2019-08-15

Causal Inference with Missingness in Confounders

Bagmar, Md. Shaddam Hossain

UNIVERSITY OF CALGARY

Causal Inference with Missingness in Confounders

by

Md. Shaddam Hossain Bagmar

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS

CALGARY, ALBERTA
AUGUST, 2019

Abstract

Causal inference is the process of uncovering causal connection between the effect variable and disease outcome in epidemiologic research. Confounders that influence both the effect variable and outcome need to be accounted for when obtaining the causal effect in observational studies. In addition, missing data often arise in the data collection procedure, working with complete cases often results in biased estimates. We consider the estimation of causal effect in the presence of missingness in the confounders under the missing at random assumption. We investigate how different estimators namely regression, G-estimation, propensity score-based estimators including matching, stratification, weighting, propensity regression and finally doubly robust estimator, perform when applying complete-case analysis or multiple imputation. Due to the uncertainty of imputation model and computational challenge for large number of imputations, we propose an expectation-maximization (EM) algorithm to estimate the expected values of the missing confounder and utilize weighting approach in the estimation of average treatment effect. Simulation studies are conducted to see whether there is any gain in estimation efficiency under the proposed method than complete case analysis and multiple imputation. The analysis identified EM method as most efficient and accurate method for dealing missingness in confounder except for propensity score matching and inverse weighting estimators. In these two estimators, multiple imputation is found as efficient, however EM is efficient for inverse weighting when the outcome is binary. Real life data application is shown for estimating the effect of adjuvant radiation treatment on patient's survival status after 10 years of breast cancer diagnosis. Under missing completely at random (MCAR) mechanism, EM is found as the most accurate method for handling missingness in confounder than multiple imputation.

Acknowledgements

First of all, I express my solemn gratitude to almighty ALLAH, who has given me the ability, strength and opportunity to perform this thesis work.

Later on, the thanks go to my amazing advisor, Dr. Hua Shen, for her immense support and extremely valuable guidance throughout my masters thesis. Dr. Hua has been very generous with her time. Throughout so many long discussions we had together, I have benefited not only from her deep insights in statistics, but also learned how to conduct research with creativeness and better vision. I am sure they will have a lasting influence on my future career as an applied statistician. Dr. Hua has been supportive all through my graduate studies with her consistent and illuminating instruction.

I would like to thank Dr. Jingjing Wu and Dr. Xuewen Lu for serving as a member in my thesis defense committee. I owe my honest gratitude to Dr. Xiaolan Feng for the permission to use the Breast Cancer (BC) data and Dr. Haocheng Li for valuable discussions regarding the BC data. I also want to thank all the faculty members and staff in our department for their invaluable help. Their kindness and help during these wonderful years at the University of Calgary is appreciated. Special thanks to Freddie Yau for his help in executing my R program on the departmental server. I want to thank the department for providing me the opportunity of education and thesis.

Finally, I am grateful to my family for their unconditional love and support throughout my whole life. I thank my beloved wife, Ajmery Jaman for all the support, care, sacrifice, and everything she have ever done and continue to do for me. Because of her, I have never felt alone or discouraged facing all the challenges in life. Thank you Ajmery for your love and trust. This thesis is never possible without you as being the emotional pillar.

August, 2019 Md. Shaddam Hossain Bagmar

To my beloved wife and our baby boy Tahmid. . .

Table of Contents

List of Figures and Illustrations

List of Tables

List of Symbols, Abbreviations and Nomenclature

Symbol or abbreviation	Definition
AIPW	Augmented inverse probability weighting
ATE	Average treatment effect
CC	Complete case
DR	Doubly robust
EM	Expectation-maximization
IPW	Inverse probability weighting
MAR	Missing at random
MCAR	Missing completely at random
MI	Multiple imputation
ML	Maximum likelihood
MNAR	Missing not at random
PS	Propensity score
α	Parameter vector for propensity score model
β	Parameter vector for outcome regression model
γ	Parameter vector for missing model
e	Propensity score

Chapter 1

Introduction

1.1 Introduction of Causal Inference

Causal inference is the process of uncovering causal connection between the effect variable and disease outcome in epidemiologic research. In this section we set out our basic framework for causal inference. We discuss three key notions underlying causal inference. The first notion is that of assumptions, when drawing causal inferences from observational studies, which we use to estimate the effect of interest. Second, we define the average treatment effect (ATE) as a measure of causal effect considering the causal assumptions. Finally, we discuss the different estimators which are used for inferring causal effects.

1.1.1 Assumptions

Causal estimation from randomized experiments is straightforward and can be done using the available statistical approaches. But in practice, randomization may not be possible due to the ethical issue and cost per observation. Observational study is the most common source of data in public health research to estimate treatment effects. However, an observational study with uncontrolled treatment assignment, require certain assumptions to mimic the design of randomized experiments (Hernán and Robins, 2019). We now discuss four causal assumptions namely, the stable unit treatment value assumption (SUTVA), ignorability, consistency and positivity. The principle goal of making these assumptions is to link potential outcomes with observed data.

Assumption 1: The stable unit treatment value assumption

In many situations it may be reasonable to assume that treatments applied to one unit do not affect the outcome for another unit. For example, if we are in different locations and have no contact with each other, it would appear reasonable to assume that whether you take an ibuprofen has no effect on the status of my fever.

The stable unit treatment value assumption, or SUTVA (Rubin, 1980) incorporates two ideas that units do not interfere with one another and the concept that for each unit there is only a single version of each treatment level. The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes. These two elements of the stability assumption enable us to exploit the presence of multiple units for estimating causal effects. The assumption of SUTVA is referred to generally as exclusion restrictions: assumptions that rely on external, substantive, information to rule out the existence of a causal effect of a particular treatment relative to an alternative.

SUTVA: No Interference

Consider, first, the no-interference component of SUTVA- the assumption that the treatment applied to one unit does not affect the outcome for other units. For example, when studying the effect of different types of fertilizers in agricultural experiments on plot yields, traditionally researchers have taken care to separate plots using "guard rows", unfertilized strips of land between fertilized areas. By controlling the leaching of different fertilizers across experimental plots, these guard rows make SUTVA more credible; without them we might suspect that the fertilizer applied to one plot affected the yields in contiguous plots.

SUTVA: No Hidden Variations of Treatments

The second component of SUTVA requires that an individual receiving a specific treatment level cannot receive different forms of that treatment. Consider the causal effect of ibuprofen on fevers. For the potential outcome with both of us taking ibuprofen, we obviously need more than one ibuprofen tablet. Suppose, however, that one of the tablets is old and no longer contains a fully effective dose, whereas the other is new and at full strength. In that case, each of us may have three treatments available: no ibuprofen, the ineffective tablet,

and the effective tablet. There are thus two forms of the active treatment, both nominally labelled "ibuprofen": ibuprofen+ and ibuprofen-. Even with no interference one can now think of there being three potential outcomes for each of us, the no ibuprofen outcome Y_i(No ibuprofen), the weak ibuprofen outcome Y_i(ibuprofen-) and the strong ibuprofen outcome Y_i(ibuprofen+), with i indexing "I" or "You". The second part of SUTVA either requires that the two ibuprofen outcomes are identical: Y_i(ibuprofen+) = Y_i(ibuprofen-), or that I can only get ibuprofen+ and you can only get ibuprofen- (or vice versa). Alternatively we can redefine the treatment as taking a randomly selected ibuprofen (either ibuprofen- or ibuprofen+). In that case SUTVA might be satisfied for the redefined stochastic treatment.

Assumption 2: Consistency

The consistency assumption in principles is a pretty obvious or simple assumption which directly linking potential outcomes and observed data. The consistency assumption states:

The potential outcome under treatment T = t, Y(t), is equal to the observed outcome if the actual treatment received is T = t and this is true for all values of T. Mathematically, $Y = TY(1) + (1 - T)Y(0)$.

In the potential outcomes framework, we imagine $Y(t)$ is the outcome that would be observed if treatment actually took the value t. So if the treatment does actually take value t, then the *consistency* assumption mentions that the observed outcome is equal to that potential outcome for treatment value $T = t$. In our ibuprofen-fever example, if I take ibuprofen then the observed outcome is equal to the potential outcome Y(ibuprofen) or vice versa.

Assumption 3: Ignorability

The ignorability assumption involves all important pre-treatment covariates or confounders (X) in the process of causal estimation. This is also referred as the *no unmeasured confounders* or *exchangeability*. A confounder is a variable that influences both treatment (T) and outcome (Y) variables. The basic idea is that treatment assignment is assumed to be independent from potential outcomes, conditional on these pre-treatment variables or confounders.

Given the confounders X, treatment (T) assignment is independent from the potential out-

comes. Mathematically, $Y(t) \perp\!\!\!\perp T|X;\ \forall\ t \in T$.

The symbol '$\perp\!\!\!\perp$' indicates conditional independence among potential outcomes $Y(t)$ and treatment assignment T, where the condition is set to confounders X. That's why the assumption is also termed as conditional ignorability or conditional exchangeability. This assumptions indicates the importance of treatment assignment and measuring all important confounders in causal effect estimation.

Assumption 4: Positivity

The positivity assumption essentially states that, for every set of values of confounders (X), treatment assignment is not deterministic. Symbolically,

$$0 < P(T = t|X = x) < 1;\ \ \forall\ t\ \in\ T,\ \text{and}\ \ x\ \in\ X.$$

The value of the above probability can not be equal to 0 or 1, which ensures the non-deterministic properties of treatment assignment. The idea is that everybody in the population has some chance of getting either treatment (T) conditional on the confounder values (X). So, at every level of X, people have a non-zero probability of getting either treatment. In other words, treatment is not deterministic as a function of confounders, X. For example, it would be a violation of the positivity if everybody who are older get treated, but no violation if older people are just more likely to get the treatment.

If for a given value of X, everybody is treated, then there's really no way to learn what would've happened if they weren't treated. But if we have some people who are treated and some who aren't within every level of X, then there's some hope of learning about the causal effects of treatment within different levels of X. We need this positivity assumption to ensure that we have some data at every level of X for people who are treated and not treated.

Positivity assumption is also helping us to identify the population of interest. If there are people who could never get the treatment, then typically we would want to exclude them and only make inference about the population of people who have some chance of getting the treatment.

1.1.2 Measures of Treatment Effect

We now turn to Rubin's causal model of observational studies. Motivated by educational-researches, Rubin (1974) argued that "the use of carefully controlled non-randomized data to estimate causal effects is a reasonable and necessary procedure in many cases". Rubin (1974) used the potential outcomes to define the causal effect of an educational treatment. This language clearly links observational studies to the more general "missing data" problem (Rubin, 1976, 1977).

In Rubin's view, the most important quantity about observational studies is the treatment *assignment mechanism*. Let i index individuals and T_i denote a treatment indicator, equal to 1 if a person is treated, and equal to 0 otherwise. Let $Y_i(0)$ denote the potential outcome that would occur when person i is not treated ($T_i = 0$) and $Y_i(1)$ the potential outcome when they are treated ($T_i = 1$). Clearly these are not both observed and one of them will be "counterfactual", an outcome that would have occurred if a different treatment had been given. The observed outcome will be

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0). \tag{1.1}$$

The individual level treatment effect (ITE) is given by

$$\Delta_{ITE} = Y_i(1) - Y_i(0), \tag{1.2}$$

which is clearly not identified, because only one of the potential outcomes is observed. There are several measures that are identified under certain conditions. In most of the traditional observation studies, the quantity of interest (estimand) is the average treatment effect (ATE), which is defined as

$$\Delta_{ATE} = E\left[\mathbf{Y}(1) - \mathbf{Y}(0)\right] = \mu_1 - \mu_0, \tag{1.3}$$

where μ_1 and μ_0 are the average potential outcome under treated and control conditions respectively. Another interesting measure is the average effect of treatment on the treated (ATT), given by

$$\Delta_{ATT} = E\left[\mathbf{Y}(1) - \mathbf{Y}(0) | \mathbf{T} = 1\right] = \mu_{1,T} - \mu_{0,T}, \tag{1.4}$$

where $\mu_{1,T}$ and $\mu_{0,T}$ are the average potential outcome under treated and control conditions in the treated group. The expectations in the above definitions are taken over the joint distribution of $(\mathbf{T}, \mathbf{Y}(1), \mathbf{Y}(0))$, or in other words over a random draw from the infinite population and a random treatment assignment. By adopting this view, we also implicitly assumed the stable unit treatment value assumption (SUTVA) of Rubin (1980), which roughly says there is no interference between units. For continuous response $\mathbf{Y}$, Δ_{ATE} is the average over the entire population of the individual treatment effects and Δ_{ATT} is the average over the sub-population of treated people of the treatment effect. In case of binary response, these measures are simply the (causal) risk difference. Based of the research purpose, one may also be interested to find the risk ratio or odds ratio as a causal effect measure. An estimate of ATE can be obtained as

$$\hat{\Delta}_{ATE} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\} = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{1}{n}\sum_{i=1}^{n} Y_i(0) = \widehat{\mu_1} - \widehat{\mu_0}, \tag{1.5}$$

an estimate of ATT is given by

$$\begin{aligned}\hat{\Delta}_{ATT} &= \frac{1}{n_1}\sum_{T_i=1}\{Y_i(1) - Y_i(0)\}, \quad n_1 = \#\{1 \le i \le n : T_i = 1\} \\ &= \frac{1}{n_1}\sum_{T_i=1} Y_i(1) - \frac{1}{n_1}\sum_{T_i=1} Y_i(0) = \widehat{\mu_{1,T}} - \widehat{\mu_{0,T}}.\end{aligned} \tag{1.6}$$

Generally, the researcher use the ATE as a measure of causal estimate whereas the ATT is useful when there is interest in: the evaluation of the effects of treatments on those who received them or the projection of potential outcomes in a target (sub-) population (Wang et al., 2017). If in addition we assume that the mean potential outcome does not depend on treatment condition T_i, i.e. $E[\mathbf{Y}(1)|\mathbf{T}] = E[\mathbf{Y}(1)]$ for treated and $E[\mathbf{Y}(0)|\mathbf{T}] = E[\mathbf{Y}(0)]$ for control individual then we will find that ATE = ATT, since

$$\begin{aligned}\Delta_{ATT} &= E[\mathbf{Y}(1) - \mathbf{Y}(0)|\mathbf{T} = 1] = E[\mathbf{Y}(1)|\mathbf{T} = 1] - E[\mathbf{Y}(0)|\mathbf{T} = 1] \\ &= E[\mathbf{Y}(1)] - E[\mathbf{Y}(0)] = \Delta_{ATE}.\end{aligned} \tag{1.7}$$

In the definitions above, selecting units occurs before the treatment assignment so the estimands are defined with respect to the fixed n units. It is often more practical to view the n units as random

samples from a large population. Theoretically, it is more convenient to assume the population is infinite, so the units are independent and identically distributed (iid) draws.

1.1.3 Randomized Experiments versus Observational Studies

Causal inference is a popular technique to infer the effect of potential interventions, such as a treatment, a factor or a policy using either randomized experiments or observational studies. A common frustration in epidemiologic research is the adage that the lack of randomized controlled trials (RCTs) does not allow us to establish causality, but merely associations. Suppose we want to evaluate the effectiveness of a treatment for a certain disease. If we have obtained data on both the treatment and control group, why not simply compare outcomes across the two groups to estimate treatment effectiveness? The intuition behind why this comparison is generally a bad strategy is that the two groups may differ systematically in ways related to the outcome. In fact, it seems reasonable that they are likely to differ precisely because one group chose to take the treatment and the other does not. Presumably, systematic differences in the study participants caused this difference in choice of treatment. For instance, one group might be older, more educated, or more motivated. This problem of unequal distributions can be eliminated with random treatment assignment.

Researchers trust randomized experiments- the randomization creates two groups that, on average, are the same. It balances the characteristics, measured and unmeasured, of the study units across treatment groups. Therefore, differences in outcomes across the two groups can be confidently attributed to the treatment without having to rely on models for the outcomes to estimate treatment effects. However, even in RCT settings it is difficult (and perhaps impossible) to control for all the factors that could possibly influence the outcome of interest. A limitation of RCTs is that when these differences are not eliminated, there is no way for researchers to quantify the error that result caused by those uncontrolled differences. Another practical shortcoming of RCTs is the ethical concern raised by some interventions. In order to completely eliminate bias, one would have to assign potentially harmful interventions to, and thus intentionally jeopardize the wellbeing of study participants.

In observational studies, investigator have no control over the treatment assignment. That may

cause a large difference in the distribution of observed covariates between the treatment and control groups, which can lead biased estimates of the treatment effects. Due to these differences in observed covariates, any difference in the outcomes could be attributed not just to treatment but could also be attributed to any other characteristics that differed among the groups. In order to account for these differences between treatment groups we must be able to measure the characteristics that differ. In addition, standard approaches to treatment effect estimation such as linear regression require us to be able to model the conditional distribution of the outcomes given these measured covariates. Under the causal assumptions, observed data can be used to obtain the causal effects as below

$$\begin{aligned} E\left[\mathbf{Y}|\mathbf{T}=t,\mathbf{X}\right] &= E\left[\mathbf{Y}(t)|\mathbf{T}=t,\mathbf{X}\right] \quad \text{(consistency)} \\ &= E\left[\mathbf{Y}(t)|\mathbf{X}\right] \quad \text{(ignorability)}, \end{aligned} \tag{1.8}$$

where $E\left[\mathbf{Y}|\mathbf{T}=t,\mathbf{X}\right]$ involves only the observed data and we ends with $E\left[\mathbf{Y}(t)|\mathbf{X}\right]$ which is the conditional average potential outcomes at treatment level $T=t$. For the marginal average $E\left[\mathbf{Y}(t)\right]$, one needs to average over the distribution of confounders $\mathbf{X}$.

1.2 Different Estimators of ATE

Data can not speak for themselves. To extract the information contained in the data, one needs to apply different statistical tools. Models are useful not only to extract information but also for future prediction. In Section 1.1, we discuss the basic frame work of causal inference, required assumptions and the estimands. In this Section, we will discuss different types of estimator for causal estimands. Mainly, we will focus on two class of estimators: based on outcome regression models and treatment assignment models (Kang et al., 2007). Outcome regression is the model where we regress response on treatment and other pre-treatment covariates (confounders). Treatment assignment model or equivalently the propensity score (PS) model estimates the probability of getting treatment by specifying a model for treatment variable on other confounders. Using these two models, various estimators of causal effect measure ATE will be discussed.

1.2.1 Outcome Regression Based Estimators

Regression estimator

Linear regression analysis is the most widely used of all statistical techniques: it is the study to determine the linear relationship between two or more variables. Let us consider Y_i as the observed outcome, T_i as a binary treatment variable taking values 0 (control) or 1 (treated), and $\mathbf{X}_i$ as a p-dimensional vector of baseline covariates or confounders for ith subject, where $i = 1, 2, \ldots, n$. A generalized form of linear regression model also symbolize as $y-$model is given by

$$g(\mu_i) = \beta_0 + \beta_t T_i + \boldsymbol{\beta}_x^\top \mathbf{X}_i, \tag{1.9}$$

where $\mu_i = E(Y_i|T_i, \mathbf{X}_i, \boldsymbol{\beta})$ is the expected response for ith subject, $\boldsymbol{\beta}_x^\top = (\beta_1, \beta_2, \ldots, \beta_p)$, $\boldsymbol{\beta} = (\beta_0, \beta_t, \boldsymbol{\beta}_x^\top)^\top$ is the vector of regression coefficients for confounders and $g(\cdot)$ is the link function which establishes the connection between expected value of the response and linear component of the assumed model. For continuous outcome with an identity link function i.e., $g(\mu_i) = \mu_i$ we can write the above model as:

$$E(Y_i|T_i, \mathbf{X}_i, \boldsymbol{\beta}) = \mu_i = \beta_0 + \beta_t T_i + \boldsymbol{\beta}_x^\top \mathbf{X}_i. \tag{1.10}$$

In binary outcome (0 vs 1) with a *logit* link we can obtain the logistic regression model as:

$$\begin{aligned} g(\mu_i) = g(p_i) = logit\,(p_i)) &= \beta_0 + \beta_t T_i + \boldsymbol{\beta}_x^\top \mathbf{X}_i \\ \Rightarrow p_i = E(Y_i|T_i, \mathbf{X}_i, \boldsymbol{\beta}) &= expit\left\{\beta_0 + \beta_t T_i + \boldsymbol{\beta}_x^\top \mathbf{X}_i\right\}, \end{aligned} \tag{1.11}$$

where $\mu_i = E(Y_i|T_i, \mathbf{X}_i, \boldsymbol{\beta}) = Pr(Y_i = 1|T_i, \mathbf{X}_i, \boldsymbol{\beta}) = p_i$ which is the probability of being $Y = 1 \in \{0, 1\}$ for ith subject given the covariates (here its T and $\mathbf{X}$). We use two popular functions *logit* and *expit* where $logit(a) = log\left(\frac{a}{1-a}\right)$ and $expit(a) = \{1 + exp(-a)\}^{-1}$.

The definition of the regression models for both continuous and binary outcomes only depends on the observed data $(\mathbf{Y}, \mathbf{T}, \mathbf{X})$. Our goal is to estimate the ATE utilising these postulated models. Given the stated assumptions in Section 1.1.1, generally we can define the regression estimator of

ATE as:

$$\begin{aligned}\Delta^{(reg)} &= E[\mathbf{Y}(1)] - E[\mathbf{Y}(0)] = E\left\{E(\mathbf{Y}|\mathbf{T}=1,\mathbf{X})\right\} - E\left\{E(\mathbf{Y}|\mathbf{T}=0,\mathbf{X})\right\} \\ &= E\left\{E(\mathbf{Y}|\mathbf{T}=1,\mathbf{X}) - E(\mathbf{Y}|\mathbf{T}=0,\mathbf{X})\right\}. \end{aligned} \tag{1.12}$$

For continuous outcome the ATE can be obtained using the model defined in Equation (1.10) as:

$$\begin{aligned}\Delta_c^{(reg)} &= E\left\{E(\mathbf{Y}|\mathbf{T}=1,\mathbf{X}) - E(\mathbf{Y}|\mathbf{T}=0,\mathbf{X})\right\} \\ &= E\left\{(\beta_0 + \beta_t * 1 + \boldsymbol{\beta}_x^\top \mathbf{X}) - (\beta_0 + \beta_t * 0 + \boldsymbol{\beta}_x^\top \mathbf{X})\right\} \qquad (1.13) \\ &= E(\beta_t) = \beta_t, \qquad (1.14)\end{aligned}$$

where in $\Delta_c^{(reg)}$, c symbolizes the continuous outcome, and the regression coefficient of the treatment variable, β_t is the ATE if the assumed y-model (1.10) is true. So, ATE for continuous outcome can be estimated directly from this model by least squares i.e., $\hat{\Delta}^{(reg)} = \hat{\beta}_t$. In case of binary outcome using the model defined in Equation (1.11) we get

$$\begin{aligned}&E(\mathbf{Y}|\mathbf{T}=1,\mathbf{X}) - E(\mathbf{Y}|\mathbf{T}=0,\mathbf{X}) \\ &= expit\left(\beta_0 + \beta_t * 1 + \boldsymbol{\beta}_x^\top \mathbf{X}\right) - expit\left(\beta_0 + \beta_t * 0 + \boldsymbol{\beta}_x^\top \mathbf{X}\right) \\ &= expit\left(\beta_0 + \beta_t + \boldsymbol{\beta}_x^\top \mathbf{X}\right) - expit\left(\beta_0 + \boldsymbol{\beta}_x^\top \mathbf{X}\right), \end{aligned} \tag{1.15}$$

which is different from continuous outcome case. The parameter estimates $\boldsymbol{\beta} = (\hat{\beta}_0, \hat{\beta}_t, \hat{\boldsymbol{\beta}}_x^\top)^\top$ can be obtained from a logistic regression model. The estimated ATE for binary outcome is defined by averaging over all $\mathbf{X}_i$ as follows:

$$\hat{\Delta}_b^{(reg)} = \frac{1}{n}\sum_{i=1}^{n}\left\{expit\left(\hat{\beta}_0 + \hat{\beta}_t + \hat{\boldsymbol{\beta}}_x^\top \mathbf{X}_i\right) - expit\left(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}_x^\top \mathbf{X}_i\right)\right\}, \tag{1.16}$$

where in $\hat{\Delta}_b^{(reg)}$, b stands for binary outcome.

G-estimation

The approach G-estimation, can be thought of as a form of standardization (Snowden et al., 2011; Vansteelandt and Keiding, 2011), on the basis of the equality $E[\mathbf{Y}(t)] = E_{\mathbf{X}}[E[\mathbf{Y}(t)|\mathbf{X}]]$ where $t \in \mathbf{T}$ can take values 0 or 1. Under the consistency and ignorability assumptions, $E[\mathbf{Y}(t)|\mathbf{X}] = E[\mathbf{Y}|\mathbf{T} = t, \mathbf{X}]$. Suppose for a continuous outcome $\mathbf{Y}$, $E[\mathbf{Y}|\mathbf{T} = t, \mathbf{X}] = \mathbf{X}\boldsymbol{\beta}^{(\top)}$ where $\mathbf{X}$ is a $(n \times p+1)$ including an extra column of 1's for the intercept term with $\boldsymbol{\beta}^{(\top)} = (\beta_0^{(t)}, \beta_1^{(t)}, \ldots, \beta_p^{(t)})^\top$. The parameter $\boldsymbol{\beta}^{(t)}$ can be estimated using a linear regression model of observed outcome $\mathbf{Y}$ on the measured confounders $\mathbf{X}$ among subjects exposed to the treatment at level $\mathbf{T} = t$. Then, $E[\mathbf{Y}(t)|\mathbf{X}]$ can be estimated from this model for each subject, whether treated ($t = 1$) or control ($t = 0$), by $\hat{\mathbf{Y}}(t) = \mathbf{X}\hat{\boldsymbol{\beta}}^{(t)}$, using hats to indicate sample estimates. To estimate μ_t, we must take the expectation of the estimated $E[\mathbf{Y}(t)|\mathbf{X}]$ over the distribution of the measured confounders, $\mathbf{X}$. This can be estimated by taking the sample average of these estimates,

$$\hat{\mu}_{t,ge} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i(t), \tag{1.17}$$

where $\hat{Y}_i(t) = \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(t)}$ is an estimate of the potential outcome $Y(t)$ for subject i at treatment level $T = t$. This is commonly used as an estimate of the treatment effect at level t in medical and epidemiological research. It does, however, make the fairly strong assumptions that (i) the effect of the treatment, $(\mathbf{Y}(1), \mathbf{Y}(0))$, is the same for all individuals and (ii) the relationship between $\mathbf{Y}(1)$ and $\mathbf{T}$ is the same as that between $\mathbf{Y}(0)$ and $\mathbf{T}$. The latter assumption can be lessened by including interaction terms between the confounders and exposure in the regression model. Finally, the ATE is obtained as

$$\hat{\Delta}_c^{(ge)} = \hat{\mu}_{1,ge} - \hat{\mu}_{0,ge} = \frac{1}{n}\sum_{i=1}^{n} \left\{\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(1)} - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(0)}\right\}. \tag{1.18}$$

For a binary outcome $\mathbf{Y}$, we fit a logistic regression model in place of linear regression model and the corresponding ATE estimates will be as follows:

$$\hat{\Delta}_b^{(ge)} = \frac{1}{n}\sum_{i=1}^{n} \left\{expit\left(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(1)}\right) - expit\left(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(0)}\right)\right\}. \tag{1.19}$$

1.2.2 Propensity Score Based Estimators

The propensity score (PS), which is defined as the conditional probability of receiving treatment given covariates, plays a central role in a variety of settings for causal inference. PS's are most commonly estimated via logistic regression (Rubin et al., 2008) as:

$$logit\,(p_i) = \mathbf{X}_i^\top \boldsymbol{\alpha}, \tag{1.20}$$

where $\mathbf{X}$ is again a data matrix including an extra column of 1's for the intercept, $p_i = Pr(T_i = 1|\mathbf{X}_i) = expit\,(\mathbf{X}_i^\top \boldsymbol{\alpha})$ is the probability of being treated given the confounders for subject i. The fitted values from this logistic regression model are the estimated PS i.e., $\hat{p}_i = \hat{e}_i$.

PS based estimators are effective approaches in reducing the confounding bias due to measured covariates (Rosenbaum and Rubin, 1983). If the treatment assignment is strongly ignorable given observed covariates, an unbiased estimate of the ATE can be obtained by adjusting for the PS alone rather than a vector of confounders, which is often of high dimension. Several estimators based on the PS have been developed and become an essential part of applied researchers' across disciplines. In particular, the PS is used to adjust for observed confounding through matching (Rosenbaum and Rubin, 1985; Rosenbaum, 1989; Abadie and Imbens, 2006), subclassification (Rosenbaum and Rubin, 1984; Rosenbaum, 1991; Hansen, 2004), weighting (Rosenbaum, 1987; Robins et al., 2000; Hirano et al., 2003), regression (Heckman et al., 1998) or their combinations (Robins et al., 1995; Ho et al., 2007; Abadie and Imbens, 2011). Imbens (2004), Lunceford and Davidian (2004), Austin (2007) and Stuart (2010) have provided comprehensive reviews of these and other estimators based on PS's.

Matching

The goal of matching on PS is to construct a subset of the population in which the PS (e) have the same distribution in both the treated and the control. Typically, the majority of control patients have covariate values very different from the treated patients'. In this situation, a matching control patient may be identified for each active treatment patient based on PS. The observed outcome of each matched control patient is used to estimate the missing potential outcome for a matched treated patient.

Any matching algorithm must first specify a distance metric d on the PS. The most commonly used distance metrics are

$$d\{\hat{e}(\mathbf{X}_i), \hat{e}(\mathbf{X}_j)\} = \begin{cases} \left|\hat{e}(\mathbf{X}_i) - \hat{e}(\mathbf{X}_j)\right| & \text{or} \\ \left|logit[\hat{e}(\mathbf{X}_i)] - logit[\hat{e}(\mathbf{X}_j)]\right|, \end{cases} \tag{1.21}$$

where $logit(a) = log[a/(1-a)]$ is the *logit* transformation. The PS, probability of getting treatment given the confounders, is bounded between 0 and 1, making many values seem similar. But *logit* transformation of the PS is unbounded and can take any value on the real line by preserving the ranks of the PS itself. For the further improvement in the matching process, one can use a *caliper* which is the maximum distance that we are willing to tolerate. Utilizing the *caliper*, the distance measures in Equation (1.21) only accept a match if:

$$d\{\hat{e}(\mathbf{X}_i), \hat{e}(\mathbf{X}_j)\} = \begin{cases} \left|\hat{e}(\mathbf{X}_i) - \hat{e}(\mathbf{X}_j)\right| \leq C & \text{or} \\ \left|logit[\hat{e}(\mathbf{X}_i)] - logit[\hat{e}(\mathbf{X}_j)]\right| \leq C, \end{cases} \tag{1.22}$$

where C is the assigned caliper, sort of threshold between an acceptable and unacceptable match. In practice, the caliper C is set to 0.2 times the standard deviation of PS or *logit* of the PS depending on the distance measures.

Once the distance metric d is selected, we can apply a matching algorithm. One of the most common, and easiest to implement and understand, methods is $k:1$ nearest neighbour (or *caliper*) matching (Rubin, 1973; Imbens, 2015). In this algorithm, each treated unit is matched to k control units that are closest (or within the *caliper*) in terms of d, and the control units that are not selected are discarded. When $k = 1$, matching is the counterpart of paired design in randomized experiments. The user can also choose whether a control unit is allowed to be used multiple times as a match (matching with replacement). After a successful matching process, one can estimate the causal effects estimand (ATE) non-parametrically from the matched sample. Following table shows, how the matching algorithm estimates the ATE:

	Propensity score		Treatment indicator		Observed outcomes	
Unit	Treated	Matched	Treated	Matched	Treated	Matched
1	$f\{e(\mathbf{X}_1)\}$	$f\left\{e^M(\mathbf{X}_1)\right\}$	1	0	$Y_1(1)$	$Y_1^M(0)$
2	$f\{e(\mathbf{X}_2)\}$	$f\left\{e^M(\mathbf{X}_2)\right\}$	1	0	$Y_2(1)$	$Y_2^M(0)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_1	$f\{e(\mathbf{X}_{n_1})\}$	$f\left\{e^M(\mathbf{X}_{n_1})\right\}$	1	0	$Y_{n_1}(1)$	$Y_{n_1}^M(0)$

In the above table, n_1 is the total number of treated ($T = 1$) subjects, $e(\mathbf{X}_i)$ and $e^M(\mathbf{X}_i)$ are the PS, and $Y_i(1)$ and $Y_i^M(0)$ are the outcomes for treated and control subjects respectively in ith matched sample. We consider a general function $f\{e(.)\}$ of PS in the matching algorithm. If the matching is done on PS directly then $f\{e(.)\} = e(.)$ or $f\{e(.)\} = logit\{e(.)\}$, if matching is done on PS after *logit* transformation. From these matched sample, we can calculate the ATE as:

$$\hat{\Delta}^{(mc)} = \left[\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i(1) - \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i^M(0)\right] = \frac{1}{n_1}\sum_{i=1}^{n_1}\left[Y_i(1) - Y_i^M(0)\right] = \widehat{E_T}\left[Y_i(1)\right] - \widehat{E_T}\left[Y_i^M(0)\right]. \quad (1.23)$$

For continuous response $\mathbf{Y}$, the above formula will estimate the usual average causal effects for matching, ATE. If the response, Y is binary then the above formula will estimate the causal risk difference. There are two popular *R* packages called '*Matching*' and '*MacthIt*' to conduct the above discussed (propensity) matching analysis.

Stratification

Although PS matching is the most common approach nowadays, but it often discards a substantial number of individuals from the analysis (Austin, 2014). The PS is a continuous variable that can take any value between 0 and 1. It is therefore unlikely that two individuals will have exactly the same value of the PS. One approach to deal with the continuous PS is to create strata that contain individuals with similar, but may not identical, values of PS. For example, the deciles of the estimated PS: individuals in the population are classified in 10 strata of approximately equal size, then the causal effect is estimated in each of the strata.

Suppose we fit a PS model ($e-$model) and define strata $s = 1, \ldots, S$ by grouping units whose

e_i's are similar. Define $c_{is} = 1$ if unit i belongs to stratum s and 0 otherwise. The $\hat{e}$-stratified estimate of ATE by a weighted average of respondents' mean in each stratum, weighted by the proportion of sample units in that stratum,

$$\hat{\Delta}^{(st)} = \sum_{s=1}^{S} \left(\frac{\sum_i c_{is}}{n} \right) \left(\frac{\sum_i c_{is} T_i Y_i}{\sum_i c_{is} T_i} - \frac{\sum_i c_{is}(1 - T_i) Y_i}{\sum_i c_{is}(1 - T_i)} \right), \tag{1.24}$$

where the summation is taken over the total S strata created from PS, first part after the summation, $\sum_s c_{is}/n$ is the stratum specific weight, and second part is the stratum specific ATE estimate. In the formula of ATE, first portion is the treatment effect among the treated subjects and second portion is the treatment effect among the control subjects. If the outcome response, $\mathbf{Y}$ is continuous then the above formula will estimate stratified version of ATE. For binary response, i.e., death or survival the above formula will estimate the causal risk difference.

In a situation, with PS to 1 for some treated subjects and close to 0 for control subjects, we may find some strata where we have either only treated subjects or only control subjects. This is also an indication of lack of overlap in the distribution of PS among treated and control groups. For those strata we can not estimate the ATE. There are two ways to handle this situation: strata exclusion and trimming tails. First approach is excluding those strata where ATE calculation is not possible. Second approach is removing subjects who have extreme values of the PS. For example, removing control subjects whose PS is less than the minimum in the treated group and removing treated subjects whose PS is greater than the maximum in the control group. Trimming the tails makes the positivity assumption (see Section 1.1.1) more reasonable. In this study, we apply the stratified version of estimators after trimming the tails of the PS distribution.

Propensity Score Regression

PS regression is a regression type estimator after including the estimated propensities as covariates in the model. In this estimator we do not need to include the covariates as the effect of covariates has already been adjusted via propensity estimation. In this scenario, the generalised version of regression model will take form as:

$$g(\mu_i) = \beta_0 + \beta_t T_i + \beta_e e\left(\mathbf{X}_i\right), \tag{1.25}$$

where $\mu_i = E(Y_i|T_i, e(\mathbf{X}_i), \boldsymbol{\beta})$ is the expected outcome with $\boldsymbol{\beta} = (\beta_0, \beta_t, \beta_e)^\top$, $e(\mathbf{X}_i)$ is the PS for ith subject, and $g(.)$ is the link function. The logic behind the PS regression is, we will retain the ignorability assumption with PS alone rather than confounders i.e., $\{\mathbf{Y}(0), \mathbf{Y}(1)\} \perp\!\!\!\perp \mathbf{T}|e(\mathbf{X})$. Using the same terminology as used in regression estimator in Section 1.2.1, we can define the ATE for PS regression estimator for continuous outcome as:

$$\begin{aligned}\Delta_c^{(psr)} &= E\left\{E(\mathbf{Y}|\mathbf{T}=1, e(\mathbf{X})) - E(\mathbf{Y}|\mathbf{T}=0, e(\mathbf{X}))\right\} \\ &= E\left\{(\beta_0 + \beta_t * 1 + \beta_e e(\mathbf{X})) - (\beta_0 + \beta_t * 0 + \beta_e e(\mathbf{X}))\right\} \\ &= E(\beta_t) = \beta_t, \end{aligned} \tag{1.26}$$

again the regression coefficient of the treatment variable β_t is the ATE if the assumed y-model (1.10) is true. In case of binary outcome the estimated ATE is defined by:

$$\hat{\Delta}_b^{(psr)} = \frac{1}{n}\sum_{i=1}^{n}\left\{expit\left(\hat{\beta}_0 + \hat{\beta}_t + \hat{\beta}_e e(\mathbf{X}_i)\right) - expit\left(\hat{\beta}_0 + \hat{\beta}_e e(\mathbf{X}_i)\right)\right\}. \tag{1.27}$$

Inverse Weighting

Inverse weighting via PS is commonly called inverse probability weighting (IPW) in the observational study literature, because that is the form of weighting to estimate the ATE. It is first developed by Horvitz and Thompson (1952) to estimate the mean of a population from a stratified random sample (a survey sampling problem), so it is also called Horvitz-Thompson estimator. IPW is applied to account for different proportions of observations within strata in the target population. If e_i is the inclusion probability of the sample Y_i, the Horvitz-Thompson estimator of population average is given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{e_i}. \tag{1.28}$$

In observational studies, the ATE, $E[\mathbf{Y}(1) - \mathbf{Y}(0)]$, can be viewed as estimating two population means. Therefore, the IPW estimator is given by the difference of two Horvitz-Thompson estimators

$$\hat{\Delta}^{(ipw)} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{e}(\mathbf{X}_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\hat{e}(\mathbf{X}_i)}, \tag{1.29}$$

where $\hat{e}()$ is the estimated PS. For continuous outcome, the above equation will estimate ATE and

for binary outcome this will be the causal risk difference. Using the consistency and exchangeability assumptions, Hernán and Robins (2019) shows that for the treated group

$$\begin{aligned} E\left[\frac{\mathbf{TY}}{e(\mathbf{X})}\right] &= E\left[E\left\{\frac{I(\mathbf{T}=1)\mathbf{Y}(1)}{e(\mathbf{X})}|\mathbf{Y}(1),\mathbf{X}\right\}\right] \\ &= E\left[\frac{\mathbf{Y}(1)}{e(\mathbf{X})}E\left\{I(\mathbf{T}=1)|\mathbf{Y}(1),\mathbf{X}\right\}\right] \\ &= E[\mathbf{Y}(1)] \end{aligned} \tag{1.30}$$

that is $E\left[\frac{\mathbf{TY}}{e(\mathbf{X})}\right]$ correctly estimate $E[\mathbf{Y}(1)]$. Similar conclusion can be made for the control group that, $E\left[\frac{(1-\mathbf{T})\mathbf{Y}}{1-e(\mathbf{X})}\right]$ correctly estimates $E[\mathbf{Y}(0)]$.

Compared to matching methods, IPW is a "cleaner" and more efficient approach, because the discrete matches are replaced by continuous weights. In a key paper, Hirano et al. (2003) showed that IPW paired with sieve PS model can achieve the semiparametric efficiency bound, which gives the theoretical reason to prefer weighting over matching or stratification (see Sections 1.2.2 and 1.2.2 for more detail). However, this also comes with a price. The inverse probability weights are more volatile and sensitive to model misspecification. If some estimated PS $\hat{e}(X_i)$ is close to 0 (for a treated unit) or 1 (for a control unit), its inverse weight can become very large and unstable, so the IPW estimator may perform poorly in finite sample. One way to mitigate this is to normalize the weights within each treatment group (Imbens, 2004). For example, the normalized IPW estimator of ATE is

$$\hat{\Delta}^{ipwn} = \left(\sum_{i=1}^{n}\frac{T_i}{\hat{e}(\mathbf{X}_i)}\right)^{-1}\sum_{i=1}^{n}\frac{T_iY_i}{\hat{e}(\mathbf{X}_i)} - \left(\sum_{i=1}^{n}\frac{1-T_i}{1-\hat{e}(\mathbf{X}_i)}\right)^{-1}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\hat{e}(\mathbf{X}_i)} \tag{1.31}$$

The IPW estimator does not completely solve the instability issue that the estimator could be largely decided by just a few observations. The instability issue was officially brought up by Kang et al. (2007), but it was perhaps well known by practitioners before that. Kang et al. (2007) constructed an artificial example in which inverse probability weights are very unstable and the IPW estimator performs poorly. Moreover, they showed that if we further augment IPW by an outcome regression (see Section 1.2.3), the estimator could perform even worse though it has the theoretical "double robustness" property.

1.2.3 Augmented Inverse Probability Weighting

Traditional regression and PS based estimators will provide unbiased estimate of the treatment effect only when the underlying models are not misspecified. Correct specification of the models of regression and PS are one of the most fundamental assumptions in statistical analysis. Outside of simulation studies, we can never know whether or not the model we have constructed accurately depicts those relationships. Thus, correct specification of the regression model is an unverifiable assumption. Augmented inverse probability weighted (AIPW) estimator has been developed in such a way that the causal effect will be correctly estimated if either of the y- or $e-$ model is correctly specified (Funk et al., 2011). The AIPW estimator does not eliminate the need for correct models but does give the double protection for estimation consistency. That's why, the AIPW estimator is also called as 'double robust' (DR) estimator.

AIPW estimation builds on the inverse probability weighting by a PS and regression modeling of the relationship between covariates and outcome for each treatment level. The AIPW estimator requires to specify regression models for the outcome and the treatment as a function of covariates. In the case of this particular AIPW estimator, we model the relations between confounders and the outcome within each treatment group i.e., $m_1(\mathbf{X}_i, \boldsymbol{\beta}^{(1)})$ and $m_0(\mathbf{X}_i, \boldsymbol{\beta}^{(0)})$ for treated and control groups respectively. The resulting parameter estimates ($\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(0)}$) are used to calculate the predicted response ($m_1(\mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(1)})$ and $m_0(\mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(0)})$) for each individual in the population under the two treatment conditions ($T = 1$ and $T = 0$) given covariate values ($\mathbf{X}$). In addition, we model the exposure as a function of covariates to estimate the PS (or predicted probability of exposure conditional on covariates, $\mathbf{X}$) for each individual using the observed data. Having estimated the PS, $m_1(\mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(1)})$ and $m_0(\mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(0)})$, the AIPW estimator efficiently uses this information in the following manner:

$$
\begin{aligned}
\hat{\Delta}^{(aipw)} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\overbrace{\frac{T_i Y_i}{\hat{e}(\mathbf{X}_i)}}^{IPW} - \overbrace{\frac{\{T_i - \hat{e}(\mathbf{X}_i)\}}{\hat{e}(\mathbf{X}_i)} m_1(\mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(1)})}^{Augmentation}\Bigg] \\
- \frac{1}{n}\sum_{i=1}^{n}\Bigg[\overbrace{\frac{(1-T_i)Y_i}{1-\hat{e}(\mathbf{X}_i)}}^{IPW} + \overbrace{\frac{\{T_i - \hat{e}(\mathbf{X}_i)\}}{1-\hat{e}(\mathbf{X}_i)} m_0(\mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(0)})}^{Augmentation}\Bigg] \\
= \hat{\mu}_{1,aipw} - \hat{\mu}_{0,aipw}. \qquad (1.32)
\end{aligned}
$$

The first line of the above equation is the AIPW estimate of treatment effects among treated subjects. So, $\hat{\mu}_{1,aipw}$ is a function of observed outcomes under treatment (T_iY_i) and predicted outcomes under treatment given covariates ($m_1(\mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(1)})$), weighted by a function of the PS. Similarly, in second line, the estimated value for $\mu_{0,aipw}$ is a function of observed outcomes under control $\{(1-T_i)Y_i\}$ combined with $m_0(\mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(0)})$ after weighting by a function of PS. Finally, the estimates of $\hat{\mu}_{1,aipw}$ and $\hat{\mu}_{0,aipw}$ are used to calculate the ATE for continuous response or causal risk difference for binary response. The AIPW estimator called doubly robust because the effect of the treatment on the outcome will be correctly estimated as long as either the $y-$ or $e-$ model is correctly specified, assuming that there are no unmeasured confounders (Glynn and Quinn, 2010; Funk et al., 2011).

Why double robust?

The AIPW estimator called doubly robust because the effect of the treatment on the outcome will be correctly estimated as long as either the $y-$ or $e-$ model is correctly specified, assuming that there are no unmeasured confounders. Glynn and Quinn (2010) and Funk et al. (2011) proof the double robustness property of the AIPW estimator defined in Equation (1.32). Here we present the proof only for the treated group; similarly it can be shown for the control group. The definition of AIPW estimator in Equation (1.32) requires estimation of two quantities $\mu_{1,aipw}$ (treated) and $\mu_{0,aipw}$ (control). We know for the treated

$$\hat{\mu}_{1,aipw} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_iY_i}{\hat{e}(\mathbf{X}_i)} - \frac{\{T_i - \hat{e}(\mathbf{X}_i)\}}{\hat{e}(\mathbf{X}_i)}m_1(\mathbf{X}_i, \hat{\beta}^{(1)})\right]. \tag{1.33}$$

By the law of large numbers, $\hat{\mu}_{1,aipw}$ estimates the mean of a term in the sum with $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{e}(\mathbf{X})$ replaced by the quantities they estimate $\beta^{(1)}$ and $e(\mathbf{X})$ respectively. In vector notation and expectation

form $\hat{\mu}_{1,aipw}$ estimates the following:

$$
\begin{aligned}
& E\left[\frac{\mathbf{TY}}{e(\mathbf{X})} - \frac{\{\mathbf{T} - e(\mathbf{X})\}}{e(\mathbf{X})} m_1\left(\mathbf{X}, \boldsymbol{\beta}^{(1)}\right)\right] \\
&= E\left[\frac{\mathbf{TY}(1)}{e(\mathbf{X})} - \frac{\{\mathbf{T} - e(\mathbf{X})\}}{e(\mathbf{X})} m_1\left(\mathbf{X}, \boldsymbol{\beta}^{(1)}\right)\right] \\
&= E\left[\mathbf{Y}(1) - \frac{e(\mathbf{X})}{e(\mathbf{X})}\mathbf{Y}(1) + \frac{\mathbf{TY}(1)}{e(\mathbf{X})} - \frac{\{\mathbf{T} - e(\mathbf{X})\}}{e(\mathbf{X})} m_1\left(\mathbf{X}, \boldsymbol{\beta}^{(1)}\right)\right] \\
&= E\left[\mathbf{Y}(1) + \frac{\{\mathbf{T} - e(\mathbf{X})\}}{e(\mathbf{X})}\mathbf{Y}(1) - \frac{\{\mathbf{T} - e(\mathbf{X})\}}{e(\mathbf{X})} m_1\left(\mathbf{X}, \boldsymbol{\beta}^{(1)}\right)\right] \\
&= E\left[\mathbf{Y}(1) + \frac{\{\mathbf{T} - e(\mathbf{X})\}}{e(\mathbf{X})}\left\{\mathbf{Y}(1) - m_1\left(\mathbf{X}, \boldsymbol{\beta}^{(1)}\right)\right\}\right] \\
&= E[\mathbf{Y}(1)] + E\left[\frac{\{\mathbf{T} - e(\mathbf{X})\}}{e(\mathbf{X})}\left\{\mathbf{Y}(1) - m_1\left(\mathbf{X}, \boldsymbol{\beta}^{(1)}\right)\right\}\right], \qquad (1.34)
\end{aligned}
$$

where $e(\mathbf{X})$ is the $e-$ model and $m_1\left(\mathbf{X}; \boldsymbol{\beta}^{(1)}\right)$ is the postulated $y-$model in the treated group. These models may or may not be exactly the true $y-$ or $e-$model. Thus, $\hat{\mu}_{1,DR}$ will correctly estimate $E[\mathbf{Y}(1)]$ only when the second term in above equation is equal to zero. Now consider the following two scenarios:

Scenario 1: Postulated $e-$ model $e(\mathbf{X})$ is correct, but $y-$model $m_1\left(\mathbf{X};\boldsymbol{\beta}^{(1)}\right)$ is incorrect, i.e., $e(\mathbf{X}) = E(\mathbf{T}|\mathbf{X}) \rightarrow E(\mathbf{T}|\mathbf{Y}(1),\mathbf{X})$ by no unmeasured confounders and $m_1\left(\mathbf{X};\boldsymbol{\beta}^{(1)}\right) \neq E(\mathbf{Y}|\mathbf{T}=1,\mathbf{X})$. Under these conditions the second term take the form as:

$$
\begin{aligned}
&E\left[\frac{\{\mathbf{T}-e(\mathbf{X})\}}{e(\mathbf{X})}\left\{\mathbf{Y}(1)-m_1\left(\mathbf{X},\boldsymbol{\beta}^{(1)}\right)\right\}\right] \\
&= E\left(E\left[\frac{\{\mathbf{T}-e(\mathbf{X})\}}{e(\mathbf{X})}\left\{\mathbf{Y}(1)-m_1\left(\mathbf{X},\boldsymbol{\beta}^{(1)}\right)\right\}|\mathbf{Y}(1),\mathbf{X}\right]\right) \\
&= E\left(\left\{\mathbf{Y}(1)-m_1\left(\mathbf{X},\boldsymbol{\beta}^{(1)}\right)\right\}E\left[\frac{\{\mathbf{T}-e(\mathbf{X})\}}{e(\mathbf{X})}|\mathbf{Y}(1),\mathbf{X}\right]\right) \\
&= E\left(\left\{\mathbf{Y}(1)-m_1\left(\mathbf{X},\boldsymbol{\beta}^{(1)}\right)\right\}\frac{\{E(\mathbf{T}|\mathbf{Y}(1),\mathbf{X})-e(\mathbf{X})\}}{e(\mathbf{X})}\right) \\
&= E\left(\left\{\mathbf{Y}(1)-m_1\left(\mathbf{X},\boldsymbol{\beta}^{(1)}\right)\right\}\frac{\{E(\mathbf{T}|\mathbf{X})-e(\mathbf{X})\}}{e(\mathbf{X})}\right) \\
&= E\left(\left\{\mathbf{Y}(1)-m_1\left(\mathbf{X},\boldsymbol{\beta}^{(1)}\right)\right\}\frac{\{e(\mathbf{X})-e(\mathbf{X})\}}{e(\mathbf{X})}\right)=0
\end{aligned}
\tag{1.35}
$$

Scenario 2: Postulated $y-$model $m_1\left(\mathbf{X},\boldsymbol{\beta}^{(1)}\right)$ is correct, but PS model $e(\mathbf{X};\boldsymbol{\alpha})$ is incorrect, i.e., $e(\mathbf{X}) \neq E(\mathbf{T}|\mathbf{X})$ and $m_1\left(\mathbf{X},\boldsymbol{\beta}^{(1)}\right) = E(\mathbf{Y}|\mathbf{T}=1,\mathbf{X}) \rightarrow [E(\mathbf{Y}(1)|\mathbf{X})]$ by no unmeasured confounders. Under these conditions the second term take the form as:

$$
\begin{aligned}
&E\left[\frac{\{\mathbf{T}-e(\mathbf{X})\}}{e(\mathbf{X})}\{\mathbf{Y}(1)-E\left(\mathbf{Y}|\mathbf{T}=1,\mathbf{X}\right))\}\right] \\
&= E\left(E\left[\frac{\{\mathbf{T}-e(\mathbf{X})\}}{e(\mathbf{X})}\{\mathbf{Y}(1)-E(\mathbf{Y}|\mathbf{T}=1,\mathbf{X}))\}\mid\mathbf{T},\mathbf{X}\right]\right) \\
&= E\left(\frac{\{\mathbf{T}-e(\mathbf{X})\}}{e(\mathbf{X})}E[\{\mathbf{Y}(1)-E(\mathbf{Y}|\mathbf{T}=1,\mathbf{X}))\}\mid\mathbf{T},\mathbf{X}]\right) \\
&= E\left(\frac{\{\mathbf{T}-e(\mathbf{X})\}}{e(\mathbf{X})}\{E(\mathbf{Y}(1)|\mathbf{T},\mathbf{X})-E(\mathbf{Y}|\mathbf{T}=1,\mathbf{X}))\}]\right) \\
&= E\left(\frac{\{\mathbf{T}-e(\mathbf{X})\}}{e(\mathbf{X})}\{E(\mathbf{Y}(1)|\mathbf{X})-E(\mathbf{Y}(1)|\mathbf{X}))\}]\right)=0
\end{aligned}
\tag{1.36}
$$

last equality stands for the no unmeasured confounders assumption, which says

$$E(\mathbf{Y}|\mathbf{T}=1,\mathbf{X}) = E(\mathbf{Y}(1)|\mathbf{T}=1,\mathbf{X}) = E(\mathbf{Y}(1)|\mathbf{X}) = E(\mathbf{Y}(1)|\mathbf{T},\mathbf{X}).$$

In both scenarios, we find the second term in Equation (1.33) is equal to zero. That is, as long as

either of the $y-$ or $e-$model is correct, $\hat{\mu}_{1,aipw}$ estimates $E[\mathbf{Y}(1)]$ correctly. Similar conclusion can be made for the control group, i.e., $\hat{\mu}_{0,aipw}$ estimates $E[\mathbf{Y}(0)]$ correctly. Finally, we can conclude that $\hat{\Delta}^{(aipw)}$ consistently estimates the true ATE when either of the model is true. This is obvious from these calculations that if both models are correct we will get consistent estimate of ATE and if both models are incorrect $\hat{\Delta}^{(aipw)}$ will not estimate true ATE correctly.

1.2.4 Variance Estimation

Statistical inference does not only aim to get the point estimate of the parameter of interest but also the uncertainty associated with the estimate. Throughout the simulation study, we use the Markov Chain Monte Carlo (MCMC) algorithm to get the empirical estimate of variation. In real life data application, we use either analytical form of variance or bootstrap algorithm to determine the variability. Abadie and Imbens (2006) and Abadie and Imbens (2008) proposed the variance estimator of matching as

$$\hat{V}^{AI}\left(\hat{\Delta}^{(mc)}\right) = \frac{1}{n_1^2}\sum_{i=1}^{n}\left\{Y_i(1) - Y_i(0) - \hat{\Delta}^{(mc)}\right\}^2 + \frac{1}{2n_1^2}\sum_{i=1}^{n}\left(\frac{1}{K_i} + \frac{1}{K_i^2}\right)(Y_i - Y_{l(i)})^2\mathbf{1}(T_i = 0) \qquad (1.37)$$

where K_i is the number of times observation i used as a match and $Y_{l(i)}$: closest matched observation with same treatment status as i. The *R* package '*Matching*' also provide the variance estimates of the PS matching estimator. In case of inverse probability weighting (IPW) estimator, '*sandwich*' package in *R* can be used to find the variance. Finally, the augmented inverse probability weighted (AIPW) estimator has the analytical form of variance as follows Glynn and Quinn (2010):

$$\hat{V}\left(\hat{\Delta}^{(aipw)}\right) = n^{-2}\sum_{i=1}^{n}\hat{I}_i^2$$

where $\hat{I}_i$ defined as below:

$$\hat{I}_i = \left[\frac{T_iY_i}{e(\mathbf{X}_i)} - \frac{\{T_i - e(\mathbf{X}_i)\}}{e(\mathbf{X}_i)}m_1\left(\mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(1)}\right)\right] - \left[\frac{(1-T_i)Y_i}{1-e(\mathbf{X}_i)} + \frac{\{T_i - e(\mathbf{X}_i)\}}{1-e(\mathbf{X}_i)}m_0\left(\mathbf{X}_i, \hat{\boldsymbol{\beta}}^{(0)}\right)\right] - \hat{\Delta}^{(aipw)}.$$

For the other causal estimators, namely regression, G-estimation, PS stratification and regression, there is no analytical form of variance or available R package. One possibility is to use statistical

theory to derive the corresponding variance estimator, which is not available in standard statistical software. A second possibility is to approximate the variance by nonparametric bootstrapping (Efron and Tibshirani, 1994; Davison and Hinkley, 1997). Many researcher in causal inference used the bootstrapping to get the variation estimate of different causal estimators (Bang and Robins, 2005; Williamson et al., 2012; Zhang et al., 2016). The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method (James et al., 2013). The bootstrap algorithm is a re-sampling method with replacement for deriving robust estimates of that uncertainty. It works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of interested parameter by the empirical standard error, denoted by $\widehat{se}_B$, where B is the number of bootstrap samples used. Each of the B bootstrap sample is the same size as the original dataset i.e., n. As a result, some samples will be represented multiple times in the bootstrap sample while others will not be selected at all. Let us consider the original data $\mathbf{D} = \{\mathbf{Y}, \mathbf{T}, \mathbf{X}\}$, then the bootstrap algorithm (Efron and Tibshirani, 1994; Davison and Hinkley, 1997) is implemented as follows:

1. Select B independent bootstrap samples $\mathbf{D}_1^*, \mathbf{D}_2^*, \ldots, \mathbf{D}_B^*$, each consisting of size n drawing with replacement from $\mathbf{D}$.

2. Calculate the ATE corresponding to each bootstrap sample

$$\widehat{ATE}_k^*(b) = f_k(\mathbf{D}_b^*); \quad b = 1, 2, \ldots, B \text{ and } k = \text{estimator type.}$$

3. Estimate the standard error of the causal effect estimator by the sample standard error of the B replicates

$$se_B(k) = \left[\frac{1}{B-1} \sum_{b=1}^{B} \left\{ \widehat{ATE}_k^*(b) - \widehat{ATE}_k^*(.) \right\}^2 \right],$$
$$\text{with } \widehat{ATE}_k^*(.) = \frac{1}{B} \sum_{b=1}^{B} \widehat{ATE}_k^*(b).$$

Here, our parameter of interest is the ATE, k stands for different estimator types (regression, G-estimation, stratification, PS regression etc.), and $f_k(.)$ is the functional form of ATE for k-type

estimator.

The bootstrap is a method of doing inference in a way that does not require assuming a parametric form for the population distribution. It does not treat the original sample as the population even those it involves sampling with replacement from the original sample. It assumes that sampling with replacement from the original sample of size n mimics taking a sample of size n from a larger population. There are some situations where the bootstrap can fail. This includes distribution that do not have finite moments, small sample sizes, estimating extreme values from the distribution and estimating variance in survey sampling (Chernick, 2011). In some cases variants of the bootstrap namely standard, bias corrected, bootstrap-t and percentile intervals can work better (DiCiccio and Efron, 1996).

A popular alternative to bootstrap is jackknife, which is based upon sequentially deleting one observation from the dataset, recomputing the estimator (Efron and Stein, 1981). The jackknife method is more conservative than the bootstrap method, that is, its estimated standard error tends to be slightly larger.

1.3 Introduction of Missing Data

Standard statistical methods have been developed to analyse rectangular data sets. Traditionally, the rows of the data matrix represent units, also called cases, observations, or subjects depending on context, and the columns represent variables measured for each unit. The entries in the data matrix are nearly always real numbers, either representing the values of essentially continuous variables or categories of response.

1.3.1 Missing Mechanisms

Missing data mechanisms are crucial since the properties of missing data methods depend very strongly on the nature of the dependencies in these mechanisms. Rubin (1976) discussed the simple device of treating the missing-data indicators as random variables and assigning them a distribution. Define the complete data $Y = (y_{ij})$ and the missing data indicator matrix $\mathbf{M} = (M_{ij})$. The missing data mechanism is characterised by the conditional distribution of $\mathbf{M}$ given $\mathbf{Y}$. If

missingness does not depend on the values of the data $\mathbf{Y}$, missing or observed, that is, if

$$f(\mathbf{M}|\mathbf{Y}, \phi) = f(\mathbf{M}|\phi) \text{ for all } \mathbf{Y}, \phi, \tag{1.38}$$

the data are called *missing completely at random* (MCAR). Note that the assumption does not mean that the pattern itself is random. Let $\mathbf{Y}_{obs}$ and $\mathbf{Y}_{mis}$ denote the observed and missing components of $\mathbf{Y}$ respectively. An assumption less restrictive than MCAR is that missingness depends only on the components $\mathbf{Y}_{obs}$ and not on $\mathbf{Y}_{mis}$. That is,

$$f(\mathbf{M}|\mathbf{Y}, \phi) = f(\mathbf{M}|\mathbf{Y}_{obs}, \phi) \text{ for all } \mathbf{Y}_{mis}, \phi. \tag{1.39}$$

The missing data mechanism is then called missing at random (MAR). The mechanism is called not missing at random (NMAR) if the distribution of $\mathbf{M}$ depends on the missing values ($\mathbf{Y}_{mis}$) in the data matrix $\mathbf{Y}$. That is

$$f(\mathbf{M}|\mathbf{Y}, \phi) = f(\mathbf{M}|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}\phi) \text{ for all } \mathbf{Y}, \phi. \tag{1.40}$$

1.3.2 Methods to Handle the Missing Data

Complete-Case Analysis

Complete case analysis is a method in which only subjects with all values recorded for all covariates are retained in the analysis (Little, 1992). A researcher using complete cases assumes that the observed complete cases are a random sample of the originally targeted sample, or in Rubin (1976) terminology, that the missing data are MCAR. When a data set has only a few missing observations, the assumption of MCAR data is more likely to apply; there is a greater chance of the complete cases representing the population when only a few cases are missing. This common and easiest approach obviously will reduce the estimation efficiency when the missingness level is high as records with missing data in any single variable are drop. It can be biased (Vach and Blettner, 1991; Little, 1992). Even when complete-subject analysis is valid, it can be very inefficient-estimates with higher variance (Rubin, 2004). The main advantage of the method is ease of implementation since the researcher can use standard methods for computing estimates for

a proposed model. One disadvantage of the method centers on the number of cases that observe all variables of interest in the data; a researcher cannot anticipate if an adequate amount of data remain for the analysis.

Multiple Imputation

These methods begin by generating multiple copies of the original data set, each with missing values replaced by values randomly generated according to some model for the distribution of incomplete regressors and its dependence on complete regressors and the outcome variable. Each of these imputed data sets is analysed as if it were complete; the different results from the data sets are then combined in a manner that takes account of the imputation variability (Heitjan and Little, 1991; Rubin and Schenker, 1991).

The principle of multiple imputation (MI) is to generate multiple sets of plausible values for the missing variables by drawing from the posterior predictive distribution of these variables given the observed data. Variables of different types are often included in the $y-$ and $e-$models. Therefore, we focused on chained equations, in which a specific imputation model is specified for each partially observed variable, rather than joint modelling to impute the missing data (Azur et al., 2011). M complete datasets are created and analysed independently to produce estimates $\hat{\boldsymbol{\theta}}_k$, $(k = 1, 2, \ldots, M)$ of $\boldsymbol{\theta}$ the vector of the parameters of interest (e.g. regression coefficients) and estimates $\mathbf{W}_k$ of their associated variance matrix. Then $\hat{\boldsymbol{\theta}}_k$ and $\mathbf{W}_k$, $(k = 1, 2, \ldots, M)$ are combined across the M imputed datasets. Rubin's rules for the mean and variance state that an overall estimate, $\hat{\boldsymbol{\theta}}_{MI}$, of $\boldsymbol{\theta}$ and an estimate of the variance of $\hat{\boldsymbol{\theta}}_{MI}$, $\widehat{Var}\left(\hat{\boldsymbol{\theta}}_{MI}\right)$, are (Carpenter and Kenward, 2012):

$$\hat{\boldsymbol{\theta}}_{MI} = \frac{1}{M}\sum_{k=1}^{M}\hat{\boldsymbol{\theta}}_k, \quad \widehat{Var}\left(\hat{\boldsymbol{\theta}}_{MI}\right) = \mathbf{W} + \left(1 + \frac{1}{M}\right)\mathbf{B}, \tag{1.41}$$

where $\mathbf{W}$ is the within-imputation variance–covariance matrix, which reflects the variability of the parameter estimates in each imputed dataset, and $\mathbf{B}$ is the between-imputation variance matrix reflecting the variability in the estimates caused by the missing information. These two components are defined as:

$$\mathbf{W} = \frac{1}{M}\sum_{k=1}^{M}\mathbf{W}_k; \quad \mathbf{B} = \frac{1}{(1-M)}\sum_{k=1}^{M}\left(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{MI}\right)\left(\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{MI}\right)^{\top}. \tag{1.42}$$

The Expectation-Maximization Algorithm

The expectation-maximization (EM) algorithm is a very general iterative algorithm for maximum-likelihood (ML) estimation in incomplete data problems. The EM algorithm formalizes a relatively old ad hoc idea for handling missing data: (1) Replace missing values by estimated values, (2) estimate parameters, (3) re-estimate the missing values assuming the new parameter estimates are correct, (4) re-estimate parameters, and so forth, iterating until convergence (Boos and Stefanski, 2013). Each iteration of EM consists of an E step (expectation step) and an M step (maximization step). The basic idea of the EM Algorithm is to view the observed data $\mathbf{Y}$ as incomplete, that somehow there is missing data $\mathbf{Z}$ that would make the problem simpler if we had it. In some cases $\mathbf{Z}$ could truly be missing data, but in others it is just additional data that we wish we had. The first step is to write down the joint likelihood of the "complete" data ($\mathbf{Y}$, $\mathbf{Z}$), call it $L_c(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})$. We always need to maximize an objective function that depends only on $\boldsymbol{\theta}$ and the observed data $\mathbf{Y}$.

The E step finds the conditional expectation of the "missing data" given the observed data and current estimated parameters, and then substitutes these expectations for the "missing data". The term "missing data" is written with quotation marks because EM does not necessarily substitute the missing values themselves. The key idea of EM, which delineates it from the ad hoc idea of filling in missing values and iterating, is that "missing data" are not $\mathbf{Z}$ but the functions of $\mathbf{Z}$ appearing in the complete-data log-likelihood $logL_c(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})$ (Little and Rubin, 2019).

Thus, the "E" step of the EM Algorithm is to compute the conditional expectation of $logL_c(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})$ given $\mathbf{Y}$ assuming the true parameter value is $\boldsymbol{\theta}^{(\nu)}$. Define

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\nu)}, \mathbf{Y}) &= E_{\boldsymbol{\theta}^{(\nu)}} \{logL_c(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z})|\mathbf{Y}\} \\ &= \int logL_c(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{z})\ f_{Z|Y}(z|\mathbf{Y}, \boldsymbol{\theta}^{(\nu)})dz, \end{aligned}$$

where we have written the expectation in the second line as if $\mathbf{Z}|\mathbf{Y}$ has a continuous density, but this is for notational convenience only. The "M" step of the EM Algorithm is to maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\nu)}, \mathbf{Y})$ with respect to $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^{(\nu)}$ fixed. This produces a new value $\boldsymbol{\theta}^{(\nu+1)}$ that is then reinserted in the "E" step and so on. The EM algorithm can be implemented using the following simple steps

Step 0: Initialization: Set an initial value for parameters, say $\boldsymbol{\theta}^0$. In many cases, this can just be a

random initialization or the estimates obtained from the complete cases.

Step 1: Expectation Step: Assume the parameters $\boldsymbol{\theta}^{(\nu)}$ from the previous step are fixed, compute the expected values of the missing variables or more often a function of the expected values of the missing variables i.e., calculate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\nu)}, \mathbf{Y})$.

Step 2: Maximization Step: Given the values computed in the last step (essentially known values for the missing variables), estimate new values for $\boldsymbol{\theta}^{(\nu+1)}$ that maximize a variant of the likelihood function i.e., $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\nu)}, \mathbf{Y})$.

Step 2: Stopping rule: If likelihood of the observations have not changed much or the difference $\left(\boldsymbol{\theta}^{(\nu+1)} - \boldsymbol{\theta}^{(\nu)}\right)$ remain with a certain limit, stop; otherwise, go back to Step 1.

In a missing data model, an estimator is doubly robust (DR) or doubly protected if it remains consistent when either a model for the missingness mechanism or a model for the distribution of the complete data is correctly specified (Bang and Robins, 2005; Rotnitzky et al., 2012; Seaman and Vansteelandt, 2018). In a causal inference model, an estimator is DR if it remains consistent when either a model for the treatment assignment mechanism or a model for counterfactual data is correctly specified (Glynn and Quinn, 2010; Funk et al., 2011). In practice, double robustness does not provide sufficient protection for estimation consistency, as it allows only one model for the PS and one for the outcome regression. Han and Wang (2013) and Han (2014) developed multiple robust estimator in causal estimation by fitting multiple models for treatment and counterfactual model. Such multiple models increase the likelihood of correct specification. They termed this properties as 'multiple robustness'. In dealing with missing data with EM algorithm, these double and multiple robustness of estimators can also be considered.

The EM algorithm allows us to estimate $\hat{\boldsymbol{\theta}}$, but it does not directly provide an estimate of variance. There are three different alternatives to get the estimate of variance, Louis' method (Louis, 1982), Supplemented EM (SEM) algorithm (Meng and Rubin, 1991) and the bootstrapping.

Louis' Method

The conditional density of missing data $\mathbf{Z}$ given the observed data $\mathbf{Y}$ can be written as

$$f(\mathbf{Z}|\mathbf{Y};\boldsymbol{\theta}) = \frac{f(\mathbf{D};\boldsymbol{\theta})}{f(\mathbf{Y};\boldsymbol{\theta})}, \tag{1.43}$$

Where $\mathbf{D} = (\mathbf{Y}, \mathbf{Z})$ is the complete data. Now taking log on both side of the above equation we get

$$\begin{aligned} \log f(\mathbf{Z}|\mathbf{Y};\boldsymbol{\theta}) &= \log f(\mathbf{D};\boldsymbol{\theta}) - \log f(\mathbf{Y};\boldsymbol{\theta}) \\ \log f(\mathbf{Z}|\mathbf{Y};\boldsymbol{\theta}) &= l_c(\boldsymbol{\theta}) - l(\boldsymbol{\theta}) \\ l(\boldsymbol{\theta}) &= l_c(\boldsymbol{\theta}) - \log f(\mathbf{Z}|\mathbf{Y};\boldsymbol{\theta}), \end{aligned} \tag{1.44}$$

where $l(\boldsymbol{\theta})$ and $l_c(\boldsymbol{\theta})$ are the observed and complete data log-likelihood respectively. Differentiating twice and negating both sides, then taking expectations over the conditional distribution of $\mathbf{D}$ given $\mathbf{Y}$,

$$\underbrace{-l''(\boldsymbol{\theta})}_{\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Y}}(\boldsymbol{\theta})} = \underbrace{E\left[-l_c''(\boldsymbol{\theta})|\mathbf{Y}\right]}_{\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{D}}(\boldsymbol{\theta})} - \underbrace{E\left[-\frac{\partial^2 \log f(\mathbf{Z}|\mathbf{Y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}|\mathbf{Y}\right]}_{\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})}, \tag{1.45}$$

where $\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Y}}(\boldsymbol{\theta})$ is the observed information, $\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{D}}(\boldsymbol{\theta})$ is the complete data information and $\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ is the missing information. Computing $\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{D}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ is sometimes easier than computing $-l''(\boldsymbol{\theta})$ directly. It can be shown that

$$\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta}) = \mathrm{Var}\left[S_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})\right], \tag{1.46}$$

where the variance is taken with respect to $\mathbf{Z}|\mathbf{Y}$ and

$$S_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta}) = \frac{\partial \log f(\mathbf{Z}|\mathbf{Y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}, \tag{1.47}$$

is the conditional score. As the expected score is zero at $\hat{\boldsymbol{\theta}}$, we have

$$\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Z}|\mathbf{Y}}(\hat{\boldsymbol{\theta}}) = \int S_{\mathbf{Z}|\mathbf{Y}}(\hat{\boldsymbol{\theta}}) S_{\mathbf{Z}|\mathbf{Y}}(\hat{\boldsymbol{\theta}})^\top \log f(\mathbf{Z}|\mathbf{Y};\hat{\boldsymbol{\theta}}) d\mathbf{z}. \tag{1.48}$$

When $\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{D}}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ cannot be computed analytically, they can be approximated by Monte Carlo simulation. First, generate simulate datasets $\mathbf{D}_j = (\mathbf{Y}, \mathbf{Z}_j)$, $j = 1, 2, \ldots, N$, where $\mathbf{Y}$ is

the observed dataset, and the $\mathbf{Z}_j$ are imputed missing datasets drawn from $f(\mathbf{Z}_j|\mathbf{Y};\boldsymbol{\theta})$. Then the approximation can be done as

$$\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{D}}(\boldsymbol{\theta}) \approx \frac{1}{N}\sum_{j=1}^{N} -\frac{\partial^2 \log f(\mathbf{D}_j;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}, \tag{1.49}$$

and $\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ is approximated by the sample variance of the values

$$\frac{\partial \log f(\mathbf{Z}_j|\mathbf{Y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}.$$

Substituting these values in Equation (1.45), we will get the observed information $\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Y}}(\boldsymbol{\theta})$. The diagonal elements of inverse of observed information matrix are the variance of parameter estimates.

The SEM Algorithm

Let $\boldsymbol{\Psi}$ denotes the EM mapping, defined by $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\Psi}\left(\boldsymbol{\theta}^{(t)}\right)$. From the convergence of EM algorithm, it is obvious that $\hat{\theta}$ is a fixed point i.e., $\hat{\theta} = \boldsymbol{\Psi}(\hat{\boldsymbol{\theta}})$. The Jacobian matrix of $\boldsymbol{\Psi}$ is the $(p \times p)$ matrix given by

$$\boldsymbol{\Psi}'(\boldsymbol{\theta}) = \left(\frac{\partial \Psi_i(\boldsymbol{\theta})}{\partial\theta_j}\right), \tag{1.50}$$

where p is the dimension of parameter vector. It can be shown that

$$\boldsymbol{\Psi}'\left(\hat{\boldsymbol{\theta}}\right) = \hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Z}|\mathbf{Y}(\boldsymbol{\theta})}\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{D}}(\hat{\boldsymbol{\theta}})^{-1}. \tag{1.51}$$

From the missing information principal in Equation (1.45), we know that

$$\begin{aligned}
\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Y}}(\hat{\boldsymbol{\theta}}) &= \hat{\boldsymbol{\mathcal{I}}}_{\mathbf{D}}(\hat{\boldsymbol{\theta}}) - \hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Z}|\mathbf{Y}}(\hat{\boldsymbol{\theta}}) \\
&= \left[\mathbf{I} - \hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Z}|\mathbf{Y}(\boldsymbol{\theta})}\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{D}}(\hat{\boldsymbol{\theta}})^{-1}\right]\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{D}}(\hat{\boldsymbol{\theta}}) \\
&= \left[\mathbf{I} - \boldsymbol{\Psi}'\left(\hat{\boldsymbol{\theta}}\right)^\top\right]\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{D}}(\hat{\boldsymbol{\theta}}),
\end{aligned} \tag{1.52}$$

where $\mathbf{I}$ is a $p-$dimensional identity matrix. Hence, $\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Y}}(\hat{\boldsymbol{\theta}})^{-1} = \hat{\boldsymbol{\mathcal{I}}}_{\mathbf{D}}(\hat{\boldsymbol{\theta}})^{-1}\left[\mathbf{I} - \boldsymbol{\Psi}'\left(\hat{\boldsymbol{\theta}}\right)^{\top}\right]^{-1}$. By the inequality, $(\mathbf{I}-\mathbf{P})^{-1} = (\mathbf{I}-\mathbf{P}+\mathbf{P})(\mathbf{I}-\mathbf{P})^{-1} = \mathbf{I}+\mathbf{P}(\mathbf{I}-\mathbf{P})^{-1}$ we can write

$$\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Y}}(\hat{\boldsymbol{\theta}})^{-1} = \hat{\boldsymbol{\mathcal{I}}}_{\mathbf{D}}(\hat{\boldsymbol{\theta}})^{-1}\left\{\mathbf{I} + \boldsymbol{\Psi}'\left(\hat{\boldsymbol{\theta}}\right)^{\top}\left[\mathbf{I} - \boldsymbol{\Psi}'\left(\hat{\boldsymbol{\theta}}\right)^{\top}\right]^{-1}\right\}. \tag{1.53}$$

Let r_{ij} be the element of $\boldsymbol{\Psi}'(\boldsymbol{\theta})$. By definition,

$$\begin{aligned} r_{ij} &= \frac{\partial \Psi_i\left(\hat{\boldsymbol{\theta}}\right)}{\partial \theta_j} \\ &= \lim_{\theta_j \to \hat{\theta}_j} \frac{\Psi_i\left(\hat{\theta}_1, \ldots, \hat{\theta}_{j-1}, \theta_j, \hat{\theta}_{j+1}, \ldots, \hat{\theta}_p\right) - \Psi_i\left(\hat{\boldsymbol{\theta}}\right)}{\theta_j - \hat{\theta}_j} \\ &= \lim_{t\to\infty} \frac{\Psi_i\left(\hat{\boldsymbol{\theta}}^{(t)}(j)\right) - \Psi_i\left(\hat{\boldsymbol{\theta}}\right)}{\theta_j^{(t)} - \hat{\theta}_j} = \lim_{t\to\infty} r_{ij}^{(t)}, \end{aligned} \tag{1.54}$$

where $\hat{\boldsymbol{\theta}}^{(t)}(j) = \left(\hat{\theta}_1, \ldots, \hat{\theta}_{j-1}, \theta_j^{(t)}, \hat{\theta}_{j+1}, \ldots, \hat{\theta}_p\right)$ and $\left(\theta_j^{(t)}\right)$, $t = 1, 2, \ldots$ is a sequence of values converging to $\hat{\theta}_j$. Compute the $r_{ij}^{(t)}$, $t = 1, 2, \ldots$ until they stabilize to some values. Then compute $\hat{\boldsymbol{\mathcal{I}}}_{\mathbf{Y}}(\hat{\boldsymbol{\theta}})^{-1}$ using Equation (1.53). Finally the SEM algorithm can be summarized with following steps:

Algorithm:

1. Run the EM algorithm to convergence, finding $\hat{\boldsymbol{\theta}}$.

2. Restart the algorithm from some $\boldsymbol{\theta}^{(0)}$ near $\hat{\boldsymbol{\theta}}$. For $t = 0, 1, 2, \ldots$

   - Take a standard E step and M step to produce $\boldsymbol{\theta}^{(t+1)}$ from $\boldsymbol{\theta}^{(t)}$.
   - For $j = 1, 2, \ldots, p$
     * Define $\boldsymbol{\theta}^{(t)}(j) = \left(\hat{\theta}_1, \ldots, \hat{\theta}_{j-1}, \theta_j^{(t)}, \hat{\theta}_{j+1}, \ldots, \hat{\theta}_p\right)$, and treating it as the current estimate of $\boldsymbol{\theta}$, run one iteration of EM to obtain $\boldsymbol{\Psi}\left(\boldsymbol{\theta}^{(t)}(j)\right)$.
     * Obtain the ratio

$$r_{ij}^{(t)} = \frac{\Psi_i\left(\hat{\boldsymbol{\theta}}^{(t)}(j)\right) - \hat{\theta}_i}{\theta_j^{(t)} - \hat{\theta}_j},$$

     for $i = 1, 2, \ldots, p$ and $\boldsymbol{\Psi}\left(\hat{\boldsymbol{\theta}}\right) = \hat{\boldsymbol{\theta}}$.

- Stop when all $r_{ij}^{(t)}$ have converged.

3. The (i, j)th element of $\boldsymbol{\Psi}'\left(\hat{\boldsymbol{\theta}}\right)$ equals $lim_{t\to\infty} r_{ij}^{(t)}$. Use the final estimate of $\boldsymbol{\Psi}'\left(\hat{\boldsymbol{\theta}}\right)$ to get the variance.

Bootstrapping

The variance of EM estimators can be obtained form the following bootstrapping steps:

1. Calculate $\hat{\boldsymbol{\theta}}_{EM}$ using EM approach applied to $\mathfrak{D} = (\mathfrak{D}_1, \mathfrak{D}_2, \ldots, \mathfrak{D}_n)$ with $\mathfrak{D}_i = (\mathbf{Y}_i, \mathbf{Z}_i)$. Let $j = 1$ and set $\hat{\boldsymbol{\theta}}_j^* = \hat{\boldsymbol{\theta}}_{EM}$.

2. Sample pseudo-data $\mathfrak{D}_j^* = \left(\mathfrak{D}_{j1}^*, \mathfrak{D}_{j2}^*, \ldots, \mathfrak{D}_{jn}^*\right)$ at random from $\mathfrak{D} = (\mathfrak{D}_1, \mathfrak{D}_2, \ldots, \mathfrak{D}_n)$ with replacement.

3. Calculate $\hat{\boldsymbol{\theta}}_j^*$ by applying the same EM approach to the pseudo-data $\mathfrak{D}_j^*$.

4. Stop if $j = B$ (typically, $B \geq 1000$); otherwise return to step 2.

The collection of parameter estimates $\hat{\boldsymbol{\theta}}_1^*, \hat{\boldsymbol{\theta}}_2^*, \ldots, \hat{\boldsymbol{\theta}}_B^*$ can be used to estimat the variance of $\hat{\boldsymbol{\theta}}$,

$$\widehat{Var}\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{B}\sum_{j=1}^{B}\left(\hat{\boldsymbol{\theta}}_j^* - \overline{\hat{\boldsymbol{\theta}}}^*\right)\left(\hat{\boldsymbol{\theta}}_j^* - \overline{\hat{\boldsymbol{\theta}}}^*\right)^\top, \tag{1.55}$$

where $\overline{\hat{\boldsymbol{\theta}}}^*$ is the sample mean of $\hat{\boldsymbol{\theta}}_1^*, \hat{\boldsymbol{\theta}}_2^*, \ldots, \hat{\boldsymbol{\theta}}_B^*$.

1.4 Literature Review of Missing Data in Causal Inference

Causal estimation from observational study is more challenging in presence of missing data. Missing data is a common problem in public health and medical research. This problem can be discussed from different angles, the mechanism that leads to missing data and the variable which is missing namely, outcome, treatment or confounder etc. Little and Rubin (2019) discussed three missing mechanisms namely, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Statistical methodologies have already been developed under the MCAR and MAR mechanisms, but little is known for the MNAR mechanism (Lu and

Ashmead, 2018). If the mechanism is MCAR, then we can restrict the causal analysis to the complete cases (CC) and still obtain valid estimates. Under missing at random (MAR) assumption, the probability of missingness depends only on the observed data, several methods has been developed using CC analysis, inverse probability weighting (Li et al., 2013; Seaman and White, 2013) and multiple imputation (MI) techniques (Mitra and Reiter, 2016; Leyrat et al., 2019).

White and Carlin (2010) compared MAR-based MI with CC analysis for missing confounders and found no universally consistent advantage of one method over the other. Mitra and Reiter (2016) and Leyrat et al. (2019) proposed three different MI approaches for missing confounders and compared within each other. However, MI approach utilizes the full information after imputing the missing values, where the validity of the estimation depends on appropriate imputation model and number of imputations. Both CC and inverse probability weighting approaches are using only the complete data either directly or after weighting the complete cases and excluding all other incomplete observations from the analysis. These two methods are often being criticized for being inefficient because most of the information from the incomplete observations is not used.

In the areas of missing data and causal inference, there is great interest in doubly robust (DR) estimators that account for both confounding and missing data (Bang and Robins, 2005; Carpenter et al., 2006). Bang and Robins (2005) discussed the DR estimation for missing outcome data using expectation-maximization (EM) algorithm after weighting complete cases by inverse of the probability of observing complete data. Kang et al. (2007) discussed different ways of constructing DR estimators in presence of outcome missing and further improved by Rotnitzky et al. (2012). Williamson et al. (2012) presented DR estimators in general situation where either of the outcome or treatment or confounder is missing. In case of confounder missing, they applied a simple algorithm involving MI approach. In this study, we will consider the expected-maximization (EM) algorithm for different causal estimators in case of confounder missing. The flexibility of our method is that we will not lose any information as it uses all observations rather the complete-cases directly or after weighting.

1.5 Introduction of Stimulating Study on Breast Cancer

Breast cancer continues to be the most frequent cancer type and the second leading cause of cancer death in females in Western Europe and North America. Although there is a significant development in diagnostic tools and therapeutic interventions in breast cancer over the last decade, breast oncologists still face paramount challenges in hormone-negative breast cancer, especially triple-negative breast cancer, the most aggressive breast cancer subtype, for which we do not yet have any molecular-targeted treatments apart from cytotoxic chemotherapy. Several large studies have identified RNA expression signatures that are prognostic for disease recurrence and metastasis early-stage hormone-positive breast cancer (HPBC) (Paik et al., 2004; Gnant et al., 2013; Mook et al., 2009). No such biomarkers exist currently for the more aggressive hormone-negative breast cancer (HNBC). In addition, mounting evidence indicates that levels of proteins that are immediately relevant to cell growth and metabolism are often not particularly well correlated to mRNA levels (Wu et al., 2013), suggesting that it may be more informative to directly measure protein levels as a indicator of biological behavior. These biomarkers likely reflect genetic signatures of early cancer formation and evolution of cancer aggressiveness.

The incidence of early stage breast cancer (stage 1-3) has been increasing over the last 20 years largely due to the introduction of nation-wide mammogram screening program (Canadian Cancer Statistics 2015). At the time of the collection of these tissues (and excluding the 20% of in-situ cases) there would likely have been around 50% of patients with stags 1, with around 40% with stage 2 and 3, the remainder being stage 4 or unknown. Although it is interesting that we do not have population data readily available for treatments given, approximately 60% of the patients with early stage breast cancer would have been given only adjuvant endocrine therapy.

Early stage hormone receptor-positive breast cancer (ES-HPBC) has a relatively good prognosis; it is, however, molecularly diverse and distinctive for late relapse. Consequently, a certain subset of patients with ES-HPBC has a poor prognosis. In this study we use a data subsampled from a clinical cohort of breast cancer patients considered by Feng et al. (2016). We considered only early stage and triple-negative breast cancer patients in our analysis.

1.6 Outline of the Thesis

Missing data is a common problem in epidemiological and clinical studies involving large numbers of individuals and large numbers of variables. Traditionally, complete case (CC) analysis and multiple imputation (MI) methods are used to handle the missing data. However, CC estimates are only unbiased when the data are missing completely at random (MCAR). MI approach is often criticized for the uncertainty of appropriate imputation model and the number of imputations. In this thesis, the outcome is two fold:

1. propose an EM algorithm to estimate the expected values of missing confounders and utilize weighting approach in the effect estimation under different causal estimators and

2. compare the performance of proposed EM method with available alternatives namely, CC analysis and MI method.

The rest of the report is organized as follows. In Chapter 2, we discuss the causal estimation in presence of confounder missing. In Chapter 3, we conduct the simulation studies to examine the performance of different missing data methods. We consider both continuous and binary outcome to estimate the ATE. Real life data application is shown in Chapter 4. Our goal is to estimate the effect of adjuvant radiation treatment on the ten years survival status of breast cancer patients. In Chapter 5 we draw some concluding remarks on causal estimation in presence of confounder missing. It contains the summary and conclusions on the performance of the method proposed. We also provide some directions for further works.

Chapter 2

Treatment Effect Estimation with Missing Confounder

Random experiment with random treatment assignment, will provide a natural balance in the distribution of confounders among treated and control groups. In practice, we can not always conduct a random experiment due to ethical issues or time restriction. The alternative way to perform causal estimation is observational studies, where we do not have any control over the study design i.e., treatment assignment. We can only observe the things that actually happened in reality. As a consequence, there is a high chance to get encountered with the problem of missing. Missing may be occurred with the outcome or treatment or confounders or with any combination of these variables. But the most probable scenario is missing in the comfounders, as the dimension of the confounders is usually high. In this Chapter we will discuss two different aspects of missing, the mechanism of missingness and how to handle the missingness in confounders. Missing mechanism implies the process that lead to missing data. In the last part of this chapter, we will discuss different estimation procedures in presence of missing in confounders as a way of handling the missingness. The failure to handle missing data appropriately may lead to inefficient or even invalid use of available data sources.

2.1 Missing Confounder Mechanisms

In missing data problem, it is important to know the mechanism that lead to missing data. Discussions of missing data frequently concentrate on three types of missing mechanism: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR; sometimes NMAR) (Little and Rubin, 2019). Suppose Y is the response, T is the binary treatment variable, Z is the set of confounder that is partially observed, $\mathbf{X} = (X_1, \ldots, X_p)^\top$ denoting the confounder vector that is always precisely observed and M indicating the missingness of Z, i.e., $M = 1$ if Z is missing, 0 if Z is observed. The $(Z, \mathbf{X})$ is the set of confounders sufficient to ensure conditional exchangeability between the treated and control. The data are said to be MCAR if the missingness of Z is independent of all variables of interest (Heitjan and Basu, 1996), symbolically

$$P(M|Y, T, Z, \mathbf{X}) = P(M), \tag{2.1}$$

which implies the distribution of M is independent of all variables of interest. Under MCAR assumption, observed data can be considered as a random sample from the full data, and thus the only effect of missing data is to reduce sample size but introduce no bias (White and Carlin, 2010; Li et al., 2013). An assumption less restrictive than MCAR is that missingness depends only on the components that are always observed and not on missing confounder Z itself. That is,

$$P(M|Y, T, Z, \mathbf{X}) = P(M|Y, T, \mathbf{X}). \tag{2.2}$$

The missing data mechanism is then called *missing at random* (MAR). MAR data are those in which the missingness of Z i.e., the distribution of M depends only on fully observed but not partially observed or unobserved variables (Rubin, 1976; Westreich et al., 2015). Finally, MNAR data are those in which missingness is related to partially observed or unobserved variables of interest; in this case, where the missingness of Z depends on the missing confounder itself. Little and Rubin (2019) denote the necessary condition for MNAR mechanism as

$$P(M|Y, T, Z, \mathbf{X}) = P(M|Z), \tag{2.3}$$

Among the three missing mechanisms MCAR requires strongest assumption and MNAR requires least restriction. Standard statistical methodologies can be implemented, directly when the missing mechanism is MCAR, and under certain assumptions and modifications for MAR. But for MNAR data the standard methodologies usually become invalid and little is known about the inference (White and Carlin, 2010; Lu and Ashmead, 2018).

2.2 Methods for Dealing Missing Confounder

When the missing mechanism is clearly identified, it is important to apply appropriate methods to deal with the missing. The literature on the analysis of partially missing data is comparatively recent (Afifi and Elashoff, 1966; Little, 1997). Methods proposed in this literature can be usefully grouped into the following categories, which are not mutually exclusive:

2.2.1 Complete Case analysis

The simplest method to handle missing covariate data is to omit subjects with any missing data from the analysis i.e., complete case (CC) analysis. CC analysis confines attention to cases where all the variables are present. CC analysis is conducted on the subset of full data where no confounder has missing values i.e., sub-setting data with $M = 0$. Advantages of this approach are (1) simplicity, since standard complete data statistical analyses can be applied without modifications i.e., we can directly get the ATE estimate from the complete data, and (2) comparability of univariate statistics, since these are all calculated on a common sample base of cases. Disadvantages stem from the potential loss of information in discarding incomplete cases. This loss of information has two aspects: loss of precision, and bias when the missing-data mechanism is not MCAR, and the complete cases are not a random sample of all the cases (Van der Heijden et al., 2006; White and Carlin, 2010). CC analysis may be justified in terms of simplicity when the loss of precision and the bias is minimal, so that the pay-off of exploiting the information in the incomplete cases will be minimal. This is more likely when the fraction of complete cases is high, but it is difficult to formulate general rules of thumb, since the degree of bias and loss of precision depends not only on the fraction of complete cases and pattern of missing data, but also on the extent to which complete and incomplete cases differ, and on the parameters of interest (Little and Rubin, 2019).

2.2.2 Multiple Imputation

Multiple imputation (MI) was developed as a general method for missing data which yields asymptotically consistent estimates of parameters when data are MAR and the imputation model is correctly specified (Rubin, 1976, 2004). When data are MAR, the observed data may be thought of as a simple random sample from the full data conditional on levels of observed values of variables, but not conditional on unobserved variables or missing values of observed variables. Multiple imputation may be used to account for data missing by chance (Wacholder, 1996; Westreich et al., 2015), a model for the mechanism by which the data became missing must be assumed.

MI refers to the procedure of replacing each missing value by a vector of $M \geq 2$ imputed values. The M values are ordered in the sense that M completed data sets can be created from the vectors of imputations; replacing each missing value by the first component in its vector of imputations creates the first completed data set, replacing each missing value by the second component in its vector creates the second completed data set, and so on. Standard complete-data methods are used to analyse each data set (Schafer, 1999; Stuart et al., 2009). When the M sets of imputations are repeated random draws from the predictive distribution of the missing values under a particular model for nonresponse, the M complete-data inferences can be combined to form one inference that properly reflects uncertainty due to nonresponse under that model (Qu and Lipkovich, 2009). Mitra and Reiter (2016) discussed PS matching for causal effect estimation after multiple imputation. They proposed two approaches namely, within and across approach to combine the imputed results. In the former approach, within each M completed data set apply PS matching method to obtain M treatment effect estimates and average out to get the MI point estimate. Under across approach, average each unit's M PSs to apply matching method based on their averaged scores, and estimate the treatment effect from this single set of matched controls. In addition to these two approaches, Leyrat et al. (2019) proposed a third method to combine MI inference called combined approach in which the PS parameters are combined rather than the PSs themselves. In this thesis, we will consider three of these approaches and also all the different estimators discussed in Section 1.2.

At first impute the missing data of confounder Z with multiple imputation using $\mathbf{Y}$, $\mathbf{T}$ and the observed confounders $\mathbf{X}$, where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$, $\mathbf{T} = (T_1, T_2, \ldots, T_n)^\top$ and $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)^\top$.

Let $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(M)}$ denote M imputed samples of $\mathbf{Z}$. Then the three MI approaches can be defined as below:

1. **Within Approach**: For each imputed sample $\mathbf{Z}^{(j)}$ with $j = 1, 2, \ldots, M$, the propensity score (PS) and outcome regression model is fitted from a generalized linear model i.e., for continuous dependent variable, fit a linear regression and for binary, fit a logistic regression model with appropriate set of covariates. Within each M completed data set, obtain M unit level estimated propensities $(\hat{e}_{i1}, \hat{e}_{i2}, \ldots, \hat{e}_{iM})$ and pair of estimated outcomes $\left\{\left(\hat{m}^1_{i1}, \hat{m}^0_{i1}\right), \left(\hat{m}^1_{i2}, \hat{m}^0_{i2}\right), \ldots, \left(\hat{m}^1_{iM}, \hat{m}^0_{iM}\right)\right\}$ for treated and control respectively, and finally M causal effect estimates $\left(\hat{\Delta}^{(k)}_1, \hat{\Delta}^{(k)}_2, \ldots, \hat{\Delta}^{(k)}_M\right)$ using different estimators, where k is the type of estimator. Finally, the point estimate under within approach, $\hat{\Delta}^{(k)}_{MI.wi}$ can be calculated after taking average of M effect estimates as follows:

$$\hat{\Delta}^{(k)}_{MI.wi} = \frac{1}{M} \sum_{j=1}^{M} \hat{\Delta}^{(k)}_j. \tag{2.4}$$

2. **Across Approach**: In the Across approach, estimate the PS for unit i, $\hat{e}^A_i$ by averaging $\hat{e}^j_i$ over the number of imputations as

$$\hat{e}^A_i = \frac{1}{M} \sum_{j=1}^{M} \hat{e}^j_i. \tag{2.5}$$

   So, $\hat{e}^A = \left(\hat{e}^A_1, \hat{e}^A_2, \ldots, \hat{e}^A_n\right)$ is the vector of unit level estimated PS after taking average of M estimates. Similarly, estimate of outcomes for treated and control can be obtained as $\hat{m}^{1A} = (1/M) \sum_{j=1}^{M} \hat{m}^1_{ij}$ and $\hat{m}^{0A} = (1/M) \sum_{j=1}^{M} \hat{m}^0_{ij}$ respectively. These combined PS's and outcomes will use to estimate ATE for different PS based estimators and regression based estimators respectively. Under across approach, we symbolize the effect estimates as $\hat{\Delta}^{(k)}_{MI.ac}$.

3. **Combined Approach**: In across method, we estimate the causal effects after averaging estimated PS or pair of outcomes from M imputed data sets. A third MI approach which is called *combined* approach, combine the model parameters rather combining PS or outcomes itself. Let us consider, $\boldsymbol{\beta} = (\beta_0, \beta_z, \beta_x, \beta_t)$ is the vector of parameters in the outcome regression model, where β_0 is the intercept, β_t, $\boldsymbol{\beta}_z$ and $\boldsymbol{\beta}_x$ are the regression coefficients for

treatment variable T, confounders Z and X respectively. Let, $\boldsymbol{\alpha} = (\alpha_0, \alpha_z, \alpha_x)$ is the vector of regression coefficient for the intercept term and confounders Z and X in PS model. In each of the M complete data set, we can fit the outcome regression and PS model and get the corresponding M parameter estimates $\left(\hat{\boldsymbol{\beta}}^1, \hat{\boldsymbol{\beta}}^2, \ldots, \hat{\boldsymbol{\beta}}^M\right)$ and $\left(\hat{\boldsymbol{\alpha}}^1, \hat{\boldsymbol{\alpha}}^2, \ldots, \hat{\boldsymbol{\alpha}}^M\right)$. Under combined approach, the average parameter estimates can be obtained by averaging M estimates as follows:

$$\hat{\bar{\boldsymbol{\beta}}}^c = \frac{1}{M}\sum_{j=1}^{M}\hat{\boldsymbol{\beta}}^j \qquad \hat{\bar{\boldsymbol{\alpha}}}^c = \frac{1}{M}\sum_{j=1}^{M}\hat{\boldsymbol{\alpha}}^j.$$

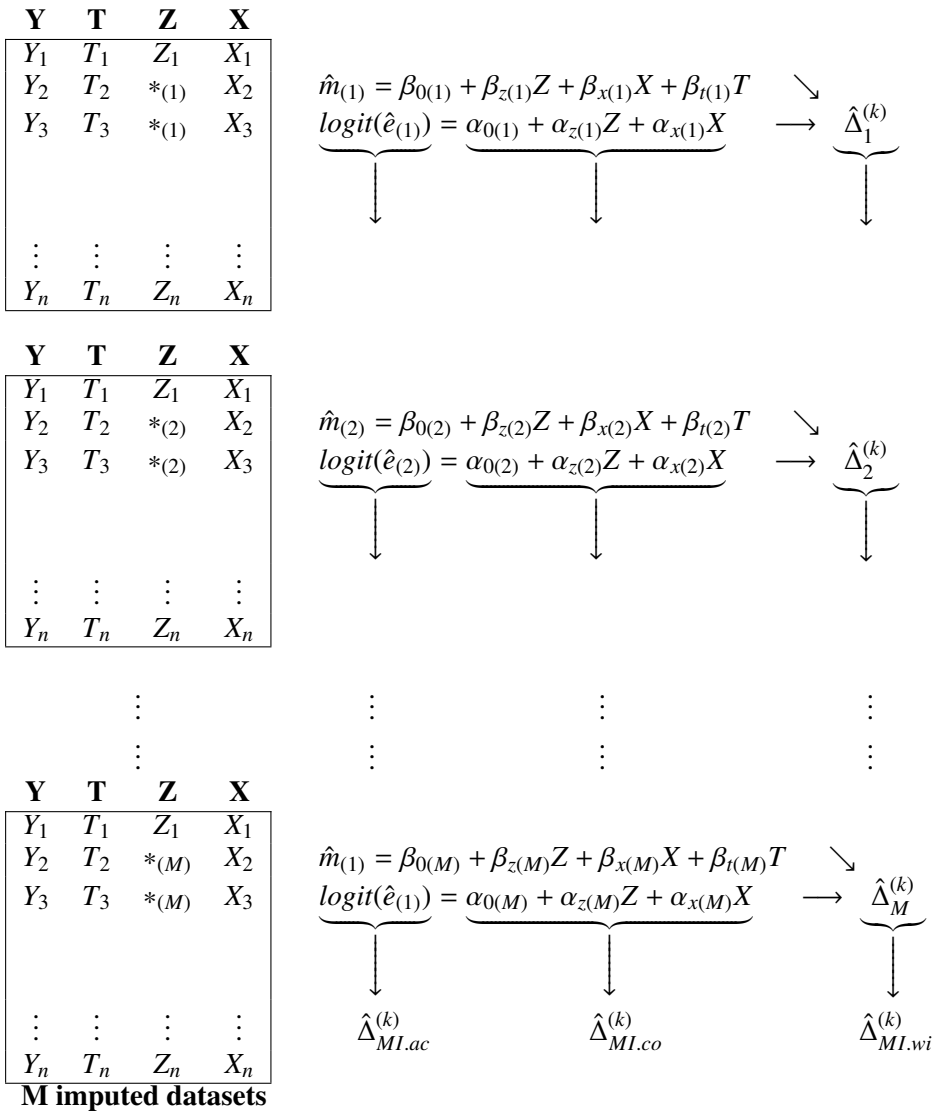
The imputed values of each individual's confounders Z_i are then averaged over the M datasets say $\bar{Z}^c$. Finally, using these averaged parameter values $\hat{\bar{\boldsymbol{\beta}}}^c = \left(\hat{\bar{\beta}}_0{}^c, \hat{\bar{\boldsymbol{\beta}}}_z{}^c, \hat{\bar{\boldsymbol{\beta}}}_x{}^c, \hat{\bar{\beta}}_t{}^c\right), \hat{\bar{\boldsymbol{\alpha}}}^c = \left(\hat{\bar{\alpha}}_0{}^c, \hat{\bar{\alpha}}_z{}^c, \hat{\bar{\alpha}}_x{}^c\right)$ and the unit level propensity and outcomes under both treatment conditions are estimated as

$$\begin{aligned}
&\hat{m}^{1c} = \hat{\bar{\beta}}_0{}^c + \hat{\bar{\boldsymbol{\beta}}}_z{}^c\bar{Z}^c + \hat{\bar{\boldsymbol{\beta}}}_x{}^c X + \hat{\bar{\beta}}_t{}^c && \hat{m}^{0c} = \hat{\bar{\beta}}_0{}^c + \hat{\bar{\boldsymbol{\beta}}}_z{}^c\bar{Z}^c + \hat{\bar{\boldsymbol{\beta}}}_x{}^c X && \text{[continuous outcome]}\\
&\hat{m}^{1c} = expit\left(\hat{\bar{\beta}}_0{}^c + \hat{\bar{\boldsymbol{\beta}}}_z{}^c\bar{Z}^c + \hat{\bar{\boldsymbol{\beta}}}_x{}^c X + \hat{\bar{\beta}}_t{}^c\right), && \hat{m}^{0c} = expit\left(\hat{\bar{\beta}}_0{}^c + \hat{\bar{\boldsymbol{\beta}}}_z{}^c\bar{Z}^c + \hat{\bar{\boldsymbol{\beta}}}_x{}^c X\right) && \text{[binary outcome]}\\
&\hat{e}^c = expit\left(\hat{\bar{\alpha}}_0{}^c + \hat{\bar{\boldsymbol{\alpha}}}_z{}^c\bar{Z}^c + \hat{\bar{\boldsymbol{\alpha}}}_x{}^c X\right) && && (2.6)
\end{aligned}$$

Using the above estimated outcomes and PS we can estimate ATE for different estimators. We denote the combined estimates of causal effects after MI as $\hat{\Delta}^{(k)}_{MI.co}$.

Table 2.1 shows the simple illustration of these three MI approaches. We consider the missing confounder set as Z observed set of confounders as X. Obviously, the outcome (Y) and treatment variable (T) are fully observed.

Table 2.1: The three approaches considered after multiple imputation (MI) of the partially observed covariates are missing values on the original dataset. $*_{(m)}, (m = 1, 2, \ldots, M)$ are imputed values in the mth imputed dataset.



After simple algebra it can be shown that under regression estimator $\hat{\Delta}_{MI.ac}^{(rg)} = \hat{\Delta}_{MI.wi}^{(rg)} = \hat{\Delta}_{MI.co}^{(rg)}$ for continuous outcome which implies the ATE estimates under across, within and combined approach are equal, we denote this as $MI_{ac.wi.co}$ and for G-estimation $\hat{\Delta}_{MI.wi}^{(gc)} = \hat{\Delta}_{MI.ac}^{(gc)}$ i.e., the ATE estimates under within and across approach are equal which we denoted by $MI_{ac.wi}$. If the outcome is binary then for both regression and G-estimation $\hat{\Delta}_{MI.ac}^{(.)} = \hat{\Delta}_{MI.wi}^{(.)}$ i.e., the estimates under within and across approach are equal and as before denoted by $MI_{ac.wi}$.

The performance of MI methods solely depends on the imputation model and number of imputations. Appropriate imputation model with large number of imputations will increase the efficiency of the method (Mitra and Reiter, 2016; Leyrat et al., 2019). In practice, appropriate model is uncertain and higher number of imputations will increase the computational time. In the next section, we will discuss the expectation-maximization (EM) algorithm to handle missingness in confounder.

2.2.3 The Expected-Maximization Algorithm

Suppose Y_i is the response, T_i is the binary treatment variable, Z_i is the confounder that is partially observed, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ denoting the confounder vector that is always precisely observed and M_i indicating the missingness of Z_i, i.e., $M_i = 1$ if Z_i is missing, 0 if Z_i is observed.Then the full data vector for i-th subject is, $D_i = (Y_i, T_i, Z_i, \mathbf{X}_i^\top, M_i)^\top$ where $Z_i = z \in \mathcal{Z}\,(0, 1)$ if $M_i = 0$ and $Z_i = \text{NA}$ if $M_i = 1$. The observed likelihood can be expressed as

$$
\begin{aligned}
L_o(\boldsymbol{\theta}) &= \prod_{i=1}^{n} P(D_i) = \prod_{i=1}^{n} P(Y_i, T_i, Z_i, \mathbf{X}_i, M_i; \boldsymbol{\theta}) \\
&\propto \prod_{i \in \mathcal{M}_0} P(M_i | Y_i, T_i, Z_i, \mathbf{X}_i; \boldsymbol{\gamma}) P(Y_i | T_i, Z_i, \mathbf{X}_i; \boldsymbol{\beta}) P(T_i | Z_i, \mathbf{X}_i; \boldsymbol{\alpha}) P(Z_i | \mathbf{X}_i; \boldsymbol{\delta}) \\
&\times \prod_{i \in \mathcal{M}_1} \sum_{z \in \mathcal{Z}} P(M_i | Y_i, T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\gamma}) P(Y_i | T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\beta}) P(T_i | Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}) P(Z_i = z | \mathbf{X}_i; \boldsymbol{\delta}),
\end{aligned}
\tag{2.7}
$$

where we have written the likelihood considering the missing confounder $\mathbf{Z}$ as categorical with $\mathcal{Z}$ is the set of all possible categories, $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\delta}^\top)^\top$ as the set of parameters, $\mathcal{M}_0 = \{i : M_i = 0\}$ is the subset of subjects with Z_i observed and $\mathcal{M}_1 = \{i : M_i = 1\}$ is the subset of subjects with Z_i missing. Our goal is to find the parameter estimates after maximizing the likelihood given by Equation (2.7). The observed likelihood includes a summation term over the binary confounder Z making it difficult to maximize over the parameter space specially when the dimension is large. Generally, EM algorithm uses complete data likelihood to find the parameter

estimates. The complete data likelihood can be expressed as

$$
\begin{aligned}
L_c(\boldsymbol{\theta}|\mathbf{D}) &= \underbrace{\prod_{i\in\mathcal{M}_0} Pr(Y_i, T_i, Z_i, \mathbf{X}_i, M_i|\boldsymbol{\theta})}_{\text{Complete-cases}} \underbrace{\prod_{i\in\mathcal{M}_1}\prod_{z\in\mathcal{Z}} Pr(Y_i, T_i, Z_i = z, \mathbf{X}_i, M_i|\theta)^{I(Z_i=z)}}_{Missing\ part} \\
&\propto \prod_{i\in\mathcal{M}_0} P(M_i|Y_i, T_i, Z_i, \mathbf{X}_i; \boldsymbol{\gamma}) P(Y_i|T_i, Z_i, \mathbf{X}_i; \boldsymbol{\beta}) P(T_i|Z_i, \mathbf{X}_i; \boldsymbol{\alpha}) P(Z_i|X_i; \boldsymbol{\delta}) \\
&\times \prod_{i\in\mathcal{M}_1}\prod_{z\in\mathcal{Z}} [P(M_i|Y_i, T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\gamma}) P(Y_i|T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\beta}) P(T_i|Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}) P(Z_i = z|\mathbf{X}_i; \boldsymbol{\delta})]^{I(Z_i=z)}
\end{aligned}
\tag{2.8}
$$

The corresponding complete data log-likelihood is given by

$$
\begin{aligned}
logL_c(\boldsymbol{\theta}|\mathbf{D}) &= \sum_{i\in\mathcal{M}_0} log\{P(M_i|Y_i, T_i, Z_i, \mathbf{X}_i; \boldsymbol{\gamma}) P(Y_i|T_i, Z_i, \mathbf{X}_i; \boldsymbol{\beta}) P(T_i|Z_i, \mathbf{X}_i; \boldsymbol{\alpha}) P(Z_i|\mathbf{X}_i; \boldsymbol{\delta})\} \\
&+ \sum_{i\in\mathcal{M}_1}\sum_{z\in\mathcal{Z}} I(Z_i = z) log\{P(M_i|Y_i, T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\gamma}) P(Y_i|T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\beta}) P(T_i|Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}) P(Z_i = z|\mathbf{X}_i; \boldsymbol{\delta})\}
\end{aligned}
\tag{2.9}
$$

The E step will calculate the expected complete data log-likelihood, more generally the functions of missing data appearing in the complete-data log-likelihood given the full observed data as

$$
\begin{aligned}
&E_{\boldsymbol{\theta}^{(\nu)}}\{logL_c(\boldsymbol{\theta}|\mathbf{D})\,|Y_i, T_i, \mathbf{X}_i, M_i\} \\
&= \sum_{i\in\mathcal{M}_0} log\left\{P(M_i|Y_i, T_i, Z_i, \mathbf{X}_i; \boldsymbol{\gamma}^{(\nu)}) P(Y_i|T_i, Z_i, \mathbf{X}_i; \boldsymbol{\beta}^{(\nu)}) P(T_i|Z_i, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)}) P(Z_i|\mathbf{X}_i; \boldsymbol{\delta}^{(\nu)})\right\} \\
&+ \sum_{i\in\mathcal{M}_1}\sum_{z\in\mathcal{Z}} E\{I(Z_i = z)|Y_i, T_i, \mathbf{X}_i\} log\left\{P(M_i|Y_i, T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\gamma}^{(\nu)}) P(Y_i|T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\beta}^{(\nu)})\right. \\
&\qquad \left. P(T_i|Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)}) P(Z_i = z|\mathbf{X}_i; \boldsymbol{\delta}^{\nu})\right\}
\end{aligned}
\tag{2.10}
$$

In the E step from the above equation, it actually computes the expectation of function of missing variables Z given the full observed variables i.e., $E\{I(Z_i = z)|Y_i, T_i, \mathbf{X}_i, M_i\}$. We know that the expectation of an indicator function is nothing but the probability. So the expectation term can be

further simplified as

$$
\begin{aligned}
E[I(Z_i = z)|Y_i, T_i, X_i, M_i; \boldsymbol{\theta}^{(\nu)}] &= P(Z_i = z|Y_i, T_i, X_i, M_i; \boldsymbol{\theta}^{(\nu)}) \\
&= \frac{P(Z_i = z, Y_i, T_i, X_i, M_i; \boldsymbol{\theta}^{(\nu)})}{P(Y_i, T_i, X_i, M_i; \boldsymbol{\theta}^{(\nu)})} \\
&= \frac{P(Y_i, Z_i = z, T_i, X_i, M_i; \boldsymbol{\theta}^{(\nu)})}{\sum_{z \in \mathcal{Z}} P(Y_i, Z_i = z, T_i, X_i, M_i; \boldsymbol{\theta}^{(\nu)})}
\end{aligned} \tag{2.11}
$$

The numerator and denominator are the same, where the denominator is marginalized over the distribution of missing confounder Z_i. Now this part can be simplified as

$$
\begin{aligned}
&P(Y_i, Z_i = z, T_i, \mathbf{X}_i, M_i; \boldsymbol{\theta}^{(\nu)}) \\
&= P(M_i|Y_i, T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\gamma}^{(\nu)}) \times P(Y_i|Z_i = z, T_i, \mathbf{X}_i; \boldsymbol{\beta}^{(\nu)}) \times P(T_i|Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)}) \times P(Z_i = z|X_i; \boldsymbol{\delta}^{(\nu)}) \times P(X_i)
\end{aligned} \tag{2.12}
$$

Substituting this results in Equation (2.11) we get

$$
\begin{aligned}
&E[I(Z_i = z)|Y_i, T_i, \mathbf{X}_i, M_i; \boldsymbol{\theta}^{(\nu)}] \\
&= \frac{P(M_i|Y_i, T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\gamma}^{(\nu)}) \times P(Y_i|Z_i = z, T_i, \mathbf{X}_i; \boldsymbol{\beta}^{(\nu)}) \times P(T_i|Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)}) \times P(Z_i = z|\mathbf{X}_i; \boldsymbol{\delta}^{(\nu)}) \times P(\mathbf{X}_i)}{\sum_{z \in \mathcal{Z}} P(M_i|Y_i, T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\gamma}^{(\nu)}) \times P(Y_i|Z_i = z, T_i, \mathbf{X}_i; \boldsymbol{\beta}^{(\nu)}) \times P(T_i|Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)}) \times P(Z_i = z|\mathbf{X}_i; \boldsymbol{\delta}^{(\nu)}) \times P(\mathbf{X}_i)} \\
&= \frac{P(M_i|Y_i, T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\gamma}^{(\nu)}) \times P(Y_i|Z_i = z, T_i, \mathbf{X}_i; \boldsymbol{\beta}^{(\nu)}) \times P(T_i|Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)}) \times P(Z_i = z|\mathbf{X}_i; \boldsymbol{\delta}^{(\nu)})}{\sum_{z \in \mathcal{Z}} P(M_i|Y_i, T_i, Z_i = z, \mathbf{X}_i; \boldsymbol{\gamma}^{(\nu)}) \times P(Y_i|Z_i = z, T_i, \mathbf{X}_i; \boldsymbol{\beta}^{(\nu)}) \times P(T_i|Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)}) \times P(Z_i = z|\mathbf{X}_i; \boldsymbol{\delta}^{(\nu)})}.
\end{aligned} \tag{2.13}
$$

Under MAR mechanism where missingness depends only on the fully observed data i.e., $P(M_i|Y_i, T_i, Z_i, \mathbf{X}_i) = P(M_i|Y_i, T_i, \mathbf{X}_i)$, the form of expected missing confounder in Equation (2.13) becomes

$$
\begin{aligned}
&E[I(Z_i = z)|Y_i, T_i, \mathbf{X}_i, M_i; \boldsymbol{\theta}^{(\nu)}] \\
&= \frac{P(M_i|Y_i, T_i, \mathbf{X}_i; \boldsymbol{\gamma}^{(\nu)}) \times P(Y_i|Z_i = z, T_i, \mathbf{X}_i; \boldsymbol{\beta}^{(\nu)}) \times P(T_i|Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)}) \times P(Z_i = z|\mathbf{X}_i; \boldsymbol{\delta}^{(\nu)}) \times P(\mathbf{X}_i)}{\sum_{z \in \mathcal{Z}} P(M_i|Y_i, T_i, \mathbf{X}_i; \boldsymbol{\gamma}^{(\nu)}) \times P(Y_i|Z_i = z, T_i, \mathbf{X}_i; \boldsymbol{\beta}^{(\nu)}) \times P(T_i|Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)}) \times P(Z_i = z|\mathbf{X}_i; \boldsymbol{\delta}^{(\nu)}) \times P(\mathbf{X}_i)} \\
&= \frac{P(Y_i|Z_i = z, T_i, \mathbf{X}_i; \boldsymbol{\beta}^{(\nu)}) \times P(T_i|Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)}) \times P(Z_i = z|\mathbf{X}_i; \boldsymbol{\delta}^{(\nu)})}{\sum_{z \in \mathcal{Z}} P(Y_i|Z_i = z, T_i, \mathbf{X}_i; \boldsymbol{\beta}^{(\nu)}) \times P(T_i|Z_i = z, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)}) \times P(Z_i = z|\mathbf{X}_i; \boldsymbol{\delta}^{(\nu)})}.
\end{aligned} \tag{2.14}
$$

In this thesis, we apply the EM algorithm under MAR mechanism. For a given set of parameters $\boldsymbol{\theta}^{(\nu)} = \left(\boldsymbol{\beta}^{(\nu)\top}, \boldsymbol{\alpha}^{(\nu)\top}, \boldsymbol{\delta}^{(\nu)\top}\right)^{\top}$, we can calculate the expected missing confounder Z using Equation (2.14) which will be used as weights in the maximization step. The M-step which maximize the

complete data log-likelihood given in Equation (2.9), will use a weighted least square approach with a weight of 1 for complete cases and the cases with missing Z assign a weight that has been calculated from Equation (2.14). The procedure is straightforward, with the initial values of parameter $\boldsymbol{\theta}^{(\nu)}$ calculate the weights from Equation (2.14). Update the parameter values for next step $\boldsymbol{\theta}^{(\nu+1)}$ using the calculated weights. Iterate this process until desired accuracy or convergence. After the convergence of EM algorithm, we will use the converged parameter estimates and corresponding weights in the estimation of average treatment effect (ATE).

Let us consider $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\alpha}}^\top, \hat{\boldsymbol{\gamma}}^\top, \hat{\boldsymbol{\delta}}^\top\right)^\top$ be the set of parameters obtained after the convergence of EM algorithm. Where $\boldsymbol{\beta} = (\beta_0, \beta_t, \beta_z, \boldsymbol{\beta}_x^\top)^\top$ is the set of parameters for outcome regression model, $\boldsymbol{\alpha} = (\alpha_0, \alpha_z, \boldsymbol{\alpha}_x^\top)^\top$ for PS model, $\boldsymbol{\gamma} = \left(\gamma_0, \boldsymbol{\gamma}_x^\top, \gamma_y, \gamma_t\right)^\top$ parameter vector for missing model and $\boldsymbol{\delta} = (\delta_0, \boldsymbol{\delta}_x^\top)^\top$ is the parameter vector for associational model between Z and X. Using the converged parameter values, assign the following weights for different categories

$$W_i^z = \begin{cases} 1, & \text{for } i \in \mathcal{M}_0, \\ P\left[I(Z_i = 0)|Y_i, T_i, \mathbf{X}_i, M_i; \boldsymbol{\theta}^{(\nu)}\right], & \text{for } i \in \mathcal{M}_1 \ \& \ z = 0, \\ 1 - P\left[I(Z_i = 0)|Y_i, T_i, \mathbf{X}_i, M_i; \boldsymbol{\theta}^{(\nu)}\right], & \text{for } i \in \mathcal{M}_1 \ \& \ z = 1. \end{cases} \tag{2.15}$$

The expected outcome under EM algorithm is calculated using the converged outcome regression model parameters and corresponding weights which is essentially the expected missing confounder as

$$\hat{m}(Y; \hat{\boldsymbol{\beta}}) = \begin{cases} E(Y|T, Z, \mathbf{X}; \hat{\boldsymbol{\beta}}), & \text{for } M = 0 \\ \sum_z E(Y|T, Z = z, \mathbf{X}; \hat{\boldsymbol{\beta}}) \times W^z, & \text{for } M = 1, \end{cases} \tag{2.16}$$

where the outcome under consideration is continuous. In case of binary outcome, $E(Y|T, Z, \mathbf{X}; \hat{\boldsymbol{\beta}}) = P\left(Y = 1|T, Z, \mathbf{X}; \hat{\boldsymbol{\beta}}\right) = expit\left(\hat{\beta}_0 + \hat{\beta}_t T + \hat{\beta}_z Z + \mathbf{X}^\top \hat{\boldsymbol{\beta}}_x\right)$ which is the probability of being $Y = 1$ given the other covariates. For regression estimator, the ATE estimate using the above estimated outcome can be defined as

$$\hat{\Delta}^{(reg)} = \frac{1}{n_1} \sum_{T_i=1} \hat{m}_i(Y; \hat{\boldsymbol{\beta}}) - \frac{1}{n_0} \sum_{T_i=0} \hat{m}_i(Y; \hat{\boldsymbol{\beta}}), \tag{2.17}$$

where n_1 and n_0 are the number of treated and untreated subjects respectively. The first part of the above equation, the summation is taken over the treated subjects ($T = 1$), whereas the second term for untreated subjects. In case of continuous outcome, ATE estimate will be $\hat{\beta}_t$ i.e., the converged

regression coefficient of treatment in outcome regression model.

In G-estimation, we need to fit two separate weighted least square for the treated and control groups including all confounders. Here the weights are the expected missing confounder obtained from E-steps of the EM algorithm using final converged parameters. Calculate the expected outcome under treated and control conditions as follows

$$m_t(Y;\hat{\boldsymbol{\beta}}^{(t)}) = \begin{cases} E(Y|Z,\mathbf{X};\hat{\boldsymbol{\beta}}^{(t)}), & \text{for } M=0 \\ \sum_z E(Y|Z=z,\mathbf{X};\hat{\boldsymbol{\beta}}^{(t)}) \times W^z, & \text{for } M=1, \end{cases} \tag{2.18}$$

where $\hat{\boldsymbol{\beta}}^{(t)} = \left(\hat{\beta}_0^{(t)}, \hat{\beta}_z^{(t)}, \hat{\boldsymbol{\beta}}_x^{(t)\top}\right)^\top$ is the parameter estimates of regression model Y on Z and $\mathbf{X}$ by weighted least square for subset of individuals with treatment $T = t$, 0 for the control and 1 for the treated group. The weights are the estimated missing confounder calculated in EM algorithm. Finally, the ATE estimate under G-estimation is given by

$$\hat{\Delta}^{(ge)} = \frac{1}{n}\sum_{i=1}^{n}\left[m_{i1}(Y;\hat{\boldsymbol{\beta}}^{(1)}) - m_{i0}(Y;\hat{\boldsymbol{\beta}}^{(0)})\right] \tag{2.19}$$

In case of PS based estimators, we need to compute propensities using the converged parameters and weights from EM algorithm. The estimated PS using EM algorithm can be obtained as

$$\hat{e}(Z,\mathbf{X};\hat{\boldsymbol{\alpha}}) = \begin{cases} P(T=1|Z,\mathbf{X};\hat{\boldsymbol{\alpha}}), & \text{for } M=0 \\ \sum_z P(T=1|Z=z,\mathbf{X};\hat{\boldsymbol{\alpha}}) \times W^z, & \text{for } M=1. \end{cases} \tag{2.20}$$

In the definition of PS based estimators defined in Section 1.2.2, replace the PS $e(\mathbf{X})$ with $e(Z,\mathbf{X};\hat{\boldsymbol{\alpha}})$ to get ATE estimates under EM algorithm. For matching estimator $\hat{\Delta}^{(mc)}$ use Equation (1.23), stratified estimator $\hat{\Delta}^{(st)}$ use Equation (1.24), IPW estimator $\hat{\Delta}^{(ipw)}$ from Equation (1.29) under both continuous and binary outcomes. In PS regression estimator, fit a linear regression and logistic regression with treatment and estimated PS $e\left(Z,\mathbf{X};\hat{\boldsymbol{\theta}}\right)$ as covariates for continuous and binary outcome respectively. In case of continuous outcome, the regression coefficient of treatment in the outcome model will be the ATE estimate of $\hat{\Delta}^{(psr)}$ (Equation (1.26)). For binary outcome use Equation (1.27) to get the ATE estimate.

Lastly, to obtain the ATE estimate using AIPW estimator $\hat{\Delta}^{(aipw)}$, modify the formula in Equa-

tion (1.32) as follows

$$\hat{\Delta}^{(aipw)} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i Y_i}{\hat{e}(Z_i, \mathbf{X}_i; \hat{\boldsymbol{\alpha}})} - \frac{\{T_i - \hat{e}(Z_i, \mathbf{X}_i; \hat{\boldsymbol{\alpha}})\}}{\hat{e}(Z_i, \mathbf{X}_i; \hat{\boldsymbol{\alpha}})} m_{i1}(Y; \hat{\boldsymbol{\beta}}^{(1)})\right]$$
$$- \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1 - T_i) Y_i}{1 - \hat{e}(Z_i, \mathbf{X}_i; \hat{\boldsymbol{\alpha}})} + \frac{\{T_i - \hat{e}(Z_i, \mathbf{X}_i; \hat{\boldsymbol{\alpha}})\}}{1 - \hat{e}(Z_i, \mathbf{X}_i; \hat{\boldsymbol{\alpha}})} m_{i0}(Y; \hat{\boldsymbol{\beta}}^{(0)})\right],$$

where $\hat{e}(Z_i, \mathbf{X}_i; \hat{\boldsymbol{\alpha}})$ is the estimated propensities for ith subject obtained from Equation (2.20), and $m_{i1}(Y; \hat{\boldsymbol{\beta}}^{(1)})$ and $m_{i0}(Y; \hat{\boldsymbol{\beta}}$ are the estimated outcomes under treated and untreated calculated using Equation (2.18).

Multiple Models Consideration

For estimating the ATE using EM algorithm, we need to obtain expected missing confounder in E-step. Equation (2.14) shows that in calculating the expectation we have to specify outcome regression and PS model but no missing model. The estimation consistency depends on the correct specification of these models. With an unknown data generating process, it is often risky to assume that the postulated models are correctly specified. Therefore, multiple models may be fitted in practice, each involving different subsets of covariates, with none of them ruling out the possibility of others. Such multiple models increase the likelihood of correct specification. To improve the robustness of EM method, we postulate J (multiple) models $\mathcal{Y} = \left\{a^j\left(T, Z, \mathbf{X}; \boldsymbol{\beta}^j\right) : \ j = 1, 2, \ldots, J\right\}$ for outcome regression and K models $\mathcal{T} = \left\{\pi^k\left(Z, X; \boldsymbol{\alpha}^k\right) : \ k = 1, 2, \ldots, K\right\}$ for PS. Here the $\boldsymbol{\beta}^j$ and $\boldsymbol{\alpha}^k$ are the vector of corresponding model parameters. Considering these multiple models, the expected missing confounder can be calculated as the average values from $(J \times K)$ multiple model combinations of outcome regression and propensities:

$$\bar{E}[I(Z_i = z_i)|Y_i, T_i, \mathbf{X}_i, M_i; \boldsymbol{\theta}^{(\nu)}]$$
$$= \left(\frac{1}{J \times K}\right) \frac{P(Z_i = z_i|\mathbf{X}_i; \boldsymbol{\delta}^{(\nu)}) \times \left\{\sum_j \sum_k a^j\left(T_i, Z_i = z_i, \mathbf{X}_i; \boldsymbol{\beta}^{(\nu)j}\right) \times \pi^k\left(Z_i = z_i, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)k}\right)\right\}}{\sum_{z_i \in \mathbf{Z}} P(Z_i = z_i|\mathbf{X}_i; \boldsymbol{\delta}^{(\nu)}) \times \left\{\sum_j \sum_k a^j\left(T_i, Z_i = z_i, \mathbf{X}; \boldsymbol{\beta}^{(\nu)j}\right) \times \pi^k\left(Z_i = z_i, \mathbf{X}_i; \boldsymbol{\alpha}^{(\nu)k}\right)\right\}}, \qquad (2.21)$$

where $\boldsymbol{\theta}^{(\nu)} = \left(\boldsymbol{\beta}^{(\nu)\top}, \boldsymbol{\alpha}^{(\nu)\top}, \boldsymbol{\delta}^{(\nu)\top}\right)^\top$ with $\boldsymbol{\beta}^{(\nu)j}$, $\boldsymbol{\alpha}^{(\nu)k}$ and $\boldsymbol{\alpha}^{(\nu)}$ are the vector of parameters for jth outcome regression model, kth PS model and associational model for missing confounder respectively in the νth step of EM algorithm. We will use Equation (2.21) to calculate the weights for

cases with missing Z values under multiple models and obtain the parameter estimates of complete data log-likelihood in Equation (2.9) via weighted least square method. We will apply the multiple models concept for the AIPW estimator to improve the double robustness property. Since there is no place for the missing models in weight calculation, we will focus on the double robustness in causal inference but not for missingness.

Chapter 3

Simulation Studies

3.1 Simulation Setup

Let Y be the outcome (continuous or binary), T the binary treatment, and three confounders X_1, X_2 and Z. Define the missing indicator M which can take a value 1 when Z is missing and 0 when observed. We define the following models with correct (C) specification:

$$
\begin{aligned}
c-\text{model} &: logit\,[P(Z=1|X_1,\boldsymbol{\delta})] = \delta_0 + \delta_1 X_1 \\
e-\text{model} &: logit\,[P(T=1|X_1,X_2,Z;\boldsymbol{\alpha})] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 Z \\
y-\text{model} &: E(Y|X_1,X_2,Z,T;\boldsymbol{\beta}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z + \beta_4 T \qquad \text{[continuous outcome]} \\
&: logit\,[P(Y=1|X_1,X_2,Z,T;\boldsymbol{\beta})] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 Z + \beta_4 T \quad \text{[binary outcome]} \\
m-\text{model} &: logit\,[P(M=1|Y,X_1,X_2,T;\boldsymbol{\gamma})] = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 Y + \gamma_4 T,
\end{aligned}
$$

where $c-$model reveals the dependency of Z on X_1 with $exp(\alpha_1)$ = odds ratio (OR), $e-$model is defined for the propensity score (PS) model, y-model is the outcome regression model, and $m-$model is defined for the missing model. Throughout the discussion, we consider the above $y-$ and $e-$ models as correctly (C) specified $y-$ and $e-$model respectively. These models are considered incorrect (I) when they excludes X_1 from their specification:

$$
\begin{aligned}
\text{Incorrect } (I) \text{ model:} \quad & y-\text{model: } \beta_0 + \beta_2 X_2 + \beta_3 Z + \beta_4 T \\
& e-\text{model: } \alpha_0 + \alpha_2 X_2 + \alpha_3 Z.
\end{aligned}
$$

The distribution of two independent confounders are $X_1 \sim Bernouli(50\%)$ and $X_2 \sim Normal(\mu = 0, \sigma = 1)$. First, we fixed the odds ratios at 1 and 4 which specifies the parameter vector for c-model as $\boldsymbol{\delta} = (\delta_0, 0)^\top$ and $\boldsymbol{\delta} = (\delta_0, log4)^\top$ respectively. Under these two odds ratio, we set different models parameter vectors as $\boldsymbol{\alpha} = (\alpha_0, 0.5, -1, 0.5)^\top$ and $\boldsymbol{\gamma} = (\gamma_0, 1, -2, 0.2, 1)^\top$ for both continuous and binary outcomes, $\boldsymbol{\beta} = (10, -2.5, 1, 4.5, 2)^\top$ for continuous, and $\boldsymbol{\beta} = (\beta_0, -1.5, 1, 1.75, 2)^\top$ for binary outcome. We solve for δ_0, α_0, β_0 (binary outcome) and γ_0 in such a way that will confirm $E(Z) = 0.5$, $E(T) = 0.3$ implies 30% treated, $E(M) = 0.4$ implies 40% missing in the confounder Z and $E(Y) = 0.4$ i.e., 40% response rate for binary outcome. Throughout the analysis we consider a sample of size $n = 500$.

Table 3.1: Parameter setup for the simulation study.

	Continuous outcome		Binary outcome	
Model	$OR = 1$	$OR = 4$	$OR = 1$	$OR = 4$
c	$\delta_0 = 0$	$\delta_0 = -0.693$	$\delta_0 = 0$	$\delta_0 = -0.693$
e	$\alpha_0 = -1.537$	$\alpha_0 = -1.544$	$\alpha_0 = -1.537$	$\alpha_0 = -1.544$
y			$\beta_0 = -1.317$	$\beta_0 = -1.283$
m	$\gamma_0 = -3.787$	$\gamma_0 = -3.791$	$\gamma_0 = -1.572$	$\gamma_0 = -1.572$

3.2 Simulation Results

We compare the performance of different causal effect estimators under different methods for dealing missingness along with the standard complete data method with no missing. Standard complete data method is identified as WM (without missing), expected-maximization method as EM, two complete case methods MCAR.cc and MAR.cc for missing completely at random (MCAR) and missing at random (MAR) respectively. Finally, three multiple imputation methods, across, within and combined are defined as MI.ac, MI.wi and MI.co respectively. The comparison is made in terms of bias, empirical standard error (ESE), root mean square error (RMSE), median absolute error (MAE) and relative bias percentage (RB). Let us consider Δ and $\hat{\Delta}$ are the true and estimate

of average treatment effects (ATE) then these measures can be defined as:

$$\text{Bias} = \left(\overline{\hat{\Delta}} - \Delta\right); \quad \text{ESE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{\Delta}_i - \overline{\hat{\Delta}}\right)^2}; \quad \text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{\Delta}_i - \Delta\right)^2}$$

$$\text{MAE} = \textit{median}\left(|\hat{\Delta}_i - \Delta|\right); \quad \%\text{RB} = \frac{\left(\overline{\hat{\Delta}} - \Delta\right)}{\Delta} \times 100. \tag{3.1}$$

Bias is a measure of closeness between the estimate and true effects. Smaller the bias better the estimator. For an unbiased estimator it is desired that ESE and RMSE should be close to each other. RB is the bias as a percentage of true effects. A useful rule of thumb is that an estimator is said to be better performed if the RB value remains within 0-5%. To examine the necessity of including highly associated variable (X_1) with the missing confounder (Z), we consider two odds ratios 1 indicates independence and 4 indicates high association.

Continuous outcome

Table 3.2 and 3.3 compares the simulation results for regression estimator and G-estimation for continuous outcome respectively. All the methods for dealing missingness under regression estimator and G-estimation are unbiased when the postulated $y-$model is correct (C) and biased when incorrect (I) for both independent and associated missing confounder Z. The RB values range within the desired accuracy (5%) for correct model and outperformed for incorrect model specification. The MAR.cc method appears least efficient, followed by MCAR.cc, MI.co, and MI.wi methods, while EM and WM methods are nearly equivalent and most efficient in terms of ESE. Based on the accuracy measures RMSE and MAE, EM method perform best and MAR.cc perform worse.

Table 3.2: Bias and efficiency measures for regression estimator under different methods for dealing missingness for continuous outcome.

		OR = 1					OR = 4				
y-		True ATE = 2					True ATE = 2				
model	Method	Bias	ESE	RMSE	MAE	%RB	Bias	ESE	RMSE	MAE	%RB
C	WM	0.005	0.108	0.108	0.075	0.25	0.002	0.107	0.107	0.070	0.10
	EM	0.005	0.100	0.100	0.068	0.25	0.006	0.105	0.105	0.071	0.30
	MCAR.cc	0.002	0.138	0.138	0.093	0.10	-0.001	0.140	0.140	0.096	-0.05
	MAR.cc	-0.054	0.170	0.178	0.117	-2.70	-0.050	0.171	0.178	0.119	-2.50
	MI.ac.wi.co	-0.002	0.129	0.129	0.082	-0.10	-0.005	0.133	0.133	0.088	-0.25
I	WM	-0.305	0.170	0.350	0.304	-15.25	-0.277	0.159	0.319	0.280	-13.85
	EM	-0.296	0.167	0.340	0.294	-14.80	-0.270	0.167	0.317	0.276	-13.50
	MCAR.cc	-0.305	0.217	0.375	0.306	-15.25	-0.272	0.216	0.347	0.274	-13.60
	MAR.cc	-0.286	0.277	0.398	0.301	-14.30	-0.262	0.264	0.371	0.280	-13.10
	MI.ac.wi.co	-0.310	0.186	0.362	0.307	-15.50	-0.287	0.173	0.335	0.292	-14.35

Table 3.3: Bias and efficiency measures for G-estimation under different methods for dealing missingness for continuous outcome.

		OR = 1					OR = 4				
y-		True ATE = 2					True ATE = 2				
model	Method	Bias	ESE	RMSE	MAE	%RB	Bias	ESE	RMSE	MAE	%RB
C	WM	0.004	0.115	0.115	0.078	0.20	0.000	0.115	0.115	0.077	0.00
	EM	0.006	0.106	0.107	0.070	0.30	0.006	0.115	0.115	0.076	0.30
	MCAR.cc	0.002	0.147	0.147	0.099	0.10	-0.000	0.150	0.150	0.100	-0.00
	MAR.cc	-0.045	0.193	0.198	0.134	-2.25	-0.043	0.193	0.197	0.134	-2.15
	MI.ac.wi	-0.003	0.133	0.133	0.085	-0.15	-0.005	0.138	0.138	0.092	-0.25
	MI.co	-0.022	0.134	0.136	0.091	-1.10	-0.024	0.140	0.142	0.096	-1.20
I	WM	-0.311	0.181	0.360	0.311	-15.55	-0.283	0.168	0.329	0.290	-14.15
	EM	-0.300	0.176	0.348	0.300	-15.00	-0.272	0.179	0.325	0.271	-13.60
	MCAR.cc	-0.310	0.234	0.388	0.315	-15.50	-0.277	0.230	0.360	0.280	-13.85
	MAR.cc	-0.295	0.300	0.421	0.308	-14.75	-0.267	0.288	0.392	0.293	-13.35
	MI.ac.wi	-0.313	0.197	0.370	0.314	-15.65	-0.291	0.187	0.345	0.294	-14.55
	MI.co	-0.332	0.198	0.387	0.331	-16.60	-0.311	0.188	0.364	0.313	-15.55

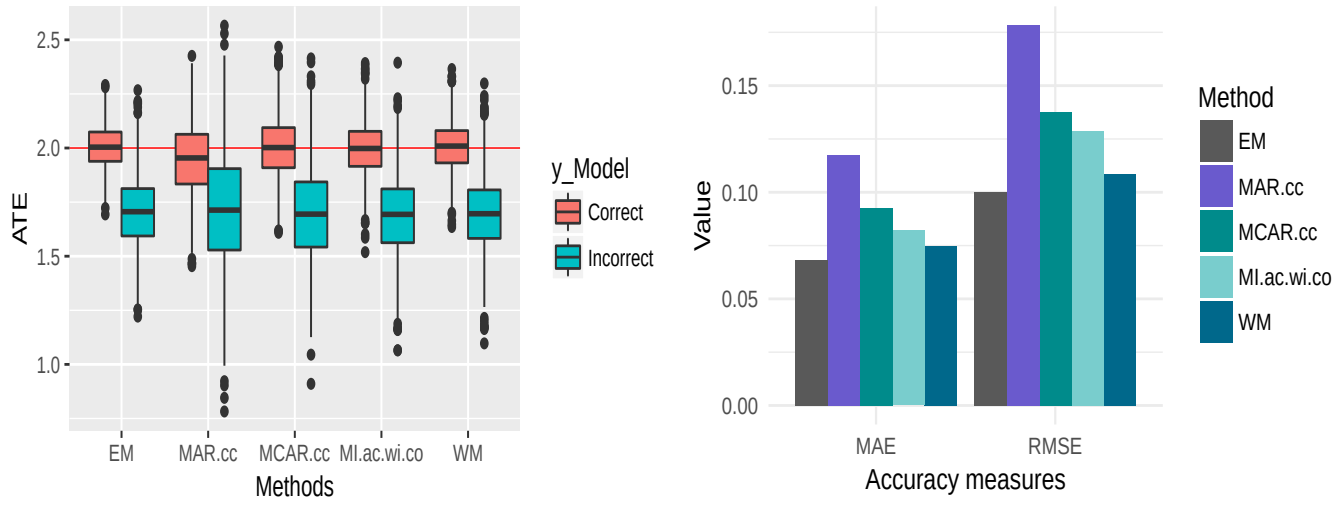


Figure 3.1: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for regression estimator with continuous outcome.

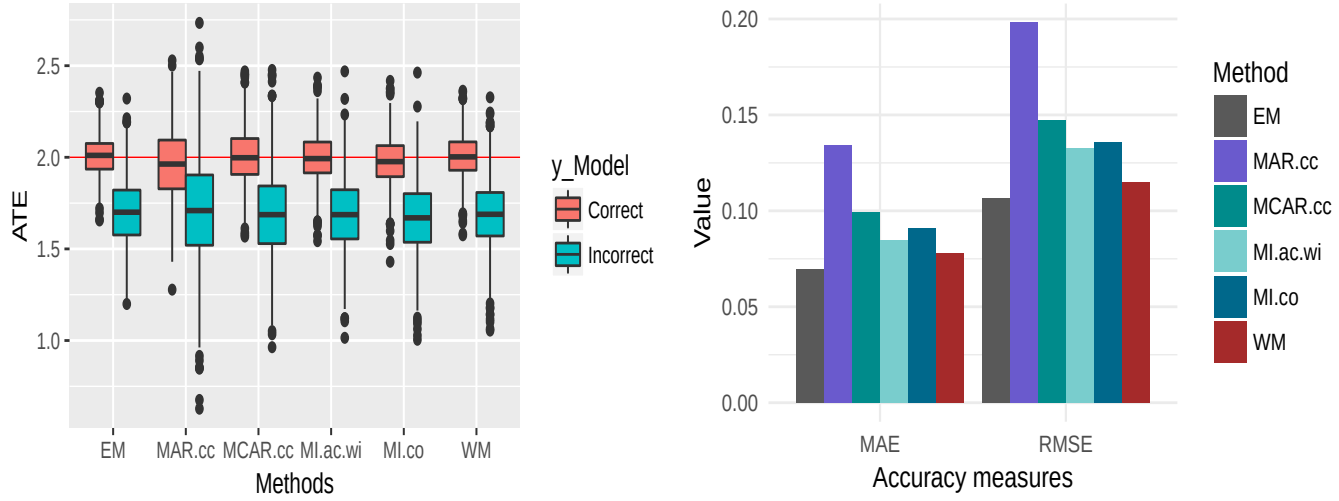


Figure 3.2: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for G-estimation with continuous outcome.

Figure 3.1 and 3.2 displays the boxplots of ATE estimates and barplots of efficiency measures RMSE and MAE based on 1000 MCMC simulations for regression estimator and G-estimation respectively considering independent missing confounder (OR = 1). The true average treatment effect is indicated by the red colored horizontal line. Boxplots depicts the fact that under correct y-model the median line of all methods are located toward the true effect at 2 i.e., unbiased and

far away from 2 for incorrect y-model i.e., biased. As we concluded earlier, it is clear from the barplots that, EM method perform best with lowest values for RMSE and MAE, whereas MAR.cc method perform worse with highest values.

Table 3.4-3.7 shows the simulation results for different PS based estimators. Under the different PS based estimators, all methods for dealing missingness (exception for MAR.cc) are unbiased when the postulated PS (e) model is correct (C) and biased when incorrect (I) for both odds ratios 1 and 4. In terms of RB, all methods for dealing missingness perform amply when the e-model is correct (C) and perform poorly for incorrect model. The poor performance is confirmed with a RB value greater than 5%. Based on efficiency measure ESE and accuracy measures RMSE and MAE, the MAR.cc method always underperformed in any combination of model specification (C or I) and odds ratios (1 or 4). The MI.ac is found as the most efficient and accurate method for IPW estimator, MI.wi for PS (logit) matching estimator, and for other PS based estimators stratified and regression, EM is the most efficient method.

Table 3.4: Bias and efficiency measures for propensity score 1-to-1 matching (logit) under different methods for dealing missingness for continuous outcome.

		OR = 1 True ATE = 2					OR = 4 True ATE = 2				
e-model	Method	Bias	ESE	RMSE	MAE	%RB	Bias	ESE	RMSE	MAE	%RB
C	WM	-0.002	0.261	0.261	0.177	-0.10	0.006	0.232	0.232	0.153	0.30
	EM	-0.001	0.258	0.258	0.175	-0.05	0.005	0.227	0.227	0.158	0.25
	MCAR.cc	-0.006	0.335	0.335	0.218	-0.30	-0.013	0.301	0.302	0.198	-0.65
	MAR.cc	-0.045	0.392	0.394	0.257	-2.25	-0.026	0.342	0.343	0.227	-1.30
	MI.ac	0.008	0.268	0.268	0.185	0.40	0.008	0.249	0.249	0.166	0.40
	MI.wi	0.007	0.238	0.238	0.164	0.35	-0.001	0.213	0.213	0.145	-0.05
	MI.co	0.006	0.268	0.268	0.179	0.30	0.010	0.248	0.249	0.168	0.50
I	WM	-0.313	0.288	0.425	0.329	-15.65	-0.275	0.257	0.376	0.286	-13.75
	EM	-0.316	0.291	0.429	0.331	-15.80	-0.273	0.259	0.377	0.290	-13.65
	MCAR.cc	-0.334	0.375	0.502	0.358	-16.70	-0.283	0.343	0.445	0.302	-14.15
	MAR.cc	-0.296	0.442	0.532	0.374	-14.80	-0.252	0.407	0.479	0.341	-12.60
	MI.ac	-0.316	0.292	0.430	0.324	-15.80	-0.281	0.270	0.390	0.287	-14.05
	MI.wi	-0.308	0.268	0.408	0.306	-15.40	-0.283	0.248	0.376	0.285	-14.15
	MI.co	-0.318	0.289	0.430	0.326	-15.90	-0.279	0.267	0.386	0.288	-13.95

Table 3.5: Bias and efficiency measures for propensity score stratified estimator (strata = 10) under different methods for dealing missingness for continuous outcome.

		OR = 1					OR = 4				
e-		True ATE = 2					True ATE = 2				
model	Method	Bias	ESE	RMSE	MAE	%RB	Bias	ESE	RMSE	MAE	%RB
	WM	-0.002	0.179	0.179	0.120	-0.10	-0.008	0.162	0.162	0.107	-0.40
	EM	0.011	0.166	0.166	0.103	0.55	0.005	0.160	0.160	0.105	0.25
	MCAR.cc	0.002	0.235	0.235	0.150	0.10	-0.003	0.220	0.220	0.140	-0.15
	MAR.cc	-0.068	0.301	0.308	0.192	-3.40	-0.065	0.280	0.287	0.188	-3.25
	MI.ac	0.008	0.191	0.191	0.122	0.40	-0.004	0.180	0.180	0.124	-0.20
	MI.wi	0.005	0.189	0.189	0.123	0.25	-0.005	0.177	0.177	0.121	-0.25
	MI.co	0.007	0.192	0.192	0.120	0.35	-0.005	0.181	0.181	0.123	-0.25
I	WM	-0.311	0.220	0.381	0.311	-15.55	-0.288	0.198	0.349	0.290	-14.40
	EM	-0.291	0.216	0.362	0.302	-14.55	-0.272	0.206	0.341	0.276	-13.60
	MCAR.cc	-0.309	0.279	0.417	0.321	-15.45	-0.271	0.274	0.386	0.289	-13.55
	MAR.cc	-0.311	0.363	0.478	0.344	-15.55	-0.281	0.329	0.433	0.317	-14.05
	MI.ac	-0.298	0.239	0.382	0.301	-14.90	-0.281	0.225	0.360	0.282	-14.05
	MI.wi	-0.299	0.236	0.380	0.300	-14.95	-0.283	0.222	0.359	0.280	-14.15
	MI.co	-0.300	0.239	0.383	0.304	-15.00	-0.282	0.225	0.361	0.281	-14.10

Table 3.6: Bias and efficiency measures for propensity score regression under different methods for dealing missingness for continuous outcome.

		OR = 1					OR = 4				
e-		True ATE = 2					True ATE = 2				
model	Method	Bias	ESE	RMSE	MAE	%RB	Bias	ESE	RMSE	MAE	%RB
C	WM	0.004	0.110	0.110	0.075	0.20	0.002	0.107	0.107	0.070	0.10
	EM	0.006	0.102	0.102	0.068	0.30	0.007	0.106	0.106	0.070	0.35
	MCAR.cc	0.001	0.139	0.139	0.094	0.05	-0.002	0.143	0.143	0.099	-0.10
	MAR.cc	-0.060	0.175	0.186	0.125	-3.00	-0.056	0.173	0.181	0.122	-2.80
	MI.ac	0.000	0.130	0.130	0.085	-0.00	-0.004	0.134	0.134	0.089	-0.20
	MI.wi	-0.002	0.130	0.130	0.086	-0.10	-0.006	0.134	0.134	0.090	-0.30
	MI.co	-0.001	0.130	0.130	0.086	-0.05	-0.005	0.134	0.134	0.090	-0.25
I	WM	-0.309	0.172	0.353	0.307	-15.45	-0.280	0.160	0.322	0.283	-14.00
	EM	-0.299	0.169	0.343	0.298	-14.95	-0.272	0.167	0.319	0.278	-13.60
	MCAR.cc	-0.309	0.219	0.379	0.308	-15.45	-0.276	0.218	0.351	0.280	-13.80
	MAR.cc	-0.291	0.279	0.403	0.303	-14.55	-0.266	0.264	0.375	0.284	-13.30
	MI.ac	-0.311	0.188	0.364	0.311	-15.55	-0.290	0.175	0.339	0.295	-14.50
	MI.wi	-0.313	0.188	0.365	0.312	-15.65	-0.291	0.175	0.340	0.296	-14.55
	MI.co	-0.312	0.188	0.365	0.312	-15.60	-0.291	0.175	0.339	0.295	-14.55

Table 3.7: Bias and efficiency measures for inverse probability weighting (IPW) estimator under different methods for dealing missingness for continuous outcome.

e-model	Method	OR = 1, True ATE = 2 Bias	ESE	RMSE	MAE	%RB	OR = 4, True ATE = 2 Bias	ESE	RMSE	MAE	%RB
C	WM	-0.008	0.212	0.212	0.135	-0.40	-0.006	0.206	0.206	0.126	-0.30
	EM	0.016	0.213	0.213	0.134	0.80	0.002	0.190	0.190	0.121	0.10
	MCAR.cc	-0.006	0.270	0.270	0.168	-0.30	-0.012	0.257	0.257	0.155	-0.60
	MAR.cc	-0.086	0.271	0.284	0.181	-4.30	-0.077	0.267	0.278	0.170	-3.85
	MI.ac	-0.016	0.142	0.143	0.093	-0.80	-0.019	0.146	0.148	0.098	-0.95
	MI.wi	-0.006	0.237	0.237	0.145	-0.30	-0.012	0.214	0.214	0.141	-0.60
	MI.co	-0.006	0.237	0.237	0.145	-0.30	-0.012	0.214	0.214	0.142	-0.60
I	WM	-0.318	0.239	0.398	0.330	-15.90	-0.288	0.223	0.364	0.302	-14.40
	EM	-0.293	0.236	0.376	0.301	-14.65	-0.276	0.223	0.355	0.287	-13.80
	MCAR.cc	-0.315	0.303	0.437	0.339	-15.75	-0.285	0.289	0.406	0.312	-14.25
	MAR.cc	-0.316	0.332	0.458	0.346	-15.80	-0.288	0.309	0.422	0.307	-14.40
	MI.ac	-0.325	0.208	0.386	0.323	-16.25	-0.304	0.200	0.364	0.304	-15.20
	MI.wi	-0.315	0.260	0.408	0.324	-15.75	-0.297	0.229	0.375	0.302	-14.85
	MI.co	-0.315	0.260	0.408	0.324	-15.75	-0.297	0.229	0.375	0.302	-14.85

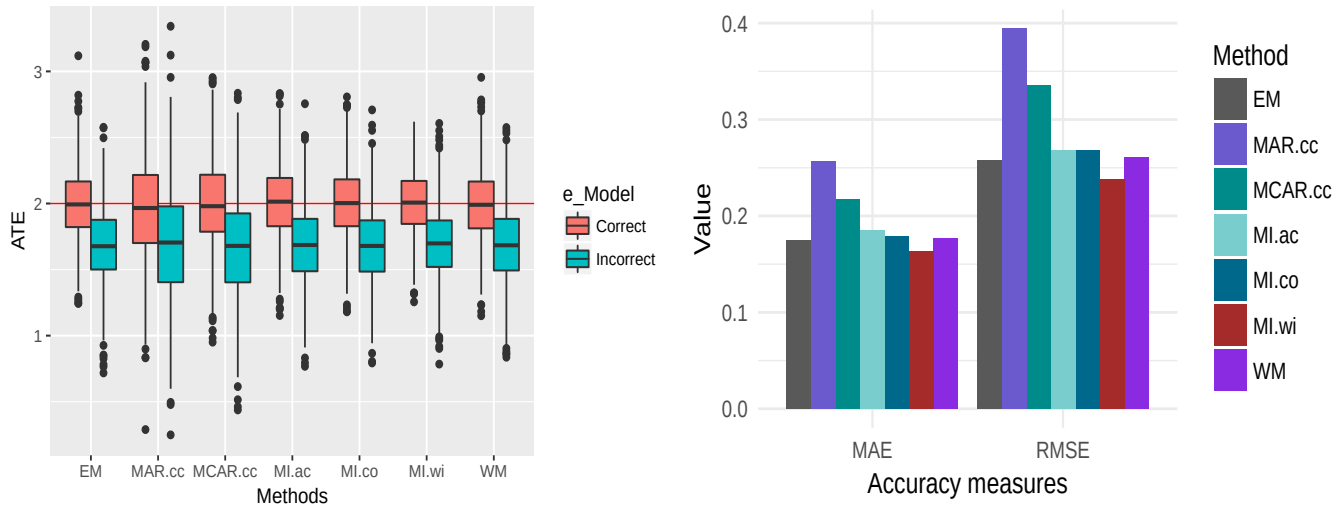


Figure 3.3: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for propensity score 1-to-1 matching (logit) estimator with continuous outcome.

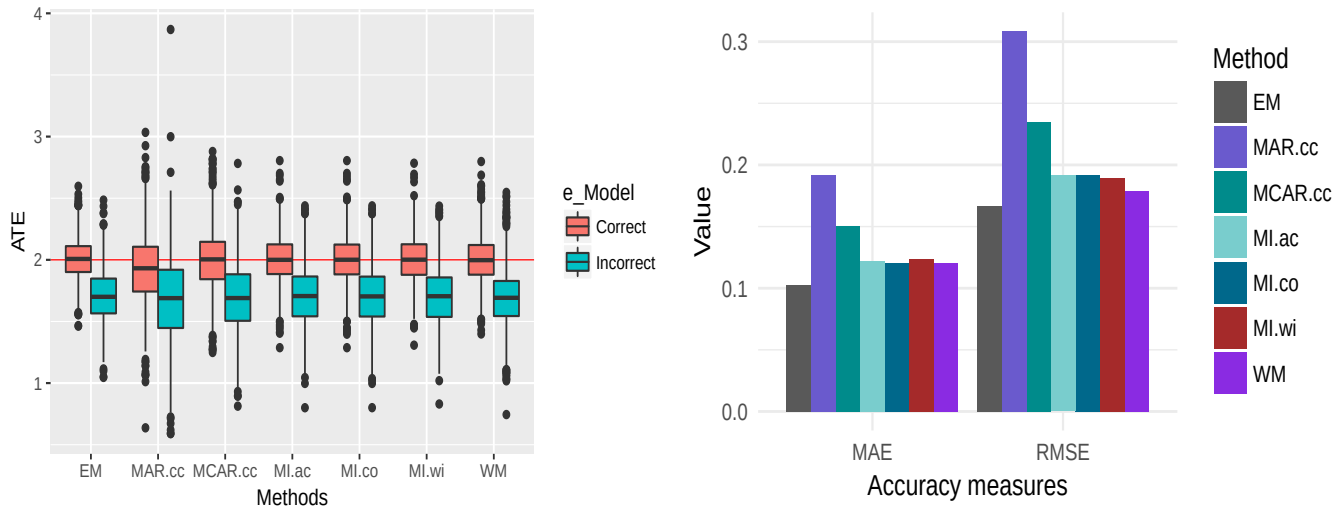


Figure 3.4: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for propensity score stratified estimator with continuous outcome.

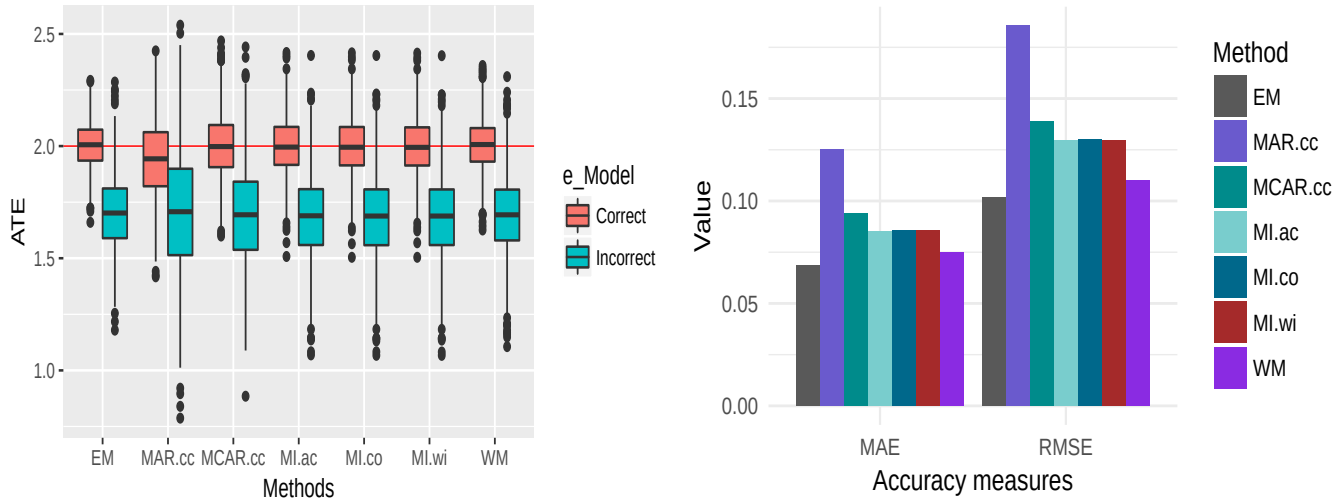


Figure 3.5: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for propensity score regression estimator with continuous outcome.

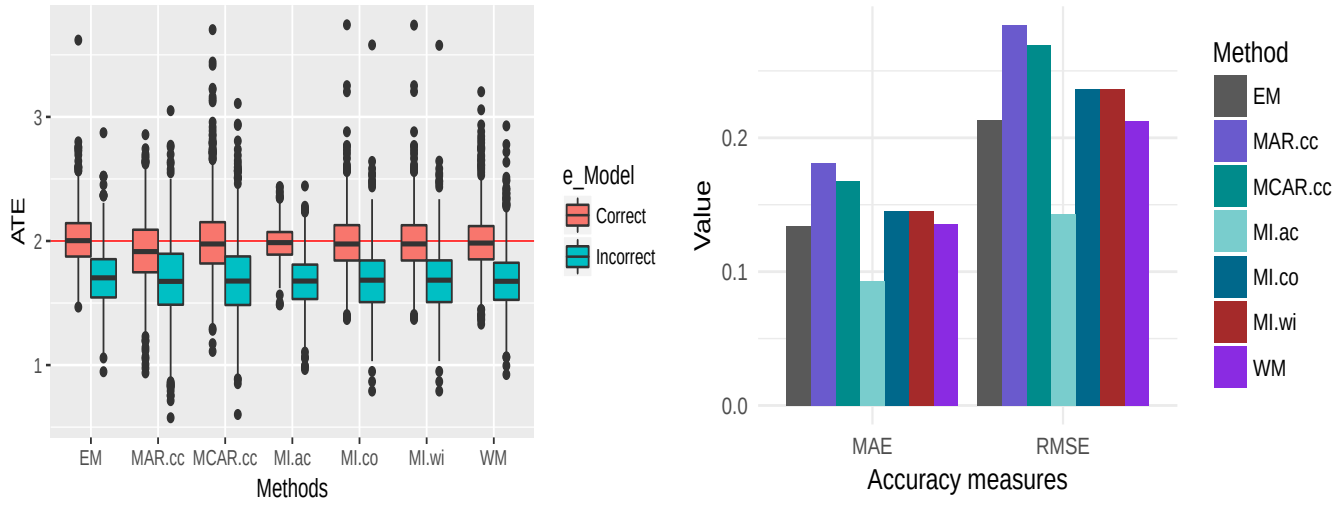


Figure 3.6: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for inverse probability weighting estimator with continuous outcome.

Figure 3.3-3.6 shows the boxplots of ATE estimates and barplots of accuracy measures for different PS based estimators with various methods for dealing missingness considering independent missing confounder (OR = 1). Boxplots depicts the unbiasedness for correct e-models and biasedness for incorrect $e-$models, while the barplots shows the accuracy of MI.ac method for IPW estimator (Figure 3.6), MI.wi for matching estimator (Figure 3.3) and EM method for stratified (3.4) and PS regression estimator (Figure 3.5).

The outcome regression and propensity score-based estimators are unbiased and efficient only when y- or e-model is correctly specified respectively. There is a question of biasedness and efficiency for the estimators to model misspecification. Augmented inverse probability weighted (AIPW) estimator provides the double protection for correct estimation of causal effects as long as either of the model is true.

Table 3.8: Bias and efficiency measures for AIPW estimator under different methods for dealing missingness for continuous outcome.

(y,e) model		OR = 1 True ATE = 2					OR = 4 True ATE = 2				
(AIPW)	Method	Bias	ESE	RMSE	MAE	%RB	Bias	ESE	RMSE	MAE	%RB
(C,C)	WM	0.003	0.124	0.124	0.082	0.15	0.001	0.125	0.125	0.080	0.05
	EM	-0.002	0.119	0.119	0.076	-0.10	-0.002	0.126	0.126	0.085	-0.10
	MCAR.cc	0.001	0.162	0.162	0.105	0.05	-0.002	0.165	0.165	0.107	-0.10
	MAR.cc	-0.047	0.195	0.201	0.134	-2.35	-0.044	0.197	0.202	0.136	-2.20
	MI.ac	-0.015	0.144	0.145	0.094	-0.75	-0.018	0.151	0.152	0.099	-0.90
	MI.wi	-0.001	0.143	0.143	0.093	-0.05	-0.004	0.149	0.149	0.096	-0.20
	MI.co	-0.015	0.144	0.145	0.094	-0.75	-0.018	0.151	0.152	0.099	-0.90
(C,I)	WM	0.003	0.123	0.123	0.081	0.15	0.001	0.124	0.124	0.079	0.05
	EM	-0.002	0.117	0.117	0.074	-0.10	-0.005	0.125	0.125	0.084	-0.25
	MCAR.cc	0.002	0.158	0.158	0.105	0.10	-0.001	0.161	0.161	0.107	-0.05
	MAR.cc	-0.046	0.195	0.200	0.135	-2.30	-0.044	0.196	0.200	0.137	-2.20
	MI.ac	-0.016	0.143	0.144	0.091	-0.80	-0.017	0.148	0.149	0.097	-0.85
	MI.wi	-0.002	0.142	0.142	0.090	-0.10	-0.004	0.146	0.146	0.095	-0.20
	MI.co	-0.016	0.143	0.144	0.091	-0.80	-0.017	0.148	0.149	0.098	-0.85
(I,C)	WM	-0.002	0.152	0.152	0.097	-0.10	-0.002	0.150	0.150	0.095	-0.10
	EM	-0.004	0.147	0.147	0.095	-0.20	0.000	0.147	0.147	0.096	0.00
	MCAR.cc	-0.006	0.195	0.195	0.123	-0.30	-0.005	0.196	0.196	0.126	-0.25
	MAR.cc	-0.072	0.223	0.234	0.156	-3.60	-0.066	0.225	0.234	0.151	-3.30
	MI.ac	-0.017	0.168	0.169	0.108	-0.85	-0.017	0.178	0.179	0.117	-0.85
	MI.wi	-0.002	0.166	0.166	0.106	-0.10	-0.005	0.176	0.176	0.116	-0.25
	MI.co	-0.017	0.168	0.169	0.108	-0.85	-0.017	0.179	0.179	0.117	-0.85
(I,I)	WM	-0.310	0.195	0.367	0.312	-15.50	-0.283	0.187	0.339	0.286	-14.15
	EM	-0.304	0.191	0.359	0.301	-15.20	-0.279	0.195	0.340	0.281	-13.95
	MCAR.cc	-0.310	0.250	0.398	0.316	-15.50	-0.275	0.251	0.372	0.287	-13.75
	MAR.cc	-0.296	0.305	0.425	0.317	-14.80	-0.268	0.291	0.396	0.293	-13.40
	MI.ac	-0.324	0.210	0.386	0.321	-16.20	-0.303	0.204	0.366	0.306	-15.15
	MI.wi	-0.310	0.210	0.374	0.310	-15.50	-0.288	0.203	0.353	0.291	-14.40
	MI.co	-0.324	0.210	0.387	0.322	-16.20	-0.303	0.205	0.366	0.306	-15.15

Table 3.8 displays the simulation results for AIPW estimator under different methods for dealing missingness. In either of the combination of y- and e-models (C,C), (C,I) and (I,C) all methods are unbiased and retain a desired value for RB within 5%, which indicates the double robustness. The MAR.cc method appears most variable, followed by MCAR.cc and EM appears least variable, while three multiple imputation methods are nearly equivalent in terms of variability. Based on the accuracy measures RMSE and MAE, the EM method dominates all other

methods. When the y-model is correct i.e., (C, C) or (C, I) there is a significant efficiency gain for all the methods compared to correct e-model only (I, C) or both models are incorrect (I, I). In the scenario when both models are incorrect (I, I), all the methods become biased, inefficient and inaccurate regardless the odds ratios between X_1 and X_2.

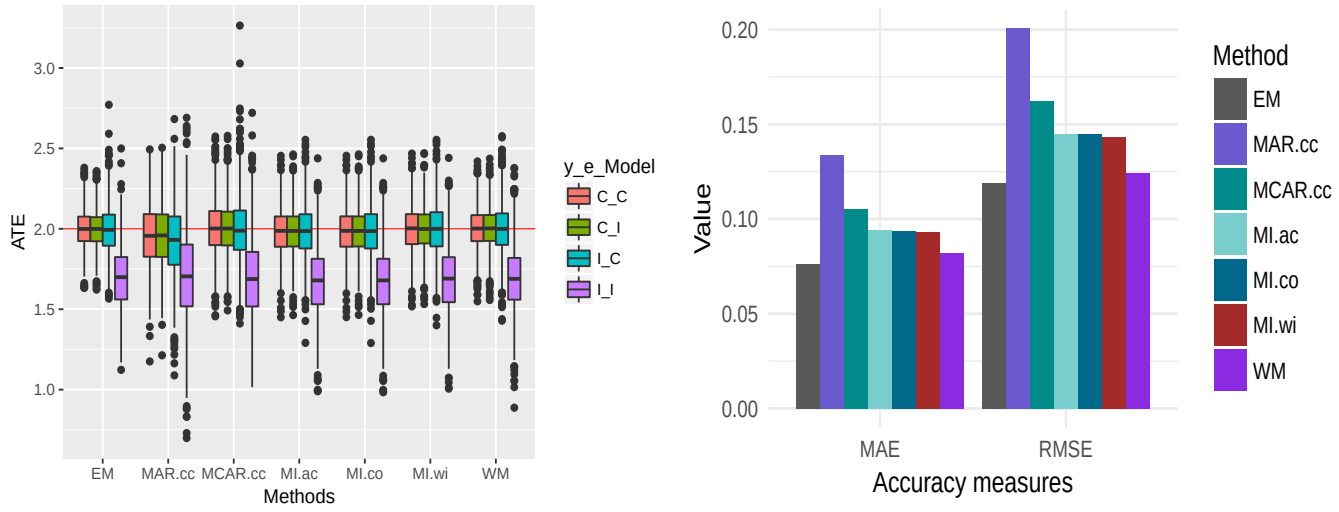


Figure 3.7: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for AIPW estimator with continuous outcome.

Figure 3.7 shows the boxplots and barplots for AIPW estimator under different methods for dealing missingness with independent missing confounder (OR = 1). From the boxplots it is evident that all the methods are unbiased except when the two models y- and e- are incorrect (I, I) implies the double robustness. Barplots depicts that the EM is the most accurate method compared to the other alternatives with a lowest values of MAE and RMSE.

In this study for different multiple imputation methods, we considered an imputation model where the outcome (Y), treatment (T) and all the confounders except missing one were used as predictors. Estimating casual effects for binary outcome, Leyrat et al. (2019) stated that multiple imputation methods are approximately unbiased as long as the outcome was included in the imputation model. Table 3.9 shows the bias values for three MI methods considering different imputation models. In all the scenarios, when the imputation model excludes either T or Y or both, all the MI methods under different estimators become biased. Much more care needs to be taken in defining imputation model, otherwise we will end with a biased estimate. Simulation results

suggests to include both treatment and outcome in the imputation model for continuous outcome.

Table 3.9: Bias in multiple imputation methods considering different imputation models for continuous outcome.

		Imputation model excludes					
Estimator	Method	T	Y	TY	T	Y	TY
Regression	MI.ac.wi.co	-0.336	0.359	0.477	-0.300	0.325	0.431
G-computation	MI.ac.wi	-0.281	0.424	0.500	-0.256	0.387	0.452
	MI.co	-0.347	0.288	0.438	-0.317	0.259	0.394
IPW	MI.ac	-0.318	0.237	0.395	-0.288	0.210	0.362
	MI.wi	-0.324	0.381	0.475	-0.296	0.335	0.437
	MI.co	-0.319	0.377	0.475	-0.292	0.331	0.436
PS Matching (1-1)	MI.ac	-0.310	0.415	0.481	-0.285	0.391	0.469
	MI.wi	-0.323	0.377	0.472	-0.294	0.365	0.447
	MI.co	-0.313	0.404	0.474	-0.288	0.382	0.454
PS Stratified	MI.ac	-0.313	0.413	0.502	-0.294	0.376	0.446
	MI.wi	-0.322	0.374	0.489	0.042	0.346	0.435
	MI.co	-0.316	0.401	0.495	-0.296	0.364	0.441
PS Regression	MI.ac	-0.327	0.435	0.500	-0.293	0.389	0.450
	MI.wi	-0.330	0.390	0.487	-0.295	0.350	0.439
	MI.co	-0.329	0.415	0.492	-0.294	0.371	0.443
AIPW	MI.ac	-0.315	0.250	0.407	-0.286	0.222	0.375
	MI.wi	-0.276	0.377	0.473	-0.249	0.339	0.436
	MI.co	-0.316	0.254	0.408	-0.288	0.226	0.376
		OR = 1			OR = 4		

The EM method results shown previously for different estimators, considered single but correct model for *y*- and *e*-, in calculating the expected value of the missing cnfounder. Table 3.10 displays the results for AIPW estimator with different specifications of *y*- and *e*- models in EM algorithm. It is obvious that a restricted version of double robust property can be maintained up to when we can specify a correct *y*-model in the EM algorithm. This is restricted to correct *y*-model in EM algorithm and need to specify either of the model correctly in the AIPW definition. In the other scenario, even both models correctly specified in AIPW definition but due to incorrect *y*-model in EM algorithm the estimator becomes biased. This limits the applicability of EM method under AIPW estimator.

Table 3.10: Augmented inverse probability weighted (AIPW) estimator with continuous outcome under EM algorithm considering single model (SM) for outcome regression and propensity score in weight calculation.

(y,e)- model in		Missing = 20%				Missing = 40%			
		OR = 1		OR = 4		OR = 1		OR = 4	
EM	AIPW	Bias	%RB	Bias	%RB	Bias	%RB	Bias	%RB
	(C,C)	-0.005	-0.23	-0.008	-0.41	-0.005	-0.26	-0.009	-0.45
(C,I)	(C,I)	-0.004	-0.22	-0.009	-0.45	-0.005	-0.27	-0.012	-0.60
	(I,C)	-0.015	-0.73	-0.012	-0.58	-0.008	-0.38	-0.016	-0.79
	(I,I)	-0.323	-16.13	-0.286	-14.29	-0.316	-15.80	-0.292	-14.59
	(C,C)	0.171	8.53	0.183	9.13	0.214	10.70	0.233	11.67
(I,C)	(C,I)	0.156	7.78	0.155	7.73	0.210	10.52	0.217	10.85
	(I,C)	0.163	8.13	0.182	9.10	0.207	10.35	0.233	11.65
	(I,I)	-0.104	-5.19	-0.089	-4.47	-0.011	-0.53	0.004	0.20
	(C,C)	0.158	7.92	0.191	9.56	0.223	11.13	0.240	11.98
(I,I)	(C,I)	0.144	7.22	0.164	8.21	0.219	10.94	0.222	11.12
	(I,C)	0.150	7.50	0.192	9.62	0.212	10.59	0.238	11.92
	(I,I)	-0.117	-5.85	-0.077	-3.83	-0.010	-0.50	0.011	0.53

The problem with single model can be overcome by considering multiple models for each y- and e- model in the EM algorithm. Table 3.11 shows the results for AIPW estimator under EM method where the expected value of missing confounder obtained with multiple model consideration namely three models for each y- and e-model. We consider three different models each y- and e- model among them one is correct and other two are incorrect:

Specification	y-model	e-model
Correct (C)	$Y \sim X_1 + X_2 + Z + T$	$T \sim X_1 + X_2 + Z$
Incorrect (I)	$Y \sim T + X_2 + Z$	$T \sim X_2 + Z$
Incorrect (I_1)	$Y \sim T + X_1 + Z$	$T \sim X_1 + Z$

Comparing the results under single model consideration in Table 3.10, we retain the original double robust property using multiple models. As long as either of the y- or e-model is correctly specified in AIPW definition, the multiple model consideration in EM method will provide unbiased estimate of treatment effects. This is obviously an advantage of using multiple models in EM algorithm, we need only to care about the correct specification of either $y-$ or $e-$ model under AIPW. The multiple model approach is better performed in moderate missing (40%) compared to

lower level of missing (20%) which is supported by the relative bias percentage values. Within the same level of missing we find that there is no significant improvements in terms of bias and variability for different odds ratios between Z and X_1. In both scenarios, single model or multiple model consideration, we see the necessity of defining y-model correctly to get unbiased and efficient estimates. The dominance of y-model in effect estimation is because the outcome is continuous. For a continuous outcome, the predicted values can vary within a wider range, which will much higher if the model specification is incorrect. AIPW estimator directly use the predicted values of the outcome in effect estimation. In incorrect situation, the high variability of the predicted outcomes will deteriorate the estimates. Similar results found for the usual regression estimator and G-estimation where the $y-$model is directly used for effect estimation.

Table 3.11: Bias and relative bias values for AIPW estimator with multiple models (MM) consideration in EM algorithm. A tick for inclusion and cross for exclusion of correct model in MM consideration.

(y,e)- model in		Missing = 20%				Missing = 40%			
		OR = 1		OR = 4		OR = 1		OR = 4	
EM	AIPW	Bias	%RB	Bias	%RB	Bias	%RB	Bias	%RB
	(C,C)	0.027	1.35	0.021	1.05	0.015	0.74	0.010	0.51
$(\surd,\surd)$	(C,I)	0.022	1.09	0.009	0.47	0.013	0.65	-0.006	-0.31
	(I,C)	0.027	1.35	0.025	1.23	0.007	0.35	0.008	0.41
	(I,I)	-0.268	-13.40	-0.254	-12.69	-0.278	-13.88	-0.271	-13.57
	(C,C)	0.026	1.28	0.023	1.16	0.014	0.71	0.004	0.22
$(\surd,\times)$	(C,I)	0.020	1.02	0.011	0.54	0.012	0.61	-0.011	-0.56
	(I,C)	0.020	1.00	0.027	1.34	0.009	0.45	0.004	0.20
	(I,I)	-0.281	-14.05	-0.255	-12.76	-0.283	-14.17	-0.275	-13.76
	(C,C)	0.058	2.92	0.069	3.44	0.055	2.76	0.043	2.16
$(\times,\surd)$	(C,I)	0.049	2.44	0.045	2.27	0.049	2.44	0.016	0.79
	(I,C)	0.056	2.78	0.071	3.54	0.049	2.46	0.046	2.32
	(I,I)	-0.241	-12.05	-0.214	-10.72	-0.233	-11.67	-0.245	-12.23
	(C,C)	0.058	2.91	0.069	3.46	0.041	2.06	0.041	2.04
$(\times,\times)$	(C,I)	0.047	2.33	0.045	2.26	0.035	1.74	0.016	0.81
	(I,C)	0.053	2.65	0.074	3.71	0.039	1.95	0.043	2.14
	(I,I)	-0.243	-12.14	-0.221	-11.04	-0.239	-11.95	-0.238	-11.92

Binary outcome

Table 3.12-3.17 displays the simulation results for different estimators with binary outcome under various methods for dealing missingness. For binary outcome, the average treatment effect $\Delta = E[Y(1)] - E[Y(0)] = Pr[Y(1) = 1] - Pr[Y(0) = 1]$ is the causal risk difference. The true risk difference for a given model can be obtained by, $\Delta = \int_{-\infty}^{\infty} Pr(Y = 1|T = 1, X = x)f(x)dx - \int_{-\infty}^{\infty} Pr(Y = 1|T = 0, X = x)f(x)dx$ or approximated by large number simulation. In our setting, we obtain the true risk difference for the given model with odds ratio 1 and 4 as 0.3343 and 0.3536 respectively. The performance of different methods for dealing missingness is compared based on the bias, ESE, and accuracy measures RMSE, MAE and RB in percentage. Overall, all methods for dealing missingness provides unbiased estimates of ATE when y-model is correct for regression based estimators and e-model is correct (C) for PS based estimators. However, they become biased for incorrect (I) model specification. There is no significant improvement in the estimation accuracy for considering highly associated covariate (X_1) with missing confounder (Z). The EM method is found unbiased, most efficient and accurate for regression estimator, G-estimation, IPW, stratified and PS regression estimators. Whereas, MI.wi is the best method for PS (logit) matching estimator.

Table 3.12: Bias and efficiency measures for regression estimator under different methods for dealing missingness for binary outcome.

		OR = 1					OR = 4				
y-		True ATE = 0.3343					True ATE = 0.3536				
model	Method	Bias	ESE	RMSE	MAE	%RB	Bias	ESE	RMSE	MAE	%RB
C	WM	0.001	0.040	0.040	0.028	0.33	-0.001	0.043	0.043	0.028	-0.23
	EM	0.001	0.039	0.039	0.025	0.29	-0.001	0.040	0.040	0.027	-0.27
	MCAR.cc	0.001	0.054	0.054	0.035	0.27	-0.002	0.051	0.051	0.036	-0.46
	MAR.cc	0.004	0.063	0.063	0.041	1.17	0.004	0.064	0.065	0.044	1.21
	MI.ac.wi	-0.002	0.043	0.043	0.030	-0.50	0.001	0.043	0.043	0.030	0.19
	MI.co	0.005	0.044	0.045	0.030	1.50	0.007	0.044	0.045	0.031	2.10
I	WM	-0.027	0.043	0.051	0.034	-8.19	-0.026	0.046	0.053	0.034	-7.28
	EM	-0.028	0.043	0.051	0.035	-8.31	-0.026	0.044	0.051	0.035	-7.42
	MCAR.cc	-0.026	0.057	0.063	0.041	-7.75	-0.026	0.056	0.062	0.043	-7.47
	MAR.cc	-0.014	0.068	0.069	0.046	-4.28	-0.012	0.069	0.070	0.046	-3.30
	MI.ac.wi	-0.029	0.046	0.054	0.036	-8.69	-0.024	0.045	0.051	0.033	-6.81
	MI.co	-0.022	0.047	0.052	0.035	-6.47	-0.020	0.045	0.049	0.033	-5.68

Table 3.13: Bias and efficiency measures for G-estimation under different methods for dealing missingness for binary outcome.

		OR = 1					OR = 4				
y-		True ATE = 0.3343					True ATE = 0.3536				
model	Method	Bias	ESE	RMSE	MAE	%RB	Bias	ESE	RMSE	MAE	%RB
C	WM	0.001	0.043	0.043	0.030	0.40	-0.001	0.045	0.045	0.029	-0.33
	EM	0.001	0.041	0.041	0.026	0.44	-0.001	0.043	0.043	0.028	-0.27
	MCAR.cc	0.000	0.056	0.056	0.039	0.11	-0.002	0.055	0.055	0.037	-0.50
	MAR.cc	0.003	0.071	0.071	0.046	0.79	0.004	0.070	0.070	0.050	1.19
	MI.ac.wi	-0.002	0.046	0.046	0.031	-0.67	-0.000	0.045	0.045	0.030	-0.01
	MI.co	0.005	0.047	0.047	0.033	1.38	0.007	0.046	0.047	0.032	1.88
I	WM	-0.028	0.046	0.054	0.036	-8.28	-0.026	0.048	0.055	0.037	-7.48
	EM	-0.028	0.044	0.053	0.037	-8.41	-0.027	0.047	0.054	0.037	-7.58
	MCAR.cc	-0.027	0.059	0.065	0.042	-8.09	-0.027	0.059	0.065	0.042	-7.60
	MAR.cc	-0.018	0.076	0.078	0.052	-5.46	-0.013	0.074	0.075	0.052	-3.65
	MI.ac.wi	-0.030	0.048	0.057	0.038	-9.04	-0.025	0.047	0.053	0.036	-7.12
	MI.co	-0.023	0.050	0.055	0.036	-6.75	-0.021	0.047	0.052	0.034	-5.92

Table 3.14: Bias and efficiency measures for propensity score 1-to-1 matching (logit) estimator under different methods for dealing missingness for binary outcome.

		OR = 1					OR = 4				
e-		True ATE = 0.3343					True ATE = 0.3536				
model	Method	Bias	ESE	RMSE	MAE	%RB	Bias	ESE	RMSE	MAE	%RB
C	WM	0.002	0.055	0.055	0.037	0.69	0.004	0.055	0.055	0.039	1.16
	EM	0.008	0.054	0.055	0.036	2.53	0.009	0.052	0.053	0.035	2.40
	MCAR.cc	0.009	0.071	0.072	0.050	2.58	0.006	0.073	0.073	0.050	1.64
	MAR.cc	0.006	0.082	0.082	0.054	1.71	0.009	0.089	0.089	0.061	2.63
	MI.ac	0.011	0.058	0.059	0.043	3.32	0.009	0.058	0.059	0.040	2.56
	MI.wi	0.005	0.051	0.051	0.034	1.55	0.005	0.051	0.051	0.033	1.50
	MI.co	0.009	0.057	0.057	0.040	2.55	0.009	0.057	0.058	0.040	2.46
I	WM	-0.030	0.058	0.065	0.043	-8.88	-0.022	0.061	0.065	0.042	-6.34
	EM	-0.022	0.057	0.061	0.041	-6.59	-0.019	0.058	0.061	0.040	-5.32
	MCAR.cc	-0.022	0.071	0.074	0.050	-6.55	-0.023	0.074	0.078	0.052	-6.63
	MAR.cc	-0.017	0.087	0.089	0.060	-4.96	-0.013	0.093	0.094	0.058	-3.62
	MI.ac	-0.019	0.060	0.063	0.042	-5.78	-0.014	0.059	0.061	0.041	-3.95
	MI.wi	-0.026	0.055	0.061	0.041	-7.66	-0.023	0.052	0.057	0.039	-6.40
	MI.co	-0.021	0.059	0.063	0.043	-6.29	-0.016	0.059	0.061	0.043	-4.47

Table 3.15: Bias and efficiency measures for propensity score stratified estimator (strata = 10) under different methods for dealing missingness for binary outcome.

		OR = 1					OR = 4				
e-		True ATE = 0.3343					True ATE = 0.3536				
model	Method	Bias	ESE	RMSE	MAE	%RB	Bias	ESE	RMSE	MAE	%RB
C	WM	0.005	0.047	0.047	0.032	1.46	0.003	0.049	0.049	0.031	0.93
	EM	0.009	0.046	0.047	0.031	2.56	0.007	0.048	0.048	0.031	1.92
	MCAR.cc	0.007	0.061	0.061	0.042	2.06	0.006	0.060	0.060	0.042	1.71
	MAR.cc	0.005	0.076	0.076	0.053	1.64	0.010	0.075	0.076	0.054	2.71
	MI.ac	0.014	0.049	0.051	0.034	4.11	0.011	0.050	0.052	0.034	3.19
	MI.wi	0.007	0.049	0.049	0.032	2.21	0.006	0.050	0.050	0.034	1.57
	MI.co	0.011	0.049	0.051	0.033	3.40	0.009	0.050	0.051	0.033	2.66
I	WM	-0.025	0.049	0.055	0.036	-7.53	-0.024	0.052	0.057	0.039	-6.86
	EM	-0.022	0.049	0.054	0.037	-6.72	-0.019	0.050	0.053	0.036	-5.31
	MCAR.cc	-0.022	0.065	0.069	0.044	-6.64	-0.022	0.063	0.066	0.044	-6.33
	MAR.cc	-0.017	0.078	0.080	0.054	-5.11	-0.011	0.079	0.080	0.053	-3.09
	MI.ac	-0.017	0.052	0.055	0.036	-5.07	-0.014	0.052	0.054	0.035	-3.92
	MI.wi	-0.023	0.051	0.056	0.037	-6.74	-0.021	0.051	0.055	0.036	-6.07
	MI.co	-0.019	0.052	0.056	0.037	-5.81	-0.016	0.052	0.055	0.035	-4.57

Table 3.16: Bias and efficiency measures for propensity score regression estimator under different methods for dealing missingness for binary outcome.

		OR = 1					OR = 4				
e-		True ATE = 0.3343					True ATE = 0.3536				
model	Method	Bias	ESE	RMSE	MAE	%RB	Bias	ESE	RMSE	MAE	%RB
C	WM	0.004	0.042	0.042	0.029	1.14	0.003	0.043	0.043	0.028	0.72
	EM	0.008	0.041	0.042	0.027	2.38	0.006	0.042	0.042	0.028	1.80
	MCAR.cc	0.004	0.055	0.055	0.037	1.28	0.002	0.053	0.053	0.036	0.61
	MAR.cc	0.008	0.064	0.064	0.042	2.36	0.008	0.065	0.066	0.046	2.31
	MI.ac	0.009	0.044	0.045	0.030	2.73	0.012	0.044	0.045	0.032	3.30
	MI.wi	0.003	0.045	0.045	0.031	0.96	0.006	0.044	0.045	0.031	1.71
	MI.co	0.007	0.044	0.045	0.030	2.02	0.009	0.044	0.045	0.031	2.65
I	WM	-0.026	0.044	0.051	0.035	-7.75	-0.024	0.046	0.052	0.035	-6.79
	EM	-0.023	0.044	0.049	0.035	-6.85	-0.018	0.045	0.048	0.033	-5.23
	MCAR.cc	-0.024	0.058	0.063	0.042	-7.26	-0.025	0.057	0.062	0.042	-6.95
	MAR.cc	-0.013	0.068	0.070	0.046	-3.80	-0.010	0.069	0.070	0.046	-2.90
	MI.ac	-0.021	0.047	0.051	0.034	-6.41	-0.015	0.045	0.047	0.032	-4.21
	MI.wi	-0.026	0.047	0.054	0.036	-7.87	-0.021	0.045	0.050	0.033	-6.06
	MI.co	-0.024	0.047	0.052	0.035	-7.10	-0.017	0.045	0.048	0.032	-4.85

Table 3.17: Bias and efficiency measures for inverse probability weighting estimator under different methods for dealing missingness for binary outcome.

e-model	Method	OR = 1, True ATE = 0.3343 Bias	ESE	RMSE	MAE	%RB	OR = 4, True ATE = 0.3536 Bias	ESE	RMSE	MAE	%RB
C	WM	0.001	0.048	0.048	0.033	0.34	-0.003	0.050	0.050	0.033	-0.72
	EM	0.000	0.048	0.048	0.031	-0.11	-0.003	0.048	0.048	0.033	-0.77
	MCAR.cc	0.000	0.063	0.063	0.045	0.00	-0.001	0.060	0.060	0.040	-0.19
	MAR.cc	0.000	0.074	0.074	0.048	0.11	0.003	0.073	0.073	0.050	0.79
	MI.ac	-0.002	0.050	0.050	0.034	-0.55	0.001	0.049	0.049	0.033	0.39
	MI.wi	-0.001	0.050	0.050	0.035	-0.40	0.002	0.049	0.049	0.033	0.58
	MI.co	-0.002	0.050	0.050	0.034	-0.50	0.002	0.049	0.049	0.033	0.49
I	WM	-0.027	0.051	0.058	0.040	-8.20	-0.028	0.053	0.060	0.039	-8.00
	EM	-0.029	0.050	0.058	0.040	-8.56	-0.028	0.050	0.057	0.039	-7.93
	MCAR.cc	-0.028	0.066	0.072	0.048	-8.23	-0.027	0.064	0.069	0.048	-7.70
	MAR.cc	-0.020	0.078	0.081	0.054	-5.96	-0.015	0.076	0.078	0.052	-4.27
	MI.ac	-0.029	0.052	0.060	0.041	-8.70	-0.025	0.050	0.056	0.037	-6.96
	MI.wi	-0.029	0.053	0.060	0.041	-8.75	-0.025	0.050	0.056	0.037	-7.01
	MI.co	-0.029	0.053	0.060	0.041	-8.79	-0.025	0.050	0.056	0.038	-7.02

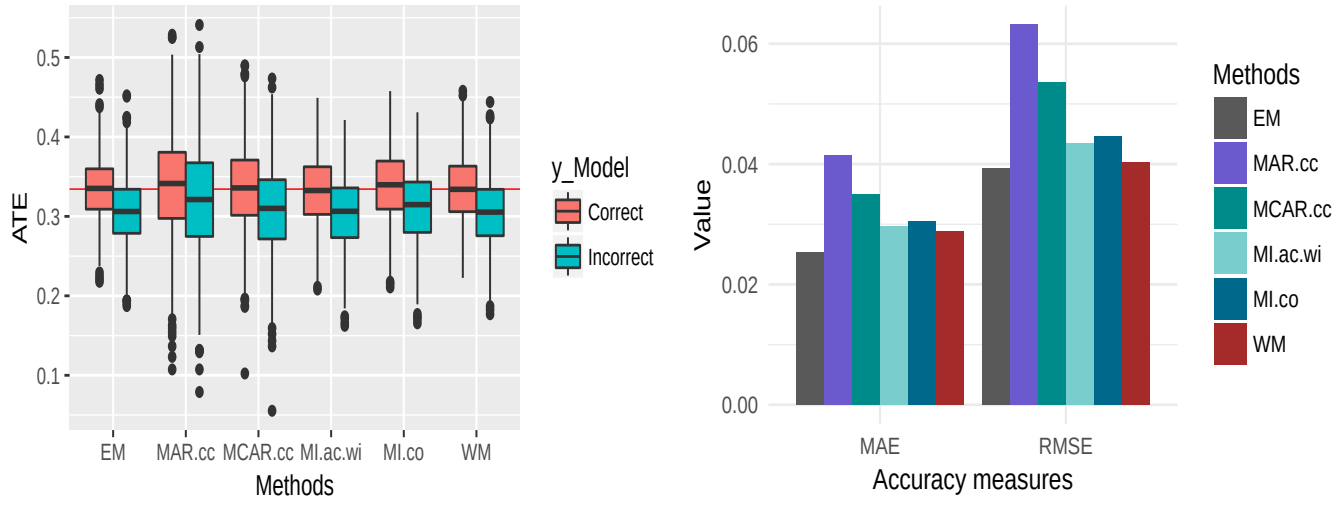


Figure 3.8: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for regression estimator with binary outcome.

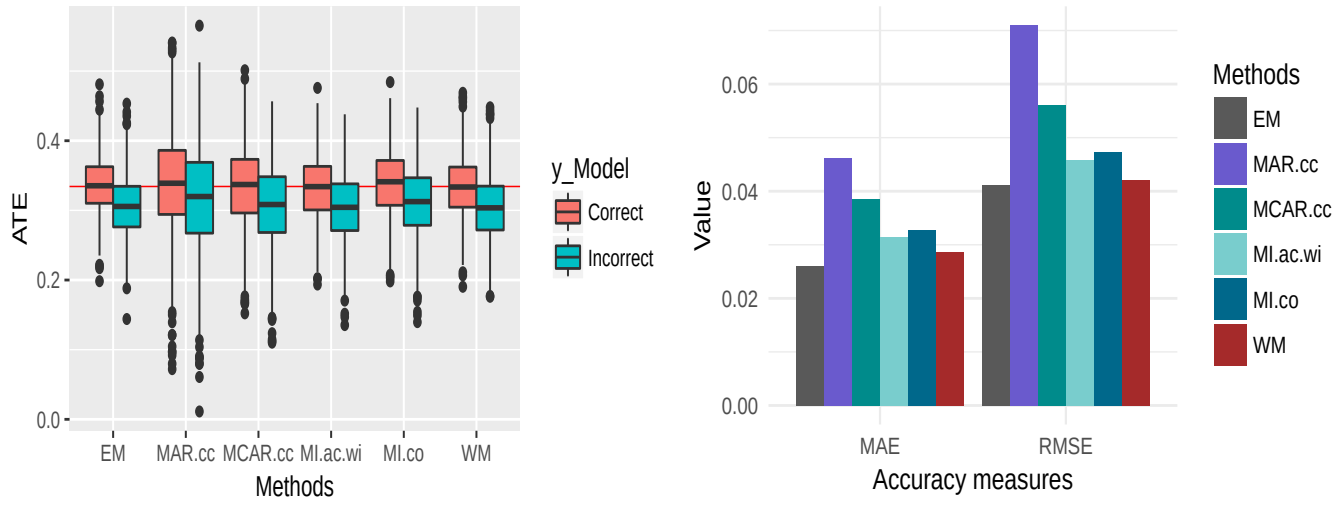


Figure 3.9: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for G-estimation with binary outcome.

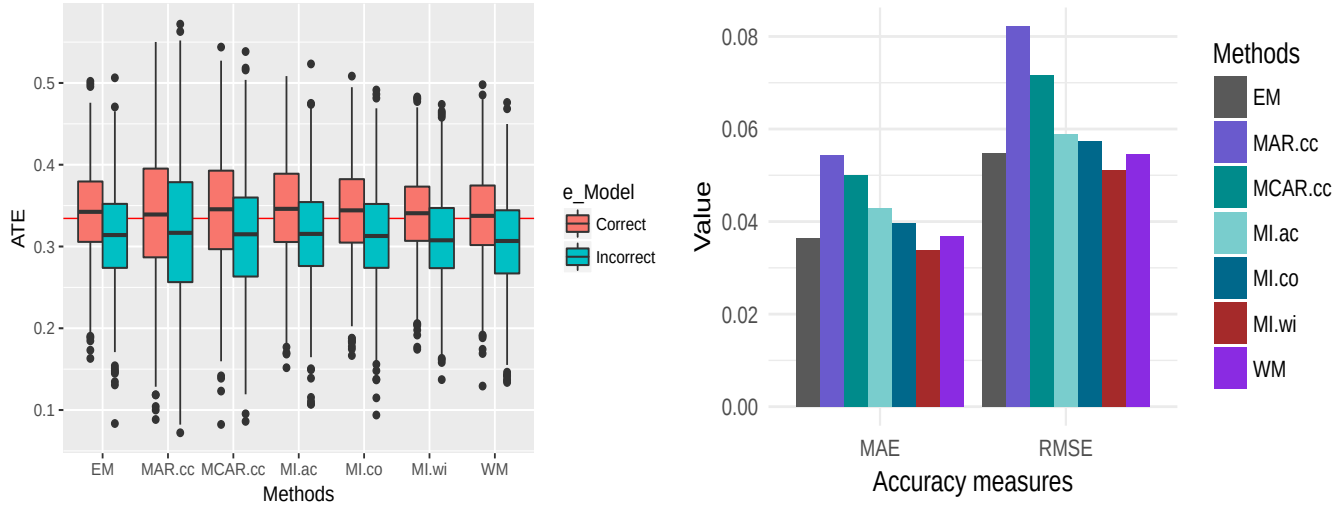


Figure 3.10: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for propensity score 1-to-1 matching (logit) estimator with binary outcome.

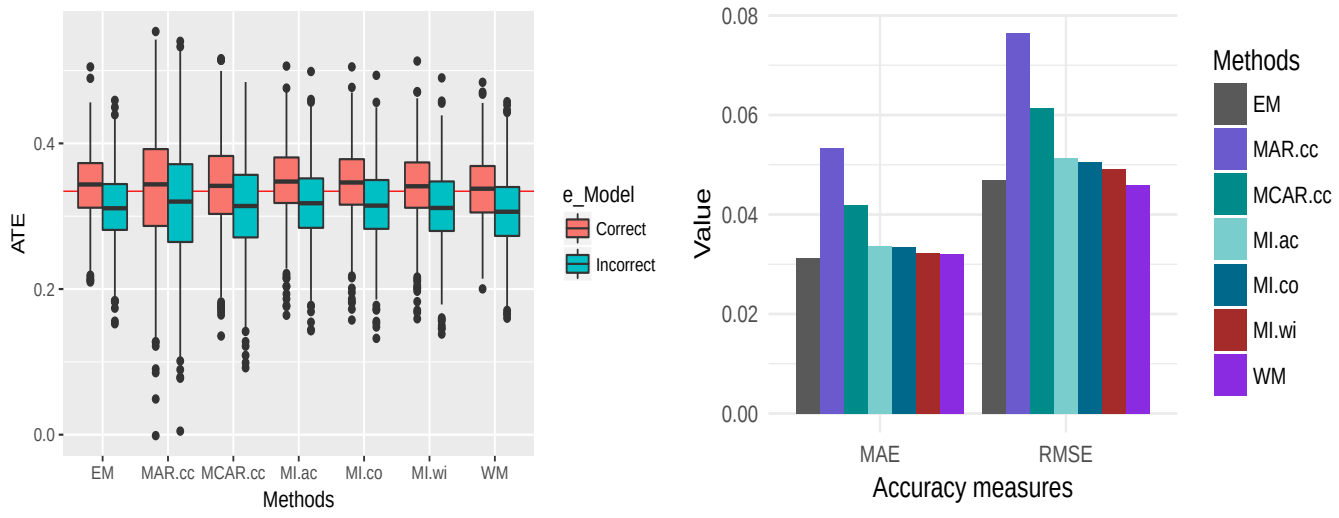


Figure 3.11: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for propensity score stratified estimator with binary outcome.

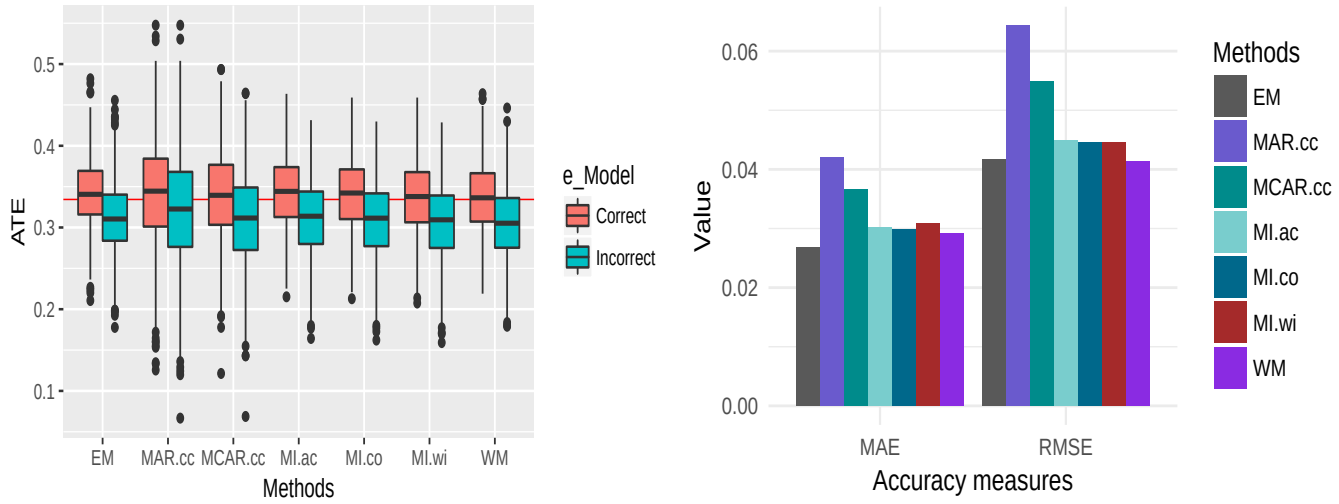


Figure 3.12: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for propensity score regression estimator with binary outcome.

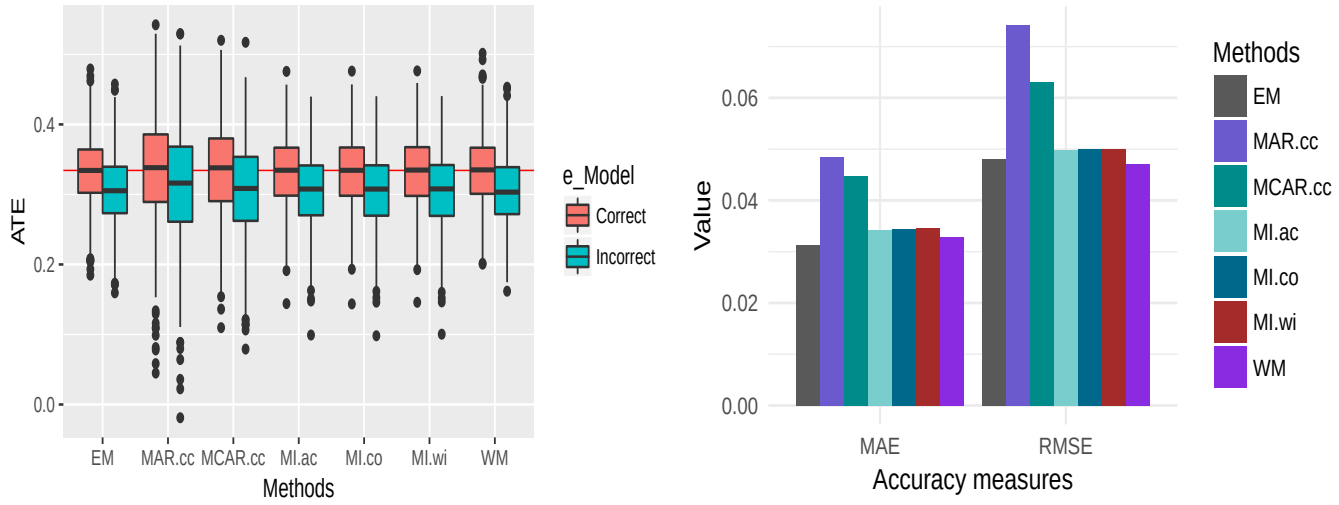


Figure 3.13: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for inverse probability weighting estimator with binary outcome.

Figure 3.8-3.13 shows the boxplots and barplots for different estimators under various methods for dealing missingness with independent missing confounder (OR = 1). Boxplots reveals the fact that under correct model specification every method is unbiased, the median line is located close to the true risk difference indicated by horizontal red line and biased for incorrect model. Barplots displays the accuracy measures RMSE and MAE for different methods for dealing missingness. It is evident from the barplots that EM is the best method to deal missing confounder for all estimators except the PS (logit) matching. The MI.wi is best performed in matching estimator.

Table 3.18: Bias and efficiency measures for AIPW estimator under different methods for dealing missingness for binary outcome.

(y,e)-Model (AIPW)	Method	OR = 1, True ATE = 0.3343					OR =4, True ATE = 0.3536				
		Bias	ESE	RMSE	MAE	%RB	Bias	ESE	RMSE	MAE	%RB
(C,C)	WM	0.002	0.045	0.045	0.031	0.73	-0.002	0.048	0.048	0.031	-0.44
	EM	-0.012	0.046	0.047	0.031	-3.48	-0.013	0.046	0.047	0.031	-3.68
	MCAR.cc	0.001	0.061	0.061	0.041	0.29	-0.001	0.057	0.057	0.039	-0.19
	MAR.cc	0.003	0.073	0.073	0.048	1.01	0.004	0.072	0.072	0.050	1.26
	MI.ac	-0.012	0.051	0.052	0.035	-3.66	-0.008	0.050	0.050	0.033	-2.38
	MI.wi	-0.002	0.049	0.049	0.033	-0.57	0.001	0.048	0.048	0.032	0.29
	MI.co	-0.014	0.051	0.053	0.035	-4.13	-0.010	0.050	0.051	0.033	-2.84
(C,I)	WM	0.002	0.045	0.045	0.029	0.74	-0.001	0.048	0.048	0.031	-0.41
	EM	-0.012	0.045	0.047	0.031	-3.62	-0.017	0.046	0.049	0.031	-4.78
	MCAR.cc	0.001	0.061	0.061	0.041	0.19	-0.001	0.058	0.058	0.040	-0.26
	MAR.cc	0.004	0.073	0.073	0.048	1.07	0.005	0.072	0.072	0.049	1.35
	MI.ac	-0.012	0.051	0.052	0.036	-3.67	-0.012	0.050	0.051	0.033	-3.33
	MI.wi	-0.002	0.048	0.048	0.033	-0.46	0.001	0.048	0.048	0.031	0.29
	MI.co	-0.013	0.051	0.053	0.036	-3.90	-0.013	0.050	0.052	0.033	-3.63
(I,C)	WM	0.002	0.046	0.046	0.031	0.65	-0.002	0.049	0.049	0.031	-0.53
	EM	-0.013	0.047	0.048	0.031	-4.03	-0.011	0.046	0.048	0.031	-3.10
	MCAR.cc	0.001	0.061	0.061	0.041	0.27	-0.000	0.058	0.058	0.040	-0.05
	MAR.cc	0.002	0.074	0.074	0.050	0.61	0.004	0.073	0.073	0.050	1.24
	MI.ac	-0.013	0.052	0.053	0.035	-4.01	-0.006	0.050	0.050	0.035	-1.68
	MI.wi	-0.002	0.049	0.049	0.034	-0.57	0.002	0.048	0.048	0.032	0.47
	MI.co	-0.014	0.052	0.054	0.036	-4.31	-0.006	0.050	0.050	0.034	-1.77
(I,I)	WM	-0.026	0.049	0.055	0.037	-7.90	-0.028	0.052	0.059	0.037	-7.79
	EM	-0.041	0.049	0.064	0.044	-12.36	-0.039	0.049	0.063	0.043	-10.90
	MCAR.cc	-0.026	0.064	0.069	0.045	-7.91	-0.026	0.062	0.068	0.046	-7.45
	MAR.cc	-0.017	0.077	0.079	0.053	-5.11	-0.013	0.075	0.076	0.052	-3.58
	MI.ac	-0.040	0.054	0.067	0.045	-11.93	-0.034	0.051	0.061	0.040	-9.48
	MI.wi	-0.029	0.051	0.059	0.039	-8.78	-0.024	0.049	0.055	0.036	-6.92
	MI.co	-0.042	0.054	0.068	0.046	-12.43	-0.035	0.051	0.062	0.041	-9.83

Table 3.18 compares the simulation results for AIPW estimator under binary outcomes with 40% missing in confounder. When both y- and e-model or either of the model is correct all methods are unbiased, but for EM, MI.ac and MI.co there is a slightly higher value of bias. However, the absolute values of bias ranges between 0.012-0.014, which seems not too much far away from zero. The relative bias percentage (%RB) which is not greater than 5, indicates that these estimators are nearly unbiased. Here the effect of interest is causal risk difference which ranges between -1 to

1, so its better to compare relative bias rather than the bias only. In all scenarios, EM method appears least variable and MAR.cc is the most, while three multiple imputation methods are nearly equivalent in terms of variability. Similar conclusion can be made based on accuracy measures RMSE and MAE. Overall, we can conclude that EM is better performed under AIPW estimator at 40% missing.

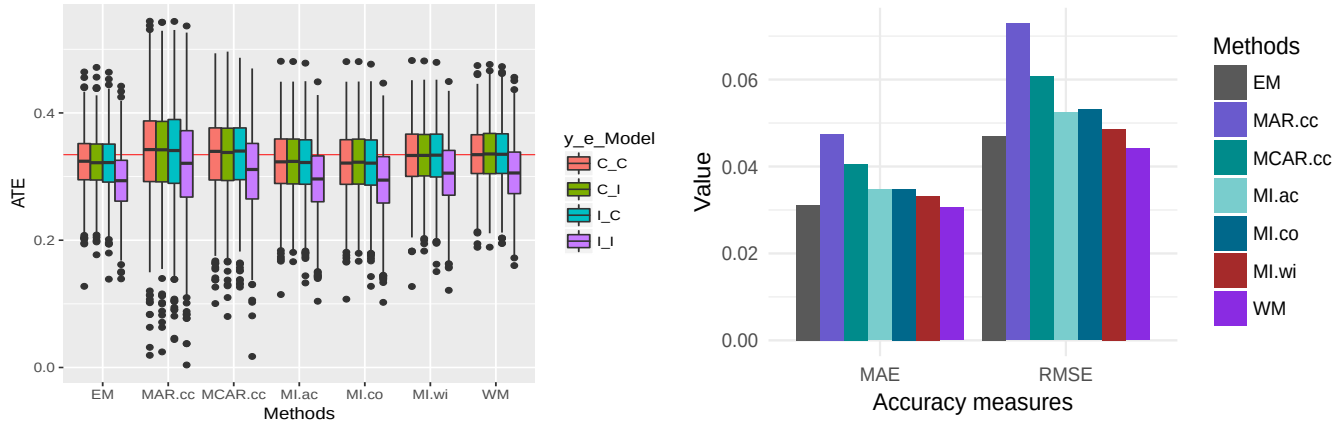


Figure 3.14: Boxplots of average treatment effect (ATE) estimates and barplots of efficiency measures based on 1000 MCMC simulations for AIPW estimator with binary outcome.

Figure 3.14 shows the boxplots of ATE estimates and barplots of accuracy measure based on 1000 MCMC simulations for AIPW estimator with independent missing confounder (OR = 1). Comparing the median line of the boxplots and the true effect line indicated by red color, we can see the slightly higher values of bias (but ignorable) for EM, MI.ac and MI.co which support the results from table. Barplots depicts the accuracy of EM method compared to all other methods for dealing missing confounder.

Table 3.19: Augmented inverse probability weighted (AIPW) estimator with binary outcome under EM algorithm considering single model (SM) for outcome regression and propensity score in weight calculation.

(y,e)- model in		Missing = 20%				Missing = 40%			
		OR = 1		OR = 4		OR = 1		OR = 4	
EM	AIPW	Bias	%RB	Bias	%RB	Bias	%RB	Bias	%RB
	(C,C)	-0.005	-1.44	-0.004	-1.10	-0.013	-3.92	-0.013	-3.78
(C,I)	(C,I)	-0.005	-1.46	-0.006	-1.58	-0.014	-4.16	-0.017	-4.84
	(I,C)	-0.006	-1.66	-0.003	-0.89	-0.015	-4.34	-0.011	-3.20
	(I,I)	-0.033	-9.94	-0.031	-8.64	-0.043	-12.93	-0.039	-11.10
	(C,C)	-0.001	-0.15	-0.003	-0.85	-0.005	-1.53	-0.009	-2.52
(I,C)	(C,I)	-0.001	-0.43	-0.006	-1.66	-0.006	-1.66	-0.014	-3.95
	(I,C)	-0.001	-0.41	-0.002	-0.55	-0.007	-2.00	-0.007	-2.11
	(I,I)	-0.029	-8.67	-0.031	-8.68	-0.033	-9.81	-0.037	-10.60
	(C,C)	-0.001	-0.16	-0.004	-1.16	-0.006	-1.77	-0.010	-2.85
(I,I)	(C,I)	-0.002	-0.49	-0.007	-1.89	-0.006	-1.89	-0.015	-4.25
	(I,C)	-0.001	-0.41	-0.003	-0.81	-0.008	-2.25	-0.009	-2.58
	(I,I)	-0.030	-9.12	-0.032	-8.92	-0.034	-10.13	-0.040	-11.22

Table 3.20: Bias and relative bias values for AIPW estimator with multiple models (MM) consideration in EM algorithm. A tick for inclusion and cross for exclusion of correct model in MM consideration.

(y,e)- model in		Missing = 20%				Missing = 40%			
		OR = 1		OR = 4		OR = 1		OR = 4	
EM	AIPW	Bias	%RB	Bias	%RB	Bias	%RB	Bias	%RB
	(C,C)	-0.004	-1.30	-0.002	-0.60	-0.010	-3.11	-0.007	-2.06
$(\surd,\surd)$	(C,I)	-0.005	-1.42	-0.004	-1.22	-0.011	-3.30	-0.012	-3.31
	(I,C)	-0.006	-1.70	-0.001	-0.25	-0.013	-3.90	-0.007	-1.95
	(I,I)	-0.033	-9.98	-0.029	-8.25	-0.041	-12.19	-0.037	-10.37
	(C,C)	-0.006	-1.68	-0.002	-0.45	-0.007	-2.20	-0.008	-2.18
$(\surd,\times)$	(C,I)	-0.006	-1.76	-0.004	-1.08	-0.008	-2.36	-0.012	-3.37
	(I,C)	-0.007	-2.15	-0.000	-0.06	-0.010	-3.10	-0.006	-1.79
	(I,I)	-0.035	-10.57	-0.028	-7.81	-0.038	-11.29	-0.035	-9.76
	(C,C)	0.000	0.06	-0.001	-0.28	-0.006	-1.82	-0.003	-0.96
$(\times,\surd)$	(C,I)	-0.000	-0.03	-0.003	-0.93	-0.007	-2.07	-0.008	-2.18
	(I,C)	-0.001	-0.23	-0.001	-0.20	-0.009	-2.83	-0.002	-0.50
	(I,I)	-0.029	-8.54	-0.029	-8.29	-0.036	-10.75	-0.032	-8.98
	(C,C)	-0.002	-0.73	-0.005	-1.29	-0.006	-1.82	-0.008	-2.28
$(\times,\times)$	(C,I)	-0.003	-0.86	-0.007	-1.94	-0.007	-2.06	-0.012	-3.49
	(I,C)	-0.003	-1.01	-0.003	-0.92	-0.009	-2.74	-0.007	-1.89
	(I,I)	-0.031	-9.36	-0.032	-8.93	-0.037	-11.02	-0.035	-10.01

Table 3.19 and 3.20 shows the simulation results for AIPW estimator under EM method considering single model and multiple models respectively. We consider the same set of three models for outcome regression and PS as in continuous outcomes. Similar conclusion can be made that EM method is better performed for lower level of missing (20%). For higher level of missing (40%), EM method incurred slightly higher but ignorable bias with a relative bias percentage less than 5%. For binary outcome, multiple model consideration in EM method does not improve the results greatly. However, multiple models consideration in the estimation process will increase the likelihood of correct specification for unknown true models.

Table 3.21: Bias in multiple imputation methods considering different imputation models for binary outcome.

		Imputation model excludes					
Estimator	Method	T	Y	TY	T	Y	TY
Regression	MI.ac.wi	-0.005	0.012	0.023	-0.005	0.011	0.022
	MI.co	0.001	0.016	0.027	0.001	0.014	0.026
G-estimation	MI.ac.wi	-0.005	0.020	0.025	-0.007	0.017	0.023
	MI.co	0.001	0.025	0.031	-0.000	0.023	0.029
IPW	MI.ac	-0.005	0.015	0.023	-0.006	0.013	0.022
	MI.wi	-0.004	0.017	0.023	-0.006	0.015	0.022
	MI.co	-0.004	0.016	0.023	-0.006	0.014	0.022
Matching (1-1)	MI.ac	0.007	0.026	0.034	0.004	0.024	0.033
	MI.wi	0.001	0.018	0.032	0.000	0.019	0.031
	MI.co	0.006	0.024	0.033	0.003	0.023	0.032
PS Stratified	MI.ac	0.005	0.025	0.034	0.005	0.025	0.030
	MI.wi	-0.000	0.017	0.032	0.000	0.018	0.028
	MI.co	0.004	0.022	0.033	0.004	0.023	0.029
PS Regression	MI.ac	0.005	0.029	0.031	0.004	0.026	0.031
	MI.wi	0.000	0.021	0.029	-0.001	0.019	0.029
	MI.co	0.004	0.026	0.030	0.003	0.024	0.030
AIPW	MI.ac	-0.017	0.012	0.022	-0.017	0.011	0.021
	MI.wi	-0.004	0.017	0.024	-0.006	0.015	0.023
	MI.co	-0.018	0.013	0.023	-0.018	0.011	0.021
		OR = 1			OR = 4		

Table 3.21 displays the bias values for different multiple imputation (MI) methods under differ-

ent imputation models consideration. We can conclude that all the MI methods are approximately unbiased if the imputation model includes outcome as one of the predictors in the model specification. Which supports Leyrat et al. (2019) findings for binary outcome situation.

Table 3.22: Efficient method for dealing missingness under different estimators for dealing missing confounders in causal estimation.

	Outcome type	
Estimator	Continuous	Binary
Regression based		
Regression	EM	EM
G-estimation	EM	EM
Propensity score based		
IPW	MI.ac	EM
Matching (1-to-1)	MI.wi	MI.wi
Stratification	EM	EM
PS Regression	EM	EM
Doubly robust estimator		
AIPW	EM	EM

Table 3.22 shows the list of efficient methods for dealing missingness under different estimators for continuous and binary outcomes. Simulation studies shows that the EM is most efficient and accurate method for all estimators except IPW under continuous outcome and matching under both continuous and binary outcomes.

Chapter 4

Application to Breast Cancer Data

In this chapter, we apply different estimators on real data intended to determine the average treatment effects (ATE) of radiation therapy (RT) on the survival of breast cancer patients. The data are a subset of a breast cancer (BC) study conducted by Dr. Xiaolan Feng from University of Calgary, Canada and her collaborators (Feng et al., 2016). In this data, a confounder is partially missing, thus, we use the methods for dealing missingness described in Chapter 2 to estimate ATE considering different estimators. We consider a binary outcome variable (Y) which is defined as the death within the next 10 years of BC diagnosis. The treatment variable is adjuvant radiation therapy (RT) again binary, 1 for getting the therapy and 0 otherwise. We use 6 potential confounders including age at diagnosis ($\geq$ 65 years vs <65 years), tumor grade (3 vs 1 or 2), tumor size ($\geq$ 2cm vs $<$ 2cm), lymph nodes (LN: $\geq$ 1 vs 0), cancer stage (1 vs 2, and 1 vs 3) and lymphovascular invasion (LVI: yes vs no). The confounder LVI is partially missing in the data.

Table 4.1: Baseline characteristics of our sub-sample of the BC study, by death, treatment and missingness status of confounder LVI.

		Death status			Treatment (RT)			Missingness of LVI			
Variables	Group	Yes	No	p^*	Yes	No	p^*	Yes	No	p^*	Overall
RT, *n* (%)	No	54 (38.0)	88 (62.0)					27 (19.0)	115 (81.0)		142 (36.1)
	Yes	61 (24.3)	190 (75.7)	0.01				54 (21.5)	197 (78.5)	0.65	251 (63.9)
Age, *n* (%)	< 65year	28 (15.1)	158 (84.9)		132 (71.0)	54 (29.0)		36 (19.4)	150 (80.6)		186 (47.3)
	≥ 65 year	87 (42.0)	120 (58.0)	0.00	119 (57.5)	88 (42.5)	0.01	45 (21.7)	162 (78.3)	0.65	207 (52.7)
Grade, *n* (%)	≤ 2	89 (26.1)	252 (73.9)		223 (65.4)	118 (34.6)		72 (21.1)	269 (78.9)		341 (86.8)
	3	26 (50.0)	26 (50.0)	0.00	28 (53.8)	24 (46.2)	0.14	9 (17.3)	43 (82.7)	0.65	52 (13.2)
Tumor size, *n* (%)	< 2 cm	51 (21.6)	185 (78.4)		162 (68.6)	74 (31.4)		49 (20.8)	187 (79.2)		236 (60.1)
	≥ 2 cm	64 (40.8)	93 (59.2)	0.00	89 (56.7)	68 (43.3)	0.02	32 (20.4)	125 (79.6)	1.00	157 (39.9)
LN, *n* (%)	0	63 (21.0)	237 (79.0)		197 (65.7)	103 (34.3)		62 (20.7)	238 (79.3)		300 (76.3)
	≥ 1	52 (55.9)	41 (44.1)	0.00	54 (58.1)	39 (41.9)	0.23	19 (20.4)	74 (79.6)	1.00	93 (23.7)
Stage, *n* (%)	1	43 (18.3)	192 (81.7)		163 (69.4)	72 (30.6)	0.02	48 (20.4)	187 (79.6)	0.94	235 (59.8)
	2	49 (37.4)	82 (62.6)		72 (55.0)	59 (45.0)		28 (21.4)	103 (78.6)		131 (33.3)
	3	4 (14.8)	23 (85.2)	0.00	16 (59.3)	11 (40.7)		5 (18.5)	22 (81.5)		27 (6.9)
LVI, *n* (%)	No	55 (22.3)	192 (77.7)		159 (64.4)	88 (35.6)					247 (79.2)
	Yes	35 (53.8)	30 (46.2)	0.00	38 (58.5)	27 (41.5)	0.46				65 (20.8)
	Overall	115 (29.3)	278 (70.7)		251 (63.9)	142 (36.1)		81 (20.6)	312 (79.4)		

* p-value from chi-square test for association

Table 4.1 shows the baseline characteristics of our sub-sample of data from the BC study by their death status, radiation treatment (RT) status and whether or not lymphovascular invasion (LVI) was measured. Overall, 29% of the 393 BC patients died within the ten years of BC diagnosis. A total of 64% of the sample received RT. Those who received RT were in good conditions to survive longer for younger age, lower tumour grade and size, minimum lymph nodes, lower stages of cancer and lower risk of lymphovascular invasion. The chi-square test result shows that all the covariates under consideration significantly affects the outcome death status, whereas age, tumor size, and stage are marginally influential for the treatment assignment i.e., propensities. The distribution of different baseline characteristics exhibits similar pattern among the categories whether LVI was measured or not. It seems likely that missingness does not depends on any baseline characteristics; thus, we might expect the missing completely at random (MCAR) mechanism.

In causal effect estimation, correct specification of the both outcome regression and propensity score (PS) models is necessary to ensure unbiased estimate of ATE. In reality, we do not know the form of the correct model. In this study, we apply model building strategy with minimum Akaike Information Criterion (AIC) to determine the possible correct model (Klein and Moeschberger,

2006). To make the model simpler, we only allowed the first order terms for different baseline covariates. Since LVI is considered as the missing confounder, so it will be included in both outcome regression and PS models. We identify the correct models as follows:

$$y - \text{model} : logit(\mu) = \beta_0 + \beta_1 RT + \beta_2 stage2 + \beta_3 stage3 + \beta_4 age + \beta_5 grade + \beta_6 lvi$$
$$e - \text{model} : logit(e) = \alpha_0 + \alpha_1 stage2 + \alpha_2 stage3 + \alpha_3 age + \alpha_4 lvi,$$

where $\mu = Pr(Y = 1|RT, lvi, age, stage, grade; \boldsymbol{\beta})$ and $e = Pr(RT = 1|stage, lvi, age; \boldsymbol{\alpha})$. Finally, the covariates stage, age and LVI are identified as potential confounders as they affects both outcome and treatment variable. The grade of tumor identified as a covariate which only influence the outcome not the treatment assignment.

Table 4.2: Point and interval estimates of average treatment effects under different methods for dealing missingness for breast cancer study.

			SE/	95% CI		Interval
Estimator	Method	Estimate	BootSE	lower	upper	width
Regression	CC	-0.097	0.048	-0.190	-0.003	0.187
	EM	-0.074	0.043	-0.158	0.010	0.167
	MI.ac.wi.co	-0.072	0.044	-0.158	0.013	0.172
G-estimation	CC	-0.086	0.047	-0.178	0.006	0.184
	EM	-0.068	0.043	-0.151	0.015	0.167
	MI.ac.wi	-0.067	0.043	-0.151	0.017	0.169
	MI.co	-0.068	0.043	-0.152	0.017	0.169
Matching (1-1)	CC	-0.099	0.034	-0.166	-0.033	0.133
	EM	-0.079	0.028	-0.134	-0.024	0.110
	MI.ac	-0.038	0.029	-0.094	0.018	0.112
	MI.wi	-0.061	0.029	-0.118	-0.005	0.113
	MI.co	-0.048	0.029	-0.105	0.009	0.114
Stratified	CC	-0.090	0.050	-0.188	0.007	0.196
	EM	-0.073	0.045	-0.160	0.015	0.175
	MI.ac	-0.070	0.045	-0.158	0.017	0.175
	MI.wi	-0.063	0.044	-0.149	0.022	0.171
	MI.co	-0.069	0.044	-0.156	0.017	0.173

Continued . . .

Estimator	Method	Estimate	SE/ BootSE	95% CI lower	95% CI upper	Interval width
PS Regression	CC	-0.099	0.048	-0.194	-0.005	0.189
	EM	-0.076	0.043	-0.160	0.008	0.168
	MI.ac	-0.073	0.044	-0.160	0.014	0.174
	MI.wi	-0.075	0.044	-0.161	0.012	0.173
	MI.co	-0.074	0.044	-0.161	0.013	0.173
IPW	CC	-0.091	0.048	-0.184	0.003	0.187
	EM	-0.071	0.043	-0.154	0.013	0.167
	MI.ac	-0.069	0.043	-0.154	0.016	0.170
	MI.wi	-0.069	0.043	-0.154	0.016	0.170
	MI.co	-0.069	0.043	-0.154	0.016	0.170
AIPW	CC	-0.086	0.048	-0.180	0.008	0.187
	EM	-0.067	0.043	-0.151	0.016	0.167
	MI.ac	-0.066	0.042	-0.149	0.017	0.166
	MI.wi	-0.066	0.043	-0.150	0.017	0.167
	MI.co	-0.066	0.043	-0.149	0.018	0.167

Table 4.2 compares the analytical results for BC study data with different methods for dealing missingness under different estimators. Given the MCAR mechanism, we might expect the complete case (CC) estimates to be unbiassed. Expectedly, the EM method under different estimators accounting for the missing data is relatively closer to the CC estimates than the estimates obtained via multiple imputation (MI). However, the distant behavior of EM and MI methods perhaps indicate that the postulated outcome regression and PS models used for the effect estimation does not fit well. The CC method appears most variable, while the other methods are nearly equivalent in terms of variability which is also supported by the 95% confidence intervals and their corresponding widths. Examining the confidence interval, we find that most of the time estimated effect measures are not significant at 5% significance level as the interval contains the null value of zero. We have already got the results from simulation that for binary outcome the EM method is outperformed under DR estimator if the missing is greater than 25%, which is not the case in our BC data (21% missing for LVI). From the point estimate we can conclude that the risk of death within next 10 years after BC diagnosis will be decreased by approximately 7% if one gets the radiation treatment compared to if one doesn't. That implies the radiation treatment has a positive impact on the survival of BC patients.

Chapter 5

Discussion and Future Work

Confounding and missing data are both likely to occur when observational data are used to estimate the causal effect of a treatment. In dealing with these problems, we face the additional challenge of not knowing the true mechanisms by which the confounding and missing data arise. Standard statistical solutions like complete case analysis, to the problems are therefore likely to create errors because of model misspecification. With a large collection of methods available for missing data and causal inference problems it is often difficult to identify the most appropriate method for a given application. This thesis aimed to identify best method for dealing with missingness in confounders under different causal estimators. Two class of estimators has been used: regression based and propensity score (PS) based estimators. We considered complete case (CC) analysis, multiple imputation (MI) and expected-maximization (EM) algorithm as methods for dealing the missingness. The performance of different methods for dealing missingness relies on the missing mechanism, model specification and type of the outcome. Simulation study has been conducted for both continuous and binary outcomes considering correct and incorrect model specifications under missing completely at random (MCAR) and missing at random (MAR) mechanisms. We applied two versions of CC method MCAR.cc and MAR.cc for MCAR and MAR mechanisms respectively, while MI and EM methods are applied only for MAR mechanism. Because the CC method is sufficient to obtain the accurate (both unbiased and efficient) estimate when the missing mechanism is MCAR. We considered three different approaches under MI method discussed by Mitra and Reiter (2016) and Leyrat et al. (2019) namely MI.ac, MI.wi and MI.co for across, within and combined approach respectively. To make the comparison better we also considered the standard analysis when there is no missing.

If the model is correctly specified most of the methods for dealing missingness for different estimators (exception for MAR.cc) are unbiased and accurate, while under incorrect model specification all methods became biased and inaccurate. Simulation studies showed that there was no significant gain in the efficiency for considering highly associated covariates with the missing confounder compared to independent assumption. The EM found as the most efficient and accurate method for the regression-based estimators namely regression estimator and G-estimation in both case of continuous and binary outcomes. In PS based estimators, the choice of method for dealing missingness differs within the estimator itself considering the types of outcome. The MI method, MI.wi is found as the most efficient for matching estimator for both continuous and binary outcomes. While the MI.ac best performed in terms of efficiency and accuracy when the outcome is continuous and under binary outcome, EM method performed best for inverse probability weighting (IPW) estimator. In other PS based estimators, stratified and PS regression, EM is the best choice to deal the missingness in confounders. The regression-based estimators are unbiased if the outcome regression model reflects the true relationship among treatment, confounders, and the outcome. In case of PS based estimators, which the treatment is modeled as a function of covariates, is unbiased if the treatment model is true. Outside of simulation studies, we can never know whether the model we had constructed accurately depicts those relationships. Thus, correct specification of the regression or PS model is an unverifiable assumption. Augmented inverse probability weighted (AIPW) estimator provided the double protection of estimation consistency, if either of the outcome regression or PS model is correct the effect will be estimated consistently. The EM is best performed for AIPW estimator under both continuous and binary outcomes.

In all scenarios, the EM and MI methods are the leading competitors to each other in dealing missingness for consistently estimate the causal effects. The performance of MI method mostly depends on the specification of imputation model and number of imputations. The simulation results showed a bias in the three MI methods when either of the outcome and treatment was excluded form the imputation model for continuous outcome, while exclusion of outcome only made the three methods biased for binary outcome. All the three MI methods became computationally challenging and even infeasible for very large datasets. However, with the recommended imputation models and 10 imputed datasets for both continuous and binary outcomes, all three MI methods are dominated by EM method except matching and IPW (in continuous outcome) estimators. The

flexibility of EM over MI method is that the EM is not computationally much demanding than MI methods. In EM method, we estimate the expected values of the missing confounder and utilize weighting approach in the effect estimation. Detailed comparison among the three MI methods is given by Leyrat et al. (2019), which is supported by our obtained results.

The AIPW estimator is the most desired estimator as it provides double protection for getting consistent estimates of the effects. In EM method, the expected value of the missing confounders is calculated after specifying the outcome regression and PS models. If these models are not correctly specified then the DR property is violated specifically for the continuous outcome. This property can be retained by allowing multiple models for both outcome regression and PS models. One of the striking advantages of using multiple models is that EM method will hold DR even when the multiple sets does not contain any of the correctly specified model. For binary outcome, multiple models in EM method there is no significant gain compare to the single model consideration in the estimation efficiency and consistency.

In real data application, we found that the data is missing completely at random (MCAR) i.e., missingness does not depends on any baseline characteristics. Under the MCAR mechanism, CC estimates will be unbiased. We found that EM estimates are more closer to CC estimates than the three MI methods. This indicates the accuracy of EM approach over MI and the results may be further improved by fitting appropriate outcome regression and propensity score models.

In this study, we have considered only the scenarios where a single categorical confounder is missing. In practice, missing can be ocurred simultaneously e.g., several confounders, or confounders with either outcome or treatment or both. The missing confounder may be continuous and can be incorporated by updating the complete data likelihood and E-step of the EM algorithm accordingly. Future research will consider the extent to which the EM method can be extended to arbitrary missing data patterns. We can directly extend the methods developed by Williamson et al. (2012). Multiple robust estimator can be considered under the EM algorithm as discussed by Han and Wang (2013). In multiple robust estimator, it allows multiple models for both outcome regression and propensity score models. The EM method can be considered in other different related topics, namely high-dimensional data, clustered data and survival data considering missing with different components. In addition, missing not at random (MNAR) can also be considered as missing mechanism to develop causal effect estimators.

Bibliography

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

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.

Afifi, A. and Elashoff, R. (1966). Missing observations in multivariate statistics i. review of the literature. *Journal of the American Statistical Association*, 61(315):595–604.

Austin, P. C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in medicine*, 26(16):3078–3094.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine*, 33(6):1057–1069.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49.

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

Boos, D. D. and Stefanski, L. (2013). *Essential Statistical Inference: Theory and Methods*, volume 120. Springer Science & Business Media.

Carpenter, J. and Kenward, M. (2012). *Multiple imputation and its application*. John Wiley & Sons.

Carpenter, J. R., Kenward, M. G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):571–584.

Chernick, M. R. (2011). *Bootstrap methods: A guide for practitioners and researchers*, volume 619. John Wiley & Sons.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*, volume 1. Cambridge university press.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical science*, pages 189–212.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Feng, X., Li, H., Kornaga, E. N., Dean, M., Lees-Miller, S. P., Riabowol, K., Magliocco, A. M., Morris, D., Watson, P. H., Enwere, E. K., et al. (2016). Low ki67/high atm protein expression in malignant tumors predicts favorable prognosis in a retrospective study of early stage hormone receptor positive breast cancer. *Oncotarget*, 7(52):85798.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767.

Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56.

Gnant, M., Filipits, M., Greil, R., Stoeger, H., Rudas, M., Bago-Horvath, Z., Mlineritsch, B., Kwasny, W., Knauer, M., Singer, C., et al. (2013). Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: using the pam50 risk of

recurrence score in 1478 postmenopausal patients of the abcsg-8 trial treated with adjuvant endocrine therapy alone. *Annals of oncology*, 25(2):339–345.

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

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

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99(467):609–618.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294.

Heitjan, D. F. and Basu, S. (1996). Distinguishing "missing at random" and "missing completely at random". *The American Statistician*, 50(3):207–213.

Heitjan, D. F. and Little, R. J. (1991). Multiple imputation for the fatal accident reporting system. *Applied Statistics*, pages 13–29.

Hernán, M. A. and Robins, J. M. (2019). *Causal inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.

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

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(3):199–236.

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

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29.

Imbens, G. W. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, 50(2):373–419.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

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

Klein, J. P. and Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J., Resche-Rigon, M., Carpenter, J. R., and Williamson, E. J. (2019). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical methods in medical research*, 28(1):3–19.

Li, L., Shen, C., Li, X., and Robins, J. M. (2013). On weighting approaches for missing data. *Statistical methods in medical research*, 22(1):14–30.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, 87(420):1227–1237.

Little, R. J. (1997). Biostatistical analysis with missing data. *Encyclopedia of Biostatistics (p. Armitage and T. Colton eds.) London: Wiley*.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. Wiley.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233.

Lu, B. and Ashmead, R. (2018). Propensity score matching analysis for causal effects with mnar covariates. *Statistica Sinica p*.

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

Meng, X.-L. and Rubin, D. B. (1991). Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86(416):899–909.

Mitra, R. and Reiter, J. P. (2016). A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical methods in medical research*, 25(1):188–204.

Mook, S., Schmidt, M. K., Weigelt, B., Kreike, B., Eekhout, I., Van de Vijver, M. J., Glas, A. M., Floore, A., Rutgers, E., and van 't Veer, L. (2009). The 70-gene prognosis signature predicts early metastasis in breast cancer patients between 55 and 70 years of age. *Annals of oncology*, 21(4):717–722.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826.

Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (mimp) approach. *Statistics in Medicine*, 28(9):1402–1414.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):597–610.

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

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

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

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, pages 159–183.

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

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

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Rubin, D. B. et al. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840.

Rubin, D. B. and Schenker, N. (1991). Multiple imputation in health-are databases: An overview and some applications. *Statistics in medicine*, 10(4):585–598.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, 8(1):3–15.

Seaman, S. R. and Vansteelandt, S. (2018). Introduction to double robust methods for incomplete data. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33(2):184.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295.

Snowden, J. M., Rose, S., and Mortimer, K. M. (2011). Implementation of g-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7):731–738.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Stuart, E. A., Azur, M., Frangakis, C., and Leaf, P. (2009). Multiple imputation with large data sets: a case study of the children's mental health initiative. *American journal of epidemiology*, 169(9):1133–1139.

Vach, W. and Blettner, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American journal of epidemiology*, 134(8):895–907.

Van der Heijden, G. J., Donders, A. R. T., Stijnen, T., and Moons, K. G. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of clinical epidemiology*, 59(10):1102–1109.

Vansteelandt, S. and Keiding, N. (2011). Invited commentary: G-computation–lost in translation? *American journal of epidemiology*, 173(7):739–742.

Wacholder, S. (1996). The case-control study as data missing by design: estimating risk differences. *Epidemiology (Cambridge, Mass.)*, 7(2):144–150.

Wang, A., Nianogo, R. A., and Arah, O. A. (2017). G-computation of average treatment effects on the treated and the untreated. *BMC medical research methodology*, 17(1):3.

Westreich, D., Edwards, J. K., Cole, S. R., Platt, R. W., Mumford, S. L., and Schisterman, E. F. (2015). Imputation approaches for potential outcomes in causal inference. *International journal of epidemiology*, 44(5):1731–1737.

White, I. R. and Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine*, 29(28):2920–2931.

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

Wu, L., Candille, S. I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. (2013). Variation and genetic control of protein abundance in humans. *Nature*, 499(7456):79.

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