathbf{\Sigma}_{\epsilon k}$ is known for now.

Model (3.8) is commonly used in the literature and is referred to as the classical additive measurement error model (e.g., Carroll et al., 2006; Yi, 2017). It characterizes situations where the observed value is an unbiased measure of the true value with additional additive noise involved.

We comment that the assumption of independency among the ϵ_{ik} is reasonable when the sources of measurement error for different visits are unlikely to be correlated. This assumption does not imply the independence among the observed measurements X_{ik}^*. To see this, consider a simple case where the $X_i(k)$ and the X_{ik}^* are all scalar. Then $cov(X_{ik}^*, X_{il}^*) = E(X_{ik}^* X_{il}^*) - E(X_{ik}^*)E(X_{il}^*) = E\{X_i(k)X_i(l)\} - E\{X_i(k)\}E\{X_i(l)\} = cov\{X_i(k), X_i(l)\}$, suggesting that X_{ik}^* and X_{il}^* are allowed to be correlated through the correlation between $X_i(k)$ and $X_i(l)$ (Freedman et al., 2015). Furthermore, the independence assumption among the ϵ_{ik} is often plausible to feature the problems in which the observed measurements at different visits are less correlated than the true measurements, because $var(X_{ik}^*) > var\{X_i(k)\}$ and $var(X_{il}^*) > var\{X_i(l)\}$ give $|corr(X_{ik}^*, X_{il}^*)| < |corr\{X_i(k), X_i(l)\}|$. When the independence assumption of the ϵ_{ik} is unreasonable, more sophisticated measurement

error models can be considered. Detailed discussions can be found in Freedman et al. (2015).

3.3 Adjusting for Measurement Error Effects on the Estimation of Causal Parameters

To carry out valid estimation of causal parameters in the presence of mismeasurements, we need to address error effects on estimation of the IPTW weights. By (3.3) or (3.5), it suffices to account for error effects on the estimation of fitted probabilities in models (3.4). In this section we describe three schemes of handling mismeasurement effects. We explore the extension of the regression calibration method (Prentice, 1982) and the conditional score method (Stefanski and Carroll, 1987) which were originally developed for non-causal settings to the settings of causal inference with measurement error. With some modifications, we also present the two types of simulation-extrapolation methods (Cook and Stefanski, 1994) which were also examined by Kyle et al. (2016).

3.3.1 Regression Calibration

The first approach is to apply the regression calibration (RC) method (Prentice, 1982) to address measurement error effects. The basic idea of the RC method is to conduct a standard analysis with the X covariates replaced by their conditional expectations given the observed data. The RC method was initiated by Prentice (1982) for analyzing survival data with time-independent covariate mismeasurements, and has proved in many non-causal inference settings to outperform the naive analysis which disregards measurement error (e.g., Carroll et al., 2006; Rosner et al., 1989, 1990).

With time-varying covariates $X_i(k)$ here, we may apply the RC method with $X_i(k)$ replaced by its expectation $E\{X_i(k)|X_{ik}^*\}$ which is conditioning on the observed value X_{ik}^*. Using the discussion of Carroll et al. (2006, Sec. 4.4.2), we estimate the conditional

expectation $E\{X_i(k)|X_{ik}^*\}$ by

$$\hat{X}_i(k) = \hat{\mu}_k + \hat{\mathbf{\Sigma}}_{Xk} \cdot (\hat{\mathbf{\Sigma}}_{Xk} + \mathbf{\Sigma}_{\epsilon k})^{-1} \cdot (X_{ik}^* - \hat{\mu}_k), \tag{3.9}$$

where $\hat{\mu}_k = \frac{\sum_{i=1}^n X_{ik}^*}{n}$ and $\hat{\mathbf{\Sigma}}_{Xk} = \frac{\sum_{i=1}^n (X_{ik}^* - \hat{\mu}_k)(X_{ik}^* - \hat{\mu}_k)^{\mathrm{T}} - (n-1)\mathbf{\Sigma}_{\epsilon k}}{n-1}$.

Replacing $X_i(k)$ with $\hat{X}_i(k)$ and fitting models (3.4) to the data gives estimates of the logistic model parameters, thus leading to the adjusted IPTW weights which will serve as the input in Step 2 of the standard IPTW estimation in Section 3.1.2.

It is interesting that this RC method and the naive analysis which replaces $X_i(k)$ with X_{ik}^* in the procedure of Section 3.1.2 produce the same point estimates for w_i and sw_i, and thus for $\boldsymbol{\beta}$. Indeed, by (3.9), the estimate of $E\{X_i(k)|X_{ik}^*\}$ is linear in X_{ik}^*, given by

$$\hat{X}_i(k) = u_k + V_k X_{ik}^*, \tag{3.10}$$

where $u_k = \hat{\mu}_k - \hat{\mathbf{\Sigma}}_{Xk} \cdot (\hat{\mathbf{\Sigma}}_{Xk} + \mathbf{\Sigma}_{\epsilon k})^{-1} \cdot \hat{\mu}_k$ and $V_k = \hat{\mathbf{\Sigma}}_{Xk} \cdot (\hat{\mathbf{\Sigma}}_{Xk} + \mathbf{\Sigma}_{\epsilon k})^{-1}$.

Maximization of the likelihood function of logistic model (3.4) with $X_i(k)$ replaced by $\hat{X}_i(k)$ yields the maximum likelihood estimator of $\boldsymbol{\gamma}$, denoted as $\hat{\boldsymbol{\gamma}}^* = (\hat{\gamma}_{0k}^*, \hat{\boldsymbol{\gamma}}^{*\mathrm{T}}_{Ak}, \hat{\boldsymbol{\gamma}}^{*\mathrm{T}}_{Zk}, \hat{\boldsymbol{\gamma}}^{*\mathrm{T}}_{Xk})^{\mathrm{T}}$.

Let $\bar{u}_k = (u_0^{\mathrm{T}}, u_1^{\mathrm{T}}, \ldots, u_k^{\mathrm{T}})^{\mathrm{T}}$ and let $\bar{V}_k = \mathrm{diag}(V_0, \ldots, V_k)$ be the diagonal block matrix. By the form of (3.10) and the logistic model (3.4), we obtain that the naive estimator $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_{0k}, \tilde{\boldsymbol{\gamma}}_{Ak}^{\mathrm{T}}, \tilde{\boldsymbol{\gamma}}_{Zk}^{\mathrm{T}}, \tilde{\boldsymbol{\gamma}}_{Xk}^{\mathrm{T}})^{\mathrm{T}}$ of $\boldsymbol{\gamma}$, obtained from fitting (3.4) with $X_i(k)$ replaced by X_{ik}^*, is related to estimator $\hat{\boldsymbol{\gamma}}^*$ as follows:

$$\tilde{\gamma}_{0k} = \hat{\gamma}_{0k}^* + \hat{\boldsymbol{\gamma}}^{*\mathrm{T}}_{Xk} \bar{u}_k, \quad \tilde{\gamma}_{Ak} = \hat{\boldsymbol{\gamma}}^*_{Ak}, \quad \tilde{\gamma}_{Zk} = \hat{\boldsymbol{\gamma}}^*_{Zk}, \qquad \text{and} \quad \tilde{\boldsymbol{\gamma}}_{Xk}^{\mathrm{T}} = \hat{\boldsymbol{\gamma}}^{*\mathrm{T}}_{Xk} \bar{V}_k. \tag{3.11}$$

Then by examining the terms in the denominators of (3.3) and (3.5) which involve error-prone covariates, we obtain the identical relationship between the two estimated counterparts derived from using estimators $\hat{\boldsymbol{\gamma}}^*$ and $\tilde{\boldsymbol{\gamma}}$:

$$\frac{1}{1 + \exp\{-\hat{\gamma}_{0k}^* - \hat{\boldsymbol{\gamma}}^{*\mathrm{T}}_{Ak} \bar{A}_i(k-1) - \hat{\boldsymbol{\gamma}}^{*\mathrm{T}}_{Zk} \bar{Z}_i(k) - \hat{\boldsymbol{\gamma}}^{*\mathrm{T}}_{Xk} \bar{\hat{X}}_i(k)\}}$$

$$= \frac{1}{1+\exp\{-\hat{\gamma}^*_{0k} - \hat{\boldsymbol{\gamma}}^{*\mathrm{T}}_{Ak}\bar{A}_i(k-1) - \hat{\boldsymbol{\gamma}}^{*\mathrm{T}}_{Zk}\bar{Z}_i(k) - \hat{\boldsymbol{\gamma}}^{*\mathrm{T}}_{Xk}(\bar{u}_k + \bar{V}_k\bar{X}^*_{ik})\}}$$
$$= \frac{1}{1+\exp\{-\tilde{\gamma}_{0k} - \tilde{\boldsymbol{\gamma}}^{\mathrm{T}}_{Ak}\bar{A}_i(k-1) - \tilde{\boldsymbol{\gamma}}^{\mathrm{T}}_{Zk}\bar{Z}_i(k) - \tilde{\boldsymbol{\gamma}}^{\mathrm{T}}_{Xk}\bar{X}^*_{ik}\}},$$

where $\bar{\hat{X}}_i(k) = (\hat{X}^{\mathrm{T}}_i(0), \ldots, \hat{X}^{\mathrm{T}}_i(k))^{\mathrm{T}}$, and (3.10) and (3.11) are, respectively, used at the first and second steps. Consequently, by (3.3) and (3.5), we conclude that RC and the naive methods produce the same estimated IPTW weights and hence, the same estimate for the causal parameter $\boldsymbol{\beta}$.

3.3.2 SIMEX Correction Methods

In association studies with error-prone variables, the simulation-extrapolation (SIMEX) method, proposed by Cook and Stefanski (1994), is another popularly used algorithm; see Yi (2008) and Yi and He (2012), among many others. Theoretical justification of this method was provided by Carroll et al. (1996).

Let B be a given positive integer, and $\Lambda = \{\lambda_1, \lambda_2 \ldots, \lambda_M\}$ be a sequence of increasing numbers, where $\lambda_1 = 0$, M is a given positive integer, and λ_M is a prespecified positive number. The SIMEX method consists of the following three steps.

Step 1 (Simulation):

For $i = 1, \ldots, n$ and $k = 0, \ldots, K$, generate $e_{ikb} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\epsilon k})$ for $b = 1, 2, \ldots, B$. For $\lambda \in \Lambda$, calculate $X^*_i(k; b, \lambda) = X^*_{ik} + \sqrt{\lambda} e_{ikb}$.

Step 2 (Estimation):

Estimate the logistic regression parameters in (3.4) with $X_i(k)$ replaced by $X^*_i(k; b, \lambda)$, and let $\hat{\boldsymbol{\gamma}}_k(b, \lambda)$ denote the resulting estimator. Calculate $\hat{\boldsymbol{\gamma}}_k(\lambda) = \frac{1}{B}\sum_{b=1}^{B} \hat{\boldsymbol{\gamma}}_k(b, \lambda)$ for $k = 0, 1, \ldots, K$.

Step 3: (Extrapolation):

For $r = 1, 2, \ldots, q_k$ where q_k is the dimension of $\boldsymbol{\gamma}_k$, fit a regression model to $\{(\lambda, \hat{\boldsymbol{\gamma}}_{kr}(\lambda)) : \lambda \in \Lambda\}$ where $\hat{\boldsymbol{\gamma}}_{kr}(\lambda)$ is the rth element of $\hat{\boldsymbol{\gamma}}_k(\lambda)$, and extrapolating back to $\lambda = -1$ gives a

predicted value $\hat{\boldsymbol{\gamma}}_{kr}(-1)$. Then $(\hat{\boldsymbol{\gamma}}_{kr}(-1), r = 1, 2, \ldots, q_k)^{\mathrm{T}}$ is the SIMEX estimator of $\boldsymbol{\gamma}_k$, denoted as $\hat{\boldsymbol{\gamma}}_k^{\dagger}$. Consequently, using (3.3) and (3.4), or, (3.4) and (3.5), with $\boldsymbol{\gamma}_k$ replaced by $\hat{\boldsymbol{\gamma}}_k^{\dagger}$ and $X_i(k)$ replaced by $\hat{X}_i(k)$ gives the adjusted IPTW weights which will serve as the input in Step 2 of the standard IPTW estimation in Section 3.1.2. Our descriptions slightly differ from Kyle et al. (2016) in that $X_i(k)$ is replaced by $\hat{X}_i(k)$ rather than X_{ik}^* at the second stage.

This method, called the *indirect SIMEX correction* method (Kyle et al., 2016) , adjusts for the measurement error effects by correcting the IPTW weights first and then substituting the resulting weights into the standard IPTW estimation procedure. Alternatively, we can directly adjust for the error effects on the estimation of causal parameter $\boldsymbol{\beta}$ by modifying Steps 2 and 3, and call it the *direct SIMEX correction* (Kyle et al., 2016) method. The procedure consists of the following three steps.

Step I (Simulation):

This step is the same as Step 1 of the indirect SIMEX correction method.

Step II (Estimation):

Estimate the causal parameter $\boldsymbol{\beta}$ by conducting the standard IPTW estimation in Section 3.1.2 with $X_i(k)$ replaced by $X_i^*(k; b, \lambda)$, and let $\hat{\boldsymbol{\beta}}(b, \lambda)$ denote the resultant estimator. Let $\hat{\boldsymbol{\beta}}_r(b, \lambda)$ be the rth component of $\hat{\boldsymbol{\beta}}(b, \lambda)$ where $r = 1, 2, \ldots, p$ and p is the dimension of $\boldsymbol{\beta}$. Calculate

$$\hat{\boldsymbol{\beta}}_r(\lambda) = \frac{1}{B} \sum_{b=1}^{B} \hat{\boldsymbol{\beta}}_r(b, \lambda).$$

Step III (Extrapolation):

For $r = 1, 2, \ldots, p$, fit a regression model to $\{(\lambda, \hat{\boldsymbol{\beta}}_r(\lambda)) : \lambda \in \Lambda\}$, and extrapolating them back to $\lambda = -1$ gives a predicted value $\hat{\boldsymbol{\beta}}_r(-1)$. Then $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_r(-1), r = 1, 2, \ldots, p)^{\mathrm{T}}$ is the SIMEX estimator of $\boldsymbol{\beta}$.

3.3.3 Refined Correction Method

Estimators derived from the SIMEX methods are simple to implement and work generally well for many settings. However, they usually are approximately consistent because there

is no knowledge of the true extrapolation function and only an approximation is used (Carroll et al., 1996). To obtain a consistent estimator of $\boldsymbol{\beta}$, we adapt the conditional score method proposed by Stefanski and Carroll (1987) for non-causal regression settings with time-independent covariates. For this purpose, we further impose

Assumption 5 (Markov Assumption):

$$P\{A(k)|\bar{A}(k-1), \bar{Z}(k), \bar{X}(k)\} = P\{A(k)|\bar{A}(k-1), \bar{Z}(k), X(k)\} \quad \text{for} \quad k = 0, \ldots, K.$$

The assumption says that given the history of error-free confounders and the treatment history, the probability of receiving the treatment at the present moment depends only on the current error-prone confounders but not their history. Assumption 5 is needed for establishing the consistent estimation of the causal parameters using the refined correction method to be discussed. This assumption holds automatically in commonly-seen time-invariant studies where $K = 0$; it can be regarded as the price paid for generalizing the development that is valid only for time-invariant settings to circumstances with time-dependent measurements. This assumption resembles the usual first-order Markov assumption and may be plausible for many settings. When this assumption is in doubt in application, one may modify the definition of $X_i(t)$ and $Z_i(t)$ by properly including current or previous confounders so that Assumption 5 may be feasible. Discussion on this strategy can be found in Miglioretti and Heagerty (2004).

Let $\Delta_i(k) = X^*_{ik} + \{A_i(k) - 1/2\}\boldsymbol{\Sigma}_{\epsilon k}\boldsymbol{\gamma}_{Xk}$. Stefanski and Carroll (1987) proposed the estimating equations

$$\sum_{i=1}^{n} \left(\left[A_i(k) - \frac{1}{1 + \exp\{-\gamma_{0k} - \boldsymbol{\gamma}_{Ak}^{\mathrm{T}}\bar{A}_i(k-1) - \boldsymbol{\gamma}_{Zk}^{\mathrm{T}}\bar{Z}_i(k) - \boldsymbol{\gamma}_{Xk}^{\mathrm{T}}\Delta_i(k)\}} \right] \begin{Bmatrix} 1 \\ \bar{A}_i(k-1) \\ \bar{Z}_i(k) \\ \Delta_i(k) \end{Bmatrix} \right) = \mathbf{0} \tag{3.12}$$

for $\boldsymbol{\gamma}_k$; Solving (3.12) for $\boldsymbol{\gamma}_k$ yields an estimate of $\boldsymbol{\gamma}_k$. Let $\hat{\boldsymbol{\gamma}}_k = (\hat{\gamma}_{0k}, \hat{\boldsymbol{\gamma}}_{Ak}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}_{Zk}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}_{Xk}^{\mathrm{T}})^{\mathrm{T}}$ be the resulting estimator of $\boldsymbol{\gamma}_k$, which is a consistent estimator of $\boldsymbol{\gamma}_k$, provided regularity

conditions (Stefanski and Carroll, 1987).

We notice that (3.12) is identical to the likelihood score function derived from the model (3.4) with $X_i(k)$ replaced by $\Delta_i(k)$. Driven by this observation, we propose to use (3.4) with $X_i(k)$ replaced by $\Delta_i(k)$ to estimate the conditional probabilities $P\{A_i(k)|\bar{A}_i(k-1), \bar{Z}_i(k), X_i(k)\}$, given by

$$\begin{aligned} &\hat{P}\{A_i(k)|\bar{A}_i(k-1), \bar{Z}_i(k), \hat{\Delta}_i(k)\} \\ &= \frac{1}{1+\exp[\{-\hat{\gamma}_{0k} - \hat{\boldsymbol{\gamma}}_{Ak}^{\mathrm{T}}\bar{A}_i(k-1) - \hat{\boldsymbol{\gamma}}_{Zk}^{\mathrm{T}}\bar{Z}_i(k) - \hat{\boldsymbol{\gamma}}_{Xk}^{\mathrm{T}}\hat{\Delta}_i(k)\}\{2A_i(k)-1\}]}, \end{aligned} \tag{3.13}$$

where $\hat{\Delta}_i(k) = X_{ik}^* + \{A_i(k) - 1/2\}\boldsymbol{\Sigma}_{\epsilon k}\hat{\boldsymbol{\gamma}}_{Xk}$.

In Appendix B, we establish the consistency for the refined estimator of the causal parameter, which is summarized as follows.

Theorem 3.1. *Suppose Assumptions 1-5 and (3.4) hold, and the measurement error model is specified as (3.8). Let*

$$\hat{w}_i = \prod_{k=0}^{K} \frac{1}{\hat{P}\{A_i(k)|\bar{A}_i(k-1), \bar{Z}_i(k), \hat{\Delta}_i(k)\}} \quad \textit{and} \quad \widehat{sw}_i = \prod_{k=0}^{K} \frac{\hat{P}\{A_i(k)|\bar{A}_i(k-1)\}}{\hat{P}\{A_i(k)|\bar{A}_i(k-1), \bar{Z}_i(k), \hat{\Delta}_i(k)\}}$$

denote the estimated weights corresponding to (3.3) and (3.5) where the denominators of $\hat{w}_i$ and $\widehat{sw}_i$ are given by (3.13), and the numerators of $\widehat{sw}_i$ are obtained from (3.6). Then, the following properties hold:

(a). Fitting model (3.2) with weight $\hat{w}_i$ yields a consistent estimator for the causal parameter $\boldsymbol{\beta}$ of model (3.1).

(b). Fitting model (3.2) with weight $\widehat{sw}_i$ yields a consistent estimator for the causal parameter $\boldsymbol{\beta}$ of model (3.1). Furthermore, misspecification of model (3.6) does not affect the consistency of the resulting estimator for $\boldsymbol{\beta}$.

Theorem 3.1 implies that, attaching weights $\hat{w}_i$ or $\widehat{sw}_i$ to subject-level data makes the data resemble those collected from randomized studies, thus the bias caused by measure-

ment error and the imbalance in confounders can be eliminated using the IPTW estimation method with the proposed correction of measurement error. We comment that (3.13) reduces to an equivalent expression considered by McCaffrey et al. (2013) when the treatment is time-invariant, where the estimand of interest is the causal mean rather than the causal parameters in marginal structural models.

To obtain variance estimates for the proposed estimators, we use the bootstrap method (Efron, 1982). Let $\hat{\beta}_r$ denote the estimator of β_r obtained by using a proposed method, where β_r is the rth element of $\boldsymbol{\beta}$. We resample the data at the individual level with replacement for L times, and form L samples each having the same size as the original sample, where L is a user-specified number. For $l = 1, \ldots L$, let $\hat{\beta}_{r,l}$ denote the estimator of β_r obtained from the lth resample. Then the bootstrap variance estimate for estimator $\hat{\beta}_r$ is given by

$$\widehat{Var}(\hat{\beta}_r) = \frac{1}{L-1}\sum_{l=1}^{L}\left\{\hat{\beta}_{r,l} - \frac{1}{L}\sum_{l=1}^{L}\hat{\beta}_{r,l}\right\}^2.$$

3.4 Numerical Studies

3.4.1 Simulation Studies

To assess finite sample performance of the three strategies developed in Section 3.3, we consider six estimation methods for $\boldsymbol{\beta}$:

• Method 1, called "Naive/RC", ignores measurement error effects and conducts IPTW estimation in Section 3.1.2 with $X_i(k)$ replaced by X^*_{ik}. This method yields the same results as RC described in Section 3.3.1.

• Method 2 is the indirect SIMEX correction method presented in Section 3.3.2; we use ISIMEX as a short name of this method.

• Method 3 is the direct SIMEX correction method presented in Section 3.3.2; we use DSIMEX as a short name of this method.

•Method 4 is a refined version of the indirect SIMEX, called "RISIMEX". This method substitutes the indirect SIMEX based $\hat{\boldsymbol{\gamma}}^{\dagger}_k$, given by Step 3 of Section 3.3.2, into (3.13) to

produce the IPTW weights, and then produces the causal estimate using Step 2 of Section 3.1.2.

• Method 5 is a refined RC version, called "RRC". This method substitutes the regression calibration based estimator $\hat{\boldsymbol{\gamma}}^*$, given by fitting models (3.4) with $X_i(k)$ replaced by $\hat{X}_i(k)$, into (3.13) to produce IPTW weights, and then yields the causal estimate using Step 2 of Section 3.1.2.

• Method 6, called "RCM", is the refined correction method described in Section 3.3.3.

For all the SIMEX-based methods, set $B = 100$, $\Lambda = \{0, 0.5, 1, 1.5, 2\}$, $M = 5$, and the quadratic regression is invoked in the extrapolation step. The number of bootstrap replicates is set as 1000. Logistic regression models (3.7) are used to estimate the numerator of (3.5). In our simulation studies we consider the case with a single confounder. Therefore, $\boldsymbol{\Sigma}_{\epsilon k}$ in (3.8) represents a variance rather than a covariance matrix, and now we denote it as $\sigma^2_{\epsilon k}$ for clarity.

Sample size of $n = 1000$ is considered and 1000 replications are run for each parameter configuration. The average relative bias in percent (ReBias%), the average bootstrap standard error (ASE), empirical standard error (ESE), mean squared error (MSE), and coverage percentage (CP%) are reported for estimation of β, where the relative bias is calculated as $\dfrac{\hat{\beta} - \beta}{\beta}$, and the coverage percentage is the percentage the 95% confidence intervals $\hat{\beta} \mp 1.96 \times \sqrt{\widehat{Var}(\hat{\beta})}$ which contain the true value β.

To measure the between-simulation variability, we use the Monte Carlo error (MCE) discussed by Koehler et al. (2009). To be specific, MCE refers to the standard deviation of the Monte Carlo estimator taken across hypothetical simulation repetitions, where each simulation is based on the same design and consists of D replications, and D is a user-specific positive integer. MCE for the estimation of β is estimated as $\widehat{MCE} = \dfrac{1}{D}\sqrt{\sum_{d=1}^{D}(\hat{\beta}^{(d)} - \bar{\hat{\beta}})^2}$, where $\bar{\hat{\beta}} = \dfrac{1}{D}\sum_{d=1}^{D}\hat{\beta}^{(d)}$ is the Monte Carlo estimator and $\hat{\beta}^{(d)}$ is the estimator of β in the dth replication. It is immediate that $\widehat{MCE} = \dfrac{\sqrt{D-1}}{D}$ESE, which equals 0.0316ESE when D is taken as 1000, as we consider here. The entries in the ESE column of Table 3.1 show that the Monte Carlo error is small.

We consider two settings. The first setting has two time points (i.e., $K = 1$) and the second setting has three time points (i.e., $K = 2$). The first setting is similar to that of Daniel et al. (2013). For the ith subject, we generate an unmeasured variable $U_i = U_i(0) = U_i(1)$ which follows a Bernoulli distribution with probability 0.3. At visit 0, the treatment $A_i(0)$ is drawn from a Bernoulli distribution with probability 0.5. At visit 1, the confounder $X_i(1)$ is generated from $N(U_i + A_i(0) - 0.8, 1)$, and the treatment $A_i(1)$ is drawn from a Bernoulli distribution with probability $1/[1+\exp\{-\gamma_{01} - \gamma_{X1}X_i(1)\}]$, where γ_{01} and γ_{X1} are parameters. The outcome is generated as

$$Y_i = 2 + \beta\{A_i(0) + A_i(1)\} - U_i,$$

where β is the parameter.

As discussed by Daniel et al. (2013), the data generating mechanism implies that the no unmeasured confounding assumption holds. Taking integration of the outcome model with respect to U gives the marginal structural model

$$E(Y_{\bar{a}}) = 2 + \beta\{a(0) + a(1)\} - 0.3 = 1.7 + \beta\{a(0) + a(1)\},$$

where β is the causal parameter of interest. Set $\beta = 1$, $\gamma_{01} = 0.5$ and $\gamma_{X1} = -1$. Note that although this setting has two time points, the confounder exists only for the time point $k = 1$ but not for $k = 0$. The measurement error model is specified as (3.8) where $\sigma_{\epsilon 1}$ is taken as 0.5 or 1.5.

The simulation results are summarized in Table 3.1. The naive analysis, or the RC method, leads to biased results, and its performance becomes worse as the degree of measurement error increases. We observe RRC produces less biased estimates than RC does, and RISIMEX produces less biased estimates than ISIMEX does. Thus, using formula (3.13) improves the performance of RC and SIMEX methods. Among all the six methods, RCM performs the best as expected, and confirms the consistency established by Theorem 3.1. Except for RCM, all other methods are not exactly consistent. As a result, they produce coverage percentages that are apart from 95% and the situation becomes worse as measurement error increases. The discrepancy between ASE and ESE is fairly small, and

the empirical coverage percentages are in close agreement with 95% for RCM. These results suggest that the bootstrap variance estimates are reliable. Stabilized and unstabilized weights show similar performance in settings we consider.

In the second setting, we consider three time points (i.e., $K = 2$). At visit 0, the confounder $X_i(0)$ is generated from the standard normal distribution $N(0, 1)$, and the treatment $A_i(0)$ is drawn from a Bernoulli distribution with probability $1/[1+\exp\{-\gamma_{X0}X_i(0)\}]$. At visit $k = 1, 2$, the confounder $X_i(k)$ is generated from $N(X_i(k-1)+A_i(k-1)-0.5, 1)$, and the treatment $A_i(k)$ is drawn from a Bernoulli distribution with probability $1/[1 + \exp\{-\boldsymbol{\gamma}_{Ak}^{\mathrm{T}}\bar{A}_i(k-1)-\gamma_{Xk}X_i(k)\}]$. The outcome is generated as $Y_i = -0.5X_i(0)+\beta\{A_i(0)+A_i(1)+A_i(2)\}+e_i$, with e_i generated from $N(0, 1)$.

Taking integration of the outcome model with respect to $X(0)$ and e gives the marginal structural model:

$$E(Y_{\bar{a}}) = \beta\{a(0) + a(1) + a(2)\},$$

where β is the causal parameter of interest. Set $\beta = 1$, $\gamma_{X0} = 0.3$, $\boldsymbol{\gamma}_{A1} = 0.2$, $\gamma_{X1} = 0.3$, $\boldsymbol{\gamma}_{A2} = (0.2, 0.2)^{\mathrm{T}}$, and $\gamma_{X2} = 0.3$. The measurement error model is taken as (3.8) with $\{\sigma_{\epsilon 0}, \sigma_{\epsilon 1}, \sigma_{\epsilon 2}\}$ set to be $\{0.5, 1, 1.5\}$ or $\{1, 1.5, 2\}$.

The simulation results for the second setting are reported in Table 3.2. Examining the ESE column of Table 3.2 shows small Monte Carlo error. Similar to results in Table 3.1, RCM outperforms other methods thanks to the consistency of its estimator. Tables 3.1 and 3.2 also reveal that the naive method may either overestimate or underestimate the causal parameter with a noticeable bias.

3.4.2 Sensitivity Analyses of NHEFS Data

As an application, we use the proposed approaches to analyze the NHEFS data, a national longitudinal study jointly initiated by the National Center for Health Statistics and the National Institute on Aging in collaboration with other agencies of the Public Health Service. The objectives are to understand complex relationships among clinical, nutritional, and behavioral factors measured in the first National Health and Nutrition Examination

Table 3.1: Average relative bias in percent (ReBias%), average bootstrap standard error (ASE), empirical standard error (ESE), mean squared error (MSE) and coverage percentage (CP%) of causal estimates with 2 time points.

		$\sigma_{\epsilon 1} = 0.5$				
Weights	Method	ReBias(%)	ASE	ESE	MSE/10^{-2}	CP%
Unstabilized	Naive/RC	2.189	0.027	0.027	0.119	88.5
	ISIMEX	2.023	0.027	0.027	0.112	87.9
	DSIMEX	0.322	0.029	0.030	0.092	93.9
	RISIMEX	0.481	0.029	0.029	0.086	94.1
	RRC	0.668	0.029	0.029	0.087	93.8
	RCM	-0.206	0.029	0.029	0.083	95.5
Stabilized	Naive/RC	2.137	0.026	0.027	0.118	87.4
	ISIMEX	1.988	0.027	0.027	0.111	87.3
	DSIMEX	0.301	0.029	0.030	0.091	93.8
	RISIMEX	0.463	0.029	0.029	0.089	93.9
	RRC	0.635	0.029	0.029	0.090	93.4
	RCM	-0.197	0.029	0.029	0.085	95.1
		$\sigma_{\epsilon 1} = 1.5$				
Unstabilized	Naive/RC	7.679	0.025	0.025	0.654	14.1
	ISIMEX	8.709	0.024	0.024	0.814	4.30
	DSIMEX	5.400	0.028	0.028	0.368	51.2
	RISIMEX	5.593	0.028	0.028	0.392	50.0
	RRC	2.396	0.041	0.045	0.263	87.2
	RCM	0.521	0.057	0.062	0.390	94.3
Stabilized	Naive/RC	7.463	0.024	0.025	0.618	13.6
	ISIMEX	8.520	0.023	0.023	0.777	3.40
	DSIMEX	5.275	0.027	0.027	0.352	51.8
	RISIMEX	5.485	0.027	0.027	0.374	49.3
	RRC	2.296	0.041	0.045	0.258	85.7
	RCM	0.459	0.058	0.062	0.391	94.1

Table 3.2: Average relative bias in percent (ReBias%), bootstrap standard error (ASE), empirical standard error (ESE), mean squared error (MSE) and coverage percentage (CP%) of causal estimates with 3 time points.

		$\{\sigma_{\epsilon 0}, \sigma_{\epsilon 1}, \sigma_{\epsilon 2}\} = \{0.5, 1, 1.5\}$				
Weights	Method	ReBias(%)	ASE	ESE	MSE/10^{-2}	CP%
Unstabilized	Naive/RC	-3.979	0.038	0.038	0.299	82.2
	ISIMEX	-4.146	0.037	0.037	0.308	80.9
	DSIMEX	-1.568	0.040	0.041	0.193	90.7
	RISIMEX	-1.351	0.040	0.040	0.177	93.2
	RRC	-1.156	0.041	0.041	0.181	93.7
	RCM	0.009	0.043	0.042	0.178	95.3
Stabilized	Naive/RC	-4.014	0.036	0.036	0.290	80.5
	ISIMEX	-4.178	0.036	0.036	0.303	78.9
	DSIMEX	-1.580	0.038	0.040	0.184	91.3
	RISIMEX	-1.382	0.039	0.039	0.167	93.7
	RRC	-1.186	0.039	0.039	0.166	93.9
	RCM	-0.017	0.042	0.041	0.172	95.4
		$\{\sigma_{\epsilon 0}, \sigma_{\epsilon 1}, \sigma_{\epsilon 2}\} = \{1, 1.5, 2\}$				
Unstabilized	Naive/RC	-6.720	0.037	0.038	0.598	56.3
	ISIMEX	-7.855	0.036	0.037	0.756	41.5
	DSIMEX	-3.816	0.039	0.038	0.291	84.3
	RISIMEX	-3.945	0.040	0.040	0.319	81.7
	RRC	-1.782	0.044	0.044	0.225	92.9
	RCM	-0.015	0.049	0.046	0.215	95.9
Stabilized	Naive/RC	-6.784	0.036	0.037	0.596	53.4
	ISIMEX	-7.911	0.035	0.036	0.758	38.8
	DSIMEX	-3.841	0.038	0.037	0.282	82.8
	RISIMEX	-4.008	0.038	0.039	0.314	80.0
	RRC	-1.850	0.042	0.042	0.212	91.2
	RCM	-0.021	0.048	0.046	0.211	95.2

Survey NHANES I and subsequent morbidity, mortality, and hospital utilization, among others. A detailed description of this study is available at

https://wwwn.cdc.gov/nchs/nhanes/nhefs/default.aspx.

The dataset consists of 1624 subjects who were cigarette smokers. The treatment and potential confounders were collected at the baseline (1971-1975) and the follow-up visit (1982). The outcome is the indicator of death by 1992. Here the treatment indicator is defined as the indicator of whether or not being a light smoker, taking value 1 if a subject smoked 10 cigarettes per day or fewer and 0 otherwise. Confounders include age, sex, exercise level, physical activity level, and SBP. The exercise level takes value 1 if a subject exercises regularly and 0 otherwise. The physical activity level is classified as binary, taking value 1 if a subject is active and 0 otherwise. We are interested in studying possible causal effects of the smoking behavior on the risk of death with confounders controlled. In other words, we want to understand the difference between the risk of death that would have been observed had the population consisted of all light smokers and the risk of death that would have been observed had the population consisted of all heavier smokers.

Using the symbols in Section 3.1, for subject i, we let Y_i be the death indicator by 1992; let $A_i(0)$ and $A_i(1)$ be the light smoker indicators at the baseline and the follow-up visit, respectively; let $Z_i(0)$ and $Z_i(1)$ be the vector of age, sex, exercise level, and physical activity level at the baseline and the follow-up visit, respectively. Since SBP involves daily and seasonal biological variabilities (e.g., Carroll et al., 2006, p.13), its measurements at the baseline and the follow-up visit generally differ from its long term average or cumulative average measurements by the visit time. It is thereby important to incorporate such discrepancies in the analysis. According to (Carroll et al., 2006, p.113), we consider the transformation of SBP, $\log(\text{SBP} - 50)$, and assume the measurement error model is (3.8). Let X_{i0}^* and X_{i1}^* be the transformed observed SBP measurements at the baseline and the follow-up visit, respectively; and let $X_i(0)$ and $X_i(1)$ be the transformed cumulative average SBP up to the baseline and the follow-up visit, respectively.

Consider the marginal structural model:

$$\text{logit } P(Y_{\bar{a}} = 1) = \beta_0 + \beta\{a(0) + a(1)\},$$

where β is the causal parameter of interest.

We apply the methods in Section 3.4.1 to analyze the NHEFS data. Confounders measured at time point k are used to model the treatment assignment at time point k. Assumption 5 is perhaps feasible here; if the cumulative average SBP up to time point k can explain the treatment at time point k, then the cumulative average SBP at earlier time point l can be ignored where $l < k$. Measurements of SBP at two visits are likely to have the same variance so we assume $\sigma_{\epsilon 0}$ and $\sigma_{\epsilon 1}$ are equal. Since there is no information on variances $\sigma^2_{\epsilon 0}$ and $\sigma^2_{\epsilon 1}$, we perform sensitivity analyses. We specify $\sigma^2_{\epsilon 0}$ and $\sigma^2_{\epsilon 1}$ as 0.0126, an estimate obtained using the data from Framingham Heart Study by Carroll et al. (2006, p.160) to characterize the degree of measurement error in SBP. Furthermore, we set $\sigma^2_{\epsilon 0}$ and $\sigma^2_{\epsilon 1}$ to be 0.02 to feature a situation with a larger degree of measurement error in SBP.

Table 3.3 summarizes the analysis results. As discussed in Section 3.3.1, RC and the naive analysis produce the same results. When measurement error effects are not accounted for, the estimated β is -0.032 with a 95% confidence interval (-0.185, 0.122) if using unstabilized weights, and is -0.028 with a 95% confidence interval (-0.183, 0.126) if using stabilized weights, both suggesting that the causal effect is statistically insignificant. This conclusion agrees with that reported by Hernán and Robins (2016) who studied causal effect of quitting smoking on the risk of death by 1992. When measurement error in SBP is taken into account, the estimated causal effects obtained from the RCM, RRC, RISIMEX and DSIMEX methods are smaller than the naive estimates. When the measurement error in SBP increases, the causal effect estimates obtained from the RCM, RRC, RISIMEX and DSIMEX methods tend to decrease. Although estimates of the causal effect β are different from method to method, from different degrees of measurement error in SBP, and from using different weights, all the analyses suggest that the causal effect is not statistically significant.

Table 3.3: Sensitivity analyses of NHEFS data with estimated causal effect (EST), bootstrap standard error (SE) and 95% confidence interval (95% CI).

		$\sigma^2_{\epsilon 0} = \sigma^2_{\epsilon 1} = 0.0126$			$\sigma^2_{\epsilon 0} = \sigma^2_{\epsilon 1} = 0.02$		
Weights	Method	EST	SE	95% CI	EST	SE	95% CI
Unstabilized	Naive/RC	-0.032	0.078	(-0.185, 0.122)	-0.032	0.078	(-0.185, 0.122)
	ISIMEX	-0.020	0.079	(-0.174, 0.135)	-0.007	0.078	(-0.161, 0.146)
	DSIMEX	-0.035	0.076	(-0.184, 0.115)	-0.039	0.077	(-0.189, 0.112)
	RISIMEX	-0.037	0.079	(-0.191, 0.117)	-0.041	0.079	(-0.195, 0.113)
	RRC	-0.049	0.079	(-0.204, 0.106)	-0.066	0.081	(-0.224, 0.093)
	RCM	-0.038	0.078	(-0.191, 0.115)	-0.044	0.079	(-0.198, 0.111)
Stabilized	Naive/RC	-0.028	0.079	(-0.183, 0.126)	-0.028	0.079	(-0.183, 0.126)
	ISIMEX	-0.019	0.079	(-0.175, 0.136)	-0.007	0.079	(-0.162, 0.149)
	DSIMEX	-0.031	0.076	(-0.180, 0.118)	-0.035	0.077	(-0.186, 0.116)
	RISIMEX	-0.037	0.079	(-0.192, 0.118)	-0.041	0.079	(-0.196, 0.115)
	RRC	-0.049	0.080	(-0.205, 0.107)	-0.066	0.082	(-0.226, 0.095)
	RCM	-0.034	0.078	(-0.187, 0.119)	-0.040	0.078	(-0.192, 0.113)

Disclaimer: Interpretations and conclusions made by the authors do not reflect the view of National Center for Health Statistics.

Chapter 4

Measurement Error in Outcomes: Bias Analysis and Estimation Methods

This chapter deals with Problem 3 discussed in Section 1.5. In Section 4.1 we describe the IPW estimation framework for error-free settings. In Sections 4.2 and 4.3, we investigate the impact of ignoring measurement error in both continuous and binary outcome variables. As an application, we analyze a real dataset from a clinical trial to examine the effectiveness of a perioperative smoking cessation program (Lee et al., 2013). In Section 4.4 we develop valid estimation procedures to accommodate measurement error effects for practical settings where either validation data or replicates of the outcome variable are available. In Section 4.5 we present simulation results to assess the performance of the proposed methods. To provide protection against model misspecification, in Section 4.6 we propose a doubly robust estimator which is consistent even when either the treatment model or the outcome model is misspecified. In Section 4.7, we extend the proposed methods to accommodating complex misclassification models.

4.1 IPW Estimation in Error-Free Settings

Suppose for an individual, T is the observed binary treatment variable, with $T = 1$ if treated and $T = 0$ if untreated; and X is a vector of pre-treatment covariates. Let Y_1 be the potential outcome that would have been observed had the subject been treated, and Y_0 be the potential outcome that would have been observed had the subject been untreated. Let Y be the observed outcome. We assume fundamental causal inference assumptions described in Section 1.1.2 for the following development.

Our goal is to estimate the average treatment effect (ATE), $E(Y_1) - E(Y_0)$, denoted as τ_0. Suppose we have a sample of size n. For $i = 1, \ldots, n$, let $Y_{i,1}$ be the potential outcome that would have been observed had subject i been treated and $Y_{i,0}$ be the potential outcome that would have been observed had subject i been untreated; T_i be the observed binary treatment variable for subject i; X_i be the observed vector of pre-treatment covariates for subject i; and Y_i be the observed outcome variable for subject i.

Since each individual can only contribute measurement of either Y_1 or Y_0 but not both, estimation of τ_0 cannot be obtained directly based on available measurements of the outcome variables. Using the observed data allows us to estimate the difference of conditional mean outcomes between the treated and untreated groups, $E(Y|T = 1) - E(Y|T = 0)$, which generally differs from ATE τ_0 because of possible imbalance of X in the treated and untreated groups. Rosenbaum and Rubin (1983) initiated the idea of using the propensity score, defined as $e = P(T = 1|X)$, to balance the distribution of X for the treated and untreated groups.

Using propensity scores, Rosenbaum (1998) proposed a consistent estimator of τ_0,

$$\hat{\tau} = \hat{E}(Y_1) - \hat{E}(Y_0), \tag{4.1}$$

where

$$\hat{E}(Y_1) = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}_i}, \ \hat{E}(Y_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{(1 - T_i)Y_i}{1 - \hat{e}_i},$$

and $\hat{e}_i$ is the propensity score for subject i where the associated parameter is replaced by a consistent estimator. Lunceford and Davidian (2004) studied the asymptotic distribution

of $\sqrt{n}(\hat{\tau} - \tau_0)$. When treatment model form for $e = P(T = 1|X)$ and its parameter are known, under regularity conditions, $\sqrt{n}(\hat{\tau} - \tau_0)$ has an asymptotic normal distribution with mean zero and variance

$$V = E\left(\frac{Y_1^2}{e} + \frac{Y_0^2}{1-e}\right) - \tau_0^2. \tag{4.2}$$

The consistency of estimator $\hat{\tau}$ given by (4.1) requires a critical condition that the outcome variable is precisely measured. However, this condition does not always hold in applications. Often, response variable Y_i cannot be measured accurately and is subject to mismeasurement; instead, a surrogate measurement, denoted as Y_i^*, is collected. Ignoring the difference between Y_i^* and Y_i and naively using (4.1) with Y_i replaced by Y_i^* usually yields biased estimation results for τ_0. In the next two sections, we explore the asymptotic bias resulted from mismeasurement in outcome variables.

4.2 IPW Estimation with Mismeasured Continuous Y

In this section, we consider the case where Y_i is a continuous variable. Suppose Y_i^* and Y_i are linked by measurement error model:

$$Y_i^* = Y_i + \alpha_1 T_i \epsilon_1 + \alpha_2 (1 - T_i)\epsilon_2 + g(X_i) + \epsilon_3, \tag{4.3}$$

where ϵ_1, ϵ_2 and ϵ_3 are mutually independent and independent of T_i given X_i, $E(\epsilon_1|X_i) = E(\epsilon_2|X_i) = E(\epsilon_3|X_i) = 0$, and α_1 and α_2 are parameters. Function $g(\cdot)$ reflects possible dependence of Y_i^* on X_i which can be linear or nonlinear; and its form is unknown. If the $g(\cdot)$ function is null, model (4.3) suggests that the surrogate measurement Y_i^* is independent of X_i, given Y_i and T_i.

Formulation (4.3) includes a general class of models in which no distributional assumptions are imposed on ϵ_1, ϵ_2 and ϵ_3. Model (4.3) features a measurement error mechanism that depends on treatment assignment and incorporates possible heterogeneity in measurement structure. For a treated individual, the error term is $\alpha_1\epsilon_1 + g(X_i) + \epsilon_3$; for an untreated individual, the error term is $\alpha_2\epsilon_2 + g(X_i) + \epsilon_3$. When $\alpha_1 = \alpha_2 = 0$, model (4.3) reduces to $Y_i^* = Y_i + g(X_i) + \epsilon_3$, showing that the treated and untreated groups share the

same magnitude of measurement error. When $\alpha_1 = 0, \alpha_2 \neq 0$, a larger measurement error variance is assumed for the untreated group; when $\alpha_2 = 0, \alpha_1 \neq 0$, treatment group has a larger measurement error variance. When $\alpha_1 = \alpha_2 = g(X_i) = 0$, model (4.3) reduces to $Y_i^* = Y_i + \epsilon_3$, the so-called classical additive measurement error model (Carroll et al., 2006) which is perhaps the most widely used measurement error model.

With the unavailability of measurements of Y_i, it is tempting to use the available measurements of Y_i^* to work out an estimator for τ_0 using formulation (4.1). That is, we replace Y_i with Y_i^* in (4.1) and obtain a naive estimator $\hat{\tau}^*$ of τ_0. Since Y_i^* is not necessarily identical to Y_i, one may expect the naive estimator $\hat{\tau}^*$ to incur bias in estimation of τ_0. However, under the additive linear structure for the measurement error model (4.3), we establish the following theorem whose proof is given in Appendix C.1.

Theorem 4.1. *Under the causal inference assumptions described in Section 1.1.2 and model (4.3), naively replacing Y_i with Y_i^* in the IPW estimator (4.1) still yields a consistent estimator, $\hat{\tau}^*$, of τ_0.*

Theorem 4.1 implies that if the measurement error process can be described by (4.3), ignoring measurement error in the analysis can still produce a consistent estimator of τ_0. We stress that the consistency of the naive estimator $\hat{\tau}^*$ relies on the additive linear structure of the measurement error model (4.3) as well as the unit coefficient of Y. If Y^* and Y are linked by a nonlinear model, the naive estimator $\hat{\tau}^*$ would commonly be biased, which is to be demonstrated by the simulation study in Section 4.5. If the coefficient of Y is not 1, even an additive linear error structure cannot guarantee the consistency of the naive estimator. For instance, suppose the measurement error model assumes a form slightly different from (4.3):

$$Y_i^* = \beta Y_i + \alpha_1 T_i \epsilon_1 + \alpha_2 (1 - T_i)\epsilon_2 + g(X_i) + \epsilon_3$$

where β is a coefficient different from 1 and 0; and other quantities follow the same descriptions as in (4.3). By analogy to the proof of Theorem 4.1, we can show that the naive estimator $\hat{\tau}^*$ converges to $\beta\tau_0$ in probability, and hence is a biased estimator of τ_0. In this instance, a consistent estimator of τ_0 can be given by $\hat{\tau}^*/\hat{\beta}$, where $\hat{\beta}$ is a consistent estimator of β which may be obtained from a validation sample.

4.3 IPW Estimation with Mismeasured Binary Y

In this section, we investigate the asymptotic bias induced from misclassification in a binary outcome and propose a closed-form consistent estimator of τ_0. We consider a useful scenario where misclassification probabilities are homogeneous in the sense that

$$P(Y^* = a|Y = b, X, T = t) = P(Y^* = a|Y = b) \tag{4.4}$$

for $a, b, t = 0, 1$. Let $p_{ab} = P(Y^* = a|Y = b)$ for $a, b = 0, 1$.

4.3.1 Estimation Method

Ignoring the difference between Y_i^* and Y_i, the naive analysis uses (4.1) to construct an estimator of τ_0:

$$\hat{\tau}^* = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i^*}{\hat{e}_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i^*}{1-\hat{e}_i}. \tag{4.5}$$

The following theorem establishes the asymptotic bias in the naive estimator whose proof is given in Appendix C.1.

Theorem 4.2. *Suppose the causal inference assumptions described in Section 1.1.2 and model (4.4) hold. Let $\hat{\tau}^*$ denote the naive estimator (4.5) of τ_0. Then*

(a). the asymptotic bias of the naive estimator $\hat{\tau}^*$ is $(p_{11} - p_{10} - 1)\tau_0$;

(b). $\hat{\tau} = \dfrac{\hat{\tau}^*}{p_{11} - p_{10}}$ is a consistent estimator of τ_0 when $p_{11} \neq p_{10}$.

Theorem 4.2(a) implies that in the presence of misclassification, the naive estimator is asymptotically biased and Theorem 4.2(b) offers us a consistent estimator of τ_0 which incorporates the misclassification effects. This estimator is conceptually easy but it requires that the misclassification probabilities are known, a condition which is rather restrictive for application (unless sensitivity analyses are conducted). In Section 4.4, we further develop valid estimation methods for practical settings. The requirement $p_{11} \neq p_{10}$ in Theorem

4.2(b) is often feasible since $P(Y^*=1|Y=1)$ is practically larger than $P(Y^*=1|Y=0)$; otherwise, the collected data are virtually useless.

4.3.2 Asymptotic Distribution

Let $\boldsymbol{\gamma}$ be the parameters for the treatment model, and $\phi(\cdot)$ be an unbiased estimating function of $\boldsymbol{\gamma}$ which is determined by the treatment model. Let $\boldsymbol{\theta}=(\tau,\boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ include all the model parameters, and $\boldsymbol{\theta}_0=(\tau_0,\boldsymbol{\gamma}_0^{\mathrm{T}})^{\mathrm{T}}$ be the true value of $\boldsymbol{\theta}$. Define

$$\Psi(Y_i^*,T_i,X_i;\boldsymbol{\theta})=\left\{\begin{array}{c}\phi(X_i,T_i;\boldsymbol{\gamma})\\ \dfrac{T_iY_i^*}{e_i}-\dfrac{(1-T_i)Y_i^*}{1-e_i}-(p_{11}-p_{10})\tau\end{array}\right\}. \tag{4.6}$$

By the result $E\left(\dfrac{TY^*}{e}\right)-E\left\{\dfrac{(1-T)Y^*}{1-e}\right\}=(p_{11}-p_{10})\tau_0$, shown in Appendix C.1, we can show that $\Psi(Y_i^*,T_i,X_i;\boldsymbol{\theta})$ is an unbiased estimating function of $\boldsymbol{\theta}$. Then solving

$$\sum_{i=1}^{n}\Psi(Y_i^*,T_i,X_i;\boldsymbol{\theta})=\mathbf{0}$$

for $\boldsymbol{\theta}$ yields an estimator of $\boldsymbol{\theta}_0$, denoted as $\hat{\boldsymbol{\theta}}=(\hat{\tau},\hat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$, where $\hat{\tau}=\dfrac{\hat{\tau}^*}{p_{11}-p_{10}}$.

By theory of estimating functions (e.g., Newey and McFadden, 1994; Heyde, 1997; Yi and Reid, 2010; Yi, 2017), we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\xrightarrow{d}N\left(\mathbf{0},A(\boldsymbol{\theta}_0)^{-1}B(\boldsymbol{\theta}_0)A(\boldsymbol{\theta}_0)^{-1\,\mathrm{T}}\right)\text{ as }n\to\infty, \tag{4.7}$$

provided regularity conditions, where $A(\boldsymbol{\theta}_0)=E\left\{-(\partial/\partial\boldsymbol{\theta}^{\mathrm{T}})\Psi(Y^*,T,X;\boldsymbol{\theta}_0)\right\}$, and $B(\boldsymbol{\theta}_0)=E\{\Psi(Y^*,T,X;\boldsymbol{\theta}_0)\Psi^{\mathrm{T}}(Y^*,T,X;\boldsymbol{\theta}_0)\}$. The variance of $\hat{\boldsymbol{\theta}}$ can then be estimated by the empirical sandwich estimator:

$$\hat{V}ar(\hat{\boldsymbol{\theta}})=\frac{1}{n}A_n(\hat{\boldsymbol{\theta}})^{-1}B_n(\hat{\boldsymbol{\theta}})A_n(\hat{\boldsymbol{\theta}})^{-1\,\mathrm{T}}, \tag{4.8}$$

where $A_n(\hat{\boldsymbol{\theta}})$ and $B_n(\hat{\boldsymbol{\theta}})$ are the empirical counterparts of $A(\boldsymbol{\theta}_0)$ and $B(\boldsymbol{\theta}_0)$, respectively.

Let $\hat{V}_{ij}$ be the element of the ith row and the jth column of $\hat{V}ar(\hat{\boldsymbol{\theta}})$. Then the variance estimate, $\hat{V}ar(\hat{\tau})$, of $\hat{\tau}$ is $\hat{V}_{11}$, and a 95% confidence interval is $\hat{\tau} \mp 1.96 \times \sqrt{\hat{V}ar(\hat{\tau})}$.

4.3.3 Efficiency Loss Caused by Misclassification

Theorem 4.2 reveals that misclassification of outcome variable produces biased estimates of ATE. In our causal inference setting, we further show that the misclassification of outcome variable can also lead to loss of efficiency.

When the treatment model and its parameter are known, the estimating function (4.6) is reduced to

$$\psi(Y_i^*, T_i, X_i; \tau) = \frac{T_i Y_i^*}{e_i} - \frac{(1-T_i)Y_i^*}{1-e_i} - (p_{11} - p_{10})\tau,$$

then solving

$$\sum_{i=1}^{n} \psi(Y_i^*, T_i, X_i; \tau) = \mathbf{0}$$

for τ yields an estimator, say $\tilde{\tau}^*$, of τ_0.

By theory of estimating functions, under regularity conditions we have

$$\sqrt{n}(\tilde{\tau}^* - \tau_0) \xrightarrow{d} N(\mathbf{0}, V_{\text{P}}) \text{ as } n \to \infty,$$

where $V_{\text{P}} = A(\tau_0)^{-1} B(\tau_0) \{A(\tau_0)^{-1}\}^{\text{T}}$, $A(\tau_0) = E\left\{-\frac{\partial}{\partial \tau}\psi(Y^*, T, X; \tau_0)\right\} = p_{11} - p_{10}$, and $B(\tau_0) = E\{\psi^2(Y^*, T, X; \tau_0)\}$.

It is reasonable to assume that $p_{11} > p_{10}$, saying that it is more likely to have $Y^* = 1$ when $Y = 1$ than when $Y = 0$. In Appendix C.2, we show $V_{\text{P}} > V$, suggesting that the misclassification reduces efficiency, where V is the asymptotic variance of $\sqrt{n}(\hat{\tau} - \tau_0)$ given by (4.2). This result is similar to that of Neuhaus (1999) who considered a misclassification problem for a non-causal setting.

4.3.4 Application to Smoking Cessation Data

To illustrate our methods, we consider the data arising from a clinical trial to assess the effectiveness of a perioperative smoking cessation program (Lee et al., 2013). One hundred sixty-eight patients were equally randomized to either the treatment group or the control group, where the treatment group was assigned to the smoking cessation intervention and the control group received standard care. Baseline variables include information on physical characteristics, type of surgery, current disease and smoking habits. Specifically, we consider gender, age, body mass index (BMI), diabetes status, hypertension, chronic obstructive pulmonary disease (COPD), cigarettes per day, the number of years of smoking, Fagerström score, and exhaled carbon monoxide (CO) level (ppm). The outcome variable of interest is the smoking cessation status for at least 7 days at the 30-day follow-up postoperatively, which was set to be 1 if the individual had quit smoking, and 0 otherwise.

Lee et al. (2013) pointed out that the follow-up data were self-reported and not confirmed by formal tests. Therefore, the smoking cessation status was subject to misclassification. Magder and Hughes (1997) illustrated that individuals who really had quit smoking were not likely to report that they still smoked, thus it is reasonable to assume $p_{11} = 100\%$. But individuals who still smoked might report that they did not smoke. Lee et al. (2013) collected smoking cessation status for at least 7 days before surgery and the exhaled CO levels, where an exhaled CO $\leq$ 10 ppm confirmed smoking cessation status (SRNT Subcommittee on Biochemical Verification, 2002). This study shows that the number of inaccurate self-reported smoking cessation records with exhaled CO of >10 ppm is 5 in the control group and 6 in the treatment group, and the number of individuals with exhaled CO of >10 ppm is 70 in the control group and 76 in the treatment group. Since the rates in the control and treated groups are close, we roughly treat the inaccurate records to be independent of intervention and pool the data to calculate a misclassification rate as $\frac{5+6}{70+76} = 7.5\%$. Then we conduct sensitivity analysis to see what the estimate of ATE is if the true probability p_{10} is specified as this misclassification rate 7.5%.

We re-analyze the data using the IPW estimation methods we propose. A logistic model is employed to link the treatment indicator and the baseline pre-treatment variables. Hypertension is found to be statistically significant in the logistic model. Our goal is to

estimate the ATE on the smoking cessation status for at least 7 days at the 30-day follow-up postoperatively. Ignoring the misclassification issue of the outcome leads to $\hat{\tau}^* = 0.170$. By using the adjusted estimator in Section 4.3.1, we obtain $\hat{\tau} = \dfrac{\hat{\tau}^*}{p_{11} - p_{10}} = \dfrac{0.170}{1 - 0.075} = 0.184$ with a variance estimate 0.390×10^{-2}, leading to a 95% confidence interval (0.061, 0.306) of τ_0, where τ_0 represents the difference between the smoking cessation rate of the population of individuals who would all have been assigned to the smoking cessation intervention and that of the population of individuals who would all have to receive standard care. This analysis suggests that there is a significant causal effect of the smoking cessation program; with misclassification in the outcome addressed, the smoking cessation intervention is very likely to reduce the smoking rate by at least 6.1% and at most 30.6%.

This analysis also reveals the attenuation effect of outcome misclassification on ATE estimation by noting that $\hat{\tau} = 0.184$ and $\hat{\tau}^* = 0.170$. Using the information on misclassification of smoking cessation from other studies, we observe the same phenomenon. For example, Magder and Hughes (1997) specified $p_{11} = 100\%$ and $p_{10} = 10\%$. Using these misclassification probabilities gives an adjusted estimated ATE $\dfrac{0.170}{1 - 0.10} = 0.189$ with a variance estimate 0.412×10^{-2}; this yields a 95% confidence interval (0.063, 0.314) of τ_0, suggesting a more substantial misclassification effect on estimation of ATE τ_0.

4.4 Estimation with Unknown Misclassification Probabilities

The development in Section 4.3 assumes that p_{11} and p_{10} are known and focuses on estimating τ_0. However, in many settings, p_{11} and p_{10} are unknown and need to be estimated from additional data sources. We consider two useful settings where validation data or replicates of the outcome variable are available.

4.4.1 Using Validation Data

Suppose among the main study subjects $\{1, \ldots, n\}$, there is a simple random internal subsample $\mathcal{V}$ which contains subjects with measurements of variables X, T, Y and Y^*. Let n_{V} be the size of $\mathcal{V}$. We estimate p_{11} and p_{10} using the validation data. Let m_{11} be the number of individuals with $Y_i = 1$ and $Y_i^* = 1$ for $i \in \mathcal{V}$; m_{10} be the number of individuals with $Y_i = 0$ and $Y_i^* = 1$ for $i \in \mathcal{V}$; m_1 be the number of individuals with $Y_i = 1$ for $i \in \mathcal{V}$; and m_0 be the number of individuals with $Y_i = 0$ for $i \in \mathcal{V}$. We then estimate p_{11} and p_{10} by

$$\hat{p}_{11} = \frac{m_{11}}{m_1} \quad \text{and} \quad \hat{p}_{10} = \frac{m_{10}}{m_0}, \tag{4.9}$$

respectively.

Note that $m_{11} = \sum_{i \in \mathcal{V}} Y_i Y_i^*$, $m_1 = \sum_{i \in \mathcal{V}} Y_i$, $m_{10} = \sum_{i \in \mathcal{V}} (1 - Y_i) Y_i^*$ and $m_0 = \sum_{i \in \mathcal{V}} (1 - Y_i)$. Estimators (4.9) are equivalently obtained from solving the estimating equations

$$\sum_{i=1}^{n} g_1(Y_i^*, Y_i; p_{11}) = 0 \quad \text{and} \quad \sum_{i=1}^{n} g_2(Y_i^*, Y_i; p_{10}) = 0,$$

where

$$g_1(Y_i^*, Y_i; p_{11}) = (Y_i Y_i^* - p_{11} Y_i) \cdot I(i \in \mathcal{V}) \cdot \frac{n}{n_{\mathrm{V}}};$$

$$g_2(Y_i^*, Y_i; p_{10}) = \{(1 - Y_i) Y_i^* - p_{10}(1 - Y_i)\} \cdot I(i \in \mathcal{V}) \cdot \frac{n}{n_{\mathrm{V}}}.$$

Let $\boldsymbol{\theta} = (\tau, \boldsymbol{\gamma}^{\mathrm{T}}, p_{11}, p_{10})^{\mathrm{T}}$. We describe three estimation methods for $\boldsymbol{\theta}$. The first method uses only validation data $\{(Y_i, X_i, T_i) : i \in \mathcal{V}\}$ to estimate τ_0. Let $\hat{\tau}_{\mathrm{V}}$ denote the resulting estimator:

$$\hat{\tau}_{\mathrm{V}} = \frac{1}{n_{\mathrm{V}}} \sum_{i \in \mathcal{V}} \frac{T_i Y_i}{\hat{e}_i} - \frac{1}{n_{\mathrm{V}}} \sum_{i \in \mathcal{V}} \frac{(1 - T_i) Y_i}{1 - \hat{e}_i}. \tag{4.10}$$

Although this approach is valid, it incurs efficiency loss, which can be large when the sample size of validation data is small. The second approach is to employ the correction method indicated by Theorem 4.2(b) using the non-validation data only, with misclassification

probabilities estimated from the validation data. Let $\hat{\tau}_{\rm N}$ denote the resulting estimator:

$$\hat{\tau}_{\rm N} = \frac{1}{\hat{p}_{11} - \hat{p}_{10}} \left\{ \frac{1}{n - n_{\rm V}} \sum_{i \notin \mathcal{V}} \frac{T_i Y_i^*}{\hat{e}_i} - \frac{1}{n - n_{\rm V}} \sum_{i \notin \mathcal{V}} \frac{(1 - T_i) Y_i^*}{1 - \hat{e}_i} \right\}. \tag{4.11}$$

The limitation of this method is that the validation data is not used. The third approach is to combine the validation data and non-validation data into the estimating function for τ. Let

$$\begin{aligned} g_\tau^{(w)}(Y_i^*, T_i, X_i, Y_i; \tau) &= wI(i \in \mathcal{V}) \left\{ \frac{T_i Y_i}{e_i} - \frac{(1 - T_i) Y_i}{1 - e_i} \right\} \\ &+ \frac{(1 - w) I(i \notin \mathcal{V})}{p_{11} - p_{10}} \left\{ \frac{T_i Y_i^*}{e_i} - \frac{(1 - T_i) Y_i^*}{1 - e_i} \right\} - \delta\tau \end{aligned} \tag{4.12}$$

where w is a weight between 0 and 1, and

$$\delta = wE\{I(i \in \mathcal{V})\} + (1 - w)E\{I(i \notin \mathcal{V})\} = \frac{wn_{\rm V}}{n} + \frac{(1 - w)(n - n_{\rm V})}{n}.$$

Define

$$\Psi(Y_i^*, T_i, X_i, Y_i; \boldsymbol{\theta}) = \begin{pmatrix} \phi(X_i, T_i; \boldsymbol{\gamma}) \\ g_1(Y_i^*, Y_i; p_{11}) \\ g_2(Y_i^*, Y_i; p_{10}) \\ g_\tau^{(w)}(Y_i^*, T_i, X_i, Y_i; \tau) \end{pmatrix},$$

then $\Psi(Y_i^*, T_i, X_i, Y_i; \boldsymbol{\theta})$ is an unbiased estimating function of $\boldsymbol{\theta}$. Under regularity conditions, solving

$$\sum_{i=1}^{n} \Psi(Y_i^*, T_i, X_i, Y_i; \boldsymbol{\theta}) = \mathbf{0}$$

for $\boldsymbol{\theta}$ yields a consistent estimator of $\boldsymbol{\theta}_0$. Let $\tilde{\tau}(w)$ denote the corresponding estimator of τ_0 which may depend on the weight w.

We comment that the estimators of τ_0 obtained from the three methods are related in a linear form:

$$\tilde{\tau}(w) = \frac{wn_{\rm V}}{wn_{\rm V} + (1 - w)(n - n_{\rm V})} \hat{\tau}_{\rm V} + \left\{ 1 - \frac{wn_{\rm V}}{wn_{\rm V} + (1 - w)(n - n_{\rm V})} \right\} \hat{\tau}_{\rm N}. \tag{4.13}$$

With weight $w = 1$, (4.13) recovers $\hat{\tau}_{\text{V}}$, which is also evidenced by the formulation of (4.12) where only the validation data are used. On the other hand, with $w = 0$, (4.13) recovers $\hat{\tau}_{\text{N}}$, and this is also suggested by (4.12) that only the non-validation data are used for estimation of τ_0. If $w = 1/2$, then (4.13) gives $\tilde{\tau}(w) = \frac{n_{\text{V}}}{n}\hat{\tau}_{\text{V}} + \frac{n - n_{\text{V}}}{n}\hat{\tau}_{\text{N}}$, which is obtained by placing equal weights on validation and non-validation data; this is the method that has been widely used in the literature for studies with validation data in addition to main study data (e.g., Yi et al., 2015a; Spiegelman et al., 2000). Let $\tilde{\tau}(0.5)$ denote the estimator with $w = 1/2$. Although the equal weight method of obtaining $\tilde{\tau}(0.5)$ is universally used in practice, this approach does not necessarily yield the most efficient estimator for τ_0.

Now we discuss an optimal choice of the weight w so that the resultant estimator $\tilde{\tau}(w)$ has the smallest variance among all the estimators of form (4.13). To do this, consider any linear combination:

$$\hat{\tau}(c) = c\hat{\tau}_{\text{V}} + (1 - c)\hat{\tau}_{\text{N}},$$

where c is a constant between 0 and 1. Noting that

$$\begin{aligned} Var\{\hat{\tau}(c)\} = &\{Var(\hat{\tau}_{\text{V}}) + Var(\hat{\tau}_{\text{N}}) - 2Cov(\hat{\tau}_{\text{V}}, \hat{\tau}_{\text{N}})\}c^2 \\ &-\{2Var(\hat{\tau}_{\text{N}}) - 2Cov(\hat{\tau}_{\text{V}}, \hat{\tau}_{\text{N}})\}c + Var(\hat{\tau}_{\text{N}}), \end{aligned}$$

we minimize $Var\{\hat{\tau}(c)\}$ at

$$c_{\text{OPT}} = \frac{Var(\hat{\tau}_{\text{N}}) - Cov(\hat{\tau}_{\text{V}}, \hat{\tau}_{\text{N}})}{Var(\hat{\tau}_{\text{V}}) + Var(\hat{\tau}_{\text{N}}) - 2Cov(\hat{\tau}_{\text{V}}, \hat{\tau}_{\text{N}})}, \tag{4.14}$$

therefore, the estimator $c_{\text{OPT}}\hat{\tau}_{\text{V}} + (1 - c_{\text{OPT}})\hat{\tau}_{\text{N}}$, denoted as $\hat{\tau}_{\text{OPT}}$, is the best estimator among the linear combination estimators of form (4.13). Using linear combinations of consistent estimators is a useful method to find an estimator with improved efficiency, see, for example, Yi and He (2006) and Braun et al. (2016).

Since c_{OPT} is unknown, in actual implementation and we estimate it by

$$\hat{c}_{\text{OPT}} = \frac{\hat{V}ar(\hat{\tau}_{\text{N}}) - \hat{C}ov(\hat{\tau}_{\text{V}}, \hat{\tau}_{\text{N}})}{\hat{V}ar(\hat{\tau}_{\text{V}}) + \hat{V}ar(\hat{\tau}_{\text{N}}) - 2\hat{C}ov(\hat{\tau}_{\text{V}}, \hat{\tau}_{\text{N}})},$$

where $\hat{V}ar(\hat{\tau}_{\rm N})$, $\hat{C}ov(\hat{\tau}_{\rm V}, \hat{\tau}_{\rm N})$ and $\hat{V}ar(\hat{\tau}_{\rm V})$ are the estimates for $Var(\hat{\tau}_{\rm N})$, $Cov(\hat{\tau}_{\rm V}, \hat{\tau}_{\rm N})$ and $Var(\hat{\tau}_{\rm V})$ obtained by combing the estimating functions constructed from the first two methods. The details are given in Appendix C.3.

We comment that the variances of $\hat{\tau}_{\rm V}$ and $\hat{\tau}_{\rm N}$ and the covariance between $\hat{\tau}_{\rm V}$ and $\hat{\tau}_{\rm N}$ are constrained with

$$\begin{aligned} & Var(\hat{\tau}_{\rm V}) + Var(\hat{\tau}_{\rm N}) - 2Cov(\hat{\tau}_{\rm V}, \hat{\tau}_{\rm N}) \\ \geq \; & 2\sqrt{Var(\hat{\tau}_{\rm V})Var(\hat{\tau}_{\rm N})} - 2\sqrt{Var(\hat{\tau}_{\rm V})Var(\hat{\tau}_{\rm N})}Cor(\hat{\tau}_{\rm V}, \hat{\tau}_{\rm N}) \geq 0, \end{aligned}$$

and $0 \leq c \leq 1$, but the empirical estimates $\hat{V}ar(\hat{\tau}_{\rm V}) + \hat{V}ar(\hat{\tau}_{\rm N}) - 2\hat{C}ov(\hat{\tau}_{\rm V}, \hat{\tau}_{\rm N})$ and $\hat{c}_{\rm OPT}$ do not necessarily satisfy these constraints. If this happens, we set $\hat{c}_{\rm OPT}$ to be 0 or 1, and $\hat{\tau}_{\rm OPT}$ is set to be either $\hat{\tau}_{\rm V}$ or $\hat{\tau}_{\rm N}$, whichever has a smaller variance. Let $\hat{\tau}_{\rm OPT}$ be the resulting optimal linear combination estimator.

4.4.2 Using Replicates

In some settings, the validation data are not available, but there are replicates of Y_i. White et al. (2001) proposed methods to estimate misclassification probabilities with replicates, and we utilize their idea here. Without loss of generality, we illustrate the case where two independent replicates of Y_i are available. Situations with more replicates can be discussed similarly.

Let $\eta = P(Y = 1)$ and let π_r be the probability of obtaining r observations of Y=1 for $r = 0, 1, 2$. Then

$$\pi_0 = \eta(1 - p_{11})^2 + (1 - \eta)(1 - p_{10})^2; \tag{4.15}$$

$$\pi_1 = 2\eta(1 - p_{11})p_{11} + 2(1 - \eta)(1 - p_{10})p_{10}; \tag{4.16}$$

$$\pi_2 = \eta p_{11}^2 + (1 - \eta)p_{10}^2. \tag{4.17}$$

The probability π_r can be estimated by $\hat{\pi}_r = d_r/n$, where d_r is the number of individuals with r observations of Y=1, and $r = 0, 1, 2$. Since $\pi_0 + \pi_1 + \pi_2 = 1$, p_{11}, p_{10} and η cannot be identified from (4.15), (4.16) and (4.17). To get around this problem, we invoke a common strategy which imposes certain constraints on the parameters (e.g., White et al., 2001).

We discuss several useful constraints. The first constraint is $p_{10} = p_{01}$, which assumes the same misclassification probability regardless of the value of Y (i.e., the sensitivity equals specificity). The second constraint is $p_{11} = 1$, which says that the outcome for those subjects with $Y = 1$ is always correctly measured. For example, it is reasonable to assume $p_{11} = 1$ in smoking cessation studies because individuals who quit smoking (i.e., $Y = 1$) are unlikely to report that they still smoke. In contrast, the third constraint is $p_{00} = 1$, which states that the outcome for those subjects with $Y = 0$ is always accurately measured. Other constraints may be considered as well. For example, by a prior knowledge, one may take the prevalence η as known. In principle, introducing suitable constraints is to reduce the parameter space to exclude those inadmissible parameter values, and a specific form of constraints is largely driven by the feature of an individual problem.

Let $Y_i^*(1)$ and $Y_i^*(2)$ be the two replicates of Y_i. Noting that $d_0 = \sum_{i=1}^n [\{1-Y_i^*(1)\}\{1-Y_i^*(2)\}]$, $d_1 = \sum_{i=1}^n [Y_i^*(1)\{1-Y_i^*(2)\} + Y_i^*(2)\{1-Y_i^*(1)\}]$ and $d_2 = \sum_{i=1}^n \{Y_i^*(1)Y_i^*(2)\}$, we obtain that estimation of η, p_{11} and p_{10} using (4.15), (4.16) and (4.17) is equivalent to solving the estimating equations

$$\begin{aligned} &\sum_{i=1}^n h_1(Y_i^*(1), Y_i^*(2); \eta, p_{11}, p_{10}) = 0; \\ &\sum_{i=1}^n h_2(Y_i^*(1), Y_i^*(2); \eta, p_{11}, p_{10}) = 0, \end{aligned} \tag{4.18}$$

where

$$h_1(Y_i^*(1), Y_i^*(2); \eta, p_{11}, p_{10}) = \{1 - Y_i^*(1)\} \cdot \{1 - Y_i^*(2)\} - \pi_0;$$

$$h_2(Y_i^*(1), Y_i^*(2); \eta, p_{11}, p_{10}) = Y_i^*(1) \cdot \{1 - Y_i^*(2)\} + Y_i^*(2) \cdot \{1 - Y_i^*(1)\} - \pi_1.$$

Consequently, to estimate $\boldsymbol{\theta} = (\tau, \boldsymbol{\gamma}^{\mathrm{T}}, \eta, p_{11}, p_{10})^{\mathrm{T}}$, we need to combine estimating functions in (4.18) with estimating functions for $\boldsymbol{\gamma}$ and τ. That is, set

$$\Psi\{Y_i^*(1), Y_i^*(2), T_i, X_i; \boldsymbol{\theta}\} = \left\{ \begin{array}{c} \phi(X_i, T_i; \boldsymbol{\gamma}) \\ h_1(Y_i^*(1), Y_i^*(2); \eta, p_{11}, p_{10}) \\ h_2(Y_i^*(1), Y_i^*(2); \eta, p_{11}, p_{10}) \\ \dfrac{T_i Y_i^*}{e_i} - \dfrac{(1-T_i)Y_i^*}{1-e_i} - (p_{11}-p_{10})\tau \end{array} \right\},$$

where $Y_i^* = \{Y_i^*(1) + Y_i^*(2)\}/2$, together with the constraint imposed for parameter identifiability.

Let $\hat{\tau}_{\text{R}}$ denote the estimator of τ_0 resulted from solving $\sum_{i=1}^n \Psi\{Y_i^*(1), Y_i^*(2), T_i, X_i; \boldsymbol{\theta}\} = \mathbf{0}$ for $\boldsymbol{\theta}$. The variance of $\hat{\tau}_{\text{R}}$ can be obtained by the same manner as described in Section 4.3.2.

4.5 Simulation Studies

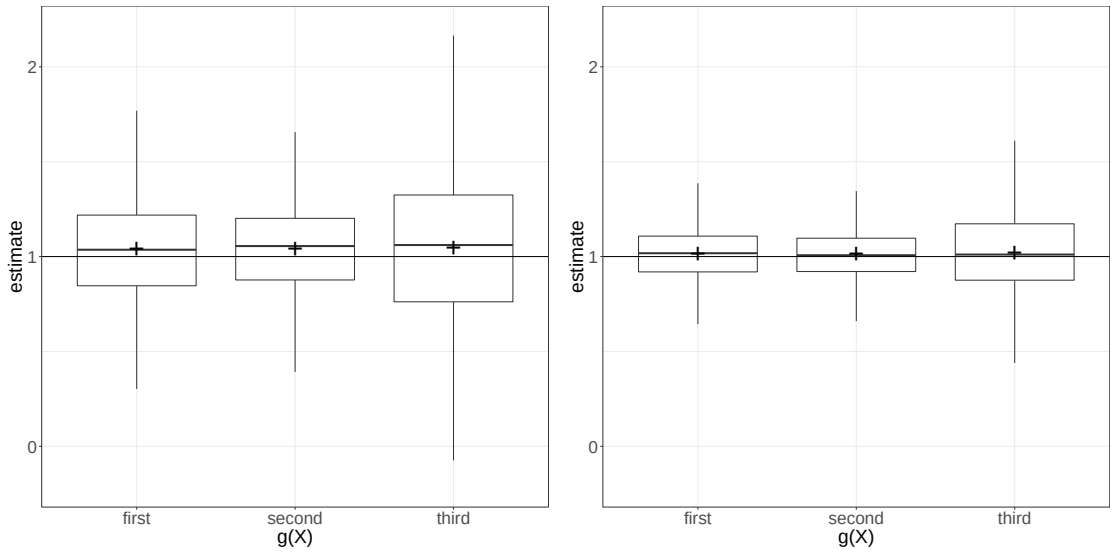
In this section, we conduct simulation studies to demonstrate the theoretical results established in Sections 4.2, 4.3 and 4.4.

4.5.1 Continuous Outcome

Let $X = (X_1, X_2, X_3)^{\text{T}}$, where X_1, X_2 and X_3 are independently generated from a standard normal distribution. The treatment T is drawn from a Bernoulli distribution with probability $1/\{1 + \exp(-0.2 - X_1 - X_2 - X_3)\}$. Let $Y = T + X_1 + X_2 + X_3^2 + Z$, where Z follows a standard normal distribution. This implies $\tau_0 = 1$, since

$$E(Y_1) - E(Y_0) = E(1 + X_1 + X_2 + X_3^2 + Z) - E(0 + X_1 + X_2 + X_3^2 + Z) = 1.$$

Surrogate measurement Y^* is generated from model (4.3), where $\alpha_1 = \alpha_2 = 1$. Function $g(X)$ is specified as 0, $\min(X_1, X_2, X_3)$, or $1/(X_1^2 + 1) + |X_2 + X_3|$, respectively. Let $\epsilon_1 = Z_1$, $\epsilon_2 = Z_2$ and $\epsilon_3 = Z_3 - 1$, where Z_1 follows a standard normal distribution, Z_2 follows a uniform distribution ranging from -1 to 1, and Z_3 is generated from a unit exponential distribution.



$n = 1000$ $n = 5000$

Figure 4.1: The performance of the naive estimator when a continuous outcome variable is subject to additive measurement error given by model (4.3) with $g(X)$ specified by three forms:

first: $g(X) = 0$*; second:* $g(X) = min(X_1, X_2, X_3)$*; third:* $g(X) = \dfrac{1}{X_1^2 + 1} + |X_2 + X_3|$.
$+$*: mean of the 1000 naive estimates.*

We consider two sample sizes with $n = 1000$ and $n = 5000$, and 1000 simulations are run for each $g(X)$. The box plots of estimates of naive analysis under each specification $g(X)$ are displayed in Figure 4.1. A horizontal line is drawn in Figure 4.1 to indicate the true value $\tau_0 = 1$. The empirical mean of the estimates is indicated by $''+''$. The box plots reveal that the naive analysis produces fairly small empirical bias despite the choice of $g(X)$ and the empirical bias becomes smaller as sample size increases, which confirms Theorem 4.1.

The consistency of the naive estimator $\hat{\tau}^*$ relies on the additive linear structure of model (4.3) as well as the unit coefficient of Y, as discussed in Section 4.2. Next we explore the effect of nonlinear measurement error models. Let X be independently generated from the

standard normal distribution. The treatment T is generated from a Bernoulli distribution with probability $1/\{1+\exp(-0.2-X)\}$. Let $Y = \tau_0 T + X + Z$, where τ_0 is the causal effect of interest and Z follows a standard normal distribution. We set

$$Y^* = \exp(\beta Y) + \epsilon, \tag{4.19}$$

where β takes values in $\{-0.1, -0.5, -1, 0.1, 0.5, 1\}$, and ϵ is independent of $\{Y, T, X\}$ and has mean 0. Noting that

$$E\left(\frac{T\epsilon}{e}\right) - E\left\{\frac{(1-T)\epsilon}{1-e}\right\} = E\left\{\frac{1}{e}E\left(T|X\right)\right\}E(\epsilon) - E\left\{\frac{1}{1-e}E\left(1-T|X\right)\right\}E(\epsilon) = 0,$$

which implies that ϵ does not cause bias for the estimation of τ_0, we consider only that ϵ follows a standard normal distribution in simulation.

We consider sample size $n = 1000$, and 1000 simulations are run for each β. Figure 4.2 displays the average bias of 1000 naive estimates versus true value τ_0 with τ_0 varying from -1 to 1, where different values of β are considered. The bias increases as the absolute value of τ_0 increases. The direction of bias depends on values of τ_0 and β. These simulation results reveal that when a continuous outcome is subject to nonlinear measurement error, naively ignoring the difference between Y_i and Y_i^* may result in seriously biased estimates of τ_0.

4.5.2 Binary Outcome with Known Misclassification Probabilities

Let X be generated from a standard normal distribution. Treatment T is drawn from a Bernoulli distribution with probability $1/\{1+\exp(-0.2-X)\}$. Outcome Y is drawn from a Bernoulli distribution with probability $1/\{1+\exp(-0.2-T-X)\}$. Model (4.4) is specified with (p_{11}, p_{10}) set to be $(0.9, 0.1)$, $(0.8, 0.2)$ or $(0.7, 0.3)$.

Unlike the linear model which gives the value of τ_0 directly, a logistic model for the binary outcome does not explicitly show τ_0 as its parameter. The coefficient of T in the outcome model is a conditional effect rather than a marginal or causal effect. To obtain

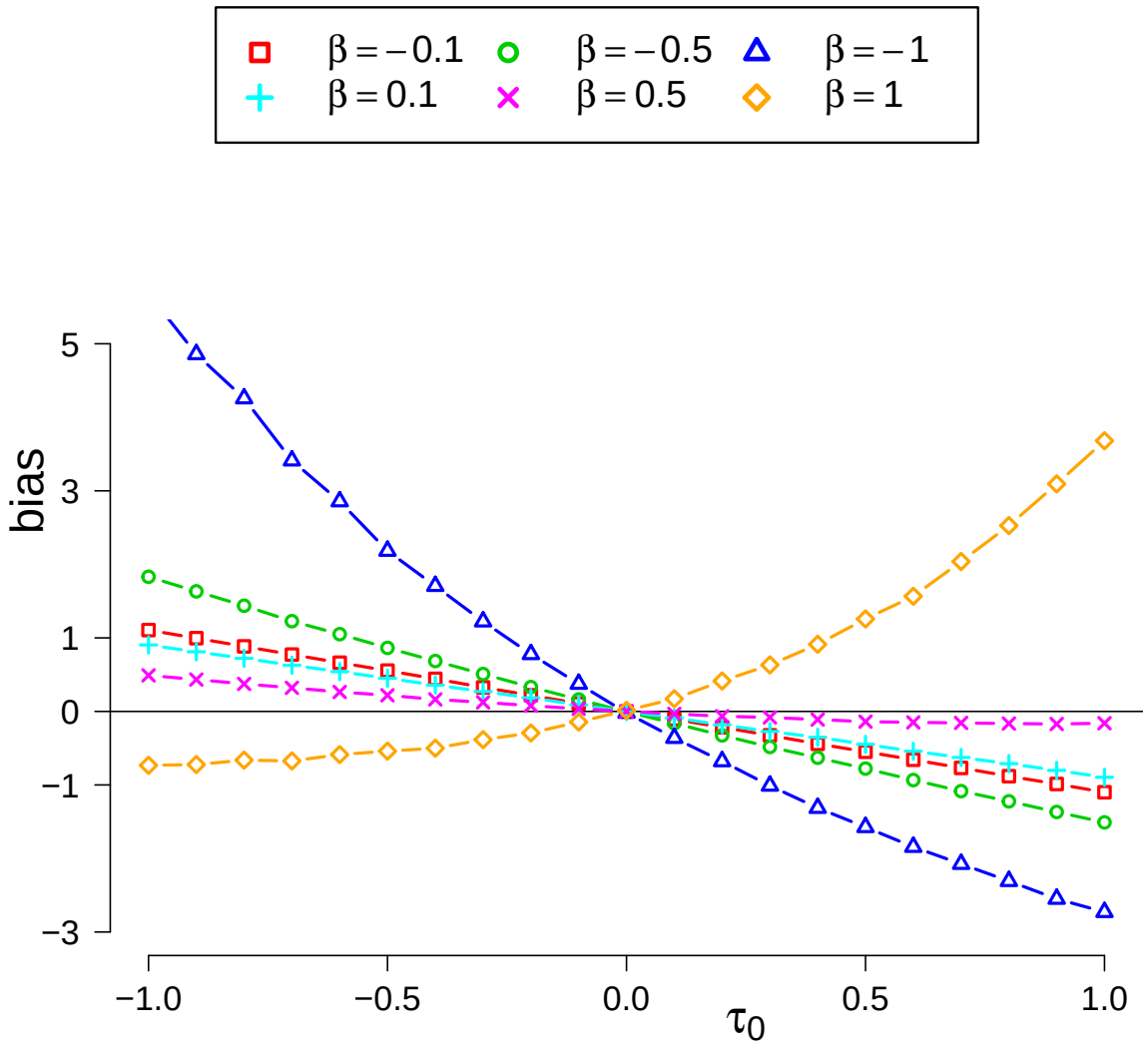


Figure 4.2: The performance of the naive estimator when a continuous outcome variable is subject to nonlinear measurement error given by model (4.19)

the value of τ_0 from the given model, one strategy is to simulate a large size of data to directly approximate $\tau_0 = P(Y_1 = 1) - P(Y_0 = 1) = \int_{-\infty}^{\infty} P(Y = 1|T = 1, X = x)f(x)dx - \int_{-\infty}^{\infty} P(Y = 1|T = 0, X = x)f(x)dx$ by $\frac{1}{N}\sum_{i=1}^{N}\{P(Y_i = 1|T_i = 1, X_i) - P(Y_i = 1|T_i = 0, X_i)\}$, where N is a sufficiently large number and $f(x)$ is the density of X. When N is large enough, the approximated value of τ_0 is almost identical to τ_0. This simulation approach was described in detail by Austin (2007). Here to obtain τ_0, we use a sample size of 50000 and replicate the process for 5000 times, and then obtain the average, which gives $\tau_0 = 0.190$.

We consider two sample sizes with $n = 1000$ and $n = 5000$, and 5000 simulations are run for each pair of (p_{11}, p_{10}). For both the corrected estimator $\hat{\tau}$ and the naive estimator $\hat{\tau}^*$, the average relative bias (ReBias), average sandwich standard error (ASE), empirical standard error (ESE) and 95% coverage percentage (CP%) are reported. For $\hat{\vartheta}$ representing $\hat{\tau}$ or $\hat{\tau}^*$, the relative bias is calculated as $\frac{\hat{\vartheta} - \tau_0}{\tau_0}$, the coverage percentage is the percentage the 95% confidence intervals $\hat{\vartheta} \mp 1.96 \times \sqrt{\hat{V}ar(\hat{\vartheta})}$ which contain τ_0.

Table 4.1 summarizes the simulation results under various misclassification probabilities. The naive analysis leads to severely biased results, and its performance becomes worse as the degree of misclassification increases. The corrected estimates demonstrate satisfactory performance in terms of bias, as assured by Theorem 4.2. The discrepancy between ASE and ESE is fairly small, and empirical coverage percentages are close to 95%, indicating that the sandwich variance estimates are reliable. By Theorem 4.2, the asymptotic relative bias (ReBias) of the naive estimator $\hat{\tau}^*$ equals $p_{11} - p_{10} - 1$. As a result, for the specification of $(p_{11}, p_{10}) = (0.9, 0.1)$, $(0.8, 0.2)$, and $(0.7, 0.3)$, theoretical asymptotic relative biases are $0.9 - 0.1 - 1 = -0.2$, $0.8 - 0.2 - 1 = -0.4$ and $0.7 - 0.3 - 1 = -0.6$, respectively. Empirical results in Table 4.1 support our expectation.

4.5.3 Binary Outcome with Validation Data

We use the same setting of Section 4.5.2 except that an internal validation sample is assumed, and (p_{11}, p_{10}) is set as $(0.8, 0.2)$ or $(0.7, 0.3)$. Take the sample size ratio of validation data and main data to be $n_{\mathrm{v}}/n = 20\%$, 40% or 60%. For four corrected estimators $\hat{\tau}_{\mathrm{v}}$,

Table 4.1: Simulation results for a binary outcome misclassified with known misclassification probabilities: the performance of the proposed IPW estimator $\hat{\tau}$ as opposed to the performance of the naive estimator $\hat{\tau}^*$

setting		$n = 1000$				$n = 5000$			
(p_{11}, p_{10})	Est.	ReBias	ASE	ESE	CP(%)	ReBias	ASE	ESE	CP(%)
$(0.9, 0.1)$	$\hat{\tau}^*$	-0.196	0.036	0.036	82.7	-0.198	0.016	0.016	35.1
	$\hat{\tau}$	0.006	0.045	0.045	95.1	0.002	0.020	0.020	95.4
$(0.8, 0.2)$	$\hat{\tau}^*$	-0.397	0.037	0.037	46.5	-0.401	0.017	0.017	0.50
	$\hat{\tau}$	0.005	0.062	0.061	95.3	-0.002	0.028	0.028	94.7
$(0.7, 0.3)$	$\hat{\tau}^*$	-0.603	0.038	0.038	13.9	-0.600	0.017	0.017	0.00
	$\hat{\tau}$	-0.007	0.094	0.095	94.6	0.001	0.042	0.043	94.6

Est.: estimator; ReBias: average relative bias; ASE: average sandwich standard error; ESE: empirical standard error; CP%: 95% *coverage percentage.*

$\hat{\tau}_{\text{N}}$, $\tilde{\tau}(0.5)$ and $\hat{\tau}_{\text{OPT}}$, the average relative bias (ReBias), average sandwich standard error (ASE), empirical standard error (ESE), 95% coverage percentage (CP%) and the average estimated relative efficiency (ARE) are reported, where the estimated relative efficiency of estimate $\hat{\vartheta}$ is calculated as $\hat{V}ar(\hat{\tau}_{\text{OPT}})/\hat{V}ar(\hat{\vartheta})$ for $\hat{\vartheta}$ set as $\hat{\tau}_{\text{V}}$, $\hat{\tau}_{\text{N}}$ or $\tilde{\tau}(0.5)$.

Table 4.2 summarizes the simulation results for the four methods described in Section 4.4.1 and the naive estimator $\hat{\tau}^*$, given by

$$\hat{\tau}^* = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i\tilde{Y}_i}{\hat{e}_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)\tilde{Y}_i}{1-\hat{e}_i}$$

where $\tilde{Y}_i = I(i \in \mathcal{V})Y_i + I(i \notin \mathcal{V})Y_i^*$. The naive analysis leads to severely biased results, and its performance becomes worse as the degree of misclassification increases. On the contrary, the four corrected estimators $\hat{\tau}_{\text{V}}$, $\hat{\tau}_{\text{N}}$, $\tilde{\tau}(0.5)$ and $\hat{\tau}_{\text{OPT}}$ all present satisfactory performance with fairly small empirical bias and empirical coverage percentages close to 95%. In terms of efficiency, the variance of $\hat{\tau}_{\text{V}}$ can be larger or smaller than that of $\hat{\tau}_{\text{N}}$, depending on the misclassification probabilities and the validation sample size. The estimator based on a linear combination of $\hat{\tau}_{\text{V}}$ and $\hat{\tau}_{\text{N}}$, $\hat{\tau}_{\text{OPT}}$, shows the best efficiency, as expected.

Figure 4.3 displays the average of 5000 estimated relative efficiency (ARE) of $\hat{\tau}_{\text{V}}$, $\hat{\tau}_{\text{N}}$ and $\tilde{\tau}(0.5)$ under various misclassification probabilities. As shown in Figure 4.3, $\hat{\tau}_{\text{OPT}}$ has the best efficiency. The ARE of $\hat{\tau}_{\text{V}}$ increases as the misclassification probabilities increase, while the ARE of $\hat{\tau}_{\text{N}}$ decreases as the misclassification probabilities increase. The ARE of $\hat{\tau}_{\text{V}}$ increases as the validation sample size increases, while the ARE of $\hat{\tau}_{\text{N}}$ decreases as the validation sample size increases. The ARE of $\tilde{\tau}(0.5)$ decreases as the misclassification probabilities increase, suggesting that the advantage of $\hat{\tau}_{\text{OPT}}$ compared to $\tilde{\tau}(0.5)$ is more significant under more substantial misclassification.

4.5.4 Binary Outcome with Replicates

We use the same setting of Section 4.5.2 except that two independent replicates of Y_i are available for each individual i. The extra assumption we use is $p_{11} = 1 - p_{10}$, where the sensitivity equals the specificity. For the corrected estimator, the average relative bias

Table 4.2: Simulation results with validation data available: the performance of proposed estimators $\hat{\tau}_{\mathrm{V}}$, $\hat{\tau}_{\mathrm{N}}$, $\tilde{\tau}(0.5)$ and $\hat{\tau}_{\mathrm{OPT}}$ compared to the naive estimator $\hat{\tau}^*$

			$p_{11}=0.8,\ p_{10}=0.2$					$p_{11}=0.7,\ p_{10}=0.3$				
n	$\frac{n_{\mathrm{V}}}{n}$	Est.	ReBias	ASE	ESE	CP%	ARE%	ReBias	ASE	ESE	CP%	ARE%
1000	20%	$\hat{\tau}^*$	-0.313	0.037	0.037	63.5	-	-0.475	0.037	0.038	32.7	-
		$\hat{\tau}_{\mathrm{V}}$	-0.001	0.121	0.123	94.9	21.2	0.022	0.121	0.121	94.9	35.8
		$\hat{\tau}_{\mathrm{N}}$	0.026	0.081	0.080	96.0	46.5	0.029	0.127	0.128	96.3	33.7
		$\tilde{\tau}(0.5)$	0.020	0.061	0.061	95.9	81.5	0.027	0.097	0.099	96.4	57.9
		$\hat{\tau}_{\mathrm{OPT}}$	0.019	0.055	0.055	95.3	-	0.014	0.071	0.071	95.2	-
	40%	$\hat{\tau}^*$	-0.237	0.036	0.036	76.4	-	-0.361	0.037	0.037	54.0	-
		$\hat{\tau}_{\mathrm{V}}$	-0.001	0.079	0.079	95.2	35.8	0.002	0.079	0.080	94.6	49.8
		$\hat{\tau}_{\mathrm{N}}$	0.013	0.099	0.100	95.1	23.0	0.007	0.151	0.152	95.4	14.3
		$\tilde{\tau}(0.5)$	0.007	0.055	0.055	95.9	75.3	0.005	0.084	0.084	95.9	46.5
		$\hat{\tau}_{\mathrm{OPT}}$	0.010	0.047	0.047	95.0	-	0.002	0.056	0.055	95.1	-
	60%	$\hat{\tau}^*$	-0.156	0.036	0.036	86.7	-	-0.240	0.036	0.036	75.6	-
		$\hat{\tau}_{\mathrm{V}}$	0.004	0.059	0.060	94.8	51.2	-0.001	0.059	0.058	95.2	62.3
		$\hat{\tau}_{\mathrm{N}}$	0.010	0.130	0.133	94.6	10.6	0.012	0.195	0.197	95.4	5.94
		$\tilde{\tau}(0.5)$	0.006	0.049	0.050	94.6	74.2	0.004	0.072	0.072	95.2	43.7
		$\hat{\tau}_{\mathrm{OPT}}$	0.009	0.042	0.043	94.7	-	0.000	0.046	0.046	95.5	-
5000	20%	$\hat{\tau}^*$	-0.319	0.016	0.016	4.36	-	-0.482	0.017	0.017	0.02	-
		$\hat{\tau}_{\mathrm{V}}$	0.002	0.054	0.056	94.7	20.4	0.000	0.055	0.056	94.3	34.4
		$\hat{\tau}_{\mathrm{N}}$	0.004	0.036	0.036	95.1	47.2	0.001	0.055	0.054	95.6	34.4
		$\tilde{\tau}(0.5)$	0.003	0.027	0.027	95.3	83.0	0.001	0.042	0.041	95.4	59.4
		$\hat{\tau}_{\mathrm{OPT}}$	0.004	0.025	0.024	94.8	-	0.000	0.032	0.032	94.8	-
	40%	$\hat{\tau}^*$	-0.238	0.016	0.016	20.2	-	-0.362	0.017	0.017	1.36	-
		$\hat{\tau}_{\mathrm{V}}$	0.002	0.036	0.036	94.4	35.7	0.002	0.036	0.035	95.2	49.5
		$\hat{\tau}_{\mathrm{N}}$	0.005	0.044	0.044	94.9	23.0	-0.008	0.066	0.067	94.7	14.3
		$\tilde{\tau}(0.5)$	0.004	0.024	0.024	95.4	76.0	-0.004	0.037	0.037	94.6	46.9
		$\hat{\tau}_{\mathrm{OPT}}$	0.004	0.021	0.021	95.0	-	-0.001	0.025	0.025	95.0	-
	60%	$\hat{\tau}^*$	-0.158	0.016	0.016	54.1	-	-0.241	0.016	0.016	18.8	-
		$\hat{\tau}_{\mathrm{V}}$	0.001	0.026	0.026	95.5	51.3	-0.002	0.026	0.027	94.9	62.2
		$\hat{\tau}_{\mathrm{N}}$	0.007	0.058	0.058	95.4	10.6	0.003	0.086	0.087	94.8	5.87
		$\tilde{\tau}(0.5)$	0.003	0.022	0.022	95.0	74.6	0.000	0.032	0.032	94.7	43.6
		$\hat{\tau}_{\mathrm{OPT}}$	0.003	0.019	0.019	94.8	-	0.001	0.021	0.021	94.9	-

Est.: estimator; ReBias: average relative bias; ASE: average sandwich standard error; ESE: empirical standard error; CP%: 95% *coverage percentage; ARE: average estimated relative efficiency.*

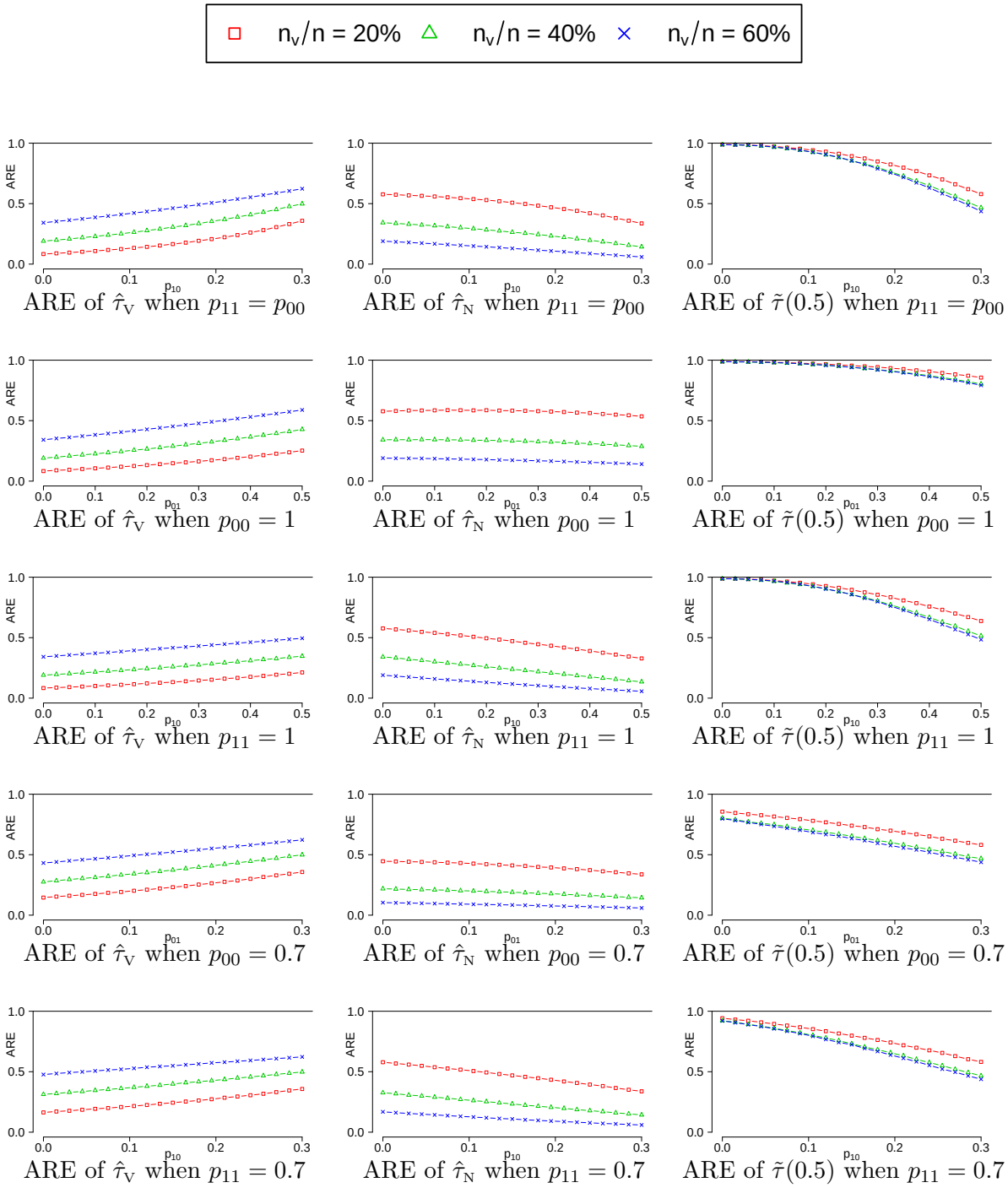


Figure 4.3: Average estimated relative efficiency (ARE) of proposed estimators $\hat{\tau}_{\mathrm{V}}$, $\hat{\tau}_{\mathrm{N}}$ and $\tilde{\tau}(0.5)$ relative to the proposed optimal estimator $\hat{\tau}_{\mathrm{OPT}}$ when validation data are available

Table 4.3: Simulation results for the misclassified binary outcome with replicates available: the performance of the proposed IPW estimator $\hat{\tau}_{\mathrm{R}}$ as opposed to the performance of the naive estimator $\hat{\tau}^*$

setting		$n = 1000$				$n = 5000$			
(p_{11}, p_{10})	Est.	ReBias	ASE	ESE	CP(%)	ReBias	ASE	ESE	CP(%)
$(0.9, 0.1)$	$\hat{\tau}^*$	-0.201	0.033	0.033	77.7	-0.200	0.015	0.015	25.9
	$\hat{\tau}_{\mathrm{R}}$	-0.002	0.041	0.041	94.9	0.000	0.018	0.018	94.6
$(0.8, 0.2)$	$\hat{\tau}^*$	-0.404	0.031	0.031	29.6	-0.401	0.014	0.014	0.00
	$\hat{\tau}_{\mathrm{R}}$	-0.004	0.051	0.053	94.8	-0.001	0.023	0.023	94.9
$(0.7, 0.3)$	$\hat{\tau}^*$	-0.605	0.029	0.030	2.90	-0.602	0.013	0.013	0.00
	$\hat{\tau}_{\mathrm{R}}$	0.002	0.076	0.078	94.9	-0.003	0.033	0.034	94.9

Est.: estimator; ReBias: average relative bias; ASE: average sandwich standard error; ESE: empirical standard error; CP%: 95% *coverage percentage.*

(ReBias), average sandwich standard error (ASE), empirical standard error (ESE) and 95% coverage percentage (CP%) are reported.

Table 4.3 summarizes the simulation results for the corrected estimator $\hat{\tau}_{\mathrm{R}}$ and the naive estimator $\hat{\tau}^*$, where $\hat{\tau}^*$ is obtained by treating $Y_i^* = \{Y_i^*(1) + Y_i^*(2)\}/2$ as the true value Y_i. The naive analysis leads to severely biased results, and its performance becomes worse as the degree of misclassification increases. The corrected estimator $\hat{\tau}_{\mathrm{R}}$ demonstrates a satisfactory performance with fairly small empirical biases and empirical coverage percentages close to 95%, as anticipated.

Finally, as discussed in Section 4.4.2, the consistency of estimator $\hat{\tau}_{\mathrm{R}}$ requires certain constraints on the model parameters. It is interesting to investigate the impact of constraint misspecification on estimation of τ_0. We examine two scenarios here. In Scenario 1, the constraint $p_{11} = 1$ is used to generate data, but the constraint $p_{11} = 1 - p_{10}$ is used to fit the data. Scenario 2 considers an opposite case; the constraint $p_{11} = 1 - p_{10}$ is imposed for the data generation but the constraint $p_{11} = 1$ is used when fitting the data. The

Table 4.4: Simulation results for the misclassified binary outcome with replicates when the constraint for identification is wrong (here $p_{11} = 1$ but we assume $p_{11} = 1 - p_{10}$): the performance of the proposed IPW estimator $\hat{\tau}_{\text{R}}$ as opposed to the performance of the naive estimator $\hat{\tau}^*$

setting		$n = 1000$				$n = 5000$			
(p_{11}, p_{10})	Est.	ReBias	ASE	ESE	CP(%)	ReBias	ASE	ESE	CP(%)
$(1, 0.1)$	$\hat{\tau}^*$	-0.099	0.033	0.033	91.5	-0.100	0.015	0.015	75.6
	$\hat{\tau}_{\text{R}}$	-0.033	0.036	0.036	94.8	-0.034	0.016	0.016	92.9
$(1, 0.2)$	$\hat{\tau}^*$	-0.198	0.032	0.033	78.2	-0.201	0.014	0.014	24.5
	$\hat{\tau}_{\text{R}}$	-0.084	0.037	0.037	92.2	-0.088	0.016	0.016	82.6
$(1, 0.3)$	$\hat{\tau}^*$	-0.298	0.031	0.031	53.9	-0.300	0.014	0.014	1.70
	$\hat{\tau}_{\text{R}}$	-0.157	0.037	0.038	87.1	-0.159	0.017	0.017	55.1

Est.: estimator; ReBias: average relative bias; ASE: average sandwich standard error; ESE: empirical standard error; CP%: 95% *coverage percentage.*

results are reported in Tables 4.4 and 4.5, respectively, for Scenarios 1 and 2, where the results of the naive analysis are also presented for comparisons. It is unsurprising that constraint misspecification may induce biased results, and biases tend to exacerbate as the degree of misclassification in the outcome variable increases. However, the performance of the estimator $\hat{\tau}_{\text{R}}$ is much better than that of the naive estimator, even in the existence of constraint misspecification.

4.6 Doubly Robust Estimator

The consistency of the estimator in Theorem 4.2 requires a correctly specified treatment model. To provide protection against model misspecification, in this section we propose a doubly robust estimator of τ_0, which is consistent even when the treatment model or the outcome model is misspecified.

Table 4.5: Simulation results for the misclassified binary outcome with replicates when the constraint for identification is wrong (here $p_{11} = 1 - p_{10}$ but we assume $p_{11} = 1$): the performance of the proposed IPW estimator $\hat{\tau}_{\mathrm{R}}$ as opposed to the performance of the naive estimator $\hat{\tau}^*$

setting		$n = 1000$				$n = 5000$			
(p_{11}, p_{10})	Est.	ReBias	ASE	ESE	CP(%)	ReBias	ASE	ESE	CP(%)
$(0.9, 0.1)$	$\hat{\tau}^*$	-0.201	0.033	0.033	78.1	-0.200	0.015	0.015	25.0
	$\hat{\tau}_{\mathrm{R}}$	0.038	0.042	0.043	94.7	0.039	0.019	0.019	92.9
$(0.8, 0.2)$	$\hat{\tau}^*$	-0.398	0.031	0.030	29.6	-0.401	0.014	0.014	0.00
	$\hat{\tau}_{\mathrm{R}}$	-0.025	0.050	0.049	95.4	-0.030	0.022	0.022	94.1
$(0.7, 0.3)$	$\hat{\tau}^*$	-0.597	0.029	0.030	2.80	-0.602	0.013	0.013	0.00
	$\hat{\tau}_{\mathrm{R}}$	-0.237	0.055	0.056	87.0	-0.246	0.025	0.024	52.8

Est.: estimator; ReBias: average relative bias; ASE: average sandwich standard error; ESE: empirical standard error; CP%: 95% *coverage percentage.*

4.6.1 Theoretical Development

Let $q_1 = P(Y = 1|T = 1, X)$ and $q_0 = P(Y = 1|T = 0, X)$ be the outcome probabilities. A consistent estimator of τ_0 can be created by the difference of a consistent estimator for $E(Y_1)$ and a consistent estimator for $E(Y_0)$ because $\tau_0 = E(Y_1) - E(Y_0)$. Therefore, we first propose augmented estimators for $E(Y_1)$ and $E(Y_0)$, respectively, given by

$$\hat{E}(Y_1) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i^*}{\hat{e}_i(p_{11} - p_{10})} - \frac{T_i - \hat{e}_i}{\hat{e}_i}\hat{q}_{i1} - \frac{T_i}{\hat{e}_i}\left(\frac{p_{10}}{p_{11} - p_{10}}\right)\right\} \tag{4.20}$$

and

$$\hat{E}(Y_0) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(1 - T_i)Y_i^*}{(1 - \hat{e}_i)(p_{11} - p_{10})} + \frac{T_i - \hat{e}_i}{1 - \hat{e}_i}\hat{q}_{i0} - \frac{1 - T_i}{1 - \hat{e}_i}\left(\frac{p_{10}}{p_{11} - p_{10}}\right)\right\}, \tag{4.21}$$

where $\hat{q}_{i1}$ is an estimate of $P(Y_i = 1|T_i = 1, X_i)$ and $\hat{q}_{i0}$ is an estimate of $P(Y_i = 1|T_i = 0, X_i)$. Then we estimate τ_0 by the augmented estimator

$$\hat{\tau}_{\text{DR}} = \hat{E}(Y_1) - \hat{E}(Y_0). \tag{4.22}$$

The augmented estimator $\hat{\tau}_{\text{DR}}$ combines the information on the treatment and outcome models. The following theorem establishes the doubly robust property of $\hat{\tau}_{\text{DR}}$ whose proof is given in Appendix C.4.

Theorem 4.3. *Assume the causal inference assumptions described in Section 1.1.2 and model (4.4) hold. Then (4.20) and (4.21) are consistent estimators of $E(Y_1)$ and $E(Y_0)$, respectively, when either the treatment model or the outcome model is correctly specified. Consequently, (4.22) is a consistent estimator of τ_0 when either the treatment model or the outcome model is correctly specified.*

Note that $\hat{q}_{i1}$ and $\hat{q}_{i0}$ cannot be directly obtained by fitting the postulated outcome models which relate Y with T and X, because the true value Y is unobserved. To obtain $\hat{q}_{i1}$ and $\hat{q}_{i0}$ with misclassification in the outcome addressed, we present a likelihood based approach using a logistic regression model to illustrate the idea; other regression model

forms can be accommodated in the same manner. Suppose the postulated outcome model is

$$\text{logit } q_1 = \beta_{01} + \boldsymbol{\beta}_{\mathrm{x}}^{\mathrm{T}} X, \tag{4.23}$$

$$\text{logit } q_0 = \beta_{00} + \boldsymbol{\beta}_{\mathrm{x}}^{\mathrm{T}} X, \tag{4.24}$$

where β_{01}, β_{00}, and $\boldsymbol{\beta}_{\mathrm{x}}$ are regression parameters. Noticing that the outcome models (4.23) and (4.24) can be rewritten in a single form as:

$$\text{logit } P(Y = 1|T, X) = \beta_{00} + (\beta_{01} - \beta_{00})T + \boldsymbol{\beta}_{\mathrm{x}}^{\mathrm{T}} X,$$

we then write the observed likelihood function contributed from subject i as

$$\begin{aligned} & L_i(\beta_{00}, \beta_{01}, \boldsymbol{\beta}_{\mathrm{x}}) \\ = & \frac{1}{1 + \exp\{-\beta_{00} - (\beta_{01} - \beta_{00})T_i - \boldsymbol{\beta}_{\mathrm{x}}^{\mathrm{T}} X_i\}} \cdot \{p_{11} Y_i^* + (1 - p_{11})(1 - Y_i^*)\} \\ + & \frac{\exp\{-\beta_{00} - (\beta_{01} - \beta_{00})T_i - \boldsymbol{\beta}_{\mathrm{x}}^{\mathrm{T}} X_i\}}{1 + \exp\{-\beta_{00} - (\beta_{01} - \beta_{00})T_i - \boldsymbol{\beta}_{\mathrm{x}}^{\mathrm{T}} X_i\}} \cdot \{p_{10} Y_i^* + (1 - p_{10})(1 - Y_i^*)\}. \end{aligned}$$

The maximization of $\prod_{i=1}^n L_i(\beta_{00}, \beta_{01}, \boldsymbol{\beta}_{\mathrm{x}})$ with respect to $(\beta_{00}, \beta_{01}, \boldsymbol{\beta}_{\mathrm{x}}^{\mathrm{T}})$ yields a consistent estimator of $(\beta_{00}, \beta_{01}, \boldsymbol{\beta}_{\mathrm{x}}^{\mathrm{T}})$, denoted as $(\hat{\beta}_{00}, \hat{\beta}_{01}, \hat{\boldsymbol{\beta}}_{\mathrm{x}}^{\mathrm{T}})$. It is immediate that

$$\hat{q}_{i1} = \hat{P}(Y_i = 1|T_i = 1, X_i) = \frac{1}{1 + \exp(-\hat{\beta}_{01} - \hat{\boldsymbol{\beta}}_{\mathrm{x}}^{\mathrm{T}} X_i)}$$

and

$$\hat{q}_{i0} = \hat{P}(Y_i = 1|T_i = 0, X_i) = \frac{1}{1 + \exp(-\hat{\beta}_{00} - \hat{\boldsymbol{\beta}}_{\mathrm{x}}^{\mathrm{T}} X_i)}.$$

4.6.2 Simulation Studies

Let X_1 follow a standard normal distribution. The treatment T is generated from a Bernoulli distribution with probability $1/\{1 + \exp(-0.1 - X_1 - 0.2X_2)\}$, where $X_2 = X_1^2$.

Write $X = (X_1, X_2)^{\mathrm{T}}$. The treatment model is a logistic regression model

$$\text{logit } P(T = 1|X) = 0.1 + X_1 + 0.2X_2, \tag{4.25}$$

and the outcome model is a logistic regression model

$$\text{logit } P(Y = 1|T, X) = -1 + T + 0.5X_1 + X_2. \tag{4.26}$$

The misclassification mechanism (4.4) is assumed with (p_{11}, p_{10}) being $(0.9, 0.1)$ or $(0.8, 0.2)$. We consider two sample sizes of $n = 2000$ and $n = 5000$, and 1000 simulations are run for each pair of (p_{11}, p_{10}). Variance estimates are obtained by bootstrapping (Efron, 1982) and the number of bootstrap replicates is set to be 1000. To evaluate the performance of the proposed doubly robust estimator $\hat{\tau}_{\text{DR}}$ in Theorem 4.3, we compare it with the corrected estimator $\hat{\tau}$ in Theorem 2 whose validity requires a correctly specified treatment model. The average relative bias (ReBias), average bootstrap standard error (ASE), empirical standard error (ESE) and 95% coverage percentage (CP%) are reported.

We specifically consider three scenarios for model specification.

1. Both the treatment model and the outcome model are correctly specified:

In this case, a logistic regression model is correctly specified to relate T with X_1 and X_2, and a logistic regression model is correctly specified to relate Y with T, X_1 and X_2. That is, we use both (4.25) and (4.26) to generate data and fit the data.

2. The treatment model is correctly specified but the outcome model is misspecified:

In this case, a logistic regression model is correctly specified to relate T with X_1 and X_2, but the outcome model is misspecified. Specifically, we use (4.25) and (4.26) to generate data, but when fitting the data, we use (4.25) and (4.26) with X_2 removed from (4.26).

3. The outcome model is correctly specified but the treatment model is misspecified:

In this case, a logistic regression model is correctly specified to relate Y with T, X_1 and X_2, but the treatment model is misspecified. Specifically, we use (4.25) and (4.26) to

generate data, but when fitting the data, we use (4.25) and (4.26) with X_2 removed from (4.25).

Table 4.6 summarizes the simulation results. Confirmed by Theorem 4.3, the doubly robust estimator $\hat{\tau}_{\text{DR}}$ presents satisfactory performance with fairly small empirical biases and the coverage rates of 95% confidence intervals close to the nominal level in all three scenarios. In comparison, estimator $\hat{\tau}$ performs well in the first scenario but produces severely biased results in the third scenario, because the validity of this estimator requires a correctly specified treatment model. It is interesting that $\hat{\tau}$ is reasonably robust to misspecification of the outcome model; $\hat{\tau}$ performs fairly well for Scenario 2 which involves misspecification of the outcome model. The reason is that the calculation of $\hat{\tau}$ does not involve information on the outcome model.

4.6.3 Analysis of Smoking Cessation Data

We re-analyze the smoking cessation data in Section 4.3.4 using the doubly robust estimator. The same treatment model is employed. We specify the outcome model as a logistic model linking the outcome and other variables (treatment and pre-treatment covariates as in Section 4.3.4). We specify $p_{11} = 100\%$ and $p_{10} = 7.5\%$ as in Section 4.3.4. By using the doubly robust estimation method, we obtain $\hat{\tau}_{\text{DR}} = 0.203$ with a bootstrap variance estimate 0.651×10^{-2}, leading to a 95% confidence interval (0.045, 0.361), where the number of bootstrap replicates is 5000. Therefore, the smoking cessation intervention reduces the smoking rate by 20.3% after adjustment for misclassification of the outcome. If we further specify $p_{11} = 100\%$ and $p_{10} = 10\%$ as in Section 4.3.4, then the adjusted estimated ATE is 0.212, with a bootstrap variance estimate 0.757×10^{-2}, leading to a 95% confidence interval (0.041, 0.382). These results indicate a statistically significant effect of the smoking cessation program on reducing smoking rate which agrees with results in Section 4.3.4.

Table 4.6: Simulation results for a binary outcome misclassified with known misclassification probabilities: the performance of doubly robust estimator $\hat{\tau}_{\text{DR}}$ in comparison with treatment model based estimator $\hat{\tau}$

			$p_{11} = 0.9,\ p_{10} = 0.1$				$p_{11} = 0.8,\ p_{10} = 0.2$			
n	Est.	Scenario	ReBias	ASE	ESE	CP%	ReBias	ASE	ESE	CP%
2000	$\hat{\tau}$	both	0.014	0.035	0.036	94.0	-0.005	0.048	0.052	94.8
		trt	-0.006	0.035	0.037	94.9	0.013	0.047	0.051	94.1
		out	0.423	0.036	0.035	31.6	0.489	0.047	0.046	42.8
	$\hat{\tau}_{\text{DR}}$	both	0.000	0.030	0.030	95.1	-0.009	0.043	0.046	94.2
		trt	-0.009	0.031	0.031	95.0	0.005	0.043	0.045	93.5
		out	0.001	0.029	0.028	96.4	0.009	0.040	0.040	94.4
5000	$\hat{\tau}$	both	0.006	0.023	0.024	93.8	0.000	0.031	0.032	95.7
		trt	0.013	0.023	0.023	95.0	0.006	0.031	0.031	94.9
		out	0.427	0.022	0.022	3.40	0.487	0.030	0.029	8.60
	$\hat{\tau}_{\text{DR}}$	both	0.001	0.019	0.020	95.5	-0.004	0.027	0.027	94.8
		trt	0.010	0.020	0.019	96.1	0.003	0.028	0.027	95.1
		out	0.003	0.018	0.018	95.4	0.003	0.026	0.025	96.0

both: both models are correctly specified;
trt: only treatment model is correctly specified;
out: only outcome model is correctly specified;
Est.: estimator; ReBias: average relative bias; ASE: average bootstrap standard error; ESE: empirical standard error; CP%: 95% *coverage percentage.*

4.7 Extensions to Complex Misclassification

For settings with a misclassified binary outcome, we assume that the misclassification probabilities are governed by the mechanism (4.4), which says that the surrogate measurement Y^* is independent of X and T if Y is controlled. This assumption is often feasible and it parallels the so-called nondifferential measurement error mechanism classified for error-contaminated covariates (Carroll et al., 2006, p.36). When the assumption (4.4) is violated, our development can be modified to generalize model (4.4) to incorporate possible dependence on X and/or T. The following theorem establishes the asymptotic bias of the naive estimator and provides a way of constructing a consistent estimator.

Theorem 4.4. *Under the causal inference assumptions described in Section 1.1.2, naively replacing Y_i with Y_i^* in the IPW estimator (4.1) yields a biased estimator of τ_0. Let $\hat{\tau}^*$ denote this naive estimator.*

(a). The asymptotic bias of the naive estimator $\hat{\tau}^*$ is

$$E\left([\{p_{111}(X)-p_{101}(X)\}P(Y_1=1|X)+p_{101}(X)]\right.$$
$$\left.-[\{p_{110}(X)-p_{100}(X)\}P(Y_0=1|X)+p_{100}(X)]\right)-\tau_0,$$

where $p_{abt}(x)=P(Y^*=a|Y=b,X=x,T=t)$ for $a,b,t=0,1$, and x being a realization of X.

(b). Let

$$\hat{\tau}=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_iY_i^*}{\hat{e}_i\{p_{111}(X_i)-p_{101}(X_i)\}}-\frac{p_{101}(X_i)}{p_{111}(X_i)-p_{101}(X_i)}\right]$$
$$-\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-T_i)Y_i^*}{(1-\hat{e}_i)\{p_{110}(X_i)-p_{100}(X_i)\}}-\frac{p_{100}(X_i)}{p_{110}(X_i)-p_{100}(X_i)}\right].$$

Then $\hat{\tau}$ is a consistent estimator of τ_0.

Similarly, we extend the doubly robust estimator (4.22) by the following theorem.

Theorem 4.5. *Assume the causal inference assumptions described in Section 1.1.2 hold. Let* $p_{abt}(x) = P(Y^* = a|Y = b, X = x, T = t)$ *for* $a, b, t = 0, 1$, *and* x *be a realization of* X. *Define* $\hat{\tau}_{\text{DR}}$ *to be*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i Y_i^*}{\hat{e}_i\{p_{111}(X_i) - p_{101}(X_i)\}} - \frac{T_i - \hat{e}_i}{\hat{e}_i}\hat{q}_{i1} - \frac{T_i}{\hat{e}_i}\left\{\frac{p_{101}(X_i)}{p_{111}(X_i) - p_{101}(X_i)}\right\}\right] -$$
$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-T_i)Y_i^*}{(1-\hat{e}_i)\{p_{110}(X_i) - p_{100}(X_i)\}} + \frac{T_i - \hat{e}_i}{1-\hat{e}_i}\hat{q}_{i0} - \frac{1-T_i}{1-\hat{e}_i}\left\{\frac{p_{100}(X_i)}{p_{110}(X_i) - p_{100}(X_i)}\right\}\right].$$

Then $\hat{\tau}_{\text{DR}}$ *is a consistent estimator of* τ_0 *when either the treatment model or the outcome model is correctly specified.*

Chapter 5

Weighted Causal Inference Methods with Mismeasured Covariates and Misclassified Outcomes

This chapter deals with Problem 4 discussed in Section 1.5. Section 5.1 describes the IPW estimation for settings with error-free data. In Section 5.2 we present the models for mismeasured covariates and outcome variables. In Section 5.3 we propose two correction methods which utilize different characteristics of the treatment model to deal with mismeasurements. Simulation studies are conducted in Section 5.4 to assess the finite sample performance of the proposed methods. In Section 5.5, we apply the proposed methods to analyze the NHEFS dataset for further illustration.

5.1 IPW Estimation with Error-Free Data

For each individual, let T be a binary treatment indicator with $T = 1$ if treated and $T = 0$ if untreated. Let Y_1 denote the potential outcome that would have been observed had the subject been treated and let Y_0 denote the potential outcome that would have been observed had the subject been untreated. Let Y be the observed binary outcome and let

X and Z be the vectors of covariates.

The goal is to estimate the average treatment effect (ATE), defined to be $\tau_0 = E(Y_1) - E(Y_0)$. Let

$$e = P(T = 1|X) \tag{5.1}$$

be the propensity score (Rosenbaum and Rubin, 1983). Assume the fundamental causal inference assumptions described in Section 1.1.2 hold.

Suppose we have a sample of size n. For subject i where $i = 1, \ldots, n$, we add subscript i to T, Y_1, Y_0, Y, X and Z for the corresponding variables of subject i and obtain T_i, $Y_{i,1}$, $Y_{i,0}$, Y_i, X_i and Z_i, respectively.

Rosenbaum (1998) proposed to estimate τ_0 by the IPW estimator, given by

$$\hat{\tau} = \hat{E}(Y_1) - \hat{E}(Y_0), \tag{5.2}$$

where

$$\hat{E}(Y_1) = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}_i}, \;\; \hat{E}(Y_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{1-\hat{e}_i},$$

and $\hat{e}_i$ is an estimated propensity score for subject i.

Under the causal inference assumptions described in Section 1.1.2 and that the treatment model for the propensity score e is correctly specified, the estimator $\hat{\tau}$ given by (5.2) is consistent. However, when the observed outcome Y and/or the covariates X are subject to measurement error, the consistency of $\hat{\tau}$ is no longer true.

5.2 Measurement Error Models

Suppose Z is precisely measured, but X is subject to measurement error. Let X^* be an observed measurement, or the surrogate of X. Suppose X^* and X are postulated by the classical additive model

$$X^* = X + \epsilon, \tag{5.3}$$

where the error term ϵ follows a normal distribution $N(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$ with covariance matrix $\mathbf{\Sigma}_\epsilon$ and ϵ is independent of $\{X, Z, T, Y, Y_1, Y_0\}$. To highlight the key idea, we assume that $\mathbf{\Sigma}_\epsilon$ is known, bearing in mind that $\mathbf{\Sigma}_\epsilon$ can be consistently estimated using extra data sources such as validation data or replicates (Carroll et al., 2006). The classical additive model (5.3) is perhaps the most commonly used measurement error model in the literature of measurement error models (Carroll et al., 2006; Yi, 2017).

Regarding misclassification in the outcome variable, we let Y^* represent an observed value of Y. To characterize the misclassification mechanisms, we assume that the misclassification probabilities satisfy

$$P(Y^* = a|Y = b, X^*, X, Z, T = t) = P(Y^* = a|Y = b, X, Z, T = t) = P(Y^* = a|Y = b) \tag{5.4}$$

for $a, b, t = 0, 1$, and let $p_{ab} = P(Y^* = a|Y = b)$ for $a, b = 0, 1$. Assumption (5.4) basically says that conditional on the true outcome variable Y, the surrogate Y^* is independent of (X, Z, T) as well as of the surrogate X^*. This assumption aligns with the outcome surrogacy discussed by Prentice (1989). To highlight the idea, we assume that p_{ab} is known for now.

5.3 Theoretical Results

In the presence of measurement error in X and misclassification in Y, estimator (5.2) cannot be directly applied because of the unavailability of X and Y. In the presence of misclassification in Y alone, in Chapter 4 we develop consistent estimators

$$\hat{E}(Y_1) = \frac{1}{n(p_{11} - p_{10})} \sum_{i=1}^{n} \frac{T_i Y_i^*}{\hat{e}_i} - \frac{p_{10}}{p_{11} - p_{10}} \tag{5.5}$$

and

$$\hat{E}(Y_0) = \frac{1}{n(p_{11} - p_{10})} \sum_{i=1}^{n} \frac{(1 - T_i) Y_i^*}{1 - \hat{e}_i} - \frac{p_{10}}{p_{11} - p_{10}} \tag{5.6}$$

for $E(Y_1)$ and $E(Y_0)$, respectively, to correct for misclassification effects in the outcome variable, where $\hat{e}_i$ is an estimate of propensity score e_i in (5.1) that is expressed in terms of X_i together with Z_i.

To further account for effects of measurement error in X_i, we need to revise the estimators (5.5) and (5.6). Noticing that when X_i is subject to error, we cannot calculate $\hat{e}_i$ by replacing X_i with X_i^* directly (Yi, 2017), we need to modify (5.5) and (5.6) in a way such that the modified estimators are expressed in terms of the observed Z_i, X_i^*, T_i and Y_i^*, so that the resulting estimator is consistent.

To this end, we propose to consider a working weight function $G(Z, X^*, T)$ of Z, X^* and T as well as of model parameters, where the dependence on the model parameters is suppressed in the notation. Let G_i be $G(Z_i, X_i^*, T_i)$, evaluated for subject i, and let $\hat{G}_i$ be G_i with unknown parameters replaced by their estimates. We propose the modified estimators

$$\hat{E}(Y_1) = \frac{1}{n(p_{11} - p_{10})} \sum_{i=1}^{n} T_i Y_i^* \hat{G}_i - \frac{p_{10}}{p_{11} - p_{10}} \tag{5.7}$$

and

$$\hat{E}(Y_0) = \frac{1}{n(p_{11} - p_{10})} \sum_{i=1}^{n} (1 - T_i) Y_i^* \hat{G}_i - \frac{p_{10}}{p_{11} - p_{10}} \tag{5.8}$$

for $E(Y_1)$ and $E(Y_0)$, respectively, to correct for the effects caused from both measurement error in X and misclassification in Y.

The inclusion of $\hat{G}_i$ in such a way mimics the inverse of the estimated propensity score $\hat{e}_i$, as shown in (5.5) and (5.6). We need to find a proper function form for $G(\cdot)$ such that the estimators (5.7) and (5.8) are consistent estimators for $E(Y_1)$ and $E(Y_0)$, respectively. The following theorem offers us a guideline of finding a $G(\cdot)$ function; the proof is deferred to Appendix D.1.

Theorem 5.1. *Suppose that the causal inference assumptions described in Section 1.1.2, measurement error model (5.3) and misclassification model (5.4) hold. Let* $\hat{\tau} = \hat{E}(Y_1) - \hat{E}(Y_0)$ *where* $\hat{E}(Y_1)$ *and* $\hat{E}(Y_0)$ *are defined by (5.7) and (5.8), respectively. Then* $\hat{\tau}$ *is a consistent estimator of* τ_0 *if function* $G(\cdot)$ *satisfies*

$$E\{G(Z, X^*, T)|X, Z, T = 1\} = \frac{1}{P(T = 1|X, Z)} \tag{5.9}$$

and

$$E\{G(Z, X^*, T)|X, Z, T = 0\} = \frac{1}{P(T = 0|X, Z)}. \tag{5.10}$$

We comment that in the absence of misclassification in Y, Theorem 5.1 recovers the results of McCaffrey et al. (2013). Within a class of logistic treatment models, the closed-form solution for $G(\cdot)$ is available (McCaffrey et al., 2013). In general settings, finding a $G(\cdot)$ function to satisfy (5.9) and (5.10) may be difficult. Now we examine this in detail in the following subsections.

5.3.1 Consistent Estimation with Logistic Treatment Models

Suppose that the treatment model (5.1) is given by a logistic regression model, a model that is widely employed in applications:

$$\text{logit}\{P(T=1|Z,X)\} = \alpha_0 + \boldsymbol{\alpha}_Z^{\mathrm{T}} Z + \boldsymbol{\alpha}_X^{\mathrm{T}} X, \tag{5.11}$$

where $\boldsymbol{\alpha} = (\alpha_0, \boldsymbol{\alpha}_Z^{\mathrm{T}}, \boldsymbol{\alpha}_X^{\mathrm{T}})^{\mathrm{T}}$ is the vector of parameters.

Under model (5.11), setting

$$G(Z, X^*, T) = 1 + \exp\{(-\alpha_0 - \boldsymbol{\alpha}_Z^{\mathrm{T}} Z - \boldsymbol{\alpha}_X^{\mathrm{T}} \Delta)(2T-1)\} \tag{5.12}$$

makes (5.9) and (5.10) hold, where $\Delta = X^* + (T - 1/2)\boldsymbol{\Sigma}_\epsilon \boldsymbol{\alpha}_X$; the justification is outlined in Appendix D.2. Consequently,

$$\hat{G}_i = 1 + \exp\{(-\hat{\alpha}_0 - \hat{\boldsymbol{\alpha}}_Z^{\mathrm{T}} Z_i - \hat{\boldsymbol{\alpha}}_X^{\mathrm{T}} \hat{\Delta}_i)(2T_i - 1)\}, \tag{5.13}$$

where $\hat{\Delta}_i = X_i^* + (T_i - 1/2)\boldsymbol{\Sigma}_\epsilon \hat{\boldsymbol{\alpha}}_X$ with a consistent estimator $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}_Z^{\mathrm{T}}, \hat{\boldsymbol{\alpha}}_X^{\mathrm{T}})^{\mathrm{T}}$ of $\boldsymbol{\alpha}$.

Now it remains to obtain a consistent estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$. This can be done by applying the conditional score method proposed by Stefanski and Carroll (1987). Specifically, we obtain $\hat{\boldsymbol{\alpha}}$ by solving the estimating equations

$$\sum_{i=1}^{n} \left\{ T_i - \frac{1}{1 + \exp(-\alpha_0 - \boldsymbol{\alpha}_Z^{\mathrm{T}} Z_i - \boldsymbol{\alpha}_X^{\mathrm{T}} \Delta_i)} \right\} \begin{pmatrix} 1 \\ Z_i \\ \Delta_i \end{pmatrix} = \mathbf{0} \tag{5.14}$$

for $\boldsymbol{\alpha}$, where $\Delta_i = X_i^* + (T_i - 1/2)\boldsymbol{\Sigma}_\epsilon \boldsymbol{\alpha}_X$. As a result, the consistent estimator $\hat{\tau}$ of τ_0 can be obtained using (5.7), (5.8) and (5.13). The associated variance estimate $\hat{V}ar(\hat{\tau})$ can be obtained using bootstrapping (Efron, 1982).

Finally, we make a comment on the proposed method. We note that the estimating equations (5.14) resembles the likelihood equation

$$\sum_{i=1}^{n} \left\{ T_i - \frac{1}{1 + \exp(-\alpha_0 - \boldsymbol{\alpha}_Z^{\mathrm{T}} Z_i - \boldsymbol{\alpha}_X^{\mathrm{T}} X_i)} \right\} \begin{pmatrix} 1 \\ Z_i \\ X_i \end{pmatrix} = \mathbf{0} \tag{5.15}$$

which is derived from the logistic model (5.11); the only difference is that X_i in (5.15) is replaced by Δ_i for (5.14). By such a similarity, we estimate the propensity score $e_i = P(T_i = 1 | Z_i, X_i)$ by

$$\hat{e}_i^* = \frac{1}{1 + \exp(-\hat{\alpha}_0 - \hat{\boldsymbol{\alpha}}_Z^{\mathrm{T}} Z_i - \hat{\boldsymbol{\alpha}}_X^{\mathrm{T}} \hat{\Delta}_i)}.$$

Replacing $\hat{e}_i$ with $\hat{e}_i^*$ in (5.5) and (5.6), we obtain the estimators

$$\hat{E}(Y_1) = \frac{1}{n(p_{11} - p_{10})} \sum_{i=1}^{n} \frac{T_i Y_i^*}{\hat{e}_i^*} - \frac{p_{10}}{p_{11} - p_{10}} \tag{5.16}$$

and

$$\hat{E}(Y_0) = \frac{1}{n(p_{11} - p_{10})} \sum_{i=1}^{n} \frac{(1 - T_i) Y_i^*}{1 - \hat{e}_i^*} - \frac{p_{10}}{p_{11} - p_{10}}. \tag{5.17}$$

Evidently, the estimators (5.16) and (5.17) are identical to the proposed estimators (5.7) and (5.8), with $\hat{G}_i$ given by (5.13). Therefore, we have shown that the proposed estimators (5.7) and (5.8) can also be obtained intuitively by adapting the idea of the conditional score method which was originally developed for the estimation of $\boldsymbol{\alpha}$.

5.3.2 Augmented Simulation-Extrapolation

The consistent estimation method described in Section 5.3.1 capitalizes on the logistic regression form (5.11) for the treatment model. With other regression forms for the treatment model, a closed-form expression for G may not be available or G does not even exist

to make conditions (5.9) and (5.10) satisfied; a similar aspect was discussed by Stefanski (1989) for a different setting. In such circumstances, we propose an augmented Simulation-Extrapolation (SIMEX) method which roots from a combination of the method developed in Chapter 4 and the SIMEX method proposed by Cook and Stefanski (1994).

Let B be a user-specified positive integer and $\Lambda = \{\lambda_1, \lambda_2 \ldots, \lambda_M\}$ be a sequence of increasingly ordered numbers, where $\lambda_1 = 0$, λ_M is positive and M is a positive integer that are user-specified. The proposed augmented SIMEX method consists of the following three steps.

Step 1 (Simulation):

For $i = 1, \ldots, n$, generate $e_{ib} \sim N(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$ for $b = 1, 2, \ldots, B$. For $\lambda \in \Lambda$, calculate

$$X_i^*(b, \lambda) = X_i^* + \sqrt{\lambda} e_{ib}.$$

Step 2 (Estimation):

For $b = 1, 2, \ldots, B$ and $\lambda \in \Lambda$, we treat $X_i^*(b, \lambda)$ as if it were the the true value and estimate ATE using (5.5) and (5.6). That is, we calculate

$$\hat{E}_{(b,\lambda)}(Y_1) = \frac{1}{n(p_{11} - p_{10})} \sum_{i=1}^{n} \frac{T_i Y_i^*}{\hat{e}_i(b, \lambda)} - \frac{p_{10}}{p_{11} - p_{10}}$$

and

$$\hat{E}_{(b,\lambda)}(Y_0) = \frac{1}{n(p_{11} - p_{10})} \sum_{i=1}^{n} \frac{(1 - T_i) Y_i^*}{1 - \hat{e}_i(b, \lambda)} - \frac{p_{10}}{p_{11} - p_{10}},$$

where $\hat{e}_i(b, \lambda)$ is the estimated propensity score for subject i with X_i replaced by $X_i^*(b, \lambda)$ in the treatment model. Then calculate

$$\hat{\tau}(b, \lambda) = \hat{E}_{(b,\lambda)}(Y_1) - \hat{E}_{(b,\lambda)}(Y_0)$$

and

$$\hat{\tau}(\lambda) = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}(b, \lambda).$$

Step 3 (Extrapolation):

Fitting a regression model to $\{(\lambda, \hat{\tau}(\lambda)) : \lambda \in \Lambda\}$ and extrapolating it back to $\lambda = -1$ gives a predicted value $\hat{\tau}(-1)$, denoted as $\tilde{\tau}$. Then $\tilde{\tau}$ is the augmented SIMEX estimator of τ_0. The associated variance estimate $\hat{Var}(\tilde{\tau})$ can be obtained using bootstrapping (Efron, 1982).

Here, λ indicates the degree of measurement error in X; the larger λ is, the larger the measurement error becomes. By gradually adding extra error to the observed X^* in the simulation step and obtaining a series of ATE estimates in the estimation step, we are able to delineate how different degrees of measurement error in X may bias the estimate of τ_0. By doing extrapolation back to $\lambda = -1$, we predict the ATE estimate without measurement error, which is of primary interest. The estimator $\tilde{\tau}$ is often not exactly consistent because the true extrapolation function is unknown and the user-specified extrapolation function is only an approximation to it.

5.4 Simulation Studies

We first conduct simulation studies to confirm the consistency of the proposed estimator $\hat{\tau}$ described in Section 5.3.1 and compare its performance to the naive analysis which ignores measurement error and misclassification. Let $\hat{\tau}^*$ denote the naive estimator, obtained by conducting IPW estimation in Section 5.1 with X^* and Y^* being treated as if they were precise measurements. Sample sizes $n = 1000$ and $n = 5000$ are considered, and 1000 simulations are run for each parameter configuration.

Consider two settings where the marginal distribution of covariates X and Z are the standard normal distribution, and the treatment T is generated from the logistic regression model (5.11) with the parameters set as $\boldsymbol{\alpha} = (0.2, 1, 1)^{\mathrm{T}}$. In Setting 1, let X and Z be independent; the outcome Y is generated from a Bernoulli distribution with probability $1/\{1 + \exp(-0.2 - T - Z - X)\}$. In Setting 2, let X and Z be dependent with correlation coefficient 0.5; the outcome Y is generated from a Bernoulli distribution with probability $1/\{1 + \exp(0.3 - T - Z + X)\}$. The measurement error model is specified as (5.3) where $\boldsymbol{\Sigma}_\epsilon$ is written as σ_ϵ^2 with σ_ϵ being 0.1, 0.5 or 1, to reflect different degrees of measurement error. The misclassification mechanism (5.4) is assumed with (p_{11}, p_{10}) being $(0.9, 0.1)$,

$(0.8, 0.2)$ and $(0.7, 0.3)$, to indicate different degrees of misclassification.

The average relative bias in percent (ReBias%), average bootstrap standard error (ASE), empirical standard error (ESE), and 95% coverage percentage (CP%) are reported. For estimator $\hat{\vartheta}$, the relative bias is defined to be $(\hat{\vartheta} - \tau_0)/\tau_0$, and the coverage percentage is defined to be the percentage of those 95% confidence intervals $\hat{\vartheta} \mp 1.96 \times \sqrt{\hat{V}ar(\hat{\vartheta})}$ which contain τ_0.

Tables 5.1 and 5.2 summarize the simulation results for Setting 1 and Setting 2, respectively. The naive estimator produces severely biased results due to the ignorance of measurement error in X and misclassification in Y. The corrected estimator $\hat{\tau}$ in Section 5.3.1 yields fairly small finite sample biases under various combinations of σ_ϵ and (p_{11}, p_{10}), and finite sample biases become smaller as the sample size increases, as expected. The discrepancy between ASE and ESE is fairly small, and the empirical coverage percentages are close to 95%, indicating that the bootstrap variance estimates are reliable.

Next, we conduct simulation studies to compare the performance of three methods. The first method is the proposed method described in Section 5.3.1, called the logistic-based correction method (LCM). The second method is the proposed method described in Section 5.3.2, called the augmented SIMEX method (ASIMEX). The third method is the naive analysis. When conducting ASIMEX, we set $B = 100$, $\Lambda = \{0, 0.5, 1, 1.5, 2\}$ and $M = 5$. The quadratic regression is used for the extrapolation step.

Two different treatment models are considered. The first treatment model is the same as before, i.e.,

$$P(T = 1|Z, X) = 1/\exp(-0.2 - Z - X). \tag{5.18}$$

The second model is

$$P(T = 1|Z, X) = \begin{cases} 0.1, & \text{if } 1/\exp(-0.2 - Z - X) \le 0.1, \\ 1/\exp(-0.2 - Z - X), & \text{if } 0.1 < 1/\exp(-0.2 - Z - X) < 0.9, \\ 0.9, & \text{if } 1/\exp(-0.2 - Z - X) \ge 0.9, \end{cases} \tag{5.19}$$

Under the two treatment models (5.18) and (5.19), the box plots of the estimates obtained from the two proposed methods and the naive analysis are displayed in Figures

Table 5.1: Simulation results for the consistent estimator $\hat{\tau}$ described in Section 5.3.1 in contrast to the naive estimator $\hat{\tau}^*$: Setting 1

			$n = 1000$				$n = 5000$			
(p_{11}, p_{10})	σ_ϵ	Est.	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
(0.9,0.1)	0.1	$\hat{\tau}^*$	-19.55	0.046	0.047	88.5	-18.68	0.020	0.020	66.5
		$\hat{\tau}$	-0.549	0.058	0.060	94.7	0.562	0.026	0.026	95.8
	0.5	$\hat{\tau}^*$	-3.786	0.043	0.044	94.5	-3.171	0.019	0.019	93.5
		$\hat{\tau}$	-1.329	0.067	0.071	95.2	-0.114	0.029	0.030	94.2
	1	$\hat{\tau}^*$	18.97	0.039	0.037	86.5	19.71	0.017	0.017	50.2
		$\hat{\tau}$	-2.475	0.132	0.133	96.3	-0.163	0.044	0.050	94.8
(0.8, 0.2)	0.1	$\hat{\tau}^*$	-40.64	0.046	0.047	68.1	-39.54	0.021	0.021	10.3
		$\hat{\tau}$	-2.138	0.078	0.078	94.9	-0.346	0.036	0.036	94.3
	0.5	$\hat{\tau}^*$	-26.21	0.044	0.045	84.7	-27.34	0.019	0.019	32.0
		$\hat{\tau}$	1.995	0.090	0.096	95.6	-0.405	0.040	0.040	94.6
	1	$\hat{\tau}^*$	-10.77	0.040	0.040	93.2	-10.38	0.018	0.017	83.3
		$\hat{\tau}$	-0.975	0.147	0.145	96.5	-0.592	0.058	0.061	95.7
(0.7, 0.3)	0.1	$\hat{\tau}^*$	-59.22	0.047	0.050	41.9	-59.22	0.021	0.022	0.40
		$\hat{\tau}$	0.853	0.120	0.127	93.7	0.840	0.054	0.055	94.8
	0.5	$\hat{\tau}^*$	-51.27	0.044	0.044	49.8	-51.37	0.020	0.020	0.60
		$\hat{\tau}$	0.963	0.132	0.135	95.7	0.527	0.061	0.062	95.6
	1	$\hat{\tau}^*$	-39.23	0.040	0.040	62.2	-39.99	0.018	0.018	2.80
		$\hat{\tau}$	2.309	0.219	0.208	95.8	0.656	0.086	0.091	94.8

Est.: estimator; ReBias%: average relative bias in percent; ASE: average bootstrap standard error; ESE: empirical standard error; CP%: 95% *coverage percentage.*

Table 5.2: Simulation results for the consistent estimator $\hat{\tau}$ described in Section 5.3.1 in contrast to the naive estimator $\hat{\tau}^*$: Setting 2

			$n = 1000$				$n = 5000$			
(p_{11}, p_{10})	σ_ϵ	Est.	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
(0.9,0.1)	0.1	$\hat{\tau}^*$	-21.22	0.057	0.062	87.2	-20.89	0.027	0.029	60.8
		$\hat{\tau}$	-0.592	0.072	0.079	95.2	-0.154	0.034	0.036	94.5
	0.5	$\hat{\tau}^*$	-33.35	0.052	0.056	71.8	-34.25	0.025	0.026	17.9
		$\hat{\tau}$	1.196	0.078	0.086	96.5	0.186	0.037	0.040	94.6
	1	$\hat{\tau}^*$	-51.70	0.049	0.050	39.6	-51.73	0.022	0.022	0.50
		$\hat{\tau}$	1.679	0.155	0.141	97.4	-0.524	0.052	0.056	95.9
(0.8, 0.2)	0.1	$\hat{\tau}^*$	-41.42	0.058	0.067	67.5	-39.84	0.027	0.026	12.6
		$\hat{\tau}$	-1.440	0.100	0.114	94.2	1.230	0.045	0.044	95.7
	0.5	$\hat{\tau}^*$	-50.75	0.053	0.056	45.6	-50.34	0.024	0.025	2.30
		$\hat{\tau}$	0.047	0.104	0.113	95.6	0.472	0.050	0.051	95.2
	1	$\hat{\tau}^*$	-64.09	0.050	0.050	23.4	-63.45	0.022	0.023	0.00
		$\hat{\tau}$	-0.212	0.185	0.180	98.2	0.513	0.069	0.079	96.1
(0.7, 0.3)	0.1	$\hat{\tau}^*$	-59.82	0.057	0.062	38.4	-60.12	0.027	0.028	2.50
		$\hat{\tau}$	1.445	0.145	0.157	95.0	0.672	0.068	0.071	95.6
	0.5	$\hat{\tau}^*$	-66.97	0.054	0.057	26.0	-67.03	0.025	0.025	0.50
		$\hat{\tau}$	0.193	0.163	0.179	95.3	-0.305	0.076	0.078	95.8
	1	$\hat{\tau}^*$	-76.24	0.050	0.051	11.4	-75.43	0.022	0.022	0.00
		$\hat{\tau}$	0.385	0.294	0.297	97.7	-0.813	0.102	0.110	95.9

Est.: estimator; ReBias%: average relative bias in percent; ASE: average bootstrap standard error; ESE: empirical standard error; CP%: 95% *coverage percentage.*

5.1 - 5.4. In each figure the horizontal line indicates the true value τ_0. The average of the estimates across 1000 runs for each method is indicated by $''+''$.

In Setting 1 with $n = 5000$, the results under models (5.18) and (5.19) are displayed in Figures 5.1 and 5.2, respectively. Figure 5.1 further demonstrates negligible empirical biases of the LCM method under the logistic treatment model (5.11). The naive analysis generally produces severely biased results. The performance of the ASIMEX method is satisfactory when $\sigma_\epsilon = 0.1$ or 0.5. When $\sigma_\epsilon = 1$, the performance of the ASIMEX method decays. Figure 5.2 also demonstrates seriously biased results of the naive analysis. Unsurprisingly, the LCM method produces biased results due to the violation of model (5.11). The ASIMEX method performs quite well when $\sigma_\epsilon = 0.1$ or 0.5, but when measurement error is severe with $\sigma_\epsilon = 1$, its performance decays as observed from Figure 5.1.

In Setting 2 with $n = 5000$, the results under models (5.18) and (5.19) are displayed in Figures 5.3 and 5.4, respectively. These figures further show the severe bias produced by ignoring mismeasurements.

We further conduct simulations for a different sample size, $n = 1000$, and a treatment model that is different from (5.18) and (5.19). The results yield similar implications to those discussed here; the detailed results are placed in the Supplementary Material.

Our simulation studies clearly show that the LCM method performs well under various combinations of measurement error and misclassification, as long as model (5.11) holds. The ASIMEX method enables us to handle general treatment models and shows promising performance with small to moderate measurement error in X. However, when the measurement error in X is remarkably large, the performance of the ASIMEX method decays.

5.5 Analysis of NHEFS Data

For illustration, we use the proposed approaches to analyze the NHEFS dataset. We are interested in studying possible causal effects of exercises on smoking cessation, with covariates age, sex, race, body mass index (BMI) and systolic blood pressure (SBP) controlled.

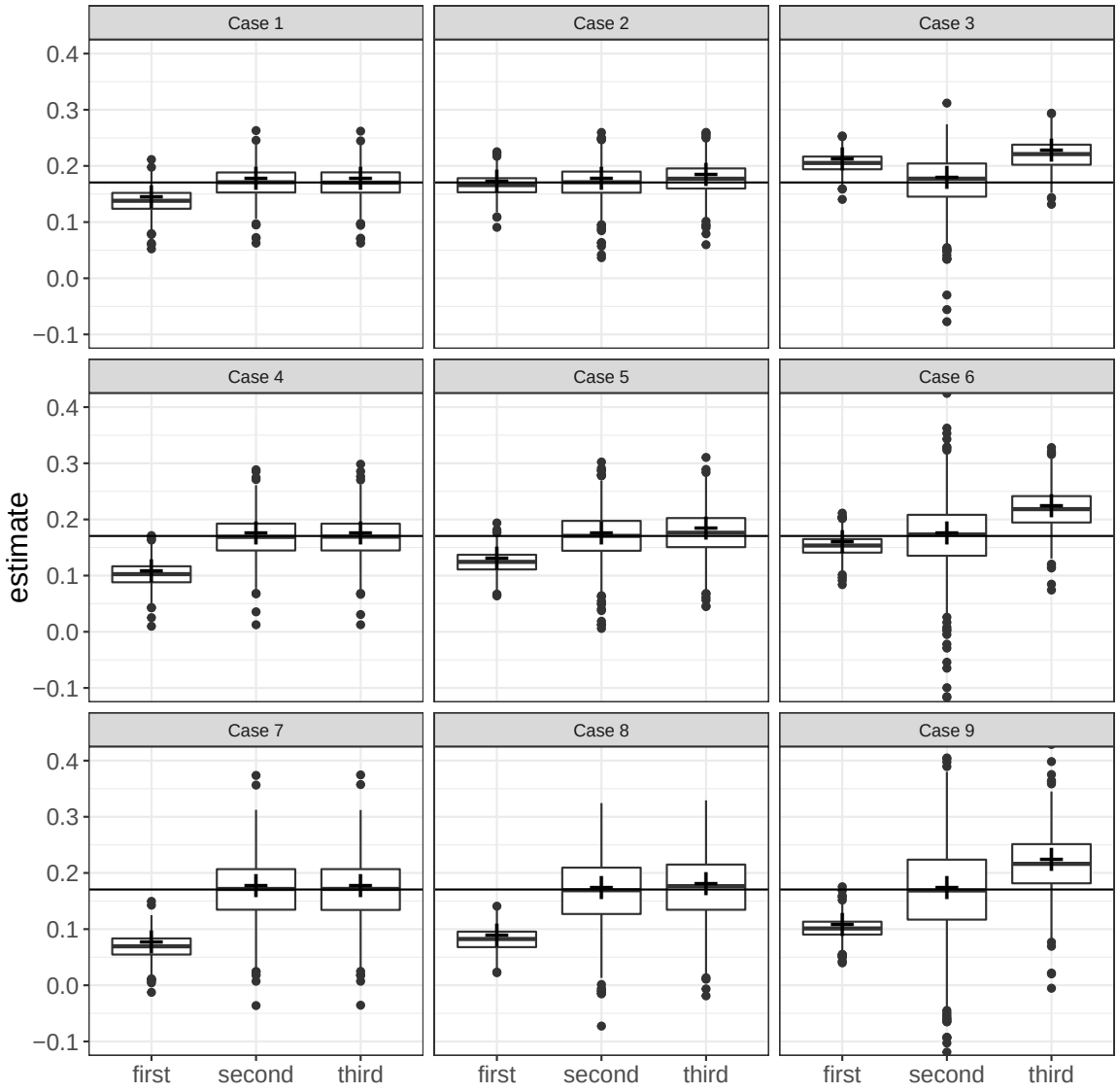


Figure 5.1: Simulation results for the comparison of the two proposed methods and the naive analysis: treatment model (5.18) in Setting 1 with $n = 5000$

first: the naive analysis; second: the proposed method in Section 5.3.1 (LCM); third: the proposed method in Section 5.3.2 (ASIMEX).
Cases 1, 2 and 3: $(p_{11}, p_{10}) = (0.9, 0.1)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 4, 5 and 6: $(p_{11}, p_{10}) = (0.8, 0.2)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 7, 8 and 9: $(p_{11}, p_{10}) = (0.7, 0.3)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
$+$*: average of estimates across 1000 runs; The horizontal solid line indicates* τ_0.

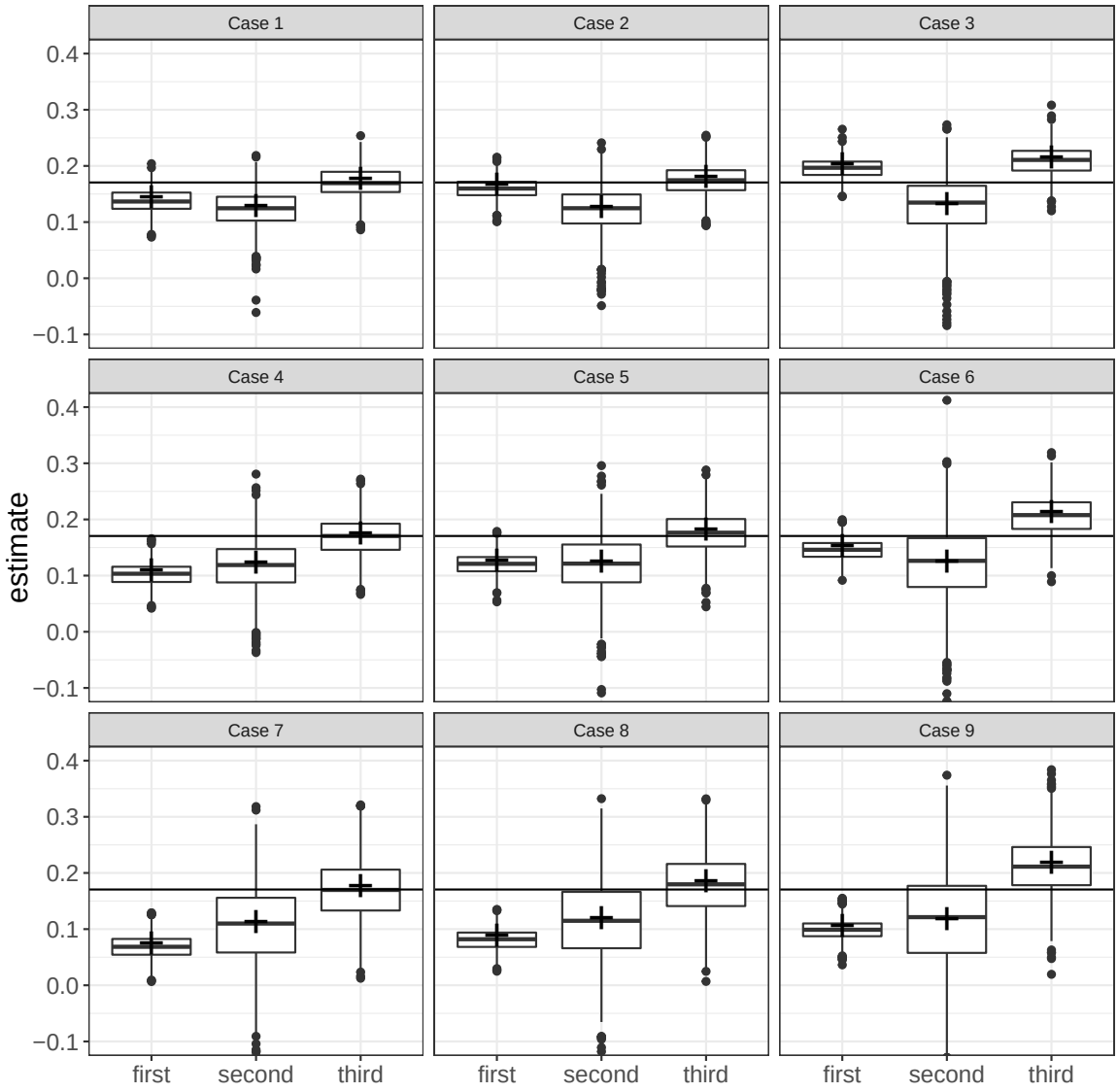


Figure 5.2: Simulation results for the comparison of the two proposed methods and the naive analysis: treatment model (5.19) in Setting 1 with $n = 5000$

first: the naive analysis; second: the proposed method in Section 5.3.1 (LCM); third: the proposed method in Section 5.3.2 (ASIMEX).
Cases 1, 2 and 3: $(p_{11}, p_{10}) = (0.9, 0.1)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 4, 5 and 6: $(p_{11}, p_{10}) = (0.8, 0.2)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 7, 8 and 9: $(p_{11}, p_{10}) = (0.7, 0.3)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
$+$*: average of estimates across 1000 runs; The horizontal solid line indicates* τ_0.

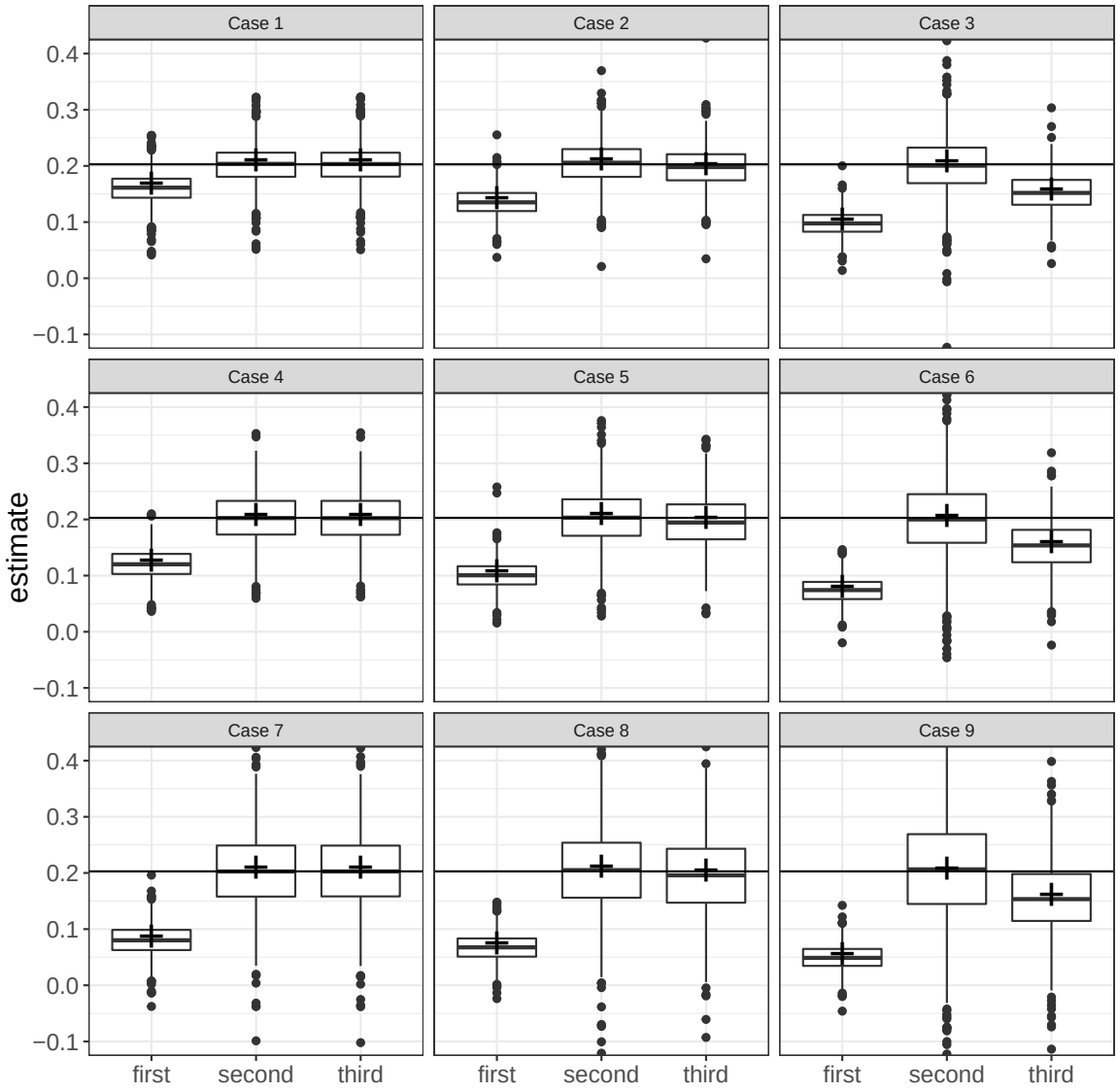


Figure 5.3: Simulation results for the comparison of the two proposed methods and the naive analysis: treatment model (5.18) in Setting 2 with $n = 5000$

first: the naive analysis; second: the proposed method in Section 5.3.1 (LCM); third: the proposed method in Section 5.3.2 (ASIMEX).
Cases 1, 2 and 3: $(p_{11}, p_{10}) = (0.9, 0.1)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 4, 5 and 6: $(p_{11}, p_{10}) = (0.8, 0.2)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 7, 8 and 9: $(p_{11}, p_{10}) = (0.7, 0.3)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
$+$*: average of estimates across 1000 runs; The horizontal solid line indicates* τ_0.

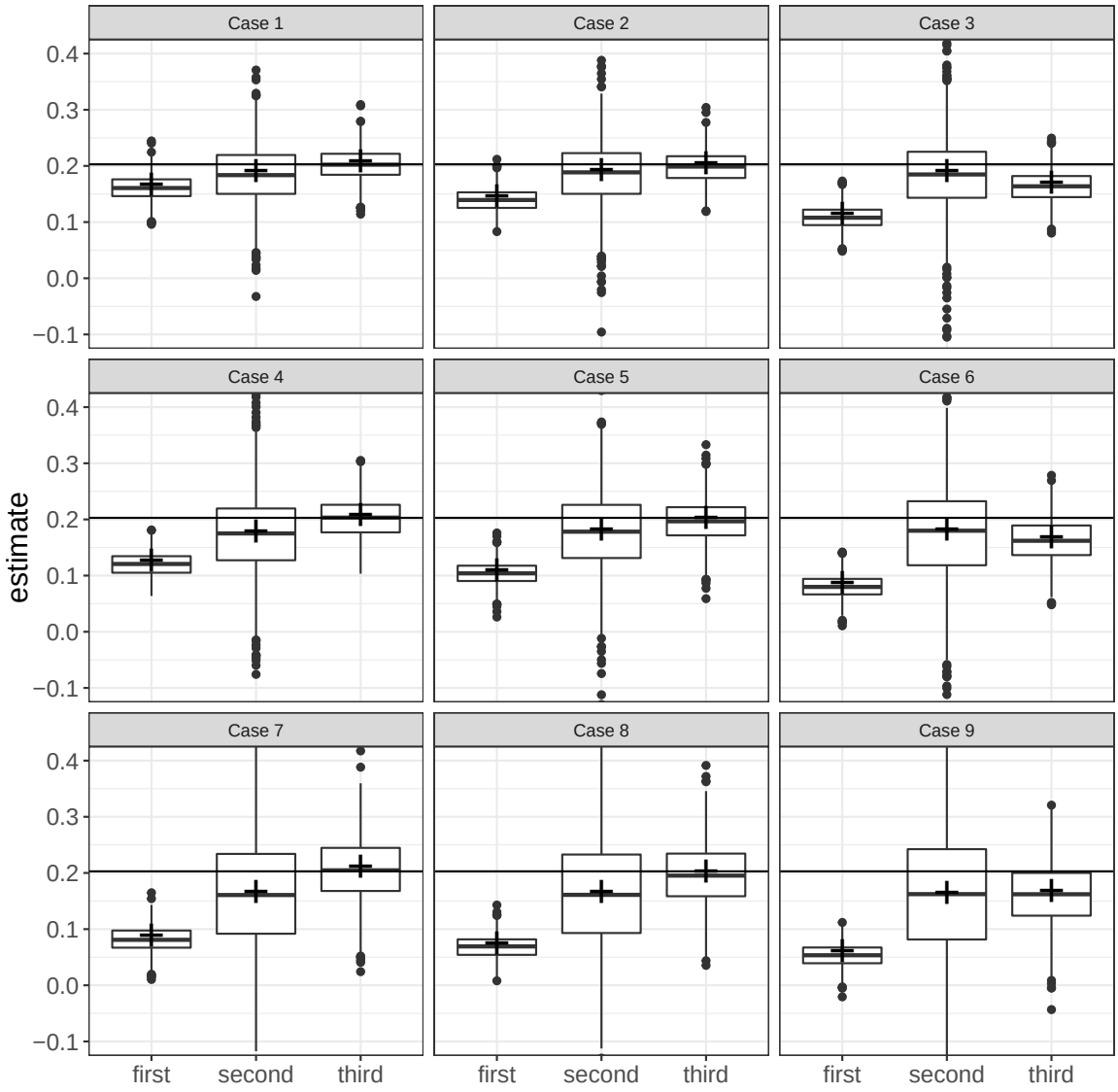


Figure 5.4: Simulation results for the comparison of the two proposed methods and the naive analysis: treatment model (5.19) in Setting 2 with $n = 5000$

first: the naive analysis; second: the proposed method in Section 5.3.1 (LCM); third: the proposed method in Section 5.3.2 (ASIMEX).
Cases 1, 2 and 3: $(p_{11}, p_{10}) = (0.9, 0.1)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 4, 5 and 6: $(p_{11}, p_{10}) = (0.8, 0.2)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 7, 8 and 9: $(p_{11}, p_{10}) = (0.7, 0.3)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
$+$*: average of estimates across 1000 runs; The horizontal solid line indicates* τ_0.

It is documented that measurement error arises inevitably in measuring the long-term average SBP due to daily and seasonal variation (Carroll et al., 2006). Moreover, self-reported smoking cessation status is subject to misclassification (Wagenknecht et al., 1992; Magder and Hughes, 1997; Lee et al., 2013). Therefore, the usual IPW estimation based on (5.2) is challenged by both covariate measurement error and outcome misclassification.

Let the treatment indicator T be a binary variable which takes value 1 if a subject has moderate or much exercise, and 0 if a subject takes little or no exercise. Precisely measured covariates Z include age, sex, race and BMI. According to Carroll et al. (2006), we consider a transformed SBP defined to be $\log(\text{SBP} - 50)$. The transformation strategy was originally described by Cornfield (1962) and then applied by Carroll et al. (1984) with a purpose to make the distribution of transformed observations reasonably approximate a normal distribution. Let X^* be the transformed observed SBP measurement and let Y^* denote the self-reported smoking cessation status.

The naive analysis and the two proposed methods are applied to analyze this dataset. One thousand bootstrap replicates are used for variance estimation. The naive analysis which ignores the presence of measurement error and misclassification yields that the estimate of ATE is -0.017 with a bootstrap standard error 0.021, leading to a 95% confidence interval $(-0.058, 0.025)$, suggesting no statistically significant causal effect of exercise on smoking cessation, at the nominal level of 5%.

We then apply the proposed methods developed in Sections 5.3.1 and 5.3.2, called LCM and ASIMEX, respectively, to analyze the data, just as described in Section 5.4. The treatment model is specified as the logistic regression model (5.11). Suppose the measurement error model (5.3) and the misclassification model (5.4) hold. Since there is no information on the degree of measurement error or misclassification, we conduct sensitivity analyses based on the information from other studies. Specifically, we take $\sigma_\epsilon^2 = 0.0126$, the value used by Carroll et al. (2006) for characterizing measurement error in SBP in analyzing the data arising from the Framingham Heart Study. In addition, we further set $\sigma_\epsilon^2 = 0.03$ to feature a scenario with a larger degree of measurement error in SBP. Magder and Hughes (1997) illustrated that subjects who really quit smoking were very likely to report that they have quit, but those who still smoked might inaccurately report having quit smoking. Therefore, assuming $p_{11} = 100\%$ and $p_{10} > 0$ is perhaps reasonable to reflect

Table 5.3: Analysis results of the NHEFS data using the proposed method in Section 5.3.1 (LCM) and the proposed method in Section 5.3.2 (ASIMEX): estimate (EST), bootstrap standard error (SE) and 95% confidence interval (95% CI)

		LCM			ASIMEX		
σ_ϵ^2	p_{10}	EST	SE	95% CI	EST	SE	95% CI
0.0126	0.05	-0.018	0.024	(-0.064, 0.029)	-0.018	0.023	(-0.063, 0.028)
	0.10	-0.019	0.025	(-0.067, 0.030)	-0.019	0.025	(-0.068, 0.031)
	0.20	-0.021	0.029	(-0.078, 0.036)	-0.021	0.028	(-0.076, 0.035)
	0.30	-0.024	0.032	(-0.087, 0.039)	-0.023	0.033	(-0.088, 0.041)
0.03	0.05	-0.018	0.023	(-0.063, 0.027)	-0.018	0.025	(-0.066, 0.030)
	0.10	-0.019	0.025	(-0.068, 0.031)	-0.018	0.024	(-0.066, 0.029)
	0.20	-0.021	0.029	(-0.079, 0.036)	-0.021	0.028	(-0.077, 0.034)
	0.30	-0.024	0.032	(-0.086, 0.038)	-0.024	0.031	(-0.085, 0.037)

Disclaimer: Interpretations and conclusions made by the authors do not reflect the view of National Center for Health Statistics.

misclassification of smoking status. Magder and Hughes (1997) specified $p_{10} = 10\%$. To study the impact of different degrees of misclassification on the estimation, here we consider $p_{10} = 5\%$, 10%, 20% or 30%.

Table 5.3 reports the estimates, the bootstrap standard errors and the 95% confidence intervals obtained from the LCM and ASIMEX methods under each combination of measurement error and misclassification. Both methods perform very similarly. As misclassification probability p_{10} increases, the resultant estimates of τ_0 decrease, although the change is small. The results are quite similar under $\sigma_\epsilon^2 = 0.0126$ and $\sigma_\epsilon^2 = 0.03$, implying a very small measurement error effect for this particular dataset. Both methods suggest no evidence of the causal effect of exercise on quitting smoking, since all the confidence intervals include 0. The proposed methods yield smaller estimates than the naive analysis, but the difference is not big. Compared with the naive analysis, the proposed methods yield larger standard errors which is consistent with the patterns in the literature.

In this example, the presence of measurement error and misclassification does not really alter the conclusion and the results given by the naive analysis are similar to those obtained by using the correction methods. However, this phenomenon does not occur in general. For example, as shown in our simulation studies, measurement error and misclassification can substantially degrade the inference results. In applications where mismeasurements are present, blindly conducting the naive analysis without a careful examination on the impact of measurement error and misclassification can yield misleading results.

Supplementary Material: Additional Simulation Results

We repeat simulations with $n = 1000$, and further consider another treatment model (5.20):

$$P(T=1|Z,X) = \begin{cases} 0.2, & \text{if } 1/\exp(-0.2-Z-X) \le 0.2, \\ 1/\exp(-0.2-Z-X), & \text{if } 0.2 < 1/\exp(-0.2-Z-X) < 0.8, \\ 0.8, & \text{if } 1/\exp(-0.2-Z-X) \ge 0.8, \end{cases} \tag{5.20}$$

The simulation results are displayed in the following figures.

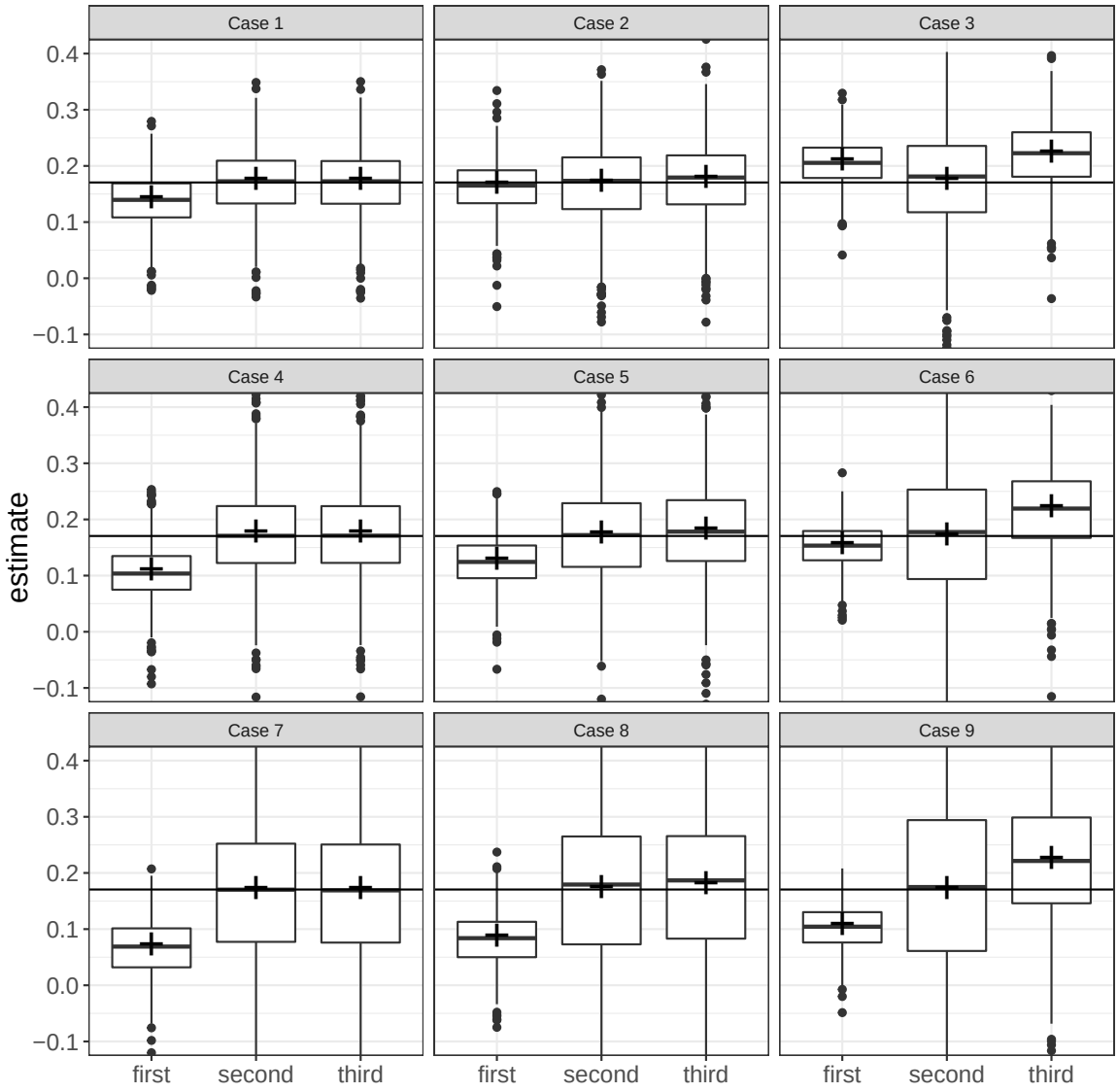


Figure 5.5: Simulation results for the comparison of the two proposed methods and the naive analysis: treatment model (5.18) in Setting 1 with $n = 1000$

first: the naive analysis; second: the proposed method in Section 5.3.1 (LCM); third: the proposed method in Section 5.3.2 (ASIMEX).
Cases 1, 2 and 3: $(p_{11}, p_{10}) = (0.9, 0.1)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 4, 5 and 6: $(p_{11}, p_{10}) = (0.8, 0.2)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 7, 8 and 9: $(p_{11}, p_{10}) = (0.7, 0.3)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
$+$*: average of estimates across 1000 runs; The horizontal solid line indicates* τ_0.

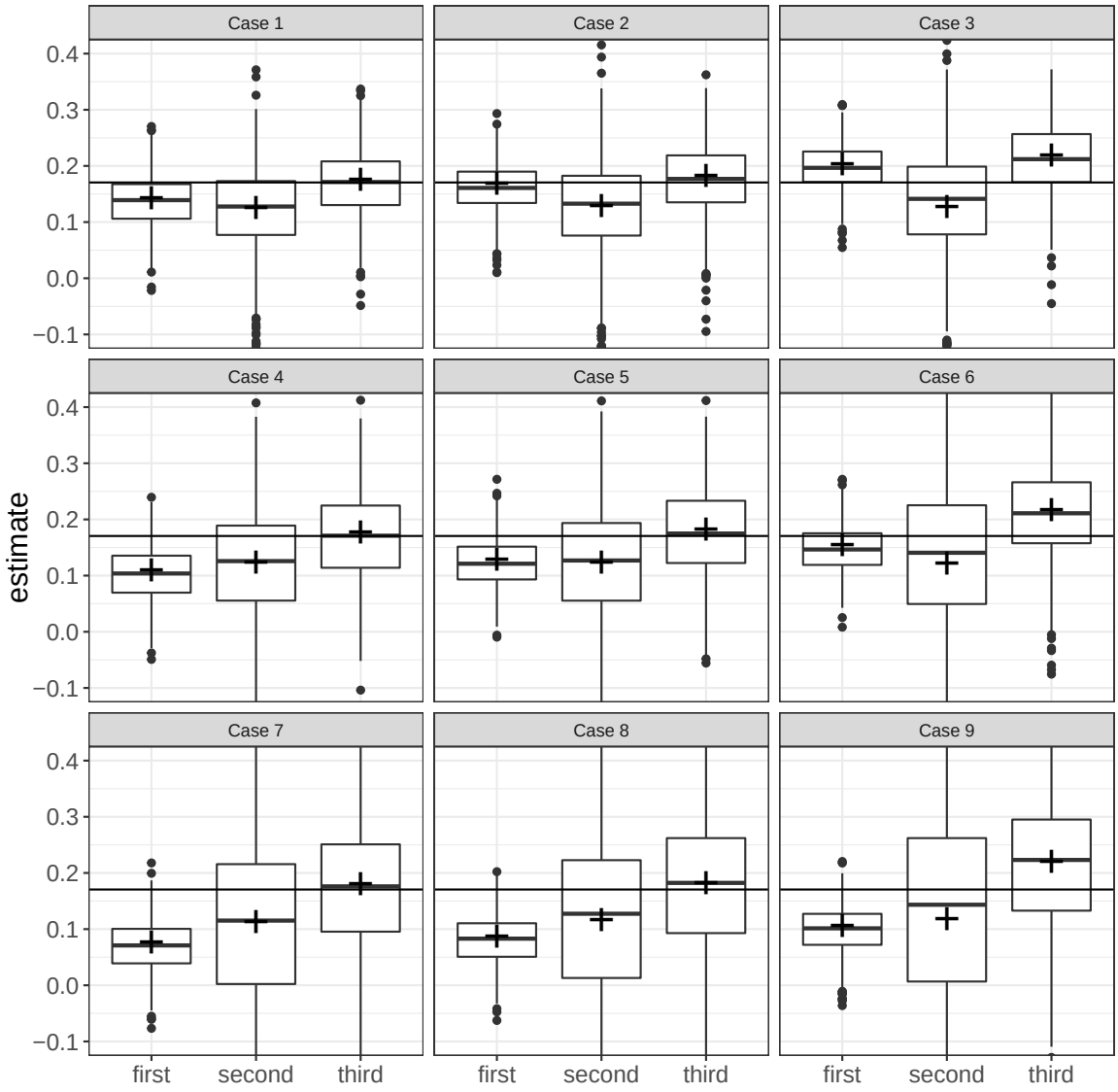


Figure 5.6: Simulation results for the comparison of the two proposed methods and the naive analysis: treatment model (5.19) in Setting 1 with $n = 1000$

first: the naive analysis; second: the proposed method in Section 5.3.1 (LCM); third: the proposed method in Section 5.3.2 (ASIMEX).
Cases 1, 2 and 3: $(p_{11}, p_{10}) = (0.9, 0.1)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 4, 5 and 6: $(p_{11}, p_{10}) = (0.8, 0.2)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 7, 8 and 9: $(p_{11}, p_{10}) = (0.7, 0.3)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
$+$*: average of estimates across 1000 runs; The horizontal solid line indicates* τ_0.

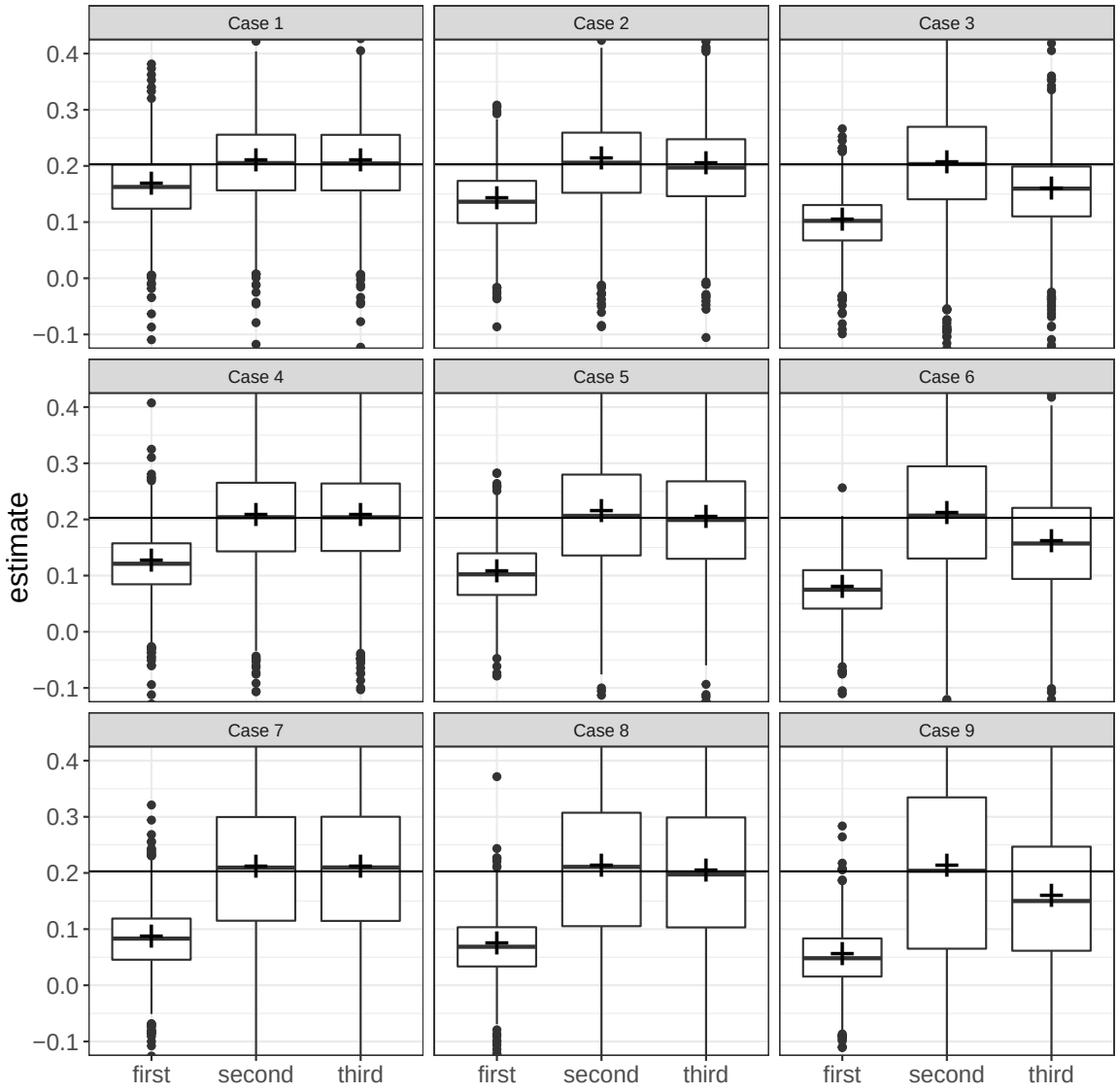


Figure 5.7: Simulation results for the comparison of the two proposed methods and the naive analysis: treatment model (5.18) in Setting 2 with $n = 1000$

first: the naive analysis; second: the proposed method in Section 5.3.1 (LCM); third: the proposed method in Section 5.3.2 (ASIMEX).
Cases 1, 2 and 3: $(p_{11}, p_{10}) = (0.9, 0.1)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 4, 5 and 6: $(p_{11}, p_{10}) = (0.8, 0.2)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 7, 8 and 9: $(p_{11}, p_{10}) = (0.7, 0.3)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
$+$*: average of estimates across 1000 runs; The horizontal solid line indicates* τ_0.

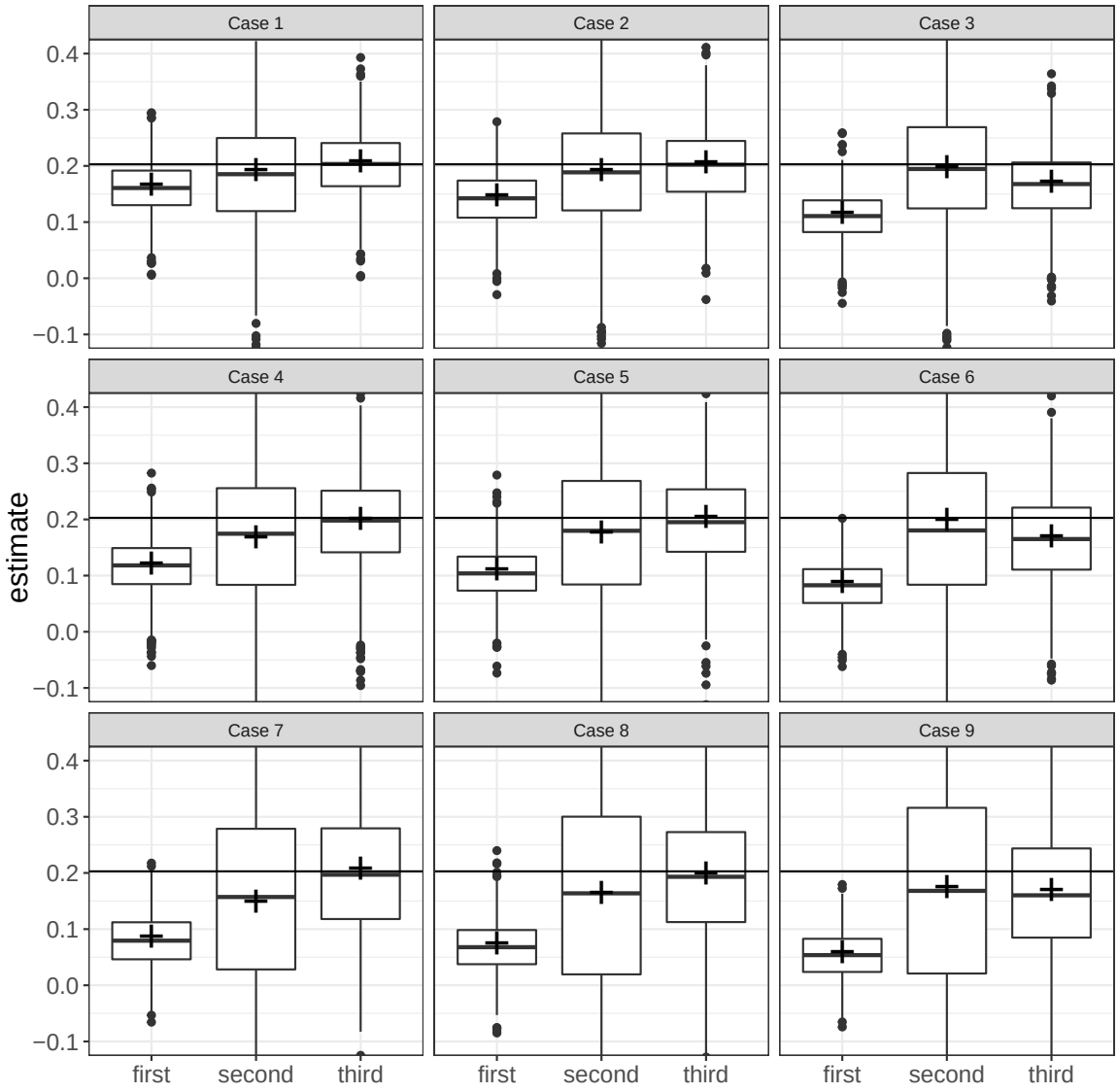


Figure 5.8: Simulation results for the comparison of the two proposed methods and the naive analysis: treatment model (5.19) in Setting 2 with $n = 1000$

first: the naive analysis; second: the proposed method in Section 5.3.1 (LCM); third: the proposed method in Section 5.3.2 (ASIMEX).
Cases 1, 2 and 3: $(p_{11}, p_{10}) = (0.9, 0.1)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 4, 5 and 6: $(p_{11}, p_{10}) = (0.8, 0.2)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 7, 8 and 9: $(p_{11}, p_{10}) = (0.7, 0.3)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
$+$*: average of estimates across 1000 runs; The horizontal solid line indicates* τ_0.

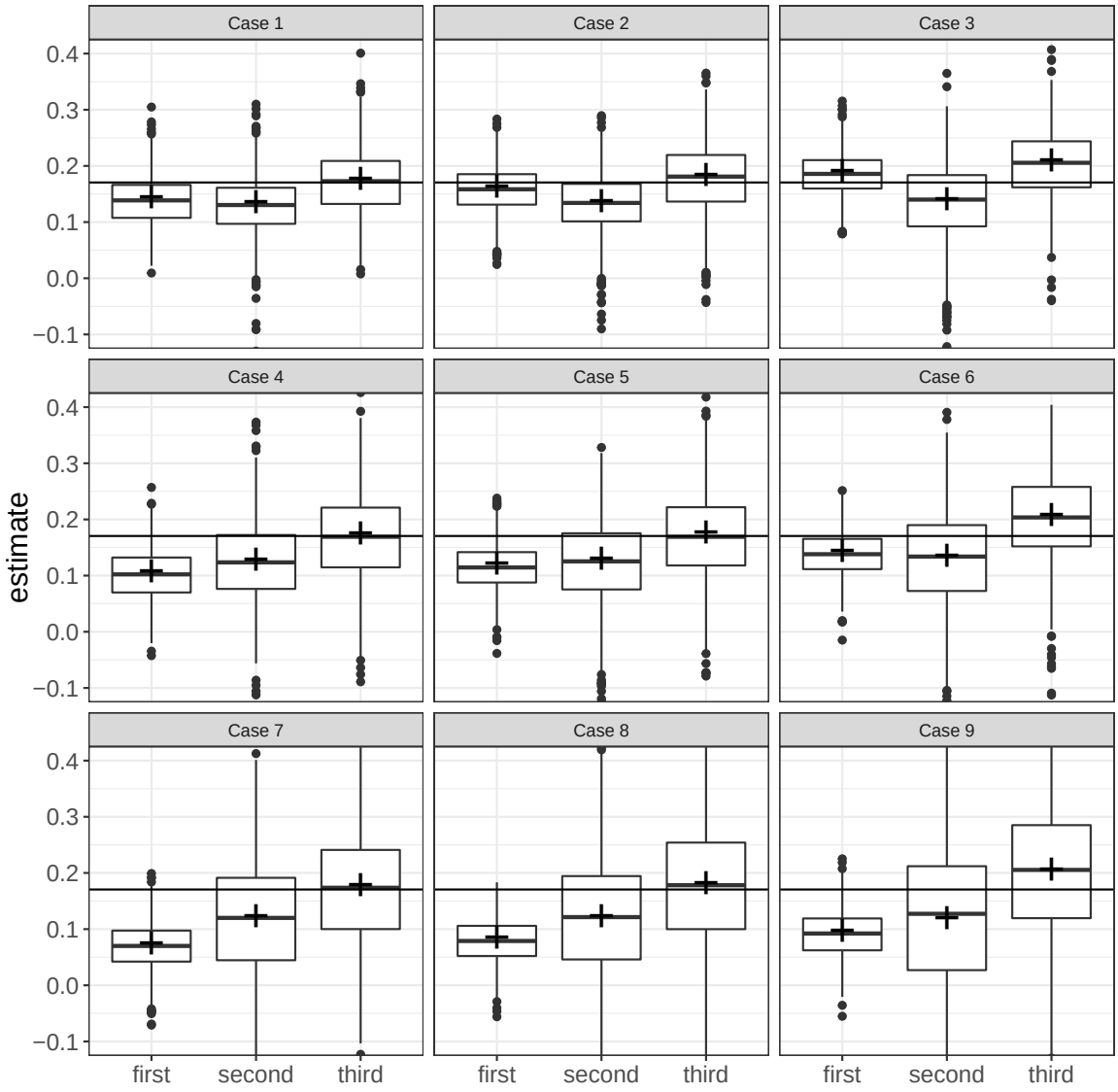


Figure 5.9: Simulation results for the comparison of the two proposed methods and the naive analysis: treatment model (5.20) in Setting 1 with $n = 1000$

first: the naive analysis; second: the proposed method in Section 5.3.1 (LCM); third: the proposed method in Section 5.3.2 (ASIMEX).
Cases 1, 2 and 3: $(p_{11}, p_{10}) = (0.9, 0.1)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 4, 5 and 6: $(p_{11}, p_{10}) = (0.8, 0.2)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 7, 8 and 9: $(p_{11}, p_{10}) = (0.7, 0.3)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
$+$*: average of estimates across 1000 runs; The horizontal solid line indicates* τ_0.

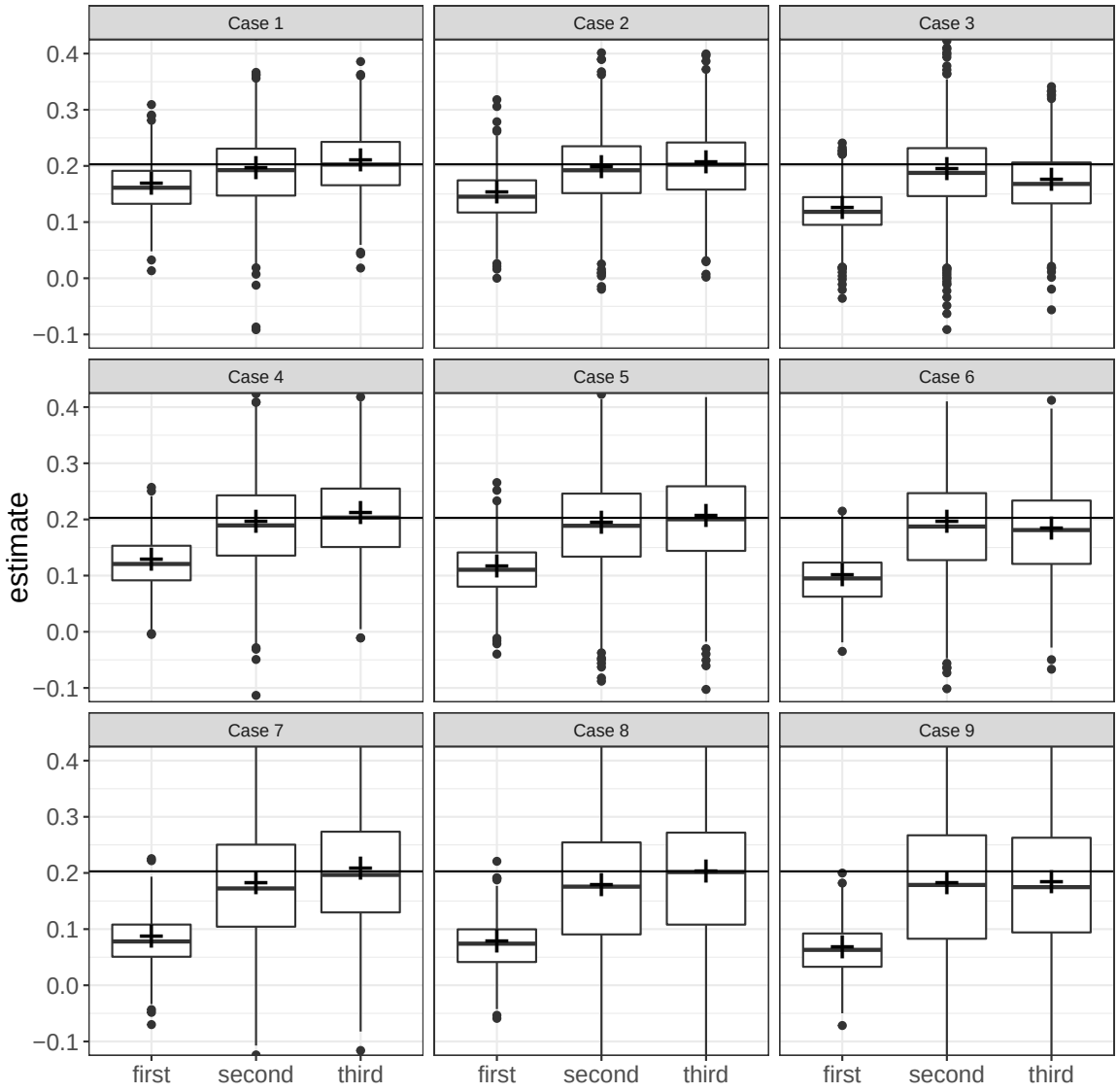


Figure 5.10: Simulation results for the comparison of the two proposed methods and the naive analysis: treatment model (5.20) in Setting 2 with $n = 1000$

first: the naive analysis; second: the proposed method in Section 5.3.1 (LCM); third: the proposed method in Section 5.3.2 (ASIMEX).
Cases 1, 2 and 3: $(p_{11}, p_{10}) = (0.9, 0.1)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 4, 5 and 6: $(p_{11}, p_{10}) = (0.8, 0.2)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 7, 8 and 9: $(p_{11}, p_{10}) = (0.7, 0.3)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
$+$*: average of estimates across 1000 runs; The horizontal solid line indicates* τ_0.

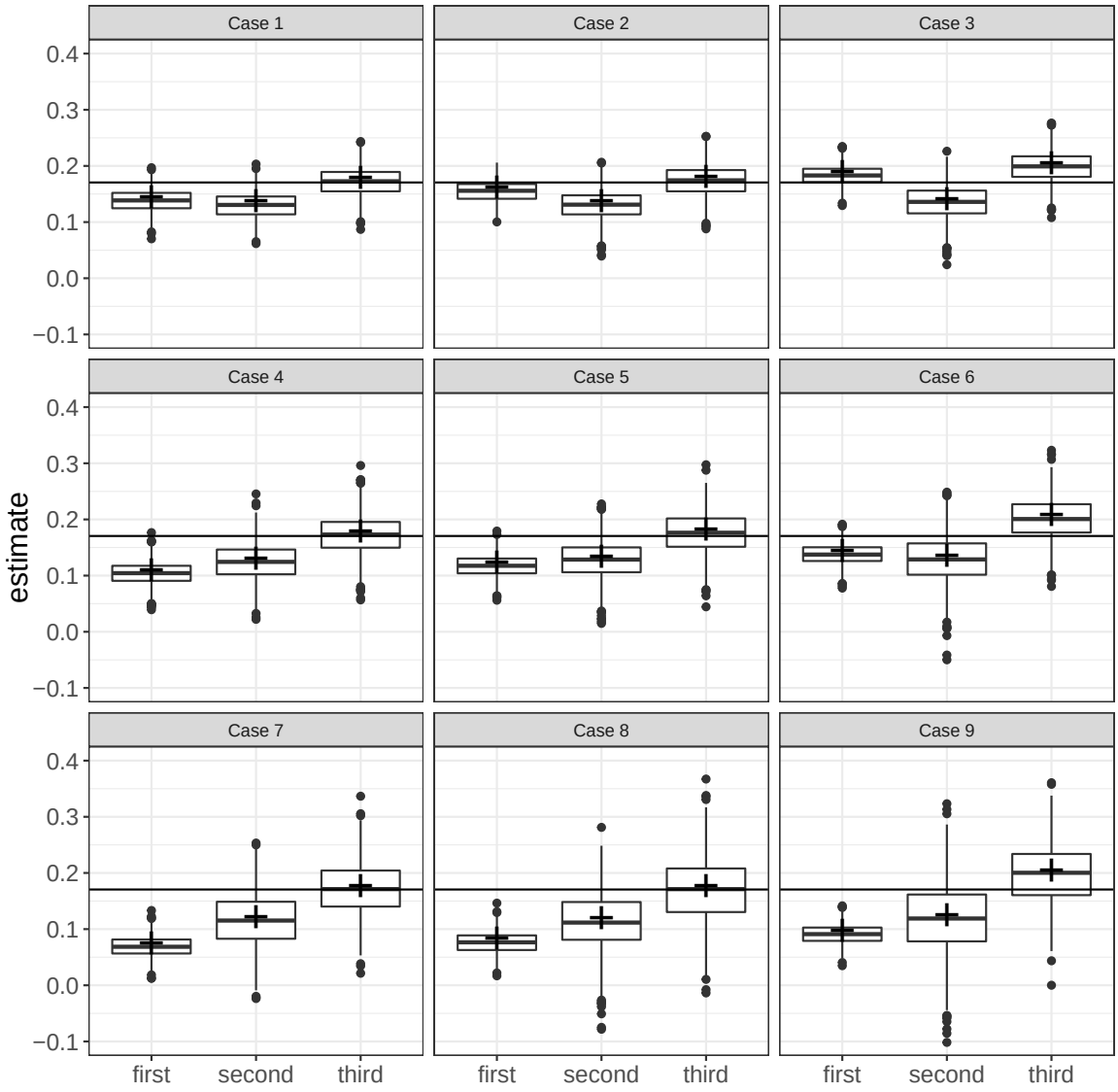


Figure 5.11: Simulation results for the comparison of the two proposed methods and the naive analysis: treatment model (5.20) in Setting 1 with $n = 5000$

first: the naive analysis; second: the proposed method in Section 5.3.1 (LCM); third: the proposed method in Section 5.3.2 (ASIMEX).
Cases 1, 2 and 3: $(p_{11}, p_{10}) = (0.9, 0.1)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 4, 5 and 6: $(p_{11}, p_{10}) = (0.8, 0.2)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 7, 8 and 9: $(p_{11}, p_{10}) = (0.7, 0.3)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
$+$*: average of estimates across 1000 runs; The horizontal solid line indicates* τ_0.

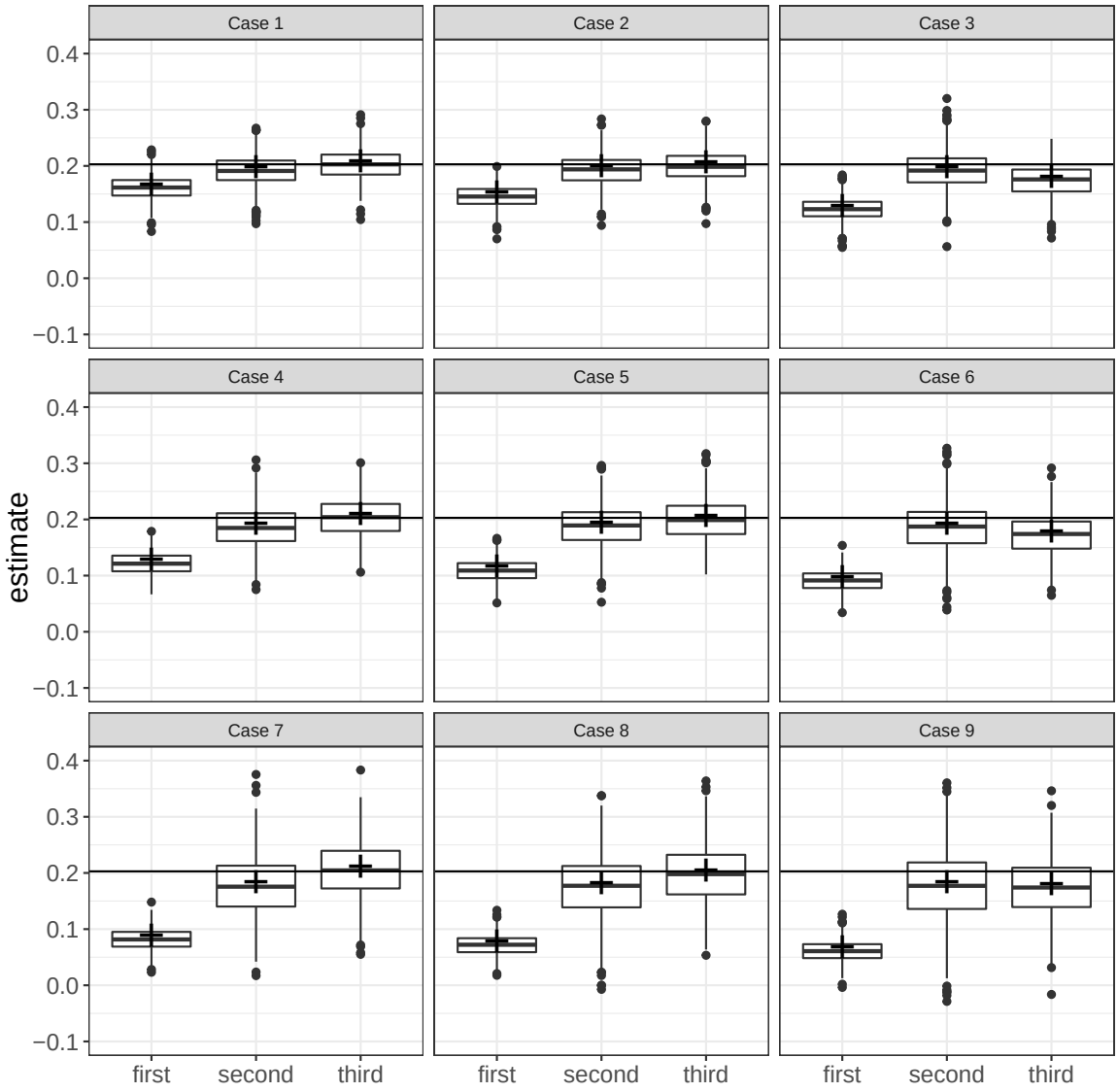


Figure 5.12: Simulation results for the comparison of the two proposed methods and the naive analysis: treatment model (5.20) in Setting 2 with $n = 5000$

first: the naive analysis; second: the proposed method in Section 5.3.1 (LCM); third: the proposed method in Section 5.3.2 (ASIMEX).
Cases 1, 2 and 3: $(p_{11}, p_{10}) = (0.9, 0.1)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 4, 5 and 6: $(p_{11}, p_{10}) = (0.8, 0.2)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
Cases 7, 8 and 9: $(p_{11}, p_{10}) = (0.7, 0.3)$, $\sigma_\epsilon = 0.1$, *0.5, and 1.0;*
$+$*: average of estimates across 1000 runs; The horizontal solid line indicates* τ_0.

Chapter 6

Weighting-Based Causal Inference with Missingness and Misclassification in Outcomes

This chapter deals with Problem 5 discussed in Section 1.5. Section 6.1 describes the IPW estimation with complete and error-free data. Section 6.2 presents the models for missingness and misclassification in outcome variables. In Section 6.3 we first derive the asymptotic bias caused by ignoring missingness, misclassification or both, and then propose two IPW based correction methods to eliminate missingness and mismeasurements effects simultaneously. In Section 6.4 we develop a doubly robust correction method to provide protection against misspecification of the treatment model. In Section 6.5 we conduct simulation studies to assess the finite sample performance of the proposed methods. As an application, in Section 6.6 we analyze a smoking cessation dataset using the proposed methods.

6.1 Notation and Framework

For any subject, let X be the vector of associated covariates and let T be the binary indicator of treatment assignment with $T = 1$ if treated and $T = 0$ if untreated. Let Y_1 be the potential binary outcome that would have been observed had the subject been treated and Y_0 be the potential binary outcome that would have been observed had the subject been untreated; let Y be the observed binary outcome. We assume the causal inference assumptions described in Section 1.1.2 for the following development.

The primary interest is to estimate the average treatment effect (ATE), defined as $\tau_0 = E(Y_1) - E(Y_0)$. With binary outcomes, the ATE can also be interpreted as the causal risk difference $P(Y_1 = 1) - P(Y_0 = 1)$.

Suppose we have a sample of size n. For $i = 1, \ldots, n$, we attach subscript i to X, T, Y_1, Y_0 and Y to denote the corresponding variables for subject i, yielding X_i, T_i, $Y_{i,1}$, $Y_{i,0}$ and Y_i, respectively. The propensity score for subject i, defined as

$$e_i = P(T_i = 1|X_i), \tag{6.1}$$

plays an important role in causal inference (Rosenbaum and Rubin, 1983), where $i = 1, \ldots, n$. Using propensity scores, Rosenbaum (1987, 1998) proposed the IPW estimator for the ATE:

$$\hat{\tau} = \hat{E}(Y_1) - \hat{E}(Y_0), \tag{6.2}$$

where

$$\hat{E}(Y_1) = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{e}_i},\ \hat{E}(Y_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\hat{e}_i},$$

and $\hat{e}_i$ is the estimated propensity score for subject i obtained by fitting the treatment model (6.1).

Under the causal inference assumptions described in Section 1.1.2 and that the treatment model (6.1) is correctly specified, $\hat{\tau}$ is a consistent estimator of τ_0 (e.g., Lunceford and Davidian, 2004). However, this consistency property of $\hat{\tau}$ also requires two critical conditions which are tacitly assumed: the variables must be measured precisely and the associated observations must be complete. When these conditions are violated, the consis-

tency of $\hat{\tau}$ no longer holds.

In subsequent sections, we discuss this issue and develop consistent estimators of τ_0 when the outcome variable is subject to both missingness and misclassification.

6.2 Missingness and Misclassification Models

Let R be the missing data indicator with $R = 1$ if the outcome variable is observed and $R = 0$ otherwise. Assume that given the covariates X and treatment variable T, the missing data indicator R and the outcome variable Y are independent, i.e., $R \perp\!\!\!\perp Y|(X, T)$. This assumption aligns with the missing at random (MAR) mechanism which is commonly considered in the non-causal framework (e.g., Little and Rubin, 2002). In other words, we assume that

$$P(R = 1|Y, X, T = t) = P(R = 1|X, T = t) \tag{6.3}$$

for $t = 0, 1$. We let $\pi_t = P(R = 1|X, T = t)$ for $t = 0, 1$.

In addition to being subject to missing, the observed outcome variable Y is subject to misclassification, and we let Y^* denote the actually observed value of Y. We consider situations where the misclassification probabilities are not affected by the covariates X nor the treatment indicator T, provided the true value Y is given, i.e.,

$$P(Y^* = a|Y = b, X, T = t) = P(Y^* = a|Y = b) \tag{6.4}$$

for $a, b, t = 0, 1$. We now write $p_{ab} = P(Y^* = a|Y = b)$ for $a, b = 0, 1$. Model (6.4) is widely used in the literature, with p_{11} and p_{00} often being referred to as sensitivity and specificity, respectively. To highlight the key idea, assume that the p_{ab} are known, but bearing in mind that unknown p_{ab} can be estimated by using validation data or replicates of outcome measurements (e.g., White et al., 2001).

6.3 Bias Analysis and Correction Methods

6.3.1 Bias Analysis

In the presence of missingness and misclassification in Y, (6.2) cannot be directly applied for estimation of τ_0. One may however, be tempted to use the available data to work out an estimator by naively using (6.2). The most naive but simple approach is to ignore both features of missingness and misclassification in the outcome variable. That is, one would apply (6.2) and replace Y_i with $Y_i R_i$ to ensure complete observations and then further replace $Y_i R_i$ with $Y_i^* R_i$ in order to use the observed measurements Y_i^*. Such a method would yield a naive estimator of τ_0:

$$\hat{\tau}^{**} = \hat{E}^{**}(Y_1) - \hat{E}^{**}(Y_0), \tag{6.5}$$

where

$$\hat{E}^{**}(Y_1) = \frac{1}{\sum_{i=1}^n R_i} \sum_{i=1}^n \frac{T_i Y_i^* R_i}{\hat{e}_i}, \ \hat{E}^{**}(Y_0) = \frac{1}{\sum_{i=1}^n R_i} \sum_{i=1}^n \frac{(1-T_i) Y_i^* R_i}{1-\hat{e}_i},$$

and $\hat{e}_i$ is an estimated propensity score for subject i, as described for (6.2).

A "less-naive" approach is to ignore missingness in Y but account for the misclassification effects. That is, we may first consider $\hat{\tau}^{**}$ and then follow Chapter 4 to incorporate misclassification effects by working on

$$\hat{\tau}^* = \hat{\tau}^{**}/(p_{11} - p_{10}). \tag{6.6}$$

Alternatively, another "less-naive" approach ignores misclassification but takes missingness effects into account. That is, one may use $\pi_{it} = P(R_i = 1 | X_i, T_i = t)$ to re-weight the contribution from subject i and then estimate τ_0 by the estimator

$$\tilde{\tau}^* = \frac{1}{n} \sum_{i=1}^n \frac{T_i Y_i^* R_i}{e_i \hat{\pi}_{i1}} - \frac{1}{n} \sum_{i=1}^n \frac{(1-T_i) Y_i^* R_i}{(1-e_i)\hat{\pi}_{i0}}, \tag{6.7}$$

where $\hat{\pi}_{it}$ is an estimated value of π_{it} for $t = 0, 1$.

To understand the differences among these naive estimators, we present the asymptotic bias of $\hat{\tau}^{**}$, $\hat{\tau}^{*}$ and $\tilde{\tau}^{*}$ in the following theorem; the proof is deferred to Appendix E.1.

Theorem 6.1. *Suppose the causal inference assumptions described in Section 1.1.2, the missingness mechanism (6.3) and the misclassification mechanism (6.4) hold. Then the following results are true.*

(a). The asymptotic bias caused by ignoring both misclassification and missingness is

$$\begin{aligned} Bias(\hat{\tau}^{**}) = & \frac{E[\pi_1\{(p_{11}-p_{10})P(Y_1=1|X)+p_{10}\}]}{P(R=1)} \\ & -\frac{E[\pi_0\{(p_{11}-p_{10})P(Y_0=1|X)+p_{10}\}]}{P(R=1)} - \tau_0; \end{aligned}$$

(b). The asymptotic bias caused by ignoring only the missingness is

$$\begin{aligned} Bias(\hat{\tau}^{*}) = & \frac{E[\pi_1\{P(Y_1=1|X)+p_{10}/(p_{11}-p_{10})\}]}{P(R=1)} \\ & -\frac{E[\pi_0\{P(Y_0=1|X)+p_{10}/(p_{11}-p_{10})\}]}{P(R=1)} - \tau_0, \end{aligned}$$

provided $p_{11} \neq p_{10}$;

(c). The asymptotic bias caused by ignoring only the misclassification is

$$Bias(\tilde{\tau}^{*}) = (p_{11}-p_{10}-1)\tau_0.$$

Theorem 6.1 gives that when $p_{11} \neq p_{10}$,

$$\text{Bias}(\hat{\tau}^{**}) = (p_{11}-p_{10})\text{Bias}(\hat{\tau}^{*}) + \text{Bias}(\tilde{\tau}^{*}). \tag{6.8}$$

Theorem 6.1 shows that the three naive methods incur different degrees of biases in estimating τ_0, and (6.8) further reveals how these biases are related. Interestingly, adding up the bias induced from missingness and the bias due to misclassification cannot fully

capture the bias caused from ignoring both missingness and misclassification simultaneously. Instead, the latter bias Bias($\hat{\tau}^{**}$) is a linear combination of Bias($\hat{\tau}^{*}$) and Bias($\tilde{\tau}^{*}$) with unequal coefficients. The unity coefficient of Bias($\tilde{\tau}^{*}$) can be understood that the bias introduced from ignoring misclassification alone amounts to part of the bias Bias($\hat{\tau}^{**}$), but the coefficient, $p_{11} - p_{10}$, of Bias($\hat{\tau}^{*}$) clearly indicates that the missingness effects interact with misclassification probabilities. The identity (6.8) also suggests that Bias($\hat{\tau}^{**}$) can be smaller than both Bias($\hat{\tau}^{*}$) and Bias($\tilde{\tau}^{*}$). That is, there are counter-intuitive situations where ignoring both missingness and misclassification can perform better than merely ignoring one feature.

In terms of the absolute magnitude, (6.8) leads to

$$|\text{Bias}(\hat{\tau}^{**})| \leq |\text{Bias}(\hat{\tau}^{*})| + |\text{Bias}(\tilde{\tau}^{*})|,$$

which says that adding up the bias in the absolute value caused from ignoring one feature can only give an upper bound of the most naive method which disregards both features.

Furthermore, Theorem 6.1(b) suggests that the asymptotic bias of $\hat{\tau}^{*}$ would be zero if the outcomes are missing completely at random (MCAR) with

$$P(R = 1|Y, X, T = t) = P(R = 1)$$

for $t = 0, 1$, which is a circumstance where the feature of missingness can be ignored for the estimation of τ_0. However, Theorem 6.1(c) uncovers that the feature of misclassification cannot be ignored when estimating τ_0 unless $\tau_0 = 0$.

6.3.2 Correction Methods

Using the results in Theorem 6.1, we develop consistent estimators of τ_0 by eliminating the effects of both missingness and misclassification. Specifically, we propose to modify a naive estimator by removing its associated bias that is quantified by Theorem 6.1.

Theorem 6.2. *Suppose that the conditions of Theorem 6.1 hold. Then*

(a). $\hat{\tau} = \hat{E}(Y_1) - \hat{E}(Y_0)$ *is a consistent estimator of* τ_0*, where*

$$\hat{E}(Y_1) = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i^* R_i}{e_i \hat{\pi}_{i1}(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} \tag{6.9}$$

and

$$\hat{E}(Y_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{(1 - T_i) Y_i^* R_i}{(1 - e_i)\hat{\pi}_{i0}(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}}; \tag{6.10}$$

(b). $\tilde{\tau} = \tilde{E}(Y_1) - \tilde{E}(Y_0)$ *is a consistent estimator of* τ_0*, where*

$$\tilde{E}(Y_1) = \left\{\sum_{i=1}^{n} \frac{R_i T_i}{\hat{e}_i \hat{\pi}_{i1}}\right\}^{-1} \sum_{i=1}^{n} \frac{T_i Y_i^* R_i}{e_i \hat{\pi}_{i1}(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} \tag{6.11}$$

and

$$\tilde{E}(Y_0) = \left\{\sum_{i=1}^{n} \frac{R_i(1 - T_i)}{(1 - \hat{e}_i)\hat{\pi}_{i0}}\right\}^{-1} \sum_{i=1}^{n} \frac{(1 - T_i) Y_i^* R_i}{(1 - e_i)\hat{\pi}_{i0}(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}}. \tag{6.12}$$

The proof of Theorem 6.2 is deferred to Appendix E.2. While Theorem 6.2 does not exhaust all consistent estimators of τ_0 which correct for missingness and misclassification in the outcome variable simultaneously, the estimators in Theorem 6.2 are motivated by modifying one of the naive estimators $\hat{\tau}^{**}$, $\hat{\tau}^*$, or $\tilde{\tau}^*$. For instance, Theorem 6.1(c) motivates an estimator of τ_0 to be given by $\tilde{\tau}^*/(p_{11} - p_{10})$, which is exactly $\hat{\tau}$ in Theorem 6.2(a). The estimator $\tilde{\tau}$ in Theorem 6.2(b) comes from the further modification of the factor $1/n$ involved in $\hat{\tau}$ in Theorem 6.2(a) by using the formulation developed by Lunceford and Davidian (2004). Noticing $E(T/e) = 1$ and $E\{(1 - T)/(1 - e)\} = 1$, Lunceford and Davidian (2004) considered to estimate τ_0 by

$$\left(\sum_{i=1}^{n} \frac{T_i}{\hat{e}_i}\right)^{-1} \sum_{i=1}^{n} \frac{T_i Y_i}{\hat{e}_i} - \left(\sum_{i=1}^{n} \frac{1 - T_i}{1 - \hat{e}_i}\right)^{-1} \sum_{i=1}^{n} \frac{(1 - T_i) Y_i}{1 - \hat{e}_i} \tag{6.13}$$

for settings without missingness nor misclassification. Modifying (6.13) by incorporating

the structures of (6.9) and (6.10) gives us the formulation of $\tilde{\tau}$.

6.4 Doubly Robust Estimator

The consistency of the proposed estimators in Theorem 6.2 requires the treatment model be correctly specified. However, in some applications specifying a suitable treatment model for e_i may be difficult, or less likely to postulate the treatment variable than to model the outcome model. To handle such problems with protection against model misspecification, we propose a doubly robust estimator of τ_0, which is consistent even when one of the treatment model and the outcome model is misspecified. To this end, we need to find suitable augmented estimators of $E(Y_1)$ and $E(Y_0)$ to incorporate the information on both the treatment model and the outcome model so that the resulting estimators of $E(Y_1)$ and $E(Y_0)$ are doubly robust.

For $i = 1, \ldots, n$, $t = 0, 1$, $q_{it} = P(Y_i = 1 | T_i = t, X_i)$ be the conditional outcome probabilities, given X_i.

Theorem 6.3. *Let*

$$\hat{E}_{\text{DR}}(Y_1) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{T_i Y_i^* R_i}{\hat{e}_i (p_{11} - p_{10}) \hat{\pi}_{i1}} - \frac{T_i - \hat{e}_i}{\hat{e}_i} \hat{q}_{i1} - \frac{T_i}{\hat{e}_i} \left(\frac{p_{10}}{p_{11} - p_{10}} \right) \right\} \tag{6.14}$$

and

$$\hat{E}_{\text{DR}}(Y_0) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{(1 - T_i) Y_i^* R_i}{(1 - \hat{e}_i)(p_{11} - p_{10}) \hat{\pi}_{i0}} + \frac{T_i - \hat{e}_i}{1 - \hat{e}_i} \hat{q}_{i0} - \frac{1 - T_i}{1 - \hat{e}_i} \left(\frac{p_{10}}{p_{11} - p_{10}} \right) \right\}, \tag{6.15}$$

where for $i = 1, \ldots, n$, $\hat{e}_i$ is an estimate of e_i, $\hat{q}_{it}$ is an estimate of q_{it} for $t = 0, 1$, and $\hat{\pi}_{it}$ is an estimate of π_{it} for $t = 0, 1$. Define

$$\hat{\tau}_{\text{DR}} = \hat{E}_{\text{DR}}(Y_1) - \hat{E}_{\text{DR}}(Y_0). \tag{6.16}$$

Suppose that conditions in Theorem 6.1 hold and the missingness model for π_{it} with $t = 0, 1$ is correctly specified. Then when either the treatment model or the outcome model is correctly specified,

(a). (6.14) and (6.15) are consistent estimators of $E(Y_1)$ and $E(Y_0)$, respectively;

(b). (6.16) is a consistent estimator of τ_0.

The proof of is deferred to Appendix E.3. To use the doubly robust estimator $\hat{\tau}_{\text{DR}}$ to estimate τ_0, we need to calculate $\hat{q}_{i1}$ and $\hat{q}_{i0}$, which however, cannot be directly calculated by fitting the postulated outcome model because of the unavailability of Y when there is missingness or misclassification in Y. To get around this issue, we present a likelihood based approach. Suppose that the outcome model $q_i = P(Y_i = 1|X_i, T_i)$ and the missingness model $\pi_i = P(R_i = 1|X_i, T_i)$ are delineated parametrically; we let $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ denote the respective parameters associated with these two models and now we write q_i as $q_i(\boldsymbol{\beta})$ and π_i as $\pi_i(\alpha)$. Then under assumption (6.3), the observed likelihood function contributed from subject i is given by

$$\begin{aligned}
& L_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\
= \; & \{P(Y_i^*|T_i, X_i; \boldsymbol{\beta})P(R_i = 1|Y_i^*, T_i, X_i; \boldsymbol{\alpha})\}^{R_i} \\
& \times \left\{\sum_y P(Y_i^* = y|T_i, X_i; \boldsymbol{\beta})P(R_i = 0|Y_i^* = y, T_i, X_i; \boldsymbol{\alpha})\right\}^{1-R_i} \\
= \; & P(Y_i^*|T_i, X_i; \boldsymbol{\beta})^{R_i} P(R_i|T_i, X_i; \boldsymbol{\alpha}) \left\{\sum_y P(Y_i^* = y|T_i, X_i; \boldsymbol{\beta})\right\}^{1-R_i} \\
= \; & S_i(\boldsymbol{\beta})^{R_i} \cdot M_i(\boldsymbol{\alpha}) \qquad (6.17)
\end{aligned}$$

where $\sum_y P(Y_i^* = y|T_i, X_i; \boldsymbol{\beta}) = 1$ is used, and $S_i(\boldsymbol{\beta})$ represents the conditional probability $P(Y_i^*|T_i, X_i; \boldsymbol{\beta})$, given by

$$\begin{aligned}
S_i(\boldsymbol{\beta}) = \; & q_i(\boldsymbol{\beta})\{p_{11}Y_i^* + (1 - p_{11})(1 - Y_i^*)\} \\
& + \{1 - q_i(\boldsymbol{\beta})\}\{p_{10}Y_i^* + (1 - p_{10})(1 - Y_i^*)\},
\end{aligned}$$

$M_i(\boldsymbol{\alpha})$ represents the conditional probability $P(R_i|T_i, X_i; \boldsymbol{\alpha})$, given by

$$M_i(\boldsymbol{\alpha}) = \pi_i(\alpha)^{R_i}\{1 - \pi_i(\alpha)\}^{1-R_i}$$

The maximum likelihood estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be obtained by maximizing $\prod_{i=1}^n L_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$

with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ jointly. Clearly, by (6.17),

$$\prod_{i=1}^{n} L_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n} S_i(\boldsymbol{\beta})^{R_i} \prod_{i=1}^{n} M_i(\boldsymbol{\alpha})$$

suggesting that this joint maximization procedure is equivalent to a two-stage approach. At the first stage, the estimator of $\boldsymbol{\alpha}$, denoted by $\hat{\boldsymbol{\alpha}}$, is obtained by maximizing $\prod_{i=1}^{n} M_i(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, and at the second stage, the estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$, is obtained by maximizing $\prod_{i=1}^{n} S_i(\boldsymbol{\beta})^{R_i}$ with respect to $\boldsymbol{\beta}$.

With $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ obtained from the two-stage procedure, it is immediate to calculate $\hat{\pi}_{it}$ and $\hat{q}_{it}$ with $t = 0, 1$ using the specified missingness model and the outcome model.

It is interesting to note that the two-stage procedure here is different from usual two-stage estimation procedures discussed in the literature for which the estimation of $\boldsymbol{\beta}$ at the second stage has to depend on the estimate of $\boldsymbol{\alpha}$ at the first stage; to have a consistent estimator of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ often has to be consistently estimated. However, the estimation of $\boldsymbol{\beta}$ at the second stage here has nothing to do with the estimation of $\boldsymbol{\alpha}$ at the first stage. Estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ can be carried out completely separately, so the consistency of the estimator of $\boldsymbol{\beta}$ is not affected by whether or not $\boldsymbol{\alpha}$ is consistently estimated, and vice versa. This property, however, does not imply that consistently estimating $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ is not important in estimating τ_0 consistently. In fact, by Theorem 6.3, consistent estimation of $\boldsymbol{\alpha}$ is required for consistent estimation of τ_0, and if $\boldsymbol{\beta}$ is not consistently estimated, then, the treatment model parameters must be consistently estimated in order to obtain a consistent estimator of τ_0. We further note that the missingness probabilities π_{it} with $t = 0, 1$ can be handled nonparametrically if there is no feasible model to delineate π_{it} parametrically; the detail can be found in Ning et al. (2018). In such an instance, one may estimate the response model parameter $\boldsymbol{\beta}$ merely using the second stage described here.

6.5 Simulation Studies

We conduct simulation studies to evaluate the finite sample performance of the proposed estimators $\hat{\tau}$, $\tilde{\tau}$ and $\hat{\tau}_{\text{DR}}$ in comparison to the naive estimators $\hat{\tau}^{**}$, $\hat{\tau}^{*}$ and $\tilde{\tau}^{*}$.

For the ith subject of the simulated data with $i = 1, \ldots, n$, let the vector of covariates $X_i = (X_{i1}, X_{i2})^{\mathrm{T}}$ be generated from the bivariate normal distribution with standard normal margins and correlation 0.5. The treatment model is specified as the logistic regression model

$$\text{logit } P(T_i = 1|X_i) = (1, X_i^{\mathrm{T}})\boldsymbol{\gamma}, \tag{6.18}$$

where $\boldsymbol{\gamma}$ is set as $(-0.2, 1.6, -0.4)^{\mathrm{T}}$ and the outcome model is given as the logistic regression model with

$$\text{logit } P(Y_i = 1|T_i, X_i) = (1, T_i, X_i^{\mathrm{T}})\boldsymbol{\beta}. \tag{6.19}$$

where $\boldsymbol{\beta}$ is taken as $(0.5, -1.5, -1, -1)^{\mathrm{T}}$.

For the (non-)missing data indicator R, we consider the model

$$\text{logit } P(R_i = 1|T_i, X_i) = (1, T_i, X_i^{\mathrm{T}})\boldsymbol{\alpha}. \tag{6.20}$$

We consider two settings where $\boldsymbol{\alpha}$ is set as $(-2.85, 1, 0.6, -0.4)^{\mathrm{T}}$ in Setting 1 and $\boldsymbol{\alpha}$ is specified as $(-1.45, 1, 0.6, -0.4)^{\mathrm{T}}$ in Setting 2. Setting 1 and Setting 2 yield about 13.6% and 38.9% probabilities of missing for subject j when $T_j = 1$, $X_j = (0, 0)^{\mathrm{T}}$. The misclassification probabilities (p_{11}, p_{10}) in (6.4) are assumed to be $(0.9, 0.1)$ or $(0.8, 0.2)$.

We consider the following three scenarios.

1. Both the treatment model and the outcome model are correctly specified:

The treatment model (6.18) and the outcome model (6.19), are, respectively, used to generate the treatment variable T and the outcome variable Y. In the estimation of τ_0, we specify (6.18) as the treatment model and (6.19) as the outcome model.

2. Only the treatment model is correctly specified:

The treatment model (6.18) and the outcome model (6.19), are, respectively, used to generate the treatment variable T and the outcome variable Y. In the estimation of τ_0, we specify (6.18) as the treatment model, but the outcome model is mistaken as (6.19) with X_i replaced by X_{i1}.

3. Only the outcome model is correctly specified:

The treatment model (6.18) and the outcome model (6.19), are, respectively, used to generate the treatment variable T and the outcome variable Y. In the estimation of τ_0, we specify (6.19) as the outcome model, but the treatment model is mistaken as (6.18) with X_i replaced by X_{i1}.

Sample sizes $n = 2000$ and $n = 5000$ are considered, and 1000 simulations are run for each parameter setting. The average relative biases in percent (ReBias%), the average bootstrap standard error (ASE), the empirical standard error (ESE) and the 95% coverage percentage (CP%) are reported, where for an estimator $\hat{\vartheta}$, the relative bias is defined to be $(\hat{\vartheta} - \tau_0)/\tau_0$, and the coverage percentage is defined to be the percentage of those 95% confidence intervals $\hat{\vartheta} \mp 1.96 \times \sqrt{\hat{V}ar(\hat{\vartheta})}$ containing τ_0.

Tables 6.1, 6.2, and 6.3 summarize the simulation results. The first type naive estimator $\hat{\tau}^{**}$ produces severely biased results due to no adjustment for the missingness and misclassification effects. The second type naive estimator $\hat{\tau}^{*}$ leads to severely biased results because of the ignorance of missingness effects, or falsely assuming MCAR when the missingness actually depends on the observed treatment and covariates. The third type naive estimator $\tilde{\tau}^{*}$ also produces biased results because of its failure of accounting for misclassification effects. Furthermore, the observed empirical biases for $\hat{\tau}^{**}$, $\hat{\tau}^{*}$ and $\tilde{\tau}^{*}$ in Tables 6.1 and 6.2 support the theoretical results established in (6.8).

When the treatment model is correctly specified, the estimators $\hat{\tau}$ and $\tilde{\tau}$ produce fairly small empirical biases and coverage percentages closed to 95% under various combinations of missing percentages and misclassification probabilities, as anticipated by their consistency. The doubly robust estimator $\hat{\tau}_{\text{DR}}$ yields fairly small empirical biases and coverage percentages closed to 95% when either the treatment model or the outcome model is correctly specified, as expected by Theorem 6.3. Discrepancies between ASE and ESE are fairly small, suggesting that the bootstrap variance estimates are reliable. Shown in Tables 6.1 and 6.2, $\tilde{\tau}$ has the smallest standard errors, $\hat{\tau}$ tends to have the largest standard errors, and $\hat{\tau}_{\text{DR}}$ seems to have in-between standard errors, although the differences are fairly small in many settings. This limited simulation suggests that $\tilde{\tau}$ may be preferred to $\hat{\tau}$ when there is no concern on the specification of the treatment model, and $\hat{\tau}_{\text{DR}}$ is recommended when the treatment model is suspected to be misspecified.

Table 6.1: Simulation results for the evaluation of performance of the proposed estimators $\hat{\tau}$, $\tilde{\tau}$ and $\hat{\tau}_{\text{DR}}$ in comparison to three types of naive estimators $\hat{\tau}^{**}$, $\hat{\tau}^{*}$ and $\tilde{\tau}^{*}$, when both the treatment model and outcome model are correctly specified.

			$n = 2000$				$n = 5000$			
(p_{11}, p_{10})	missing	Est.	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
(0.9,0.1)	Setting 1	$\hat{\tau}^{**}$	-42.27	0.032	0.032	10.6	-42.59	0.020	0.021	0.70
		$\hat{\tau}^{*}$	-27.84	0.040	0.040	60.8	-28.24	0.026	0.026	24.7
		$\tilde{\tau}^{*}$	-19.62	0.033	0.034	65.1	-20.08	0.021	0.022	35.5
		$\hat{\tau}$	0.478	0.042	0.043	96.1	-0.099	0.027	0.028	94.6
		$\tilde{\tau}$	-0.177	0.035	0.034	95.5	-0.234	0.022	0.023	94.9
		$\hat{\tau}_{\text{DR}}$	0.484	0.037	0.037	95.7	-0.134	0.024	0.024	94.3
	Setting 2	$\hat{\tau}^{**}$	-82.36	0.039	0.040	0.60	-82.32	0.025	0.024	0.00
		$\hat{\tau}^{*}$	-77.95	0.049	0.050	4.00	-77.90	0.031	0.031	0.00
		$\tilde{\tau}^{*}$	-20.20	0.043	0.046	70.9	-20.23	0.028	0.028	51.5
		$\hat{\tau}$	-0.251	0.054	0.057	94.1	-0.289	0.035	0.036	94.5
		$\tilde{\tau}$	-0.726	0.041	0.043	94.9	-0.213	0.027	0.026	95.3
		$\hat{\tau}_{\text{DR}}$	-0.217	0.051	0.053	95.2	-0.134	0.033	0.033	95.6
(0.8,0.2)	Setting 1	$\hat{\tau}^{**}$	-61.74	0.034	0.035	1.60	-61.67	0.022	0.022	0.00
		$\hat{\tau}^{*}$	-36.23	0.056	0.058	66.5	-36.12	0.036	0.036	32.0
		$\tilde{\tau}^{*}$	-40.10	0.035	0.036	20.2	-39.86	0.022	0.023	2.70
		$\hat{\tau}$	-0.165	0.058	0.059	95.5	0.234	0.037	0.038	95.7
		$\tilde{\tau}$	-0.552	0.050	0.051	95.4	-0.004	0.032	0.032	95.0
		$\hat{\tau}_{\text{DR}}$	-0.477	0.053	0.054	95.5	0.143	0.034	0.033	95.9
	Setting 2	$\hat{\tau}^{**}$	-100.9	0.041	0.042	0.20	-101.1	0.026	0.026	0.00
		$\hat{\tau}^{*}$	-101.4	0.068	0.069	5.10	-101.9	0.043	0.043	0.00
		$\tilde{\tau}^{*}$	-40.04	0.044	0.050	33.5	-40.31	0.028	0.028	9.00
		$\hat{\tau}$	-0.060	0.074	0.083	95.1	-0.509	0.047	0.047	95.6
		$\tilde{\tau}$	0.143	0.058	0.059	94.2	-0.429	0.038	0.038	95.5
		$\hat{\tau}_{\text{DR}}$	0.063	0.071	0.076	95.0	-0.453	0.044	0.046	94.6

Est.: estimator; ReBias%: average relative bias in percent; ASE: average bootstrap standard error; ESE: empirical standard error; CP%: 95% *coverage percentage.*

Table 6.2: Simulation results for the evaluation of performance of the proposed estimators $\hat{\tau}$, $\tilde{\tau}$ and $\hat{\tau}_{\text{DR}}$ in comparison to three types of naive estimators $\hat{\tau}^{**}$, $\hat{\tau}^{*}$ and $\tilde{\tau}^{*}$, when only the treatment model is correctly specified.

			$n = 2000$				$n = 5000$			
(p_{11}, p_{10})	missing	Est.	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
(0.9,0.1)	Setting 1	$\hat{\tau}^{**}$	-42.28	0.032	0.033	12.8	-42.44	0.020	0.021	0.30
		$\hat{\tau}^{*}$	-27.85	0.040	0.041	58.5	-28.04	0.026	0.026	24.9
		$\tilde{\tau}^{*}$	-19.50	0.033	0.034	65.6	-19.90	0.021	0.021	36.0
		$\hat{\tau}$	0.626	0.042	0.043	94.0	0.120	0.027	0.027	95.0
		$\tilde{\tau}$	0.522	0.035	0.035	94.1	0.111	0.022	0.022	95.5
		$\hat{\tau}_{\text{DR}}$	0.343	0.038	0.038	94.4	0.084	0.024	0.023	95.5
	Setting 2	$\hat{\tau}^{**}$	-81.89	0.039	0.039	0.20	-81.92	0.025	0.024	0.10
		$\hat{\tau}^{*}$	-77.36	0.048	0.048	3.10	-77.40	0.031	0.031	0.10
		$\tilde{\tau}^{*}$	-20.18	0.042	0.045	70.6	-19.78	0.028	0.030	51.2
		$\hat{\tau}$	-0.220	0.053	0.056	94.6	0.271	0.034	0.037	94.1
		$\tilde{\tau}$	-0.104	0.041	0.042	94.7	0.400	0.026	0.027	94.1
		$\hat{\tau}_{\text{DR}}$	-0.085	0.051	0.053	96.0	0.545	0.033	0.034	94.4
(0.8,0.2)	Setting 1	$\hat{\tau}^{**}$	-61.82	0.034	0.035	2.70	-61.94	0.022	0.022	0.10
		$\hat{\tau}^{*}$	-36.36	0.056	0.059	63.9	-36.57	0.036	0.036	29.9
		$\tilde{\tau}^{*}$	-40.02	0.035	0.037	19.4	-40.15	0.022	0.023	2.40
		$\hat{\tau}$	-0.028	0.058	0.061	94.9	-0.255	0.037	0.038	95.0
		$\tilde{\tau}$	-0.079	0.050	0.051	94.3	0.098	0.032	0.031	95.4
		$\hat{\tau}_{\text{DR}}$	-0.030	0.053	0.054	94.6	0.019	0.034	0.033	95.0
	Setting 2	$\hat{\tau}^{**}$	-100.8	0.041	0.041	0.10	-100.4	0.026	0.026	0.00
		$\hat{\tau}^{*}$	-101.3	0.068	0.068	5.50	-100.6	0.043	0.043	0.10
		$\tilde{\tau}^{*}$	-39.79	0.045	0.048	36.7	-39.64	0.028	0.030	11.2
		$\hat{\tau}$	0.346	0.074	0.081	94.4	0.600	0.047	0.049	94.6
		$\tilde{\tau}$	-0.005	0.059	0.061	94.3	0.499	0.037	0.038	95.1
		$\hat{\tau}_{\text{DR}}$	0.671	0.071	0.075	95.3	0.629	0.045	0.047	95.3

Est.: estimator; ReBias%: average relative bias in percent; ASE: average bootstrap standard error; ESE: empirical standard error; CP%: 95% *coverage percentage.*

Table 6.3: Simulation results for the evaluation of performance of the proposed estimators $\hat{\tau}$, $\tilde{\tau}$ and $\hat{\tau}_{\text{DR}}$ in comparison to three types of naive estimators $\hat{\tau}^{**}$, $\hat{\tau}^{*}$ and $\tilde{\tau}^{*}$, when only the outcome model is correctly specified.

			$n = 2000$				$n = 5000$			
(p_{11}, p_{10})	missing	Est.	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
(0.9,0.1)	Setting 1	$\hat{\tau}^{**}$	-27.29	0.030	0.031	39.1	-26.03	0.019	0.020	11.1
		$\hat{\tau}^{*}$	-9.110	0.038	0.039	90.4	-7.533	0.024	0.025	86.6
		$\tilde{\tau}^{*}$	-7.783	0.030	0.031	89.1	-6.640	0.019	0.020	84.9
		$\hat{\tau}$	15.27	0.038	0.039	82.5	16.70	0.024	0.025	60.7
		$\tilde{\tau}$	15.47	0.032	0.033	76.9	16.69	0.021	0.021	48.1
		$\hat{\tau}_{\text{DR}}$	-0.709	0.035	0.035	95.2	0.346	0.022	0.023	94.3
	Setting 2	$\hat{\tau}^{**}$	-61.11	0.037	0.039	3.60	-61.47	0.024	0.024	0.00
		$\hat{\tau}^{*}$	-51.38	0.047	0.049	24.6	-51.84	0.030	0.029	1.20
		$\tilde{\tau}^{*}$	-6.300	0.039	0.041	90.8	-7.147	0.025	0.025	87.0
		$\hat{\tau}$	17.13	0.049	0.052	89.4	16.07	0.031	0.031	79.0
		$\tilde{\tau}$	16.45	0.038	0.039	80.9	16.07	0.024	0.025	63.2
		$\hat{\tau}_{\text{DR}}$	0.686	0.047	0.048	95.4	-0.220	0.030	0.030	94.5
(0.8,0.2)	Setting 1	$\hat{\tau}^{**}$	-48.91	0.033	0.032	4.20	-48.83	0.021	0.020	0.10
		$\hat{\tau}^{*}$	-14.85	0.054	0.053	91.0	-14.71	0.035	0.034	83.0
		$\tilde{\tau}^{*}$	-29.72	0.033	0.032	37.7	-29.79	0.021	0.020	7.30
		$\hat{\tau}$	17.13	0.054	0.054	89.4	17.02	0.034	0.034	78.5
		$\tilde{\tau}$	16.23	0.048	0.047	86.1	16.66	0.030	0.030	74.0
		$\hat{\tau}_{\text{DR}}$	0.301	0.051	0.049	95.4	0.628	0.032	0.032	95.7
	Setting 2	$\hat{\tau}^{**}$	-83.83	0.040	0.041	0.60	-83.31	0.025	0.025	0.00
		$\hat{\tau}^{*}$	-73.05	0.066	0.069	23.8	-72.18	0.042	0.042	1.30
		$\tilde{\tau}^{*}$	-30.09	0.040	0.043	49.9	-29.62	0.026	0.026	20.5
		$\hat{\tau}$	16.52	0.067	0.071	92.6	17.30	0.043	0.043	85.2
		$\tilde{\tau}$	16.47	0.055	0.056	86.7	16.88	0.035	0.035	77.5
		$\hat{\tau}_{\text{DR}}$	-0.223	0.066	0.069	95.4	0.682	0.041	0.042	95.2

Est.: estimator; ReBias%: average relative bias in percent; ASE: average bootstrap standard error; ESE: empirical standard error; CP%: 95% *coverage percentage.*

6.6 Application to Smoking Cessation Data

As an application, we analyze a smoking cessation data using the proposed methods. The dataset was collected from a study on the effectiveness of a perioperative smoking cessation program (Lee et al., 2013) for which 168 patients were recruited with 30-day follow up. We consider the baseline covariates gender, age, body mass index, diabetes status, hypertension, chronic obstructive pulmonary disease, cigarettes per day, the number of years of smoking, and the exhaled carbon monoxide (CO) level. The indicator of the smoking cessation intervention is taken as the treatment variable and the outcome variable is the indicator of smoking cessation for previous 7 days at the 30-day follow-up postoperatively.

The follow-up data were collected from telephone interview. Lee et al. (2013) pointed out that the collected outcomes were self-reported which were likely to be subject to misclassification. In addition, outcome measurements for 18 patients, 7 in the treated group and 11 in the control group, were missing. Our primary interest is to estimate the ATE (i.e., τ_0) on smoking cessation for previous 7 days at the follow-up.

The naive analysis without correcting for missingness and misclassification effects yields $\hat{\tau}^{**} = 0.178$ with standard error 0.063, and hence a 95% confidence interval for τ_0 is (0.055, 0.301). These results suggest that there is a significant causal effect of the smoking cessation intervention on reducing smoking.

To investigate the effects of missingness and misclassification effects, we apply the proposed methods to analyze the data. Because individuals who had quit smoking were unlikely to report that they still smoked, so we assume that $p_{11} = 1$ (e.g., Magder and Hughes, 1997). To see what misclassification probability p_{10} might be, we take the information available in Lee et al. (2013) as a reference point. Lee et al. (2013) collected self-reported data on the smoking cessation status for seven days preoperatively and the exhaled CO levels.

In total, there were 146 patients with exhaled CO levels greater than 10 ppm. Among them, 11 patients self-reported no smoking. If we treat exhaled CO level <=10 ppm as a gold standard for smoking cessation, then these data yield a misclassification rate $11/146 = 7.5\%$. Using 7.5% as a reference, we consider multiple values around 7.5%

for p_{10} to see how sensitive the results will be under various degrees of misclassification. Specifically, we take p_{10} to be a value of 5.0%, 7.5%, 10.0% or 15.0%.

The missingness model is specified as the logistic model (6.20) and the treatment model is specified as the logistic model (6.18). When applying the doubly robust method, the outcome model is specified as the logistic model (6.19).

Table 6.4 summarizes the analyses results which can be obtained from the proposed estimators $\hat{\tau}$,$\tilde{\tau}$ and $\hat{\tau}_{\text{DR}}$ as well as the three types of naive estimators $\hat{\tau}^{**}$, $\hat{\tau}^{*}$ and $\tilde{\tau}^{*}$. The results obtained from the proposed estimators $\hat{\tau}$ and $\tilde{\tau}$ are similar to that obtained from the estimator $\hat{\tau}^{*}$ which ignores missingness but differ from those obtained from $\hat{\tau}^{**}$ and $\tilde{\tau}^{*}$ with the differences become larger as p_{10} increases, suggesting missingness effects are negligible whereas misclassification effects are not. The doubly robust estimator $\hat{\tau}_{\text{DR}}$ provides larger estimates and standard errors than $\hat{\tau}$ and $\tilde{\tau}$ do. While the results are different for different estimators, all the methods reveal evidence of significant causal effects of the intervention on smoking cessation.

Table 6.4: Analysis results of the smoking cessation data using proposed estimators $\hat{\tau}$, $\tilde{\tau}$ and $\hat{\tau}_{\text{DR}}$ and naive estimators $\hat{\tau}^{**}$, $\hat{\tau}^{*}$ and $\tilde{\tau}^{*}$: estimate (EST), bootstrap standard error (SE) and 95% confidence interval (95% CI)

	$p_{10} = 5\%$			$p_{10} = 7.5\%$		
Method	EST	SE	95% CI	EST	SE	95% CI
$\hat{\tau}^{**}$	0.178	0.063	(0.055, 0.301)	0.178	0.063	(0.055, 0.301)
$\hat{\tau}^{*}$	0.188	0.066	(0.058, 0.318)	0.193	0.071	(0.054, 0.331)
$\tilde{\tau}^{*}$	0.172	0.066	(0.042, 0.301)	0.172	0.066	(0.042, 0.301)
$\hat{\tau}$	0.181	0.067	(0.050, 0.311)	0.185	0.076	(0.037, 0.334)
$\tilde{\tau}$	0.185	0.066	(0.055, 0.315)	0.190	0.071	(0.050, 0.330)
$\hat{\tau}_{\text{DR}}$	0.194	0.072	(0.053, 0.335)	0.204	0.080	(0.047, 0.361)
	$p_{10} = 10\%$			$p_{10} = 15\%$		
Method	EST	SE	95% CI	EST	SE	95% CI
$\hat{\tau}^{**}$	0.178	0.063	(0.055, 0.301)	0.178	0.063	(0.055, 0.301)
$\hat{\tau}^{*}$	0.198	0.074	(0.054, 0.342)	0.210	0.074	(0.064, 0.356)
$\tilde{\tau}^{*}$	0.172	0.066	(0.042, 0.301)	0.172	0.066	(0.042, 0.301)
$\hat{\tau}$	0.191	0.074	(0.046, 0.335)	0.202	0.077	(0.051, 0.353)
$\tilde{\tau}$	0.195	0.074	(0.050, 0.341)	0.207	0.076	(0.058, 0.355)
$\hat{\tau}_{\text{DR}}$	0.207	0.082	(0.047, 0.368)	0.226	0.084	(0.062, 0.391)

Chapter 7

Multiply Robust Estimation of Causal Effects with Outcomes Subject to Both Misclassification and Missingness

This chapter deals with Problem 6 discussed in Section 1.5. Section 7.1 describes the notation and framework in the absence of misclassification and missingness. Sections 7.2-7.4 propose multiple robust estimation methods in situations where only misclassification occurs, only missingness occurs, and both misclassification and missingness occur, respectively. In Section 7.5, simulation studies are conducted to evaluate the finite sample performance of the proposed methods. Section 7.6 presents an application of the proposed method to the smoking cessation data.

7.1 Notation and Framework

For any subject, let X be the vector of covariates and T be the binary indicator of treatment assignment with $T = 1$ if treated and $T = 0$ otherwise. Let Y be the observed binary

outcome, let Y_1 be the potential outcome that would have been observed had the subject been treated and Y_0 be the potential outcome that would have been observed had the subject been untreated. Assume fundamental causal inference assumptions described in Section 1.1.2 hold.

The quantity of primary interest is the average treatment effect (ATE) defined as follows:

$$\tau_0 = E(Y_1) - E(Y_0). \tag{7.1}$$

With binary outcome, τ_0 can also be interpreted as as the causal risk difference. Naturally, it is also of interest to estimate other commonly-used causal effect measures such as the causal risk ratio and the causal odds ratio given by

$$\psi_{\text{RR}} = \frac{E(Y_1)}{E(Y_0)}, \tag{7.2}$$

and

$$\psi_{\text{OR}} = \frac{E(Y_1)/\{1 - E(Y_1)\}}{E(Y_0)/\{1 - E(Y_0)\}}, \tag{7.3}$$

respectively.

To consistently estimate τ_0, ψ_{RR} and ψ_{OR}, it suffices to consistently estimate $E(Y_1)$ and $E(Y_0)$ in (7.1), (7.2) and (7.3). Therefore, although this chapter focuses on the ATE, our development covers the consistent estimation of $E(Y_1)$ and $E(Y_0)$, which can be immediately applied to estimate the causal risk ratio and the causal odds ratio.

Suppose we have a sample of size n. For $i = 1, \ldots, n$, subscript i in notations will be used to denote the corresponding variables for subject i. Without loss of generality, suppose subjects $i = 1, \ldots, m$ are assigned to take the treatment while subjects $i = m + 1, \ldots, n$ are untreated, where m is the size of treatment group.

Han and Wang (2013) proposed an empirical likelihood (Qin and Lawless, 1994; Owen, 2001) based estimation methods for the estimation of sample mean with missing data. Their method can be directly applied to estimate $E(Y_1)$ and $E(Y_0)$ and enjoys the property called multiple robustness, i.e., the resulting causal estimators are consistent when either the set of multiple postulated treatment models or the set of postulated outcome models

contains a correctly specified model. The estimation procedure is briefly described as follows.

Let $e(X) = P(T = 1|X)$ be the true treatment model, also termed as the propensity score (Rosenbaum and Rubin, 1983). Let $q_t(X) = P(Y = 1|X, T = t)$ be the true outcome probabilities for $t = 0, 1$. Suppose $\mathcal{E} = \{e^j(\boldsymbol{\gamma}^j; X), j = 1 \ldots, J\}$ is a set of J postulated treatment models, where $\boldsymbol{\gamma}^j$ is the regression parameter for the jth treatment model. Suppose $\mathcal{Q} = \{q_t^k(\boldsymbol{\beta}^k; X), k = 1 \ldots, K\}$ is a set of K postulated outcome models, where $\boldsymbol{\beta}^k$ is the regression parameter for the kth outcome model. Let $\hat{\boldsymbol{\gamma}}^j$ be the estimator of $\boldsymbol{\gamma}^j$ obtained by fitting the jth treatment model, and $\hat{\boldsymbol{\beta}}^k$ be the estimator of $\boldsymbol{\beta}^k$ obtained by fitting the kth outcome model. Write $\hat{\boldsymbol{\gamma}} = \{\hat{\boldsymbol{\gamma}}^j : j = 1, \ldots, J\}$ and $\hat{\boldsymbol{\beta}} = \{\hat{\boldsymbol{\beta}}^k : j = 1, \ldots, K\}$.

Define $\hat{\theta}^j = n^{-1} \sum_{i=1}^n e^j(\hat{\boldsymbol{\gamma}}^j; X_i)$ for $j = 1, \ldots, J$, $\hat{\eta}_1^k = n^{-1} \sum_{i=1}^n q_1^k(\hat{\boldsymbol{\beta}}^k; X_i)$ for $k = 1, \ldots, K$, and

$$\hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) = (e^1(\hat{\boldsymbol{\gamma}}^1; X_i) - \hat{\theta}^1, \ldots, e^J(\hat{\boldsymbol{\gamma}}^J; X_i) - \hat{\theta}^J, q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) - \hat{\eta}_1^1, \ldots, q_1^K(\hat{\boldsymbol{\beta}}^K; X_i) - \hat{\eta}_1^K)^{\mathrm{T}}.$$

By the proposed method of Han and Wang (2013), $E(Y_1)$ can be estimated by

$$\hat{E}(Y_1) = \sum_{i=1}^m \hat{w}_i Y_i, \tag{7.4}$$

where

$$\hat{w}_i = \arg\max_{w_i} \prod_{i=1}^m w_i$$

subject to constraints

$$\sum_{i=1}^m w_i = 1, \quad \sum_{i=1}^m w_i \hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) = \mathbf{0},$$

Han and Wang (2013) showed that for $i = 1 \ldots, m$,

$$\hat{w}_i = \left\{ \frac{1}{m} \frac{1}{1 + \hat{\rho}^{\mathrm{T}} \hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})} \right\} \Big/ \left\{ \frac{1}{m} \sum_{i=1}^m \frac{1}{1 + \hat{\rho}^{\mathrm{T}} \hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})} \right\} \tag{7.5}$$

where $\hat{\rho} = (\hat{\rho}_1, \ldots, \hat{\rho}_{J+K})^{\mathrm{T}}$ is a $(J+K) \times 1$ vector solving the following equation for ρ:

$$\sum_{i=1}^{m} \frac{\hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})}{1 + \rho^{\mathrm{T}} \hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})} = \mathbf{0}. \tag{7.6}$$

Similarly, we define $\hat{\eta}_0^k = n^{-1} \sum_{i=1}^{n} q_0^k(\hat{\boldsymbol{\beta}}^k; X_i)$ for $k = 1, \ldots, K$ and

$$\hat{h}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) = (\hat{\theta}^1 - e^1(\hat{\boldsymbol{\gamma}}^1; X_i), \ldots, \hat{\theta}^J - e^J(\hat{\boldsymbol{\gamma}}^J; X_i), q_0^1(\hat{\boldsymbol{\beta}}^1; X_i) - \hat{\eta}_0^1, \ldots, q_0^K(\hat{\boldsymbol{\beta}}^K; X_i) - \hat{\eta}_0^K)^{\mathrm{T}}.$$

By the symmetry between $E(Y_1)$ and $E(Y_0)$, $E(Y_0)$ can be estimated by

$$\hat{E}(Y_0) = \sum_{i=m+1}^{n} \tilde{w}_i Y_i, \tag{7.7}$$

where

$$\tilde{w}_i = \arg\max_{w_i} \prod_{i=m+1}^{n} w_i$$

subject to constraints

$$\sum_{i=m+1}^{n} w_i = 1, \quad \sum_{i=m+1}^{n} w_i \hat{h}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) = 0,$$

Similarly, for $i = m+1 \ldots, n$,

$$\tilde{w}_i = \left\{ \frac{1}{n-m} \frac{1}{1 + \hat{\delta}^{\mathrm{T}} \hat{h}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})} \right\} \Bigg/ \left\{ \frac{1}{n-m} \sum_{i=m+1}^{n} \frac{1}{1 + \hat{\delta}^{\mathrm{T}} \hat{h}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})} \right\} \tag{7.8}$$

where $\hat{\delta} = (\hat{\delta}_1, \ldots, \hat{\delta}_{J+K})^{\mathrm{T}}$ is a $(J+K) \times 1$ vector solving the following equation for δ:

$$\sum_{i=m+1}^{n} \frac{\hat{h}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})}{1 + \delta^{\mathrm{T}} \hat{h}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})} = \mathbf{0}. \tag{7.9}$$

Under regularity conditions, (7.4) is a consistent estimator of $E(Y_1)$ when either $\mathcal{E}$ or $\mathcal{Q}$ contains a correctly specified model (Han and Wang, 2013). Similarly, (7.7) is a consistent estimator of $E(Y_0)$ when either $\mathcal{E}$ or $\mathcal{Q}$ contains a correctly specified model. This approach provides more protection against model misspecification than the doubly robust estimation;

this property is called multiple robustness.

However, this property of multiple robustness also requires a critical condition: the variables are measured completely and precisely. In many applications, measurement error problems and/or missing data frequently occur and may jeopardize the multiple robustness of (7.4) and (7.7).

In subsequent sections, we develop multiply robust estimators for $E(Y_1)$ and $E(Y_0)$ when the outcome variable is subject to only the misclassification, only the missingness, and both.

7.2 Multiply Robust Estimation Accommodating Outcome Misclassification

In the section, we consider situations where only the outcome misclassification occurs.

7.2.1 Misclassification Model

Suppose the outcome variable is subject to misclassification, and let Y^* be the actually observed value of Y. We consider situations where the misclassification probabilities are not determined by the covariates X nor the treatment indicator T, conditioning on the true value Y, that is

$$P(Y^* = a|Y = b, X, T = t) = P(Y^* = a|Y = b) \tag{7.10}$$

for $a, b, t = 0, 1$. Write $p_{ab} = P(Y^* = a|Y = b)$ for $a, b = 0, 1$. Model (7.10) is widely used in the literature, with p_{11} and p_{00} often being referred to as sensitivity and specificity, respectively. To highlight the key idea, assume that the p_{ab} are known, but bearing in mind that unknown p_{ab} can be estimated by using validation data or replicates of outcome measurements (e.g., White et al., 2001).

7.2.2 Correction Method

In the presence of outcome misclassification, we observe that (7.4) and (7.7) cannot be applied directly because the true value Y is unobserved. Moreover, by (7.5), (7.6), (7.8) and (7.9), the weights $\hat{w}_i$ for $i = 1, \ldots m$ and $\tilde{w}_i$ for $i = m+1, \ldots n$ cannot be directly calculated because we cannot fit the postulated outcome models in $\mathcal{Q}$ directly with unobserved Y. As a result, the estimation of $\hat{\boldsymbol{\beta}}^k$, $(k = 1, \ldots, K)$ is challenged by the misclassification effect. Fortunately, the estimation of $\hat{\boldsymbol{\gamma}}^j$, $(j = 1, \ldots, J)$ is unaffected by misclassification, because the treatment model does not involve Y.

To eliminate the misclassification effect, we propose a five-step correction approach, where efforts are taken in Steps 2 and 4 to correct for outcome misclassification.

Step 1 (Obtain $\hat{\boldsymbol{\gamma}}^j$ for $j = 1, \ldots, J$):

We obtain $\hat{\boldsymbol{\gamma}}^j$ by fitting the jth postulated treatment model directly.

Step 2 (Obtain $\hat{\boldsymbol{\beta}}^k$ for $k = 1, \ldots, K$):

We obtain $\hat{\boldsymbol{\beta}}^k$ by maximizing the observed likelihood rather than naively fitting the postulated outcome model with Y^* regarded as the true value Y. For the kth postulated outcome model, the observed likelihood function contributed from subject i is

$$\begin{aligned} L_i^k(\boldsymbol{\beta}^k) &= q_{T_i}^k(\boldsymbol{\beta}^k; X_i)\{p_{11}Y_i^* + (1-p_{11})(1-Y_i^*)\} \\ &\quad +\{1 - q_{T_i}^k(\boldsymbol{\beta}^k; X_i)\}\{p_{10}Y_i^* + (1-p_{10})(1-Y_i^*)\}, \end{aligned}$$

Maximizing $\prod_{i=1}^n L_i^k(\boldsymbol{\beta}^k)$ gives $\hat{\boldsymbol{\beta}}^k$.

Step 3 (Obtain weights $\hat{w}_i$ for $i = 1, \ldots m$ and $\tilde{w}_i$ for $i = m+1, \ldots n$):

We calculate $\hat{w}_i$ and $\tilde{w}_i$ using (7.5) and (7.8), respectively, with $\hat{\boldsymbol{\gamma}}^j$ obtained in Step 1 and $\hat{\boldsymbol{\beta}}^k$ obtained in Step 2.

Step 4 (Obtain $\hat{E}(Y_1)$ and $\hat{E}(Y_0)$):

We modify (7.4) and (7.7) and propose to estimate $E(Y_1)$ and $E(Y_0)$ by

$$\hat{E}(Y_1) = \sum_{i=1}^{m} \hat{w}_i \left(\frac{Y_i^*}{p_{11} - p_{10}} \right) - \frac{p_{10}}{p_{11} - p_{10}} \tag{7.11}$$

and

$$\hat{E}(Y_0) = \sum_{i=m+1}^{n} \tilde{w}_i \left(\frac{Y_i^*}{p_{11} - p_{10}} \right) - \frac{p_{10}}{p_{11} - p_{10}}, \tag{7.12}$$

respectively, where $\hat{w}_i$ for $i = 1, \ldots m$ and $\tilde{w}_i$ for $i = m+1, \ldots n$ are obtained in Step 3.

Step 5 (Estimate causal effect):

We estimate the causal risk difference by

$$\hat{\tau} = \hat{E}(Y_1) - \hat{E}(Y_0), \tag{7.13}$$

where $\hat{E}(Y_1)$ and $\hat{E}(Y_0)$ are given by (7.11) and (7.12), respectively.

The following theorem establishes the multiple robustness of the proposed estimators and the proof is deferred to Appendix F.1.

Theorem 7.1. *Suppose the causal inference assumptions described in Section 1.1.2 and the misclassification mechanism (7.10) hold. When either $\mathcal{E}$ or $\mathcal{Q}$ contains a correctly specified model,*

(a). $\hat{E}(Y_1)$ given by (7.11) is a consistent estimator of $E(Y_1)$;

(b). $\hat{E}(Y_0)$ given by (7.12) is a consistent estimator of $E(Y_0)$;

(c). $\hat{\tau}$ given by (7.13) is a consistent estimator of τ_0.

7.3 Multiply Robust Estimation Accommodating Outcome Missingness

In the section, we consider situations where only the outcome missingness occurs.

7.3.1 Missingness Model

Let R be the missing data indicator with $R = 1$ if the outcome variable is observed and $R = 0$ otherwise. Assume that given the covariates X and treatment variable T, the missing data indicator R and the outcome variable Y are independent, i.e., $R \perp\!\!\!\perp Y|(X,T)$. This assumption aligns with the missing at random (MAR) mechanism which is commonly considered in the non-causal framework (e.g., Little and Rubin, 2002). In other words, we assume that

$$P(R = r|Y, X, T = t) = P(R = r|X, T = t) \tag{7.14}$$

for $t, r = 0, 1$. We let $\pi_{rt}(\boldsymbol{\alpha}; X) = P(R = r|X, T = t)$ for $r, t = 0, 1$, where $\boldsymbol{\alpha}$ is the regression parameter of missingness model.

7.3.2 Correction Method

In the presence of outcome missingness, we observe that (7.4) and (7.7) cannot be applied directly because the true value Y is subject to missingness. Moreover, The estimation of $\hat{\boldsymbol{\beta}}^k$, $(k = 1, \ldots, K)$ is affected by missingness because the postulated outcome models cannot be fit using data of all subjects. Fortunately, the estimation of $\hat{\boldsymbol{\gamma}}^j$, $(j = 1, \ldots, J)$ is unaffected by missingness, because the treatment model does not involve Y.

Observe that the joint likelihood is

$$\prod_{i=1}^{n} L_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \{\pi_{1T_i}(\boldsymbol{\alpha}; X_i) P(Y_i|T_i, X_i; \boldsymbol{\beta})\}^{R_i} \cdot \pi_{0T_i}(\boldsymbol{\alpha}; X_i)^{1-R_i}$$

can be factorized as

$$\prod_{i=1}^{n} L_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_{R_i T_i}(\boldsymbol{\alpha}; X_i) \prod_{i=1}^{n} P(Y_i|T_i, X_i; \boldsymbol{\beta})^{R_i}.$$

Therefore, the estimation of $\hat{\boldsymbol{\beta}}^k$ can be conducted by directly fitting the postulated outcome model using complete cases, and the estimation of $\boldsymbol{\alpha}$ can be carried out by fitting the missingness model using all sample data.

To eliminate the missingness effect, we propose a five-step correction approach, where efforts are taken in Steps 2 and 4 to correct for outcome missingness.

Step 1 (Obtain $\hat{\boldsymbol{\gamma}}^j$ for $j = 1, \ldots, J$):

We obtain $\hat{\boldsymbol{\gamma}}^j$ by fitting the jth postulated treatment model directly.

Step 2 (Obtain $\hat{\boldsymbol{\beta}}^k$ for $k = 1, \ldots, K$):

We obtain $\hat{\boldsymbol{\beta}}^k$ by fitting the kth postulated outcome model directly using complete cases data.

Step 3 (Obtain weights $\hat{w}_i$ for $i = 1, \ldots m$ and $\tilde{w}_i$ for $i = m+1, \ldots n$):

We calculate $\hat{w}_i$ and $\tilde{w}_i$ using (7.5) and (7.8), respectively, with $\hat{\boldsymbol{\gamma}}^j$ obtained in Step 1 and $\hat{\boldsymbol{\beta}}^k$ obtained in Step 2.

Step 4 (Obtain $\hat{E}(Y_1)$ and $\hat{E}(Y_0)$):

We modify (7.4) and (7.7) and propose to estimate $E(Y_1)$ and $E(Y_0)$ by

$$\hat{E}(Y_1) = \sum_{i=1}^{m} \hat{w}_i Y_i R_i / \pi_{11}(\hat{\boldsymbol{\alpha}}; X_i) \tag{7.15}$$

and

$$\hat{E}(Y_0) = \sum_{i=m+1}^{n} \tilde{w}_i Y_i R_i / \pi_{10}(\hat{\boldsymbol{\alpha}}; X_i) \tag{7.16}$$

respectively, where $\hat{w}_i$ for $i = 1, \ldots m$ and $\tilde{w}_i$ for $i = m+1, \ldots n$ are obtained in Step 3, $\hat{\boldsymbol{\alpha}}$ is obtained by fitting the missingness model (7.14).

Step 5 (Estimate causal effect):

We estimate the causal risk difference by

$$\hat{\tau} = \hat{E}(Y_1) - \hat{E}(Y_0), \tag{7.17}$$

where $\hat{E}(Y_1)$ and $\hat{E}(Y_0)$ are given by (7.15) and (7.16), respectively.

The following theorem establishes the multiple robustness of the proposed estimators and the proof is deferred to Appendix F.2.

Theorem 7.2. *Suppose the causal inference assumptions described in Section 1.1.2 and the missingness mechanism (7.14) hold. When either $\mathcal{E}$ or $\mathcal{Q}$ contains a correctly specified model,*

(a). $\hat{E}(Y_1)$ given by (7.15) is a consistent estimator of $E(Y_1)$;

(b). $\hat{E}(Y_0)$ given by (7.16) is a consistent estimator of $E(Y_0)$;

(c). $\hat{\tau}$ given by (7.17) is a consistent estimator of τ_0.

7.4 Multiply Robust Estimation with Both Misclassification and Missingness Effects Incorporated

In the section, we consider situations where both outcome misclassification and outcome missingness occur. Suppose the misclassification model (7.10) and the missing model (7.14) both hold.

When the outcome variable is subject to both misclassification and missingness, we observe that (7.4) and (7.7) cannot be applied directly. The estimation of $\hat{\boldsymbol{\beta}}^k$, $(k = 1, \ldots, K)$ is affected by missingness and misclassification because the postulated outcome models cannot be fit using data of all subjects, nor using complete cases whose outcome measurements are misclassified. The estimation of $\hat{\boldsymbol{\gamma}}^j$, $(j = 1, \ldots, J)$ is unaffected, because the treatment model does not involve Y.

Observe that the joint likelihood is

$$\prod_{i=1}^{n} L_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \{\pi_{1T_i}(\boldsymbol{\alpha}; X_i) S_i(\boldsymbol{\beta})\}^{R_i} \cdot \pi_{0T_i}(\boldsymbol{\alpha}; X_i)^{1-R_i},$$

where

$$S_i(\boldsymbol{\beta}) = P(Y_i = 1|T_i, X_i; \boldsymbol{\beta})\{p_{11}Y_i^* + (1-p_{11})(1-Y_i^*)\} + P(Y_i = 0|T_i, X_i; \boldsymbol{\beta})\{p_{10}Y_i^* + (1-p_{10})(1-Y_i^*)\}.$$

Note that the joint likelihood can be factorized by

$$\prod_{i=1}^{n} L_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_{R_i T_i}(\boldsymbol{\alpha}; X_i) \prod_{i=1}^{n} S_i(\boldsymbol{\beta})^{R_i}.$$

Therefore, the estimation of $\hat{\boldsymbol{\beta}}^k$ can be conducted by maximizing $\prod_{i=1}^{n} S_i(\boldsymbol{\beta})^{R_i}$, and the estimation of $\boldsymbol{\alpha}$ can be carried out by fitting the missingness model using all sample data.

To eliminate both the misclassification and missingness effects, we propose a five-step correction approach, where efforts are taken in Steps 2 and 4 to correct for outcome misclassification and missingness.

Step 1 (Obtain $\hat{\boldsymbol{\gamma}}^j$ for $j = 1, \ldots, J$):

We obtain $\hat{\boldsymbol{\gamma}}^j$ by fitting the jth postulated treatment model directly.

Step 2 (Obtain $\hat{\boldsymbol{\beta}}^k$ for $k = 1, \ldots, K$):

We obtain $\hat{\boldsymbol{\beta}}^k$ by maximizing $\prod_{i=1}^{n} S_i^k(\boldsymbol{\beta}^k)^{R_i}$, where

$$S_i^k(\boldsymbol{\beta}^k) = q_{T_i}^k(\boldsymbol{\beta}^k; X_i)\{p_{11}Y_i^* + (1-p_{11})(1-Y_i^*)\} + \{1 - q_{T_i}^k(\boldsymbol{\beta}^k; X_i)\}\{p_{10}Y_i^* + (1-p_{10})(1-Y_i^*)\},$$

with $q_{T_i}^k(\boldsymbol{\beta}^k; X_i)$ given by the model form of the kth postulated outcome model.

Step 3 (Obtain weights $\hat{w}_i$ for $i = 1, \ldots m$ and $\tilde{w}_i$ for $i = m+1, \ldots n$):

We calculate $\hat{w}_i$ and $\tilde{w}_i$ using (7.5) and (7.8), respectively, with $\hat{\boldsymbol{\gamma}}^j$ obtained in Step 1 and $\hat{\boldsymbol{\beta}}^k$ obtained in Step 2.

Step 4 (Obtain $\hat{E}(Y_1)$ and $\hat{E}(Y_0)$):

We modify (7.4) and (7.7) and propose to estimate $E(Y_1)$ and $E(Y_0)$ by

$$\hat{E}(Y_1) = \sum_{i=1}^{m} \frac{\hat{w}_i Y_i^* R_i}{\pi_{11}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} \tag{7.18}$$

and

$$\hat{E}(Y_0) = \sum_{i=m+1}^{n} \frac{\tilde{w}_i Y_i^* R_i}{\pi_{10}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} \tag{7.19}$$

respectively, where $\hat{w}_i$ for $i = 1, \ldots m$ and $\tilde{w}_i$ for $i = m+1, \ldots n$ are obtained in Step 3, $\hat{\boldsymbol{\alpha}}$ is obtained by fitting the missingness model (7.14).

Step 5 (Estimate causal effect):

We estimate the causal risk difference by

$$\hat{\tau} = \hat{E}(Y_1) - \hat{E}(Y_0), \tag{7.20}$$

where $\hat{E}(Y_1)$ and $\hat{E}(Y_0)$ are given by (7.18) and (7.19), respectively.

The following theorem establishes the multiple robustness of the proposed estimators and the proof is deferred to Appendix F.3.

Theorem 7.3. *Suppose the causal inference assumptions described in Section 1.1.2, the misclassification mechanism (7.10) and the missingness mechanism (7.14) hold. When either $\mathcal{E}$ or $\mathcal{Q}$ contains a correctly specified model,*

(a). $\hat{E}(Y_1)$ given by (7.18) is a consistent estimator of $E(Y_1)$;

(b). $\hat{E}(Y_0)$ given by (7.19) is a consistent estimator of $E(Y_0)$;

(c). $\hat{\tau}$ given by (7.20) is a consistent estimator of τ_0.

7.5 Simulation Studies

In this section, we conduct simulation studies to assess the finite sample performance of correction methods developed in Sections 7.2-7.4.

7.5.1 Simulation Setup

Let $X = (X_1, X_2)^{\mathrm{T}}$, where X_1 and X_2 are independent and following the uniform distribution ranging from -2 to 2 and the standard normal distribution, respectively. The true

treatment model is a logistic regression model with

$$\text{logit } P(T = 1|X_1, X_2) = 0.2 + X_1 + 0.5X_2,$$

and the true outcome model is a logistic regression model with

$$\text{logit } P(Y = 1|T, X_1, X_2) = -0.5 + T - 0.3X_1 + X_2.$$

Consider the following three scenarios for model specification.

I. one postulated treatment model is correctly specified and none of the postulated outcome models is correctly specified:

The two postulated treatment models are

$$\text{logit } P(T = 1|X_1, X_2) = (1, X_1, X_2)\boldsymbol{\gamma}^1 \quad (\checkmark)$$

and

$$\text{logit } P(T = 1|X_1) = (1, X_1)\boldsymbol{\gamma}^2 \quad (\text{✗})$$

The two postulated outcome models are

$$\text{logit } P(Y = 1|T, X_1) = (1, T, X_1)\boldsymbol{\beta}^1 \quad (\text{✗})$$

and

$$\text{logit } P(Y = 1|T, X_2) = (1, T, X_2)\boldsymbol{\beta}^2 \quad (\text{✗})$$

II. one postulated outcome model is correctly specified and none of the postulated treatment models is correctly specified:

The two postulated treatment models are

$$\text{logit } P(T = 1|X_1) = (1, X_1)\boldsymbol{\gamma}^1 \quad (\text{✗})$$

and

$$\text{logit } P(T = 1|X_2) = (1, X_2)\boldsymbol{\gamma}^2 \quad (\text{✗})$$

The two postulated outcome models are

$$\text{logit } P(Y=1|T,X_1,X_2)=(1,T,X_1,X_2)\boldsymbol{\beta}^1 \quad (\checkmark)$$

and

$$\text{logit } P(Y=1|T,X_1)=(1,T,X_1)\boldsymbol{\beta}^2 \quad (\text{✗})$$

III. one postulated treatment model and one postulated outcome model are correctly specified:

The two postulated treatment models are

$$\text{logit } P(T=1|X_1,X_2)=(1,X_1,X_2)\boldsymbol{\gamma}^1 \quad (\checkmark)$$

and

$$\text{logit } P(T=1|X_2)=(1,X_2)\boldsymbol{\gamma}^2 \quad (\text{✗})$$

The two postulated outcome models are

$$\text{logit } P(Y=1|T,X_1,X_2)=(1,T,X_1,X_2)\boldsymbol{\beta}^1 \quad (\checkmark)$$

and

$$\text{logit } P(Y=1|T,X_2)=(1,T,X_2)\boldsymbol{\beta}^2 \quad (\text{✗})$$

The subsequent sections 7.5.2, 7.5.3 and 7.5.4 conduct simulations for situations where only misclassification occurs, only missingness occurs, and both misclassification and missingness occur, respectively. Consider sample sizes $n = 2000$ and $n = 5000$, and 1000 simulations runs for each configuration. The average relative bias in percent (ReBias%), average bootstrap standard error (ASE), empirical standard error (ESE) and 95% coverage percentage (CP%) are reported. For estimator $\hat{\vartheta}$, the relative bias in percent is defined to be $(\hat{\vartheta}-\tau_0)/\tau_0 \times 100\%$, and the coverage percentage is defined to be the percentage of the 95% bootstrap percentile confidence intervals which contain τ_0.

7.5.2 Only Misclassification Occurs

We compare the performance of the proposed estimator $\hat{\tau}$ developed in Section 7.2 to the naive estimator $\hat{\tau}^*$ which ignores misclassification effects and regards the Y_i^* as if they were true data. The misclassification probabilities (p_{11}, p_{10}) are specified as $(0.9, 0.1)$, $(0.8, 0.2)$ and $(0.7, 0.3)$ to cover different degrees of misclassification.

Table 7.1 summarizes the simulation results under various combinations of misclassification probabilities and scenarios for model specification. The naive analysis produces severely biased results, and its performance becomes worse with the degree of misclassification. These results suggest that ignoring the misclassification effect can result in the loss of multiple robustness. The proposed estimator presents satisfactory performance with negligible finite sample bias for all combinations of misclassification probabilities and scenarios for model specification, as anticipated from its multiple robustness. The discrepancy between ASE and ESE is fairly small, and empirical coverage percentages are close to 95%, indicating that the bootstrap variance estimates and the bootstrap percentile confidence intervals are reliable.

7.5.3 Only Missingness Occurs

We compare the performance of the proposed estimator $\hat{\tau}$ developed in Section 7.3 to the naive estimator $\hat{\tau}^*$ which ignores missingness effects. Specifically, $\hat{\tau}^*$ shares the same estimation procedures with $\hat{\tau}$ except for Step 4, where $E(Y_1)$ and $E(Y_0)$ are estimated by

$$\hat{E}^*(Y_1) = \left(\sum_{i=1}^{m} \hat{w}_i R_i \right)^{-1} \sum_{i=1}^{m} \hat{w}_i Y_i R_i$$

and

$$\hat{E}^*(Y_0) = \left(\sum_{i=m+1}^{n} \tilde{w}_i R_i \right)^{-1} \sum_{i=m+1}^{n} \tilde{w}_i Y_i R_i.$$

Then calculate $\hat{\tau}^* = \hat{E}^*(Y_1) - \hat{E}^*(Y_0)$.

The missingness mechanism is specified as logit $P(R = 1|X, T = t) = \alpha_0 - T + 0.5X_1 -$

Table 7.1: Simulation results comparing the proposed estimator $\hat{\tau}$ in Section 7.2 to the naive estimator $\hat{\tau}^*$, when only outcome misclassification occurs

			$n = 2000$				$n = 5000$			
(p_{11}, p_{10})	Setting	Est	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
(0.9,0.1)	I	$\hat{\tau}^*$	-19.57	0.026	0.026	69.8	-19.94	0.017	0.017	32.4
		$\hat{\tau}$	0.549	0.033	0.033	94.7	0.075	0.021	0.021	94.7
	II	$\hat{\tau}^*$	-19.99	0.028	0.029	68.8	-19.88	0.018	0.019	41.0
		$\hat{\tau}$	0.011	0.035	0.036	94.2	0.186	0.023	0.024	94.4
	III	$\hat{\tau}^*$	-19.36	0.026	0.026	67.9	-20.34	0.016	0.016	29.8
		$\hat{\tau}$	0.843	0.032	0.032	95.0	-0.423	0.020	0.020	95.1
(0.8,0.2)	I	$\hat{\tau}^*$	-40.01	0.027	0.027	14.1	-42.05	0.017	0.017	0.20
		$\hat{\tau}$	-1.716	0.045	0.045	95.0	-0.092	0.028	0.028	94.8
	II	$\hat{\tau}^*$	-40.78	0.028	0.029	17.7	-40.07	0.019	0.020	0.80
		$\hat{\tau}$	-1.155	0.047	0.049	95.7	-0.043	0.032	0.034	95.3
	III	$\hat{\tau}^*$	-40.69	0.027	0.027	13.8	-40.04	0.017	0.017	0.30
		$\hat{\tau}$	-1.063	0.045	0.044	94.3	-0.045	0.028	0.030	93.1
(0.7,0.3)	I	$\hat{\tau}^*$	-59.76	0.028	0.028	1.50	-60.39	0.017	0.019	0.10
		$\hat{\tau}$	0.641	0.069	0.071	93.6	-0.897	0.044	0.046	93.7
	II	$\hat{\tau}^*$	-60.32	0.029	0.030	1.50	-60.47	0.019	0.020	0.00
		$\hat{\tau}$	-0.484	0.073	0.076	94.9	-0.591	0.048	0.051	94.2
	III	$\hat{\tau}^*$	-60.16	0.028	0.028	0.60	-60.21	0.017	0.018	0.00
		$\hat{\tau}$	-0.276	0.068	0.068	94.5	-0.526	0.043	0.044	94.0

Est: estimator; ReBias%: average relative bias in percent; ASE: average bootstrap standard error; ESE: empirical standard error; CP%: 95% *coverage percentage;*
I: one postulated treatment model is correct and no postulated outcome models are correct;
II: one postulated outcome model is correct and no postulated treatment models are correct;
III: one postulated treatment model and one postulated outcome model are correct.

$0.6X_2$. The parameter α_0 is set to be 3, 1.5 and 0.5 such that there are approximately 10%, 30% and 50% subjects with missing outcomes, respectively.

Table 7.2 summarizes the simulation results under various combinations of missingness rates and scenarios for model specification. The naive analysis leads to biased results and the performance becomes worse as the degree of missingness increases. These results demonstrate that ignoring the missingness effect can result in the loss of multiple robustness. The proposed estimator presents satisfactory performance with negligible finite sample bias for all combinations of missingness models and scenarios for model specification, as expected from its multiple robustness. The discrepancy between ASE and ESE is fairly small and the empirical coverage percentages are close to 95%, indicating that the bootstrap variance estimates and the bootstrap percentile confidence intervals are reliable.

7.5.4 Both Misclassification and Missingness Occur

We compare the performance of the proposed estimator $\hat{\tau}$ developed in Section 7.4 to the naive estimator $\hat{\tau}^*$ which ignores misclassification and missingness effects. Specifically, the first three steps in estimating $\hat{\tau}^*$ is Steps 1 to 3 in Section 7.3.2 with Y_i replaced by Y_i^*. In the fourth step, $E(Y_1)$ and $E(Y_0)$ are estimated by

$$\hat{E}^*(Y_1) = \left(\sum_{i=1}^{m} \hat{w}_i R_i\right)^{-1} \sum_{i=1}^{m} \hat{w}_i Y_i^* R_i$$

and

$$\hat{E}^*(Y_0) = \left(\sum_{i=m+1}^{n} \tilde{w}_i R_i\right)^{-1} \sum_{i=m+1}^{n} \tilde{w}_i Y_i^* R_i.$$

Then calculate $\hat{\tau}^* = \hat{E}^*(Y_1) - \hat{E}^*(Y_0)$.

The misclassification probabilities (p_{11}, p_{10}) are specified as $(0.9, 0.1)$ and $(0.8, 0.2)$. The missingness mechanism is the same as in Section 7.5.3, with α_0 set to be 3 and 1.5 to yield approximately 10% and 30% missingness rates, respectively.

Table 7.3 summarizes the simulation results under various combinations of misclassification probabilities, missingness rates and scenarios for model specification. The naive

Table 7.2: Simulation results comparing the proposed estimator $\hat{\tau}$ in Section 7.3 to the naive estimator $\hat{\tau}^*$, when only outcome missingness occurs

			$n = 2000$				$n = 5000$			
Missing	Setting	Est	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
10%	I	$\hat{\tau}^*$	-4.808	0.027	0.028	93.3	-4.683	0.017	0.017	91.3
		$\hat{\tau}$	0.520	0.027	0.028	94.5	0.589	0.017	0.017	94.5
	II	$\hat{\tau}^*$	-5.336	0.028	0.030	92.0	-5.129	0.019	0.019	90.6
		$\hat{\tau}$	0.006	0.029	0.031	94.4	0.283	0.019	0.020	95.2
	III	$\hat{\tau}^*$	-4.902	0.027	0.026	93.1	-5.461	0.017	0.017	90.0
		$\hat{\tau}$	0.320	0.027	0.026	95.5	-0.263	0.017	0.017	94.8
30%	I	$\hat{\tau}^*$	-11.82	0.031	0.031	87.1	-11.11	0.020	0.020	78.6
		$\hat{\tau}$	-0.214	0.035	0.035	95.9	0.369	0.022	0.023	93.9
	II	$\hat{\tau}^*$	-11.75	0.033	0.035	88.5	-12.33	0.022	0.024	78.3
		$\hat{\tau}$	0.201	0.037	0.039	94.9	-0.066	0.024	0.027	94.2
	III	$\hat{\tau}^*$	-10.82	0.031	0.031	88.6	-11.38	0.019	0.020	75.8
		$\hat{\tau}$	1.152	0.035	0.035	94.1	0.042	0.022	0.022	94.4
50%	I	$\hat{\tau}^*$	-15.09	0.037	0.038	88.0	-14.84	0.024	0.024	75.2
		$\hat{\tau}$	-0.076	0.051	0.051	95.3	-0.018	0.032	0.031	94.9
	II	$\hat{\tau}^*$	-15.18	0.040	0.043	86.3	-15.69	0.026	0.029	76.4
		$\hat{\tau}$	-0.018	0.053	0.054	94.1	-0.356	0.035	0.037	94.3
	III	$\hat{\tau}^*$	-14.18	0.037	0.037	87.6	-14.72	0.023	0.025	73.7
		$\hat{\tau}$	0.173	0.050	0.051	94.7	0.437	0.032	0.033	93.5

Est: estimator; ReBias%: average relative bias in percent; ASE: average bootstrap standard error; ESE: empirical standard error; CP%: 95% *coverage percentage;*
I: one postulated treatment model is correct and no postulated outcome models are correct;
II: one postulated outcome model is correct and no postulated treatment models are correct;
III: one postulated treatment model and one postulated outcome model are correct.

analysis leads to severely biased results, suggesting that the effects of misclassification and missingness can degrade the validity of multiply robust estimation. The proposed estimator demonstrate satisfactory performance with reasonably small finite sample bias for all combinations of misclassification probabilities, missingness degrees and scenarios for model specification, as expected from its multiple robustness. The discrepancy between ASE and ESE is fairly small and the empirical coverage percentages are close to 95%, indicating that the bootstrap variance estimates and the bootstrap percentile confidence intervals are reliable.

7.6 Application to Smoking Cessation Data

For illustration, in this section we apply the proposed method to the smoking cessation data arising from a study on the effectiveness of a perioperative smoking cessation program (Lee et al., 2013). The dataset includes 168 patients with baseline covariates gender, age, body mass index, diabetes status, chronic obstructive pulmonary disease, hypertension, cigarettes per day, the number of years of smoking, and the exhaled carbon monoxide (CO) level. The treatment variable is an indicator of taking the smoking cessation intervention. The outcome variable is the indicator of smoking cessation for previous 7 days at the 30-day follow-up postoperatively. Our primary goal is to estimate the causal effect of treatment on smoking cessation for previous 7 days at the 30-day follow-up postoperatively.

Among the 168 patients, 18 (10.7%) of them had missing outcomes. For the rest of the patients who had complete outcome data, misclassification is a real concern. The outcome measurements at the 30-day follow-up were self-reported via telephone interview without verification by biochemical tests (Lee et al., 2013). As a result, the smoking cessation status data were subject to misclassification. The effects of missingness and misclassification, put together, present a challenge to the causal inference. In order to obtain valid inference results, causal estimation should take both the missingness and misclassification into account.

We postulate two treatment models both linking the treatment variable to the baseline covariates. The first postulated treatment model assumes the logit link, while the second

Table 7.3: Simulation results comparing the proposed estimator $\hat{\tau}$ in Section 7.4 to the naive estimator $\hat{\tau}^*$, when both outcome misclassification and outcome missingness occur

				$n = 2000$				$n = 5000$			
(p_{11}, p_{10})	Missing	Setting	Est	ReBias%	ASE	ESE	CP%	ReBias%	ASE	ESE	CP%
(0.9,0.1)	10%	I	$\hat{\tau}^*$	-24.90	0.028	0.027	57.2	-24.23	0.018	0.017	19.1
			$\hat{\tau}$	-0.695	0.035	0.034	95.6	-0.123	0.022	0.022	94.3
		II	$\hat{\tau}^*$	-23.57	0.029	0.031	63.0	-24.46	0.019	0.021	30.0
			$\hat{\tau}$	0.871	0.037	0.040	93.5	-0.254	0.025	0.028	94.2
		III	$\hat{\tau}^*$	-24.56	0.028	0.028	56.2	-23.84	0.017	0.017	21.6
			$\hat{\tau}$	-0.360	0.035	0.036	94.7	0.418	0.022	0.022	95.6
(0.9,0.1)	30%	I	$\hat{\tau}^*$	-29.21	0.032	0.033	54.7	-28.91	0.020	0.020	16.7
			$\hat{\tau}$	0.414	0.045	0.045	95.5	0.418	0.028	0.028	94.2
		II	$\hat{\tau}^*$	-28.60	0.034	0.035	57.9	-30.01	0.022	0.023	22.0
			$\hat{\tau}$	1.194	0.047	0.049	94.2	-0.494	0.031	0.033	94.2
		III	$\hat{\tau}^*$	-29.65	0.032	0.032	54.7	-29.03	0.020	0.022	16.6
			$\hat{\tau}$	-0.646	0.044	0.044	95.3	0.314	0.028	0.028	94.8
(0.8,0.2)	10%	I	$\hat{\tau}^*$	-42.76	0.029	0.029	14.4	-43.27	0.018	0.018	0.00
			$\hat{\tau}$	0.703	0.048	0.048	94.1	-0.172	0.030	0.030	94.1
		II	$\hat{\tau}^*$	-43.45	0.030	0.032	17.4	-43.34	0.020	0.021	1.20
			$\hat{\tau}$	-0.497	0.051	0.054	94.1	-0.009	0.034	0.037	93.6
		III	$\hat{\tau}^*$	-43.14	0.029	0.029	14.4	-43.32	0.018	0.018	0.20
			$\hat{\tau}$	-0.054	0.048	0.047	94.9	-0.426	0.030	0.030	95.3
(0.8,0.2)	30%	I	$\hat{\tau}^*$	-46.75	0.033	0.033	17.4	-46.07	0.021	0.021	0.70
			$\hat{\tau}$	0.157	0.060	0.060	94.1	1.111	0.038	0.037	94.9
		II	$\hat{\tau}^*$	-48.53	0.034	0.037	19.3	-47.84	0.022	0.024	0.90
			$\hat{\tau}$	-1.507	0.063	0.067	93.8	-0.383	0.042	0.047	93.9
		III	$\hat{\tau}^*$	-47.10	0.033	0.033	16.5	-47.46	0.021	0.021	0.10
			$\hat{\tau}$	0.211	0.060	0.059	95.7	-1.060	0.038	0.038	94.5

Est: estimator; ReBias%: average relative bias in percent; ASE: average bootstrap standard error; ESE: empirical standard error; CP%: 95% *coverage percentage;*
I: one postulated treatment model is correct and no postulated outcome models are correct;
II: one postulated outcome model is correct and no postulated treatment models are correct;
III: one postulated treatment model and one postulated outcome model are correct.

postulated treatment model assumes the complementary log-log link. Similarly, we postulate two outcome models both relating the outcome variable to the treatment variable and the baseline covariates, where one model assumes the logit link and the other assumes the complementary log-log link.

We first analyze the data using the naive method which ignores both the missingness and misclassification in outcomes. The resultant estimated causal risk difference is $\hat{\tau}^* = 0.171$ with bootstrap standard error 0.063 and 95% bootstrap percentile confidence interval (0.046, 0.310), suggesting a significant causal effect of the smoking cessation intervention on reducing smoking rate.

To correct for both the missingness and misclassification effects, we apply the proposed method in Section 7.4 to this dataset. It is reasonable to assume $p_{11} = 1$ given that those who had quit smoking were unlikely to report that they still smoked. However, it is likely that $p_{10} > 0$ because subjects who still smoked might report that they had quit smoking (e.g., Magder and Hughes, 1997). Lee et al. (2013) collected self-reported smoking cessation data preoperatively along with the exhaled CO levels, with an exhaled CO of $\leq$10 ppm confirming the self-reported smoking cessation. Among the 146 patients with exhaled CO of >10 ppm, 11 patients had exhaled CO of >10 ppm despite self-reported smoking cessation. These preoperatively collected data yield a misclassification rate $11/146 = 0.075$. It is reasonable to assume the preoperative misclassification mechanism is similar to the postoperative misclassification mechanism. Therefore, we specify $p_{10} = 0.075$ when applying the proposed method.

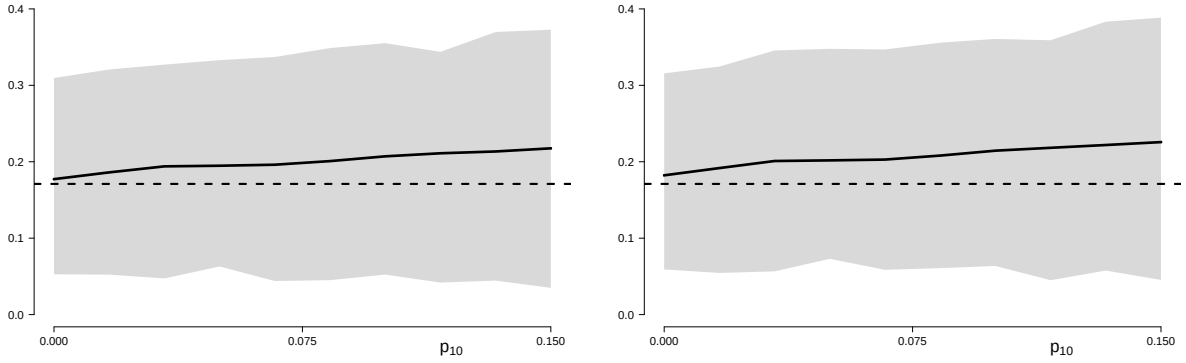
The missingness model is fit using the logistic model relating non-missingness indicator to the treatment variable and baseline covariates. After adjustment for the missingness and misclassification effects, the proposed method yields estimated causal risk difference $\hat{\tau} = 0.202$ with bootstrap standard error 0.084 and 95% bootstrap percentile confidence interval (0.050, 0.348), confirming the statistical significance of the causal effect of the smoking cessation intervention. The comparison of the results obtained using the naive method and those obtained using the proposed method reveals an attenuation effect of missingness and misclassification. Magder and Hughes (1997) specified the falsely reported smoking cessation probability $p_{10} = 0.1$ in the data analysis for another study. If we borrow this value for our analysis, the estimated causal risk difference will be $\hat{\tau} = 0.207$ with

bootstrap standard error 0.084 and 95% bootstrap percentile confidence interval (0.034, 0.356). These results reveal the attenuation effect of misclassification.

We further repeat the adjustment assuming the missing completely at random (MCAR) mechanism $P(R = 1|X, T, Y) = P(R = 1)$, or equivalently, the missingness model is fit using the logistic model relating non-missingness indicator R to 1. If we specify $p_{10} = 0.075$, the resulting estimated causal risk difference $\hat{\tau} = 0.209$ with bootstrap standard error 0.072 and 95% bootstrap percentile confidence interval (0.068, 0.346). If we specify $p_{10} = 0.1$, the resulting estimated causal risk difference $\hat{\tau} = 0.214$ with bootstrap standard error 0.082 and 95% bootstrap percentile confidence interval (0.057, 0.362).

We compare the adjustment fitting logit $P(R = 1|X, T) \sim T + X$ and that fitting logit $P(R = 1|X, T) \sim 1$. The latter produces larger estimates, but the difference is small. Therefore, inference results are similar under MAR and MCAR.

Figure 7.1 further displays the adjustment results under MAR and MCAR, for P_{10} between 0 and 0.15. We use the solid line to denote the adjusted estimate and the grey region to represent the 95% confidence interval. The causal estimate becomes larger as the misclassification probability p_{10} increases. The confidence interval becomes wider as well, indicated by the shape of grey regions. The dashed line of 0.171 (i.e., naive estimate) is below the solid line (adjusted estimates), suggesting the attenuation effect of misclassification and missingness. The grey region is above 0, indicating the statistical significance of the causal effect. Comparing the left and right plots demonstrates that assuming MAR and MCAR yields similar results.



assume MAR: fit logit $P(R = 1|X, T) \sim T + X$ assume MCAR: fit logit $P(R = 1|X, T) \sim 1$

Figure 7.1: Sensitivity analyses of the smoking cessation data using the proposed estimator $\hat{\tau}$ developed in Section 7.4 under various p_{10}. The solid line represents the estimates and the grey region represents the 95% confidence intervals. The dashed line represents the naive estimate which ignores misclassification and missingness.

Chapter 8

ipwErrorY: An R Package for Estimating Average Treatment Effects with Outcome Misclassification

8.1 Introduction

In Chapter 4 we explore estimation of the average treatment effect (ATE) with outcomes subject to measurement error. We derive the asymptotic bias caused by misclassification and developed consistent estimation methods to eliminate the misclassification effects. The development covers practical scenarios where (1) the misclassification probabilities are known, or (2) the misclassification probabilities are unknown but validation data or replicates of outcome measurements are available for their estimation. We further propose a doubly robust estimator to provide protection against possible misspecification of the treatment model.

These methods enjoy wide applications, because misclassified outcome data arise commonly in practice. For example, the self-reported smoking status without being confirmed

by biochemical tests is subject to misclassification; results of screening tests are often subject to false positive error and/or false negative error. For datasets with outcome misclassification, ignoring misclassification effects may lead to severely biased results. To expedite the application of the correction methods for general users, we develop an R (R Core Team, 2017) package called **ipwErrorY** (Shu and Yi, 2018c), to implement the methods developed in Chapter 4 for practical settings where the commonly-used logistic regression model is employed for the treatment model and the outcome model.

The remainder is organized as follows. Section 8.2 presents the implementation details. Section 8.3 illustrates the use of the package with examples. The developed R package **ipwErrorY** is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=ipwErrorY`.

8.2 Implementation in R

In this section we describe our developed R package. The developed package imports R packages **stats** (R Core Team, 2017) and **nleqslv** (Hasselman, 2016).

8.2.1 Implementation with Known Error

The function `KnownError` produces the ATE estimate using the correction method described in Section 4.3 along with the standard error and $(1-\alpha)100\%$ confidence interval. The details of function `KnownError` are given by

```
KnownError(data, indA, indYerror, indX, sensitivity, specificity,
           confidence=0.95)
```

with arguments described as follows:

- `data`: the dataset to be analyzed in the form of R data frame;
- `indA`: the column name which indicates the treatment variable;

- `indYerror`: the column name which indicates the misclassified outcome variable;
- `indX`: the vector of column names of covariates included in the treatment model;
- `sensitivity`: the specified sensitivity (i.e., p_{11}) between 0 and 1;
- `specificity`: the specified specificity (i.e., $1 - p_{10}$) between 0 and 1;
- `confidence`: The confidence level (i.e., $1 - \alpha$) between 0 and 1; the default is 0.95 corresponding to a 95% confidence interval.

8.2.2 Implementation with Validation Data

The function `EstValidation` produces the results for the method described in Section 4.4.1; they include the optimal linear combination estimate, the standard error, $(1-\alpha)100\%$ confidence interval and the estimated sensitivity and specificity. The details of function `EstValidation` are given by

```
EstValidation(maindata, validationdata, indA, indYerror, indX, indY,
              confidence=0.95)
```

with arguments described as follows:

- `maindata`: the non-validation main data in the form of R data frame;
- `validationdata`: the validation data in the form of R data frame;
- `indA`: the column name which indicates the treatment variable;
- `indYerror`: the column name which indicates the misclassified outcome variable;
- `indX`: the vector of column names of covariates included in the treatment model;
- `indY`: the column name which indicates the precisely measured outcome variable;
- `confidence`: The confidence level (i.e., $1 - \alpha$) between 0 and 1; the default is 0.95 corresponding to a 95% confidence interval.

8.2.3 Implementation with Replicates

The function `Est2Replicates` produces the results for the method described in Section 4.4.2 with a constraint imposed; they include the estimate, the standard error, $(1-\alpha)100\%$ confidence interval, and the imposed constraint(s), and the information on sensitivity and specificity. The details of function `Est2Replicates` are given by

```
Est2Replicates(data, indA, indYerror, indX, constraint=c("sensitivity equals
 specificity", "known sensitivity", "known specificity", "known prevalence"),
 sensitivity = NULL, specificity = NULL, prevalence = NULL, confidence=0.95)
```

with arguments described as follows:

- `data`: the dataset to be analyzed in the form of R data frame;
- `indA`: the column name which indicates the treatment variable;
- `indYerror`: the vector of two column names indicating the replicates of the outcome;
- `indX`: the vector of column names of covariates included in the treatment model;
- `constraint`: the imposed constraint with sensitivity equals specificity by default;
- `sensitivity`: the specified sensitivity between 0 and 1 when imposing the constraint that sensitivity is known, and the default is set to be `NULL`;
- `specificity`: the specified specificity between 0 and 1 when imposing the constraint that specificity is known, and the default is set to be `NULL`;
- `prevalence`: the specified prevalence between 0 and 1 when imposing the constraint that prevalence is known, and the default is set to be `NULL`;
- `confidence`: The confidence level (i.e., $1-\alpha$) between 0 and 1; the default is 0.95 corresponding to a 95% confidence interval.

8.2.4 Implementation of Doubly Robust Estimation

The function `KnownErrorDR` produces the ATE estimate using the doubly robust correction method described in Section 4.6 along with the standard error and $(1-\alpha)100\%$ confidence interval. The details of function `KnownErrorDR` are given by

```
KnownErrorDR(data, indA, indYerror, indXtrt, indXout, sensitivity,
    specificity, numBoot, sharePara=FALSE, confidence=0.95)
```

with arguments described as follows:

- `data`: the dataset to be analyzed in the form of R data frame;
- `indA`: the column name which indicates the treatment variable;
- `indYerror`: the column name which indicates the misclassified outcome variable;
- `indXtrt`: the vector of column names indicating the covariates that are included in the treatment model;
- `indXout`: the vector of column names indicating the covariates that are included in the outcome model;
- `sensitivity`: the specified sensitivity (i.e., p_{11}) between 0 and 1;
- `specificity`: the specified specificity (i.e., $1-p_{10}$) between 0 and 1;
- `numBoot`: the specified number of bootstrap replicates for variance estimation;
- `sharePara=FALSE`: if the treated and untreated groups share parameters for X in the logistic outcome model (i.e., assuming $Y \sim T + X$), then set `sharePara=TRUE`; if not (i.e., modeling $Y \sim X$ for the treated and untreated groups separately), then set `sharePara=FALSE`. By default, `sharePara=FALSE`;
- `confidence`: The confidence level (i.e., $1-\alpha$) between 0 and 1; the default is 0.95 corresponding to a 95% confidence interval.

8.3 Examples

To illustrate the use of the developed R package **ipwErrorY**, for each method, we simulate a dataset and then apply a function described in Section 8.2 to analyze the dataset. To make sure users can reproduce the results, we use the function `set.seed` to generate data. Moreover, the simulated data provide users a clear sense about the data structure.

8.3.1 Example with Known Error

We first load the package in R:

```
R> library("ipwErrorY")
```

Using sensitivity 0.95 and specificity 0.85, we create `da`, a dataset of size 2000 with "`X1`", "`A`" and "`Yast`" being the column names for the covariate, treatment and misclassified outcome, respectively:

```
R> set.seed(100)
R> X1 = rnorm(2000)
R> A = rbinom(2000, 1, 1/(1 + exp(-0.2 - X1)))
R> Y = rbinom(2000, 1, 1/(1 + exp(-0.2 - A - X1)))
R> y1 = which(Y == 1)
R> y0 = which(Y == 0)
R> Yast = Y
R> Yast[y1] = rbinom(length(y1), 1, 0.95)
R> Yast[y0] = rbinom(length(y0), 1, 0.15)
R> da = data.frame(X1 = X1, A = A,Yast = Yast)
```

By using the function `head`, we print the first six observations of dataset `da` so that the data structure is clearly shown as follows:

```
R> head(da)
```

```
           X1 A Yast
1 -0.50219235 1    1
2  0.13153117 1    1
3 -0.07891709 1    1
4  0.88678481 0    1
5  0.11697127 1    1
6  0.31863009 1    1
```

To apply the method described in Section 4.3 with sensitivity 0.95 and specificity 0.85, we call the developed function `KnownError` and obtain a list of the estimate, the standard error and a 95% confidence interval:

```
R> KnownError(data = da, indA = "A", indYerror = "Yast", indX = "X1",
+     sensitivity = 0.95, specificity = 0.85, confidence=0.95)
$Estimate
[1] 0.1702513

$Std.Error
[1] 0.02944824

$`95% Confidence Interval`
[1] 0.1125338 0.2279688
```

8.3.2 Example with Validation Data

Using sensitivity 0.95 and specificity 0.85, we create `mainda` which is the non-validation main data of size 1200, and `validationda` which is the validation data of size 800:

```
R> set.seed(100)
R> X1= rnorm(1200)
R> A = rbinom(1200, 1, 1/(1 + exp(-0.2 - X1)))
R> Y= rbinom(1200, 1, 1/(1 + exp(-0.2 - A - X1)))
```

```
R> y1 = which(Y == 1)
R> y0 = which(Y==0)
R> Yast = Y
R> Yast[y1] = rbinom(length(y1), 1, 0.95)
R> Yast[y0] = rbinom(length(y0), 1, 0.15)
R> mainda = data.frame(A = A, X1 = X1, Yast = Yast)
R> X1 = rnorm(800)
R> A = rbinom(800, 1, 1/(1 + exp(-0.2 - X1)))
R> Y = rbinom(800, 1, 1/(1 + exp(-0.2 - A - X1)))
R> y1 = which(Y == 1)
R> y0 = which(Y == 0)
R> Yast = Y
R> Yast[y1] = rbinom(length(y1), 1, 0.95)
R> Yast[y0] = rbinom(length(y0), 1, 0.15)
R> validationda = data.frame(A = A, X1 = X1, Y = Y, Yast = Yast)
```

We print the first six observations of non-validation data `mainda` and validation data `validationda`:

```
R> head(mainda)
  A          X1 Yast
1 1 -0.50219235    0
2 0  0.13153117    0
3 1 -0.07891709    1
4 1  0.88678481    1
5 0  0.11697127    1
6 1  0.31863009    1
R> head(validationda)
  A            X1 Y Yast
1 0 -0.0749961081 0    0
2 1 -0.9470827924 1    1
3 1  0.0003758095 1    1
```

```
4 0 -1.5249574007 0    0
5 1  0.0983516474 0    0
6 0 -1.5266078213 1    1
```

The preceding output clearly reveals that the non-validation data and validation data differ in the data structure. The non-validation data `mainda` record measurements of the treatment, covariate and misclassified outcome, indicated by the column names "`A`", "`X1`" and "`Yast`", respectively. In comparison, the validation data `validationda` record measurements of the treatment, covariate, misclassified outcome and the true outcome, indicated by the column names "`A`", "`X1`", "`Yast`", and "`Y`", respectively.

To apply the optimal linear combination method described in Section 4.4.1, we call the developed function `EstValidation` and obtain a list of the estimate, the standard error, a 95% confidence interval, and the estimated sensitivity and specificity:

```
R> EstValidation(maindata = mainda, validationdata = validationda, indA = "A",
+     indYerror = "Yast",  indX = "X1",  indY = "Y", confidence=0.95)
$Estimate
[1] 0.1714068

$Std.Error
[1] 0.02714957

$`95% Confidence Interval`
[1] 0.1181946 0.2246189

$`estimated sensitivity and estimated specificity`
[1] 0.9482072 0.8557047
```

8.3.3 Example with Replicates

Using sensitivity 0.95 and specificity 0.85, we create `da`, a dataset of size 2000 with "`A`", "`X1`", and { "`Yast1`", "`Yast2`"} being the column names for the treatment, covariate, and

two replicates of outcome, respectively:

```
R> set.seed(100)
R> X1 = rnorm(2000)
R> A = rbinom(2000, 1, 1/(1 + exp(-0.2 - X1)))
R> Y = rbinom(2000, 1, 1/(1 + exp(-0.2 - A - X1)))
R> y1 = which(Y == 1)
R> y0 = which(Y == 0)
R> Yast1 = Y
R> Yast1[y1] = rbinom(length(y1), 1, 0.95)
R> Yast1[y0] = rbinom(length(y0), 1, 0.15)
R> Yast2 = Y
R> Yast2[y1] = rbinom(length(y1), 1, 0.95)
R> Yast2[y0] = rbinom(length(y0), 1, 0.15)
R> da = data.frame(A = A, X1 = X1, Yast1 = Yast1, Yast2 = Yast2)
```

By using the function `head`, we print the first six observations of dataset `da` so that the data structure is clearly shown as follows:

```
R> head(da)
  A          X1 Yast1 Yast2
1 1 -0.50219235     1     1
2 1  0.13153117     1     1
3 1 -0.07891709     1     1
4 0  0.88678481     1     0
5 1  0.11697127     1     1
6 1  0.31863009     1     1
```

To apply the method described in Section 4.4.2 with the imposed constraint that specificity equals 0.85, we call the developed function `Est2Replicates` and obtain a list of the estimate, the standard error, a 95% confidence interval, the imposed constraint and the information on sensitivity and specificity:

```
R> Est2Replicates(data = da,  indA = "A", indYerror = c("Yast1", "Yast2"),
+     indX = "X1", constraint = "known specificity", sensitivity = NULL,
+     specificity = 0.85, prevalence = NULL, confidence=0.95)
$Estimate
[1] 0.1908935

$Std.Error
[1] 0.02687287

$`95% Confidence Interval`
[1] 0.1382236 0.2435634

$`imposed constraint`
[1] "known specificity"

$`estimated sensitivity and assumed specificity`
[1] 0.95 0.85
```

8.3.4 Example of Doubly Robust Estimation

Using sensitivity 0.95 and specificity 0.85, we create `da`, a dataset of size 2000 with "`A`", {"`X`", "`xx`"} and "`Yast`" being the column names for the treatment, covariates and misclassified outcome, respectively:

```
R> set.seed(100)
R> X = rnorm(2000)
R> xx = X^2
R> A = rbinom(2000, 1, 1/(1 + exp(-0.1 - X - 0.2*xx)))
R> Y = rbinom(2000, 1, 1/(1 + exp(1 - A - 0.5*X - xx)))
R> y1 = which(Y == 1)
R> y0 = which(Y == 0)
R> Y[y1] = rbinom(length(y1), 1, 0.95)
```

```
R> Y[y0] = rbinom(length(y0), 1, 0.15)
R> Yast = Y
R> da = data.frame(A = A, X = X, xx = xx, Yast = Yast)
```

By using the function `head`, we print the first six observations of dataset `da` so that the data structure is clearly shown as follows:

```
R> head(da)
  A           X          xx Yast
1 1 -0.50219235 0.252197157    1
2 1  0.13153117 0.017300447    1
3 1 -0.07891709 0.006227907    1
4 0  0.88678481 0.786387298    0
5 1  0.11697127 0.013682278    1
6 1  0.31863009 0.101525133    0
```

When applying the doubly robust method described in Section 4.6 with sensitivity 0.95 and specificity 0.85, covariates indicated by column names "`X`" and "`xx`" are both included in the treatment model and the outcome model. The number of bootstrap replicates is specified as 50 for illustrative purpose, but bearing in mind a larger number such as 1000 may be used in applications. Let the outcome model be fitted for the treated and untreated groups separately. We call the developed function `KnownErrorDR` and obtain a list of the estimate, the standard error, and a 95% confidence interval:

```
R> set.seed(100)
R> KnownErrorDR(data = da, indA = "A", indYerror = "Yast",
+    indXtrt = c("X", "xx"),  indXout = c("X", "xx"), sensitivity = 0.95,
+    specificity = 0.85, numBoot = 50, sharePara = FALSE, confidence=0.95)
$Estimate
[1] 0.2099162

$Std.Error
```

```
[1] 0.02696814

$`95% Confidence Interval`
[1] 0.1570597 0.2627728
```

Note that we call the function `set.seed` before the developed function `KnownErrorDR` in order to make sure users can reproduce the results. If `set.seed` is not used, then the variance estimates obtained by different users can differ due to the inner randomness of the bootstrap method.

Chapter 9

Summary and Discussion

In this thesis, we investigate several important research problems regarding causal inference with measurement error as well as other features such as missing values. The results in this thesis have been or will be wrapped up as papers for dissemination. The results in Chapters 2 and 4, respectively, come from the publications by Shu and Yi (2018a) and Shu and Yi (2017); the results in Chapter 3 and Chapter 5, respectively, come from the papers by Shu and Yi (2016) and Shu and Yi (2018d) which are now under revision; the results in Chapters 6 and 7 will be wrapped up for publication shortly; and the material in Chapter 8 has been wrapped up as Shu and Yi (2018b) and submitted for publication. Below we present a summary for each chapter with discussions.

Chapter 2

The odds ratio, the risk ratio and the risk difference are important measures in epidemiological studies to assess the comparative effectiveness of available treatment plans. However, their estimation becomes complicated when confounders are subject to measurement error. In this chapter, we examine the measurement error effects by simulation studies and develop valid estimation methods to correct for measurement error effects on estimation of causal effect measures when confounders, time-independent or time-dependent, are error-contaminated. The proposed methods are easy to implement and are robust in the sense that the distribution of the unobserved true variables is left unspecified.

In principle, all the proposed methods may be used in practice, as they all provide consistent estimators. Theoretical justifications demonstrate that the estimator obtained from the linear combination method has the smallest *asymptotic* variance. However, our numerical studies show that the finite sample performance of the adaptive conditional score method is similar to that of the linear combination method, and therefore can be used as well in applications.

The variance estimation of the linear combination estimator is approximate in a sense that uncertainty in the estimation of c_{opt} is not accounted for. Bootstrap procedures may be helpful to handle this.

Although the measurement error models we consider are commonly used, they cannot cover all practical problems which may have complex underlying measurement error mechanisms. It would be interesting to generalize our development here to feature other measurement error models. A discussion of various measurement error mechanisms in the context of causal inference was given by Hernán and Robins (2016).

Our methods are developed under the assumption that the covariance matrix in the measurement error model is known. However this assumption can be relaxed. The proposed approaches can be extended to accommodating settings where the covariance matrix is unknown and estimated by using additional data, such as repeated measurements and validation data. When the covariance matrix in the measurement error model is unknown and no extra data are available to estimate it, sensitivity analyses can be carried out to assess how sensitive the estimates are to different degrees of measurement error, by considering a series of representative values of the covariance matrix. Analyses can be carried out along the lines of Small and Rosenbaum (2008) and Baiocchi et al. (2010) in combinations with our development here.

Chapter 3

We explore a number of methods for adjusting for measurement error effects in causal inference with time-dependent and error-contaminated confounders. Our simulation studies demonstrate that ignoring measurement error effects can produce biased estimates for causal parameters. The regression calibration method is equivalent to the naive analysis in our setting. The direct SIMEX and refined indirect SIMEX approaches (i.e., DSIMEX

and RISIMEX in Section 3.4) outperform the indirect SIMEX method (i.e., ISIMEX in Section 3.4). The refined regression calibration method (i.e., RRC in Section 3.4) tends to perform better than the SIMEX-based methods (i.e., ISIMEX, DSIMEX and RISIMEX). The refined correction method (i.e., RCM in Section 3.4) performs the best. In application, we recommend to implement the RCM method if Assumption 5 is regarded plausible. The method RRC is preferred to RISIMEX due to its smaller finite sample biases and much less computational burdens. If Assumption 5 is unreasonable, then methods based on it (i.e., RISIMEX, RRC and RCM) may not be preferred. Instead, we can use the DSIMEX method whose implementation is the most straightforward, though time-consuming just like the ISIMEX method.

When the treatment assignment is driven by the observed measurements of confounders instead of the underlying true values, addressing effects of measurement error in confounders may be unnecessary.

The development here assumes that the covariance matrix of measurement error $\mathbf{\Sigma}_{\epsilon k}$ is known, which is applicable when conducting sensitivity analyses, as illustrated in Section 3.4.2. In application, this assumption may not be true, then we need to estimate $\mathbf{\Sigma}_{\epsilon k}$ using additional data sources such as validation data and repeated measurements.

When an external validation sample data is available, i.e., both measurements for $X_i(k)$ and X_{ik}^* are available for some subjects. Using the measurement error model (3.8), $var(X_{ik}^*) = var\{X_i(k)\} + \mathbf{\Sigma}_{\epsilon k}$, we can estimate the unknown covariance matrix $\mathbf{\Sigma}_\epsilon$ as $\widehat{var}(X_{ik}^*) - \widehat{var}\{X_i(k)\}$, where $\widehat{var}(X_{ik}^*)$ and $\widehat{var}\{X_i(k)\}$ are the sample covariance matrices for X_{ik}^* and $X_i(k)$ which can be obtained from the validation data.

In the case where repeated surrogate measurements are available, let X_{ikj}^* denote the jth independent replicate of observed measurements for $X_i(k)$, where $j = 1, 2, \cdots, l_k$ and $l_k \geq 2$. Let $\bar{X}_{ik}^* = l_k^{-1} \sum_{j=1}^{l_k} X_{ikj}^*$. By Carroll et al. (2006), the covariance matrix $\mathbf{\Sigma}_\epsilon$ can be estimated as

$$\hat{\mathbf{\Sigma}}_{\epsilon k} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{l_k} (X_{ikj}^* - \bar{X}_{ik}^*)(X_{ikj}^* - \bar{X}_{ik}^*)^{\mathrm{T}}}{n(l_k - 1)}.$$

With the availability of repeated measurements of confounders, the empirical SIMEX algorithm developed by Devanarayan and Stefanski (2002) can also be adapted for our setting.

Chapter 4

We explore IPW estimation of ATE in the presence of mismeasurement in outcome variables. We investigate the impact of measurement error on the estimation of ATE and develop valid statistical methods to adjust for measurement error by constructing unbiased estimating functions. Several useful results are established. When a continuous outcome variable is mismeasured under the additive measurement error model (4.3), we reveal that the naive analysis still yields a consistent estimator of ATE. When the outcome is binary with misclassification, we find that the naive analysis leads to a biased estimator of ATE and then identify a closed-form for the resulting asymptotic bias. To address measurement error effects, we develop valid estimation procedures for settings where either internal validation data or replicates of outcome variable are available. With validation data, we propose an efficient method for estimation of ATE. The efficiency gain can be substantial relative to usual methods of using validation data. To provide protection against model misspecification, we propose a doubly robust estimator which is consistent even when the treatment model or the outcome model is misspecified. In such development, we assume that the validation subsample is a simple random sample of the main study. When this assumption is not true, the proposed procedures need to be modified properly. Let $S = 1$ if a subject is included in the validation subsample and $S = 0$ otherwise. Since

$$E(Y_1 - Y_0) = P(S = 1)E(Y_1 - Y_0|S = 1) + P(S = 0)E(Y_1 - Y_0|S = 0),$$

to obtain a consistent estimator of ATE, the IPW estimation procedure needs to be stratified by the membership of being in the validation sample or the non-validation sample, and propensity score models have to be formed to reflect the membership information by further conditionally on S, in addition to X.

In the development with binary outcomes, we focus on studying the asymptotic bias for ATE which can be interpreted as causal risk difference $P(Y_1 = 1) - P(Y_0 = 1)$. The same investigation applies to other causal effect measures, such as causal odds ratio $\frac{E(Y_1)/\{1-E(Y_1)\}}{E(Y_0)/\{1-E(Y_0)\}}$ and causal risk ratio $E(Y_1)/E(Y_0)$, to accommodate outcome misclassification. However, the asymptotic bias for the naive estimators of those measures does not possess the same transparent analytic expressions as what we obtain for ATE. Con-

sistent estimation of these measures hinges on consistent estimation of $E(Y_1)$ and $E(Y_0)$, which is immediate by using (C.8) and (C.9) in Appendix C.

Chapter 5

We propose estimation methods to simultaneously address measurement error in covariates and misclassification in outcomes. Under a class of logistic treatment models (5.11) that is usually employed to model the treatment assignment in practice, we develop a fully consistent estimation method. If the treatment model assumes a form different from model (5.11), we further develop an augmented SIMEX method to account for measurement error in covariates and misclassification in outcomes. These proposed methods are straightforward to implement and have a broad scope of applications.

Our development in the previous sections assumes that $\mathbf{\Sigma}_\epsilon$ in the measurement error model (5.3) and p_{ab} in the misclassification model (5.4) are known, which is suitable for performing sensitivity analyses, as illustrated in Section 5.5. In practice, when $\mathbf{\Sigma}_\epsilon$ and p_{ab} are unknown, we can estimate them using validation data or repeated measurements.

Finally, we comment that following our development lines, other estimators alternative to (5.7) and (5.8) can also be developed. For instance, examining $E\{TG(Z, X^*, T)|X, Z\}$ and $E\{(1-T)G(Z, X^*, T)|X, Z\}$ using the conditions (5.9) and (5.10) gives $E\{TG(Z, X^*, T)\} = 1$ and $E\{(1-T)G(Z, X^*, T)\} = 1$, which suggests that

$$\hat{E}(Y_1) = \frac{1}{(p_{11} - p_{10}) \sum_{i=1}^n T_i \hat{G}_i} \sum_{i=1}^n T_i Y_i^* \hat{G}_i - \frac{p_{10}}{p_{11} - p_{10}} \tag{9.1}$$

and

$$\hat{E}(Y_0) = \frac{1}{(p_{11} - p_{10}) \sum_{i=1}^n (1-T_i) \hat{G}_i} \sum_{i=1}^n (1-T_i) Y_i^* \hat{G}_i - \frac{p_{10}}{p_{11} - p_{10}} \tag{9.2}$$

can be used to estimate $E(Y_1)$ and $E(Y_0)$, respectively.

Analogously, noticing $E(T/e) = 1$ and $E\{(1-T)/(1-e)\} = 1$, we obtain alternatives to (5.5) and (5.6), given by

$$\hat{E}(Y_1) = \frac{1}{(p_{11} - p_{10})} \left(\sum_{i=1}^n \frac{T_i}{\hat{e}_i} \right)^{-1} \sum_{i=1}^n \frac{T_i Y_i^*}{\hat{e}_i} - \frac{p_{10}}{p_{11} - p_{10}} \tag{9.3}$$

and

$$\hat{E}(Y_0) = \frac{1}{(p_{11} - p_{10})} \left\{ \sum_{i=1}^{n} \frac{1 - T_i}{1 - \hat{e}_i} \right\}^{-1} \sum_{i=1}^{n} \frac{(1 - T_i) Y_i^*}{1 - \hat{e}_i} - \frac{p_{10}}{p_{11} - p_{10}}, \tag{9.4}$$

respectively. Then the augmented SIMEX method based on (9.3) and (9.4) can be conducted in the same manner as in Section 5.3.2.

Chapter 6

We consider causal inference on ATE with missing and misclassified outcome variable. We analytically investigate the impact of ignoring outcome missingness and/or outcome misclassification on IPW estimation of ATE and establish intrinsic connections between missingness effects and misclassification effects. By using suitable weighting strategies, we develop valid correction methods to eliminate the effects of missingness and misclassification on causal inference. We further propose a doubly robust correction method which yields consistent estimators even when either the treatment model or the outcome model is misspecified. Our simulation studies show that ignoring missingness and misclassification effects can result in severely biased results. The proposed methods present satisfactory finite sample performance in our simulation studies.

There are two types of weighting in our proposed estimators. The first type is based on the propensity score to conduct causal inference, as in the IPW estimator (6.2) with complete and error-free data. The second type corresponds to the missingness probability to account for missingness. Our proposed methods are simple in that we develop the correction methods by adding an additional layer of weighting to the IPW estimators without completely breaking the existing structure of the original IPW estimators.

The development is directed to estimation of ATE. It can be readily modified to handle estimation of other causal effect measures such as the causal odds ratio and the causal risk ratio.

Chapter 7

We consider multiple robust estimation of causal treatment effects with missing and/or misclassified outcome variable. We develop three estimation methods applicable to situations where only misclassification occurs, only missingness occurs, or both misclassification and missingness occur. Our proposed estimators are consistent when either the set

of multiple postulated treatment models or the set of multiple postulated outcome models contains a correctly specified model. They provide even more protection against model misspecification than the doubly robust estimation methods.

Our simulation studies not only reveal that ignoring missingness and/or misclassification effects can lead to severely biased results, but also show satisfactory finite sample performance of the proposed methods under various settings.

The current development assumes both the misclassification mechanism (7.10) and the missingness mechanism (7.14) hold. However, in many applications, the underlying mechanisms for missingness and misclassification could be much more complex. It will be interesting to consider more complex mechanisms for missingness and misclassification.

Chapter 8

Misclassified outcome data arise frequently in practice and present a challenge in conducting causal inference. Discussion on addressing this issue is rather limited in the literature. Chapter 4 develops the IPW estimation methods for ATE with mismeasured outcome effects incorporated. To expedite the application of these correction methods, we develop an R package **ipwErrorY**. For practical settings where the treatment model and the outcome model are specified as logistic regression models, we implement the correction methods developed in Chapter 4 for settings with known misclassification probabilities, validation data, or replicates of the outcome data as well as the doubly robust method with known misclassification probabilities. Our package offers a useful and convenient tool for data analysts to perform valid inference about ATE when the binary outcome variable is subject to misclassification.

The source code for the developed R package **ipwErrorY** is available for download at `https://CRAN.R-project.org/package=ipwErrorY`.

In summary, causal inference with measurement error is a new and rapidly growing research area. While important contributions have been made in the last decade, many exciting research problems remain unexplored. The problems we have studied in this thesis are of practical importance as well as theoretical values.

References

Austin, P. C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine*, 26:3078–3094.

Babanezhad, M., Vansteelandt, S., and Goetghebeur, E. (2010). Comparison of causal effect estimators under exposure misclassification. *Journal of Statistical Planning and Inference*, 140:1306–1319.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33:2297–2340.

Baiocchi, M., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105:1285–1296.

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

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association*, 45:164–180.

Blackwell, M., Honaker, J., and King, G. (2017). A unified approach to measurement error and missing data: overview and applications. *Sociological Methods & Research*, 46:303–341.

Blakely, T., McKenzie, S., and Carter, K. (2013). Misclassification of the mediator matters when estimating indirect effects. *Journal of Epidemiology and Community Health*, 67:1–9.

Braun, D., Gorfine, M., Parmigiani, G., Arvold, N. D., Dominici, F., and Zigler, C. (2017). Propensity scores with misclassified treatment assignment: a likelihood-based adjustment. *Biostatistics*, 18:695–710.

Braun, D., Zigler, C., Dominici, F., and Gorfine, M. (2016). Using validation data to adjust the inverse probability weighting estimator for misclassified treatment. *Harvard University Biostatistics Working Paper Series. Working Paper 201*, pages 1–19.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. Boca Raton: Chapman & Hall/CRC.

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*, 96:723–734.

Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the simex estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, 91:242–250.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, 2nd Edition*. Boca Raton: Chapman & Hall/CRC.

Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T., and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika*, 71:19–25.

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

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

Cole, S. R., Chu, H., and Greenland, S. (2006). Multiple-imputation for measurement-error correction. *International Journal of Epidemiology*, 35:1074–1081.

Cole, S. R. and Frangakis, C. E. (2009). The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20:3–5.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89:1314–1328.

Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Federation Proceedings*, 21:59–61.

Daniel, R. M., Cousens, S. N., De Stavola, B. L., Kenward, M. G., and Sterne, J. A. C. (2013). Methods for dealing with time-dependent confounding. *Statistics in Medicine*, 32:1584–1618.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Devanarayan, V. and Stefanski, L. A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics & Probability Letters*, 59:219–225.

Edwards, J. K., Cole, S. R., and Westreich, D. (2015a). All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. *International Journal of Epidemiology*, 44:1452–1459.

Edwards, J. K., Cole, S. R., Westreich, D., Crane, H., Eron, J. J., Mathews, W. C., Moore, R., Boswell, S. L., Lesko, C. R., and Mugavero, M. J. (2015b). Multiple imputation to account for measurement error in marginal structural models. *Epidemiology*, 26:645–652.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, volume 38. Philadelphia: SIAM.

Fan, J., Imai, K., Liu, H., Ning, Y., and Yang, X. (2016). Improving covariate balancing propensity score: a doubly robust and efficient approach. Technical report, Princeton University.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58:21–29.

Freedman, L. S., Fainberg, V., Kipnis, V., Midthune, D., and Carroll, R. J. (2004). A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics*, 60:172–181.

Freedman, L. S., Midthune, D., Carroll, R. J., and Kipnis, V. (2008). A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine*, 27:5195–5216.

Freedman, L. S., Midthune, D., Dodd, K. W., Carroll, R. J., and Kipnis, V. (2015). A statistical model for measurement error that incorporates variation over time in the target measure, with application to nutritional epidemiology. *Statistics in Medicine*, 34:3590–3605.

Fuller, W. A. (1987). *Measurement Error Models*. New York: John Wiley & Sons.

Gravel, C. A. and Platt, R. W. (2018). Weighted estimation for confounded binary outcomes subject to misclassification. *Statistics in Medicine*, 37:425–436.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10:37–48.

Gustafson, P. (2003). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Boca Raton: Chapman & Hall/CRC.

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

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

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

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

Hasselman, B. (2016). *nleqslv: Solve Systems of Nonlinear Equations*. R package version 3.0.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65:261–294.

Hernán, M. A., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, 11:561–570.

Hernán, M. A. and Cole, S. R. (2009). Invited commentary: causal diagrams and measurement bias. *American Journal of Epidemiology*, 170:959–962.

Hernán, M. A. and Robins, J. M. (2016). *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation*. New York: Springer.

Hill, A. B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58:295–300.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. *In: Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, Gelman, A., Meng X.-L. (eds.) Hobokon: John Wiley & Sons*, pages 73–84.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15:199–236.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960.

Hong, H., Rudolph, K. E., and Stuart, E. A. (2017). Bayesian approach for addressing differential covariate measurement error in propensity score methods. *Psychometrika*, 82:1078–1096.

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

Huang, Y. and Wang, C. Y. (2001). Consistent functional methods for logistic regression with errors in covariates. *Journal of the American Statistical Association*, 96:1469–1482.

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

Imai, K. and Yamamoto, T. (2010). Causal inference with differential measurement error: nonparametric identification and sensitivity analysis. *American Journal of Political Science*, 54:543–560.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87:706–710.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press.

Kahan, B. C., Jairath, V., Doré, C. J., and Morris, T. P. (2014). The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*, 15:139.

Koehler, E., Brown, E., and Haneuse, S. J.-P. (2009). On the assessment of monte carlo error in simulation-based statistical analyses. *The American Statistician*, 63:155–162.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101:423–437.

Kyle, R. P., Moodie, E. E. M., Klein, M. B., and Abrahamowicz, M. (2016). Correcting for measurement error in time-varying covariates in marginal structural models. *American Journal of Epidemiology*, 184:249–258.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29:337–346.

Lee, S. M., Landry, J., Jones, P. M., Buhrmann, O., and Morley-Forster, P. (2013). The effectiveness of a perioperative smoking cessation program: a randomized clinical trial. *Anesthesia & Analgesia*, 117:605–613.

Lenis, D., Ebnesajjad, C. F., and Stuart, E. A. (2016). A doubly robust estimator for the average treatment effect in the context of a mean-reverting measurement error. *Biostatistics*, 18:325–337.

Lewbel, A. (2007). Estimation of average treatment effects with misclassification. *Econometrica*, 75:537–551.

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

Lockwood, J. R. and McCaffrey, D. F. (2016). Matching and weighting with functions of error-prone covariates for causal inference. *Journal of the American Statistical Association*, 111:1831–1839.

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*, 23:2937–2960.

Magder, L. S. and Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, 146:195–203.

McCaffrey, D. F., Lockwood, J. R., and Setodji, C. M. (2013). Inverse probability weighting with error-prone covariates. *Biometrika*, 100:671–680.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9:403–425.

Miglioretti, D. L. and Heagerty, P. J. (2004). Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics*, 5:381–398.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, 86:843–855.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *In: Handbook of Econometrics, Engle, R., McFadden, D. (eds). Amsterdam: Elsevier*, 4:2111–2245.

Neyman, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: essai des principes. English translation and edition by Dabrowska, D. M. and Speed, T. P. (1990). *Statistical Science*, 5:465–472.

Ning, Y., Yi, G. Y., and Reid, N. (2018). A class of weighted estimating equations for semiparametric transformation models with missing covariates. *Scandinavian Journal of Statistics*, 45:87–109.

Ogburn, E. L. and VanderWeele, T. J. (2012). Analytic results on the bias due to non-differential misclassification of a binary mediator. *American Journal of Epidemiology*, 176:555–561.

Owen, A. B. (2001). *Empirical Likelihood*. New York: Chapman & Hall/CRC.

Pearl, J. (2009a). *Causality: Models, Reasoning and Inference, 2nd Edition*. Cambridge: Cambridge University Press.

Pearl, J. (2009b). On measurement bias in causal inference. *Technical Report R-357, Department of Computer Science, University of California, Los Angeles.*

Polis, C. B., Westreich, D., Balkus, J. E., and Heffron, R. (2013). Assessing the effect of hormonal contraception on hiv acquisition in observational data: challenges and recommended analytic approaches. *AIDS*, 27(Suppl 1):S35–43.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69:331–342.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8:431–440.

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

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

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*, 69:101–122.

Raykov, T. (2012). Propensity score analysis with fallible covariates: a note on a latent variable modeling approach. *Educational and Psychological Measurement*, 72:715–733.

Regier, M. D., Moodie, E. E. M., and Platt, R. W. (2014). The effect of error-in-confounders on the estimation of the causal parameter when using marginal structural models and inverse probability-of-treatment weights: a simulation study. *The International Journal of Biostatistics*, 10:1–15.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *In: Health Service Research Methodology: A Focus on AIDS, Sechrest L., Freeman H., Mulley A. (eds). Washington, D.C.: U.S. Public Health Service*, pages 113–159.

Robins, J. M. (1997). Structural nested failure time models. *In: The Encyclopedia of Biostatistics, Armitage, P., Colton, T. (eds). Chichester: John Wiley & Sons*, pages 4372–4389.

Robins, J. M. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine*, 17:269–302.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. *In: Statistical Models in Epidemiology: The Environment and Clinical Trials, Hallora, M. E., Berry, D. A. (eds). New York: Springer*, pages 95–134.

Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of aids patients. *Epidemiology*, 3:319–336.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560.

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*, 89:846–866.

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*, 90:106–121.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82:387–394.

Rosenbaum, P. R. (1998). Propensity score. *In: Encyclopedia of Biostatistics, Armitage, P., Colton, T. (eds). New York: John Wiley & Sons*, 5:3551–3555.

Rosenbaum, P. R. (2017). *Observation and Experiment: An Introduction to Causal Inference*. Cambridge: Harvard University Press.

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

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*, 79:516–524.

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*, 39:33–38.

Rosner, B., Spiegelman, D., and Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology*, 132:734–745.

Rosner, B., Willett, W. C., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8:1051–1069.

Rothman, K. J., Greenland, S., and Lash, T. L. (2008). *Modern Epidemiology, 3rd Edition*. Philadelphia: Lippincott Williams & Wilkins.

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

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489.

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

Shu, D. and Yi, G. Y. (2016). Inverse-probability-of-treatment weighted estimation of causal parameters in the presence of error-contaminated and time-dependent confounders. *Biometrical Journal. Under Revision*.

Shu, D. and Yi, G. Y. (2017). Causal inference with measurement error in outcomes: bias analysis and estimation methods. *Statistical Methods in Medical Research, doi: 10.1177/0962280217743777*.

Shu, D. and Yi, G. Y. (2018a). Estimation of causal effect measures in the presence of measurement error in confounders. *Statistics in Biosciences, doi: 10.1007/s12561-018-9213-8*.

Shu, D. and Yi, G. Y. (2018b). ipwErrorY: an R package for estimating average treatment effects with outcome misclassification. *Submitted*.

Shu, D. and Yi, G. Y. (2018c). *ipwErrorY: Inverse Probability Weighting Estimation of Average Treatment Effect with Outcome Misclassification*. R package version 1.0.

Shu, D. and Yi, G. Y. (2018d). Weighted causal inference methods with mismeasured covariates and misclassified outcomes. *Statistics in Medicine. Under Revision*.

Small, D. S. and Rosenbaum, P. R. (2008). War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association*, 103:924–933.

Spiegelman, D., Rosner, B., and Logan, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association*, 95:51–61.

Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement-error models. *Communications in Statistics-Theory and Methods*, 18:4335–4358.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, 74:703–716.

Stürmer, T., Schneeweiss, S., Avorn, J., and Glynn, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology*, 162:279–289.

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

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

SRNT Subcommittee on Biochemical Verification (2002). Biochemical verification of tobacco use and cessation. *Nicotine & Tobacco Research*, 4:149–159.

Tsiatis, A. A. (2007). *Semiparametric Theory and Missing Data*. New York: Springer.

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer.

Wagenknecht, L. E., Burke, G. L., Perkins, L. L., Haley, N. J., and Friedman, G. D. (1992). Misclassification of smoking status in the cardia study: a comparison of self-report with serum cotinine levels. *American Journal of Public Health*, 82:33–36.

Webb-Vargas, Y., Rudolph, K. E., Lenis, D., Murakami, P., and Stuart, E. A. (2017). An imputation-based solution to using mismeasured covariates in propensity score analysis. *Statistical Methods in Medical Research*, 26:1824–1837.

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63:826–833.

White, I., Frost, C., and Tokunaga, S. (2001). Correcting for measurement error in binary and continuous variables using replicates. *Statistics in Medicine*, 20:3441–3457.

Williamson, E. J., Forbes, A., and White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine*, 33:721–737.

Yi, G. Y. (2008). A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates. *Biostatistics*, 9:501–512.

Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. New York: Springer.

Yi, G. Y. and He, W. (2006). Methods for bivariate survival data with mismeasured covariates under an accelerated failure time model. *Communications in Statistics-Theory and Methods*, 35:1539–1554.

Yi, G. Y. and He, W. (2012). Bias analysis and the simulation-extrapolation method for survival data with covariate measurement error under parametric proportional odds models. *Biometrical Journal*, 54:343–360.

Yi, G. Y., Liu, W., and Wu, L. (2011). Simultaneous inference and bias analysis for longitudinal data with covariate measurement error and missing responses. *Biometrics*, 67:67–75.

Yi, G. Y., Ma, Y., and Carroll, R. J. (2012). A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*, 99:151–165.

Yi, G. Y., Ma, Y., Spiegelman, D., and Carroll, R. J. (2015a). Functional and structural methods with mixed measurement error and misclassification in covariates. *Journal of the American Statistical Association*, 110:681–696.

Yi, G. Y. and Reid, N. (2010). A note on mis-specified estimating functions. *Statistica Sinica*, 20:1749–1769.

Yi, G. Y., Tan, X., and Li, R. (2015b). Variable selection and inference procedures for marginal analysis of longitudinal data with missing observations and covariate measurement error. *Canadian Journal of Statistics*, 43:498–518.

APPENDICES

Appendix A

Proofs for the Results in Chapter 2

Let $\Delta(k) = X_k^* + \{A(k) - 1/2\}\boldsymbol{\Sigma}_{\epsilon k}\boldsymbol{\gamma}_{\mathrm{X}k}$, $G_k = 1 + \exp[\{-\gamma_{0k} - \boldsymbol{\gamma}_{\mathrm{A}k}^{\mathrm{T}}\bar{A}(k-1) - \boldsymbol{\gamma}_{\mathrm{Z}k}^{\mathrm{T}}Z(k) - \boldsymbol{\gamma}_{\mathrm{X}k}^{\mathrm{T}}\Delta(k)\}\{2A(k)-1\}]$ and $\hat{G}_i(k) = 1+\exp[\{-\hat{\gamma}_{0k}-\hat{\boldsymbol{\gamma}}_{\mathrm{A}k}^{\mathrm{T}}\bar{A}_i(k-1)-\hat{\boldsymbol{\gamma}}_{\mathrm{Z}k}^{\mathrm{T}}Z_i(k)-\hat{\boldsymbol{\gamma}}_{\mathrm{X}k}^{\mathrm{T}}\hat{\Delta}_i(k)\}\{2A_i(k)-1\}]$. Then $\hat{w}_i = \prod_{k=0}^{K}\hat{G}_i(k)$. We first show $E\left\{YI(\bar{A}=\bar{a})\prod_{k=0}^{K}G_k\right\} = E(Y_{\bar{a}})$.

$$
\begin{aligned}
& E\left\{YI(\bar{A}=\bar{a})\prod_{k=0}^{K}G_k\right\} \\
= \;& P(\bar{A}=\bar{a})\iiiint y_{\bar{a}}\prod_{k=0}^{K}g_k f(y_{\bar{a}},\bar{z},\bar{x},\bar{x}^*|\bar{A}=\bar{a})d\bar{x}^*d\bar{z}d\bar{x}dy_{\bar{a}} \\
= \;& P(\bar{A}=\bar{a})\iiint y_{\bar{a}}\left\{\int\prod_{k=0}^{K}g_k f(\bar{x}^*|\bar{z},\bar{x},y_{\bar{a}},\bar{A}=\bar{a})d\bar{x}^*\right\}f(\bar{z},\bar{x},y_{\bar{a}}|\bar{A}=\bar{a})d\bar{z}d\bar{x}dy_{\bar{a}} \\
= \;& P(\bar{A}=\bar{a})\iiint y_{\bar{a}}\left\{\prod_{k=0}^{K}\int g_k f\{x_k^*|x(k)\}dx_k^*\right\}f(\bar{z},\bar{x},y_{\bar{a}}|\bar{A}=\bar{a})d\bar{z}d\bar{x}dy_{\bar{a}} \\
= \;& P(\bar{A}=\bar{a})\iiint y_{\bar{a}}\prod_{k=0}^{K}\frac{1}{P\{a(k)|,\bar{a}(k-1),z(k),x(k)\}}f(\bar{z},\bar{x},y_{\bar{a}}|\bar{A}=\bar{a})d\bar{z}d\bar{x}dy_{\bar{a}} \\
= \;& P(\bar{A}=\bar{a})\iiint y_{\bar{a}}\prod_{k=0}^{K}\frac{1}{P\{a(k)|,\bar{a}(k-1),z(k),x(k)\}}\frac{P(\bar{A}=\bar{a}|\bar{z},\bar{x},y_{\bar{a}})f(\bar{z},\bar{x},y_{\bar{a}})}{P(\bar{A}=\bar{a})}d\bar{z}d\bar{x}dy_{\bar{a}} \\
= \;& \iiint y_{\bar{a}}\frac{1}{P(\bar{A}=\bar{a}|\bar{z},\bar{x})}P(\bar{A}=\bar{a}|\bar{z},\bar{x})f(\bar{z},\bar{x},y_{\bar{a}})d\bar{z}d\bar{x}dy_{\bar{a}} \\
= \;& E(Y_{\bar{a}}).
\end{aligned}
$$

Using similar arguments, it can then be shown that

$$
\begin{aligned}
& E\left\{I(\bar{A}=\bar{a})\prod_{k=0}^{K}G_k\right\} \\
= \;& P(\bar{A}=\bar{a})\iiiint\prod_{k=0}^{K}g_k f(y_{\bar{a}},\bar{z},\bar{x},\bar{x}^*|\bar{A}=\bar{a})d\bar{x}^*d\bar{z}d\bar{x}dy_{\bar{a}} \\
= \;& P(\bar{A}=\bar{a})\iiint\left\{\int\prod_{k=0}^{K}g_k f(\bar{x}^*|\bar{z},\bar{x},y_{\bar{a}},\bar{A}=\bar{a})d\bar{x}^*\right\}f(\bar{z},\bar{x},y_{\bar{a}}|\bar{A}=\bar{a})d\bar{z}d\bar{x}dy_{\bar{a}} \\
= \;& P(\bar{A}=\bar{a})\iiint\left\{\prod_{k=0}^{K}\int g_k f\{x_k^*|x(k)\}dx_k^*\right\}f(\bar{z},\bar{x},y_{\bar{a}}|\bar{A}=\bar{a})d\bar{z}d\bar{x}dy_{\bar{a}} \\
= \;& P(\bar{A}=\bar{a})\iiint\prod_{k=0}^{K}\frac{1}{P\{a(k)|,\bar{a}(k-1),z(k),x(k)\}}f(\bar{z},\bar{x},y_{\bar{a}}|\bar{A}=\bar{a})d\bar{z}d\bar{x}dy_{\bar{a}} \\
= \;& P(\bar{A}=\bar{a})\iiint\prod_{k=0}^{K}\frac{1}{P\{a(k)|,\bar{a}(k-1),z(k),x(k)\}}\frac{P(\bar{A}=\bar{a}|\bar{z},\bar{x},y_{\bar{a}})f(\bar{z},\bar{x},y_{\bar{a}})}{P(\bar{A}=\bar{a})}d\bar{z}d\bar{x}dy_{\bar{a}} \\
= \;& \iiint\frac{1}{P(\bar{A}=\bar{a}|\bar{z},\bar{x})}P(\bar{A}=\bar{a}|\bar{z},\bar{x})f(\bar{z},\bar{x},y_{\bar{a}})d\bar{z}d\bar{x}dy_{\bar{a}} \\
= \;& 1.
\end{aligned}
$$

Therefore, the causal mean $E(Y_{\bar{a}})$ can be consistently estimated by

$$\frac{\sum_{i=1}^{n}\hat{w}_i Y_i I(\bar{A}_i=\bar{a})}{n}$$

or

$$\frac{\sum_{i=1}^{n}\hat{w}_i Y_i I(\bar{A}_i=\bar{a})}{n}\Big/\frac{\sum_{i=1}^{n}\hat{w}_i I(\bar{A}_i=\bar{a})}{n}=\frac{\sum_{i=1}^{n}\hat{w}_i Y_i I(\bar{A}_i=\bar{a})}{\sum_{i=1}^{n}\hat{w}_i I(\bar{A}_i=\bar{a})}.$$

Appendix B

Proofs for the Results in Chapter 3

Let $B_i(k) = -\gamma_{0k} - \boldsymbol{\gamma}_{Ak}^{\mathrm{T}} \bar{A}_i(k-1) - \boldsymbol{\gamma}_{Zk}^{\mathrm{T}} \bar{Z}_i(k) - \boldsymbol{\gamma}_{Xk}^{\mathrm{T}} \Delta_i(k)$, and $G_i(k) = 1 + \exp[B_i(k)\{2A_i(k) - 1\}]$. Let $\hat{B}_i(k)$ and $\hat{G}_i(k)$ be $B_i(k)$ and $G_i(k)$ with $(\gamma_{0k}, \boldsymbol{\gamma}_{Ak}^{\mathrm{T}}, \boldsymbol{\gamma}_{Zk}^{\mathrm{T}}, \boldsymbol{\gamma}_{Xk}^{\mathrm{T}})^{\mathrm{T}}$ replaced by the consistent estimator $(\hat{\gamma}_{0k}, \hat{\boldsymbol{\gamma}}_{Ak}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}_{Zk}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}_{Xk}^{\mathrm{T}})^{\mathrm{T}}$ which is obtained by solving (3.12). Weight $\hat{w}_i$ in Theorem 3.1 can be expressed by $\prod_{k=0}^{K} \hat{G}_i(k)$.

We first show that

$E\left\{Y I(\bar{A} = \bar{a}) \prod_{k=0}^{K} G_k\right\} = E(Y_{\bar{a}})$.

By Assumption 2, we obtain

$$E\left\{Y I(\bar{A} = \bar{a}) \prod_{k=0}^{K} G_k\right\} = E\left\{Y_{\bar{A}} I(\bar{A} = \bar{a}) \prod_{k=0}^{K} G_k\right\},$$

which equals

$$\begin{aligned} & P(\bar{A} = \bar{a}) E\left\{Y_{\bar{A}} I(\bar{A} = \bar{a}) \prod_{k=0}^{K} G_k | \bar{A} = \bar{a}\right\} + P(\bar{A} \neq \bar{a}) E\left\{Y_{\bar{A}} I(\bar{A} = \bar{a}) \prod_{k=0}^{K} G_k | \bar{A} \neq \bar{a}\right\} \\ = \; & P(\bar{A} = \bar{a}) E\left(Y_{\bar{a}} \prod_{k=0}^{K} G_k | \bar{A} = \bar{a}\right). \end{aligned}$$

Let g_k denote a realization of G_k, and $y_{\bar{a}}$ denote a realization of $Y_{\bar{a}}$, we write

$$
\begin{aligned}
& P(\bar{A}=\bar{a})E\left(Y_{\bar{a}}\prod_{k=0}^{K}G_k|\bar{A}=\bar{a}\right) \\
&= P(\bar{A}=\bar{a})\iiiint y_{\bar{a}}\prod_{k=0}^{K}g_k f(y_{\bar{a}},\bar{z},\bar{x},\bar{x}^*|\bar{A}=\bar{a})d\bar{x}^*d\bar{z}d\bar{x}dy_{\bar{a}} \\
&= P(\bar{A}=\bar{a})\iiiint y_{\bar{a}}\prod_{k=0}^{K}g_k f(\bar{x}^*|\bar{z},\bar{x},y_{\bar{a}},\bar{A}=\bar{a})f(\bar{z},\bar{x},y_{\bar{a}}|\bar{A}=\bar{a})d\bar{x}^*d\bar{z}d\bar{x}dy_{\bar{a}} \\
&= P(\bar{A}=\bar{a})\iiint y_{\bar{a}}\left\{\int\prod_{k=0}^{K}g_k f(\bar{x}^*|\bar{z},\bar{x},y_{\bar{a}},\bar{A}=\bar{a})d\bar{x}^*\right\}f(\bar{z},\bar{x},y_{\bar{a}}|\bar{A}=\bar{a})d\bar{z}d\bar{x}dy_{\bar{a}}.
\end{aligned}
$$

By (3.8), $\int\prod_{k=0}^{K}g_k f(\bar{x}^*|\bar{z},\bar{x},y_{\bar{a}},\bar{A}=\bar{a})d\bar{x}^*=\int\prod_{k=0}^{K}g_k f(\bar{x}^*|\bar{x})d\bar{x}^*$.

Since $X^*_{k_1}\perp\!\!\!\perp X^*_{k_2}|\bar{X}$ for $k_1\neq k_2$, above becomes

$$\prod_{k=0}^{K}\int g_k f(x^*_k|\bar{x})dx^*_k=\prod_{k=0}^{K}\int g_k f\{x^*_k|x(k)\}dx^*_k.$$

By (3.8), $X^*_k\sim N(X(k),\boldsymbol{\Sigma}_{\epsilon k})$ and using the moment generating function of normal distributions, we obtain that above equals

$$\prod_{k=0}^{K}\left(1+\exp[h_k\{2a(k)-1\}-\{2a(k)-1\}\boldsymbol{\gamma}^{\mathrm{T}}_{Xk}x(k)+\{2a(k)-1\}^2\boldsymbol{\gamma}^{\mathrm{T}}_{Xk}\boldsymbol{\Sigma}_{\epsilon k}\boldsymbol{\gamma}_{Xk}/2]\right),$$

where $h_k=-\gamma_{0k}-\boldsymbol{\gamma}^{\mathrm{T}}_{Ak}\bar{a}(k-1)-\boldsymbol{\gamma}^{\mathrm{T}}_{Zk}\bar{z}(k)-\{a(k)-1/2\}\boldsymbol{\gamma}^{\mathrm{T}}_{Xk}\boldsymbol{\Sigma}_{\epsilon k}\boldsymbol{\gamma}_{Xk}$. Since $\{2a(k)-1\}^2$=1 for both $a(k)=0$ and 1, above equals

$$
\begin{aligned}
&\prod_{k=0}^{K}\left(1+\exp[\{-\gamma_{0k}-\boldsymbol{\gamma}^{\mathrm{T}}_{Ak}\bar{a}(k-1)-\boldsymbol{\gamma}^{\mathrm{T}}_{Zk}\bar{z}(k)-\boldsymbol{\gamma}^{\mathrm{T}}_{Xk}x(k)\}\{2a(k)-1\}]\right) \\
&=\prod_{k=0}^{K}\frac{1}{P\{a(k)|,\bar{a}(k-1),\bar{z}(k),x(k)\}},
\end{aligned}
$$

where the last identity comes from (3.4) and Assumption 5. Therefore,

$$
\begin{aligned}
& E\left\{ Y I(\bar{A}=\bar{a}) \prod_{k=0}^{K} G_k \right\} \\
= \;& P(\bar{A}=\bar{a}) \iiint y_{\bar{a}} \prod_{k=0}^{K} \frac{1}{P\{a(k)|, \bar{a}(k-1), \bar{z}(k), x(k)\}} f(\bar{z}, \bar{x}, y_{\bar{a}}|\bar{A}=\bar{a}) d\bar{z} d\bar{x} dy_{\bar{a}} \\
= \;& P(\bar{A}=\bar{a}) \iiint y_{\bar{a}} \prod_{k=0}^{K} \frac{1}{P\{a(k)|, \bar{a}(k-1), \bar{z}(k), x(k)\}} \frac{P(\bar{A}=\bar{a}|\bar{z}, \bar{x}, y_{\bar{a}}) f(\bar{z}, \bar{x}, y_{\bar{a}})}{P(\bar{A}=\bar{a})} d\bar{z} d\bar{x} dy_{\bar{a}} \\
= \;& \iiint y_{\bar{a}} \prod_{k=0}^{K} \frac{1}{P\{a(k)|, \bar{a}(k-1), \bar{z}(k), x(k)\}} P(\bar{A}=\bar{a}|\bar{z}, \bar{x}, y_{\bar{a}}) f(\bar{z}, \bar{x}, y_{\bar{a}}) d\bar{z} d\bar{x} dy_{\bar{a}} \\
= \;& \iiint y_{\bar{a}} \prod_{k=0}^{K} \frac{1}{P\{a(k)|, \bar{a}(k-1), \bar{z}(k), x(k)\}} P(\bar{A}=\bar{a}|\bar{z}, \bar{x}) f(\bar{z}, \bar{x}, y_{\bar{a}}) d\bar{z} d\bar{x} dy_{\bar{a}} \\
= \;& \iiint y_{\bar{a}} \frac{1}{P(\bar{A}=\bar{a}|\bar{z}, \bar{x})} P(\bar{A}=\bar{a}|\bar{z}, \bar{x}) f(\bar{z}, \bar{x}, y_{\bar{a}}) d\bar{z} d\bar{x} dy_{\bar{a}} \\
= \;& \iiint y_{\bar{a}} f(\bar{z}, \bar{x}, y_{\bar{a}}) d\bar{z} d\bar{x} dy_{\bar{a}} \\
= \;& E(Y_{\bar{a}}),
\end{aligned}
$$

where the fourth last identity is due to the Assumption 3 and the third last identity is due to Assumption 5.

Let $w = \prod_{k=0}^{K} G_k$. Similarly, it can be shown that

$$
E\left\{ I(\bar{A}=\bar{a}) w \right\} = E\left\{ I(\bar{A}=\bar{a}) \prod_{k=0}^{K} G_k \right\} = \iiint f(\bar{z}, \bar{x}, y_{\bar{a}}) d\bar{z} d\bar{x} dy_{\bar{a}} = 1.
$$

Now we consider the pseudo-population as a result of assigning weights w and let Y^p denote the observations of the pseudo-population. Fitting model (3.2) to the pseudo-population implies

$$
E(Y^p|\bar{A}=\bar{a}) = \frac{E(Yw|\bar{A}=\bar{a})}{E(w|\bar{A}=\bar{a})} = \frac{E(Yw|\bar{A}=\bar{a})}{E\{wI(\bar{A}=\bar{a})\}/P(\bar{A}=\bar{a})} = P(\bar{A}=\bar{a}) E(Yw|\bar{A}=\bar{a}),
$$

where the first identity can be understood by the fact that $E(Y^p|\bar{A}=\bar{a})$ is consistently

estimated by $\frac{\sum_{i=1}^n y_i w_i | \bar{A}_i = \bar{a}}{\sum_{i=1}^n w_i | \bar{A}_i = \bar{a}}$, which converges in probability to $\frac{E(Yw|\bar{A}=\bar{a})}{E(w|\bar{A}=\bar{a})}$. Note that

$$P(\bar{A}=\bar{a})E(Yw|\bar{A}=\bar{a}) = E\left\{YI(\bar{A}=\bar{a})w\right\} = E(Y_{\bar{a}}) = h(\bar{a};\boldsymbol{\beta}).$$

So $E(Y^p|\bar{A}=\bar{a}) = h(\bar{a};\boldsymbol{\beta})$. It follows that fitting model (3.2) with weight $\hat{w}_i$ which is w_i with $(\gamma_{0k}, \boldsymbol{\gamma}_{Ak}^{\mathrm{T}}, \boldsymbol{\gamma}_{Zk}^{\mathrm{T}}, \boldsymbol{\gamma}_{Xk}^{\mathrm{T}})^{\mathrm{T}}$ replaced by a consistent estimator $(\hat{\gamma}_{0k}, \hat{\boldsymbol{\gamma}}_{Ak}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}_{Zk}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}_{Xk}^{\mathrm{T}})^{\mathrm{T}}$ yields a consistent estimator for the causal parameter $\boldsymbol{\beta}$ in model (3.1).

Weight sw can be expressed as $w \cdot l(\bar{A})$ no matter the model for $A(k)$ given $\bar{A}(k-1)$ is misspecified or not, where $l(\cdot)$ is a function of $\bar{A}$. Similar to the above development, we can show $E\left\{YI(\bar{A}=\bar{a})sw\right\} = l(\bar{a})E(Y_{\bar{a}})$ and $E\left\{I(\bar{A}=\bar{a})sw\right\} = l(\bar{a})$.

Now we consider the pseudo-population as a result of assigning weights sw and let Y^p denote the observations of the pseudo-population. Fitting model (3.2) to the pseudo-population implies

$$E(Y^p|\bar{A}=\bar{a}) = \frac{E(Ysw|\bar{A}=\bar{a})}{E(sw|\bar{A}=\bar{a})} = \frac{E(Ysw|\bar{A}=\bar{a})}{E\{swI(\bar{A}=\bar{a})\}/P(\bar{A}=\bar{a})} = P(\bar{A}=\bar{a})E(Ysw|\bar{A}=\bar{a})/l(\bar{a}).$$

Note that

$$P(\bar{A}=\bar{a})E(Ysw|\bar{A}=\bar{a})/l(\bar{a}) = E\left\{YI(\bar{A}=\bar{a})sw\right\}/l(\bar{a}) = E(Y_{\bar{a}}) = h(\bar{a};\boldsymbol{\beta}).$$

So $E(Y^p|\bar{A}=\bar{a}) = h(\bar{a};\boldsymbol{\beta})$. It follows that fitting model (3.2) with weight $\widehat{sw}_i$ which is sw_i with $(\gamma_{0k}, \boldsymbol{\gamma}_{Ak}^{\mathrm{T}}, \boldsymbol{\gamma}_{Zk}^{\mathrm{T}}, \boldsymbol{\gamma}_{Xk}^{\mathrm{T}})^{\mathrm{T}}$ replaced by a consistent estimator $(\hat{\gamma}_{0k}, \hat{\boldsymbol{\gamma}}_{Ak}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}_{Zk}^{\mathrm{T}}, \hat{\boldsymbol{\gamma}}_{Xk}^{\mathrm{T}})^{\mathrm{T}}$ yields a consistent estimator for the causal parameter $\boldsymbol{\beta}$ in model (3.1).

Appendix C

Proofs for the Results in Chapter 4

C.1 Proofs of Theorems 4.1 and 4.2

Let $\boldsymbol{\gamma}$ be the parameters for the treatment model, and $\phi(\cdot)$ be an unbiased estimating function of $\boldsymbol{\gamma}$ which is determined by the treatment model. Let $\boldsymbol{\theta} = (\boldsymbol{\gamma}^{\mathrm{T}}, \tau)^{\mathrm{T}}$, then the estimating function for the naive analysis is

$$\Psi^*(Y_i^*, T_i, X_i; \boldsymbol{\theta}) = \left\{ \begin{array}{c} \phi(X_i, T_i, \boldsymbol{\gamma}) \\ \dfrac{T_i Y_i^*}{e_i} - \dfrac{(1-T_i)Y_i^*}{1-e_i} - \tau \end{array} \right\} \tag{C.1}$$

and solving

$$\sum_{i=1}^{n} \Psi^*(Y_i^*, T_i, X_i; \boldsymbol{\theta}) = \mathbf{0} \tag{C.2}$$

for $\boldsymbol{\theta}$ yields the naive estimator $\hat{\tau}^* = n^{-1} \sum_{i=1}^{n} \dfrac{T_i Y_i^*}{\hat{e}_i} - n^{-1} \sum_{i=1}^{n} \dfrac{(1-T_i)Y_i^*}{1-\hat{e}_i}$.

Proof of Theorem 4.1:

By Yi and Reid (2010), the solution of (C.2) converges in probability to $\boldsymbol{\theta}_0^*$ which solves $E\{\Psi^*(Y^*, T, X; \boldsymbol{\theta}_0^*)\} = \mathbf{0}$. We now show that $\boldsymbol{\theta}_0^* = (\boldsymbol{\gamma}_0^{\mathrm{T}}, \tau_0)^{\mathrm{T}}$. Since $\phi(\cdot)$ is an unbiased estimating function of $\boldsymbol{\gamma}$, by the form of (C.1), it suffices to show that $E\left(\dfrac{TY^*}{e}\right) -$

$E\left\{\dfrac{(1-T)Y^*}{1-e}\right\} = \tau_0$. Let $\epsilon = \alpha_1 T\epsilon_1 + \alpha_2(1-T)\epsilon_2 + \epsilon_3$, then

$$E\left(\frac{TY^*}{e}\right) = E\left[\frac{T\{Y+\epsilon+g(X)\}}{e}\right] = E\left(\frac{TY}{e}\right) + E\left(\frac{T\epsilon}{e}\right) + E\left\{\frac{Tg(X)}{e}\right\}. \qquad \text{(C.3)}$$

By Lunceford and Davidian (2004), $E\left(\dfrac{TY}{e}\right) = E(Y_1)$ under causal inference assumptions. Then, (C.3) becomes

$$E\left(\frac{TY^*}{e}\right) = E(Y_1) + E\left(\frac{T\epsilon}{e}\right) + E\left\{\frac{Tg(X)}{e}\right\}. \qquad \text{(C.4)}$$

Noticing that $T \cdot T = T$ and $T \cdot (1-T)$=0, we obtain that

$$\begin{aligned} E\left(\frac{T\epsilon}{e}\right) &= E\left[\frac{T\{\alpha_1 T\epsilon_1 + \alpha_2(1-T)\epsilon_2 + \epsilon_3\}}{e}\right] \\ &= E\left\{\frac{T(\alpha_1\epsilon_1 + \epsilon_3)}{e}\right\} \\ &= \alpha_1 E\left(\frac{T\epsilon_1}{e}\right) + E\left(\frac{T\epsilon_3}{e}\right). \end{aligned} \qquad \text{(C.5)}$$

Because ϵ_1 and ϵ_3 are both independent of T given X, and $E(\epsilon_1|X) = E(\epsilon_3|X) = 0$, we obtain that

$$E\left(\frac{T\epsilon_1}{e}\right) = E\left\{E\left(\frac{T\epsilon_1}{e}\middle| X\right)\right\} = E\left\{\frac{1}{e}E(T|X)E(\epsilon_1|X)\right\} = 0,$$

and

$$E\left(\frac{T\epsilon_3}{e}\right) = E\left\{E\left(\frac{T\epsilon_3}{e}\middle| X\right)\right\} = E\left\{\frac{1}{e}E(T|X)E(\epsilon_3|X)\right\} = 0.$$

By (C.5), $E\left\{\dfrac{T\epsilon}{e}\right\} = 0$. Since

$$E\left\{\frac{Tg(X)}{e}\right\} = E\left[E\left\{\frac{Tg(X)}{e}\middle| X\right\}\right] = E\left\{\frac{g(X)}{e}E(T|X)\right\} = E\{g(X)\},$$

(C.4) reduces to

$$E\left(\frac{TY^*}{e}\right) = E(Y_1) + E\{g(X)\}. \tag{C.6}$$

Similarly, by Lunceford and Davidian (2004), $E\left\{\frac{(1-T)Y}{1-e}\right\} = E(Y_0)$ under causal inference assumptions and we obtain

$$E\left\{\frac{(1-T)Y^*}{1-e}\right\} = E(Y_0) + E\{g(X)\}. \tag{C.7}$$

Combining (C.6) and (C.7) yields that

$$E\left(\frac{TY^*}{e}\right) - E\left\{\frac{(1-T)Y^*}{1-e}\right\} = E(Y_1) - E(Y_0),$$

which is the ATE τ_0.

Proof of Theorem 4.2:

By Yi and Reid (2010), the solution of (C.2) converges in probability to $\boldsymbol{\theta}_0^*$ which solves $E\{\Psi^*(Y^*, T, X; \boldsymbol{\theta}_0^*)\} = \mathbf{0}$. We now show that $\boldsymbol{\theta}_0^* = (\boldsymbol{\gamma}_0^{\mathrm{T}}, (p_{11} - p_{10})\tau_0)^{\mathrm{T}}$. Since $\phi(\cdot)$ is an unbiased estimating function of $\boldsymbol{\gamma}$, by the form of (C.1), it suffices to show that $E\left(\frac{TY^*}{e}\right) - E\left\{\frac{(1-T)Y^*}{1-e}\right\} = (p_{11} - p_{10})\tau_0$.

Noting that $E\left(\frac{TY^*}{e}\right) = E\left\{E\left(\frac{TY^*}{e}\middle| X\right)\right\}$, we derive that

$$\begin{aligned}
E\left(\frac{TY^*}{e}\middle| X\right) &= \frac{1}{e}E(TY^*|X) \\
&= \frac{1}{e} \cdot P(T=1, Y^*=1|X) \\
&= \frac{1}{e} \cdot P(T=1|X) \cdot P(Y^*=1|T=1, X) \\
&= P(Y^*=1|T=1, X) \\
&= P(Y=1, Y^*=1|T=1, X) + P(Y=0, Y^*=1|T=1, X) \\
&= P(Y^*=1|Y=1)P(Y=1|T=1, X) + P(Y^*=1|Y=0)P(Y=0|T=1, X)
\end{aligned}$$

$$
\begin{aligned}
&= p_{11}P(Y=1|T=1,X) + p_{10}P(Y=0|T=1,X) \\
&= (p_{11}-p_{10})P(Y=1|T=1,X) + p_{10},
\end{aligned}
$$

where the homogeneous misclassification probabilities assumption is used in the third last step. So

$$
\begin{aligned}
E\left(\frac{TY^*}{e}\right) &= E\left\{E\left(\frac{TY^*}{e}\middle|X\right)\right\} \\
&= (p_{11}-p_{10})\cdot E[\{P(Y=1|T=1,X)\}|X] + p_{10} \\
&= (p_{11}-p_{10})\cdot E[\{P(Y_1=1|T=1,X)\}|X] + p_{10} \\
&= (p_{11}-p_{10})\cdot E[\{P(Y_1=1|X)\}|X] + p_{10} \\
&= (p_{11}-p_{10})E(Y_1) + p_{10}, \qquad \text{(C.8)}
\end{aligned}
$$

where the consistency assumption is used in the third step and the no unmeasured confounding assumption is used in the second last step.

Similarly,

$$
E\left\{\frac{(1-T)Y^*}{1-e}\right\} = (p_{11}-p_{10})E(Y_0) + p_{10}. \qquad \text{(C.9)}
$$

Therefore, by (C.8) and (C.9), we obtain that

$$
\begin{aligned}
E\left(\frac{TY^*}{e}\right) - E\left\{\frac{(1-T)Y^*}{1-e}\right\} &= (p_{11}-p_{10})\cdot\{E(Y_1)-E(Y_0)\} \\
&= (p_{11}-p_{10})\tau_0, \qquad \text{(C.10)}
\end{aligned}
$$

thus the result in (a) follows. Conclusion in (b) follows immediately form the convergence theorem.

C.2 Proof of Efficiency Loss Caused by Misclassification

First, we obtain that

$$A(\tau_0) = E\left\{-\frac{\partial}{\partial\tau}\psi(Y^*, T, X; \tau_0)\right\} = p_{11} - p_{10}.$$

Now we calculate that

$$\begin{aligned}
B(\tau_0) &= E\{\psi^2(Y^*, T, X; \tau_0)\} \\
&= E\left[\left\{\frac{TY^*}{e} - \frac{(1-T)Y^*}{1-e} - (p_{11} - p_{10})\tau_0\right\}^2\right] \\
&= E\left[\frac{TY^*}{e^2} + \frac{(1-T)Y^*}{(1-e)^2} + (p_{11} - p_{10})^2\tau_0^2 - 2(p_{11} - p_{10})\tau_0\left\{\frac{TY^*}{e} - \frac{(1-T)Y^*}{1-e}\right\}\right] \\
&= E\left\{\frac{TY^*}{e^2} + \frac{(1-T)Y^*}{(1-e)^2}\right\} + (p_{11} - p_{10})^2\tau_0^2 - 2(p_{11} - p_{10})\tau_0 E\left\{\frac{TY^*}{e} - \frac{(1-T)Y^*}{1-e}\right\} \\
&= E\left\{\frac{TY^*}{e^2} + \frac{(1-T)Y^*}{(1-e)^2}\right\} - (p_{11} - p_{10})^2\tau_0^2, \qquad \text{(C.11)}
\end{aligned}$$

where we use that $Y^* \cdot Y^* = Y^*$, $T \cdot T = T$ and $(1-T)\cdot(1-T) = 1-T$ in the third step and we use

$$E\left\{\frac{TY^*}{e} - \frac{(1-T)Y^*}{1-e}\right\} = (p_{11} - p_{10})\tau_0$$

in the last step.

Let $q_1 = P(Y = 1|T = 1, X)$ and $q_0 = P(Y = 1|T = 0, X)$, then

$$E\left(\frac{TY^*}{e^2}\right) = E\left\{E\left(\frac{TY^*}{e^2}\right)\middle|X\right\} = E\left\{\frac{1}{e^2}P(T = 1, Y^* = 1|X)\right\} = E\left\{\frac{q_1 p_{11} + (1-q_1)p_{10}}{e}\right\},$$

and similarly,

$$E\left\{\frac{(1-T)Y^*}{(1-e)^2}\right\} = E\left\{\frac{1}{(1-e)^2}P(T = 0, Y^* = 1|X)\right\} = E\left\{\frac{q_0 p_{11} + (1-q_0)p_{10}}{1-e}\right\}.$$

Therefore, (C.11) becomes

$$E\{\psi^2(Y^*,T,X;\tau_0)\} = E\left\{\frac{q_1p_{11}+(1-q_1)p_{10}}{e}\right\}+E\left\{\frac{q_0p_{11}+(1-q_0)p_{10}}{1-e}\right\}-(p_{11}-p_{10})^2\tau_0^2.$$

Consequently, the asymptotic variance of $\sqrt{n}(\tilde{\tau}^*-\tau_0)$ is

$$\begin{aligned} V_{\text{P}} &= A(\tau_0)^{-1}B(\tau_0)\{A(\tau_0)^{-1}\}^{\text{T}} \\ &= \frac{1}{(p_{11}-p_{10})^2}\left[E\left\{\frac{q_1p_{11}+(1-q_1)p_{10}}{e}\right\}+E\left\{\frac{q_0p_{11}+(1-q_0)p_{10}}{1-e}\right\}\right]-\tau_0^2. \end{aligned}$$

When there is no misclassification of Y, the asymptotic variance of $\sqrt{n}(\hat{\tau}-\tau_0)$ can be written as

$$V = E\left(\frac{q_1}{e}\right)+E\left(\frac{q_0}{1-e}\right)-\tau_0^2.$$

Now we compare V_{P} and V, assuming that $0<p_{11}-p_{10}<1$. Noticing that

$$E\left\{\frac{q_1p_{11}+(1-q_1)p_{10}}{e}\right\} = E\left\{\frac{q_1(p_{11}-p_{10})+p_{10}}{e}\right\} \geq E\left\{\frac{q_1(p_{11}-p_{10})}{e}\right\}$$

and

$$E\left\{\frac{q_0p_{11}+(1-q_0)p_{10}}{1-e}\right\} = E\left\{\frac{q_0(p_{11}-p_{10})+p_{10}}{1-e}\right\} \geq E\left\{\frac{q_0(p_{11}-p_{10})}{1-e}\right\},$$

we obtain that

$$\frac{1}{(p_{11}-p_{10})^2}E\left\{\frac{q_1p_{11}+(1-q_1)p_{10}}{e}\right\} \geq \frac{1}{(p_{11}-p_{10})}E\left(\frac{q_1}{e}\right) > E\left(\frac{q_1}{e}\right)$$

and

$$\frac{1}{(p_{11}-p_{10})^2}E\left\{\frac{q_0p_{11}+(1-q_0)p_{10}}{1-e}\right\} \geq \frac{1}{(p_{11}-p_{10})}E\left(\frac{q_0}{1-e}\right) > E\left(\frac{q_0}{1-e}\right).$$

Therefore, $V_{\text{P}} > V$, suggesting that the misclassification reduces the efficiency.

C.3 Estimates of $Var(\hat{\tau}_{\text{V}})$, $Var(\hat{\tau}_{\text{N}})$ and $Cov(\hat{\tau}_{\text{V}}, \hat{\tau}_{\text{N}})$

To work out covariance between $\hat{\tau}_{\text{V}}$ and $\hat{\tau}_{\text{N}}$, we jointly examine the estimation procedures for $\hat{\tau}_{\text{V}}$ and $\hat{\tau}_{\text{N}}$ by combining associated estimating functions. Directly stacking the estimating functions for $\hat{\tau}_{\text{V}}$ and $\hat{\tau}_{\text{N}}$ creates estimating function with a dimension twice of the parameter τ, which generates a scenario with over-constrained estimating functions. To overcome this problem, we artificially enlarge the parameter space by using τ_{V} and τ_{N} to respectively replace τ in the estimating functions of $\hat{\tau}_{\text{V}}$ and $\hat{\tau}_{\text{N}}$, where the true value of τ_{V} and τ_{N} is required to be identical to that of τ (i.e., τ_0). Specifically, in combination with the parameters for the treatment model and the misclassification models, we let $\boldsymbol{\theta} = (\boldsymbol{\gamma}^{\text{T}}, p_{11}, p_{10}, \tau_{\text{V}}, \tau_{\text{N}})^{\text{T}}$, and consider the combined unbiased estimating functions

$$\Psi_c(Y_i^*, T_i, X_i, Y_i; \boldsymbol{\theta}) = \begin{pmatrix} \phi(X_i, T_i; \boldsymbol{\gamma}) \\ g_1(Y_i^*, Y_i; p_{11}) \\ g_2(Y_i^*, Y_i; p_{10}) \\ \left\{ \dfrac{T_i Y_i}{e_i} - \dfrac{(1-T_i)Y_i}{1-e_i} - \tau_{\text{V}} \right\} \cdot I(i \in \mathcal{V}) \cdot \dfrac{n}{n_{\text{V}}} \\ \left\{ \dfrac{T_i Y_i^*}{e_i} - \dfrac{(1-T_i)Y_i^*}{1-e_i} - (p_{11}-p_{10})\tau_{\text{N}} \right\} I(i \notin \mathcal{V}) \cdot \dfrac{n}{n-n_{\text{V}}} \end{pmatrix},$$

where $I(\cdot)$ is the indicator function. Then solving $\sum_{i=1}^n \Psi_c(Y_i^*, T_i, X_i, Y_i; \boldsymbol{\theta}) = \mathbf{0}$ for $\boldsymbol{\theta}$ yields an estimator of θ, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\gamma}}^{\text{T}}, \hat{p}_{11}, \hat{p}_{10}, \hat{\tau}_{\text{V}}, \hat{\tau}_{\text{N}})^{\text{T}}$.

By estimating function theory, the variance of $\hat{\boldsymbol{\theta}}$ can then be estimated by the empirical sandwich estimator:

$$\hat{Var}(\hat{\boldsymbol{\theta}}) = \frac{1}{n} A_n(\hat{\boldsymbol{\theta}})^{-1} B_n(\hat{\boldsymbol{\theta}}) \{A_n(\hat{\boldsymbol{\theta}})^{-1}\}^{\text{T}},$$

where $A_n(\hat{\boldsymbol{\theta}}) = \dfrac{1}{n} \sum_{i=1}^n \left\{ -\dfrac{\partial}{\partial \boldsymbol{\theta}^{\text{T}}} \Psi_c(Y_i^*, T_i, X_i, Y_i; \hat{\boldsymbol{\theta}}) \right\}$,

and $B_n(\hat{\boldsymbol{\theta}}) = \dfrac{1}{n} \sum_{i=1}^n \Psi_c(Y_i^*, T_i, X_i, Y_i; \hat{\boldsymbol{\theta}}) \Psi_c^{\text{T}}(Y_i^*, T_i, X_i, Y_i; \hat{\boldsymbol{\theta}})$. Let $\hat{V}_{i,j}$ be the element of the ith row and the jth column of $\hat{Var}(\hat{\boldsymbol{\theta}})$. Then $\hat{Var}(\hat{\tau}_{\text{V}}) = \hat{V}_{d+3,d+3}$, $\hat{Cov}(\hat{\tau}_{\text{V}}, \hat{\tau}_{\text{N}}) = \hat{V}_{d+3,d+4}$, and $\hat{Var}(\hat{\tau}_{\text{N}}) = \hat{V}_{d+4,d+4}$, where d is the dimension of $\boldsymbol{\gamma}$.

C.4 Proof of Theorem 4.3

Let l_1 and l_0 be $\hat{q}_1$ and $\hat{q}_0$ with estimated parameters of the postulated outcome models replaced by the true parameters of the postulated outcome models. It suffices to show

$$E\left\{\frac{TY^*}{e(p_{11}-p_{10})}-\frac{T-e}{e}l_1-\frac{T}{e}\left(\frac{p_{10}}{p_{11}-p_{10}}\right)\right\}=E(Y_1)$$

and

$$E\left\{\frac{(1-T)Y^*}{(1-e)(p_{11}-p_{10})}+\frac{T-e}{1-e}l_0-\frac{1-T}{1-e}\left(\frac{p_{10}}{p_{11}-p_{10}}\right)\right\}=E(Y_0).$$

Note

$$\begin{aligned}
& E\left\{\frac{TY^*}{e(p_{11}-p_{10})}\right\} \\
= \; & E\left[E\left\{\frac{TY^*}{e(p_{11}-p_{10})}\middle| X\right\}\right] \\
= \; & \frac{1}{p_{11}-p_{10}}E\left\{\frac{P(T=1,Y^*=1|X)}{e}\right\} \\
= \; & \frac{1}{p_{11}-p_{10}}E\left\{\frac{P(T=1|X)\{q_1p_{11}+(1-q_1)p_{10}\}}{e}\right\} \\
= \; & E\left\{\frac{P(T=1|X)q_1}{e}\right\}+\frac{p_{10}}{p_{11}-p_{10}}E\left\{\frac{P(T=1|X)}{e}\right\},
\end{aligned}$$

$$\begin{aligned}
& E\left(\frac{T-e}{e}l_1\right) \\
= \; & E\left\{E\left(\frac{T-e}{e}l_1\middle| X\right)\right\} \\
= \; & E\left\{\frac{P(T=1|X)-e}{e}l_1\right\},
\end{aligned}$$

and

$$\begin{aligned} & E\left\{\frac{T}{e}\left(\frac{p_{10}}{p_{11}-p_{10}}\right)\right\} \\ = & \frac{p_{10}}{p_{11}-p_{10}} E\left[E\left(\frac{T}{e}\middle| X\right)\right] \\ = & \frac{p_{10}}{p_{11}-p_{10}} E\left\{\frac{P(T=1|X)}{e}\right\}. \end{aligned}$$

So

$$\begin{aligned} & E\left\{\frac{TY^*}{e(p_{11}-p_{10})} - \frac{T-e}{e} l_1 - \frac{T}{e}\left(\frac{p_{10}}{p_{11}-p_{10}}\right)\right\} \\ = & E\left\{\frac{P(T=1|X)q_1}{e}\right\} - E\left\{\frac{P(T=1|X)-e}{e} l_1\right\} \\ = & \begin{cases} E(q_1), \text{ when treatment model is correct}, i.e., e = P(T=1|X) \\ E(q_1), \text{ when outcome model is correct}, i.e., l_1 = q_1, l_0 = q_0 \end{cases}, \end{aligned}$$

which equals $E(Y_1)$, since $E(q_1) = E\{P(Y=1|T=1,X)\} = E(Y_1)$.

Similarly, we calculate

$$\begin{aligned} & E\left\{\frac{(1-T)Y^*}{(1-e)(p_{11}-p_{10})}\right\} \\ = & E\left[E\left\{\frac{(1-T)Y^*}{(1-e)(p_{11}-p_{10})}\middle| X\right\}\right] \\ = & \frac{1}{p_{11}-p_{10}} E\left\{\frac{P(T=0, Y^*=1|X)}{1-e}\right\} \\ = & \frac{1}{p_{11}-p_{10}} E\left\{\frac{P(T=0|X)\{q_0 p_{11} + (1-q_0)p_{10}\}}{1-e}\right\} \\ = & E\left\{\frac{P(T=0|X)q_0}{1-e}\right\} + \frac{p_{10}}{p_{11}-p_{10}} E\left\{\frac{P(T=0|X)}{1-e}\right\}, \end{aligned}$$

$$E\left(\frac{T-e}{1-e} l_0\right)$$

$$
\begin{aligned}
&= E\left\{E\left(\frac{T-e}{1-e}l_0\middle|X\right)\right\} \\
&= E\left\{\frac{P(T=1|X)-e}{1-e}l_0\right\} \\
&= E\left\{\frac{1-P(T=0|X)-e}{1-e}l_0\right\},
\end{aligned}
$$

and

$$
\begin{aligned}
& E\left\{\frac{1-T}{1-e}\left(\frac{p_{10}}{p_{11}-p_{10}}\right)\right\} \\
&= \frac{p_{10}}{p_{11}-p_{10}}E\left[E\left(\frac{1-T}{1-e}\middle|X\right)\right] \\
&= \frac{p_{10}}{p_{11}-p_{10}}E\left\{\frac{P(T=0|X)}{1-e}\right\}.
\end{aligned}
$$

So

$$
\begin{aligned}
& E\left\{\frac{(1-T)Y^*}{(1-e)(p_{11}-p_{10})}+\frac{T-e}{1-e}l_0-\frac{1-T}{1-e}\left(\frac{p_{10}}{p_{11}-p_{10}}\right)\right\} \\
&= E\left\{\frac{P(T=0|X)q_0}{1-e}\right\}+E\left\{\frac{1-P(T=0|X)-e}{1-e}l_0\right\} \\
&= \begin{cases} E(q_0), \text{ when treatment model is correct}, i.e., e=P(T=1|X) \\ E(q_0), \text{ when outcome model is correct}, i.e., l_1=q_1, l_0=q_0 \end{cases},
\end{aligned}
$$

which equals $E(Y_0)$, since $E(q_0)=E\{P(Y=1|T=0,X)\}=E(Y_0)$.

Appendix D

Proofs for the Results in Chapter 5

D.1 Proof of Theorem 5.1

To show the conclusion, it suffices to show

$$\frac{1}{(p_{11}-p_{10})}E\{TY^*G(Z,X^*,T)\}-\frac{p_{10}}{p_{11}-p_{10}}=E(Y_1)$$

and

$$\frac{1}{(p_{11}-p_{10})}E\{(1-T)Y^*G(Z,X^*,T)\}-\frac{p_{10}}{p_{11}-p_{10}}=E(Y_0).$$

Let $q_1=P(Y=1|T=1,X,Z)$ and $q_0=P(Y=1|T=0,X,Z)$. Noting that

$$E\{TY^*G(Z,X^*,T)\}=E[E\{TY^*G(Z,X^*,T)|X,Z\}], \tag{D.1}$$

we evaluate

$$\begin{aligned}
& E\{TY^*G(Z,X^*,T)|X,Z\}\\
= \;& P(T=1|X,Z)E\{Y^*G(Z,X^*,T)|X,Z,T=1\}\\
= \;& P(T=1|X,Z)[P(Y=1|X,Z,T=1)E\{Y^*G(Z,X^*,T)|X,Z,T=1,Y=1\}\\
& +P(Y=0|X,Z,T=1)E\{Y^*G(Z,X^*,T)|X,Z,T=1,Y=0\}]
\end{aligned}$$

$$
\begin{aligned}
&= P(T=1|X,Z)[q_1 E\{Y^* G(Z,X^*,T)|X,Z,T=1,Y=1\} \\
&\quad +(1-q_1)E\{Y^* G(Z,X^*,T)|X,Z,T=1,Y=0\}] \\
&= P(T=1|X,Z)[q_1 P(Y^*=1|X,Z,T=1,Y=1)E\{G(Z,X^*,T)|X,Z,T=1,Y=1,Y^*=1\} \\
&\quad +(1-q_1)P(Y^*=1|X,Z,T=1,Y=0)E\{G(Z,X^*,T)|X,Z,T=1,Y=0,Y^*=1\}] \\
&= P(T=1|X,Z)[q_1 p_{11} E\{G(Z,X^*,T)|X,Z,T=1,Y=1,Y^*=1\} \\
&\quad +(1-q_1)p_{10}E\{G(Z,X^*,T)|X,Z,T=1,Y=0,Y^*=1\}], \qquad \text{(D.2)}
\end{aligned}
$$

where the second equality comes from that $E(U) = \sum_{k=0}^{1} P(Y=k)E(U|Y=k)$ for a random variable U, and the last equality is due to (5.4).

Next, we show that

$$
E\{G(Z,X^*,T)|X,Z,T=1,Y,Y^*\} = E\{G(Z,X^*,T)|X,Z,T=1\}. \qquad \text{(D.3)}
$$

Indeed, let $f(\cdot|\cdot)$ represent the conditional probability density or mass function for the corresponding random variables indicated by the arguments. Then we have that

$$
\begin{aligned}
&E\{G(Z,X^*,T)|X,Z,T=1,Y,Y^*\} \\
&= \int G(Z,x^*,1) f(x^*|X,Z,T=1,Y,Y^*)dx^* \\
&= \int G(Z,x^*,1)\cdot \frac{f(x^*,X,Z,T=1,Y,Y^*)}{f(X,Z,T=1,Y,Y^*)}dx^* \\
&= \int G(Z,x^*,1)\cdot \frac{f(X,Z,T=1)f(x^*|X,Z,T=1)f(Y|x^*,X,Z,T=1)f(Y^*|Y,x^*,X,Z,T=1)}{f(X,Z,T=1)f(Y|X,Z,T=1)f(Y^*|Y,X,Z,T=1)}dx^* \\
&= \int G(Z,x^*,1)\cdot \frac{f(x^*|X,Z,T=1)f(Y|X,Z,T=1)f(Y^*|Y)}{f(Y|X,Z,T=1)f(Y^*|Y)}dx^* \\
&= \int G(Z,x^*,1) f(x^*|X,Z,T=1)dx^* \\
&= E\{G(Z,X^*,T)|X,Z,T=1\},
\end{aligned}
$$

where the fourth equality is due to (5.3) and (5.4).

Then applying (5.9) and (D.3) to (D.2), we obtain that

$$
E\{TY^* G(Z,X^*,T)|X,Z\}
$$

$$
\begin{aligned}
&= P(T=1|X,Z)\left\{q_1 p_{11}\frac{1}{P(T=1|X,Z)} + (1-q_1)p_{10}\frac{1}{P(T=1|X,Z)}\right\} \\
&= q_1 p_{11} + (1-q_1)p_{10}.
\end{aligned}
$$

Therefore, using (D.1) gives

$$
\begin{aligned}
& E\{TY^*G(Z,X^*,T)\} \\
= \;& (p_{11}-p_{10})E(q_1) + p_{10} \\
= \;& (p_{11}-p_{10})E\{P(Y=1|T=1,X,Z)\} + p_{10} \\
= \;& (p_{11}-p_{10})E(Y_1) + p_{10},
\end{aligned}
$$

and hence

$$
\frac{1}{(p_{11}-p_{10})}E\{TY^*G(Z,X^*,T)\} - \frac{p_{10}}{p_{11}-p_{10}} = E(Y_1).
$$

Similarly, examining $E\{(1-T)Y^*G(Z,X^*,T)|X,Z\}$ yields

$$
\frac{1}{(p_{11}-p_{10})}E\{(1-T)Y^*G(Z,X^*,T)\} - \frac{p_{10}}{p_{11}-p_{10}} = E(Y_0).
$$

D.2 Justification for the Use of (5.12)

We now show that (5.12) meets the two conditions (5.9) and (5.10). By (5.12),

$$
\begin{aligned}
& E\{G(Z,X^*,T)|X,Z,T=1\} \\
= \;& 1 + E[\exp\{(-\alpha_0 - \boldsymbol{\alpha}_Z^{\mathrm{T}}Z - \boldsymbol{\alpha}_X^{\mathrm{T}}\Delta)(2T-1)\}|X,Z,T=1] \\
= \;& 1 + E\{\exp([-\alpha_0 - \boldsymbol{\alpha}_Z^{\mathrm{T}}Z - \boldsymbol{\alpha}_X^{\mathrm{T}}\{X^* + (T-1/2)\boldsymbol{\Sigma}_\epsilon\boldsymbol{\alpha}_X\}])|X,Z,T=1\} \\
= \;& 1 + E\{\exp(-\alpha_0 - \boldsymbol{\alpha}_Z^{\mathrm{T}}Z - \boldsymbol{\alpha}_X^{\mathrm{T}}X^* - \boldsymbol{\alpha}_X^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon\boldsymbol{\alpha}_X/2)|X,Z,T=1\} \\
= \;& 1 + \exp(-\alpha_0 - \boldsymbol{\alpha}_Z^{\mathrm{T}}Z - \boldsymbol{\alpha}_X^{\mathrm{T}}\boldsymbol{\Sigma}_\epsilon\boldsymbol{\alpha}_X/2)E\{\exp(-\boldsymbol{\alpha}_X^{\mathrm{T}}X^*|X,Z,T=1)\}. \qquad \text{(D.4)}
\end{aligned}
$$

By (5.3), $X^*|(X, Z, T) \sim N(X, \boldsymbol{\Sigma}_\epsilon)$. Then using the moment generating function of normal distributions, we obtain that (D.4) equals

$$\begin{aligned} & 1 + \exp(-\alpha_0 - \boldsymbol{\alpha}_Z^{\mathrm{T}} Z - \boldsymbol{\alpha}_X^{\mathrm{T}} \boldsymbol{\Sigma}_\epsilon \boldsymbol{\alpha}_X / 2) \exp(-\boldsymbol{\alpha}_X^{\mathrm{T}} X + \boldsymbol{\alpha}_X^{\mathrm{T}} \boldsymbol{\Sigma}_\epsilon \boldsymbol{\alpha}_X / 2) \\ = \; & 1 + \exp(-\alpha_0 - \boldsymbol{\alpha}_Z^{\mathrm{T}} Z - \boldsymbol{\alpha}_X^{\mathrm{T}} X) \\ = \; & \frac{1}{P(T = 1|X, Z)}, \end{aligned}$$

where the last step is due to model (5.11).

Similarly,

$$E\{G(Z, X^*, T)|X, Z, T = 0\} = \frac{1}{P(T = 0|X, Z)}.$$

Therefore, (5.12) meets the two conditions (5.9) and (5.10).

Appendix E

Proofs for the Results in Chapter 6

E.1 Proof of Theorem 6.1

Asymptotic Bias of $\hat{\tau}^{**}$:

Note that $\hat{\tau}^{**}$ can be re-written as

$$\begin{aligned}\hat{\tau}^{**} &= \frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} \frac{T_i Y_i^* R_i}{\hat{e}_i} - \frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} \frac{(1-T_i) Y_i^* R_i}{1-\hat{e}_i} \\ &= \frac{n}{\sum_{i=1}^{n} R_i} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{T_i Y_i^* R_i}{\hat{e}_i} \right\} - \frac{n}{\sum_{i=1}^{n} R_i} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{(1-T_i) Y_i^* R_i}{1-\hat{e}_i} \right\},\end{aligned}$$

$$\hat{\tau}^{**} \xrightarrow{p} \frac{1}{E(R)} E\left\{ \frac{T Y^* R}{e} \right\} - \frac{1}{E(R)} E\left\{ \frac{(1-T) Y^* R}{1-e} \right\}.$$

Calculate

$$\begin{aligned}& E\left(\frac{T Y^* R}{e} \middle| X \right) \\ =\ & \frac{1}{e} E(T Y^* R | X) \\ =\ & \frac{1}{e} P(T=1, Y^*=1, R=1 | X)\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{e}P(T=1|X)P(R=1|T=1,X)P(Y^*=1|T=1,R=1,X) \\
&= P(R=1|T=1,X)P(Y^*=1|T=1,R=1,X) \\
&= P(R=1|T=1,X)\{P(Y=1,Y^*=1|T=1,R=1,X)+P(Y=0,Y^*=1|T=1,R=1,X)\} \\
&= P(R=1|T=1,X)\{p_{11}P(Y=1|T=1,R=1,X)+p_{10}P(Y=0|T=1,R=1,X)\} \\
&= P(R=1|T=1,X)\{(p_{11}-p_{10})P(Y=1|T=1,X)+p_{10}\} \\
&= P(R=1|T=1,X)\{(p_{11}-p_{10})P(Y_1=1|T=1,X)+p_{10}\} \\
&= P(R=1|T=1,X)\{(p_{11}-p_{10})P(Y_1=1|X)+p_{10}\} \qquad \text{(E.1)}
\end{aligned}
$$

By (E.1),

$$
E\left\{\frac{TY^*R}{e}\right\} = E\left\{E\left(\frac{TY^*R}{e}\middle|X\right)\right\} = E[P(R=1|T=1,X)\{(p_{11}-p_{10})P(Y_1=1|X)+p_{10}\}],
$$

Similarly,

$$
\begin{aligned}
&E\left\{\frac{(1-T)Y^*R}{1-e}\middle|X\right\} \\
&= \frac{1}{1-e}E\{(1-T)Y^*R|X\} \\
&= \frac{1}{1-e}P(T=0,Y^*=1,R=1|X) \\
&= \frac{1}{1-e}P(T=0|X)P(R=1|T=0,X)P(Y^*=1|T=0,R=1,X) \\
&= P(R=1|T=0,X)P(Y^*=1|T=0,R=1,X) \\
&= P(R=1|T=0,X)\{P(Y=1,Y^*=1|T=0,R=1,X)+P(Y=0,Y^*=1|T=0,R=1,X)\} \\
&= P(R=1|T=0,X)\{p_{11}P(Y=1|T=0,R=1,X)+p_{10}P(Y=0|T=0,R=1,X)\} \\
&= P(R=1|T=0,X)\{(p_{11}-p_{10})P(Y=1|T=0,X)+p_{10}\} \\
&= P(R=1|T=0,X)\{(p_{11}-p_{10})P(Y_0=1|T=0,X)+p_{10}\} \\
&= P(R=1|T=0,X)\{(p_{11}-p_{10})P(Y_0=1|X)+p_{10}\} \qquad \text{(E.2)}
\end{aligned}
$$

By (E.2),

$$\begin{aligned} & E\left\{\frac{(1-T)Y^*R}{1-e}\right\} = E\left[E\left\{\frac{(1-T)Y^*R}{1-e}\middle| X\right\}\right] \\ = \;& E[P(R=1|T=0,X)\{(p_{11}-p_{10})P(Y_0=1|X)+p_{10}\}]. \end{aligned}$$

Therefore,

$$\begin{aligned} \hat{\tau}^{**} \xrightarrow{p} & \frac{E[P(R=1|T=1,X)\{(p_{11}-p_{10})P(Y_1=1|X)+p_{10}\}]}{E(R)} \\ & -\frac{E[P(R=1|T=0,X)\{(p_{11}-p_{10})P(Y_0=1|X)+p_{10}\}]}{E(R)} \end{aligned}$$

and the asymptotic bias of $\hat{\tau}^{**}$ is

$$\begin{aligned} \text{Bias}(\hat{\tau}^{**}) = & \frac{E[P(R=1|T=1,X)\{(p_{11}-p_{10})P(Y_1=1|X)+p_{10}\}]}{E(R)} \\ & -\frac{E[P(R=1|T=0,X)\{(p_{11}-p_{10})P(Y_0=1|X)+p_{10}\}]}{E(R)} - \tau_0. \end{aligned}$$

Asymptotic Bias of $\hat{\tau}^*$:

Since $\hat{\tau}^* = \hat{\tau}^{**}/(p_{11}-p_{10})$, it is immediate that the asymptotic bias of $\hat{\tau}^*$ is

$$\begin{aligned} \text{Bias}(\hat{\tau}^{*}) = & \frac{E[P(R=1|T=1,X)\{P(Y_1=1|X)+p_{10}/(p_{11}-p_{10})\}]}{E(R)} \\ & -\frac{E[P(R=1|T=0,X)\{P(Y_0=1|X)+p_{10}/(p_{11}-p_{10})\}]}{E(R)} - \tau_0. \end{aligned}$$

Asymptotic Bias of $\tilde{\tau}^*$:

Since

$$\tilde{\tau}^* = \frac{1}{n}\sum_{i=1}^{n}\frac{T_iY_i^*R_i}{e_i\hat{P}(R_i=1|T_i=1,X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i^*R_i}{(1-e_i)\hat{P}(R_i=1|T_i=0,X_i)},$$

we obtain

$$\tilde{\tau}^* \xrightarrow{p} E\left\{\frac{TY^*R}{eP(R=1|T=1,X)}\right\} - E\left\{\frac{(1-T)Y^*R}{(1-e)P(R=1|T=0,X)}\right\}.$$

Similarly, we can show

$$E\left(\frac{TY^*R}{eP(R=1|T=1,X)}\middle| X\right) = (p_{11}-p_{10})P(Y_1=1|X) + p_{10} \tag{E.3}$$

and

$$E\left\{\frac{(1-T)Y^*R}{(1-e)P(R=1|T=0,X)}\middle| X\right\} = (p_{11}-p_{10})P(Y_0=1|X) + p_{10}. \tag{E.4}$$

Therefore

$$\begin{aligned}\tilde{\tau}^* &\xrightarrow{p} E\{(p_{11}-p_{10})P(Y_1=1|X)+p_{10}\} - E\{(p_{11}-p_{10})P(Y_0=1|X)+p_{10}\}\\ &= (p_{11}-p_{10})[E\{P(Y_1=1|X)\} - E\{P(Y_0=1|X)\}]\\ &= (p_{11}-p_{10})\tau_0\end{aligned}$$

and the asymptotic bias of $\tilde{\tau}^*$ is $\text{Bias}(\tilde{\tau}^*) = (p_{11}-p_{10}-1)\tau_0$.

E.2 Proof of Theorem 6.2

Theorem 6.2(a):

It suffices to show

$$E\left\{\frac{TY^*R}{eP(R=1|T=1,X)(p_{11}-p_{10})}\right\} - \frac{p_{10}}{p_{11}-p_{10}} = E(Y_1)$$

and

$$E\left\{\frac{(1-T)Y^*R}{(1-e)P(R=1|T=0,X)(p_{11}-p_{10})}\right\} - \frac{p_{10}}{p_{11}-p_{10}} = E(Y_0),$$

which is immediate by using (E.3) and (E.4).

Theorem 6.2(b):

Note that

$$\begin{aligned} & E\left(\frac{RT}{eP(R=1|T=1,X)}\middle|X\right) \\ = & \frac{E(RT|X)}{eP(R=1|T=1,X)} \\ = & \frac{P(R=1,T=1|X)}{eP(R=1|T=1,X)} \\ = & \frac{P(T=1|X)P(R=1|T=1,X)}{eP(R=1|T=1,X)} \\ = & 1, \end{aligned}$$

and

$$\begin{aligned} & E\left(\frac{R(1-T)}{(1-e)P(R=1|T=0,X)}\middle|X\right) \\ = & \frac{P(R=1,T=0|X)}{(1-e)P(R=1|T=0,X)} \\ = & \frac{P(T=0|X)P(R=1|T=0,X)}{(1-e)P(R=1|T=0,X)} \\ = & 1. \end{aligned}$$

So

$$E\left(\frac{RT}{eP(R=1|T=1,X)}\right)^{-1} E\left\{\frac{TY^*R}{eP(R=1|T=1,X)(p_{11}-p_{10})}\right\} - \frac{p_{10}}{p_{11}-p_{10}} = E(Y_1)$$

and

$$E\left(\frac{R(1-T)}{(1-e)P(R=1|T=0,X)}\right)^{-1} E\left\{\frac{(1-T)Y^*R}{(1-e)P(R=1|T=0,X)(p_{11}-p_{10})}\right\} - \frac{p_{10}}{p_{11}-p_{10}} = E(Y_0),$$

and the consistency is justified.

E.3 Proof of Theorem 6.3

Let $q_1 = P(Y = 1|T = 1, X)$ and $q_0 = P(Y = 1|T = 0, X)$, l_1 and l_0 be $\hat{q}_1$ and $\hat{q}_0$ with estimated parameters replaced by the true parameters. It suffices to show

$$E\left\{\frac{TY^*R}{e(p_{11}-p_{10})P(R=1|T=1,X)} - \frac{T-e}{e}l_1 - \frac{T}{e}\left(\frac{p_{10}}{p_{11}-p_{10}}\right)\right\} = E(Y_1)$$

and

$$E\left\{\frac{(1-T)Y^*R}{(1-e)(p_{11}-p_{10})P(R=1|T=0,X)} + \frac{T-e}{1-e}l_0 - \frac{1-T}{1-e}\left(\frac{p_{10}}{p_{11}-p_{10}}\right)\right\} = E(Y_0).$$

Note

$$\begin{aligned}
& E\left\{\frac{TY^*R}{e(p_{11}-p_{10})P(R=1|T=1,X)}\right\} \\
= \; & E\left[E\left\{\frac{TY^*R}{e(p_{11}-p_{10})P(R=1|T=1,X)}\middle|X\right\}\right] \\
= \; & \frac{1}{p_{11}-p_{10}}E\left\{\frac{P(T=1,Y^*=1,R=1|X)}{eP(R=1|T=1,X)}\right\} \\
= \; & \frac{1}{p_{11}-p_{10}}E\left\{\frac{P(T=1|X)P(R=1|X,T=1)\{q_1p_{11}+(1-q_1)p_{10}\}}{eP(R=1|T=1,X)}\right\} \\
= \; & E\left\{\frac{P(T=1|X)q_1}{e}\right\} + \frac{p_{10}}{p_{11}-p_{10}}E\left\{\frac{P(T=1|X)}{e}\right\},
\end{aligned}$$

$$\begin{aligned}
& E\left(\frac{T-e}{e}l_1\right) \\
= \; & E\left\{E\left(\frac{T-e}{e}l_1\middle|X\right)\right\} \\
= \; & E\left\{\frac{P(T=1|X)-e}{e}l_1\right\},
\end{aligned}$$

and

$$\begin{aligned} & E\left\{\frac{T}{e}\left(\frac{p_{10}}{p_{11}-p_{10}}\right)\right\} \\ = & \frac{p_{10}}{p_{11}-p_{10}} E\left[E\left(\frac{T}{e}\middle| X\right)\right] \\ = & \frac{p_{10}}{p_{11}-p_{10}} E\left\{\frac{P(T=1|X)}{e}\right\}. \end{aligned}$$

So

$$\begin{aligned} & E\left\{\frac{TY^*R}{e(p_{11}-p_{10})P(R=1|T=1,X)} - \frac{T-e}{e} l_1 - \frac{T}{e}\left(\frac{p_{10}}{p_{11}-p_{10}}\right)\right\} \\ = & E\left\{\frac{P(T=1|X)q_1}{e}\right\} - E\left\{\frac{P(T=1|X)-e}{e} l_1\right\} \\ = & \begin{cases} E(q_1), \text{ when treatment model is correct}, i.e., e = P(T=1|X) \\ E(q_1), \text{ when outcome model is correct}, i.e., l_1 = q_1, l_0 = q_0 \end{cases}, \end{aligned}$$

which equals $E(Y_1)$, since $E(q_1) = E\{P(Y=1|T=1,X)\} = E(Y_1)$.

Similarly, we calculate

$$\begin{aligned} & E\left\{\frac{(1-T)Y^*R}{(1-e)(p_{11}-p_{10})P(R=1|T=0,X)}\right\} \\ = & E\left[E\left\{\frac{(1-T)Y^*R}{(1-e)(p_{11}-p_{10})P(R=1|T=0,X)}\middle| X\right\}\right] \\ = & \frac{1}{p_{11}-p_{10}} E\left\{\frac{P(T=0,Y^*=1,R=1|X)}{(1-e)P(R=1|T=0,X)}\right\} \\ = & \frac{1}{p_{11}-p_{10}} E\left[\frac{P(T=0|X)P(R=1|T=0,X)\{q_0 p_{11}+(1-q_0)p_{10}\}}{(1-e)P(R=1|T=0,X)}\right] \\ = & E\left\{\frac{P(T=0|X)q_0}{1-e}\right\} + \frac{p_{10}}{p_{11}-p_{10}} E\left\{\frac{P(T=0|X)}{1-e}\right\}, \end{aligned}$$

$$E\left(\frac{T-e}{1-e} l_0\right)$$

$$
\begin{aligned}
&= E\left\{E\left(\frac{T-e}{1-e}l_0\middle|X\right)\right\} \\
&= E\left\{\frac{P(T=1|X)-e}{1-e}l_0\right\} \\
&= E\left\{\frac{1-P(T=0|X)-e}{1-e}l_0\right\},
\end{aligned}
$$

and

$$
\begin{aligned}
& E\left\{\frac{1-T}{1-e}\left(\frac{p_{10}}{p_{11}-p_{10}}\right)\right\} \\
&= \frac{p_{10}}{p_{11}-p_{10}}E\left[E\left(\frac{1-T}{1-e}\middle|X\right)\right] \\
&= \frac{p_{10}}{p_{11}-p_{10}}E\left\{\frac{P(T=0|X)}{1-e}\right\}.
\end{aligned}
$$

So

$$
\begin{aligned}
& E\left\{\frac{(1-T)Y^*R}{(1-e)(p_{11}-p_{10})P(R=1|T=0,X)}+\frac{T-e}{1-e}l_0-\frac{1-T}{1-e}\left(\frac{p_{10}}{p_{11}-p_{10}}\right)\right\} \\
= \; & E\left\{\frac{P(T=0|X)q_0}{1-e}\right\}+E\left\{\frac{1-P(T=0|X)-e}{1-e}l_0\right\} \\
= \; & \begin{cases} E(q_0), \text{ when treatment model is correct}, i.e., e=P(T=1|X) \\ E(q_0), \text{ when outcome model is correct}, i.e., l_1=q_1, l_0=q_0 \end{cases},
\end{aligned}
$$

which equals $E(Y_0)$, since $E(q_0)=E\left\{P(Y=1|T=0,X)\right\}=E(Y_0)$.

Appendix F

Proofs for the Results in Chapter 7

F.1 Proof of Theorem 7.1

Consistency when $\mathcal{E}$ contains a correctly specified model:

Suppose the set of postulated treatment models $\mathcal{E}$ contains a correctly specified model. Without loss of generality, let the first model $e^1(\boldsymbol{\gamma}^1; X)$ be correctly specified. Let $\boldsymbol{\gamma}_0^1$ be the true value of $\boldsymbol{\gamma}^1$, i.e., $e^1(\boldsymbol{\gamma}_0^1; X) = e(X) = P(T = 1|X)$.

By arguments of Han and Wang (2013), we have

$$\hat{w}_i = \frac{1}{m} \frac{\hat{\theta}^1 / e^1(\hat{\boldsymbol{\gamma}}^1; X_i)}{1 + \hat{\lambda}^{\mathrm{T}} \hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) / e^1(\hat{\boldsymbol{\gamma}}^1; X_i)},$$

where $\hat{\lambda} = O_p(n^{-1/2})$ and $\hat{\theta}^1 = n^{-1} \sum_{i=1}^n e^1(\hat{\boldsymbol{\gamma}}^1; X_i)$. Then $1 + \hat{\lambda}^{\mathrm{T}} \hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) / e^1(\hat{\boldsymbol{\gamma}}^1; X_i) \xrightarrow{p} 0$ and $\hat{\theta}^1 \xrightarrow{p} E\{e^1(\boldsymbol{\gamma}_0^1; X)\} = P(T = 1)$. As a nonparametric estimator of $P(T = 1)$, m/n well approximates $P(T = 1)$ as sample becomes larger. Thus, with regularity conditions satisfied,

$$\hat{E}(Y_1) = \sum_{i=1}^m \frac{\hat{w}_i Y_i^*}{p_{11} - p_{10}} - \frac{p_{10}}{p_{11} - p_{10}}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i^*}{p_{11}-p_{10}}\frac{n}{m}\frac{\hat{\theta}^1/e^1(\hat{\boldsymbol{\gamma}}^1;X_i)}{1+\hat{\lambda}^{\mathrm{T}}\hat{g}_i(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}})/e^1(\hat{\boldsymbol{\gamma}}^1;X_i)} - \frac{p_{10}}{p_{11}-p_{10}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i^*}{e^1(\hat{\boldsymbol{\gamma}}^1;X_i)(p_{11}-p_{10})} - \frac{p_{10}}{p_{11}-p_{10}} + o_p(1).
\end{aligned}
$$

We observe that

$$
\frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i^*}{e^1(\hat{\boldsymbol{\gamma}}^1;X_i)(p_{11}-p_{10})} - \frac{p_{10}}{p_{11}-p_{10}}
$$

is consistent estimator for $\hat{E}(Y_1)$ by Chapter 4. Therefore, the consistency of $\hat{E}(Y_1)$ is established.

Similarly, with regularity conditions satisfied, we have

$$
\hat{E}(Y_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i^*}{\{1-e^1(\hat{\boldsymbol{\gamma}}^1;X_i)\}(p_{11}-p_{10})} - \frac{p_{10}}{p_{11}-p_{10}} + o_p(1).
$$

Observe that

$$
\frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i^*}{\{1-e^1(\hat{\boldsymbol{\gamma}}^1;X_i)\}(p_{11}-p_{10})} - \frac{p_{10}}{p_{11}-p_{10}}
$$

is a consistent estimator $\hat{E}(Y_0)$ by Chapter 4. Therefore, the consistency of $\hat{E}(Y_0)$ is established.

Consistency when $\mathcal{Q}$ contains a correctly specified model:

Suppose the set of postulated outcome models $\mathcal{Q}$ contains a correctly specified model. Without loss of generality, let the first model $q_t^1(\boldsymbol{\beta}^1;X)$ be correctly specified and $\boldsymbol{\beta}_0^1$ be the true value of $\boldsymbol{\beta}^1$. Then $q_t^1(\boldsymbol{\beta}_0^1;X) = q_t(X) = P(Y=1|X,T=t)$.

By constraint $\sum_{i=1}^{m}\hat{w}_i = 1$ and $\sum_{i=m+1}^{n}\tilde{w}_i = 1$, we observe that $\hat{E}(Y_1)$ and $\hat{E}(Y_1)$ can be re-written as

$$
\hat{E}(Y_1) = \sum_{i=1}^{m}\hat{w}_i\left(\frac{Y_i^*}{p_{11}-p_{10}}\right) - \frac{p_{10}}{p_{11}-p_{10}} = \sum_{i=1}^{m}\hat{w}_i\left(\frac{Y_i^*-p_{10}}{p_{11}-p_{10}}\right)
$$

and

$$
\hat{E}(Y_0) = \sum_{i=m+1}^{n}\tilde{w}_i\left(\frac{Y_i^*}{p_{11}-p_{10}}\right) - \frac{p_{10}}{p_{11}-p_{10}} = \sum_{i=m+1}^{n}\tilde{w}_i\left(\frac{Y_i^*-p_{10}}{p_{11}-p_{10}}\right).
$$

We have

$$\sum_{i=1}^{m} \hat{w}_i \left(\frac{Y_i^* - p_{10}}{p_{11} - p_{10}} \right) = \sum_{i=1}^{m} \hat{w}_i \left\{ \left(\frac{Y_i^* - p_{10}}{p_{11} - p_{10}} \right) - q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) \right\} + \sum_{i=1}^{m} \hat{w}_i q_1^1(\hat{\boldsymbol{\beta}}^1; X_i).$$

The constraint

$$\sum_{i=1}^{m} \hat{w}_i \{ q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) - n^{-1} \sum_{i=1}^{n} q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) \} = 0$$

gives

$$\sum_{i=1}^{m} \hat{w}_i q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) = \sum_{i=1}^{m} \hat{w}_i \frac{1}{n} \sum_{i=1}^{n} q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) = \frac{1}{n} \sum_{i=1}^{n} q_1^1(\hat{\boldsymbol{\beta}}^1; X_i),$$

where

$$\frac{1}{n} \sum_{i=1}^{n} q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) \xrightarrow{p} E\{q_1^1(\boldsymbol{\beta}_0^1; X)\} = E(Y_1).$$

Thus it remains to show $\sum_{i=1}^{m} \hat{w}_i \left\{ \left(\frac{Y_i^* - p_{10}}{p_{11} - p_{10}} \right) - q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) \right\} \xrightarrow{p} 0.$

$$\begin{aligned}
& \sum_{i=1}^{m} \hat{w}_i \left\{ \left(\frac{Y_i^* - p_{10}}{p_{11} - p_{10}} \right) - q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) \right\} \\
= \; & \frac{1}{m} \sum_{i=1}^{m} \frac{(Y_i^* - p_{10})/(p_{11} - p_{10}) - q_1^1(\hat{\boldsymbol{\beta}}^1; X_i)}{1 + \hat{\rho}^{\mathrm{T}} \hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})} \Bigg/ \left\{ \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1 + \hat{\rho}^{\mathrm{T}} \hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})} \right\} \\
= \; & \frac{n}{m} \frac{1}{n} \sum_{i=1}^{n} \frac{T_i \{(Y_i^* - p_{10})/(p_{11} - p_{10}) - q_1^1(\hat{\boldsymbol{\beta}}^1; X_i)\}}{1 + \hat{\rho}^{\mathrm{T}} \hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})} \Bigg/ \left\{ \frac{n}{m} \frac{1}{n} \sum_{i=1}^{n} \frac{T_i}{1 + \hat{\rho}^{\mathrm{T}} \hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})} \right\} \\
\xrightarrow{p} \; & \frac{1}{P(T=1)} E \left[\frac{T\{(Y^* - p_{10})/(p_{11} - p_{10}) - q_1^1(\boldsymbol{\beta}_0^1; X)\}}{1 + \rho_*^{\mathrm{T}} g(\boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)} \right] \Bigg/ \left[\frac{1}{P(T=1)} E \left\{ \frac{T}{1 + \rho_*^{\mathrm{T}} g(\boldsymbol{\gamma}_*, \boldsymbol{\beta}_*))} \right\} \right],
\end{aligned}$$

where ρ_*, $\boldsymbol{\gamma}_*$ and $\boldsymbol{\beta}_*$ are the limiting values of ρ, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, respectively. Note

$$\begin{aligned}
& \frac{1}{P(T=1)} E \left(E \left[\frac{T\{(Y^* - p_{10})/(p_{11} - p_{10}) - q_1^1(\boldsymbol{\beta}_0^1; X)\}}{1 + \rho_*^{\mathrm{T}} g(\boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)} \Bigg| X \right] \right) \\
= \; & \frac{1}{P(T=1)} E \left(\frac{1}{1 + \rho_*^{\mathrm{T}} g(\boldsymbol{\gamma}_*, \boldsymbol{\beta}_*)} E \left[T\{(Y^* - p_{10})/(p_{11} - p_{10}) - q_1^1(\boldsymbol{\beta}_0^1; X)\} \Big| X \right] \right),
\end{aligned}$$

where

$$
\begin{aligned}
& E\left[T\{(Y^*-p_{10})/(p_{11}-p_{10})-q_1^1(\boldsymbol{\beta}_0^1;X)\}\middle|X\right] \\
= \;& P(T=1|X)E\left\{\frac{Y_i^*}{(p_{11}-p_{10})}-\frac{p_{10}}{(p_{11}-p_{10})}-q_1^1(\boldsymbol{\beta}_0^1;X)\middle|X,T=1\right\} \\
= \;& P(T=1|X)\left\{\frac{P(Y_i^*=1|X,T=1)}{(p_{11}-p_{10})}-\frac{p_{10}}{(p_{11}-p_{10})}-q_1^1(\boldsymbol{\beta}_0^1;X)\right\} \\
= \;& P(T=1|X)\left[\frac{p_{11}q_1(X)+p_{10}\{1-q_1(X)\}}{(p_{11}-p_{10})}-\frac{p_{10}}{(p_{11}-p_{10})}-q_1^1(\boldsymbol{\beta}_0^1;X)\right] \\
= \;& P(T=1|X)\{q_1(X)-q_1^1(\boldsymbol{\beta}_0^1;X)\} \\
= \;& 0.
\end{aligned}
$$

Therefore $\hat{E}(Y_1)\xrightarrow{p}E(Y_1)$. Similarly, $\hat{E}(Y_0)\xrightarrow{p}E(Y_0)$. The consistency is established.

F.2 Proof of Theorem 7.2

Consistency when $\mathcal{E}$ contains a correctly specified model:

By arguments in Appendix F.1, we have

$$
\begin{aligned}
& \hat{E}(Y_1)=\sum_{i=1}^m \hat{w}_i Y_i R_i/\pi_{11}(\hat{\boldsymbol{\alpha}};X_i) \\
= \;& \frac{1}{n}\sum_{i=1}^n \frac{T_i R_i Y_i}{\pi_{11}(\hat{\boldsymbol{\alpha}};X_i)}\frac{n}{m}\frac{\hat{\theta}^1/e^1(\hat{\boldsymbol{\gamma}}^1;X_i)}{1+\hat{\lambda}^{\mathrm{T}}\hat{g}_i(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}})/e^1(\hat{\boldsymbol{\gamma}}^1;X_i)} \\
= \;& \frac{1}{n}\sum_{i=1}^n \frac{T_i R_i Y_i}{e^1(\hat{\boldsymbol{\gamma}}^1;X_i)\pi_{11}(\hat{\boldsymbol{\alpha}};X_i)}+o_p(1).
\end{aligned}
$$

We observe that

$$
\begin{aligned}
& E\left\{\frac{TRY}{e^1(\boldsymbol{\gamma}_0^1;X)\pi_{11}(\boldsymbol{\alpha};X)}\right\} \\
= \;& E\left[E\left\{\frac{TRY}{e^1(\boldsymbol{\gamma}_0^1;X)\pi_{11}(\boldsymbol{\alpha};X)}\right\}\middle|X\right]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[\frac{P(T=1,R=1,Y=1|X)}{e^1(\boldsymbol{\gamma}_0^1;X)\pi_{11}(\boldsymbol{\alpha};X)}\right] \\
&= E\left[\frac{P(T=1|X)P(Y=1|T=1,X)P(R=1|Y=1,T=1,X)}{e^1(\boldsymbol{\gamma}_0^1;X)\pi_{11}(\boldsymbol{\alpha};X)}\right] \\
&= E\{P(Y=1|T=1,X)\} \qquad \text{(by MAR assumption)} \\
&= E(Y_1).
\end{aligned}
$$

Therefore, the consistency of $\hat{E}(Y_1)$ is established.

Similarly, with regularity conditions satisfied, we have

$$\hat{E}(Y_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)R_iY_i}{\{1-e^1(\hat{\boldsymbol{\gamma}}^1;X_i)\}\pi_{10}(\hat{\boldsymbol{\alpha}};X_i)} + o_p(1),$$

and

$$E\left[\frac{(1-T)RY}{\{1-e^1(\boldsymbol{\gamma}_0^1;X)\}\pi_{10}(\boldsymbol{\alpha};X)}\right] = E(Y_0).$$

Therefore, the consistency of $\hat{E}(Y_0)$ is established.

Consistency when $\mathcal{Q}$ contains a correctly specified model:

$$\hat{E}(Y_1) = \sum_{i=1}^{m} \hat{w}_i Y_i R_i/\pi_{11}(\hat{\boldsymbol{\alpha}};X_i) = \sum_{i=1}^{m} \hat{w}_i\{Y_iR_i/\pi_{11}(\hat{\boldsymbol{\alpha}};X_i) - q_1^1(\hat{\boldsymbol{\beta}}^1;X_i\} + \hat{w}_i q_1^1(\hat{\boldsymbol{\beta}}^1;X_i).$$

The constraint

$$\sum_{i=1}^{m} \hat{w}_i\{q_1^1(\hat{\boldsymbol{\beta}}^1;X_i) - n^{-1}\sum_{i=1}^{n} q_1^1(\hat{\boldsymbol{\beta}}^1;X_i)\} = 0$$

gives

$$\sum_{i=1}^{m} \hat{w}_i q_1^1(\hat{\boldsymbol{\beta}}^1;X_i) = \sum_{i=1}^{m} \hat{w}_i \frac{1}{n}\sum_{i=1}^{n} q_1^1(\hat{\boldsymbol{\beta}}^1;X_i) = \frac{1}{n}\sum_{i=1}^{n} q_1^1(\hat{\boldsymbol{\beta}}^1;X_i),$$

where

$$\frac{1}{n}\sum_{i=1}^{n} q_1^1(\hat{\boldsymbol{\beta}}^1;X_i) \xrightarrow{p} E\{q_1^1(\boldsymbol{\beta}_0^1;X)\} = E(Y_1).$$

Thus it remains to show $\sum_{i=1}^m \hat{w}_i\{Y_iR_i/\pi_{11}(\hat{\boldsymbol{\alpha}};X_i) - q_1^1(\hat{\boldsymbol{\beta}}^1;X_i\} \xrightarrow{p} 0$. Calculate

$$
\begin{aligned}
& \sum_{i=1}^m \hat{w}_i\{Y_iR_i/\pi_{11}(\hat{\boldsymbol{\alpha}};X_i) - q_1^1(\hat{\boldsymbol{\beta}}^1;X_i\} \\
= \;& \frac{1}{m}\sum_{i=1}^m \frac{Y_iR_i/\pi_{11}(\hat{\boldsymbol{\alpha}};X_i) - q_1^1(\hat{\boldsymbol{\beta}}^1;X_i)}{1+\hat{\rho}^{\mathrm{T}}\hat{g}_i(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}})} \Big/ \left\{\frac{1}{m}\sum_{i=1}^m \frac{1}{1+\hat{\rho}^{\mathrm{T}}\hat{g}_i(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}})}\right\} \\
= \;& \frac{n}{m}\frac{1}{n}\sum_{i=1}^n \frac{T_i\{R_iY_i/\pi_{11}(\hat{\boldsymbol{\alpha}};X_i) - q_1^1(\hat{\boldsymbol{\beta}}^1;X_i)\}}{1+\hat{\rho}^{\mathrm{T}}\hat{g}_i(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}})} \Big/ \left\{\frac{n}{m}\frac{1}{n}\sum_{i=1}^n \frac{T_i}{1+\hat{\rho}^{\mathrm{T}}\hat{g}_i(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}})}\right\} \\
\xrightarrow{p} \;& \frac{1}{P(T=1)}E\left[\frac{T\{RY/\pi_{11}(\boldsymbol{\alpha};X) - q_1^1(\boldsymbol{\beta}_0^1;X)\}}{1+\rho_*^{\mathrm{T}}g(\boldsymbol{\gamma}_*,\boldsymbol{\beta}_*)}\right] \Big/ \left[\frac{1}{P(T=1)}E\left\{\frac{T}{1+\rho_*^{\mathrm{T}}g(\boldsymbol{\gamma}_*,\boldsymbol{\beta}_*))}\right\}\right],
\end{aligned}
$$

where ρ_*, $\boldsymbol{\gamma}_*$ and $\boldsymbol{\beta}_*$ are the limiting values of ρ, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, respectively. Note

$$
\begin{aligned}
& E\left[\frac{T\{RY/\pi_{11}(\boldsymbol{\alpha};X) - q_1^1(\boldsymbol{\beta}_0^1;X)\}}{1+\rho_*^{\mathrm{T}}g(\boldsymbol{\gamma}_*,\boldsymbol{\beta}_*)}\middle| X\right] \\
= \;& \frac{P(T=1|X)E\{RY/\pi_{11}(\boldsymbol{\alpha};X) - q_1^1(\boldsymbol{\beta}_0^1;X)|X,T=1\}}{1+\rho_*^{\mathrm{T}}g(\boldsymbol{\gamma}_*,\boldsymbol{\beta}_*)} \\
= \;& \frac{P(T=1|X)}{1+\rho_*^{\mathrm{T}}g(\boldsymbol{\gamma}_*,\boldsymbol{\beta}_*)}[E\{RY/\pi_{11}(\boldsymbol{\alpha};X)|X,T=1\} - q_1^1(\boldsymbol{\beta}_0^1;X)] \\
= \;& \frac{P(T=1|X)}{1+\rho_*^{\mathrm{T}}g(\boldsymbol{\gamma}_*,\boldsymbol{\beta}_*)}[P(R=1|X,Y=1,T=1)P(Y=1|X,T=1)/\pi_{11}(\boldsymbol{\alpha};X) - q_1^1(\boldsymbol{\beta}_0^1;X)] \\
= \;& \frac{P(T=1|X)}{1+\rho_*^{\mathrm{T}}g(\boldsymbol{\gamma}_*,\boldsymbol{\beta}_*)}[P(Y=1|X,T=1) - q_1^1(\boldsymbol{\beta}_0^1;X)] \qquad \text{(by MAR assumption)} \\
= \;& 0.
\end{aligned}
$$

Therefore $\hat{E}(Y_1) \xrightarrow{p} E(Y_1)$. Similarly, $\hat{E}(Y_0) \xrightarrow{p} E(Y_0)$. The consistency is established.

F.3 Proof of Theorem 7.3

Consistency when $\mathcal{E}$ contains a correctly specified model:

By arguments in Appendix F.1, we have

$$\begin{aligned}
\hat{E}(Y_1) &= \sum_{i=1}^{m} \frac{\hat{w}_i Y_i^* R_i}{\pi_{11}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{T_i R_i Y_i^*}{\pi_{11}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} \frac{n}{m} \frac{\hat{\theta}^1 / e^1(\hat{\boldsymbol{\gamma}}^1; X_i)}{1 + \hat{\lambda}^{\mathrm{T}} \hat{g}_i(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}) / e^1(\hat{\boldsymbol{\gamma}}^1; X_i)} - \frac{p_{10}}{p_{11} - p_{10}} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{T_i R_i Y_i^*}{e^1(\hat{\boldsymbol{\gamma}}^1; X_i) \pi_{11}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} + o_p(1).
\end{aligned}$$

Caculate

$$\begin{aligned}
& E\left\{ \frac{TRY^*}{e^1(\boldsymbol{\gamma}_0^1; X) \pi_{11}(\boldsymbol{\alpha}; X)(p_{11} - p_{10})} \right\} - \frac{p_{10}}{p_{11} - p_{10}} \\
= \; & \frac{1}{p_{11} - p_{10}} E\left[E\left\{ \frac{TRY^*}{e^1(\boldsymbol{\gamma}_0^1; X) \pi_{11}(\boldsymbol{\alpha}; X)} \right\} \middle| X \right] - \frac{p_{10}}{p_{11} - p_{10}} \\
= \; & \frac{1}{p_{11} - p_{10}} E\left\{ \frac{P(T=1, R=1, Y^*=1|X)}{e^1(\boldsymbol{\gamma}_0^1; X) \pi_{11}(\boldsymbol{\alpha}; X)} \right\} - \frac{p_{10}}{p_{11} - p_{10}} \\
= \; & \frac{1}{p_{11} - p_{10}} E\left\{ \frac{P(T=1|X) P(R=1|X, T=1) P(Y^*=1|X, T=1, R=1)}{e^1(\boldsymbol{\gamma}_0^1; X) \pi_{11}(\boldsymbol{\alpha}; X)} \right\} - \frac{p_{10}}{p_{11} - p_{10}} \\
= \; & \frac{1}{p_{11} - p_{10}} E\{ p_{11} P(Y=1|X, T=1, R=1) + p_{10} P(Y=0|X, T=1, R=1) \} - \frac{p_{10}}{p_{11} - p_{10}} \\
= \; & E\{ P(Y=1|T=1, X) \} \qquad \text{(by MAR assumption)} \\
= \; & E(Y_1).
\end{aligned}$$

Therefore, the consistency of $\hat{E}(Y_1)$ is established.

Similarly, with regularity conditions satisfied, we have

$$\hat{E}(Y_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{(1 - T_i) R_i Y_i^*}{\{1 - e^1(\hat{\boldsymbol{\gamma}}^1; X_i)\} \pi_{10}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} + o_p(1),$$

and

$$E\left[\frac{(1 - T) R Y^*}{\{1 - e^1(\boldsymbol{\gamma}_0^1; X)\} \pi_{10}(\boldsymbol{\alpha}; X)(p_{11} - p_{10})} \right] - \frac{p_{10}}{p_{11} - p_{10}} = E(Y_0).$$

Therefore, the consistency of $\hat{E}(Y_0)$ is established.

Consistency when $\mathcal{Q}$ contains a correctly specified model:

By constraint $\sum_{i=1}^{m} \hat{w}_i = 1$ and $\sum_{i=m+1}^{n} \tilde{w}_i = 1$, we observe that $\hat{E}(Y_1)$ and $\hat{E}(Y_1)$ can be re-written as

$$\hat{E}(Y_1) = \sum_{i=1}^{m} \frac{\hat{w}_i Y_i^* R_i}{\pi_{11}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} = \sum_{i=1}^{m} \hat{w}_i \left\{ \frac{Y_i^* R_i}{\pi_{11}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} \right\}$$

and

$$\hat{E}(Y_0) = \sum_{i=m+1}^{n} \frac{\tilde{w}_i Y_i^* R_i}{\pi_{10}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} = \sum_{i=m+1}^{n} \tilde{w}_i \left\{ \frac{Y_i^* R_i}{\pi_{10}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} \right\}.$$

We have

$$\begin{aligned} & \sum_{i=1}^{m} \hat{w}_i \left\{ \frac{Y_i^* R_i}{\pi_{11}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} \right\} \\ = & \sum_{i=1}^{m} \hat{w}_i \left\{ \frac{Y_i^* R_i}{\pi_{11}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} - q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) \right\} + \sum_{i=1}^{m} \hat{w}_i q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) \end{aligned}$$

The constraint

$$\sum_{i=1}^{m} \hat{w}_i \{ q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) - n^{-1} \sum_{i=1}^{n} q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) \} = 0$$

gives

$$\sum_{i=1}^{m} \hat{w}_i q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) = \sum_{i=1}^{m} \hat{w}_i \frac{1}{n} \sum_{i=1}^{n} q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) = \frac{1}{n} \sum_{i=1}^{n} q_1^1(\hat{\boldsymbol{\beta}}^1; X_i),$$

where

$$\frac{1}{n} \sum_{i=1}^{n} q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) \xrightarrow{p} E\{q_1^1(\boldsymbol{\beta}_0^1; X)\} = E(Y_1).$$

Thus it remains to show $\sum_{i=1}^{m} \hat{w}_i \left\{ \frac{Y_i^* R_i}{\pi_{11}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} - q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) \right\} \xrightarrow{p} 0.$

Denote $Z_i = \frac{Y_i^* R_i}{\hat{\pi}_{11}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}}$, then

$$\sum_{i=1}^{m} \hat{w}_i \left\{ \frac{Y_i^* R_i}{\pi_{11}(\hat{\boldsymbol{\alpha}}; X_i)(p_{11} - p_{10})} - \frac{p_{10}}{p_{11} - p_{10}} - q_1^1(\hat{\boldsymbol{\beta}}^1; X_i) \right\}$$

$$
\begin{aligned}
&= \frac{1}{m}\sum_{i=1}^{m}\frac{Z_i - q_1^1(\hat{\boldsymbol{\beta}}^1; X_i)}{1+\hat{\rho}^{\mathrm{T}}\hat{g}_i(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}})}\Bigg/\left\{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{1+\hat{\rho}^{\mathrm{T}}\hat{g}_i(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}})}\right\} \\
&= \frac{n}{m}\frac{1}{n}\sum_{i=1}^{n}\frac{T_i\{Z_i - q_1^1(\hat{\boldsymbol{\beta}}^1; X_i)\}}{1+\hat{\rho}^{\mathrm{T}}\hat{g}_i(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}})}\Bigg/\left\{\frac{n}{m}\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{1+\hat{\rho}^{\mathrm{T}}\hat{g}_i(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}})}\right\} \\
&\xrightarrow{p} \frac{1}{P(T=1)}E\left[\frac{T\{Z - q_1^1(\boldsymbol{\beta}_0^1; X)\}}{1+\rho_*^{\mathrm{T}}g(\boldsymbol{\gamma}_*,\boldsymbol{\beta}_*)}\right]\Bigg/\left[\frac{1}{P(T=1)}E\left\{\frac{T}{1+\rho_*^{\mathrm{T}}g(\boldsymbol{\gamma}_*,\boldsymbol{\beta}_*))}\right\}\right],
\end{aligned}
$$

where ρ_*, $\boldsymbol{\gamma}_*$ and $\boldsymbol{\beta}_*$ are the limiting values of ρ, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Note

$$
\begin{aligned}
& E\left[\frac{T\{Z - q_1^1(\boldsymbol{\beta}_0^1; X)\}}{1+\rho_*^{\mathrm{T}}g(\boldsymbol{\gamma}_*,\boldsymbol{\beta}_*)}\middle| X\right] \\
= \;& \frac{E[T\{Z - q_1^1(\boldsymbol{\beta}_0^1; X)\}|X]}{1+\rho_*^{\mathrm{T}}g(\boldsymbol{\gamma}_*,\boldsymbol{\beta}_*)} \\
= \;& \frac{P(T=1|X)E\{Z - q_1^1(\boldsymbol{\beta}_0^1; X)|X, T=1\}}{1+\rho_*^{\mathrm{T}}g(\boldsymbol{\gamma}_*,\boldsymbol{\beta}_*)},
\end{aligned}
$$

where

$$
\begin{aligned}
& E\{Z - q_1^1(\boldsymbol{\beta}_0^1; X)|X, T=1\} \\
= \;& E\left\{\frac{Y^*R}{\pi_{11}(\boldsymbol{\alpha}; X)(p_{11}-p_{10})} - \frac{p_{10}}{p_{11}-p_{10}} - q_1^1(\boldsymbol{\beta}_0^1; X)\middle| X, T=1\right\} \\
= \;& \frac{P(R=1|X,T=1)\{p_{11}P(Y=1|X,T=1,R=1) + p_{10}P(Y=0|X,T=1,R=1)\}}{\pi_{11}(\boldsymbol{\alpha}; X)(p_{11}-p_{10})} \\
& -\frac{p_{10}}{p_{11}-p_{10}} - q_1^1(\boldsymbol{\beta}_0^1; X) \\
= \;& P(Y=1|X,T=1) + \frac{p_{10}}{p_{11}-p_{10}} - \frac{p_{10}}{p_{11}-p_{10}} - q_1^1(\boldsymbol{\beta}_0^1; X) \quad \text{(by MAR assumption)} \\
= \;& 0
\end{aligned}
$$

Therefore $\hat{E}(Y_1) \xrightarrow{p} E(Y_1)$. Similarly, $\hat{E}(Y_0) \xrightarrow{p} E(Y_0)$. The consistency is established.