Marginal Causal Sub-Group Analysis with Incomplete Covariate Data

by

Meaghan S. Cuerden
(Meaghan Cuerden Knight)

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

Waterloo, Ontario, Canada, 2018

Examining Committee Membership

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

External Examiner:	Dr. Gerarda Darlington Professor, University of Guelph
Supervisors:	Dr. Richard J. Cook Professor
	Dr. Cecilia A. Cotton Associate Professor
	Dr. Liqun Diao Assistant Professor
Internal Member:	Dr. Steve Brown Professor Emeritus

Internal Member: Dr. Joel A. Dubin
Associate Professor

Internal-External Member: Dr. Carrie McAiney
Associate Professor

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

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

Abstract

Incomplete data arises frequently in health research studies designed to investigate the causal relationship between a treatment or exposure, and a response of interest. Statistical methods for conditional causal effect parameters in the setting of incomplete data have been developed, and we expand upon these methods for estimating marginal causal effect parameters. This thesis focuses on the estimation of marginal causal odds ratios, which are distinct from conditional causal odds ratios in logistic regression models; marginal causal odds ratios are frequently of interest in population studies. We introduce three methods for estimating the marginal causal odds ratio of a binary response for different levels of a subgroup variable, where the subgroup variable is incomplete. In each chapter, the subgroup variable, exposure variable and the response variable are binary and the subgroup variable is missing at random.

In Chapter 2, we begin with an overview of inverse probability weighted methods for confounding in an observational setting where data are complete. We also briefly review methods to deal with incomplete data in a randomized setting. We then introduce a doubly inverse probability weighted estimating equation approach to estimate marginal causal odds ratios in an observational setting, where an important subgroup variable is incomplete. One inverse probability weight accounts for the incomplete data, and the other weight accounts for treatment selection. Only complete cases are included in the response model. Consistency results are derived, and a method to obtain estimates of the asymptotic standard error is introduced; the extra variability introduced by estimating two weights is incorporated in the estimation of the asymptotic standard error. We give a method for hypothesis testing and calculation of confidence intervals. Simulation studies show that the doubly weighted estimating equation approach is effective in a non-ignorable missingness setting with confounding, and it is straightforward to implement. It also performs well when the missing data process is ignorable, and/or when confounding is not present.

In Chapter 3, we begin with an overview of an EM algorithm approach for estimating conditional causal effect parameters in the setting of incomplete covariate data, in both randomized and observational settings. We then propose the use of a doubly weighted EM-type algorithm approach to estimate the marginal causal odds ratio in the setting of

missing subgroup data. In this method, instead of using complete case analysis in the response model, all available data is used and the incomplete subgroup variable is filled in using a maximum likelihood approach. Two inverse probability weights are used here as well, to account for confounding and incomplete data. The weight which accounts for the incomplete data is needed, even though an EM approach is being used, because the marginal causal odds ratio is of interest. A method to obtain asymptotic standard error estimates is given where the extra variability introduced by estimating the two inverse probability weights, as well as the variability introduced by estimating the conditional expectation of the incomplete subgroup variable, is incorporated. Simulation studies show that this method is effective in terms of obtaining consistent estimates of the parameters of interest; however it is difficult to implement, and in certain settings there is a loss of efficiency in comparison to the methods introduced in Chapter 2.

In Chapter 4, we begin by reviewing multiple imputation methods in randomized and observational settings, where estimation of the conditional causal odds ratio is of interest. We then propose the use of multiple imputation with one inverse probability weight to account for confounding in an observational setting where the subgroup variable is incomplete. We discuss methods to correctly specify the imputation model in the setting where the conditional causal odds ratio is of interest, as well as in the setting where the marginal causal odds ratio is of interest. We use standard methods for combining the estimates of the marginal log odds ratios from each imputed dataset. We propose a method for estimating the asymptotic standard error of the estimates, which incorporates both the estimation of the parameters in the weight for confounding, and the multiply imputed datasets. We give a method for hypothesis testing and calculation of confidence intervals. Simulation studies show that this method is efficient and straightforward to implement, but correct specification of the imputation model is necessary.

In Chapter 5, the three methods that have been introduced are used in an application to an observational cohort study of 418 colorectal cancer patients. We compare patients who received an experimental chemotherapy with patients who received standard chemotherapy; of interest is estimation of the marginal causal odds ratio of a thrombotic event during the course of treatment or 30 days after treatment is discontinued. The important subgroups are (i) patients receiving first line of treatment, and (ii) patients receiving second line of

treatment.

In Chapter 6, we compare and contrast the three methods proposed. We also discuss extensions to different response models, models for missing response data, and weighted models in the longitudinal data setting.

Acknowledgements

I am so grateful to my supervisors Dr. Richard Cook, Dr. Cecilia Cotton and Dr. Liqun Diao. Thank you for believing in me, and thank you for your support and encouragement. Your example and advice has been invaluable, and I will carry it with me always. I want to say a special thank you to Liqun, my friend and former office mate - it has been inspiring to work with you and learn from you.

I thank Dr. Steve Brown, Dr. Joel Dubin, Dr. Gerarda Darlington and Dr. Carrie McAiney for serving as my committee members and for offering guidance.

To my mom, thank you for your generosity, support and guidance. Thank you to my mother-in-law for always being there for me and the kids. Thank you Dad and Linda for your support over the years.

I want to say a special thank you to my (other) office mates, Di, Narges and Nathalie. You are all so wonderful and talented! I am blessed and honoured to have been able to get to know you, and to call you my friends.

Thank you Mary Lou Dufton for supporting my peers and I with kindness and good guidance. Thank you to Ker-Ai Lee for your advice with statistical computing. And thank you to Marg Feeney for your help over the years.

I want to thank my colleagues at London Health Sciences Centre. Thank you Dr. Amit Garg for your support - I have learned so much during my time with the Kidney Clinical Research Unit. I want to say a special thank you to Heather Thiessen-Philbrook who trained me when I was first employed at LHSC - thank you for your advice and example.

I would like to thank Dr. Pierre Major and Dr. Humaid Al-Shamsi for permission to use the data from the cancer study.

Thank you to Drs. Marcus Pivato and Michelle Boué who taught me at Trent University and encouraged me to continue studying at the graduate level.

To my husband, Brendan - I dedicate this thesis to you. Thank you for believing in me and supporting me every day, consistently, unwaveringly. I dedicate this thesis to our children as well, who came into our lives during my studies. William and Rosie, you are my guiding light.

Table of Contents

List of Tables xiv

List of Figures xvi

1 Introduction 1

1.1 Statistical Methods for Causal Analysis 1

1.1.1 Counterfactual Notation, Causal Effects and Study Design 2

1.1.2 The Propensity Score 7

1.1.3 Inverse Probability of Treatment Weighted Estimating Equations 9

1.1.4 Subgroup Effects 10

1.2 Statistical Methods for Incomplete Covariates 12

1.2.1 Inverse Probability Weighted Estimating Equations 13

1.2.2 The Expectation-Maximization Algorithm 14

1.2.3 Multiple Imputation 15

1.3 Causal Inference in the Setting of Missing Data 16

1.4 Incidence of Thrombotic Events in Metastatic Colorectal Cancer Patients . 17

1.5 Outline of Thesis 18

2 Doubly Weighted Estimating Equations for Causal Inference with an Incomplete Subgroup Variable **20**

2.1 Notation and Models . . . 21

2.1.1 Subgroup Effects in a Randomized Setting . . . 21

2.1.2 Subgroup Effects in an Observational Setting . . . 25

2.2 Inference with Incomplete Subgroup Data . . . 29

2.2.1 Estimation in the Randomized Setting . . . 29

2.2.2 Estimation in an Observational Setting . . . 32

2.3 Statistical Inference . . . 34

2.3.1 Consistency of Doubly Weighted Estimators for $\boldsymbol{\beta}$ . . . 34

2.3.2 Derivation of the Asymptotic Covariance . . . 37

2.3.3 Some Results on Misspecified Estimating Functions . . . 39

2.4 Simulation Studies . . . 41

2.4.1 Parameter Settings . . . 42

2.4.2 Discussion of Simulation Results . . . 47

2.5 Discussion . . . 51

3 Causal Inference with Incomplete Data Via a Weighted EM Algorithm **52**

3.1 Estimation of Conditional Causal Effects in a Randomized Setting . . . 53

3.1.1 Notation and Models . . . 53

3.1.2 An EM Algorithm for Incomplete Covariates . . . 55

3.2 Estimation of Conditional Causal Effects in an Observational Setting . . . 59

3.2.1 Notation and Models . . . 59

3.2.2 The EM Algorithm with Observational Data . . . 61

3.3 Estimation of Marginal Causal Effects in a Randomized Setting . . . 62

3.3.1 Notation and Models . . . 62
3.3.2 A Weighted EM-Type Algorithm for Incomplete Data . . . 63
3.4 Estimation of Marginal Causal Effects in an Observational Setting . . . 66
3.4.1 Notation and Models . . . 66
3.4.2 Doubly Weighted EM-Type Algorithm . . . 67
3.4.3 Consistency Results . . . 70
3.4.4 Derivation of the Asymptotic Variance . . . 74
3.5 Simulation Studies . . . 77
3.5.1 Parameter Settings . . . 77
3.5.2 Discussion of Simulation Study Results . . . 80
3.6 Discussion . . . 80

4 Multiple Imputation for Causal Inference with Incomplete Data 82

4.1 Estimation of Conditional Causal Effects in a Randomized Setting . . . 83
4.1.1 Notation and Models . . . 83
4.1.2 Imputation Model . . . 84
4.1.3 Estimation of Response Model Coefficients . . . 87
4.1.4 Inference . . . 89
4.2 Estimation of Conditional Causal Effects in an Observational Setting . . . 90
4.2.1 Notation and Models . . . 90
4.2.2 Imputation Model and Estimation of Causal Effects . . . 91
4.3 Estimation of Marginal Causal Effects in a Randomized Setting . . . 92
4.3.1 Notation and Models . . . 92
4.3.2 Imputation Model . . . 92

4.3.3 Estimation of Response Model Coefficients . . . 93
4.3.4 Inference . . . 94
4.4 Estimation of Marginal Causal Effects in an Observational Setting . . . 94
4.4.1 Imputation Model . . . 95
4.4.2 Estimation of Response Model Coefficients . . . 95
4.4.3 Inference . . . 97
4.5 Simulation Specifications and Results . . . 98
4.5.1 Parameter Settings . . . 99
4.5.2 Discussion of Simulation Studies . . . 106
4.6 Summary . . . 107

5 Incidence of Thrombotic Events in the Treatment of Metastatic Colorectal Cancer 108

5.1 Background, Exclusions and Data Summary . . . 108
5.2 Variable Selection and Notation . . . 110
5.3 Conditional Response Model . . . 119
5.4 Marginal Causal Effect Estimation . . . 122
5.5 Additional Estimates of Marginal Causal Effects . . . 125
5.6 Discussion and Conclusions . . . 126

6 Discussion, Conclusions and Future Work 127

6.1 Summary and Discussion . . . 127
6.2 Future Work . . . 130
6.2.1 Extension to Other Response Models . . . 130
6.2.2 Stabilized Weights . . . 131
6.2.3 Extension to Missing Response Data . . . 132
6.2.4 Extension to Longitudinal Data . . . 132

References **135**

APPENDICES **146**

A Empirical Power Calculations **147**

A.1 Marginal Causal Effects 147

A.1.1 Discussion 148

A.2 Conditional Causal Effects 149

A.2.1 Discussion 151

B Simulation Study to Explore the Effect of a Subgroup Variable That Is Also a Confounder **152**

B.1 Discussion 153

List of Tables

2.1 Simulation study, percentage of missing data is 20% 45

2.2 Simulation study, percentage of missing data is 40% 46

3.1 Layout of a dataset for the EM algorithm, at the $(k+1)$st M-step. 56

3.2 Layout of a dataset for the weighted EM-type algorithm, at the $(k+1)$st M-step. 65

3.3 Layout of a dataset for the doubly weighted EM-type algorithm, at the $(k+1)$st M-step. 69

3.4 Simulation study: estimating $\boldsymbol{\vartheta}$ using the EM algorithm 78

3.5 Simulation study: estimating $\boldsymbol{\beta}$ using the (doubly) weighted EM-type algorithm 79

4.1 Simulation study: estimating $\boldsymbol{\vartheta}$ using multiple imputation, 20% missingness 101

4.2 Simulation study: estimating $\boldsymbol{\vartheta}$ using multiple imputation, 40% missingness 102

4.3 Simulation study: estimating $\boldsymbol{\beta}$ using multiple imputation, 20% missingness 103

4.4 Simulation study: estimating $\boldsymbol{\beta}$ using multiple imputation, 40% missingness 104

4.5 Simulation study: estimating $\boldsymbol{\beta}$ using Chapter 3 parameters 105

5.1 Year of start of therapy for the BV plus FOLFIRI study participants . . . 109

5.2 Exclusions for the BV plus FOLFIRI study 110

5.3 Baseline characteristics for FOLFIRI study patients 111
5.4 Summary of thrombotic events for the BV plus FOLFIRI study 112
5.5 Building a propensity score model in the BV plus FOLFIRI study 115
5.6 Logistic regression model for the subgroup variable in the BV plus FOLFIRI study 116
5.7 Logistic regression model for the observed data indicator variable in the BV plus FOLFIRI study 117
5.8 Probability of a thrombotic event for second line of treatment patients . . 117
5.9 Probability of a thrombotic event for first line of treatment patients 118
5.10 Logistic regression models for the response in the BV plus FOLFIRI study 120
5.11 Conditional logistic regression model for the response in the BV plus FOLFIRI study 121
5.12 Complete case conditional response model in the BV plus FOLFIRI study 121
5.13 Application of the doubly weighted estimating equation method in the BV plus FOLFIRI study 123
5.14 Application of the doubly weighted EM-type algorithm method in the BV plus FOLFIRI study 124
5.15 Application of the weighted multiple imputation method in the BV plus FOLFIRI study 125

B.1 Simulation study, omitting S from propensity score 153

List of Figures

2.1 Directed acyclic graph for exposure, response, and auxiliary variables . . . 22

2.2 Directed acyclic graph for exposure, response, confounder, and auxiliary variables . . . 26

2.3 Directed acyclic graph for exposure, response, and auxiliary variables . . . 30

2.4 Directed acyclic graph for exposure, response, confounder, and auxiliary variables . . . 33

2.5 Asymptotic bias under model misspecification, 20% missing data . . . 49

2.6 Asymptotic bias under model misspecification, 40% missing data . . . 50

3.1 Directed acyclic graph in the randomized setting . . . 54

3.2 Directed acyclic graph in an observational setting . . . 60

5.1 Directed acyclic graph for the BV plus FOLFIRI study . . . 113

A.1 Power for Subgroup Effects in a Marginal Model, Chapter 2 Methods . . . 148

A.2 Power for Subgroup Effects in a Marginal Model, Chapter 3 Methods . . . 149

A.3 Power for Subgroup Effects in a Marginal Model, Chapter 4 Methods . . . 150

A.4 Power Calculations for Subgroup Effects in a Conditional Causal Model . . . 150

Chapter 1

Introduction

This thesis focuses on causal analysis in observational settings for a binary response where subgroup data is incompletely observed. Doubly inverse probability weighted estimating equations are developed to account for missing data in the causal inference setting, a doubly weighted EM-type algorithm is proposed, and multiple imputation is explored as a third technique.

In this chapter, statistical methods for causal inference, including an introduction to counterfactual notation, are briefly described. This is followed by a brief review of methods for incomplete data.

1.1 Statistical Methods for Causal Analysis

Methods for causal inference are used to investigate whether something - a treatment or exposure - is causally linked to an outcome of interest. The terms 'treatment' and 'exposure' will be used interchangeably throughout this thesis. In a medical setting, it is often of interest to make causal statements about the effect of an exposure on a health-related outcome (Papanikolaou et al., 2006; Schulz et al., 2010). For example, interest may lie in determining whether patients undergoing non-cardiac surgery who are exposed to

low-dose aspirin are at a lower risk of post-operative acute kidney injury compared with patients who do not take aspirin (Garg et al., 2014).

Throughout this thesis, we focus on estimation of the odds ratio in logistic regression models. The odds ratio is a popular summary measure in comparisons of probabilities between groups as it is easily estimable using standard statistical software (Bland and Altman, 2000; Greenland, 1987). When the probability of success is less than 10% in the comparison group, the odds ratio is a reasonable approximation of the relative risk, which is an easily interpretable summary measure (Viera, 2008).

In this section, we define the causal odds ratio, discuss assumptions for estimation of the causal odds ratio, and discuss issues about collapsibility and subgroup effect estimation in both randomized and observational settings.

1.1.1 Counterfactual Notation, Causal Effects and Study Design

Let Y denote a response of interest, and let X denote a treatment or exposure. For example, Y_i could indicate whether subject i had an acute kidney injury within 14 days after surgery, and X_i could indicate whether the subject was treated with low-dose aspirin. Let Y_{1i} denote the response under treatment $X = 1$ for subject i and Y_{0i} denote the response under treatment $X = 0$. Then the average causal effect of X on Y is defined as

$$E(Y_{1i}) - E(Y_{0i}) \tag{1.1}$$

(Rubin, 1974; Greenland et al., 1999). In the context of a binary response, where interest lies in quantifying the difference in the probability of an event between the two treatments with the odds ratio, the causal odds ratio is defined as

$$\frac{P(Y_{1i} = 1)/[1 - P(Y_{1i} = 1)]}{P(Y_{0i} = 1)/[1 - P(Y_{0i} = 1)]} \ . \tag{1.2}$$

The causal model notation above assumes that treatment $X = 1$ can be administered to subject i, and then treatment $X = 0$ can be applied to the same individual under the exact same conditions. This is not possible in practice; (Y_1, Y_0) are called *counterfactuals*

or *potential* responses because only one can be observed, and observation of the other is therefore 'counter to fact' (Neyman, 1923; Rubin, 1990). Counterfactual notation can be used as a framework for describing relationships in observational studies where causal, rather than associative, questions are of interest (Pearl, 2010).

In practice, treatment $X = 1$ is administered to one group of subjects, and treatment $X = 0$ is administered to another group of subjects, and the difference between the average responses for each group is used to approximate the causal effect (1.1). Care must be taken to ensure that there are no systematic differences between the treatment groups, other than the treatment itself. The gold standard for making causal inference in a clinical setting is the randomized controlled trial, where the treatment or exposure is randomly allocated to study participants and therefore no participant factors, measured or unmeasured, can influence both treatment selection and the response of interest. Because the treatment allocation is not related to participant characteristics in a randomized trial, participant characteristics will be balanced across the treatment groups, and any differences in outcomes between those who are treated and those who are untreated can be attributed to the treatment (Greenland, 1990; Greenland et al., 1999; Rubin, 1974).

Nonrandomized, or observational, studies are also useful for causal analysis, although analysis and interpretation of observational data is more complex due to systematic differences that often exist between those who are treated and those who are untreated; these differences may, at least in part, explain any differences in the average response between the treatment groups (Black, 1996; von Elm et al., 2007). In other words, characteristics that are different between the treatment groups can *confound* our ability to attribute any differences in the response to the treatment. We define a *confounding variable* or *confounder* as a variable that is both related to treatment selection (i.e. whether a subject receives treatment $X = 1$ or $X = 0$), and the response of interest (Greenland et al., 1999), and we let $\mathbf{Z}_1$ denote a vector of measured confounders. Note that a confounder in this setting is a variable that is not on the causal pathway between the treatment and the response; we assume that confounders are measured at the time that treatment is selected/randomly allocated.

We define the following assumptions for causal analysis: (i) strong ignorability, (ii) positivity, (iii) consistency, and (iv) the stable unit treatment value assumption. These

assumptions are made implicitly in all chapters of this thesis.
(i) *Strong Ignorability:* Strong ignorability holds when treatment assignment (whether $X = 1$ or $X = 0$) and response (Y_1, Y_0) are independent, conditional on measured confounders $\mathbf{Z}_1$. This is also referred to as the *no unmeasured confounders* assumption, or *conditional exchangeability* (Cole and Hernán, 2008; Rosenbaum and Rubin, 1983). In the randomized setting where $X \perp \mathbf{Z}_1$, treatment selection is strongly ignorable.
(ii) *Positivity:* The positivity assumption holds when the conditional probability of receiving treatment $X = 1$ or $X = 0$ is non-zero, i.e. $0 < P(X = 1|\mathbf{Z}_1) < 1$ (Petersen et al., 2010).
(iii) *Consistency:* The consistency assumption holds when a subject's observed outcome is equal to their counterfactual outcome under the observed exposure (Cole and Hernán, 2008). Formally, using counterfactual notation, consistency holds when

$$Y_x = Y \ ,$$

for $x = 0, 1$.
(iv) *Stable Unit Treatment Value Assumption (SUTVA):* The stable unit treatment value assumption holds when: (a) all individuals in the same exposure group have identical treatment, (b) the potential response of one individual does not influence that of another individual, (c) the exposure status of one individual does not influence the potential responses of another individual, and (d) identical repetitions of the exposure would result in identical responses (Cotton, 2009; Pearl, 2003).

The strong ignorability, or no unmeasured confounders, assumption is difficult to test in a study setting. In practice, if the set of covariates collected is rich enough and contains clinically relevant information about the response process, the strong ignorability assumption may be appropriate. Brumback et al. (2004) developed a method for attempting to quantify unmeasured confounding in marginal structural model analyses, but we do not explore these methods here.

These assumptions, particularly consistency and strong ignorability, are useful in both the randomized and observational study settings for making causal inferences. First, we show that we can obtain unbiased estimates of the average causal difference in the response (equation (1.1)) using randomized data. Next, we show that the same is true in an obser-

vational setting, provided there are no unmeasured confounders. In a randomized trial, the expected value of the difference in the average response is equal to the causal difference in means:

$$\begin{aligned} & E[E(Y_i|X_i = 1) - E(Y_i|X_i = 0)] \\ = \; & E[E(Y_{1i}|X_i = 1) - E(Y_{0i}|X_i = 0)] \text{ (consistency)} \\ = \; & E[E(Y_{1i}) - E(Y_{0i})] \text{ (strong ignorability, randomized trial)} \\ = \; & E(Y_{1i}) - E(Y_{0i}) \end{aligned}$$

where the strong ignorability assumption holds because there are no confounders in a randomized setting. Before showing that observational data can be used to estimate the average causal difference in the response of interest, we introduce new notation for the treatment variable in the observational setting; let W_i denote the treatment variable in an observational study for subject i. We distinguish between X_i and W_i because the distribution of the treatment variable in a randomized setting is independent of all subject characteristics, whereas the distribution of the treatment variable is often dependent upon subject characteristics in observational settings. In an observational setting with no unmeasured confounders, the expected value of the difference in the average response is

$$\begin{aligned} & E[E(Y_i|W_i = 1, \mathbf{Z}_{1i}) - E(Y_i|W_i = 0, \mathbf{Z}_{1i})] \\ = \; & E[E(Y_{1i}|W_i = 1, \mathbf{Z}_{1i}) - E(Y_{0i}|W_i = 0, \mathbf{Z}_{1i})] \text{ (consistency)} \\ = \; & E[E(Y_{1i}|\mathbf{Z}_{1i}) - E(Y_{0i}|\mathbf{Z}_{1i})] \text{ (strong ignorability)} \\ = \; & E(Y_{1i}) - E(Y_{0i}) \end{aligned} \tag{1.3}$$

where the strong ignorability assumption holds because confounder vector $\mathbf{Z}_{1i}$ contains information on all important confounders.

The causal odds ratio is defined in (1.2), and we show here that randomized study data

can be used to estimate the causal odds ratio where the response of interest is binary:

$$\begin{aligned}
& \frac{E(Y_i|X_i=1)/(1-E(Y_i|X_i=1))}{E(Y_i|X_i=0)/(1-E(Y_i|X_i=0))} \\
= & \frac{E(Y_{1i}|X_i=1)/(1-E(Y_{1i}|X_i=1))}{E(Y_{0i}|X_i=0)/(1-E(Y_{0i}|X_i=0))} \text{ (consistency)} \\
= & \frac{E(Y_{1i})/(1-E(Y_{1i}))}{E(Y_{0i})/(1-E(Y_{0i}))} \text{ (strong ignorability, randomized trial)} \\
= & \frac{P(Y_{1i}=1)/(1-P(Y_{1i}=1))}{P(Y_{0i}=1)/(1-P(Y_{0i}=1))} \, .
\end{aligned}$$

However, we cannot use observational data to estimate the causal odds ratio in the same way as in (1.3) (i.e. by conditioning on the confounders) and this is because the odds ratio is subject to non-collapsibility (Greenland and Robins, 1986; Martinussen and Vansteelandt, 2013; Miettinen and Cook, 1981).

A model is said to be *collapsible* if the *marginal causal effect* is equal to the *conditional causal effect* (Austin et al., 2007b; Greenland et al., 1999). A conditional response model is one where all variables that are associated with the response are included (conditioned on). An example of a conditional generalized linear model is

$$g[\mu_i(X_i, \mathbf{Z}_{1i}, \mathbf{V}_i)] = \vartheta_0 + \vartheta_1 X_i + \mathbf{Z}_{1i}\boldsymbol{\vartheta}_2 + \mathbf{V}_i\boldsymbol{\vartheta}_3 \, ,$$

where $\mu_i(X_i, \mathbf{Z}_{1i}, \mathbf{V}_i) = E(Y_i|X_i, \mathbf{Z}_{1i}, \mathbf{V}_i)$ and $\mathbf{V}_i$ denotes a vector of auxiliary variables measured at the time that treatment is administered that are associated with the response, but independent of treatment selection. The vector of regression coefficients in the conditional response model is $\boldsymbol{\vartheta} = (\vartheta_0, \vartheta_1, \boldsymbol{\vartheta}_2^T, \boldsymbol{\vartheta}_3^T)^T$. A marginal response model is a causal model where only the treatment variable is included as an independent variable. For example,

$$g[\mu_i(X_i)] = \beta_0 + \beta_1 X_i \, , \tag{1.4}$$

where $\mu_i(X_i) = E(Y_i|X_i)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is a vector of regression coefficients. A model is said to be *collapsible* if the marginal treatment effect is equal to the conditional treatment effect, i.e. $\beta_1 = \vartheta_1$. In summary,
i) $\beta_1 = \vartheta_1$ when $X \perp \mathbf{Z}_1$ and the model is collapsible;

ii) $\beta_1 \neq \vartheta_1$ when $X \perp \mathbf{Z}_1$ and the model is not collapsible;
and iii) $\beta_1 \neq \vartheta_1$ when $X \not\perp \mathbf{Z}_1$ regardless of whether the model is collapsible.
Note that for simplicity, we have used X and W interchangeably here, and we have let $X \perp \mathbf{Z}_1$ denote a randomized setting, and $X \not\perp \mathbf{Z}_1$ in an observational setting. The X and W notation will be used to distinguish between a randomized setting and an observational setting throughout the rest of this work. An example of a collapsible regression model is the linear regression model where $g[\mu] = \mu$, and an example of a non-collapsible regression model is the logistic regression model with logit link function $g[\mu] = \log(\mu/(1-\mu))$ (Austin et al., 2007b; Greenland, 1987; Greenland et al., 1999).

The *conditional causal odds ratio* is defined as

$$\frac{P(Y_{1i} = 1|\mathbf{Z}_{1i}, \mathbf{V}_i)/[1 - P(Y_{1i} = 1|\mathbf{Z}_{1i}, \mathbf{V}_i)]}{P(Y_{0i} = 1|\mathbf{Z}_{1i}, \mathbf{V}_i)/[1 - P(Y_{0i} = 1|\mathbf{Z}_{1i}, \mathbf{V}_i)]} \ . \tag{1.5}$$

In the following subsections, we discuss methods for estimation of the marginal causal odds ratio (1.2) that do not include conditioning on measured confounders in an observational setting.

1.1.2 The Propensity Score

The propensity score is the conditional probability that an individual receives treatment $X = 1$ (or $W = 1$) given their covariates (Lunceford and Davidian, 2004; Rosenbaum and Rubin, 1983). In this section, we focus on the observational setting where treatment is denoted by W. Note that in the randomized setting, the propensity score is the same for each subject in the trial and the probability of receiving $X = 1$ is typically 50% (i.e. $P(X_i = 1|\mathbf{Z}_{1i}) = P(X_i = 1) = 0.5$).

Let $\mathbf{Z}_{2i}$ denote a vector of auxiliary variables that are associated with treatment selection (i.e. the distribution of $\mathbf{Z}_{2i}$ is not balanced between treatment groups), but $\mathbf{Z}_{2i}$ is not related to the response variable, conditional on other observed confounders and auxiliary variables, i.e. $\mathbf{Z}_{2i} \perp Y_i|(W_i, \mathbf{Z}_{1i}, \mathbf{V}_i))$. In this thesis, the propensity score for subject i is defined as

$$\pi_i(\boldsymbol{\xi}_1) = P(W_i = 1|\mathbf{Z}_{1i}, \mathbf{Z}_{2i}; \boldsymbol{\xi}_1) \ , \tag{1.6}$$

where $\boldsymbol{\xi}_1$ is a vector of regression coefficients. Auxiliary variable $\mathbf{V}_i$ is not included here because $W_i \perp \mathbf{V}_i$ by definition. Note that some researchers suggest only including variables that are both predictive of treatment selection and the response in the propensity score (i.e. $\mathbf{Z}_{1i}$ only), while others recommend including all variables that are either predictive of treatment selection or response (i.e. $\mathbf{Z}_{1i}, \mathbf{Z}_{2i}, \mathbf{V}_i$) (Austin et al., 2007a).

The propensity score is a balancing score (Rosenbaum and Rubin, 1983). A balancing score is a function of the data where, conditional on the balancing score, treatment selection and baseline variables are independent, i.e. covariates are balanced between treatment groups. Formally, if $b(\mathbf{Z}_{1i}, \mathbf{Z}_{2i})$ is a balancing score, then

$$(\mathbf{Z}_{1i}, \mathbf{Z}_{2i}) \perp W_i \mid b(\mathbf{Z}_{1i}, \mathbf{Z}_{2i}) \ .$$

Intuitively, this means that subjects with a similar balancing score have a similar covariate distribution. In other words, if covariates $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are summarized among subjects with a similar balancing score within each treatment group, one would expect the summary statistics (e.g. sample means) to be similar across the treatment groups on average.

Rosenbaum and Rubin (1983) showed that if treatment selection is strongly ignorable given $\mathbf{Z}_1$ and $\mathbf{Z}_2$, then treatment selection is strongly ignorable given $\pi(\boldsymbol{\xi}_1)$. Formally, this can be expressed as

$$(Y_{1i}, Y_{0i}) \ \perp \ W_i \mid (\mathbf{Z}_{1i}, \mathbf{Z}_{2i}) \ \Rightarrow \ (Y_{1i}, Y_{0i}) \ \perp \ W_i \mid \pi_i(\boldsymbol{\xi}_1) \ .$$

In estimation of the difference in the average response between the treatment groups (1.1), this is useful, especially when $\mathbf{Z}_1$ is of high dimension (Dehejia and Wahba, 2002); instead of conditioning on $\mathbf{Z}_1$ in the causal model for the response, one can condition on the propensity score to estimate the causal difference in the average response:

$$\begin{aligned}
& E[E(Y_i|W_i = 1, \pi_i(\boldsymbol{\xi}_1)) - E(Y_i|W_i = 0, \pi_i(\boldsymbol{\xi}_1))] \\
= \ & E[E(Y_{1i}|W_i = 1, \pi_i(\boldsymbol{\xi}_1)) - E(Y_{0i}|W_i = 0, \pi_i(\boldsymbol{\xi}_1))] \text{ (consistency)} \\
= \ & E[E(Y_{1i}|\pi_i(\boldsymbol{\xi}_1)) - E(Y_{0i}|\pi_i(\boldsymbol{\xi}_1))] \text{ (strong ignorability)} \\
= \ & E(Y_{1i}) - E(Y_{0i}) \ .
\end{aligned}$$

Because the odds ratio is non-collapsible, we cannot condition on the propensity score to adjust for confounding to ensure consistent estimation of causal odds ratio (1.2) (Austin

et al., 2007b). However, the propensity score can be used as a balancing score in the logistic regression setting in order to obtain consistent estimates of the marginal causal odds ratio by stratification on the propensity score, propensity score matching, and inverse probability of treatment weighting using the propensity score (Austin, 2009; Loux et al., 2017). In this thesis, we focus on inverse probability weighting and introduce this approach in Section 1.1.3.

Before proceeding, we note that in observational studies, the true propensity score for each individual is unknown. Propensity scores can be estimated using the observed data; estimated propensity scores produce sample balance for the probability of exposure and can therefore be used for causal inference (Rosenbaum and Rubin, 1983; Rosenbaum, 1987). Typically, a logistic regression model is used where the independent variables are the confounders, and the dependent variable is the exposure (Rosenbaum and Rubin, 1984).

1.1.3 Inverse Probability of Treatment Weighted Estimating Equations

The inverse probability weighted estimating equation method is used to fit a marginal structural model (MSM). This approach was developed by Robins and colleagues (Hernán et al., 2000; Robins, 2000; Robins et al., 2000) for causal analysis in an observational setting with longitudinal data. In an inverse probability weighted estimating equation, each subject is given a weight which is the inverse of the propensity score for those who are treated, and the inverse of the complement of the propensity score for those who are untreated. Under the causal assumptions listed in Section 1.1.1, and given that the model used to estimate the weights is not misspecified, a weighted model creates a pseudo-population where treatment selection is independent of measured confounders (Cole and Hernán, 2008; Hernán et al., 2000). One can then use observational data to estimate marginal parameters that would be of interest in a randomized trial setting (Tsai et al., 2010).

The following is an example of an inverse probability weighted estimating function with a logistic regression response model, where the response for subject i is denoted by Y_i, the

treatment variable is denoted by W_i and the propensity score is denoted by $\pi_i(\boldsymbol{\xi}_1)$ from (1.6):

$$\mathbf{U}_i(\boldsymbol{\beta};\boldsymbol{\xi}_1) \;=\; \sum_{l=0}^{1} \frac{I(W_i = l)}{\pi_i(\boldsymbol{\xi}_1)^l(1-\pi_i(\boldsymbol{\xi}_1))^{1-l}}\, \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\,(Y_i - \mu_i(\boldsymbol{\beta}))$$

where $\boldsymbol{\beta}$ is a vector of regression parameters, $\boldsymbol{\xi}_1$ are the parameters in the propensity score model, $\mu_i(\boldsymbol{\beta}) = P(Y_i = 1|W_i;\boldsymbol{\beta})$, $\mathbf{D}_i(\boldsymbol{\beta}) = \partial\mu_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, and $V_i(\boldsymbol{\beta}) = \mu_i(\boldsymbol{\beta})(1-\mu_i(\boldsymbol{\beta}))$. Stabilized weights can be used, where the denominator of the weight is unchanged, and a numerator is added which may be, for example, the marginal probability of receiving the treatment or the control. Stabilized weights are less variable, and resulting estimators provide a good basis for inference (Hernán et al., 2000, 2002) since they are often more efficient, but we do not explore them here. Standard statistical packages can be used to implement the inverse probability weighted method (van der Wal and Geskus, 2011).

In the inverse probability weighted method, consistency of the parameter estimates is subject to the selection of an appropriate propensity score model. *Doubly robust* methods, or *augmented inverse probability weighted* methods for controlling for confounding combines an inverse probability weight for confounding and regression modeling so that if either the weight or the regression model is correctly specified, the parameter estimates are consistent (Bang and Robins, 2005), however we do not explore doubly robust methods here.

1.1.4 Subgroup Effects

This thesis focuses on estimation of causal effects within sub-populations. A *sub-group* or *subgroup* is defined as group of patients characterized by a set of common parameters measured at the time that treatment is administered (Yusuf et al., 1991). Subgroup variables are also referred to as *moderators* in the literature (Aguinis et al., 2005). When the causal effect of a treatment depends on the level of, or the value of, another variable, this is referred to as *effect modification* or *interaction* (Altman and Matthews, 1996; Greenland and Morgenstern, 1989). Marginal treatment effects for different levels of a subgroup variable are often the effects of interest in randomized trials (Mazer et al., 2018; Raison et al., 2013; Rothwell, 2005). For example, in a study of acute kidney injury in the setting

of non-cardiac surgery where the treatment of interest is receipt of low-dose aspirin, an important subgroup variable is pre-surgery chronic kidney disease, defined as an estimated glomerular filtration rate (eGFR) < 60 mL/min per 1.73 m^2 (Garg et al., 2014).

Let S denote a binary subgroup variable. The average causal effect of treatment represented by a difference in means (or proportions) within the subgroup of the population with $S_i = 1$ can be defined as followed:

$$E(Y_{1i}|S_i = 1) - E(Y_{0i}|S_i = 1) \ ,$$

and similarly, the average causal effect within the subgroup of the population with $S_i = 0$ is defined as

$$E(Y_{1i}|S_i = 0) - E(Y_{0i}|S_i = 0) \ .$$

In the setting where Y_i is binary and the causal odds ratio within subgroups is of interest, the causal odds ratio for $S_i = 1$ is defined as

$$\frac{P(Y_{1i} = 1|S_i = 1)/[1 - P(Y_{1i} = 1|S_i = 1)]}{P(Y_{0i} = 1|S_i = 1)/[1 - P(Y_{0i} = 1|S_i = 1)]} \ ,$$

and the causal odds ratio for $S_i = 0$ is

$$\frac{P(Y_{1i} = 1|S_i = 0)/[1 - P(Y_{1i} = 1|S_i = 0)]}{P(Y_{0i} = 1|S_i = 0)/[1 - P(Y_{0i} = 1|S_i = 0)]} \ .$$

(Hernán and Robins, 2006; VanderWeele, 2009).

Issues have been raised with multiple testing and inflated type I error rates in studies where multiple subgroup effects are tested (Schulz and Grimes, 2005). To avoid finding spurious subgroup effects, it is recommended that subgroup analyses should be pre-specified and restricted to only one or two clinically relevant subgroups, and in the event that many comparisons are made within multiple subgroups, multiple comparisons must be taken into account (Yusuf et al., 1991). As well, subgroup analyses should be restricted to the primary outcome (Assman et al., 2000).

Randomized trials and observational studies are not typically powered to detect subgroup effects (Brookes et al., 2004). For example, in a simulation study designed to assess

the power of a study under different parameter settings, a significant subgroup effect was detected in between 7% and 64% of simulations (Brookes et al., 2004). Brookes et al. (2004) recommended that, in order to detect an interaction effect of the same magnitude as the overall effect, the sample size should be approximately four times the size required to detect the overall effect.

1.2 Statistical Methods for Incomplete Covariates

Incomplete data frequently arises in both randomized and observational study settings. Incomplete data can arise by design, where subsets of individuals are sampled in order to minimize costs (e.g. in the case of collecting biomarker data); incomplete data can also arise due to participant or investigator non-compliance. Problems can arise when the missing data process is related to the response model. Before describing existing methods to analyze data when covariates are incomplete, we begin with a description of the different types of missing data patterns.

Data are missing completely at random (MCAR) when the missingness process is not dependent on any variables, including the incomplete variable itself. In other words, the subjects with missing covariate data are a random subset of the population of interest. Data are missing at random (MAR) when the missingness process is dependent on variables that are fully observed, but not dependent on the missing variable itself (Bhaskaran and Smeeth, 2014; Sterne et al., 2009). Data are missing not at random (MNAR) when the missingness process is dependent upon unobserved data (Sterne et al., 2009).

A *complete case* analysis is one where only data from individuals with fully observed data are included, and subjects with at least one missing variable are excluded; this is also referred to as *list-wise deletion* (Allison, 2000). Complete case analysis can produce consistent estimates in some settings, although there is typically a loss of information (Austin and Escobar, 2005; Jones, 1996; Paik and Tsai, 1997; Robins et al., 1994). For example, Vach (1994) used a complete case analysis approach in a logistic regression response model with two covariates where one was subject to missing data; when the missing data process was independent of the response, the parameter estimates were consistent.

The term *ignorable missing value mechanism* was introduced by Rubin (1976) to refer to data that is MCAR or MAR (Vach and Blettner, 1991). The term *ignorable* can also be used to describe missing data settings where a complete case analysis is a valid approach, and *non-ignorable* can refer to settings where a complete case analysis would result in biased estimates (Carpenter and Kenward, 2013); we apply this usage throughout the thesis.

Methods to obtain consistent estimates of regression parameters in the setting of non-ignorable missing data have been developed, including inverse probability weighted estimating equations, the EM-algorithm approach, and multiple imputation (Ibrahim et al., 1999; McIsaac and Cook, 2017). In the following sections, we briefly describe these methods.

1.2.1 Inverse Probability Weighted Estimating Equations

Inverse probability weights can be used to account for the effect of missing data in a similar way to inverse weighting for confounding when data are MAR (Chen et al., 2010; Robins et al., 1995; Rotnitzky and Robins, 1995; Seaman and White, 2013). Carpenter et al. (2006) give an intuitive example of the use of inverse probability weighting for missingness. Essentially, inverse probability weighting involves the use of a complete case analysis in fitting the response model, where each individual is given a weight which is the inverse of the conditional probability of their data being observed. Subjects who are less likely to have fully observed data are given more weight while those who are more likely to have observed data are given lower weight. This effectively creates a pseudo-population where the distribution of the covariates explaining the dependence between the outcome and the and the missing data indicator is like that of the original sample. When it is possible to condition on covariates in the response model so that missingness is ignorable, but estimation of conditional causal parameters is not of interest, inverse probability weighted models are useful.

An inverse probability weighted estimating function in a logistic regression response model, where the response for subject i is denoted by Y_i, the covariates are denoted by $\mathbf{X}_i$, the missing data indicator is denoted by R_i, and $\mathcal{D}_i$ represents variables that are predictive

of the missing data process, can take the form

$$\mathbf{U}_i(\boldsymbol{\beta};\boldsymbol{\rho}) \;=\; \frac{R_i}{\pi_i(\boldsymbol{\rho})}\,\mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\,(Y_i-\mu_i(\boldsymbol{\beta})) \tag{1.7}$$

where $\mu_i(\boldsymbol{\beta}) = P(Y_i = 1|\mathbf{X}_i;\boldsymbol{\beta})$, $\boldsymbol{\beta}$ is a vector of regression parameters, $\pi_i(\boldsymbol{\rho}) = P(R_i = 1|\mathcal{D}_i;\boldsymbol{\rho})$, $\boldsymbol{\rho}$ are the parameters in the inverse probability weight model, $\mathbf{D}_i(\boldsymbol{\beta}) = \partial\mu_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, and $V_i(\boldsymbol{\beta}) = \mu_i(\boldsymbol{\beta})(1-\mu_i(\boldsymbol{\beta}))$. Standard statistical packages can be used to implement the inverse probability weighted method (van der Wal and Geskus, 2011). Stabilized weights can be used here as well, where the denominator of the weight is the conditional probability of having observed data, and a numerator is added which may be, for example, the marginal probability of observed data. Stabilized weights are less extreme, and have good properties in terms of estimation and hypothesis testing (Robins et al., 2000), but we do not explore them here.

The parameters in the inverse probability weight are estimated using observed data, therefore a model for the missing data process must be chosen and fitted, and the variability in the parameter estimates must be incorporated in the variance estimate. In Section 2.3.2 we derive the formula for the asymptotic standard error of estimators obtained from inverse probability weighted estimating equations in a general setting, and this derivation can be applied to more simple settings, for example, the above estimating equation (1.7).

Consistency of parameter estimates is subject to the selection of an appropriate missing data model. 'Augmented' inverse probability weighted models, or 'doubly robust' models, are a generalization of inverse probability weighted methods and make use of data from subjects with partially incomplete data. In augmented inverse probability weighting, an inverse probability weight is used as well as an estimate of the conditional expectation of the response for those with incomplete data. Consistency can be achieved as long as one or both of the missing data process model or the conditional expectation of the response are correctly specified (Seaman and White, 2013).

1.2.2 The Expectation-Maximization Algorithm

When covariate data are incomplete and the missing data is ignorable, an expectation-maximization (EM) algorithm can be used to estimate causal treatment effects in the

setting of missing data (Ibrahim et al., 2005). Unlike in the inverse probability weighted method, the EM algorithm makes use of data from subjects with incomplete data. At the E-step, the expectation of the complete data likelihood over the conditional distribution of the missing data is computed given the current estimate of the parameters of interest. In the M-step, the expectation of the E-step is maximized with respect to the parameters of interest to obtain updated estimates. This process is repeated until pre-specified convergence criteria are met. Standard errors of the treatment effect parameters can be estimated using Louis's formula (Louis, 1982; Ibrahim et al., 2005).

When the incomplete covariate is binary, and when there are very few missing covariates, the EM algorithm method is straightforward. However in the setting where the missing covariate is continuous or the dimensionality is high, the EM algorithm method is not ideal since valid inference depends on the correct specification of the conditional covariate distribution.

When covariate data are non-ignorably missing, the EM-algorithm method can be adapted with the addition of a *selection model*, which is a parametric model for the missing data process (Ibrahim et al., 2005).

1.2.3 Multiple Imputation

Multiple imputation is described in detail in Rubin (1987). Multiple imputation involves specifying a distribution for the missing variable(s), imputing, or 'filling-in', the missing values using this distribution while allowing for variability, and repeating this imputation process K times to create K complete datasets. The response model is fitted for each of the complete datasets and the average of the K estimates is computed as the final parameter estimate. A simple formula can be used to compute the variance which incorporates both the variance of each estimate, as well as the the variability between the K estimates. In multiple imputation methods, there may be somewhat of a separation between the imputation model and the analysis model, which is in contrast to EM-algorithm based methods (Schafer, 2003). Also, multiple imputation is different from inverse probability weighting methods because it is not necessary to model the missing data process; this is

useful in some settings, since surveys are not typically designed to capture adequate data for modeling the missing data process.

Selection of variables to include in the imputation model has been widely studied (Carpenter and Kenward, 2013; Collins et al., 2001; Penning de Vries and Groenwold, 2016). The *substantive* response model is defined as the response model of interest. Rubin (1996) offers the following guidelines for choosing imputation covariates: (i) variables that are associated with the missing variable, and (ii) all of the variables that are included in the substantive response model. Generally, any interaction and non-linear terms that will be used in the final analysis model should be included in the imputation model, as well as variables that are predictive of the missing data process (Collins et al., 2001; Rezvan et al., 2015).

The addition of packages for multiple imputation in standard statistical software, including fully conditional specification (FCS) multiple imputation which is useful when there are many incomplete variables and the missing data pattern is non-monotone, has helped multiple imputation become a popular method to deal with missing data.

Multiple imputation methods are designed for the setting where data are MAR (or MCAR). When data are MNAR, multiple imputation methods can be augmented to accommodate the missing data process (Schafer, 2003), or weighted to account for the degree of the departure from the MAR assumption (Carpenter et al., 2007).

1.3 Causal Inference in the Setting of Missing Data

In this thesis, we focus on methods for estimation of marginal causal parameters in non-collapsible response models where an important subgroup variable is incomplete.

Moodie et al. (2008) use a "full-weight" approach to causal analysis with an incomplete confounding variable in a longitudinal treatment setting. The full-weight is the product of two weights: one which accounts for confounding and one which accounts for missingness. In Chapter 2, we apply this method to a single treatment setting with an incomplete subgroup variable (which is not a confounder) and give a method for variance estimation.

Hill (2004) introduced methods for estimating propensity scores within multiply imputed datasets, and Mitra and Reiter (2016) extended this work in comparing two approaches for causal analysis using multiple imputation and propensity score matching. In the first approach, Mitra and Reiter (2016) estimated the propensity score and individuals were matched based on the propensity score within each imputed dataset; the treatment effect was estimated within each dataset and standard methods to combine the treatment effects were used. In the second approach, the treatment effect was estimated within each dataset and then was averaged across each imputed dataset for each subject. Leyrat et al. (2017) further extended these methods in a comparison of propensity score-based models within a multiple imputation setting in analysis of a binary response variable. Research involving the use of multiple imputation for missing covariates in propensity score models, with inverse probability weighted methods for estimating marginal causal effects, has been conducted (Crowe et al., 2010; Eulenburg et al., 2016; Leyrat et al., 2017; Moodie et al., 2008; Qu and Lipkovich, 2009; Seaman and White, 2014). In Chapter 4 we investigate the use of multiple imputation to account for a missing subgroup variable in a causal setting where the covariates in the propensity score are complete and therefore the propensity score is estimated before imputation.

1.4 Incidence of Thrombotic Events in Metastatic Colorectal Cancer Patients

In 2006, bevacizumab (BV, Avastin®) became a publicly funded treatment in Ontario in combination with chemotherapy FOLFIRI (leucovorin, fluorouracil and irinotican) for those with metastatic colorectal cancer (mCRC). Previous studies have shown a survival benefit in patients receiving BV in combination with standard chemotherapy (Hurwitz et al., 2005; Saif and Mehra, 2006; Saltz et al., 2008), however the effect is modest and BV has been associated with increased treatment-related adverse events (Ranpura et al., 2011). Interest lies in comparing the rate of adverse events, including thrombotic events, in mCRC patients receiving BV plus FOLFIRI in comparison to patients receiving FOLFIRI alone. Registry data from patients treated between 2004 and 2011 at the Juravinski Cancer

Centre in Hamilton, Ontario, Canada were used to study the causal treatment effect of BV plus FOLFIRI; see Al-Shamsi et al. (2015) for more information. BV plus FOLFIRI can be used as a second line of therapy or a first line of therapy. Interest lies in estimating the causal effect of BV plus FOLFIRI on the risk of a thrombotic event within each line of treatment subgroup. In Chapter 5, we apply the methods proposed in Chapters 2, 3 and 4 to this registry data in estimating the causal odds ratio of a thrombotic event in each subgroup.

1.5 Outline of Thesis

In Chapter 2, we introduce a doubly inverse probability weighted estimating equation approach to estimate marginal causal odds ratios in an observational setting, where an important subgroup variable is incomplete. One inverse probability weight accounts for the incomplete data, and the other weight accounts for treatment selection. Only complete cases are included in the response model. Consistency results are derived, and a method to obtain estimates of the asymptotic standard error is introduced; the extra variability introduced by estimating two weights is incorporated in the estimation of the asymptotic standard error. We give a method for hypothesis testing and calculation of confidence intervals. Simulation studies show that the doubly weighted estimating equation approach is effective in a non-ignorable missingness setting with confounding, and it is straightforward to implement. It also performs well when the missing data process is ignorable, and/or when confounding is not present.

In Chapter 3, we propose the use of a doubly weighted EM-type algorithm approach to estimating the marginal causal odds ratio in the setting of missing subgroup data. The two inverse probability weights are used to account for confounding and incomplete data. A method to obtain asymptotic standard error estimates is given where the extra variability introduced by estimating the two inverse probability weights, as well as the variability introduced by estimating the parameters of the conditional distribution of the incomplete subgroup variable, is incorporated. Simulation studies show that this method is effective in terms of obtaining consistent estimates of the parameters of interest; however it is difficult

to implement and in certain settings there is a loss of efficiency in comparison to the methods introduced in Chapter 2.

In Chapter 4, we explore the use of multiple imputation with one inverse probability weight for confounding in an observational setting where the subgroup variable is incomplete. We discuss methods to correctly specify the imputation model in the setting where the conditional causal odds ratio is of interest, as well as in the setting where the marginal causal odds ratio is of interest. Standard methods for combining the estimates of the marginal log odds ratios are used. We propose a method for estimating the asymptotic standard error of the estimates, which incorporates both the estimation of the parameters in the weight for confounding, and the multiply imputed datasets. We give a method for hypothesis testing and calculation of confidence intervals. Simulation studies show that this method is efficient and straightforward to implement, but correct specification of the imputation model is necessary.

In Chapter 5, the three methods are used in an application to a cohort study of 418 metastatic colorectal cancer patients. We compare patients who received an experimental chemotherapy with patients who received standard chemotherapy; of interest is estimation of the marginal causal odds ratio of a thrombotic event during the course of treatment or 30 days after treatment is discontinued. The important subgroups are (i) patients who received treatment as a first line strategy, and (ii) patients who received treatment as a second line strategy.

In Chapter 6, we compare and contrast the three methods proposed. We also discuss extensions of the methods presented.

Chapter 2

Doubly Weighted Estimating Equations for Causal Inference with an Incomplete Subgroup Variable

In this chapter, we introduce a doubly weighted estimating equation approach to estimate coefficients of a marginal regression model in an observational setting, where an important subgroup variable is incomplete.

This chapter is organized as follows. In Section 2.1, the notation and models are defined in a complete data setting. Estimation of subgroup effects in a marginal regression model is described in randomized and observational settings. In Section 2.2, estimation of subgroup effects in a marginal regression model is described in randomized and observational settings with an incomplete subgroup variable; the method in the observational setting is the doubly weighted estimating equation approach. In Section 2.3, the consistency and the asymptotic distribution of the estimators from the doubly weighted estimating equation approach are discussed. We give a method for estimating standard errors of marginal regression coefficient estimates obtained using the doubly weighted estimating equation approach. Limiting values of estimators from estimating functions with misspecified weights are derived. In Section 2.4, we describe simulation studies and graphically study the bias introduced by misspecifying the weights in the weighted estimating functions. Concluding

remarks are in Section 2.5.

2.1 Notation and Models

Let Y_i denote a binary response variable, X_i denote a binary treatment variable in a randomized setting, and S_i denote a binary subgroup variable, for subject i in a random sample of n individuals, $i = 1, ..., n$. Variable S is a subgroup variable (or effect modifier) in the sense that the effect of the treatment on the response is dependent upon the level of S. Let $\mathbf{Z}_{1i} = (Z_{11i}, ..., Z_{1p_1i})$ denote a vector of variables that are directly associated with response, and $\mathbf{Z}_{2i} = (Z_{21i}, ..., Z_{2p_2i})$ denote a vector of variables that do not have direct effects on the response. We refer to $\mathbf{Z}_{1i}$ and $\mathbf{Z}_{2i}$ as 'auxiliary' variables; in an observational setting, we further distinguish between $\mathbf{Z}_{1i}$ and $\mathbf{Z}_{2i}$.

We begin by introducing the models and specifying conditional independence assumptions in general terms in the setting of a randomized controlled trial in Section 2.1.1, and proceed to describe the models in the setting of an observational study in Section 2.1.2.

2.1.1 Subgroup Effects in a Randomized Setting

We make the following conditional independence assumptions in a randomized controlled trial setting:

A.1	$Y_i \perp \mathbf{Z}_{2i} \mid (X_i, S_i, \mathbf{Z}_{1i})$
A.2	$X_i \perp (S_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i})$ (randomized setting)
A.3	$\mathbf{Z}_{1i} \perp \mathbf{Z}_{2i}$.

Response variable Y_i is independent of $\mathbf{Z}_{2i}$ given all other variables, which distinguishes $\mathbf{Z}_{2i}$ from $\mathbf{Z}_{1i}$ in the sense that $\mathbf{Z}_{2i}$ is not independently associated with Y_i. Because of randomization, X_i is independent of all other variables that are measured at or a short time before treatment selection. We also assume that $\mathbf{Z}_{1i} \perp \mathbf{Z}_{2i}$ which is an assumption that can later be relaxed. See Figure 2.1 for a causal directed acyclic graph (DAG) which gives a visual representation of the relationships between the variables. In a causal DAG,

an arrow connecting variable X to Y indicates that X is a *parent* of Y, and Y is a *child* of X. The term *directed* indicates that all edges connecting variables are arrows, and the term *acyclic* indicates that none of the directed paths form a closed loop. The DAG in Figure 2.1 is *causal* because every arrow represents the presence of an effect of the parent variable on the child variable (Greenland and Brumback, 2002).

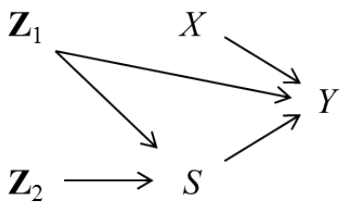

Figure 2.1: Simple causal DAG for treatment or exposure variable X, response Y, and auxiliary variables $\mathbf{Z}_1$ and $\mathbf{Z}_2$, in the context of a randomized controlled trial. Variables $\mathbf{Z}_1$ are causally related to the response, and may be causally related to the subgroup variable S, but are not independently associated with X in the randomized setting. Variables $\mathbf{Z}_2$ are not causally related to the response, and are not associated with treatment in the randomized setting, but may be causally related to the subgroup variable.

The assumptions above allow us to factor the joint probability of $Y_i, X_i, S_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}$ for subject i as

$$\begin{aligned} & P(Y_i, X_i, S_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}) \\ = \; & P(Y_i|X_i, S_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i})P(X_i|S_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i})P(S_i|\mathbf{Z}_{1i}, \mathbf{Z}_{2i})P(\mathbf{Z}_{1i}, \mathbf{Z}_{2i}) \\ = \; & P(Y_i|X_i, S_i, \mathbf{Z}_{1i})P(X_i)P(S_i|\mathbf{Z}_{1i}, \mathbf{Z}_{2i})P(\mathbf{Z}_{1i})P(\mathbf{Z}_{2i}) \end{aligned} \tag{2.1}$$

For simplicity, we consider the setting where $\mathbf{Z}_{1i}$ and $\mathbf{Z}_{2i}$ are scalars denoted by Z_{1i} and Z_{2i}.

Define $\mu_i(\boldsymbol{\vartheta}) = E(Y_i|X_i, S_i, Z_{1i}) = P(Y_i = 1|X_i, S_i, Z_{1i}; \boldsymbol{\vartheta})$. We consider a logistic regression model for the response whereby

$$g(\mu_i(\boldsymbol{\vartheta})) \;\; = \;\; \vartheta_0 + \vartheta_1 X_i + \vartheta_2 S_i + \vartheta_3 X_i S_i + \vartheta_4 X_i Z_{1i} \; . \tag{2.2}$$

Here, $g(\mu) = \log(\mu/(1-\mu))$ is the logit link function, and $\boldsymbol{\vartheta} = (\vartheta_0, \vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4)^T$ is a vector of regression coefficients. We assume that there is no main effect of Z_{1i} on Y_i, but this constraint can be relaxed.

When Z_1 is a subgroup variable, the causal treatment effect depends on the level of Z_1. However, we are in a setting where the investigators do not wish to estimate the effect of Z_1 as a subgroup variable; the only subgroups of interest are $S = 1$ and $S = 0$. This is not uncommon. Often, the true response model is a complex combination of baseline variables, with many subgroup effects. However, investigators generally focus on a small number of subgroup effects due to sample size limitations, and interpretability of causal effects.

Let $E(X_i) = P(X_i = 1) = 0.5$, so that the percentage of treated subjects in the randomized setting is 50%. For the auxiliary variables, let $E(Z_{ji}) = P(Z_{ji} = 1; \zeta_j) = \text{expit}(\zeta_j)$, for $j = 1, 2$, where expit is the inverse of the logit link function. The subgroup variable is modeled as follows:

$$E(S_i|Z_{1i}, Z_{2i}) \;=\; P(S_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_2) \;=\; \text{expit}(\xi_{20} + \xi_{21}Z_{1i} + \xi_{22}Z_{2i}) \;, \tag{2.3}$$

where $\boldsymbol{\xi}_2 = (\xi_{20}, \xi_{21}, \xi_{22})^T$ is a vector of regression coefficients.

Although Z_{1i} is directly causally related to response Y_i through (2.2), we are interested in estimating the marginal causal effect of treatment (X_i) on response (Y_i) for each subgroup defined by values of S_i without conditioning on any other variables. As discussed in Chapter 1, a logistic regression model that includes Z_1 will produce conditional odds ratios of treatment effects for each subgroup rather than marginal odds ratios. We are interested in estimating the parameters from the following marginal logistic regression model for the response:

$$\mu_i(\boldsymbol{\beta}) \;=\; \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 S_i + \beta_3 X_i S_i) \;, \tag{2.4}$$

where $\mu_i(\boldsymbol{\beta}) = E(Y_i|X_i, S_i) = P(Y_i = 1|X_i, S_i; \boldsymbol{\beta})$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ is a vector of regression coefficients. The marginal causal odds ratio when $S = 0$ is $\exp(\beta_1)$ and the marginal causal odds ratio when $S = 1$ is $\exp(\beta_1 + \beta_3)$. The marginal model can be written in terms of the conditional model by taking the expectation of the conditional response

model over Z_{1i} given X_i and S_i:

$$\begin{aligned} P(Y_i = 1|X_i, S_i; \boldsymbol{\beta}) &= E_{Z_{1i}|X_i,S_i}[P(Y_i = 1|X_i, S_i, Z_{1i}; \boldsymbol{\vartheta})] \\ &= E_{Z_{1i}|S_i}[P(Y_i = 1|X_i, S_i, Z_{1i}; \boldsymbol{\vartheta})] \qquad \text{A.2 (randomized setting)} \\ &= \sum_{z_1=0}^{1} P(Y_i = 1|X_i, S_i, z_1; \boldsymbol{\vartheta}) P(Z_{1i} = z_1|S_i; \boldsymbol{\xi}_2, \zeta_1, \zeta_2) \qquad (2.5) \end{aligned}$$

where

$$P(Z_{1i} = z_1|S_i; \boldsymbol{\xi}_2, \zeta_1, \zeta_2) = \frac{\sum_{z_2=0}^{1} P(S_i|Z_{1i} = z_1, Z_{2i} = z_2; \boldsymbol{\xi}_2) P(Z_{1i} = z_1; \zeta_1) P(Z_{2i} = z_2; \zeta_2)}{P(S_i; \boldsymbol{\xi}_2, \zeta_1, \zeta_2)} .$$

Marginal model parameters $\boldsymbol{\beta}$ can be estimated by solving the following estimating equation

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{0} \qquad (2.6)$$

where

$$\mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}(Y_i - \mu_i(\boldsymbol{\beta})) \quad ,$$

with $\mathbf{D}_i(\boldsymbol{\beta}) = \partial\mu_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $V_i(\boldsymbol{\beta}) = \text{var}(Y_i|X_i, S_i) = \mu_i(\boldsymbol{\beta})[1 - \mu_i(\boldsymbol{\beta})]$ (McCullagh and Nelder, 1989). The resulting estimator is denoted by $\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the solution to (2.6).

Under mild regularity conditions, $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal due to the unbiasedness of estimating functions $\mathbf{U}(\boldsymbol{\beta})$, i.e.

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \text{MVN}(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta})^{-1})$$

where

$$\mathcal{I}(\boldsymbol{\beta}) = E[-\partial\mathbf{U}_i(\boldsymbol{\beta})/\boldsymbol{\beta}^T] .$$

To test whether the interaction between treatment and the subgroup variable is significant, one would use standard statistical methods to test the null hypothesis that β_3 is equal to zero. The estimate of the causal odds ratio comparing those who were treated

to those who were not treated, when $S = 0$, is $\exp(\hat{\beta}_1)$. The odds ratio when $S = 1$ is $\exp(\hat{\beta}_1 + \hat{\beta}_3)$.

We have described the notation and models for estimating subgroup effects in a randomized setting. Next, we will describe the notation and models in an observational setting.

2.1.2 Subgroup Effects in an Observational Setting

We now consider the models in the context of an observational study. As noted before, we use W_i to denote the treatment or exposure variable for subject i in an observational setting, instead of X_i. We make the distinction between the two variables because their probability distributions are different. In an observational setting, we assume that the treatment variable, W_i, is not independent of auxiliary variables Z_{1i} and Z_{2i}, as treatment (or exposure) selection is often dependent upon subject characteristics; for simplicity, we continue to assume that the auxiliary variables are scalars. We make the assumption that W_i is independent of subgroup variable S_i conditional on Z_{1i} and Z_{2i}. This assumption is appropriate in settings where, for example, the subgroup variable S_i denotes the presence or absence of some genetic factor, which is information that may not be readily available to the treating physician and hence could not directly affect treatment decisions.

Suppose we make the same conditional independence assumptions as in the randomized setting, with the exception of the assumption for the treatment variable:

B.1	$Y_i \perp Z_{2i} \mid (W_i, S_i, Z_{1i})$	
B.2	$W_i \perp S_i \mid (Z_{1i}, Z_{2i})$	(observational setting)
B.3	$Z_{1i} \perp Z_{2i}$.	

See Figure 2.2 for a causal DAG of the variables in the observational study setting.

The assumptions above allow us to factor the joint probability of $Y_i, W_i, S_i, Z_{1i}, Z_{2i}$ for subject i as

$$\begin{aligned} & P(Y_i, W_i, S_i, Z_{1i}, Z_{2i}) \\ = \; & P(Y_i|W_i, S_i, Z_{1i}, Z_{2i})P(W_i|S_i, Z_{1i}, Z_{2i})P(S_i|Z_{1i}, Z_{2i})P(Z_{1i}, Z_{2i}) \\ = \; & P(Y_i|W_i, S_i, Z_{1i})P(W_i|Z_{1i}, Z_{2i})P(S_i|Z_{1i}, Z_{2i})P(Z_{1i})P(Z_{2i}) \; . \end{aligned}$$

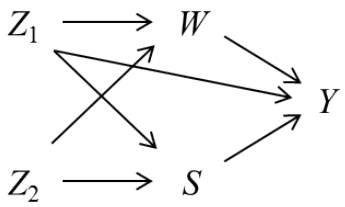

Figure 2.2: Simple directed acyclic graph for treatment variable (or exposure variable) W, response Y, and auxiliary variables Z_1 and Z_2, in the context of an observational study. Variable Z_1 is directly associated with response, whereas Z_2 is not directly associated with response.

We can re-write the logistic regression model for the response using the notation for the treatment variable in an observational setting: $\mu_i(\boldsymbol{\vartheta}) = E(Y_i|W_i, S_i, Z_{1i}) = P(Y_i = 1|W_i, S_i, Z_{1i}; \boldsymbol{\vartheta})$. As before, we consider a logistic regression model for the response whereby

$$\text{logit}(\mu_i(\boldsymbol{\vartheta})) \quad = \quad \vartheta_0 + \vartheta_1 W_i + \vartheta_2 S_i + \vartheta_3 W_i S_i + \vartheta_4 W_i Z_{1i} \ , \tag{2.7}$$

where $\boldsymbol{\vartheta} = (\vartheta_0, \vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4)^T$ is a vector of regression coefficients. We further assume that the probability distribution of the response stays the same if we condition on the same observed values of exposure, subgroup variable and potential confounder. In other words, the conditional odds ratios of treatment are the same whether we are in a randomized setting or in an observational setting, as discussed in Chapter 1. This can be written as

$$P(Y_i = 1|X_i = w, S_i, Z_{1i}; \boldsymbol{\vartheta}) \quad = \quad P(Y_i = 1|W_i = w, S_i, Z_{1i}; \boldsymbol{\vartheta}) \ . \tag{2.8}$$

The parametric model forms of S_i, Z_{1i} and Z_{2i} are the same as in the randomized setting. Since the treatment selection process for subject i depends on auxiliary variables Z_{1i} and Z_{2i}, we assume

$$\pi_i(\boldsymbol{\xi}_1) = P(W_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_1) = \text{expit}(\xi_{10} + \xi_{11} Z_{1i} + \xi_{12} Z_{2i})$$

where $\boldsymbol{\xi}_1 = (\xi_{10}, \xi_{11}, \xi_{12})^T$ are regression parameters. Here, $\pi_i(\boldsymbol{\xi}_1)$ is the propensity score for subject i.

Next, we compute the marginal conditional distribution for response Y given only the treatment, subgroup variable, and their interaction, which is the model that we are interested in fitting. In the observational setting, we have

$$\begin{aligned} & P(Y_i = 1|W_i, S_i) \\ = \; & E_{Z_{1i}|W_i,S_i}\{P(Y_i = 1|W_i, S_i, Z_{1i}; \boldsymbol{\vartheta})\} \\ = \; & \sum_{z_1=0}^{1} P(Y_i = 1|W_i, S_i, Z_{1i} = z_1; \boldsymbol{\vartheta}) P(Z_{1i} = z_1|W_i, S_i) \; . \end{aligned} \tag{2.9}$$

Comparing the forms of $P(Y|W, S)$ from (2.9) and $P(Y|X, S)$ in (2.5), we see that they are not equal in general.

Marginal regression parameters $\boldsymbol{\beta}$ from (2.4) can be estimated in an observational setting using the following weighted estimating equation

$$\widetilde{\mathbf{U}}_1(\boldsymbol{\beta}; \boldsymbol{\xi}_1) \;\; = \;\; \sum_{i=1}^{n} \widetilde{\mathbf{U}}_{1i}(\boldsymbol{\beta}; \boldsymbol{\xi}_1) \tag{2.10}$$

(Robins et al., 2000), where

$$\widetilde{\mathbf{U}}_{1i}(\boldsymbol{\beta}; \boldsymbol{\xi}_1) \;\; = \;\; \sum_{l=0}^{1} \frac{I(W_i = l)}{\pi_i(\boldsymbol{\xi}_1)^l (1 - \pi_i(\boldsymbol{\xi}_1))^{1-l}} \mathbf{D}_i(\boldsymbol{\beta}) [V_i(\boldsymbol{\beta})]^{-1} (Y_i - \mu_i\big(\boldsymbol{\beta}\big)\big) \tag{2.11}$$

with $\mathbf{D}_i(\boldsymbol{\beta}) = \partial \mu_i(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ and $V_i(\boldsymbol{\beta}) = \text{var}(Y_i|W_i, S_i) = \mu_i(\boldsymbol{\beta})[1 - \mu_i(\boldsymbol{\beta})]$. The tilde indicates that a weighted estimating equation is used with a weight for confounding. Let $\tilde{\boldsymbol{\beta}}$ denote the solution to $\widetilde{\mathbf{U}}_1(\boldsymbol{\beta}; \boldsymbol{\xi}_1) = \mathbf{0}$ for fixed $\boldsymbol{\xi}_1$.

An auxiliary estimating function is required to estimate $\boldsymbol{\xi}_1$ and we specify this as

$$\mathbf{U}_2(\boldsymbol{\xi}_1) = \sum_{i=1}^{n} \mathbf{U}_{2i}(\boldsymbol{\xi}_1)$$

where

$$\mathbf{U}_{2i}(\boldsymbol{\xi}_1) = \sum_{i=1}^{n} \mathbf{D}_i(\boldsymbol{\xi}_1) [V_i(\boldsymbol{\xi}_1)]^{-1} (W_i - \pi_i\big(\boldsymbol{\xi}_1\big)\big) \; ,$$

$\mathbf{D}_i(\boldsymbol{\xi}_1) = \partial\pi_i(\boldsymbol{\xi}_1)/\partial\boldsymbol{\xi}_1$ and $V_i(\boldsymbol{\xi}_1) = \text{var}(W_i|Z_{1i}, Z_{2i}) = \pi_i(\boldsymbol{\xi}_1)(1-\pi_i(\boldsymbol{\xi}_1))$. Then let

$$\widetilde{\mathbf{U}}_i(\boldsymbol{\gamma}) = \begin{pmatrix} \widetilde{\mathbf{U}}_{1i}(\boldsymbol{\beta}) \\ \mathbf{U}_{2i}(\boldsymbol{\xi}_1) \end{pmatrix} ,$$

where $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \boldsymbol{\xi}_1^T)^T$. In practice, we use the data to obtain an estimate for $\boldsymbol{\xi}_1$, denoted by $\hat{\boldsymbol{\xi}}_1$, and replace the weight in (2.11) with its estimated counterpart $\pi_i(\hat{\boldsymbol{\xi}}_1) = P(W_i = 1|Z_{1i}, Z_{2i}; \hat{\boldsymbol{\xi}}_1)$. Let $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\xi}}_1^{\ T})^T$.

Under mild regularity conditions and provided the propensity score model is not misspecified, $\tilde{\boldsymbol{\gamma}}$ is consistent and asymptotically normal due to the unbiasedness of estimating functions, i.e.

$$\sqrt{n}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \xrightarrow{D} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\gamma})) .$$

The asymptotic covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\gamma})$ takes the form

$$\boldsymbol{\Sigma}(\boldsymbol{\gamma}) = \mathcal{I}^{-1}(\boldsymbol{\gamma})\mathbf{C}(\boldsymbol{\gamma}) \left[\mathcal{I}^{-1}(\boldsymbol{\gamma})\right]^T ,$$

where

$$\mathcal{I}(\boldsymbol{\gamma}) = E[-\partial\mathbf{U}_i(\boldsymbol{\gamma})/\boldsymbol{\gamma}^T]$$

and

$$\mathbf{C}(\boldsymbol{\gamma}) = E\left[\mathbf{U}_i(\boldsymbol{\gamma})\mathbf{U}_i(\boldsymbol{\gamma})^T\right] .$$

The above covariance matrix takes into account the estimation of $\boldsymbol{\xi}_1$. See Section 2.3.2 for details on estimation of $\boldsymbol{\Sigma}(\boldsymbol{\gamma})$ in a more general setting.

The subgroup effects are estimated in the same way as in the randomized setting, which is described at the end of Section 2.1.1. We have discussed estimation of treatment effects in an observational setting with an important subgroup variable using a weighted estimating function approach. In the next section, we discuss estimation of $\boldsymbol{\beta}$ in the setting where the subgroup variable is not observed for some subjects.

2.2 Inference with Incomplete Subgroup Data

2.2.1 Estimation in the Randomized Setting

In the randomized controlled trial setting, we may experience missing data. We consider the setting where the subgroup variable is not observed for some subjects. A weighted estimating equation approach can be used in this setting.

The subgroup variable S_i is incomplete for some individuals and so we let $R_i = I(S_i \text{ is observed})$. We make the following assumptions in the randomized setting when the subgroup variable is missing for some subjects.

A.0	$R_i \perp (Y_i, X_i, S_i) \mid (Z_{1i}, Z_{2i})$.
A.1	$Y_i \perp Z_{2i} \mid (X_i, S_i, Z_{1i})$
A.2	$X_i \perp (S_i, Z_{1i}, Z_{2i})$ (randomized setting)
A.3	$Z_{1i} \perp Z_{2i}$

As before, response variable Y_i is independent of Z_{2i} given all other variables, which distinguishes Z_{2i} from Z_{1i} in the sense that Z_{2i} is not independently associated with response. The missing data indicator R_i depends on Z_{1i} and Z_{2i} only, so that conditioned on (Z_{1i}, Z_{2i}), R_i is independent of all other variables. In particular, R_i is conditionally independent of the subgroup variable S_i, which means that we do not have a MNAR missing data setting. Because of randomization, X_i is independent of all variables measured at the time of randomization (S_i, Z_{1i}, Z_{2i}). As before, we also assume that $Z_{1i} \perp Z_{2i}$, but note that this assumption can be relaxed, and the methods in this section will still be valid. See Figure 2.3 for a causal DAG which summarizes the conditional independence assumptions between the variables.

The models for generating Y_i, X_i, S_i and Z_{1i} and Z_{2i} are the same as in Section 2.1.1. The logistic regression model for generating R_i is

$$\pi_i(\boldsymbol{\rho}) \;=\; E(R_i|Z_{1i}, Z_{2i}) \;=\; P(R_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\rho}) = \text{expit}(\rho_0 + \rho_1 Z_{1i} + \rho_2 Z_{2i}) \;, \quad (2.12)$$

where $\boldsymbol{\rho} = (\rho_0, \rho_1, \rho_2)^T$ is a vector of regression parameters.

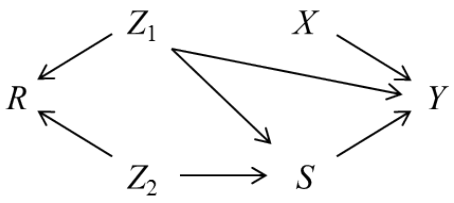

Figure 2.3: Simple causal DAG for treatment X, response Y, a variable that is independently associated with response Z_1, and auxiliary variable Z_2, in the context of a randomized controlled trial. Variable R indicates whether subgroup variable S is observed.

Assumptions A.0 - A.3 allow us to factor the joint probability of $R_i, Y_i, X_i, S_i, Z_{1i}, Z_{2i}$ for subject i as

$$\begin{aligned} P(R_i, Y_i, X_i, S_i, Z_{1i}, Z_{2i}) &= P(R_i|Y_i, X_i, S_i, Z_{1i}, Z_{2i})P(Y_i|X_i, S_i, Z_{1i}, Z_{2i}) \\ &\quad \cdot P(X_i|S_i, Z_{1i}, Z_{2i})P(S_i|Z_{1i}, Z_{2i})P(Z_{1i}, Z_{2i}) \\ &= P(R_i|Z_{1i}, Z_{2i})P(Y_i|X_i, S_i, Z_{1i})P(X_i)P(S_i|Z_{1i}, Z_{2i}) \\ &\quad \cdot \ P(Z_{1i})P(Z_{2i}) \ . \end{aligned} \tag{2.13}$$

Since we are interested in estimating the marginal response model parameters $\boldsymbol{\beta}$, we write the joint probability omitting Z_{1i} as

$$\begin{aligned} P(R_i, Y_i, X_i, S_i, Z_{2i}) &= P(R_i|Y_i, X_i, S_i, Z_{2i})P(Y_i|X_i, S_i, Z_{2i}) \\ &\quad \cdot P(X_i|S_i, Z_{2i})P(S_i|Z_{2i})P(Z_{2i}) \\ &= P(R_i|Y_i, X_i, S_i, Z_{2i})P(Y_i|X_i, S_i)P(X_i)P(S_i|Z_{2i})P(Z_{2i}) \ . \end{aligned} \tag{2.14}$$

We see that the missing data process is not independent of the response model parameters because we cannot factor Y_i from the conditional probability for R_i, and therefore when we omit Z_{1i} from the response model the missingness is non-ignorable.

To handle the missing data problem, we use a weighted estimating function method, with an inverse probability weight for missing data. Only complete cases are included in

the estimation. The weighted estimating function for estimating $\boldsymbol{\beta}$ from (2.4) is given by

$$\overline{\mathbf{U}}_1(\boldsymbol{\beta};\boldsymbol{\rho}) \;=\; \sum_{i=1}^{n} \overline{\mathbf{U}}_{1i}(\boldsymbol{\beta};\boldsymbol{\rho}) \tag{2.15}$$

where

$$\overline{\mathbf{U}}_{1i}(\boldsymbol{\beta};\boldsymbol{\rho}) \;=\; \frac{R_i}{\pi_i(\boldsymbol{\rho})}\,\mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\big(Y_i-\mu_i(\boldsymbol{\beta})\big)\;,$$

with $\mathbf{D}_i(\boldsymbol{\beta}) = \partial\mu_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $V_i(\boldsymbol{\beta}) = \text{var}(Y_i|X_i,S_i) = \mu_i(\boldsymbol{\beta})(1-\mu_i(\boldsymbol{\beta}))$. The overbar indicates the use of an inverse probability weight for missingness. Let $\bar{\boldsymbol{\beta}}$ denote the solution to $\overline{\mathbf{U}}_1(\boldsymbol{\beta};\boldsymbol{\rho}) = \mathbf{0}$ for fixed $\boldsymbol{\rho}$.

Let

$$\mathbf{U}_2(\boldsymbol{\rho}) \;=\; \sum_{i=1}^{n} \mathbf{U}_{2i}(\boldsymbol{\rho})$$

where

$$\mathbf{U}_{2i}(\boldsymbol{\rho}) = \mathbf{D}_i(\boldsymbol{\rho})[V_i(\boldsymbol{\rho})]^{-1}(R_i-\pi_i\big(\boldsymbol{\rho})\big)\;,$$

with $\mathbf{D}_i(\boldsymbol{\rho}) = \partial\pi_i(\boldsymbol{\rho})/\partial\boldsymbol{\rho}$ and $V_i(\boldsymbol{\rho}) = \text{var}(R_i|Z_{1i},Z_{2i}) = \pi_i(\boldsymbol{\rho})(1-\pi_i(\boldsymbol{\rho}))$. Then let

$$\overline{\mathbf{U}}_i(\boldsymbol{\eta}) = \begin{pmatrix} \overline{\mathbf{U}}_{1i}(\boldsymbol{\beta}) \\ \mathbf{U}_{2i}(\boldsymbol{\rho}) \end{pmatrix}\;,$$

where $\boldsymbol{\eta} = (\boldsymbol{\beta}^T,\boldsymbol{\rho}^T)^T$. In practice, we use the data to obtain an estimate for $\boldsymbol{\rho}$, denoted by $\hat{\boldsymbol{\rho}}$, and replace the weight in (2.15) with its estimated counterpart, $\pi_i(\hat{\boldsymbol{\rho}}) = P(R_i = 1|Z_{1i},Z_{2i};\hat{\boldsymbol{\rho}})$.

Under mild regularity conditions, $\overline{\boldsymbol{\eta}} = (\overline{\boldsymbol{\beta}}^T,\hat{\boldsymbol{\rho}}^T)^T$ is consistent and asymptotically normal due to the unbiasedness of estimating functions, i.e.

$$\sqrt{n}(\overline{\boldsymbol{\eta}}-\boldsymbol{\eta}) \;\xrightarrow{D}\; \text{MVN}(\mathbf{0},\boldsymbol{\Sigma}(\boldsymbol{\eta}))\;, \tag{2.16}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\eta})$ takes the form

$$\boldsymbol{\Sigma}(\boldsymbol{\eta}) \;=\; \mathcal{I}^{-1}(\boldsymbol{\eta})\mathbf{C}(\boldsymbol{\eta})\left[\mathcal{I}^{-1}(\boldsymbol{\eta})\right]^T\;.$$

Here,

$$\mathcal{I}(\boldsymbol{\eta}) \;=\; E[-\partial\overline{\mathbf{U}}_i(\boldsymbol{\eta})/\boldsymbol{\eta}^T]$$

and

$$\mathbf{C}(\boldsymbol{\eta}) \;=\; E\left[\overline{\mathbf{U}}_i(\boldsymbol{\eta})\overline{\mathbf{U}}_i(\boldsymbol{\eta})^T\right] \;.$$

The above covariance matrix takes into account the estimation of $\boldsymbol{\rho}$. See Section 2.3.2 for details on estimating $\boldsymbol{\Sigma}(\boldsymbol{\eta})$ in a generalized setting.

We have discussed the use of a single weight in the randomized trial setting, where the subgroup variable is missing for some subjects. Next, we introduce the use of a double inverse probability weight in an observational setting with a missing subgroup variable and confounding.

2.2.2 Estimation in an Observational Setting

Suppose we make the same conditional independence assumptions as in the randomized setting, with the exception of the conditional assumptions for the treatment variable:

B.0 $R_i \perp (Y_i, W_i, S_i) \mid (Z_{1i}, Z_{2i})$
B.1 $Y_i \perp Z_{2i} \mid (W_i, S_i, Z_{1i})$
B.2 $W_i \perp S_i \mid (Z_{1i}, Z_{2i})$ (observational setting)
B.3 $Z_{1i} \perp Z_{2i}$

See Figure 2.4 for a causal DAG that summarizes the relationships between the variables. The models to generate Y_i, W_i, S_i, Z_{1i} and Z_{2i} are the same as in Section 2.1.2 and R_i is generated in the same way was described in Section 2.2.1.

The assumptions allow us to factor the joint probability of $R_i, Y_i, W_i, S_i, Z_{1i}, Z_{2i}$ for subject i as

$$\begin{aligned} P(R_i, Y_i, W_i, S_i, Z_{1i}, Z_{2i}) \;&=\; P(R_i|Y_i, W_i, S_i, Z_{1i}, Z_{2i})P(Y_i|W_i, S_i, Z_{1i}, Z_{2i}) \\ &\quad \cdot\; P(W_i|S_i, Z_{1i}, Z_{2i})P(S_i|Z_{1i}, Z_{2i})P(Z_{1i}, Z_{2i}) \\ &=\; P(R_i|Z_{1i}, Z_{2i})P(Y_i|W_i, S_i, Z_{1i}) \\ &\quad \cdot\; P(W_i|Z_{1i}, Z_{2i})P(S_i|Z_{1i}, Z_{2i})P(Z_{1i})P(Z_{2i}) \qquad (2.17) \end{aligned}$$

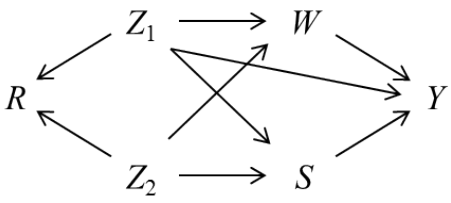

Figure 2.4: Simple causal DAG for treatment (W), response (Y), confounding variable (Z_1), and an auxiliary variable (Z_2), in the context of an observational study. Variable R indicates whether the subgroup variable is observed.

However, interest lies in fitting a marginal model, where Z_{1i} is not included in the response model. If we omit Z_{1i} from the joint probability (2.17) and factor $P(R_i, Y_i, W_i, S_i, Z_{2i})$ as in Section 2.2.1, we see that the missingness is non-ignorable.

Because the missing data is non-ignorable in the marginal causal model setting, we use a doubly inverse probability weighted estimating equation approach for estimating parameters in a marginal model using observational data. As in the previous section, we restrict attention to individuals with complete data and use inverse probability weights for both missingness and treatment selection. The doubly weighted estimating equation is

$$\widetilde{\overline{\mathbf{U}}}_1(\boldsymbol{\beta};\boldsymbol{\psi}) \;=\; \sum_{i=1}^{n} \widetilde{\overline{\mathbf{U}}}_{1i}(\boldsymbol{\beta};\boldsymbol{\psi}) \tag{2.18}$$

where

$$\widetilde{\overline{\mathbf{U}}}_{1i}(\boldsymbol{\beta};\boldsymbol{\psi}) \;=\; \sum_{l=0}^{1} \frac{R_i}{\pi_i(\boldsymbol{\rho})} \, \frac{I(W_i = l)}{\pi_i(\boldsymbol{\xi}_1)^l (1-\pi_i(\boldsymbol{\xi}_1))^{1-l}} \; \mathbf{D}_i(\boldsymbol{\beta}) [V_i(\boldsymbol{\beta})]^{-1} \big(Y_i - \mu_i(\boldsymbol{\beta})\big) \;,$$

with $\boldsymbol{\psi} = (\boldsymbol{\rho}^T, \boldsymbol{\xi}_1^T)^T$, $\mathbf{D}_i(\boldsymbol{\beta}) = \partial \mu_i(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$ and $V_i(\boldsymbol{\beta}) = \text{var}(Y_i | X_i, S_i) = \mu_i(\boldsymbol{\beta})(1 - \mu_i(\boldsymbol{\beta}))$. The tilde and overbar denote that we are weighting for treatment selection and missingness, respectively. Let $\widetilde{\overline{\boldsymbol{\beta}}}$ denote the solution to $\widetilde{\overline{\mathbf{U}}}_1(\boldsymbol{\beta};\boldsymbol{\psi}) = \mathbf{0}$ for fixed $\boldsymbol{\psi}$.

Let

$$\mathbf{U}_2(\boldsymbol{\psi}) = \sum_{i=1}^{n} \mathbf{U}_{2i}(\boldsymbol{\psi})$$

denote an unbiased estimating function for $\boldsymbol{\psi}$ where

$$\mathbf{U}_{2i}(\boldsymbol{\psi}) = \begin{pmatrix} \mathbf{U}_{2i}(\boldsymbol{\rho}) \\ \mathbf{U}_{2i}(\boldsymbol{\xi}_1) \end{pmatrix} . \tag{2.19}$$

Let $\hat{\boldsymbol{\psi}}$ denote the solution to $\mathbf{U}_2(\boldsymbol{\psi}) = \mathbf{0}$. Let $\boldsymbol{\omega} = (\boldsymbol{\beta}^T, \boldsymbol{\psi}^T)^T$ be the vector containing all parameters with $\widetilde{\widetilde{\boldsymbol{\omega}}} = (\widetilde{\widetilde{\boldsymbol{\beta}}}^T, \hat{\boldsymbol{\psi}}^T)^T$, and write

$$\widetilde{\widetilde{\mathbf{U}}}(\boldsymbol{\omega}) = \sum_{i=1}^{n} \widetilde{\widetilde{\mathbf{U}}}_i(\boldsymbol{\omega}) = \sum_{i=1}^{n} \begin{pmatrix} \widetilde{\widetilde{\mathbf{U}}}_{1i}(\boldsymbol{\beta}; \boldsymbol{\psi}) \\ \mathbf{U}_{2i}(\boldsymbol{\psi}) \end{pmatrix} = \mathbf{0} . \tag{2.20}$$

In practice, to obtain an estimate for $\boldsymbol{\beta}$, we first estimate $\boldsymbol{\xi}_1$ using the full dataset (including subjects where S_i is missing, since only W_i and Z_{1i}, Z_{2i} are needed) and replace $\pi_i(\boldsymbol{\xi}_1)$ with its estimated counterpart, $\pi_i(\hat{\boldsymbol{\xi}}_1) = P(W_i = 1 | Z_{1i}, Z_{2i}; \hat{\boldsymbol{\xi}}_1)$. Second, we estimate $\boldsymbol{\rho}$ using the full dataset, and replace $\pi_i(\boldsymbol{\rho})$ with its estimated counterpart $\pi_i(\hat{\boldsymbol{\rho}}) = P(R_i = 1 | Z_{1i}, Z_{2i}; \hat{\boldsymbol{\rho}})$. Standard software can be used to solve for $\widetilde{\widetilde{\boldsymbol{\beta}}}$. For example, in `R`, we use the `glm` function, with 'family=binomial(link="logit")' for logistic regression (R Core Team, 2018). We can create a new variable which is the product of the two inverse probability weights, and specify the weight using the 'weights' option.

In the following section, we show that

$$E_{R_i, Y_i, X_i, S_i, Z_{1i}, Z_{2i}}[\widetilde{\widetilde{\mathbf{U}}}_{1i}(\boldsymbol{\beta}; \boldsymbol{\psi})] = \mathbf{0} , \tag{2.21}$$

and therefore $\widetilde{\widetilde{\boldsymbol{\beta}}}$ is a consistent estimator of $\boldsymbol{\beta}$ from (2.4). We also derive the asymptotic distribution for $\widetilde{\widetilde{\boldsymbol{\beta}}}$.

2.3 Statistical Inference

2.3.1 Consistency of Doubly Weighted Estimators for $\boldsymbol{\beta}$

Theorem 1. *$\widetilde{\widetilde{\boldsymbol{\beta}}}$, the solution to $\widetilde{\widetilde{\mathbf{U}}}_1(\boldsymbol{\beta}; \boldsymbol{\psi}) = \sum_{i=1}^{n} \widetilde{\widetilde{\mathbf{U}}}_i(\boldsymbol{\beta}; \boldsymbol{\psi}) = \mathbf{0}$, is a consistent estimator of $\boldsymbol{\beta}$ from (2.4).*

Proof: Consistency results can be established by showing that the doubly inverse probability weighted estimating function is an unbiased estimating function. Note that, although we have observed treatment W_i for subject i, we write the conditional expectation of Y_i with respect to X_i because we are estimating $\boldsymbol{\beta}$ from (2.4); in other words, we write

$$\mu_i(\boldsymbol{\beta}) = E(Y_i|X_i, S_i; \boldsymbol{\beta}) = P(Y_i = 1|X_i, S_i; \boldsymbol{\beta}) \ .$$

First, we take the conditional expectation of $\widetilde{\widetilde{\mathbf{U}}}_{1i}(\boldsymbol{\beta}; \boldsymbol{\psi})$ from (2.18) with respect to R_i given $Y_i, W_i, S_i, Z_{1i}, Z_{2i}$ which by assumption B.0 gives

$$\sum_{l=0}^{1} \frac{P(R_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\rho})}{\pi_i(\boldsymbol{\rho})} \ \frac{I(W_i = l)}{\pi_i(\boldsymbol{\xi}_1)^l(1 - \pi_i(\boldsymbol{\xi}_1))^{1-l}} \ \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\Big\{Y_i - P(Y_i = 1|X_i = l, S_i; \boldsymbol{\beta})\Big\} \ .$$

Since $\pi_i(\boldsymbol{\rho}) = P(R_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\rho})$, this simplifies to

$$\sum_{l=0}^{1} \frac{I(W_i = l)}{\pi_i(\boldsymbol{\xi}_1)^l(1 - \pi_i(\boldsymbol{\xi}_1))^{1-l}} \ \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\Big\{Y_i - P(Y_i = 1|X_i = l, S_i; \boldsymbol{\beta})\Big\} \ .$$

Next, we take the conditional expectation of this with respect to $Y_i|W_i, S_i, Z_{1i}, Z_{2i}$ which by assumption B.1 gives

$$\sum_{l=0}^{1} \frac{I(W_i = l)}{\pi_i(\boldsymbol{\xi}_1)^l(1 - \pi_i(\boldsymbol{\xi}_1))^{1-l}} \ \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\Big\{P(Y_i = 1|W_i, S_i, Z_{1i}; \boldsymbol{\vartheta}) - P(Y_i = 1|X_i = l, S_i; \boldsymbol{\beta})\Big\} \ .$$

Next, we take the conditional expectation of this with respect to $W_i|S_i, Z_{1i}, Z_{2i}$ which by assumption B.2 gives

$$\sum_{w=0}^{1}\sum_{l=0}^{1} \frac{I(w = l)P(W_i = w|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_1)}{\pi_i(\boldsymbol{\xi}_1)^l(1 - \pi_i(\boldsymbol{\xi}_1))^{1-l}} \ \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\Big\{P(Y_i = 1|W_i = w, S_i, Z_{1i}; \boldsymbol{\vartheta}) - P(Y_i = 1|X_i = l, S_i; \boldsymbol{\beta})\Big\} \ .$$

The indicator $I(W_i = l)$ ensures only one term is retained (when $w = l$ for $w = 0, 1$).

Therefore we have

$$\sum_{w=0}^{1} \frac{P(W_i = w|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_1)}{\pi_i(\boldsymbol{\xi}_1)^w(1-\pi_i(\boldsymbol{\xi}_1))^{1-w}} \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\Big\{P(Y_i = 1|W_i = w, S_i, Z_{1i}; \boldsymbol{\vartheta}) - P(Y_i = 1|X_i = w, S_i; \boldsymbol{\beta})\Big\}$$

$$= \sum_{w=0}^{1} \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\Big\{P(Y_i = 1|W_i = w, S_i, Z_{1i}; \boldsymbol{\vartheta}) - P(Y_i = 1|X_i = w, S_i; \boldsymbol{\beta})\Big\} .$$

Finally, we take the expectation of this with respect to S_i, Z_{1i}, Z_{2i} which by assumption B.3 gives

$$\sum_{z_2=0}^{1}\sum_{z_1=0}^{1}\sum_{s=0}^{1}\sum_{w=0}^{1} \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\Big\{P(Y_i = 1|W_i = w, S_i = s, Z_{1i} = z_1; \boldsymbol{\vartheta}) - P(Y_i = 1|X_i = w, S_i = s; \boldsymbol{\beta})\Big\} \cdot P(S_i = s|Z_{1i} = z_1, Z_{2i} = z_2; \boldsymbol{\xi}_2)P(Z_{1i} = z_1; \zeta_1)P(Z_{2i} = z_2; \zeta_2) .$$

From (2.8) we can replace W_i with X_i in the conditional probability of the response, and therefore this can be re-written as

$$\sum_{z_2=0}^{1}\sum_{z_1=0}^{1}\sum_{s=0}^{1}\sum_{w=0}^{1} \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\Big\{P(Y_i = 1|X_i = w, S_i = s, Z_{1i} = z_1; \boldsymbol{\vartheta}) - P(Y_i = 1|X_i = w, S_i = s; \boldsymbol{\beta})\Big\} \cdot P(S_i = s|Z_{1i} = z_1, Z_{2i} = z_2; \boldsymbol{\xi}_2)P(Z_{1i} = z_1; \zeta_1)P(Z_{2i} = z_2; \zeta_2) .$$

From (2.5), we have

$$P(Y_i = 1|X_i, S_i; \boldsymbol{\beta}) = \sum_{z_2=0}^{1}\sum_{z_1=0}^{1} P(Y_i = 1|X_i, S_i, Z_{1i} = z_1; \boldsymbol{\vartheta}) \cdot \frac{P(S_i|Z_{1i} = z_1, Z_{2i} = z_2; \boldsymbol{\xi}_2)P(Z_{1i} = z_1; \zeta_1)P(Z_{2i} = z_2; \zeta_2)}{P(S_i; \boldsymbol{\xi}_2, \zeta_1, \zeta_2)}$$

in the context of a randomized controlled trial. Therefore,

$$\begin{aligned}
&\sum_{z_2=0}^{1}\sum_{z_1=0}^{1} P(Y_i=1|X_i,S_i,Z_{1i}=z_1;\boldsymbol{\vartheta})P(S_i|Z_{1i}=z_1,Z_{2i}=z_2;\boldsymbol{\xi}_2)\\
&\quad\cdot P(Z_{1i}=z_1;\zeta_1)P(Z_{2i}=z_2;\zeta_2)\\
=\;& P(Y_i=1|X_i,S_i;\boldsymbol{\beta})P(S_i)\ .
\end{aligned}$$

Then, the conditional expectation of the estimating equation can be written as

$$\begin{aligned}
&\sum_{s=0}^{1}\sum_{w=0}^{1}\mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}P(Y_i=1|X_i=w,S_i=s;\boldsymbol{\beta})P(S_i=s)\\
&-\sum_{s=0}^{1}\sum_{w=0}^{1}\mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}P(Y_i=1|X_i=w,S_i=s;\boldsymbol{\beta})\\
&\cdot\sum_{z_1=0}^{1}\sum_{z_2=0}^{1}P(S_i=s|Z_{1i}=z_1,Z_{2i}=z_2;\boldsymbol{\xi}_2)P(Z_{1i}=z_1;\zeta_1)P(Z_{2i}=z_2;\zeta_2)\\
=\;&\sum_{s=0}^{1}\sum_{w=0}^{1}\mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\Big\{P(Y_i=1|X_i=w,S_i=s;\boldsymbol{\beta})P(S_i=s)\\
&-P(Y_i=1|X_i=w,S_i=s;\boldsymbol{\beta})P(S_i=s)\Big\}\\
=\;&\mathbf{0}\ .
\end{aligned}$$

Therefore, $\widetilde{\widetilde{\boldsymbol{\beta}}}$ is a consistent estimator of $\boldsymbol{\beta}$ from (2.4).

2.3.2 Derivation of the Asymptotic Covariance

The following asymptotic variance derivation is based on the theory introduced by Newey and McFadden (1994) and Robins et al. (1995). We present a method for deriving the asymptotic covariance matrix for $\widetilde{\widetilde{\boldsymbol{\beta}}}$ which is the solution to $\widetilde{\widetilde{\mathbf{U}}}_1(\boldsymbol{\beta};\boldsymbol{\psi})=\mathbf{0}$. Recall that $\boldsymbol{\omega}=(\boldsymbol{\beta}^T,\boldsymbol{\psi}^T)^T$ is the vector containing all parameters with $\widetilde{\widetilde{\boldsymbol{\omega}}}=(\widetilde{\widetilde{\boldsymbol{\beta}}}^T,\hat{\boldsymbol{\psi}}^T)^T$. Since

$$\widetilde{\widetilde{\mathbf{U}}}(\widetilde{\widetilde{\boldsymbol{\omega}}})=\widetilde{\widetilde{\mathbf{U}}}(\boldsymbol{\omega})+\frac{\partial\widetilde{\widetilde{\mathbf{U}}}(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}^T}(\widetilde{\widetilde{\boldsymbol{\omega}}}-\boldsymbol{\omega})+o_p\left(\frac{1}{\sqrt{n}}\right)$$

then

$$\sqrt{n}(\widetilde{\widetilde{\boldsymbol{\omega}}} - \boldsymbol{\omega}) = \left[-\frac{1}{n} \frac{\partial \widetilde{\widetilde{\mathbf{U}}}(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}^T} \right]^{-1} \left[\frac{1}{\sqrt{n}} \widetilde{\widetilde{\mathbf{U}}}(\boldsymbol{\omega}) \right] + o_p(1) \tag{2.22}$$

where

$$-\frac{1}{n} \frac{\partial \widetilde{\widetilde{\mathbf{U}}}(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}^T} = -\frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} \partial \widetilde{\widetilde{\mathbf{U}}}_{1i}(\boldsymbol{\beta}; \boldsymbol{\psi}) / \partial \boldsymbol{\beta}^T & \partial \widetilde{\widetilde{\mathbf{U}}}_{1i}(\boldsymbol{\beta}; \boldsymbol{\psi}) / \partial \boldsymbol{\psi}^T \\ \mathbf{0} & \partial \mathbf{U}_{2i}(\boldsymbol{\psi}) / \partial \boldsymbol{\psi}^T \end{bmatrix} . \tag{2.23}$$

As $n \to \infty$, (2.23) converges in probability to

$$E(-\partial \widetilde{\widetilde{\mathbf{U}}}_i(\boldsymbol{\omega}) / \partial \boldsymbol{\omega}^T) = \begin{bmatrix} E(-\partial \widetilde{\widetilde{\mathbf{U}}}_{1i}(\boldsymbol{\beta}; \boldsymbol{\psi}) / \partial \boldsymbol{\beta}^T) & E(-\partial \widetilde{\widetilde{\mathbf{U}}}_{1i}(\boldsymbol{\beta}; \boldsymbol{\psi}) / \partial \boldsymbol{\psi}^T) \\ & E(-\partial \widetilde{\widetilde{\mathbf{U}}}_{i2}(\boldsymbol{\psi}) / \partial \boldsymbol{\psi}^T) \end{bmatrix}$$

$$= \begin{bmatrix} \mathcal{I}_{11}(\boldsymbol{\omega}) & \mathcal{I}_{12}(\boldsymbol{\omega}) \\ 0 & \mathcal{I}_{22}(\boldsymbol{\omega}) \end{bmatrix} = \mathcal{I}(\boldsymbol{\omega}) .$$

Let

$$\mathbf{C}(\boldsymbol{\omega}) = E\left[\widetilde{\widetilde{\mathbf{U}}}_i(\boldsymbol{\omega}) \widetilde{\widetilde{\mathbf{U}}}_i(\boldsymbol{\omega})^T \right] .$$

Then

$$\sqrt{n}(\widetilde{\widetilde{\boldsymbol{\omega}}} - \boldsymbol{\omega}) \xrightarrow{D} MVN(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\omega}))$$

where $\boldsymbol{\Sigma}(\boldsymbol{\omega}) = \mathcal{I}^{-1}(\boldsymbol{\omega}) \mathbf{C}(\boldsymbol{\omega}) \left[\mathcal{I}^{-1}(\boldsymbol{\omega}) \right]^T$ (Cook et al., 2013; Newey and McFadden, 1994; Robins et al., 1995). Variance estimates are computed by replacing the expectations with their empirical counterparts to obtain

$$\boldsymbol{\Sigma}(\widetilde{\widetilde{\boldsymbol{\omega}}}) = I^{-1}(\widetilde{\widetilde{\boldsymbol{\omega}}}) \mathbf{C}(\widetilde{\widetilde{\boldsymbol{\omega}}}) \left[I^{-1}(\widetilde{\widetilde{\boldsymbol{\omega}}}) \right]^T \tag{2.24}$$

where

$$I(\widetilde{\widetilde{\boldsymbol{\omega}}}) = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} \frac{\partial \widetilde{\widetilde{\mathbf{U}}}_{1i}(\boldsymbol{\beta}; \boldsymbol{\psi})}{\partial \boldsymbol{\beta}^T} & \frac{\partial \widetilde{\widetilde{\mathbf{U}}}_{1i}(\boldsymbol{\beta}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}^T} \\ \mathbf{0} & \frac{\partial \mathbf{U}_{2i}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}^T} \end{bmatrix} \Bigg|_{\boldsymbol{\omega} = \tilde{\tilde{\boldsymbol{\omega}}}}$$

and

$$\mathbf{C}(\widetilde{\widetilde{\boldsymbol{\omega}}}) = \frac{1}{n} \sum_{i=1}^{n} \widetilde{\widetilde{\mathbf{U}}}_i(\boldsymbol{\omega}) \widetilde{\widetilde{\mathbf{U}}}_i^T(\boldsymbol{\omega}) \Bigg|_{\boldsymbol{\omega} = \tilde{\tilde{\boldsymbol{\omega}}}} .$$

The upper block of matrix $\boldsymbol{\Sigma}(\widetilde{\widetilde{\boldsymbol{\omega}}})$ from (2.24) is the estimated covariance matrix for $\widetilde{\widetilde{\boldsymbol{\beta}}}$.

We have derived the asymptotic distribution of this estimator, and we have given a method for computing an estimate of the variance. We can use this information for inference for $\boldsymbol{\beta}$. Suppose we are interested in testing the hypothesis that the main effect for the treatment variable is equal to zero. To do this, we compute a Wald statistic of the form

$$\frac{\widetilde{\widetilde{\beta}}_1 - 0}{\text{s.e.}(\widetilde{\widetilde{\beta}}_1)}$$

where

$$\text{s.e.}(\widetilde{\widetilde{\beta}}_1) = \sqrt{\boldsymbol{\Sigma}(\widetilde{\widetilde{\boldsymbol{\omega}}})_{[2,2]}/n}$$

and assess significance using a reference standard normal distribution which this follows asymptotically under the null hypothesis. We can also construct a $100(1-\alpha)\%$ confidence interval as

$$\left(\widetilde{\widetilde{\beta}}_1 - z_{\alpha/2}\,\text{s.e.}(\widetilde{\widetilde{\beta}}_1)\ ,\ \widetilde{\widetilde{\beta}}_1 + z_{\alpha/2}\,\text{s.e.}(\widetilde{\widetilde{\beta}}_1)\right)$$

where z is a variable with standard normal distribution and $P(z < -|z_\alpha|) = \alpha$. Other coefficients can be tested in a likewise fashion.

2.3.3 Some Results on Misspecified Estimating Functions

We next derive the limiting values of estimators of $\boldsymbol{\beta}$ in an observational setting with a missing subgroup variable, where we omit all weights from the estimating function.

To calculate the limiting values, we compute the expectation of the misspecified estimating function over all of the data. When we fail to weight for confounding and missingness, and use only complete cases, the estimating function is

$$\mathbf{U}(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n} R_i\, \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\big(Y_i - \mu_i(\boldsymbol{\beta})\big)\ . \tag{2.25}$$

By taking the expectation of the misspecified estimating function in a similar way to Section 2.3.1 (omitting the i subscript), we obtain the following expression:

$$\sum_{z_1=0}^{1}\sum_{z_2=0}^{1}\sum_{s=0}^{1}\sum_{w=0}^{1} \mathbf{D}(\boldsymbol{\beta})[V(\boldsymbol{\beta})]^{-1}\big[P(Y=1|W=w,s,z_1;\boldsymbol{\vartheta}) - P(Y=1|X=w,s;\boldsymbol{\beta})\big] \\ \cdot P(R=1|z_1,z_2;\boldsymbol{\rho})P(w|z_1,z_2;\boldsymbol{\xi}_1)P(s|z_1,z_2;\boldsymbol{\xi}_2)P(z_1;\zeta_1)P(z_2;\zeta_2) \quad .$$

Let $\boldsymbol{\beta}^\dagger$ be the solution to setting the above expectation equal to $\mathbf{0}$. We can solve for $\boldsymbol{\beta}^\dagger$ in terms of the other parameters. By considering the setting with $w = s = 0$, we can solve for $\beta_0^\dagger$:

$$\sum_{z_1=0}^{1}\sum_{z_2=0}^{1} P(R=1|z_1,z_2;\boldsymbol{\rho})\ \mathbf{D}(\boldsymbol{\beta})[V(\boldsymbol{\beta})]^{-1}\ \big[\text{expit}(\vartheta_0) - \text{expit}(\beta_0^\dagger)\big] \\ \cdot P(W=0|z_1,z_2;\boldsymbol{\xi}_1)P(S=0|z_1,z_2;\boldsymbol{\xi}_2)P(z_1;\zeta_1)P(z_2;\zeta_2) = \mathbf{0} \quad ,$$

therefore

$$\beta_0^\dagger = \vartheta_0 \ .$$

By setting $w = 1$ and $s = 0$, we can solve for $\beta_1^\dagger$:

$$\beta_1^\dagger \ = \ \text{logit}\left\{\frac{Q_1}{Q_2}\right\} - \beta_0^\dagger \ ,$$

where

$$Q_1 \ = \ \sum_{z_1=0}^{1}\sum_{z_2=0}^{1}\Big\{\text{expit}(\vartheta_0+\vartheta_1+\vartheta_4 z_1)\ \text{expit}(\rho_0+\rho_1 z_1+\rho_2 z_2)\ \text{expit}(\xi_{10}+\xi_{11}z_1+\xi_{12}z_2) \\ \cdot (1-\text{expit}(\xi_{20}+\xi_{21}z_1+\xi_{22}z_2))P(z_1;\zeta_1)P(z_2;\zeta_2)\Big\}$$

and

$$Q_2 \ = \ \sum_{z_1=0}^{1}\sum_{z_2=0}^{1}\Big\{\text{expit}(\rho_0+\rho_1 z_1+\rho_2 z_2)\ \text{expit}(\xi_{10}+\xi_{11}z_1+\xi_{12}z_2) \\ \cdot (1-\text{expit}(\xi_{20}+\xi_{21}z_1+\xi_{22}z_2))\ P(z_1;\zeta_1)P(z_2;\zeta_2)\Big\} \ .$$

By setting $w = 0$ and $s = 1$, we can solve for $\beta_2^\dagger$:

$$\beta_2^\dagger = \vartheta_2 \ .$$

And finally, by setting $w = s = 1$, we can solve for $\beta_3^\dagger$:

$$\beta_3^\dagger \ = \ \text{logit}\left\{\frac{Q_3}{Q_4}\right\} - (\beta_0^\dagger + \beta_1^\dagger + \beta_2^\dagger) \ ,$$

where

$$\begin{aligned} Q_3 \ = \ & \sum_{z_1=0}^{1}\sum_{z_2=0}^{1} \text{expit}(\vartheta_0 + \vartheta_1 + \vartheta_2 + \vartheta_3 + \vartheta_4 z_1)\ \text{expit}(\rho_0 + \rho_1 z_1 + \rho_2 z_2) \\ & \text{expit}(\xi_{10} + \xi_{11} z_1 + \xi_{12} z_2)\ \text{expit}(\xi_{20} + \xi_{21} z_1 + \xi_{22} z_2) P(z_1;\zeta_1) P(z_2;\zeta_2) \end{aligned}$$

and

$$\begin{aligned} Q_4 \ = \ & \sum_{z_1=0}^{1}\sum_{z_2=0}^{1} \text{expit}(\rho_0 + \rho_1 z_1 + \rho_2 z_2)\ \text{expit}(\xi_{10} + \xi_{11} z_1 + \xi_{12} z_2) \\ & \text{expit}(\xi_{20} + \xi_{21} z_1 + \xi_{22} z_2)\ P(z_1;\zeta_1) P(z_2;\zeta_2) \ . \end{aligned}$$

In a similar way, we can solve for the limiting values when we (i) weight for treatment selection only (i.e. solve for $\tilde{\boldsymbol{\beta}}^\dagger$) or (ii) weight for missingness only (i.e. solve for $\overline{\boldsymbol{\beta}}^\dagger$). When we weight for both treatment selection and missingness, $\tilde{\overline{\boldsymbol{\beta}}}^\dagger = \boldsymbol{\beta}$ when the weights are correctly specified (see Section 2.3.1).

2.4 Simulation Studies

Here, we describe simulation studies conducted in order to explore the degree of bias introduced by misspecifying the weights in weighted estimating equations. We are most interested in assessing the consistency and relative efficiency of our estimates of β_1 and $\beta_1 + \beta_3$ from (2.4); β_1 is the marginal causal log odds ratio of response comparing those who received the treatment to those who did not when $S = 0$ and $\beta_1 + \beta_3$ is the log odds ratio when $S = 1$.

The parameters are set so that we are able to examine the effect of misspecified weighted estimating functions in four distinct settings: (i) treatment is randomized and the missing data process is ignorable, (ii) treatment is randomized and the missing data process is non-ignorable, (iii) observational data with ignorable missing data, (iv) observational data with non-ignorable missing data.

In our simulation studies, we use four methods to analyze the data in each of the parameter settings: complete case logistic regression analysis with (i) no weights, (ii) one weight for missing data, (iii) one weight for treatment selection, and (iv) weights for both missing data and treatment selection. When we use method (iv), we are using the doubly weighted estimating equation $\widetilde{\widetilde{\mathbf{U}}}(\boldsymbol{\beta};\boldsymbol{\psi})$ from (2.18). When we use method (i), we are conducting a naïve complete case analysis.

In addition to our simulation studies, we give a visual representation of the difference between the true values, and the limiting values, using estimating equations with misspecified weights based on the material in the previous section. The parameter settings described in the following section are used in the simulation studies as well as the limiting values figures.

2.4.1 Parameter Settings

The parameter specifications for our simulation studies and limiting values figures are given here. The model for the response, written in terms of $\boldsymbol{\beta}$ from (2.4), is given by

$$\mu_i(\boldsymbol{\beta}) \;=\; \text{expit}(\text{logit}(0.5) + \log(0.8)X_i + \log(1.2)S_i + \log(1.2)X_iS_i) \;.$$

The model that is used to generate the response variable is the conditional response model (2.2) which contains parameters $\boldsymbol{\vartheta}$, however, because we are interested in estimating $\boldsymbol{\beta}$, we specify $\boldsymbol{\beta}$ first, and solve for $\boldsymbol{\vartheta}$. The non-linear equations can be solved using a Newton method, for example by using the 'nleqslv' package in R (Hasselman, 2017). To do this, we must specify one of the $\boldsymbol{\vartheta}$ coefficients, therefore, we set

$$\vartheta_4 \;=\; \log(1.25) \;,$$

so that the interaction between X and Z_1 is non-zero in the response model. The parameters are chosen this way so that the causal effect of treatment is to decrease the probability of the response, when the subgroup variable is $S = 0$. When the subgroup variable is $S = 1$, the effect is lessened, and the causal effect is closer to the null value.

In the randomized setting,

$$P(X_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_1) \quad = \quad \text{expit}(\text{logit}(0.5) + \log(1.0)Z_{1i} + \log(1.0)Z_{2i}) \ ,$$

so that $P(X_i = 1) = 0.5$. In an observational setting,

$$P(W_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_1) \quad = \quad \text{expit}(\text{logit}(0.2) + \log(4.0)Z_{1i} + \log(4.0)Z_{2i}) \ ,$$

so that the treatment selection process for subject i is dependent upon auxiliary variables Z_{1i} and Z_{2i}, and $P(W_i = 1) = 0.5$. In the observational setting, the parameters are such that the probability of being treated is higher when $Z_1 = 1$ and/or $Z_2 = 1$.

The coefficients for generating the subgroup variable are

$$P(S_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_2) \quad = \quad \text{expit}(\text{logit}(0.5) + \log(0.9)Z_{1i} + \log(1.2)Z_{2i}) \ .$$

The parameters are set so that the probability of being in subgroup $S = 1$ is lower for those with $Z_1 = 1$ and higher for those with $Z_2 = 1$.

For the confounding and auxiliary variables, we set ζ_1 and ζ_2 so that $P(Z_1 = 1; \zeta_1) = 0.4$ and $P(Z_2 = 1; \zeta_2) = 0.6$.

To explore the effect of the percentage of missing subgroup data on bias and efficiency, we chose parameters $\boldsymbol{\rho}$ from (2.12) so that $P(R = 1) = 0.8$ and $P(R = 1) = 0.6$. In the first setting, when the percentage of missing subgroup data is 20%,

$$\begin{aligned} \text{ignorable:} \quad \pi_i(\boldsymbol{\rho}) \quad &= \quad \text{expit}(\text{logit}(0.76) + \log(1.0)Z_{1i} + \log(1.5)Z_{2i}) \\ \text{non-ignorable:} \quad \pi_i(\boldsymbol{\rho}) \quad &= \quad \text{expit}(\text{logit}(0.73) + \log(1.5)Z_{1i} + \log(1.5)Z_{2i}) \end{aligned}$$

In the second setting, when the percentage of missing subgroup data is 40%, $\boldsymbol{\rho}$ is set to

$$\begin{aligned} \text{ignorable:} \quad \pi_i(\boldsymbol{\rho}) \quad &= \quad \text{expit}(\text{logit}(0.55) + \log(1.0)Z_{1i} + \log(1.5)Z_{2i}) \\ \text{non-ignorable:} \quad \pi_i(\boldsymbol{\rho}) \quad &= \quad \text{expit}(\text{logit}(0.50) + \log(1.5)Z_{1i} + \log(1.5)Z_{2i}) \ . \end{aligned}$$

Empirical bias (EBias) is defined as the average difference between the estimated log odds ratio and the true log odds ratio. The average asymptotic standard error (ASE) is calculated as the average of the estimated standard errors using formula (2.24). The empirical standard error (ESE) is defined as the standard deviation of the m log odds ratio estimates, where m is the number of simulated datasets. The empirical coverage probability (ECP) is the empirical probability that the true log odds ratio is included in the 95% confidence interval; for each simulated dataset, the confidence interval is computed using the methods described in Section 2.3.2.

For each simulation study, the number of subjects per dataset is 2000 and the number of simulated datasets is 5000. The simulated datasets are independent, and the seed value for the first simulated dataset is the same across the simulation studies.

Table 2.1: Empirical bias and efficiency of estimated marginal regression coefficients, using the doubly inverse probability weighted estimating equation approach. In this table, parameters are chosen so that the overall probability of missing subgroup data is 20%.

Weighted for		β_1				$\beta_1 + \beta_3$			
X/W	R	EBias	ASE	ESE	ECP	EBias	ASE	ESE	ECP
Randomized setting, ignorable missing data									
No	No	0.0002	0.1448	0.1436	0.9530	0.0030	0.1398	0.1417	0.9496
No	Yes	0.0003	0.1450	0.1436	0.9526	0.0030	0.1400	0.1419	0.9498
Yes	No	0.0002	0.1451	0.1436	0.9534	0.0030	0.1401	0.1417	0.9506
Yes	Yes	0.0003	0.1453	0.1437	0.9548	0.0030	0.1404	0.1419	0.9500
Randomized setting, non-ignorable missing data									
No	No	0.0051	0.1448	0.1443	0.9512	0.0068	0.1400	0.1424	0.9460
No	Yes	0.0009	0.1451	0.1443	0.9508	0.0027	0.1403	0.1426	0.9464
Yes	No	0.0051	0.1451	0.1443	0.9514	0.0068	0.1403	0.1423	0.9468
Yes	Yes	0.0009	0.1455	0.1444	0.9500	0.0027	0.1406	0.1426	0.9480
Observational setting, ignorable missing data									
No	No	0.0327	0.1447	0.1443	0.9454	0.0324	0.1399	0.1404	0.9392
No	Yes	0.0331	0.1449	0.1445	0.9460	0.0329	0.1401	0.1405	0.9380
Yes	No	0.0005	0.1637	0.1627	0.9510	0.0018	0.1578	0.1568	0.9492
Yes	Yes	0.0005	0.1642	0.1630	0.9502	0.0020	0.1584	0.1574	0.9498
Observational setting, non-ignorable missing data									
No	No	0.0368	0.1447	0.1448	0.9412	0.0366	0.1401	0.1404	0.9378
No	Yes	0.0332	0.1450	0.1450	0.9438	0.0330	0.1404	0.1406	0.9404
Yes	No	0.0050	0.1638	0.1629	0.9508	0.0059	0.1581	0.1570	0.9522
Yes	Yes	0.0008	0.1644	0.1634	0.9510	0.0021	0.1587	0.1576	0.9516

Abbreviations: EBias Empirical bias, ASE asymptotic standard error,
ESE empirical standard error, ECP empirical coverage probability

Table 2.2: Empirical bias and efficiency of estimated marginal regression coefficients, using the doubly inverse probability weighted estimating equation approach. In this table, parameters are chosen so that the overall probability of missing subgroup data is 40%.

Weighted for		β_1				$\beta_1 + \beta_3$			
X/W	R	EBias	ASE	ESE	ECP	EBias	ASE	ESE	ECP
Randomized setting, ignorable missing data									
No	No	-0.0006	0.1662	0.1660	0.9528	0.0017	0.1603	0.1619	0.9486
No	Yes	-0.0004	0.1670	0.1665	0.9522	0.0017	0.1611	0.1626	0.9492
Yes	No	-0.0006	0.1666	0.1661	0.9522	0.0017	0.1607	0.1619	0.9494
Yes	Yes	-0.0004	0.1674	0.1665	0.9522	0.0017	0.1614	0.1625	0.9502
Randomized setting, non-ignorable missing data									
No	No	0.0077	0.1674	0.1677	0.9496	0.0096	0.1618	0.1614	0.9526
No	Yes	-0.0007	0.1688	0.1691	0.9506	0.0009	0.1631	0.1625	0.9544
Yes	No	0.0077	0.1678	0.1678	0.9506	0.0096	0.1622	0.1614	0.9542
Yes	Yes	-0.0006	0.1692	0.1691	0.9514	0.0009	0.1635	0.1625	0.9544
Observational setting, ignorable missing data									
No	No	0.0315	0.1661	0.1679	0.9452	0.0296	0.1605	0.1614	0.9438
No	Yes	0.0325	0.1669	0.1684	0.9442	0.0306	0.1612	0.1620	0.9446
Yes	No	-0.0009	0.1877	0.1878	0.9506	-0.0007	0.1807	0.1793	0.9524
Yes	Yes	-0.0007	0.1892	0.1890	0.9492	-0.0003	0.1823	0.1807	0.9518
Observational setting, non-ignorable missing data									
No	No	0.0384	0.1674	0.1680	0.9474	0.0399	0.1622	0.1605	0.9456
No	Yes	0.0311	0.1686	0.1690	0.9478	0.0326	0.1633	0.1612	0.9512
Yes	No	0.0067	0.1893	0.1882	0.9550	0.0082	0.1828	0.1800	0.9564
Yes	Yes	-0.0018	0.1913	0.1901	0.9550	0.0003	0.1845	0.1815	0.9570

Abbreviations: EBias Empirical bias, ASE asymptotic standard error,
ESE empirical standard error, ECP empirical coverage probability

2.4.2 Discussion of Simulation Results

When the missingness is ignorable (i.e. $\rho_1 = 0$), and treatment is randomized, any of the four weighted estimating function approaches (unweighted, weighted for missing data only, weighted for confounding only, and doubly weighted) yield consistent estimates for $\boldsymbol{\beta}$. See Tables 2.1 and 2.2, rows 1-4.

When the missingness is non-ignorable and treatment is randomized, only one weight for missingness is necessary, i.e. the use of estimating function $\overline{\mathbf{U}}_1(\boldsymbol{\beta}; \boldsymbol{\rho})$ from (2.15) will yield consistent estimates. In rows 5-8 of Tables 2.1 and 2.2, the parameters are set so that the missing data is non-ignorable (i.e. $\rho_1 \neq 0$) and treatment is randomized; we see that when we weight for missing data only (row 6), our estimates of β_1 and $\beta_1 + \beta_3$ are consistent, as expected. When we use a doubly weighted estimating function approach (row 8), our estimates are consistent as well, which shows that if a weight for confounding is included when it is not needed, consistent estimates are still obtained. However, when we do not weight for missing data (rows 5 and 7), we see that consistency is affected, although the degree of bias is relatively small. We also note that when a weight for missingness is omitted, the bias is larger when the percentage of missing data is higher, as expected (i.e. the EBias in Table 2.2 is larger than in Table 2.1).

In rows 9-12 of Tables 2.1 and 2.2, the parameters are set so that the missing data is ignorable (i.e. $\rho_1 = 0$) and treatment selection depends on a confounding variable that is also independently associated with the response (Z_1). In row 11, a single weight for confounding is used, and we see that our estimates of β_1 and $\beta_1 + \beta_3$ are consistent, as expected. In row 12, when a doubly weighted estimating equation approach is used, estimates are consistent, however, the estimates are slightly less efficient. In rows 9 and 10, a weight for confounding is not used, and we see that our estimates of β_1 and $\beta_1 + \beta_3$ are biased.

When the missing data is non-ignorable and treatment is not randomized, we have shown that the use of estimating function $\widetilde{\overline{\mathbf{U}}}_1(\boldsymbol{\beta}; \boldsymbol{\psi})$ from (2.18) will yield consistent estimates, and the simulation results reflect this. In rows 13-16 of Tables 2.1 and 2.2, the parameters are set so that the missing data is non-ignorable (i.e. $\rho_1 \neq 0$) and treatment selection is dependent upon confounding variable Z_{1i}. Row 16 shows the empirical bias,

average asymptotic standard error, average empirical standard error and the 95% confidence interval coverage probability for estimates of β_1 and $\beta_1 + \beta_3$ under a doubly weighted estimating equation approach. We see that the bias is small and the coverage probability is very close to 0.95, as expected. However in rows 13-15 when we fail to weight for either confounding or missing data or both, our estimates are not consistent. When a weight for confounding is used but the weight for missing data is omitted, the bias is relatively small, whereas when there is only one weight for missing data and the weight for confounding is omitted, the bias is much larger.

The limiting values of the marginal odds ratios of treatment for each of the subgroups is shown in Figures 2.5 and 2.6. These figures show that as the strength of the relationship between the response and confounding variable increases (i.e. the further $\exp(\vartheta_4)$ is away from 1), the degree of bias in estimates of β_1 and $\beta_1 + \beta_3$ increases. Also we see that when the weight for missing data is omitted but the weight for confounding is included, the bias is relatively small, which is also reflected in our simulation studies.

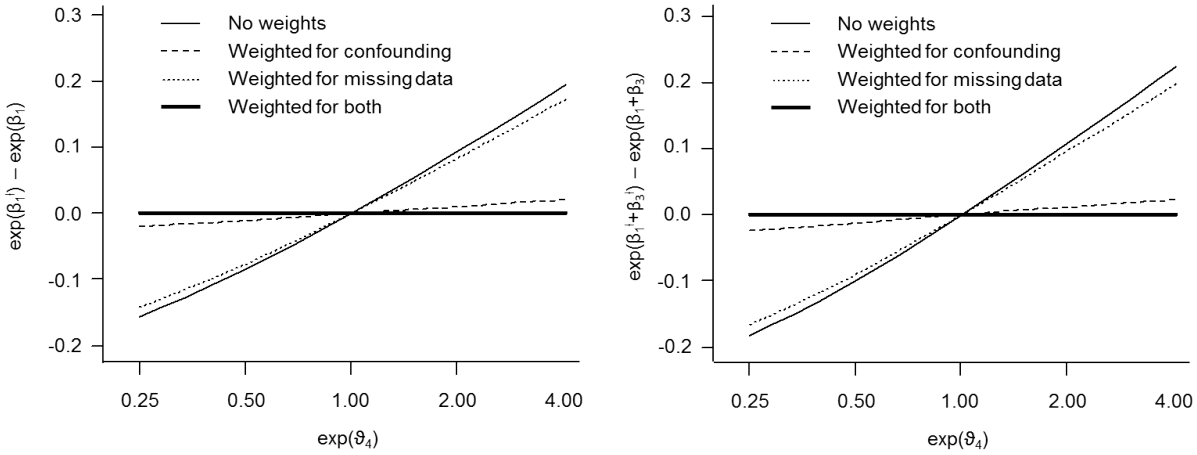

Figure 2.5: Asymptotic bias under model misspecification, in an observational setting where the missing data is non-ignorable. Bias is calculated as the difference between $\exp(\beta_1^\dagger)$ and $\exp(\beta_1)$ where $\exp(\beta_1^\dagger)$ is the estimate for $\exp(\beta_1)$. $\exp(\beta_1)$ is the odds ratio of the response, comparing exposed to unexposed subjects, for those with subgroup variable $S = 0$. $\exp(\beta_1 + \beta_3)$ is the odds ratio when $S = 1$. See Section 2.4.1 for parameter specifications for parameters other than ϑ_4. The percentage of subjects with a missing subgroup variable measurement is 20%.

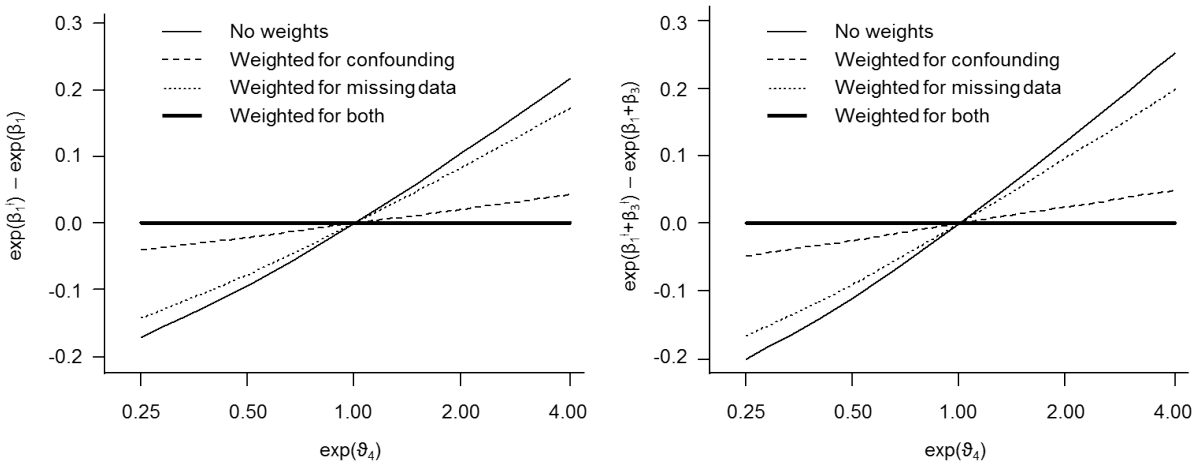

Figure 2.6: Asymptotic bias under model misspecification, in an observational setting where the missing data is non-ignorable. Bias is calculated as the difference between $\exp(\beta_1^\dagger)$ and $\exp(\beta_1)$ where $\exp(\beta_1^\dagger)$ is the estimate for $\exp(\beta_1)$. $\exp(\beta_1)$ is the odds ratio of the response, comparing exposed to unexposed subjects, for those with subgroup variable $S = 0$. $\exp(\beta_1 + \beta_3)$ is the odds ratio when $S = 1$. See Section 2.4.1 for parameter specifications for parameters other than ϑ_4. The percentage of subjects with a missing subgroup variable measurement is 40%.

2.5 Discussion

In this chapter, methods for estimating marginal causal effects in both randomized and observational settings with missing covariate data are given. In the observational setting, we describe the use of an inverse probability weight to account for confounding. In a randomized setting where the subgroup variable of interest is incomplete, an inverse probability weight to account for missing data is used. A doubly weighted estimating equation approach for estimating marginal regression coefficients is introduced for an observational setting where the subgroup variable is incomplete. This is an application of the 'full-weight' method introduced by Moodie et al. (2008) to a single treatment setting where the incomplete variable is an important subgroup variable. Consistency of the estimator from the doubly weighted estimating equation is shown, and a derivation of the asymptotic variance is given.

Limiting values under misspecified estimating functions, where the weight is misspecified, are calculated. Results of simulation studies are presented, and the bias introduced by misspecifying the weights in the weighted estimating functions is shown in figures for a variety of parameter settings.

We explore the impact of omitting one or both of the inverse probability weights in the estimating functions for estimating marginal regression parameters. There are other ways to misspecify the estimating functions, for example, omitting a variable from the propensity score model which may impact our ability to adjust for confounding. (See Appendix B for a simulation study where an important confounder is omitted from the propensity score model.) This is true for the weight for missing data as well. Variables can be included unnecessarily in the weight models, which may affect efficiency.

Data from subjects with incomplete subgroup data are omitted in the doubly inverse probability weighted method, which may impact the efficiency of the estimates. To address this, in the following chapter, we propose a weighted EM-type algorithm method for estimating marginal causal parameters using all available data, including data from subjects who are missing the subgroup variable.

Chapter 3

Causal Inference with Incomplete Data Via a Weighted EM Algorithm

In Chapter 2 we introduce a complete case doubly weighted estimating equation approach for estimating marginal causal odds ratios based on data from an observational study. In this chapter, we introduce an alternative approach based on an expectation-maximization (EM) algorithm which makes use of all available data (i.e. it is not restricted to individuals with complete data).

We first describe the EM approach for estimating conditional causal effects. In this framework, we then describe a weighted EM-type algorithm for estimating marginal treatment effects by subgroup in the randomized setting where the subgroup variable is incompletely observed. Finally, we describe a weighted EM-type algorithm for dealing with an incompletely observed subgroup variable in an observational setting.

3.1 Estimation of Conditional Causal Effects in a Randomized Setting

3.1.1 Notation and Models

Suppose we have a setting similar to the one in Section 2.2.1 where Y_i denotes a binary response variable, X_i a binary treatment variable in a randomized setting, and S_i a binary subgroup variable for subject i in a random sample of n individuals, $i = 1, ..., n$. Let $\mathbf{Z}_{1i} = (Z_{11i}, ..., Z_{1p_1i})^T$ denote a vector of auxiliary variables that are directly associated with the response, and $\mathbf{Z}_{2i} = (Z_{21i}, ..., Z_{2p_2i})^T$ denote a vector of auxiliary variables that are not independently associated with the response. For simplicity, we again consider the setting where $\mathbf{Z}_{1i}$ and $\mathbf{Z}_{2i}$ are scalars denoted by Z_{1i} and Z_{2i}. The subgroup variable S_i may be incomplete for some individuals, so we let $R_i = I(S_i \text{ is observed})$ as before.

We wish to consider a setting where only one auxiliary variable is predictive of the subgroup variable. We make the following conditional independence assumptions in a randomized controlled trial setting:

A.0	$R_i \perp (Y_i, X_i, S_i) \mid (Z_{1i}, Z_{2i})$
A.1	$Y_i \perp Z_{2i} \mid (X_i, S_i, Z_{1i})$
A.2	$X_i \perp (S_i, Z_{1i}, Z_{2i})$ (randomized setting)
A.3	$S_i \perp Z_{1i} \mid Z_{2i}$
A.4	$Z_{1i} \perp Z_{2i}$.

The assumptions here are the same as in Chapter 2, with the exception of the additional assumption that $S_i \perp Z_{1i}|Z_{2i}$ which is made for convenience throughout this chapter. The joint probability of $R_i, Y_i, X_i, S_i, Z_{1i}, Z_{2i}$ for subject i can be factorized as before (see (2.1)):

$$\begin{aligned} P(R_i, Y_i, X_i, S_i, Z_{1i}, Z_{2i}) &= P(R_i|Y_i, X_i, S_i, Z_{1i}, Z_{2i})P(Y_i|X_i, S_i, Z_{1i}, Z_{2i}) \\ &\quad \cdot P(X_i|S_i, Z_{1i}, Z_{2i})P(S_i|Z_{1i}, Z_{2i})P(Z_{1i}, Z_{2i}) \\ &= P(R_i|Z_{1i}, Z_{2i})P(Y_i|X_i, S_i, Z_{1i})P(X_i)P(S_i|Z_{2i})P(Z_{1i})P(Z_{2i}) \ . \end{aligned} \tag{3.1}$$

Conditional on X_i, S_i and Z_{1i}, the response model parameters and the missing data mechanism parameters are distinct, therefore the missing data mechanism is ignorable.

Figure 3.1 contains a causal DAG that summarizes the relationships between the variables.

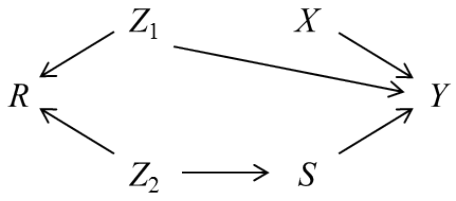

Figure 3.1: Simple causal DAG for treatment X, response Y, a variable that is independently associated with response Z_1, and auxiliary variable Z_2, in the context of a randomized controlled trial. Variable R indicates whether the subgroup variable is observed.

The conditional expectation of the binary response can be written as

$$\mu_i(\boldsymbol{\vartheta}) = E(Y_i|X_i, S_i, Z_{1i}) = P(Y_i = 1|X_i, S_i, Z_{1i}; \boldsymbol{\vartheta}) \ .$$

We again consider a logistic regression model whereby

$$\mu_i(\boldsymbol{\vartheta}) \ = \ \text{expit}(\vartheta_0 + \vartheta_1 X_i + \vartheta_2 S_i + \vartheta_3 X_i S_i + \vartheta_4 X_i Z_{1i}) \ , \tag{3.2}$$

and $\boldsymbol{\vartheta} = (\vartheta_0, \vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4)^T$ is a vector of regression coefficients. Although we do not include a main effect for the auxiliary variable Z_1 to simplify the model, the methods described in this chapter are applicable in the setting where the coefficient for the main effect of Z_1 is non-zero. We consider this as the 'true' data generating model for the response and deal with methods for fitting it here; we deal with the observational setting subsequently.

In the randomized setting, we let $E(X_i) = P(X_i = 1; \xi_1) = 0.5$, so that the percentage of treated subjects is 50%. For the auxiliary variables, $E(Z_{ji}) = P(Z_{ji} = 1; \zeta_j) = \text{expit}(\zeta_j)$, for $j = 1, 2$.

The subgroup variable is modeled as follows:

$$\pi_i(\boldsymbol{\xi}_2) = E(S_i|Z_{1i}, Z_{2i}) = P(S_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_2) = \text{expit}(\xi_{20} + \xi_{21} Z_{1i} + \xi_{22} Z_{2i}) \ , \tag{3.3}$$

where $\boldsymbol{\xi}_2 = (\xi_{20}, \xi_{21}, \xi_{22})^T$ is a vector of regression coefficients. Because we have assumed that $S_i \perp Z_{1i}|Z_{2i}$, we set $\xi_{21} = 0$ and so write this as $P(S_i = 1|Z_{2i}; \boldsymbol{\xi}_2)$ in what follows.

The logistic regression model for generating R_i is

$$\pi_i(\boldsymbol{\rho}) \;=\; E(R_i|Z_{1i}, Z_{2i}) \;=\; P(R_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\rho}) = \text{expit}(\rho_0 + \rho_1 Z_{1i} + \rho_2 Z_{2i}) \;, \qquad (3.4)$$

where $\boldsymbol{\rho} = (\rho_0, \rho_1, \rho_2)^T$ is a vector of regression parameters.

In the following section, we describe an EM algorithm for estimating $\boldsymbol{\vartheta}$.

3.1.2 An EM Algorithm for Incomplete Covariates

If interest lies in estimating $\boldsymbol{\vartheta}$ from (3.2), an EM algorithm method for dealing with missing covariate data can be used (Ibrahim et al., 2005).

Let $\boldsymbol{\delta} = (\boldsymbol{\vartheta}^T, \boldsymbol{\xi}_2^T)^T$, and let

$$\mathcal{D}_{1i} = \{Y_i, X_i, S_i, Z_{1i}, Z_{2i} \text{ for } R_i = 1; \text{ or } Y_i, X_i, Z_{1i}, Z_{2i} \text{ for } R_i = 0\}$$

denote the observed data for individual i, with $\mathcal{D}_1 = \{\mathcal{D}_{1i},\ i = 1, ..., n\}$. In the E-step, the conditional expectation of the loglikelihood for $\boldsymbol{\delta}$ with respect to $\mathcal{D}_1$ at the kth iteration is written as:

$$\begin{aligned} Q(\boldsymbol{\delta}; \hat{\boldsymbol{\delta}}^{(k)}) &= Q_1(\boldsymbol{\vartheta}; \hat{\boldsymbol{\delta}}^{(k)}) + Q_2(\boldsymbol{\xi}_2; \hat{\boldsymbol{\delta}}^{(k)}) \\ &= \sum_{i=1}^{n} \left(Q_{1i}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\delta}}^{(k)}) + Q_{2i}(\boldsymbol{\xi}_2; \hat{\boldsymbol{\delta}}^{(k)}) \right) \end{aligned} \qquad (3.5)$$

where

$$\begin{aligned} Q_{1i}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\delta}}^{(k)}) &= R_i \log P(Y_i|X_i, S_i, Z_{1i}; \boldsymbol{\vartheta}) \\ &\quad + (1 - R_i) \sum_{j=0}^{1} \log P(Y_i|X_i, S_i = j, Z_{1i}; \boldsymbol{\vartheta}) P(S_i = j|\mathcal{D}_{1i}; \hat{\boldsymbol{\delta}}^{(k)}) \end{aligned} \qquad (3.6)$$

and

$$\begin{aligned} Q_{2i}(\boldsymbol{\xi}_2; \hat{\boldsymbol{\delta}}^{(k)}) &= R_i \log P(S_i|Z_{2i}; \boldsymbol{\xi}_2) \\ &\quad + (1 - R_i) \sum_{j=0}^{1} \log P(S_i = j|Z_{2i}; \boldsymbol{\xi}_2) P(S_i = j|\mathcal{D}_{1i}; \hat{\boldsymbol{\delta}}^{(k)}) \;. \end{aligned} \qquad (3.7)$$

For subjects with $R_i = 0$,

$$P(S_i = j|\mathcal{D}_{1i}; \hat{\boldsymbol{\delta}}^{(k)}) = \frac{P(Y_i|X_i, S_i = j, Z_{1i}; \hat{\boldsymbol{\vartheta}}^{(k)})P(S_i = j|Z_{2i}; \hat{\boldsymbol{\xi}}_2^{(k)})}{\sum_{s=0}^{1} P(Y_i|X_i, S_i = s, Z_{1i}; \hat{\boldsymbol{\vartheta}}^{(k)})P(S_i = s|Z_{2i}; \hat{\boldsymbol{\xi}}_2^{(k)})} . \quad (3.8)$$

In the M-step, we maximize $Q_1(\boldsymbol{\vartheta}; \hat{\boldsymbol{\delta}}^{(k)})$ and $Q_2(\boldsymbol{\xi}_2; \hat{\boldsymbol{\delta}}^{(k)})$ to obtain $\hat{\boldsymbol{\delta}}^{(k+1)}$. To do this, a weighted GLM can be fitted using a 'long' dataset. The long dataset is constructed as follows. Each subject has two rows, one for $j = 1$ and one for $j = 0$; when S_i is observed, the weight is 1 when $S_i = j$ and 0 otherwise. When S_i is missing, the weight is computed using equation (3.8). See Table 3.1 for an example of how to set up such a dataset.

Table 3.1: Layout of a dataset for the EM algorithm, at the $(k+1)$st M-step.

i	Z_{1i}	Z_{2i}	X_i	S_i	Y_i	R_i	j	weight
1	1	0	0	1	1	1	1	1
1	1	0	0	1	1	1	0	0
2	1	1	0	.	0	0	1	$P(S_i = 1\|Z_{1i} = 1, Z_{2i} = 1, X_i = 0, Y_i = 0; \hat{\boldsymbol{\delta}}^{(k)})$
2	1	1	0	.	0	0	0	$P(S_i = 0\|Z_{1i} = 1, Z_{2i} = 1, X_i = 0, Y_i = 0; \hat{\boldsymbol{\delta}}^{(k)})$
3	0	1	1	0	0	1	1	0
3	0	1	1	0	0	1	0	1
4	1	0	1	.	1	0	1	$P(S_i = 1\|Z_{1i} = 1, Z_{2i} = 0, X_i = 1, Y_i = 1; \hat{\boldsymbol{\delta}}^{(k)})$
4	1	0	1	.	1	0	0	$P(S_i = 0\|Z_{1i} = 1, Z_{2i} = 0, X_i = 1, Y_i = 1; \hat{\boldsymbol{\delta}}^{(k)})$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The estimating function for obtaining $\hat{\boldsymbol{\vartheta}}^{(k+1)}$ at the $(k+1)$st M-step of the EM algorithm is

$$\mathbf{U}_1(\boldsymbol{\vartheta}; \hat{\boldsymbol{\delta}}^{(k)}) = \sum_{i=1}^{n} \mathbf{U}_{1i}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\delta}}^{(k)}) \quad (3.9)$$

where

$$\begin{aligned}\mathbf{U}_{1i}(\boldsymbol{\vartheta};\hat{\boldsymbol{\delta}}^{(k)}) &= R_i\,\mathbf{D}_i(\boldsymbol{\vartheta})[V_i(\boldsymbol{\vartheta})]^{-1}\big(Y_i-\mu_i(\boldsymbol{\vartheta})\big)\\ &\quad+(1-R_i)\sum_{j=0}^{1}\mathbf{D}_i(\boldsymbol{\vartheta})[V_i(\boldsymbol{\vartheta})]^{-1}\Big\{\big(Y_i-\mu_i(\boldsymbol{\vartheta})\big)P(S_i=j|\mathcal{D}_{1i};\hat{\boldsymbol{\delta}}^{(k)})\Big\}\,,\end{aligned}$$

with $\mathbf{D}_i(\boldsymbol{\vartheta})=\partial\mu_i(\boldsymbol{\vartheta})/\partial\boldsymbol{\vartheta}$ and $V_i(\boldsymbol{\vartheta})=\mu_i(\boldsymbol{\vartheta})(1-\mu_i(\boldsymbol{\vartheta}))$. Similarly, the estimating function for obtaining $\hat{\boldsymbol{\xi}}_2^{(k+1)}$ at the $(k+1)$st M-step of the EM algorithm is

$$\mathbf{U}_2(\boldsymbol{\xi}_2;\hat{\boldsymbol{\delta}}^{(k)}) = \sum_{i=1}^{n}\mathbf{U}_{2i}(\boldsymbol{\xi}_2;\hat{\boldsymbol{\delta}}^{(k)}) \tag{3.10}$$

where

$$\begin{aligned}\mathbf{U}_{2i}(\boldsymbol{\xi}_2;\hat{\boldsymbol{\delta}}^{(k)}) &= \mathbf{D}_i(\boldsymbol{\xi}_2)[V_i(\boldsymbol{\xi}_2)]^{-1}\Big[R_i\,\big(S_i-\pi_i(\boldsymbol{\xi}_2)\big)\\ &\quad+(1-R_i)\sum_{j=0}^{1}\Big\{\big(j-\pi_i(\boldsymbol{\xi}_2)\big)P(S_i=j|\mathcal{D}_{1i};\hat{\boldsymbol{\delta}}^{(k)})\Big\}\Big]\,,\end{aligned}$$

with $\mathbf{D}_i(\boldsymbol{\xi}_2)=\partial\pi_i(\boldsymbol{\xi}_2)/\partial\boldsymbol{\xi}_2$ and $V_i(\boldsymbol{\xi}_2)=\mathrm{var}(S_i|Z_{2i})=\pi_i(\boldsymbol{\xi}_2)(1-\pi_i(\boldsymbol{\xi}_2))$. The expectation and maximization steps are repeated iteratively until $|\hat{\boldsymbol{\delta}}^{(k)}-\hat{\boldsymbol{\delta}}^{(k+1)}|<\epsilon$ for a fixed tolerance ϵ. In other words, we (i) obtain initial estimates for $\boldsymbol{\delta}=(\boldsymbol{\vartheta}^T,\boldsymbol{\xi}_2^T)^T$ denoted by $\hat{\boldsymbol{\delta}}^{(1)}$; (ii) calculate the weights for each individual; (iii) fit each weighted GLM to obtain $\hat{\boldsymbol{\delta}}^{(2)}$. This is repeated until the convergence criteria are met. One way to obtain initial estimates for $\boldsymbol{\vartheta}$ and $\boldsymbol{\xi}_2$ is to fit two separate unweighted GLMs using complete cases only. The EM algorithm described here is simply a computational device for obtaining the maximum likelihood estimate of $\boldsymbol{\delta}$.

If $\hat{\boldsymbol{\delta}}$ denotes the estimate of $\boldsymbol{\delta}$ at convergence, the estimate of the asymptotic covariance matrix of $\boldsymbol{\vartheta}$ is the upper block of matrix $[I(\hat{\boldsymbol{\delta}})]^{-1}$ where $I(\hat{\boldsymbol{\delta}})$ is the observed information

given by

$$
\begin{aligned}
I(\hat{\boldsymbol{\delta}}) &= -\ddot{Q}(\hat{\boldsymbol{\delta}};\hat{\boldsymbol{\delta}}) \\
&\quad -\Bigg\{\sum_{i=1}^{n}\Bigg(R_i\, \mathbb{S}_i(\mathcal{D}_{1i};\hat{\boldsymbol{\delta}})\, \mathbb{S}_i^T(\mathcal{D}_{1i};\hat{\boldsymbol{\delta}}) \\
&\quad +(1-R_i)\sum_{j=0}^{1}\mathbb{S}_i(\mathcal{D}_{1i},S_i=j;\hat{\boldsymbol{\delta}})\, \mathbb{S}_i^T(\mathcal{D}_{1i},S_i=j;\hat{\boldsymbol{\delta}})P(S_i=j|\mathcal{D}_{1i};\hat{\boldsymbol{\delta}})\Bigg) \\
&\quad -\sum_{i=1}^{n}\dot{Q}_i(\hat{\boldsymbol{\delta}};\hat{\boldsymbol{\delta}})\, \dot{Q}_i^T(\hat{\boldsymbol{\delta}};\hat{\boldsymbol{\delta}})\Bigg\}
\end{aligned}
\tag{3.11}
$$

(Louis, 1982). The contribution to (3.11) when $R_i = 1$ involves

$$
\mathbb{S}_i(\mathcal{D}_{1i};\boldsymbol{\delta}) = \begin{pmatrix} \mathbb{S}_{1i}(Y_i,X_i,S_i,Z_{1i};\boldsymbol{\vartheta}) \\ \mathbb{S}_{2i}(S_i,Z_{2i};\boldsymbol{\xi}_2) \end{pmatrix}
$$

where

$$
\mathbb{S}_{1i}(Y_i,X_i,S_i,Z_{1i};\boldsymbol{\vartheta}) = \begin{pmatrix} 1 \\ X_i \\ S_i \\ X_iS_i \\ X_iZ_{1i} \end{pmatrix}(Y_i-\mu_i(\boldsymbol{\vartheta}))
$$

and

$$
\mathbb{S}_{2i}(S_i,Z_{2i};\boldsymbol{\xi}_2) = \begin{pmatrix} 1 \\ Z_{2i} \end{pmatrix}(S_i-\pi_i(\boldsymbol{\xi}_2))\ .
$$

Likewise when $R_i = 0$, the contributions involve the terms

$$
\mathbb{S}_i(\mathcal{D}_{1i},S_i=j;\boldsymbol{\delta}) = \begin{pmatrix} \mathbb{S}_{1i}(Y_i,X_i,S_i=j,Z_{1i};\boldsymbol{\vartheta}) \\ \mathbb{S}_{2i}(S_i=j,Z_{2i};\boldsymbol{\xi}_2) \end{pmatrix}\ .
$$

Finally we define $\dot{Q}_i(\boldsymbol{\delta};\boldsymbol{\delta})$ and $\ddot{Q}(\boldsymbol{\delta};\boldsymbol{\delta})$ as

$$
\dot{Q}_i(\boldsymbol{\delta};\boldsymbol{\delta}) = R_i\, \mathbb{S}_i(\mathcal{D}_{1i};\boldsymbol{\delta}) + (1-R_i)\sum_{j=0}^{1}\mathbb{S}_i(\mathcal{D}_{1i},S_i=j;\boldsymbol{\delta})P(S_i=j|\mathcal{D}_{1i};\boldsymbol{\delta})
$$

and

$$\ddot{Q}(\boldsymbol{\delta};\boldsymbol{\delta}) \;=\; \sum_{i=1}^{n}\left(R_i \,\frac{\partial \mathbb{S}_i(\mathcal{D}_{1i};\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^T} + (1-R_i)\sum_{j=0}^{1} \frac{\partial \mathbb{S}_i(\mathcal{D}_{1i}, S_i=j;\boldsymbol{\delta})}{\partial \boldsymbol{\delta}^T} P(S_i=j|\mathcal{D}_{1i};\boldsymbol{\delta})\right) .$$

In this section we have described an EM algorithm for estimating regression coefficients in the randomized setting with an incomplete subgroup variable. Next, we review the notation and models for an observational setting, and show that the EM algorithm for estimating regression coefficients in the correctly specified conditional model is the same as in the randomized setting.

3.2 Estimation of Conditional Causal Effects in an Observational Setting

Here, we briefly show that the EM algorithm described in Section 3.1 also works in an observational setting where confounding is an issue provided we control for the confounding variable in the response model by conditioning on Z_{1i}. We first review the notation and models in an observational setting which are presented in Sections 2.1.2 and 2.2.2.

3.2.1 Notation and Models

Suppose we make the same conditional independence assumptions as in the randomized setting, with the exception of the assumptions for the treatment variable:

B.0	$R_i \perp (Y_i, W_i, S_i) \mid (Z_{1i}, Z_{2i})$
B.1	$Y_i \perp Z_{2i} \mid (W_i, S_i, Z_{1i})$
B.2	$W_i \perp S_i \vert (Z_{1i}, Z_{2i})$ (observational setting)
B.3	$S_i \perp Z_{1i} \vert Z_{2i}$
B.4	$Z_{1i} \perp Z_{2i}$.

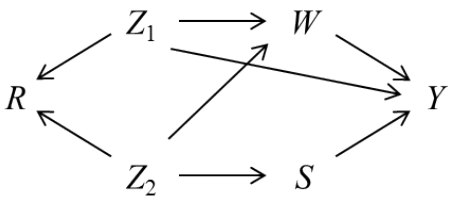

Figure 3.2: Simple causal DAG for treatment W, response Y, a variable that is independently associated with response Z_1 and treatment selection, and auxiliary variable Z_2 that is associated with treatment selection, in the context of an observational study. Variable R indicates whether subgroup variable S is observed.

The joint probability of $R_i, Y_i, W_i, S_i, Z_{1i}, Z_{2i}$ for subject i can be factorized as

$$\begin{aligned} P(R_i, Y_i, W_i, S_i, Z_{1i}, Z_{2i}) &= P(R_i|Y_i, W_i, S_i, Z_{1i}, Z_{2i})P(Y_i|W_i, S_i, Z_{1i}, Z_{2i}) \\ &\quad \cdot P(W_i|S_i, Z_{1i}, Z_{2i})P(S_i|Z_{1i}, Z_{2i})P(Z_{1i}, Z_{2i}) \\ &= P(R_i|Z_{1i}, Z_{2i})P(Y_i|W_i, S_i, Z_{1i}) \\ &\quad \cdot P(W_i|Z_{1i}, Z_{2i})P(S_i|Z_{2i})P(Z_{1i})P(Z_{2i}) \ . \end{aligned} \tag{3.12}$$

See Figure 3.2 for a causal DAG that summarizes the relationships between the variables.

The conditional expectation of the response can be written as $\mu_i(\boldsymbol{\vartheta}) = E(Y_i|W_i, S_i, Z_{1i})$. We again consider a logistic regression model for the response whereby

$$\mu_i(\boldsymbol{\vartheta}) = \text{expit}(\vartheta_0 + \vartheta_1 W_i + \vartheta_2 S_i + \vartheta_3 W_i S_i + \vartheta_4 W_i Z_{1i}) \ , \tag{3.13}$$

and $\boldsymbol{\vartheta} = (\vartheta_0, \vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4)^T$ is a vector of regression coefficients. Note that $\boldsymbol{\vartheta}$ is the same vector of regression coefficients as in equation (3.2) regardless of whether we are in a setting where treatment is randomized, or an observational setting.

When treatment is not randomized, the conditional probability of treatment given (Z_{1i}, Z_{2i}) for subject i is assumed to have the form

$$\pi_i(\boldsymbol{\xi}_1) = P(W_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_1) = \text{expit}(\xi_{10} + \xi_{11} Z_{1i} + \xi_{12} Z_{2i}) \ , \tag{3.14}$$

where $\boldsymbol{\xi}_1 = (\xi_{10}, \xi_{11}, \xi_{12})^T$ is a vector of regression coefficients. The remaining variables, R_i, S_i, Z_{1i} and Z_{2i}, are generated as in Section 3.1.1.

3.2.2 The EM Algorithm with Observational Data

Here we show that EM algorithm described in Section 3.1.2 can be used in an observational setting.

Let

$$\mathcal{D}_{2i} = \{Y_i, W_i, S_i, Z_{1i}, Z_{2i} \text{ for } R_i = 1; \text{ or } Y_i, W_i, Z_{1i}, Z_{2i} \text{ for } R_i = 0\}$$

denote the observed data for individual i, with $\mathcal{D}_2 = \{\mathcal{D}_{2i},\ i = 1, ..., n\}$. At the E-step, the conditional expectation of the loglikelihood for $\boldsymbol{\delta}$ with respect to $\mathcal{D}_2$ at the kth iteration is written as:

$$\begin{aligned} Q(\boldsymbol{\delta}; \hat{\boldsymbol{\delta}}^{(k)}) &= Q_1(\boldsymbol{\vartheta}; \hat{\boldsymbol{\delta}}^{(k)}) + Q_2(\boldsymbol{\xi}_2; \hat{\boldsymbol{\delta}}^{(k)}) \\ &= \sum_{i=1}^{n} \left(Q_{1i}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\delta}}^{(k)}) + Q_{2i}(\boldsymbol{\xi}_2; \hat{\boldsymbol{\delta}}^{(k)}) \right) \end{aligned}$$

where

$$\begin{aligned} Q_{1i}(\boldsymbol{\vartheta}; \hat{\boldsymbol{\delta}}^{(k)}) &= R_i \log P(Y_i | W_i, S_i, Z_{1i}; \boldsymbol{\vartheta}) \\ &\quad + (1 - R_i) \sum_{j=0}^{1} \log P(Y_i | W_i, S_i = j, Z_{1i}; \boldsymbol{\vartheta}) P(S_i = j | \mathcal{D}_{2i}; \hat{\boldsymbol{\delta}}^{(k)}) \end{aligned}$$

and

$$\begin{aligned} Q_{2i}(\boldsymbol{\xi}_2; \hat{\boldsymbol{\delta}}^{(k)}) &= R_i \log P(S_i | Z_{2i}; \boldsymbol{\xi}_2) \\ &\quad + (1 - R_i) \sum_{j=0}^{1} \log P(S_i = j | Z_{2i}; \boldsymbol{\xi}_2) P(S_i = j | \mathcal{D}_{2i}; \hat{\boldsymbol{\delta}}^{(k)}) \ . \end{aligned}$$

For subjects with $R_i = 0$,

$$P(S_i = j | \mathcal{D}_{2i}; \hat{\boldsymbol{\delta}}^{(k)}) = \frac{P(Y_i | W_i, S_i = j, Z_{1i}; \hat{\boldsymbol{\vartheta}}^{(k)}) P(S_i = j | Z_{2i}; \hat{\boldsymbol{\xi}}_2^{(k)})}{\sum_{s=0}^{1} P(Y_i | W_i, S_i = s, Z_{1i}; \hat{\boldsymbol{\vartheta}}^{(k)}) P(S_i = s | Z_{2i}; \hat{\boldsymbol{\xi}}_2^{(k)})} \ .$$

Because we condition on Z_{1i} in the conditional response model, confounding is not an issue, and we proceed with the methods given in Section 3.1.2.

We have shown how to use an EM algorithm to estimate conditional causal odds ratios in both randomized and observational settings with a missing subgroup variable.

3.3 Estimation of Marginal Causal Effects in a Randomized Setting

In this section, we extend the EM algorithm for incomplete covariate data in a conditional response model for estimation of marginal causal effects.

3.3.1 Notation and Models

See Section 3.1.1 for a list of assumptions and models for random variables R, Y, X, S, Z_1 and Z_2 in a randomized setting. A key assumption used in Chapter 2, which is used here as well, is:

$$P(Y_i = 1|W_i = w, S_i = s, Z_{1i} = z_1; \boldsymbol{\vartheta}) \;=\; P(Y_i = 1|X_i = w, S_i = s, Z_{1i} = z_1; \boldsymbol{\vartheta}) \; . \quad (3.15)$$

This means that the full conditional distribution of the response given the treatment variable and S_i, Z_{1i} is the same whether we are in a randomized or an observational setting. This holds because we are conditioning on all of the variables that are (independently) associated with response.

In the marginal model setting, we wish to estimate $\boldsymbol{\beta}$ defined as the logistic regression coefficients in the following response model:

$$P(Y_i = 1|X_i, S_i; \boldsymbol{\beta}) = \mu_i(\boldsymbol{\beta}) = \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 S_i + \beta_3 X_i S_i) \; , \quad (3.16)$$

where

$$\begin{aligned} P(Y_i = 1|X_i, S_i; \boldsymbol{\beta}) &= E_{Z_{1i}|X_i,S_i}[P(Y_i = 1|X_i, S_i, Z_{1i}; \boldsymbol{\vartheta})] \\ &= \sum_{z_1=0}^{1} P(Y_i = 1|X_i, S_i, z_1; \boldsymbol{\vartheta}) P(Z_{1i} = z_1; \zeta_1) \end{aligned}$$

by assumptions A.2, A.3 and A.4.

3.3.2 A Weighted EM-Type Algorithm for Incomplete Data

We begin by writing the joint probability in the context of estimating $\boldsymbol{\beta}$. Because we are not interested in estimating the effect of Z_1 on Y, we do not include Z_1 in the model of interest (3.16), but we note that Z_1 is an important variable which will be used in the discussion of bias and limiting values. We include Z_2 since it contains information about the missing subgroup variable and is conditionally independent of Y. The joint distribution of the variables can be factored as follows:

$$\begin{aligned} P(R_i, Y_i, X_i, S_i, Z_{2i}) &= P(R_i|Y_i, X_i, S_i, Z_{2i})P(Y_i|X_i, S_i, Z_{2i}) \\ &\quad \cdot\ P(X_i|S_i, Z_{2i})P(S_i|Z_{2i})P(Z_{2i}) \\ &= P(R_i|Y_i, X_i, S_i, Z_{2i})P(Y_i|X_i, S_i)P(X_i)P(S_i|Z_{2i})P(Z_{2i})\ . \end{aligned}$$

We cannot factor Y_i from the missing data model when we do not include Z_{1i} in the joint probability. Omitting Z_{1i} from the model for Y creates non-ignorable missingness. A selection model can be incorporated into the likelihood to account for the non-ignorably missing data (Ibrahim et al., 2005), however, instead we propose the use of an inverse probability weight when using the EM algorithm to estimate parameters of the marginal model. A weighted EM-type estimating function is described in what follows using weights based on the inverse of $P(R_i = r|Z_{1i}, Z_{2i}; \boldsymbol{\rho})$.

Let

$$\mathcal{D}_{3i} = \{Y_i, X_i, S_i, Z_{2i} \text{ for } R_i = 1; \text{ or } Y_i, X_i, Z_{2i} \text{ for } R_i = 0\}\ ,$$

$\mathcal{D}_3 = \{\mathcal{D}_{3i},\ i = 1, ..., n\}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\xi}_2^T)^T$. At the E-step, the conditional expectation of a working loglikelihood for $\boldsymbol{\theta}$ with respect to $\mathcal{D}_3$ at the kth iteration is written as:

$$\begin{aligned} \overline{Q}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\rho}) &= \overline{Q}_1(\boldsymbol{\beta}; \hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\rho}) + \overline{Q}_2(\boldsymbol{\xi}_2; \hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\rho}) \\ &= \sum_{i=1}^{n} \left(\overline{Q}_{1i}(\boldsymbol{\beta}; \hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\rho}) + \overline{Q}_{2i}(\boldsymbol{\xi}_2; \hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\rho}) \right) \end{aligned} \tag{3.17}$$

where

$$\overline{Q}_{1i}(\boldsymbol{\beta};\hat{\boldsymbol{\theta}}^{(k)},\boldsymbol{\rho}) = \frac{R_i}{\pi_i(\boldsymbol{\rho})}\log P(Y_i|X_i,S_i;\boldsymbol{\beta}) + \frac{(1-R_i)}{(1-\pi_i(\boldsymbol{\rho}))}\sum_{j=0}^{1}\log P(Y_i|X_i,S_i=j;\boldsymbol{\beta})P(S_i=j|\mathcal{D}_{3i};\hat{\boldsymbol{\theta}}^{(k)})$$

and

$$\overline{Q}_{2i}(\boldsymbol{\xi}_2;\hat{\boldsymbol{\theta}}^{(k)},\boldsymbol{\rho}) = \frac{R_i}{\pi_i(\boldsymbol{\rho})}\log P(S_i|Z_{2i};\boldsymbol{\xi}_2) + \frac{(1-R_i)}{(1-\pi_i(\boldsymbol{\rho}))}\sum_{j=0}^{1}\log P(S_i=j|Z_{2i};\boldsymbol{\xi}_2)P(S_i=j|\mathcal{D}_{3i};\hat{\boldsymbol{\theta}}^{(k)}) \ .$$

For subjects with $R_i = 0$,

$$P(S_i=j|\mathcal{D}_{3i};\hat{\boldsymbol{\theta}}^{(k)}) = \frac{P(Y_i|X_i,S_i=j;\hat{\boldsymbol{\beta}}^{(k)})P(S_i=j|Z_{2i};\hat{\boldsymbol{\xi}}_2^{(k)})}{\sum_{s=0}^{1}P(Y_i|X_i,S_i=s;\hat{\boldsymbol{\beta}}^{(k)})P(S_i=s|Z_{2i};\hat{\boldsymbol{\xi}}_2^{(k)})} \ .$$

In the M-step, we fit two weighted GLMs to maximize $\overline{Q}_1(\boldsymbol{\beta};\hat{\boldsymbol{\theta}}^{(k)},\boldsymbol{\rho})$ and $\overline{Q}_2(\boldsymbol{\xi}_2;\hat{\boldsymbol{\theta}}^{(k)},\boldsymbol{\rho})$ to obtain $\hat{\boldsymbol{\theta}}^{(k+1)}$. We do this in the same way as described for estimating $\boldsymbol{\vartheta}$ in Section 3.1, however, the first step is to estimate $\boldsymbol{\rho}$ using the full dataset. A logistic regression model can be fitted to obtain $\hat{\boldsymbol{\rho}}$. Then, a "long" dataset is constructed as before with two rows for each subject, setting $j = 1$ in the first row and $j = 0$ in the second row. See Table 3.2 for an example of a dataset structure in this context. The weight for subject i with $R_i = 1$ is $1/\pi_i(\hat{\boldsymbol{\rho}})$ when $S_i = j$ and 0 otherwise. The weight for subject i with $R_i = 0$ is $P(S_i = j|\mathcal{D}_{3i};\hat{\boldsymbol{\theta}}^{(k)})/(1-\pi_i(\hat{\boldsymbol{\rho}}))$.

The weighted estimating function at the M-step for obtaining $\hat{\boldsymbol{\beta}}^{(k+1)}$ is

$$\overline{\mathbf{U}}_1(\boldsymbol{\beta};\hat{\boldsymbol{\theta}}^{(k)},\boldsymbol{\rho}) = \sum_{i=1}^{n}\overline{\mathbf{U}}_{1i}(\boldsymbol{\beta};\hat{\boldsymbol{\theta}}^{(k)},\boldsymbol{\rho})$$

where

$$\overline{\mathbf{U}}_{1i}(\boldsymbol{\beta};\hat{\boldsymbol{\theta}}^{(k)},\boldsymbol{\rho}) = \frac{R_i}{\pi_i(\boldsymbol{\rho})}\mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\big(Y_i-\mu_i(\boldsymbol{\beta})\big) + \frac{(1-R_i)}{1-\pi_i(\boldsymbol{\rho})}\sum_{j=0}^{1}\mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\Big\{\big(Y_i-\mu_i(\boldsymbol{\beta})\big)P(S_i=j|\mathcal{D}_{3i};\hat{\boldsymbol{\theta}}^{(k)})\Big\} \ ,$$

Table 3.2: Layout of a dataset for the weighted EM-type algorithm, at the $(k+1)$st M-step.

i	Z_{1i}	Z_{2i}	X_i	S_i	Y_i	R_i	j	weight
1	1	0	0	1	1	1	1	$1/\pi_i(\hat{\boldsymbol{\rho}})$
1	1	0	0	1	1	1	0	0
2	1	1	0	.	0	0	1	$\frac{P(S_i=1\|Z_{2i}=1,X_i=0,Y_i=0;\hat{\boldsymbol{\theta}}^{(k)})}{1-\pi_i(\hat{\boldsymbol{\rho}})}$
2	1	1	0	.	0	0	0	$\frac{P(S_i=0\|Z_{2i}=1,X_i=0,Y_i=0;\hat{\boldsymbol{\theta}}^{(k)})}{1-\pi_i(\hat{\boldsymbol{\rho}})}$
3	0	1	1	0	0	1	1	0
3	0	1	1	0	0	1	0	$1/\pi_i(\hat{\boldsymbol{\rho}})$
4	1	0	1	.	1	0	1	$\frac{P(S_i=1\|Z_{2i}=0,X_i=1,Y_i=1;\hat{\boldsymbol{\theta}}^{(k)})}{1-\pi_i(\hat{\boldsymbol{\rho}})}$
4	1	0	1	.	1	0	0	$\frac{P(S_i=0\|Z_{2i}=0,X_i=1,Y_i=1;\hat{\boldsymbol{\theta}}^{(k)})}{1-\pi_i(\hat{\boldsymbol{\rho}})}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

where $\mathbf{D}_i(\boldsymbol{\beta}) = \partial\mu_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $V_i(\boldsymbol{\beta}) = \mu_i(\boldsymbol{\beta})(1-\mu_i(\boldsymbol{\beta}))$.

Similarly, the estimating function for the M-step for estimating $\boldsymbol{\xi}_2^{(k+1)}$ is

$$\overline{\mathbf{U}}_2(\boldsymbol{\xi}_2; \hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\rho}) \;=\; \sum_{i=1}^{n} \overline{\mathbf{U}}_{2i}(\boldsymbol{\xi}_2; \hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\rho})$$

where

$$\begin{aligned}\overline{\mathbf{U}}_{2i}(\boldsymbol{\xi}_2; \hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\rho}) \;=\; & \mathbf{D}_i(\boldsymbol{\xi}_2)[V_i(\boldsymbol{\xi}_2)]^{-1}\Bigg[\frac{R_i}{\pi_i(\boldsymbol{\rho})}\left(S_i - \pi_i(\boldsymbol{\xi}_2)\right) \\ & + \frac{(1-R_i)}{(1-\pi_i(\boldsymbol{\rho}))}\sum_{j=0}^{1}\Big\{\left(j - \pi_i(\boldsymbol{\xi}_2)\right)P(S_i = j|\mathcal{D}_{3i}; \hat{\boldsymbol{\theta}}^{(k)})\Big\}\Bigg],\end{aligned}$$

where $\mathbf{D}_i(\boldsymbol{\xi}_2) = \partial\pi_i(\boldsymbol{\xi}_2)/\partial\boldsymbol{\xi}_2$ and $V_i(\boldsymbol{\xi}_2) = \pi_i(\boldsymbol{\xi}_2)(1-\pi_i(\boldsymbol{\xi}_2))$. If we iterate between the E-step and the M-step until $|\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}}^{(k+1)}| < \epsilon$ for a fixed tolerance ϵ, the solution $\hat{\boldsymbol{\theta}}$ is obtained. In other words, we (i) obtain initial estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_2$ denoted by $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\xi}}_2^{(1)}$; (ii) calculate the weights for each row for each individual using $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\theta}}^{(1)}$; (iii) fit each weighted GLM to obtain $\hat{\boldsymbol{\beta}}^{(2)}$ and $\hat{\boldsymbol{\xi}}_2^{(2)}$. Steps (ii) and (iii) are repeated until the

convergence criteria are met. Initial estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_2$ can be obtained by fitting two separate unweighted GLMs using complete cases only.

Let $\boldsymbol{\phi} = (\boldsymbol{\theta}^T, \boldsymbol{\rho}^T)^T$. If $\overline{\boldsymbol{\phi}}$ denotes the estimate of $\boldsymbol{\phi}$ at convergence, the estimate of the asymptotic covariance matrix of $\overline{\boldsymbol{\beta}}$ is the upper 4×4 block of matrix $\Sigma(\boldsymbol{\phi})$ which has the sandwich form. See Section 3.4.4 for details on variance estimation in a more generalized setting.

We have proposed a type of weighted EM algorithm that can be used for handling missing data in the randomized setting when marginal regression coefficients are of interest. An inverse probability weight for missing data is introduced to account for non-ignorably missing subgroup data.

3.4 Estimation of Marginal Causal Effects in an Observational Setting

Here we introduce a doubly inverse probability weighted EM-type algorithm for estimating marginal causal effects in an observational setting where the subgroup variable is incompletely observed.

3.4.1 Notation and Models

In Section 3.2.1 a list of conditional independence assumptions and the logistic regression models that generate the variables are provided in an observational setting. As before, the binary response Y_i for subject i is generated by the conditional probability

$$P(Y_i = 1 | W_i, S_i, Z_{1i}; \boldsymbol{\vartheta}) \ ,$$

but interest lies in the marginal model $Y_i | X_i, S_i$ where the causal effect of X_i at different levels of the subgroup variable S_i is of primary interest. Therefore, interest lies in estimating $\boldsymbol{\beta}$ from (3.16). However in this observational setting, the treatment is not randomized, and we make the assumption that treatment selection depends on confounder Z_1 and auxiliary variable Z_2.

3.4.2 Doubly Weighted EM-Type Algorithm

The joint probability for $R_i, Y_i, W_i, S_i, Z_{2i}$ for subject i can be factored as

$$\begin{aligned} P(R_i, Y_i, W_i, S_i, Z_{2i}) &= P(R_i|Y_i, W_i, S_i, Z_{2i})P(Y_i|W_i, S_i, Z_{2i}) \\ &\quad \cdot P(W_i|S_i, Z_{2i})P(S_i|Z_{2i})P(Z_{2i}) \ . \end{aligned} \tag{3.18}$$

As in Section 3.3.2, we cannot factor the response variable from the conditional probability of the missing data indicator when we do not condition on Z_1, therefore the missing subgroup data is non-ignorable. As well, confounding is an issue since we are in an observational setting where treatment selection is dependent upon confounder Z_1.

In Section 3.3.2, we introduced a 'fix' for estimating marginal regression coefficients which involved weighting the estimating function at the M-step by the inverse of $P(R_i = r|Z_{1i}, Z_{2i}; \boldsymbol{\rho})$. Here, in the observational data setting, we propose the use of two inverse probability weights for both missing data and confounding.

Let

$$\mathcal{D}_{4i} = \{Y_i, W_i, S_i, Z_{2i} \text{ for } R_i = 1; \text{ or } Y_i, W_i, Z_{2i} \text{ for } R_i = 0\} \ ,$$

$\mathcal{D}_4 = \{\mathcal{D}_{4i},\ i = 1, ..., n\}$, and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\xi}_2^T)^T$ as before. Also, let $\boldsymbol{\psi} = (\boldsymbol{\rho}^T, \boldsymbol{\xi}_1^T)^T$ represent the parameters in the inverse probability weight for missing data and confounding.

At the E-step, the conditional expectation of a working loglikelihood for $\boldsymbol{\theta}$ with respect to $\mathcal{D}_4$ at the kth iteration is written as:

$$\widetilde{\widetilde{Q}}_i(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\psi}) = \widetilde{\widetilde{Q}}_{1i}(\boldsymbol{\beta}; \hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\psi}) + \widetilde{\widetilde{Q}}_{2i}(\boldsymbol{\xi}_2; \hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\psi}) \tag{3.19}$$

where

$$\begin{aligned} \widetilde{\widetilde{Q}}_{1i}(\boldsymbol{\beta}; \hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{\psi}) &= \sum_{l=0}^{1} \frac{I(W_i = l)}{\pi_i(\boldsymbol{\xi}_1)^l (1 - \pi_i(\boldsymbol{\xi}_1))^{1-l}} \Bigg[\frac{R_i}{\pi_i(\boldsymbol{\rho})} \log P(Y_i|W_i, S_i; \boldsymbol{\beta}) \\ &\quad + \frac{1 - R_i}{1 - \pi_i(\boldsymbol{\rho})} \sum_{j=0}^{1} \log P(Y_i|W_i, S_i = j; \boldsymbol{\beta}) P(S_i = j|\mathcal{D}_{4i}; \hat{\boldsymbol{\theta}}^{(k)}) \Bigg] \end{aligned}$$

and

$$\begin{aligned}\widetilde{\widetilde{Q}}_{2i}(\boldsymbol{\xi}_2;\hat{\boldsymbol{\theta}}^{(k)},\boldsymbol{\psi}) &= \sum_{l=0}^{1}\frac{I(W_i=l)}{\pi_i(\boldsymbol{\xi}_1)^l(1-\pi_i(\boldsymbol{\xi}_1))^{1-l}}\Bigg[\frac{R_i}{\pi_i(\boldsymbol{\rho})}\log P(S_i|Z_{2i};\boldsymbol{\xi}_2)\\ &\quad+\frac{1-R_i}{1-\pi_i(\boldsymbol{\rho})}\sum_{j=0}^{1}\log P(S_i=j|Z_{2i};\boldsymbol{\xi}_2)P(S_i=j|\mathcal{D}_{4i};\hat{\boldsymbol{\theta}}^{(k)})\Bigg]\ .\end{aligned}$$

For subjects with $R_i=0$,

$$P(S_i=j|\mathcal{D}_{4i};\hat{\boldsymbol{\theta}}^{(k)})=\frac{P(Y_i|W_i,S_i=j;\hat{\boldsymbol{\beta}}^{(k)})P(S_i=j|Z_{2i};\hat{\boldsymbol{\xi}}_2^{(k)})}{\sum_{s=0}^{1}P(Y_i|W_i,S_i=s;\hat{\boldsymbol{\beta}}^{(k)})P(S_i=s|Z_{2i};\hat{\boldsymbol{\xi}}_2^{(k)})}\ .$$

In the M-step, we fit two weighted GLMs to maximize $\widetilde{\widetilde{Q}}_1(\boldsymbol{\beta};\hat{\boldsymbol{\theta}}^{(k)})$ and $\widetilde{\widetilde{Q}}_2(\boldsymbol{\xi}_2;\hat{\boldsymbol{\theta}}^{(k)})$ to obtain $\hat{\boldsymbol{\theta}}^{(k+1)}$. The weighted estimating equation for the M-step of the weighted EM-type algorithm for obtaining estimate $\hat{\boldsymbol{\beta}}^{(k+1)}$ is

$$\begin{aligned}&\widetilde{\widetilde{\mathbf{U}}}_1(\boldsymbol{\beta};\hat{\boldsymbol{\theta}}^{(k)},\boldsymbol{\psi}) \qquad (3.20)\\ =\ &\sum_{i=1}^{n}\sum_{l=0}^{1}\frac{I(W_i=l)}{\pi_i(\boldsymbol{\xi}_1)^l(1-\pi_i(\boldsymbol{\xi}_1))^{1-l}}\Bigg[\frac{R_i}{\pi_i(\boldsymbol{\rho})}\mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\big(Y_i-\mu_i(\boldsymbol{\beta})\big)\\ &+\ \frac{1-R_i}{1-\pi_i(\boldsymbol{\rho})}\sum_{j=0}^{1}\mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\Big\{\big(Y_i-\mu_i(\boldsymbol{\beta})\big)P(S_i=j|\mathcal{D}_{4i};\hat{\boldsymbol{\theta}}^{(k)})\Big\}\Bigg]\ ,\end{aligned}$$

where $\mathbf{D}_i(\boldsymbol{\beta})=\partial\mu_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, and $V_i(\boldsymbol{\beta})=\mu_i(\boldsymbol{\beta})(1-\mu_i(\boldsymbol{\beta}))$.

Similarly, the weighted estimating equation at the M-step for obtaining $\hat{\boldsymbol{\xi}}_2^{(k+1)}$ is

$$\begin{aligned}&\widetilde{\widetilde{\mathbf{U}}}_2(\boldsymbol{\xi}_2;\hat{\boldsymbol{\theta}}^{(k)},\boldsymbol{\psi}) \qquad (3.21)\\ =\ &\sum_{i=1}^{n}\sum_{l=0}^{1}\mathbf{D}_i(\boldsymbol{\xi}_2)[V_i(\boldsymbol{\xi}_2)]^{-1}\frac{I(W_i=l)}{\pi_i(\boldsymbol{\xi}_1)^l(1-\pi_i(\boldsymbol{\xi}_1))^{1-l}}\Bigg[\frac{R_i}{\pi_i(\boldsymbol{\rho})}\big(S_i-\pi_i(\boldsymbol{\xi}_2)\big)\\ &+\frac{1-R_i}{1-\pi_i(\boldsymbol{\rho})}\sum_{j=0}^{1}\Big\{\big(j-\pi_i(\boldsymbol{\xi}_2)\big)P(S_i=j|\mathcal{D}_{4i};\hat{\boldsymbol{\theta}}^{(k)})\Big\}\Bigg]\ ,\end{aligned}$$

where $\mathbf{D}_i(\boldsymbol{\xi}_2)=\partial\pi_i(\boldsymbol{\xi}_2)/\partial\boldsymbol{\xi}_2$ and $V_i(\boldsymbol{\xi}_2)=\pi_i(\boldsymbol{\xi}_2)(1-\pi_i(\boldsymbol{\xi}_2))$.

Table 3.3: Layout of a dataset for the doubly weighted EM-type algorithm, at the $(k+1)$st M-step.

i	Z_{1i}	Z_{2i}	W_i	S_i	Y_i	R_i	j	weight
1	1	0	0	1	1	1	1	$\frac{1}{\pi_i(\hat{\boldsymbol{\rho}})(1-\pi_i(\hat{\boldsymbol{\xi}}_1))}$
1	1	0	0	1	1	1	0	0
2	1	1	0	.	0	0	1	$\frac{P(S_i=1\|Z_{2i}=1,X_i=0,Y_i=0;\hat{\boldsymbol{\theta}}^{(k)})}{(1-\pi_i(\hat{\boldsymbol{\rho}}))(1-\pi_i(\hat{\boldsymbol{\xi}}_1))}$
2	1	1	0	.	0	0	0	$\frac{P(S_i=0\|Z_{2i}=1,X_i=0,Y_i=0;\hat{\boldsymbol{\theta}}^{(k)})}{(1-\pi_i(\hat{\boldsymbol{\rho}}))(1-\pi_i(\hat{\boldsymbol{\xi}}_1))}$
3	0	1	1	0	0	1	1	0
3	0	1	1	0	0	1	0	$\frac{1}{\pi_i(\hat{\boldsymbol{\rho}})\pi_i(\boldsymbol{\xi}_1)}$
4	1	0	1	.	1	0	1	$\frac{P(S_i=1\|Z_{2i}=0,X_i=1,Y_i=1;\hat{\boldsymbol{\theta}}^{(k)})}{(1-\pi_i(\hat{\boldsymbol{\rho}}))\pi_i(\hat{\boldsymbol{\xi}}_1)}$
4	1	0	1	.	1	0	0	$\frac{P(S_i=0\|Z_{2i}=0,X_i=1,Y_i=1;\hat{\boldsymbol{\theta}}^{(k)})}{(1-\pi_i(\hat{\boldsymbol{\rho}}))\pi_i(\hat{\boldsymbol{\xi}}_1)}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

In practice, the inverse probability weights are unknown, so the first step is to fit a logistic regression model using the full dataset to obtain $\hat{\boldsymbol{\rho}}$. The second step is to fit another logistic regression model to obtain $\hat{\boldsymbol{\xi}}_1$, again using the full dataset. The inverse probability weights are then replaced with their estimated counterparts. The inverse probability weight estimates are unchanged throughout the iterations and only $\boldsymbol{\theta}$ is updated at each iteration.

For estimation we iterate between the E-step and the M-step until $|\hat{\boldsymbol{\theta}}^{(k)} - \hat{\boldsymbol{\theta}}^{(k+1)}| < \epsilon$, for fixed ϵ. We do this in the same way as described for estimating $\boldsymbol{\vartheta}$. A "long" dataset is constructed with two rows for each subject. In the first row, a new variable j is set to 1, and in the second row j is set to 0. The weight in rows with $R_i = 0$ is $P(S_i = j|Y_i, W_i, Z_{2i}; \hat{\boldsymbol{\theta}}^{(k)})/[(1-\pi_i(\hat{\boldsymbol{\rho}}))\pi_i(\hat{\boldsymbol{\xi}}_1)^{W_i}(1-\pi_i(\hat{\boldsymbol{\xi}}_1))^{1-W_i}]$. For rows with $R_i = 1$, the weight is 0 when $j \neq S_i$, and $1/[\pi_i(\hat{\boldsymbol{\rho}})\pi_i(\hat{\boldsymbol{\xi}}_1)^{W_i}(1-\pi_i(\hat{\boldsymbol{\xi}}_1))^{1-W_i}]$ when $j = S_i$. See Table 3.3 for an example of a "long" dataset with weights.

Let $\tilde{\tilde{\boldsymbol{\beta}}}$ denote the estimate of $\boldsymbol{\beta}$ when convergence criteria are met, and likewise let $\tilde{\tilde{\boldsymbol{\theta}}}$ denote the estimate of $\boldsymbol{\theta}$.

In the following sections, we show that $\tilde{\tilde{\boldsymbol{\beta}}}$ is a consistent estimator of $\boldsymbol{\beta}$ by showing that

the estimating function (3.20) is unbiased.

3.4.3 Consistency Results

Theorem 2. *Let* $\tilde{\tilde{\boldsymbol{\beta}}}$ *be the solution to* $\widetilde{\widetilde{\mathbf{U}}}_1(\boldsymbol{\beta};\boldsymbol{\theta},\boldsymbol{\psi}) = \mathbf{0}$. *Then* $\tilde{\tilde{\boldsymbol{\beta}}}$ *is a consistent estimator of* $\boldsymbol{\beta}$ *from (3.16).*

Proof: Consistency results can be established by showing that estimating function (3.20) is an unbiased estimating function for $\boldsymbol{\beta}$. Note that, although we observe the treatment variable in the context of an observational study, the response model is written in terms of Y, X, S since we are interested in estimating the effects that we would estimate using randomized data where the treatment selection process is randomized and X is independent of (Z_1, Z_2). Also, in a logistic regression setting, $\mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1} = (1, W_i, S_i, W_iS_i)^T$.

First, we take the conditional expectation with respect to $R|Y, W, S, Z_1, Z_2$ (dropping the i subscript) which by Assumption B.0 of Section 3.2.1 gives

$$\begin{aligned}
&\sum_{l=0}^{1} \frac{I(W=l)}{\pi(\boldsymbol{\xi}_1)^l(1-\pi(\boldsymbol{\xi}_1))^{(1-l)}} \Bigg\{ \frac{P(R=1|Z_1,Z_2;\boldsymbol{\rho})}{\pi(\boldsymbol{\rho})}(1,W,S,WS)^T\big(Y - P(Y=1|X=l,S;\boldsymbol{\beta})\big) \\
&+ \frac{P(R=0|Z_1,Z_2;\boldsymbol{\rho})}{1-\pi(\boldsymbol{\rho})} \sum_{j=0}^{1} \Bigg[(1,W,j,Wj)^T\big(Y - P(Y=1|X=l,S=j;\boldsymbol{\beta})\big) \\
&\cdot P(S=j|Y,X=l,Z_2;\boldsymbol{\theta}) \Bigg] \Bigg\} \\
= &\sum_{l=0}^{1} \frac{I(W=l)}{\pi(\boldsymbol{\xi}_1)^l(1-\pi(\boldsymbol{\xi}_1))^{(1-l)}} \Bigg\{ (1,W,S,WS)^T\big(Y - P(Y=1|X=l,S;\boldsymbol{\beta})\big) \\
&+ \sum_{j=0}^{1} (1,W,j,Wj)^T \Big[\big(Y - P(Y=1|X=l,S=j;\boldsymbol{\beta})\big) P(S=j|Y,X=l,Z_2;\boldsymbol{\theta}) \Big] \Bigg\} .
\end{aligned}$$

Next, we take the conditional expectation of this with respect to $Y|W, S, Z_1, Z_2$ which by

Assumption B.1 of Section 3.2.1 gives

$$
\begin{aligned}
&\sum_{l=0}^{1} \frac{I(W=l)}{\pi(\boldsymbol{\xi}_1)^l(1-\pi(\boldsymbol{\xi}_1))^{(1-l)}} \Big\{ (1,W,S,WS)^T \big(P(Y=1|W,S,Z_1;\boldsymbol{\vartheta}) - P(Y=1|X=l,S;\boldsymbol{\beta})\big) \\
&+ \sum_{j=0}^{1} (1,W,j,Wj)^T \Big[\sum_{y=0}^{1} \big(y - P(Y=1|X=l,S=j;\boldsymbol{\beta})\big) \\
&\cdot P(S=j|Y=y,X=l,Z_2;\boldsymbol{\theta}) P(Y=y|W,S,Z_1;\boldsymbol{\vartheta}) \Big] \Big\} \\
= &\sum_{l=0}^{1} \frac{I(W=l)}{\pi(\boldsymbol{\xi}_1)^l(1-\pi(\boldsymbol{\xi}_1))^{(1-l)}} \Big\{ (1,W,S,WS)^T \big(P(Y=1|W,S,Z_1;\boldsymbol{\vartheta}) - P(Y=1|X=l,S;\boldsymbol{\beta})\big) \\
&+ \sum_{j=0}^{1} (1,W,j,Wj)^T \Big[P(S=j|Y=1,X=l,Z_2;\boldsymbol{\theta}) P(Y=1|W,S,Z_1;\boldsymbol{\vartheta}) \\
&- P(Y=1|X=l,S=j;\boldsymbol{\beta}) \sum_{y=0}^{1} P(S=j|Y=y,X=l,Z_2;\boldsymbol{\theta}) P(Y=y|W,S,Z_1;\boldsymbol{\vartheta}) \Big] \Big\} .
\end{aligned}
$$

Next, we take the conditional expectation of this with respect to $W|S,Z_1,Z_2$ which by

Assumption B.2 of Section 3.2.1 gives

$$
\begin{aligned}
&\sum_{w=0}^{1}\sum_{l=0}^{1}\frac{I(w=l)P(W=w|Z_1,Z_2;\boldsymbol{\xi}_1)}{\pi(\boldsymbol{\xi}_1)^l(1-\pi(\boldsymbol{\xi}_1))^{(1-l)}}\Big\{(1,w,S,wS)^T\big(P(Y=1|W=w,S,Z_1;\boldsymbol{\vartheta})\\
&\quad -P(Y=1|X=l,S;\boldsymbol{\beta})\big)+\\
&\sum_{j=0}^{1}(1,w,j,wj)^T\Big[P(S=j|Y=1,X=l,Z_2;\boldsymbol{\theta})P(Y=1|W=w,S,Z_1;\boldsymbol{\vartheta})\\
&-P(Y=1|X=l,S=j;\boldsymbol{\beta})\sum_{y=0}^{1}P(S=j|Y=y,X=l,Z_2;\boldsymbol{\theta})P(Y=y|W=w,S,Z_1;\boldsymbol{\vartheta})\Big]\Big\}\\
=\;&\sum_{w=0}^{1}\frac{P(W=w|Z_1,Z_2;\boldsymbol{\xi}_1)}{\pi(\boldsymbol{\xi}_1)^w(1-\pi(\boldsymbol{\xi}_1))^{(1-w)}}\Big\{(1,w,S,wS)^T\big(P(Y=1|W=w,S,Z_1;\boldsymbol{\vartheta})\\
&-P(Y=1|X=w,S;\boldsymbol{\beta})\big)\\
&+\sum_{j=0}^{1}(1,w,j,wj)^T\Big[P(S=j|Y=1,X=w,Z_2;\boldsymbol{\theta})P(Y=1|W=w,S,Z_1;\boldsymbol{\vartheta})\\
&-P(Y=1|X=w,S=j;\boldsymbol{\beta})\\
&\cdot\sum_{y=0}^{1}P(S=j|Y=y,X=w,Z_2;\boldsymbol{\theta})P(Y=y|W=w,S,Z_1;\boldsymbol{\vartheta})\Big]\Big\}\\
=\;&\sum_{w=0}^{1}\Big\{(1,w,S,wS)^T\big(P(Y=1|W=w,S,Z_1;\boldsymbol{\vartheta})-P(Y=1|X=w,S;\boldsymbol{\beta})\big)\\
&+\sum_{j=0}^{1}(1,w,j,wj)^T\Big[P(S=j|Y=1,X=w,Z_2;\boldsymbol{\theta})P(Y=1|W=w,S,Z_1;\boldsymbol{\vartheta})\\
&-P(Y=1|X=w,S=j;\boldsymbol{\beta})\\
&\cdot\sum_{y=0}^{1}P(S=j|Y=y,X=w,Z_2;\boldsymbol{\theta})P(Y=y|W=w,S,Z_1;\boldsymbol{\vartheta})\Big]\Big\}
\end{aligned}
$$

since the indicator $I(W_i = l)$ ensures only one term is retained (when $l = w$ for $w = 0, 1$). Finally, we take the expectation of this with respect to S, Z_1, Z_2 which by Assumptions

B.3 and B.4 of Section 3.2.1 gives

$$\sum_{z_2=0}^{1}\sum_{z_1=0}^{1}\sum_{s=0}^{1}\sum_{w=0}^{1}\Big\{(1,w,s,ws)^T\big(P(Y=1|W=w,S=s,Z_1=z_1;\boldsymbol{\vartheta})$$
$$-P(Y=1|X=w,S=s;\boldsymbol{\beta})\big)$$
$$+\sum_{j=0}^{1}(1,w,j,wj)^T\Big[P(S=j|Y=1,X=w,z_2;\boldsymbol{\theta})P(Y=1|W=w,S=s,z_1;\boldsymbol{\vartheta})$$
$$-P(Y=1|X=w,S=j;\boldsymbol{\beta})$$
$$\cdot\sum_{y=0}^{1}P(S=j|Y=y,X=w,z_2;\boldsymbol{\theta})P(Y=y|W=w,S=s,z_1;\boldsymbol{\vartheta})\Big]\Big\}$$
$$\cdot P(S=s|Z_2=z_2;\boldsymbol{\xi}_2)P(Z_1=z_1;\zeta_1)P(Z_2=z_2;\zeta_2)\ .$$

Since

$$P(Y=1|W=w,S=s,Z_1=z_1;\boldsymbol{\vartheta})=P(Y=1|X=w,S=s,Z_1=z_1;\boldsymbol{\vartheta})\ ,$$

the expectation of the estimating equation can be written as

$$\sum_{z_2=0}^{1}\sum_{s=0}^{1}\sum_{w=0}^{1}\Big\{(1,w,s,ws)^T\big(P(Y=1|X=w,S=s;\boldsymbol{\beta})P(S=s|Z_2=z_2;\boldsymbol{\xi}_2)P(Z_2=z_2;\zeta_2)$$
$$-P(Y=1|X=w,S=s;\boldsymbol{\beta})P(S=s|Z_2=z_2;\boldsymbol{\xi}_2)P(Z_2=z_2;\zeta_2)\big)\Big\}$$
$$+\sum_{z_2=0}^{1}\sum_{z_1=0}^{1}\sum_{s=0}^{1}\sum_{w=0}^{1}\Big\{\sum_{j=0}^{1}(1,w,j,wj)^T$$
$$\cdot\Big[P(S=j|Y=1,X=w,z_2;\boldsymbol{\theta})P(Y=1|X=w,S=s,z_1;\boldsymbol{\vartheta})$$
$$-P(Y=1|X=w,S=j;\boldsymbol{\beta})$$
$$\cdot\sum_{y=0}^{1}P(S=j|Y=y,X=w,z_2;\boldsymbol{\theta})P(Y=y|X=w,S=s,z_1;\boldsymbol{\vartheta})\Big]\Big\}$$
$$\cdot P(S=s|Z_2=z_2;\boldsymbol{\xi}_2)P(Z_1=z_1;\zeta_1)P(Z_2=z_2;\zeta_2)$$

since

$$\sum_{z_1=0}^{1}P(Y=1|X=w,S=s,Z_1=z_1;\boldsymbol{\vartheta})P(Z_1=z_1;\zeta_1)=P(Y=1|X=w,S=s;\boldsymbol{\beta})\ .$$

We can then write the expectation as

$$
\begin{aligned}
= & \sum_{z_2=0}^{1}\sum_{z_1=0}^{1}\sum_{s=0}^{1}\sum_{w=0}^{1}\Bigg\{\sum_{j=0}^{1}(1,w,j,wj)^T\Bigg[\frac{P(Y=1|X=w,S=j;\boldsymbol{\beta})P(S=j|z_2;\boldsymbol{\xi}_2)}{\sum_{r=0}^{1}P(Y=1|X=w,S=r;\boldsymbol{\beta})P(S=r|z_2;\boldsymbol{\xi}_2)} \\
& \cdot P(Y=1|X=w,S=s,z_1;\boldsymbol{\vartheta})P(S=s|Z_2=z_2;\boldsymbol{\xi}_2)P(Z_1=z_1;\zeta_1)P(Z_2=z_2;\zeta_2) \\
& -P(Y=1|X=w,S=j;\boldsymbol{\beta})\sum_{y=0}^{1}\frac{P(Y=y|X=w,S=j;\boldsymbol{\beta})P(S=j|z_2;\boldsymbol{\xi}_2)}{\sum_{r=0}^{1}P(Y=y|X=w,S=r;\boldsymbol{\beta})P(S=r|z_2;\boldsymbol{\xi}_2)} \\
& \cdot P(Y=y|X=w,S=s,z_1;\boldsymbol{\vartheta})P(S=s|Z_2=z_2;\boldsymbol{\xi}_2)P(Z_1=z_1;\zeta_1)P(Z_2=z_2;\zeta_2)\Bigg]\Bigg\} \\
= & \sum_{z_2=0}^{1}\sum_{w=0}^{1}\Bigg\{\sum_{j=0}^{1}(1,w,j,wj)^T\Bigg[P(Y=1|X=w,S=j;\boldsymbol{\beta})P(S=j|z_2;\boldsymbol{\xi}_2)P(Z_2=z_2;\zeta_2) \\
& -P(Y=1|X=w,S=j;\boldsymbol{\beta}) \\
& \cdot \sum_{y=0}^{1}P(Y=y|X=w,S=j;\boldsymbol{\beta})P(S=j|z_2;\boldsymbol{\xi}_2)P(Z_2=z_2;\zeta_2)\Bigg]\Bigg\} \\
= & \mathbf{0}
\end{aligned}
$$

We have thus shown that $\tilde{\tilde{\boldsymbol{\beta}}}$ is a consistent estimator of $\boldsymbol{\beta}$.

3.4.4 Derivation of the Asymptotic Variance

Let $\mathbf{U}_3(\boldsymbol{\rho})$ represent the standard logistic regression estimating function for estimating $\boldsymbol{\rho}$, and let $\mathbf{U}_4(\boldsymbol{\xi}_1)$ represent the standard logistic regression estimating function for estimating $\boldsymbol{\xi}_1$. Then

$$
\tilde{\tilde{\mathbf{U}}}(\boldsymbol{\omega};\boldsymbol{\theta}) = \begin{pmatrix} \tilde{\tilde{\mathbf{U}}}_1(\boldsymbol{\beta};\boldsymbol{\theta},\boldsymbol{\psi}) \\ \tilde{\tilde{\mathbf{U}}}_2(\boldsymbol{\xi}_2;\boldsymbol{\theta},\boldsymbol{\psi}) \\ \mathbf{U}_3(\boldsymbol{\rho}) \\ \mathbf{U}_4(\boldsymbol{\xi}_1) \end{pmatrix} \tag{3.22}
$$

is the joint estimating function for $\boldsymbol{\omega} = (\boldsymbol{\theta}^T, \boldsymbol{\psi}^T)^T$ obtained by replacing the $\hat{\boldsymbol{\theta}}^{(k)}$ terms in (3.20) and (3.21) with $\boldsymbol{\theta}$.

Here we derive the asymptotic covariance matrix for $\tilde{\tilde{\boldsymbol{\beta}}}$ by generalizing the Louis (1982) method (Section 3.1.2) to incorporate doubly inverse probability weights (Section 2.3.2).

The asymptotic covariance matrix for $\widetilde{\widetilde{\boldsymbol{\omega}}}$ takes the form

$$\mathcal{I}(\boldsymbol{\omega})^{-1}\mathcal{C}(\boldsymbol{\omega})[\mathcal{I}(\boldsymbol{\omega})^{-1}]^T \ , \tag{3.23}$$

where

$$\mathcal{I}(\boldsymbol{\omega}) = \begin{bmatrix} \mathcal{I}_{11}(\boldsymbol{\omega}) & \mathcal{I}_{12}(\boldsymbol{\omega}) \\ \mathbf{0} & \mathcal{I}_{22}(\boldsymbol{\omega}) \end{bmatrix} \ ,$$

and

$$\mathcal{C}(\boldsymbol{\omega}) = E[\widetilde{\widetilde{\mathbf{U}}}_i(\boldsymbol{\omega}), \widetilde{\widetilde{\mathbf{U}}}_i(\boldsymbol{\omega})^T] \ .$$

$\mathcal{I}_{11}(\boldsymbol{\omega})$ is the expected value of the observed information $I_{11}(\boldsymbol{\omega})$ given by

$$\begin{aligned} I_{11}(\boldsymbol{\omega}) \ = \ & -\ddot{Q}(\boldsymbol{\theta};\boldsymbol{\theta},\boldsymbol{\psi}) \\ & -\Bigg\{\sum_{i=1}^{n}\sum_{l=0}^{1}\frac{I(W_i=l)}{\pi_i(\boldsymbol{\xi}_1)^l(1-\pi_i(\boldsymbol{\xi}_1))^{1-l}}\Bigg(\frac{R_i}{\pi_i(\boldsymbol{\rho})}\mathbb{S}_i(\mathcal{D}_{4i};\boldsymbol{\theta})\,\mathbb{S}_i^T(\mathcal{D}_{4i};\boldsymbol{\theta}) \\ & +\frac{(1-R_i)}{1-\pi_i(\boldsymbol{\rho})}\sum_{j=0}^{1}\mathbb{S}_i(\mathcal{D}_{4i},S_i=j;\boldsymbol{\theta})\,\mathbb{S}_i^T(\mathcal{D}_{4i},S_i=j;\boldsymbol{\theta})P(S_i=j|\mathcal{D}_{4i};\boldsymbol{\theta})\Bigg) \\ & -\sum_{i=1}^{n}\dot{Q}_i(\boldsymbol{\theta};\boldsymbol{\theta},\boldsymbol{\psi})\,\dot{Q}_i^T(\boldsymbol{\theta};\boldsymbol{\theta},\boldsymbol{\psi})\Bigg\} \ . \end{aligned}$$

For $R_i = 1$,

$$\mathbb{S}_i(\mathcal{D}_{4i};\boldsymbol{\theta}) = \begin{pmatrix} \mathbb{S}_{1i}(Y_i,W_i,S_i;\boldsymbol{\beta}) \\ \mathbb{S}_{2i}(S_i,Z_{2i};\boldsymbol{\xi}_2) \end{pmatrix}$$

where

$$\mathbb{S}_{1i}(Y_i,W_i,S_i;\boldsymbol{\beta}) = \begin{pmatrix} 1 \\ W_i \\ S_i \\ W_iS_i \end{pmatrix}(Y_i-\mu_i(\boldsymbol{\beta}))$$

and

$$\mathbb{S}_{2i}(S_i,Z_{2i};\boldsymbol{\xi}_2) = \begin{pmatrix} 1 \\ Z_{2i} \end{pmatrix}(S_i-\pi_i(\boldsymbol{\xi}_2)) \ .$$

Likewise when $R_i = 0$,

$$\mathbb{S}_i(\mathcal{D}_{4i}, S_i = j; \boldsymbol{\theta}) = \begin{pmatrix} \mathbb{S}_{1i}(Y_i, W_i, S_i = j; \boldsymbol{\beta}) \\ \mathbb{S}_{2i}(S_i = j, Z_{2i}; \boldsymbol{\xi}_2) \end{pmatrix} .$$

Also,

$$\begin{aligned} \dot{Q}_i(\boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\psi}) &= \sum_{l=0}^{1} \frac{I(W_i = l)}{\pi_i(\boldsymbol{\xi}_1)^l (1 - \pi_i(\boldsymbol{\xi}_1))^{1-l}} \Bigg(\frac{R_i}{\pi_i(\boldsymbol{\rho})} \mathbb{S}_i(\mathcal{D}_{4i}; \boldsymbol{\theta}) \\ &\quad + \frac{(1 - R_i)}{1 - \pi_i(\boldsymbol{\rho})} \sum_{j=0}^{1} \mathbb{S}_i(\mathcal{D}_{4i}, S_i = j; \boldsymbol{\theta}) P(S_i = j | \mathcal{D}_{4i}; \boldsymbol{\theta}) \Bigg) , \end{aligned}$$

and

$$\begin{aligned} \ddot{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}, \boldsymbol{\psi}) &= \sum_{i=1}^{n} \sum_{l=0}^{1} \frac{I(W_i = l)}{\pi_i(\boldsymbol{\xi}_1)^l (1 - \pi_i(\boldsymbol{\xi}_1))^{1-l}} \Bigg(\frac{R_i}{\pi_i(\boldsymbol{\rho})} \frac{\partial \mathbb{S}_i(\mathcal{D}_{4i}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \\ &\quad + \frac{1 - R_i}{1 - \pi_i(\boldsymbol{\rho})} \sum_{j=0}^{1} \frac{\partial \mathbb{S}_i(\mathcal{D}_{4i}, S_i = j; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} P(S_i = j | \mathcal{D}_{4i}; \boldsymbol{\theta}) \Bigg) . \end{aligned}$$

Next, we define the remaining sub-matrices of $\mathcal{I}(\boldsymbol{\omega})$:

$$\mathcal{I}_{12}(\boldsymbol{\omega}) = E[-\partial \widetilde{\overline{\mathbf{U}}}_i(\boldsymbol{\omega}; \boldsymbol{\theta}) / \partial \boldsymbol{\psi}^T]$$

and

$$\mathcal{I}_{22}(\boldsymbol{\omega}) = E[-\partial \mathbf{U}_{\boldsymbol{\psi},i}(\boldsymbol{\psi}) / \partial \boldsymbol{\psi}^T]$$

where

$$\mathbf{U}_{\boldsymbol{\psi},i}(\boldsymbol{\psi}) = \begin{pmatrix} \mathbf{U}_{3i}(\boldsymbol{\rho}) \\ \mathbf{U}_{4i}(\boldsymbol{\xi}_1) \end{pmatrix} .$$

As in Chapter 2, we replace $\boldsymbol{\omega}$ by its estimated counterpart $\widetilde{\overline{\boldsymbol{\omega}}}$ to obtain variance estimates. We have given a method for obtaining the asymptotic covariance matrix for $\widetilde{\overline{\boldsymbol{\beta}}}$.

3.5 Simulation Studies

Here, we describe simulation studies conducted in order to explore the bias and relative efficiency of estimates obtained from the EM-type algorithm with inverse probability weights described in this chapter. In the marginal response model, we are most interested in assessing the consistency and relative efficiency of our estimates of β_1 and $\beta_1 + \beta_3$, where β_1 is the log odds ratio of response comparing those who received the treatment to those who did not when $S = 0$; $\beta_1 + \beta_3$ is the log odds ratio when $S = 1$.

We also compare the methods introduced in this chapter to the doubly inverse probability weighted estimating equation approach introduced in Chapter 2.

3.5.1 Parameter Settings

Refer to Section 2.4.1 for a list of parameters for our simulation studies. In this setting, where variable Z_1 is not independently associated with the subgroup variable, the model that generates the subgroup variable is the following logistic regression model:

$$\pi_i(\boldsymbol{\xi}_2) = \text{expit}(\text{logit}(0.5) + \log(1.0)Z_{1i} + \log(1.2)Z_{2i}) \quad .$$

As in Chapter 2, we vary the percentage of missing data to be 20% and 40%. Results of analyses using the EM algorithm described in Section 3.2 are given in Table 3.4. Results of the methods proposed in Sections 3.3 and 3.4 are given in Table 3.5. Also included in Table 3.5 is a comparison to the doubly inverse probability weighted estimating function method introduced in Chapter 2 where the same parameter settings are used in order to compare the efficiency of both methods.

Empirical bias (EBias) is defined as the average difference between the estimated log odds ratio and the true log odds ratio. The average asymptotic standard error (ASE) is calculated as the average of the estimated standard errors using equation (3.23). The empirical standard error (ESE) is defined as the standard deviation of the m log odds ratio estimates, where m is the number of simulated datasets. The empirical coverage probability (ECP) is the empirical probability that the true log odds ratio is included in

Table 3.4: Empirical biases and standard errors of estimated conditional causal effects using an EM algorithm in the setting of an incomplete subgroup variable.

	ϑ_1 [a]				$\vartheta_1 + \vartheta_3$ [a]			
	EBias	ASE	ESE	ECP	EBias	ASE	ESE	ECP
				$P(R=1)=0.8$				
Randomized setting								
EM algorithm	0.0001	0.1580	0.1567	0.9506	-0.0016	0.1502	0.1511	0.9482
Complete case	0.0027	0.1703	0.1688	0.9504	-0.0047	0.1627	0.1624	0.9488
Observational setting								
EM algorithm	-0.0035	0.1609	0.1601	0.9516	0.0007	0.1528	0.1533	0.9510
Complete case	-0.0013	0.1733	0.1735	0.9508	-0.0027	0.1655	0.1659	0.9540
				$P(R=1)=0.6$				
Randomized setting								
EM algorithm	-0.0013	0.1698	0.1679	0.9494	-0.0005	0.1600	0.1603	0.9512
Complete case	-0.0001	0.1991	0.1981	0.9486	-0.0048	0.1903	0.1923	0.9478
Observational setting								
EM algorithm	-0.0032	0.1726	0.1713	0.9534	0.0005	0.1626	0.1634	0.9510
Complete case	-0.0017	0.2024	0.2003	0.9524	-0.0036	0.1934	0.1966	0.9466

Abbreviations: EBias Empirical bias, ASE asymptotic standard error, ESE empirical standard error, ECP empirical coverage probability, EM expectation-maximization

[a] Conditional causal effect for $Z_1 = 0$.

the 95% confidence interval; for each simulated dataset, the confidence interval is computed using the estimated asymptotic standard deviation from equation (3.23), using a normal approximation.

For each simulation study, the number of subjects per dataset is 2000 and the number of simulated datasets is 5000. The simulated datasets are independent, and the seed value for the first simulated dataset is the same across the simulation studies.

Table 3.5: Empirical biases and standard errors of estimated marginal causal effects using a (doubly) inverse probability weighted EM-type algorithm in the setting of an incomplete subgroup variable.

	β_1				$\beta_1 + \beta_3$			
	EBias	ASE	ESE	ECP	EBias	ASE	ESE	ECP
				$P(R=1)=0.8$				
Randomized setting								
Weighted EM	0.0003	0.1558	0.1537	0.9514	-0.0013	0.1501	0.1513	0.9478
Weighted EE [a]	0.0010	0.1468	0.1458	0.9490	0.0026	0.1388	0.1407	0.9476
Observational setting								
Weighted EM	-0.0008	0.1759	0.1763	0.9470	0.0033	0.1692	0.1714	0.9460
Weighted EE [a]	0.0016	0.1662	0.1656	0.9498	0.0024	0.1571	0.1563	0.9492
				$P(R=1)=0.6$				
Randomized setting								
Weighted EM	0.0008	0.1546	0.1522	0.9570	0.0016	0.1448	0.1439	0.9546
Weighted EE [a]	-0.0005	0.1708	0.1711	0.9514	0.0009	0.1613	0.1606	0.9546
Observational setting								
Weighted EM	0.0001	0.1746	0.1743	0.9504	0.0020	0.1633	0.1628	0.9508
Weighted EE [a]	-0.0010	0.1936	0.1923	0.9556	0.0008	0.1826	0.1804	0.9546

Abbreviations: EBias Empirical bias, ASE asymptotic standard error, ESE empirical standard error, ECP empirical coverage probability, EM expectation-maximization, EE estimating equation

[a] The (doubly) weighted EE method introduced in Chapter 2 with a weight for missingness, and a weight for confounding in the observational setting.

3.5.2 Discussion of Simulation Study Results

In Table 3.4, the estimates for ϑ_1 and $\vartheta_1 + \vartheta_3$ are consistent using the EM algorithm described in Sections 3.1.2 and 3.2.2 for the randomized and observational settings respectively. The asymptotic standard error is larger in the observational setting, compared to the randomized setting. As expected, the asymptotic standard errors for $\hat{\vartheta}_1$ and $\hat{\vartheta}_1 + \hat{\vartheta}_3$ are smaller when the percentage of missingness is lower. The complete case analysis approach results in consistent estimates, however there is a loss of efficiency compared to the EM algorithm method.

In Table 3.5, the (doubly) inverse probability weighted EM-type algorithm performs well in terms of consistency, and the method is more efficient in the randomized setting. Interestingly, the weighted EM algorithm is as efficient when the percentage of missing data is high (40%) as when the percentage is lower (20%). In comparison to the weighted estimating equation approach introduced in Chapter 2, the weighted EM-type algorithm method is more efficient in the setting where there is more missing data (40%), and less efficient in the setting where there is less missing data (20%).

3.6 Discussion

In this chapter, we review the EM algorithm for estimating conditional regression parameters in randomized and observational settings with ignorably missing covariate data. We expanded upon this method in the setting where estimation of marginal regression parameters is of interest; a single inverse probability weight is used to account for non-ignorably missing covariate data in the randomized setting, and a double inverse probability weight is used to account for both missing data and confounding in an observational setting.

In simulation studies, the doubly weighted EM-type algorithm method performs well in terms of consistency. In estimating the marginal regression parameters, when the percentage of missingness is high, the doubly weighted EM-type algorithm method is more efficient than the doubly weighted estimating equation approach introduced in Chapter 2. Further study of the relative efficiency of the two approaches could be carried out using

the asymptotic variances analogous to the way relative efficiency was studied in McIsaac and Cook (2017), but we do not consider that in detail here.

The methods presented in this chapter are not ideal when the missing variable is continuous, or when more than one variable is missing. Also, variance estimation is not straightforward, even in the simplest setting. In the following chapter, we introduce a weighted multiple imputation method that is conceptually similar to the method presented in this chapter, but that is easier to use and is more generalizable.

Chapter 4

Multiple Imputation for Causal Inference with Incomplete Data

In Chapter 2, we investigate the use of complete case analysis with inverse probability weighted estimating equations for estimation of marginal regression coefficients to address (i) treatment selection and (ii) missing subgroup data. The weight for treatment selection is based on the propensity score, while the weight for missingness accounts for non-ignorably missing subgroup data in a marginal model setting. In Chapter 3, we explore the use of an EM algorithm approach, with two inverse probability weights for treatment selection and missing data. The doubly weighted method introduced in Chapter 2 is relatively straightforward, but requires making assumptions about the missing data process, which can be complex when there is more than one missing variable. The weighted EM-type algorithm introduced in Chapter 3 requires modeling the incomplete covariate and has the potential to be more efficient, but generalization to accommodate continuous or higher dimensional missing covariates is challenging. In this chapter, we introduce another approach that is easy to implement for estimation of marginal causal effects using partially incomplete observational data: multiple imputation with one inverse probability weight for treatment selection (confounding).

4.1 Estimation of Conditional Causal Effects in a Randomized Setting

4.1.1 Notation and Models

Here we adopt the notation and models introduced in the randomized setting of Chapter 2, Section 2.2.1. We let Y_i denote a binary response variable, X_i denote a binary treatment variable in a randomized setting, and S_i denote a binary subgroup variable, for subject i in a random sample of n individuals, $i = 1, ..., n$. Let Z_{1i} denote a confounding variable that is associated with treatment selection and the response, and Z_{2i} denote an auxiliary variable that does not have a direct effect on the response. The subgroup variable S_i is unknown for some individuals and so we let $R_i = I(S_i \text{ is observed})$ as before. Let $N_{\text{obs}} = \sum_{i=1}^{n} R_i$ be the number of subjects with an observed subgroup variable, and $N_{\text{mis}} = n - N_{\text{obs}}$ be the number of subjects with a missing subgroup variable. We let

$$\mathcal{R} = \{i : R_i = 1\} \text{ and } \mathcal{R}^c = \{i : R_i = 0\} \quad .$$

We make the following conditional independence assumptions in a randomized controlled trial setting:

A.0	$R_i \perp (Y_i, X_i, S_i) \mid Z_{1i}, Z_{2i}$
A.1	$Y_i \perp Z_{2i} \mid (X_i, S_i, Z_{1i})$
A.2	$X_i \perp (S_i, Z_{1i}, Z_{2i})$ (randomized setting)
A.3	$Z_{1i} \perp Z_{2i}$.

The assumptions allow us to factor the joint probability of $R_i, Y_i, X_i, S_i, Z_{1i}, Z_{2i}$ for subject i as

$$\begin{aligned} & P(R_i, Y_i, X_i, S_i, Z_{1i}, Z_{2i}) \\ = \; & P(R_i|Y_i, X_i, S_i, Z_{1i}, Z_{2i})P(Y_i|X_i, S_i, Z_{1i}, Z_{2i})P(X_i|S_i, Z_{1i}, Z_{2i})P(S_i|Z_{1i}, Z_{2i})P(Z_{1i}, Z_{2i}) \\ = \; & P(R_i|Z_{1i}, Z_{2i})P(Y_i|X_i, S_i, Z_{1i})P(X_i)P(S_i|Z_{1i}, Z_{2i})P(Z_{1i})P(Z_{2i}) \; . \end{aligned} \tag{4.1}$$

We define $\mu_i(\boldsymbol{\vartheta}) = E(Y_i|X_i, S_i, Z_{1i}) = P(Y_i = 1|X_i, S_i, Z_{1i}; \boldsymbol{\vartheta})$ as before and consider a logistic regression model for the response whereby

$$\mu_i(\boldsymbol{\vartheta}) \;\; = \;\; \text{expit}(\vartheta_0 + \vartheta_1 X_i + \vartheta_2 S_i + \vartheta_3 X_i S_i + \vartheta_4 X_i Z_{1i}) \; , \tag{4.2}$$

where $\boldsymbol{\vartheta} = (\vartheta_0, \vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4)^T$ is a vector of regression coefficients. Note that we have assumed that there is no main effect of Z_{1i} on Y_i, but this constraint can be relaxed.

We let $E(X_i) = P(X_i = 1) = 0.5$, so that the percentage of treated subjects in the randomized setting is 50%. For the auxiliary variables, $E(Z_{ji}) = P(Z_{ji} = 1; \zeta_j) = \text{expit}(\zeta_j)$, for $j = 1, 2$. The subgroup variable is modeled as follows:

$$\pi_i(\boldsymbol{\xi}_2) = E(S_i|Z_{1i}, Z_{2i}) = P(S_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_2) = \text{expit}(\xi_{20} + \xi_{21} Z_{1i} + \xi_{22} Z_{2i}) \; , \tag{4.3}$$

where $\boldsymbol{\xi}_2 = (\xi_{20}, \xi_{21}, \xi_{22})^T$ is a vector of regression coefficients. The model for the missing subgroup indicator is

$$\pi_i(\boldsymbol{\rho}) \;\; = \;\; E(R_i|Z_{1i}, Z_{2i}) \;\; = \;\; P(R_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\rho}) = \text{expit}(\rho_0 + \rho_1 Z_{1i} + \rho_2 Z_{2i}) \; ,$$

where $\boldsymbol{\rho} = (\rho_0, \rho_1, \rho_2)^T$ is a vector of regression coefficients.

Assume here that we are interested in estimating $\boldsymbol{\vartheta}$ from the full response model (4.2). In the multiple imputation literature, the *substantive model* is the response model of interest (Carpenter and Kenward, 2013); in this setting, the substantive model is the conditional response model in equation (4.2). In the substantive model, we condition on Z_{1i} so that the missing data process and the response model can be modeled separately, which means that the missingness is ignorable, see (4.1). Therefore it is not critical that we use multiple imputation, however we may wish to use multiple imputation in this setting to increase efficiency.

4.1.2 Imputation Model

Here, we describe the imputation model for the missing subgroup variable S_i to be used for subjects with $R_i = 0$. We use a *sequential imputation* approach to impute the missing

subgroup variable (Carpenter and Kenward, 2013). Sequential imputation involves modeling and imputing each missing variable separately. Rather than using a global multivariate distribution for all variables, each individual missing variable is modeled separately, typically using a regression analysis with one response variable. In our setting, there is one incomplete variable, therefore only one imputation model is necessary.

In terms of selecting an appropriate imputation model for the subgroup variable, the true conditional distribution of S_i for subject i given all observed data is

$$P(S_i|Y_i, X_i, Z_{1i}, Z_{2i}) = \frac{P(Y_i|X_i, S_i, Z_{1i}; \boldsymbol{\vartheta})P(S_i|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_2)}{\sum_{s=0}^{1} P(Y_i|X_i, S_i = s, Z_{1i}; \boldsymbol{\vartheta})P(S_i = s|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_2)} \ . \tag{4.4}$$

In practice, to obtain an estimate of the relationship between S_i and the fully observed variables, we use data from subjects with $R_i = 1$ to fit a logistic regression model to approximate (4.4), where S_i is the response variable, and we condition on all observed variables. The logistic regression model to impute the missing subgroup variable has the following general structure:

$$P(S_i = 1|Y_i, X_i, Z_{1i}, Z_{2i}; \boldsymbol{\xi}_3) \quad = \quad \text{expit}(\mathbf{A}_i\boldsymbol{\xi}_3) \ , \tag{4.5}$$

where $\boldsymbol{\xi}_3$ is the column vector of imputation model regression parameters, and $\mathbf{A}_i$ is a row vector of variables which are a combination of the observed variables Y_i, X_i, Z_{1i}, Z_{2i}, which can include main effects and interaction terms.

A *congenial imputation model* is one that is compatible with the substantive response model (Meng, 1994). This means that any variable and any interaction term that is included in the response model, is included in the imputation model. Rubin (1996) offers the following guidelines for choosing imputation covariates: (i) variables that are included in the model that generates the missing variable (in this setting, these variables are Z_1 and Z_2), see (4.3); and (ii) all of the variables that are included in the substantive response model (in our setting, these variables are Y, X, S and Z_1) (Rubin, 1996). Barnard and Meng (1999) also recommend that the imputation model should include information about the missing data process, while avoiding over-fitting.

Given the above recommendations, we propose a formal method that can be used to choose the two-way interaction terms necessary in the imputation model. To approximate

the true conditional distribution of S using a logistic regression model, we examine the odds ratio of S for a particular variable, at at different levels of another variable, using the true conditional distribution (4.4). For example, to assess whether we need a Z_1Z_2 interaction term, we look at the odds of S comparing $Z_2 = 1$ to $Z_2 = 0$, for both levels of Z_1 (i.e. when $Z_1 = 1$ and when $Z_1 = 0$). The odds ratio of $S = 1$ versus $S = 0$ as a function of Z_2 when $Z_1 = 0$ is

$$\frac{P(S = 1|Y, X, Z_1 = 0, Z_2 = 1)/P(S = 0|Y, X, Z_1 = 0, Z_2 = 1)}{P(S = 1|Y, X, Z_1 = 0, Z_2 = 0)/P(S = 0|Y, X, Z_1 = 0, Z_2 = 0)}, \tag{4.6}$$

but

$$P(S = 1|Y, X, Z_1, Z_2 = 1) = \frac{P(Y|X, S = 1, Z_1)P(S = 1|Z_1, Z_2 = 1)}{P(Y|X, Z_1)}$$

so the odds in the numerator is

$$\frac{P(S = 1|Y, X, Z_1, Z_2 = 1)}{P(S = 0|Y, X, Z_1, Z_2 = 1)} = \frac{P(Y|X, S = 1, Z_1)P(S = 1|Z_1, Z_2 = 1)}{P(Y|X, S = 0, Z_1)P(S = 0|Z_1, Z_2 = 1)}.$$

As a result (4.6) is equal to

$$\frac{P(S = 1|Z_1 = 0, Z_2 = 1)/P(S = 0|Z_1 = 0, Z_2 = 1)}{P(S = 1|Z_1 = 0, Z_2 = 0)/P(S = 0|Z_1 = 0, Z_2 = 0)}$$

which does not depend on Z_1 since there is no Z_1Z_2 interaction term in the model for S.

To show that a YX interaction term is needed in the imputation model S, we write the odds ratio of S comparing $Y = 1$ to $Y = 0$ in terms of the true model (4.4) for two settings: (i) when $X = 0$ and (ii) when $X = 1$. When $X = 0$,

$$\begin{aligned} & \frac{P(S = 1|Y = 1, X = 0, Z_1, Z_2)/P(S = 0|Y = 1, X = 0, Z_1, Z_2)}{P(S = 1|Y = 0, X = 0, Z_1, Z_2)/P(S = 0|Y = 0, X = 0, Z_1, Z_2)} \\ = & \frac{P(Y = 1|X = 0, S = 1, Z_1)/P(Y = 1|X = 0, S = 0, Z_1)}{P(Y = 0|X = 0, S = 1, Z_1)/P(Y = 0|X = 0, S = 0, Z_1)} \end{aligned}$$

This ratio of odds is different when $X = 1$ because there is an XS interaction term in the conditional probability for Y: $P(Y = 1|X, S, Z_1) = \vartheta_0 + \vartheta_1 X + \vartheta_2 S + \vartheta_3 XS + \vartheta_4 XZ_1$. A similar method can be used to show that the following two-way interaction terms are

required in the logistic regression model for imputing S: YX, YZ_1, and XZ_1. Therefore an imputation model that adequately describes the relationship between S and (Y, X, Z_1, Z_2) is one for which

$$\mathbf{A}_i = (1, Y_i, X_i, Z_{1i}, Z_{2i}, Y_iX_i, Y_iZ_{1i}, X_iZ_{1i}) \tag{4.7}$$

in equation (4.5). In practice, the true conditional distribution of S (equation (4.4)) is unknown, and investigators must make assumptions about the true distribution in order to select an appropriate imputation model.

From imputation model (4.5) with $\mathbf{A}_i$ given by (4.7), we obtain the maximum likelihood estimate $\hat{\boldsymbol{\xi}}_3$ with covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\xi}_3)$. The estimated conditional probability of $S_i = 1$ given $\mathbf{A}_i$ is

$$\pi_i(\hat{\boldsymbol{\xi}}_3) = \text{expit}(\mathbf{A}_i\hat{\boldsymbol{\xi}}_3) \ .$$

By $\pi(\hat{\boldsymbol{\xi}}_3)$, we mean the estimated probability using the estimated parameters, and we will use this convention throughout the chapter. Having fitted the imputation model, the next step is to draw from the (estimated) posterior distribution $N(\hat{\boldsymbol{\xi}}_3, \boldsymbol{\Sigma}(\hat{\boldsymbol{\xi}}_3))$ (Carpenter and Kenward, 2013); for the first imputation we let $\tilde{\boldsymbol{\xi}}_3^1$ denote the drawn sample. For each subject i with $R_i = 0$ we then draw S_i from the Bernoulli distribution with probability

$$\pi_i(\tilde{\boldsymbol{\xi}}_3^1) = \text{expit}(\mathbf{A}_i\tilde{\boldsymbol{\xi}}_3^1) \ ,$$

and let S_i^1 denote the realization for $i \in \mathcal{R}^c$. This process is repeated K times, starting with the draw from the posterior distribution to get a new $\tilde{\boldsymbol{\xi}}_3^k$ at the kth sample, to form K independent imputed datasets. With the use of the sampled values for S_i, for N_{mis} individuals with $i \in \mathcal{R}^c$, each of the K imputed datasets is 'complete'.

In the following section, we discuss methods to estimate the conditional causal parameters using the multiply imputed datasets.

4.1.3 Estimation of Response Model Coefficients

In this section, we discuss estimation of the parameters in the conditional response model ($\boldsymbol{\vartheta}$) using multiply imputed data in a randomized study setting. Once the imputation

is complete and we have K complete datasets to work with, we fit a logistic regression model to each imputed dataset separately. Let S_i^k denote the imputed value of S from the kth imputation for subjects with $R_i = 0$; for subjects with $R_i = 1$, $S_i^k = S_i$. For the kth imputed dataset, conditional model parameters $\boldsymbol{\vartheta}$ can be estimated by solving the following standard estimating equation

$$\begin{aligned} \mathbf{U}^k(\boldsymbol{\vartheta}) &= \sum_{i=1}^{n} \mathbf{U}_i^k(\boldsymbol{\vartheta}) \\ &= \sum_{i=1}^{n} \mathbf{D}_i^k(\boldsymbol{\vartheta})[V_i^k(\boldsymbol{\vartheta})]^{-1}(Y_i - \mu_i^k(\boldsymbol{\vartheta})) \ , \end{aligned}$$

where $\mu_i^k(\boldsymbol{\vartheta}) = P(Y_i = 1|X_i, S_i^k, Z_{1i}; \boldsymbol{\vartheta})$, $\mathbf{D}_i^k(\boldsymbol{\vartheta}) = \partial\mu_i^k(\boldsymbol{\vartheta})/\partial\boldsymbol{\vartheta}$ and $V_i^k(\boldsymbol{\vartheta}) = \mu_i^k(\boldsymbol{\vartheta})[1 - \mu_i^k(\boldsymbol{\vartheta})]$. To obtain estimator $\hat{\boldsymbol{\vartheta}}^k$ for the kth dataset, solve $\mathbf{U}^k(\boldsymbol{\vartheta}) = \mathbf{0}$. Let $\boldsymbol{\Sigma}^k(\hat{\boldsymbol{\vartheta}}^k)$ be the corresponding standard estimated covariance matrix for $\hat{\boldsymbol{\vartheta}}^k$.

To obtain the overall estimate of $\boldsymbol{\vartheta}$, we take the average of the K estimates, so

$$\hat{\boldsymbol{\vartheta}}^{\mathrm{MI}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\boldsymbol{\vartheta}}^k \ .$$

An estimate of the covariance matrix for $\hat{\boldsymbol{\vartheta}}^{\mathrm{MI}}$ is then given by

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\vartheta}}^{\mathrm{MI}}) = \mathbf{W}(\hat{\boldsymbol{\vartheta}}^{\mathrm{MI}}) + \left(1 + \frac{1}{K}\right)\mathbf{B}(\hat{\boldsymbol{\vartheta}}^{\mathrm{MI}}) \ , \tag{4.8}$$

where

$$\mathbf{W}(\hat{\boldsymbol{\vartheta}}^{\mathrm{MI}}) = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\Sigma}^k(\hat{\boldsymbol{\vartheta}}^k) \ ,$$

$$\mathbf{B}(\hat{\boldsymbol{\vartheta}}^{\mathrm{MI}}) = \frac{1}{K-1}\sum_{k=1}^{K} (\hat{\boldsymbol{\vartheta}}^k - \hat{\boldsymbol{\vartheta}}^{\mathrm{MI}})(\hat{\boldsymbol{\vartheta}}^k - \hat{\boldsymbol{\vartheta}}^{\mathrm{MI}})^T \ ,$$

and the $(1 + K^{-1})$ term in (4.8) accounts for the finite number of imputations (Carpenter and Kenward, 2013; Rubin, 1987).

4.1.4 Inference

Let $\hat{\vartheta}_1^{\mathrm{MI}}$ be the second component of $\hat{\boldsymbol{\vartheta}}^{\mathrm{MI}}$ which is the multiple imputation estimate for ϑ_1 (the conditional causal odds ratio of treatment when $S = Z_1 = 0$). Let $\hat{\sigma}^2_{\vartheta_1}$ be the variance estimate for $\hat{\vartheta}_1^{\mathrm{MI}}$. Tests and confidence intervals for ϑ_1 can be based on a Student's t approximation:

$$\frac{(\hat{\vartheta}_1^{\mathrm{MI}} - \vartheta_1)}{\sqrt{\hat{\sigma}^2_{\vartheta_1}}} \sim t_v \ ,$$

with degrees of freedom

$$v = (K-1)\left[1 + \frac{\hat{\vartheta}_1^{\mathrm{MI}}}{(1+K^{-1})B(\hat{\vartheta}_1^{\mathrm{MI}})}\right]^2$$

where $B(\hat{\vartheta}_1^{\mathrm{MI}})$ is the $[2,2]$ entry of matrix $\mathbf{B}(\hat{\boldsymbol{\vartheta}}^{\mathrm{MI}})$ that corresponds to ϑ_1 (Schafer, 1999). In our simulation study, we use a normal approximation to build confidence intervals. A normal approximation is appropriate in the case where, for example, $\vartheta_1 = \log(1.25)$, $K = 5$ and $B(\hat{\vartheta}_1^{\mathrm{MI}}) = 0.1$, so that $v > 30$.

In this section, estimation of the conditional causal odds ratio is of interest in the randomized setting where the subgroup variable is incomplete. A method for imputing the missing subgroup variable is given, along with a method for choosing two-way interaction terms. Then, a way to combine estimates from the imputed datasets, as well as a method to compute an estimate of the variability, is given. A method for making inferences about the estimates, along with confidence intervals, is given. In the following sections, we show that multiple imputation methods for estimation of conditional causal parameters in an observational setting are the same as in the randomized setting. We then present methods for estimation of marginal causal parameters in both the randomized and observational settings.

4.2 Estimation of Conditional Causal Effects in an Observational Setting

In this section, we show that estimation of conditional causal parameters in the observational setting with an incomplete subgroup variable is the same as in the randomized setting when multiple imputation is used.

4.2.1 Notation and Models

We refer to the notation and models introduced in the observational setting in Chapter 2, Section 2.1.2. To recap, W_i denotes the treatment variable for subject i in an observational setting, instead of X_i. We make the distinction between the two variables because their probability distributions are different. In an observational setting, we assume that the treatment variable, W_i, is not independent of auxiliary variables Z_{1i} and Z_{2i}, as treatment selection is often dependent upon subject characteristics. We make the assumption that W_i is independent of subgroup variable S_i conditional on Z_{1i} and Z_{2i}. This assumption is appropriate in settings where the subgroup variable S_i denotes the presence or absence of some genetic factor, which is information that may not be readily available to the treating physician.

Suppose we have the same conditional independence assumptions as in the randomized setting, with the exception of the assumption for the treatment variable:

B.0	$R_i \perp (Y_i, W_i, S_i) \mid (Z_{1i}, Z_{2i})$	
B.1	$Y_i \perp Z_{2i} \mid (W_i, S_i, Z_{1i})$	
B.2	$W_i \perp S_i \mid (Z_{1i}, Z_{2i})$	(observational setting)
B.3	$Z_{1i} \perp Z_{2i}$.	

The assumptions allow us to factor the joint probability of $R_i, Y_i, W_i, S_i, Z_{1i}, Z_{2i}$ for subject i as

$$
\begin{aligned}
& P(R_i, Y_i, W_i, S_i, Z_{1i}, Z_{2i}) \\
= \; & P(R_i|Y_i, W_i, S_i, Z_{1i}, Z_{2i})P(Y_i|W_i, S_i, Z_{1i}, Z_{2i})P(W_i|S_i, Z_{1i}, Z_{2i})P(S_i|Z_{1i}, Z_{2i})P(Z_{1i}, Z_{2i}) \\
= \; & P(R_i|Z_{1i}, Z_{2i})P(Y_i|W_i, S_i, Z_{1i})P(W_i|Z_{1i}, Z_{2i})P(S_i|Z_{1i}, Z_{2i})P(Z_{1i})P(Z_{2i}) \; .
\end{aligned}
$$

As before, in an observational setting, the conditional expectation of response Y_i is denoted by $\mu_i(\boldsymbol{\vartheta}) = E(Y_i|W_i, S_i, Z_{1i}) = P(Y_i = 1|W_i, S_i, Z_{1i}; \boldsymbol{\vartheta})$. A logistic regression model is used for the response whereby

$$\mu_i(\boldsymbol{\vartheta}) \quad = \quad \text{expit}(\vartheta_0 + \vartheta_1 W_i + \vartheta_2 S_i + \vartheta_3 W_i S_i + \vartheta_4 W_i Z_{1i}) \ , \tag{4.9}$$

where $\boldsymbol{\vartheta} = (\vartheta_0, \vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4)^T$ is the same vector of regression coefficients as in (4.2).

When treatment is not randomized, the conditional probability of treatment given (Z_{1i}, Z_{2i}) for subject i is

$$\pi_i(\boldsymbol{\xi}_1) = P(W_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_1) = \text{expit}(\xi_{10} + \xi_{11} Z_{1i} + \xi_{12} Z_{2i}) \ , \tag{4.10}$$

where $\boldsymbol{\xi}_1 = (\xi_{10}, \xi_{11}, \xi_{12})^T$ is a vector of regression coefficients.

The remaining variables, R_i, S_i, Z_{1i} and Z_{2i} are generated as in Section 4.1.1.

4.2.2 Imputation Model and Estimation of Causal Effects

By following the methods given in Section 4.1.2, the imputation model from (4.5) with variables listed in (4.7) is the same in an observational setting. This is because the model for the treatment variable is not a factor in terms of determining the main effects and two-way interaction terms included in the imputation model.

Similarly, the methods described in Section 4.1.3 for estimating $\boldsymbol{\vartheta}$ in the randomized setting using multiply imputed data, are the same for estimating $\boldsymbol{\vartheta}$ in an observational setting. Again, this is because the distribution of the treatment variable does not change the methods given.

4.3 Estimation of Marginal Causal Effects in a Randomized Setting

4.3.1 Notation and Models

Although Z_{1i} is directly causally related to response Y_i through (4.2), we are interested in the marginal causal effect of treatment (X_i) on response (Y_i) for each subgroup defined by values of S_i; that is, we are interested in fitting the model with an interaction between the treatment and the important subgroup variable. Refer to Section 4.1.1 for notation and models in the randomized setting. Estimation of the parameters from the following marginal logistic regression model for the response is of interest:

$$\mu_i(\boldsymbol{\beta}) \;=\; \text{expit}(\beta_0 + \beta_1 X_i + \beta_2 S_i + \beta_3 X_i S_i)\;, \tag{4.11}$$

where $\mu_i(\boldsymbol{\beta}) = E(Y_i|X_i, S_i) = P(Y_i = 1|X_i, S_i; \boldsymbol{\beta})$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ is a vector of regression coefficients. The conditional expectation $\mu_i(\boldsymbol{\beta})$ can be computed by taking the expectation of the true response model over Z_{1i} given X_i and S_i:

$$\begin{aligned} P(Y_i = 1|X_i, S_i; \boldsymbol{\beta}) &= E_{Z_{1i}|X_i,S_i}[P(Y_i = 1|X_i, S_i, Z_{1i}; \boldsymbol{\vartheta})] \\ &= E_{Z_{1i}|S_i}[P(Y_i = 1|X_i, S_i, Z_{1i}; \boldsymbol{\vartheta})] \qquad \text{A.2 (randomized setting)} \\ &= \sum_{z_1=0}^{1} P(Y_i = 1|X_i, S_i, z_1; \boldsymbol{\vartheta}) P(Z_{1i} = z_1|S_i; \boldsymbol{\xi}_2, \zeta_1, \zeta_2) \end{aligned} \tag{4.12}$$

where

$$P(Z_{1i} = z_1|S_i; \boldsymbol{\xi}_2, \zeta_1, \zeta_2) = \frac{\sum_{z_2=0}^{1} P(S_i|z_1, z_2; \boldsymbol{\xi}_2) P(Z_{1i} = z_1; \zeta_1) P(Z_{2i} = z_2; \zeta_2)}{P(S_i; \boldsymbol{\xi}_2, \zeta_1, \zeta_2)}\;.$$

4.3.2 Imputation Model

The imputation model used in the setting of estimating conditional causal parameters $\boldsymbol{\vartheta}$ will be used for estimating $\boldsymbol{\beta}$. Although Z_{1i} is not incorporated into the response model, we use Z_{1i} in the imputation model, because it is independently associated with S. When

the imputation model contains more terms and/or higher order terms than the substantive model (the substantive model being (4.11) in the marginal causal setting), it is said to be a 'richer' model (Carpenter and Kenward, 2013). The imputation model from Section 4.1.2 can be used in the context of the marginal substantive model because it is an imputation method for a richer model (i.e. one containing an extra term, XZ_1). A richer imputation model that follows the imputation model rules listed in Section 4.1.2 but that includes extra terms that are not included in the response model is a valid imputation model (Rubin, 1996). Therefore, we use the imputation model described in Section 4.1.2.

4.3.3 Estimation of Response Model Coefficients

Here, we discuss the use of multiply imputed data to estimate $\boldsymbol{\beta}$ from equation (4.11). As in Section 4.1.3, we obtain an estimate of $\boldsymbol{\beta}$ using a logistic regression model to obtain $\hat{\boldsymbol{\beta}}^k$, $k = 1, ..., K$. A consistent estimator for $\boldsymbol{\beta}$ can be obtained as a solution to the estimating equation for the kth imputed dataset:

$$\mathbf{U}^k(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n} \mathbf{U}_i^k(\boldsymbol{\beta})$$

where

$$\mathbf{U}_i^k(\boldsymbol{\beta}) \;=\; \mathbf{D}_i^k(\boldsymbol{\beta})[V_i^k(\boldsymbol{\beta})]^{-1}(Y_i - \mu_i^k(\boldsymbol{\beta}))\ ,$$

for $k = 1, ..., K$, provided that the imputation model is appropriate. Here, $\mu_i^k(\boldsymbol{\beta}) = P(Y_i = 1|X_i, S_i^k; \boldsymbol{\beta})$, $\mathbf{D}_i^k(\boldsymbol{\beta}) = \partial\mu_i^k(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $V_i^k(\boldsymbol{\beta}) = \text{var}(Y_i|X_i, S_i^k) = \mu_i^k(\boldsymbol{\beta})(1 - \mu_i^k(\boldsymbol{\beta}))$.

The estimate of $\boldsymbol{\beta}$ and its covariance matrix from the kth imputed dataset are denoted by $\hat{\boldsymbol{\beta}}^k$ and $\boldsymbol{\Sigma}^k(\hat{\boldsymbol{\beta}}^k)$ respectively. To obtain the multiple imputation estimate of $\boldsymbol{\beta}$, we take the average of the K estimates:

$$\hat{\boldsymbol{\beta}}^{\text{MI}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\boldsymbol{\beta}}^k\ .$$

An estimate of the covariance matrix for $\hat{\boldsymbol{\beta}}^{\text{MI}}$ is

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}^{\text{MI}}) = \mathbf{W}(\hat{\boldsymbol{\beta}}^{\text{MI}}) + \left(1 + \frac{1}{K}\right)\mathbf{B}(\hat{\boldsymbol{\beta}}^{\text{MI}})\ , \tag{4.13}$$

where

$$\mathbf{W}(\hat{\boldsymbol{\beta}}^{\mathrm{MI}}) = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}^{k}(\hat{\boldsymbol{\beta}}^{k}) \ ,$$

and

$$\mathbf{B}(\hat{\boldsymbol{\beta}}^{\mathrm{MI}}) = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\boldsymbol{\beta}}^{k} - \hat{\boldsymbol{\beta}}^{\mathrm{MI}})(\hat{\boldsymbol{\beta}}^{k} - \hat{\boldsymbol{\beta}}^{\mathrm{MI}})^{T} \ .$$

4.3.4 Inference

Tests and confidence intervals for β_1, the parameter that corresponds to the log odds ratio of the response comparing treated to untreated when $S = 0$, can be based on a Student's t approximation:

$$\frac{(\hat{\beta}_1^{\mathrm{MI}} - \beta_{10})}{\sqrt{\hat{\sigma}^2_{\beta_1}}} \sim t_v \ ,$$

with

$$v = (K-1)\left[1 + \frac{\hat{\beta}_1}{(1+K^{-1})B(\hat{\beta}_1^{\mathrm{MI}})}\right]^2 .$$

degrees of freedom (Schafer, 1999). Here, $\hat{\sigma}^2_{\beta_1}$ is the component of $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}^{\mathrm{MI}})$ from (4.13) that corresponds to $\hat{\beta}_1$, and $B(\hat{\beta}_1^{\mathrm{MI}})$ is the same for $\mathbf{B}(\hat{\boldsymbol{\beta}}^{\mathrm{MI}})$. In our simulation study, we use a normal approximation to build confidence intervals. A normal approximation is appropriate in the case where, for example, $\beta_1 = \log(1.25)$, $K = 5$ and $B(\hat{\boldsymbol{\beta}}_1^{\mathrm{MI}}) = 0.1$, so that $v > 30$.

4.4 Estimation of Marginal Causal Effects in an Observational Setting

Refer to Section 4.2.1 for notation and models in an observational setting.

4.4.1 Imputation Model

We are interested in estimating $\boldsymbol{\beta}$ from equation (4.11) using observational data. We base our imputation model on (i) variables that are included in the model that generates the missing variable (Z_1, Z_2); (ii) variables that are included in the substantive (i.e. response) model (Y, W); and (iii) variables that are predictive of the missing data indicator (which are already listed: Z_1 and Z_2). The variables included in the imputation model are the same as in Section 4.1.2, which is a randomized setting with treatment variable X, instead of W. However, we note that the distribution of the treatment variable does not change the imputation model, i.e.

$$P(S_i = j|Y_i, W_i, Z_{1i}, Z_{2i}) \quad = \quad \frac{P(Y_i|W_i, S_i = j, Z_{1i}; \boldsymbol{\vartheta})P(S_i = j|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_2)}{\sum_{s=0}^{1} P(Y_i|W_i, S_i = s, Z_{1i}; \boldsymbol{\vartheta})P(S_i = s|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_2)} \ .$$

where $P(Y_i|W_i, S_i = j, Z_{1i}; \boldsymbol{\vartheta}) = P(Y_i|X_i, S_i = j, Z_{1i}; \boldsymbol{\vartheta})$. Therefore the imputation model derived in Section 4.1.2 in a randomized setting is valid here as well in an observational setting.

4.4.2 Estimation of Response Model Coefficients

In this section, we describe the use of multiply imputed datasets to estimate $\boldsymbol{\beta}$ in an observational setting, where systematic differences exist between treatment groups, and those differences can be explained by variables Z_1 and Z_2. We use an inverse probability weight which is based on the propensity score to account for the differences between the treatment groups. We make the assumption that the true propensity score model is given by the following logistic regression model:

$$\pi(\boldsymbol{\xi}_1) = P(W = 1|Z_1, Z_2; \boldsymbol{\xi}_1) = \text{expit}(\xi_{10} + \xi_{11}z_1 + \xi_{12}z_2) \ . \tag{4.14}$$

The estimate $\hat{\boldsymbol{\xi}}_1$ that is obtained from fitting this model to the full dataset is used to create an estimated weight. The propensity score is estimated using the entire dataset, since treatment variable W and auxiliary variables Z_1 and Z_2 are observed for all subjects.

For each of the K imputed datasets, we estimate $\boldsymbol{\beta}$ using a logistic regression model. A consistent estimator for $\boldsymbol{\beta}$ can be obtained as a solution to the estimating equation for

the kth imputed dataset

$$\widetilde{\mathbf{U}}_1^k(\boldsymbol{\beta};\boldsymbol{\xi}_1) \;=\; \sum_{i=1}^{n} \widetilde{\mathbf{U}}_{1i}^k(\boldsymbol{\beta};\boldsymbol{\xi}_1)$$

where

$$\widetilde{\mathbf{U}}_{1i}^k(\boldsymbol{\beta};\boldsymbol{\xi}_1) \;=\; \sum_{l=0}^{1} \frac{I(W_i=l)}{\pi_i(\boldsymbol{\xi}_1)^l(1-\pi_i(\boldsymbol{\xi}_1))^{1-l}} \mathbf{D}_i^k(\boldsymbol{\beta})[V_i^k(\boldsymbol{\beta})]^{-1}(Y_i-\mu_i^k(\boldsymbol{\beta}))\ ,$$

for $k = 1, ..., K$, provided the imputation model is appropriate and the weight model is properly specified. Here, $\mu_i^k(\boldsymbol{\beta}) = P(Y_i = 1|W_i, S_i^k; \boldsymbol{\beta})$, $\mathbf{D}_i^k(\boldsymbol{\beta}) = \partial\mu_i^k(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $V_i^k(\boldsymbol{\beta}) = \text{var}(Y_i|X_i, S_i^k) = \mu_i^k(\boldsymbol{\beta})(1-\mu_i^k(\boldsymbol{\beta}))$.

In practice, we estimate $\boldsymbol{\xi}_1$ first using the full dataset to obtain $\hat{\boldsymbol{\xi}}_1$, and replace $\pi_i(\boldsymbol{\xi}_1)$ with its estimated counterpart $\pi_i(\hat{\boldsymbol{\xi}}_1)$. The estimate of $\boldsymbol{\beta}$ for the kth imputed dataset is denoted by $\tilde{\boldsymbol{\beta}}^k$, where the tilde denotes the inverse probability weight for confounding. The multiple imputation estimate of $\boldsymbol{\beta}$ is

$$\tilde{\boldsymbol{\beta}}^{\text{MI}} = \frac{1}{K}\sum_{k=1}^{K} \tilde{\boldsymbol{\beta}}^k\ .$$

The covariance matrix of the estimate is computed in a similar way as described in Section 4.3.3. The difference in this weighted method is that we need to incorporate the fact that we have used an (estimated) inverse probability weight. Let

$$\mathbf{U}_2(\boldsymbol{\xi}_1) = \sum_{i=1}^{n} \mathbf{U}_{2i}(\boldsymbol{\xi}_1)$$

denote an unbiased estimating function for $\boldsymbol{\xi}_1$ corresponding to (4.14). Let $\boldsymbol{\omega} = (\boldsymbol{\beta}^T, \boldsymbol{\xi}_1^T)^T$ be the vector containing all parameters, and write

$$\widetilde{\mathbf{U}}^k(\boldsymbol{\omega}) = \sum_{i=1}^{n} \mathbf{U}_i^k(\boldsymbol{\omega}) = \sum_{i=1}^{n} \begin{pmatrix} \widetilde{\mathbf{U}}_{1i}^k(\boldsymbol{\beta};\boldsymbol{\xi}_1) \\ \mathbf{U}_{2i}(\boldsymbol{\xi}_1) \end{pmatrix} = \mathbf{0}\ . \tag{4.15}$$

Let $\widetilde{\boldsymbol{\omega}}^k = ([\tilde{\boldsymbol{\beta}}^k]^T, \hat{\boldsymbol{\xi}}_1^T)^T$ denote the solution to (4.15), and let $\widetilde{\boldsymbol{\omega}} = ([\tilde{\boldsymbol{\beta}}^{\text{MI}}]^T, \hat{\boldsymbol{\xi}}_1^T)^T$. From the results derived in Chapter 2 which are based on theory given by Newey and McFadden

(1994) and Robins et al. (1995), we let

$$\mathcal{I}^k(\boldsymbol{\omega}) = \begin{bmatrix} \mathcal{I}^k_{11}(\boldsymbol{\omega}) & \mathcal{I}^k_{12}(\boldsymbol{\omega}) \\ \mathbf{0} & \mathcal{I}^k_{22}(\boldsymbol{\omega}) \end{bmatrix} = \begin{bmatrix} E(-\partial\widetilde{\mathbf{U}}^k_{1i}(\boldsymbol{\beta};\boldsymbol{\xi}_1)/\partial\boldsymbol{\beta}^T) & E(-\partial\widetilde{\mathbf{U}}^k_{1i}(\boldsymbol{\beta};\boldsymbol{\xi}_1)/\partial\boldsymbol{\xi}_1^T) \\ & E(-\partial\mathbf{U}_{2i}(\boldsymbol{\xi}_1)/\partial\boldsymbol{\xi}_1^T) \end{bmatrix}$$

and

$$\mathbf{C}^k(\boldsymbol{\omega}) = E\left[\widetilde{\mathbf{U}}^k_i(\boldsymbol{\omega})\widetilde{\mathbf{U}}^k_i(\boldsymbol{\omega})^T\right] \ .$$

Then

$$\sqrt{n}(\widetilde{\boldsymbol{\omega}}^k - \boldsymbol{\omega}) \ \xrightarrow{D} \ MVN(\mathbf{0}, \boldsymbol{\Sigma}^k(\boldsymbol{\omega}))$$

where $\boldsymbol{\Sigma}^k(\boldsymbol{\omega}) = [\mathcal{I}^k(\boldsymbol{\omega})]^{-1}\mathbf{C}^k(\boldsymbol{\omega})\left[[\mathcal{I}^k(\boldsymbol{\omega})]^{-1}\right]^T$. We compute variance estimates by replacing the expectations with their empirical counterparts to obtain estimated covariance matrix $\boldsymbol{\Sigma}^k(\widetilde{\boldsymbol{\omega}}^k)$ for the estimates from the kth imputed dataset.

Then,

$$\mathbf{W}(\tilde{\boldsymbol{\beta}}^{\mathrm{MI}}) = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\Sigma}^k(\tilde{\boldsymbol{\beta}}^k) \ ,$$

where $\boldsymbol{\Sigma}^k(\tilde{\boldsymbol{\beta}}^k)$ is the upper block of the $\boldsymbol{\Sigma}^k(\widetilde{\boldsymbol{\omega}}^k)$ matrix which corresponds to the estimates for $\boldsymbol{\beta}$.

Next, an estimate of the covariance matrix for $\tilde{\boldsymbol{\beta}}^{\mathrm{MI}}$ is

$$\boldsymbol{\Sigma}(\tilde{\boldsymbol{\beta}}^{\mathrm{MI}}) = \mathbf{W}(\tilde{\boldsymbol{\beta}}^{\mathrm{MI}}) + \left(1 + \frac{1}{K}\right)\mathbf{B}(\tilde{\boldsymbol{\beta}}^{\mathrm{MI}}) \ ,$$

where

$$\mathbf{B}(\tilde{\boldsymbol{\beta}}^{\mathrm{MI}}) = \frac{1}{K-1}\sum_{k=1}^{K}(\tilde{\boldsymbol{\beta}}^k - \tilde{\boldsymbol{\beta}}^{\mathrm{MI}})(\tilde{\boldsymbol{\beta}}^k - \tilde{\boldsymbol{\beta}}^{\mathrm{MI}})^T \ .$$

4.4.3 Inference

Tests and confidence intervals for β_1, the parameter that corresponds to the marginal causal log odds ratio of the response when $S = 0$, is based on a Student's t approximation:

$$\frac{(\tilde{\beta}_1 - \beta_1)}{\sqrt{\tilde{\sigma}_1^2}} \sim t_v \ ,$$

with

$$v = (K-1)\left[1 + \frac{\tilde{\beta}_1}{(1+K^{-1})B(\tilde{\beta}_1)}\right]^2$$

degrees of freedom (Schafer, 1999). Here, $\tilde{\sigma}_1^2$ and $B(\tilde{\beta}_1)$ are the components of $\mathbf{\Sigma}(\tilde{\boldsymbol{\beta}}^{\mathrm{MI}})$ and $\mathbf{B}(\tilde{\boldsymbol{\beta}}^{\mathrm{MI}})$ that correspond to $\tilde{\beta}_1$, respectively. In our simulation study, we use a normal approximation to build confidence intervals. A normal approximation is appropriate in the case where, for example, $\beta_1 = \log(1.25)$, $K = 5$ and $B(\tilde{\beta}_1) = 0.1$, so that $v > 30$. Methods for inference for $\beta_1 + \beta_3$, the log odds ratio when $S = 1$, are similar.

4.5 Simulation Specifications and Results

Here, we describe simulation studies conducted in order to explore the degree of bias introduced by misspecifying the imputation model. We are interested in assessing the consistency and relative efficiency of our estimates of the conditional causal parameters (ϑ_1 and $\vartheta_1 + \vartheta_3$), and the marginal causal parameters (β_1 and $\beta_1 + \beta_3$). Here, ϑ_1 is the log odds ratio of the response for those with $S = Z_1 = 0$ conditional on all variables that are independently associated with the response, and $\vartheta_1 + \vartheta_3$ is the same for $S = 1$,$Z_1 = 0$. Similarly, β_1 is the marginal log odds ratio of response when $S = 0$; $\beta_1 + \beta_3$ is the log odds ratio when $S = 1$.

The parameters are set so that we are able to examine the effect of different imputation models in an observational setting, where the missingness is ignorable in the conditional causal model where we condition on Z_1, and non-ignorable in the marginal causal model where we omit Z_1.

Recall that the imputation model takes the form

$$g(P(S_i = 1|Y_i, X_i, Z_{1i}, Z_{2i}; \boldsymbol{\xi}_3)) \quad = \quad \mathbf{A}_i\boldsymbol{\xi}_3 \ .$$

Because two-way interaction terms YW, YZ_1, and WZ_1 are identified as important terms to include in the imputation model for S, in simulation, we explore the following variable

combinations for $\mathbf{A}_i$:

$$
\begin{aligned}
\text{Main effects only:} \quad & \mathbf{A}_i = (1, Y_i, W_i, Z_{1i}, Z_{2i}) \\
YW \text{ interaction:} \quad & \mathbf{A}_i = (1, Y_i, W_i, Z_{1i}, Z_{2i}, Y_i W_i) \\
YZ_1 \text{ interaction:} \quad & \mathbf{A}_i = (1, Y_i, W_i, Z_{1i}, Z_{2i}, Y_i Z_{1i}) \\
WZ_1 \text{ interaction:} \quad & \mathbf{A}_i = (1, Y_i, W_i, Z_{1i}, Z_{2i}, W_i Z_{1i}) \\
YW \text{ and } YZ_1 \text{ interaction:} \quad & \mathbf{A}_i = (1, Y_i, W_i, Z_{1i}, Z_{2i}, Y_i W_i, Y_i Z_{1i}) \\
YW \text{ and } WZ_1 \text{ interaction:} \quad & \mathbf{A}_i = (1, Y_i, W_i, Z_{1i}, Z_{2i}, Y_i W_i, W_i Z_{1i}) \\
YZ_1 \text{ and } WZ_1 \text{ interaction:} \quad & \mathbf{A}_i = (1, Y_i, W_i, Z_{1i}, Z_{2i}, Y_i Z_{1i}, W_i Z_{1i}) \\
\text{All two-way interactions:} \quad & \mathbf{A}_i = (1, Y_i, W_i, Z_{1i}, Z_{2i}, Y_i W_i, Y_i Z_{1i}, Y_i Z_{2i}, W_i Z_{1i}, W_i Z_{2i}, Z_{1i} Z_{2i}) \\
\text{All three-way interactions:} \quad & \mathbf{A}_i = (1, Y_i, W_i, Z_{1i}, Z_{2i}, Y_i W_i, Y_i Z_{1i}, Y_i Z_{2i}, W_i Z_{1i}, \\
& \qquad W_i Z_{2i}, Z_{1i} Z_{2i}, Y_i W_i Z_{1i}, Y_i W_i Z_{2i}, W_i Z_{1i} Z_{2i}) \ .
\end{aligned}
$$

The methods described in Section 4.2.2 are used for estimating $\boldsymbol{\vartheta}$ in Tables 4.1 and 4.2. The methods described in Section 4.4.2 are used for estimating $\boldsymbol{\beta}$ in Tables 4.3 and 4.3, where a single inverse probability weight for confounding is used. We also perform complete case analysis without the use of imputation for comparison. In the marginal model without the use of multiple imputation where complete case analysis is used, one inverse probability weight for confounding is used (i.e. only the missingness was ignored). We also present the results from the Chapter 2 simulation study (the doubly weighted estimating equation approach introduced in Section 2.2.2) and the results of the simulation study in Chapter 3 (the weighted EM algorithm introduced in Section 3.4.2) for comparison.

4.5.1 Parameter Settings

See Section 2.4.1 for parameter settings.

We use multiple imputation by fully conditional specification (FCS) which is implemented by the `mice` procedure in `R` (van Buuren and Groothuis-Oudshoorn, 2011); the methods described in Section 4.1.2 are performed in this procedure. The number of imputations in each simulation is $K = 5$. Recently, others have suggested that a larger number of imputations should be used (Carpenter and Kenward, 2013), however we use $K = 5$ for

these simulation studies, and note that a larger number of imputed datasets is practical in an applied setting with one dataset of interest.

Empirical bias (EBias) is defined as the average difference between the estimated log odds ratio and the true log odds ratio. The average asymptotic standard error (ASE) is calculated as the average of the estimated standard errors using methods described in Section 4.4.3. The empirical standard error (ESE) is defined as the standard deviation of the m log odds ratio estimates, where m is the number of simulated datasets. The empirical coverage probability (ECP) is the average of the indicator variables which indicate whether the true log odds ratio is included in the 95% confidence interval. For each simulated dataset, the confidence interval is computed using the methods described in Section 4.4.3.

For each simulation study, the number of subjects per dataset is 2000 and the number of simulated datasets is 5000. The simulated datasets are independent, and the seed value for the first simulated dataset is the same across the simulation studies.

Table 4.1: Empirical biases and standard errors of conditional causal parameters in an observational setting with incomplete subgroup data using multiple imputation, unless otherwise specified. Imputation model interaction terms are listed under "Imputation Model." The number of imputed datasets is $K = 5$ for each simulated dataset.

	ϑ_1 [a]				$\vartheta_1 + \vartheta_3$ [a]			
Imputation Model	EBias	ASE	ESE	ECP	EBias	ASE	ESE	ECP
				$P(R=1)=0.8$				
No imputation [b]	-0.0045	0.1733	0.1736	0.9510	0.0002	0.1655	0.1658	0.9546
Main effects only	0.0192	0.1605	0.1473	0.9640	-0.0185	0.1526	0.1410	0.9654
YW	-0.0007	0.1622	0.1595	0.9554	0.0010	0.1541	0.1553	0.9474
YZ_1	0.0178	0.1606	0.1473	0.9672	-0.0155	0.1528	0.1445	0.9580
WZ_1	0.0194	0.1605	0.1460	0.9680	-0.0171	0.1527	0.1438	0.9606
YW and YZ_1	-0.0036	0.1623	0.1619	0.9506	-0.0018	0.1540	0.1545	0.9522
YW and WZ_1	-0.0039	0.1623	0.1619	0.9508	-0.0018	0.1540	0.1544	0.9516
YZ_1 and WZ_1	0.0149	0.1606	0.1494	0.9624	-0.0183	0.1527	0.1438	0.9632
YW, YZ_1, WZ_1	-0.0049	0.1622	0.1619	0.9506	0.0018	0.1540	0.1535	0.9538
All two-way [c]	-0.0023	0.1623	0.1616	0.9520	0.0023	0.1540	0.1546	0.9470
All three-way [c]	0.0011	0.1624	0.1645	0.9442	0.0008	0.1543	0.1540	0.9508

Abbreviations: EBias Empirical bias, ASE asymptotic standard error, ESE empirical standard error, ECP empirical coverage probability

[a] Conditional log odds ratio for $Z_1 = 0$.

[b] Complete case analysis, where subjects with a missing subgroup variable are omitted.

[c] All two-way/three-way interactions are included in the subgroup variable imputation model.

Table 4.2: Empirical biases and standard errors of conditional causal parameters in an observational setting with incomplete subgroup data using multiple imputation, unless otherwise specified. Imputation model interaction terms are listed under "Imputation Model." The number of imputed datasets is $K = 5$ for each simulated dataset.

	ϑ_1 [a]				$\vartheta_1 + \vartheta_3$ [a]			
Imputation Model	EBias	ASE	ESE	ECP	EBias	ASE	ESE	ECP
				$P(R=1)=0.6$				
No imputation [b]	-0.0048	0.2023	0.2003	0.9524	-0.0006	0.1933	0.1962	0.9492
Main effects only	0.0380	0.1672	0.1441	0.9704	-0.0350	0.1584	0.1419	0.9648
YW	0.0001	0.1760	0.1777	0.9472	-0.0028	0.1656	0.1642	0.9490
YZ_1	0.0370	0.1681	0.1466	0.9690	-0.0352	0.1590	0.1399	0.9704
WZ_1	0.0398	0.1675	0.1439	0.9730	-0.0383	0.1585	0.1377	0.9710
YW and YZ_1	-0.0013	0.1755	0.1754	0.9484	-0.0017	0.1655	0.1645	0.9476
YW and WZ_1	-0.0016	0.1755	0.1758	0.9480	-0.0020	0.1655	0.1648	0.9472
YZ_1 and WZ_1	0.0348	0.1678	0.1463	0.9686	-0.0342	0.1591	0.1407	0.9672
YW, YZ_1, WZ_1	-0.0012	0.1759	0.1758	0.9456	-0.0054	0.1656	0.1687	0.9464
All two-way [c]	-0.0033	0.1756	0.1755	0.9472	-0.0033	0.1655	0.1688	0.9434
All three-way [c]	-0.0003	0.1762	0.1789	0.9444	0.0024	0.1657	0.1659	0.9506

Abbreviations: EBias Empirical bias, ASE asymptotic standard error, ESE empirical standard error, ECP empirical coverage probability

[a] Conditional log odds ratio for $Z_1 = 0$.

[b] Complete case analysis, where subjects with a missing subgroup variable are omitted.

[c] All two-way/three-way interactions are included in the subgroup variable imputation model.

Table 4.3: Empirical biases and standard errors of marginal causal parameters in an observational setting with incomplete subgroup data using multiple imputation, unless otherwise specified. Imputation interaction terms are listed under "Imputation Model." The overall percentage of missing data is 20%. The number of imputed datasets is $K = 5$ for each simulated dataset.

Imputation	β_1				$\beta_1 + \beta_3$			
Model	EBias	ASE	ESE	ECP	EBias	ASE	ESE	ECP
No imputation [a]	0.0050	0.1638	0.1629	0.9508	0.0059	0.1581	0.1570	0.9522
Main effects	0.0185	0.1551	0.1405	0.9668	-0.0195	0.1494	0.1365	0.9636
YW	-0.0018	0.1569	0.1550	0.9516	0.0014	0.1510	0.1495	0.9478
YZ_1	0.0179	0.1552	0.1412	0.9674	-0.0170	0.1495	0.1370	0.9654
WZ_1	0.0178	0.1552	0.1411	0.9660	-0.0170	0.1494	0.1370	0.9652
YW and YZ_1	-0.0009	0.1570	0.1545	0.9502	-0.0006	0.1510	0.1495	0.9504
YW and WZ_1	-0.0009	0.1568	0.1528	0.9536	-0.0007	0.1509	0.1480	0.9552
YZ_1 and WZ_1	0.0189	0.1552	0.1390	0.9696	-0.0191	0.1494	0.1356	0.9676
YW, YZ_1, WZ_1	-0.0016	0.1570	0.1555	0.9524	0.0033	0.1511	0.1516	0.9484
All two-way [b]	-0.0055	0.1570	0.1566	0.9502	-0.0033	0.1510	0.1490	0.9514
All three-way [b]	-0.0020	0.1574	0.1531	0.9532	-0.0001	0.1514	0.1514	0.9472
DWEE [c]	0.0008	0.1644	0.1634	0.9510	0.0021	0.1587	0.1576	0.9516

Abbreviations: EBias Empirical bias, ASE asymptotic standard error, ESE empirical standard error, ECP empirical coverage probability, DWEE doubly weighted estimating equation

[a] Complete case analysis, where subjects with a missing subgroup variable are omitted.

[b] All two-way/three-way interactions are included in the subgroup variable imputation model.

[c] Method proposed in Chapter 2, Section 2.2.2.

Table 4.4: Empirical biases and standard errors of marginal causal parameters in an observational setting with incomplete subgroup data using multiple imputation, unless otherwise specified. Imputation interaction terms are listed under "Imputation Model." The overall percentage of missing data is 40%. The number of imputed datasets is $K = 5$ for each simulated dataset.

Imputation	β_1				$\beta_1 + \beta_3$			
Model	EBias	ASE	ESE	ECP	EBias	ASE	ESE	ECP
No imputation [a]	0.0067	0.1893	0.1882	0.9550	0.0082	0.1828	0.1800	0.9564
Main effects	0.0369	0.1633	0.1344	0.9782	-0.0373	0.1570	0.1292	0.9762
YW	-0.0002	0.1714	0.1679	0.9480	0.0004	0.1642	0.1602	0.9530
YZ_1	0.0393	0.1634	0.1349	0.9748	-0.0365	0.1570	0.1306	0.9764
WZ_1	0.0395	0.1633	0.1349	0.9776	-0.0367	0.1569	0.1306	0.9754
YW and YZ_1	0.0003	0.1720	0.1679	0.9512	-0.0059	0.1647	0.1631	0.9514
YW and WZ_1	-0.0004	0.1714	0.1651	0.9530	-0.0055	0.1641	0.1607	0.9510
YZ_1 and WZ_1	0.0382	0.1632	0.1326	0.9788	-0.0414	0.1568	0.1310	0.9734
YW, YZ_1, WZ_1	0.0015	0.1721	0.1694	0.9532	-0.0005	0.1649	0.1588	0.9556
All two-way [b]	-0.0005	0.1729	0.1719	0.9478	-0.0024	0.1655	0.1668	0.9458
All three-way [b]	0.0012	0.1739	0.1709	0.9520	-0.0001	0.1665	0.1644	0.9496
DWEE [c]	-0.0018	0.1913	0.1901	0.9550	0.0003	0.1845	0.1815	0.9570

Abbreviations: EBias Empirical bias, ASE asymptotic standard error, ESE empirical standard error, ECP empirical coverage probability, DWEE doubly weighted estimating equation

[a] Complete case analysis, where subjects with a missing subgroup variable are omitted.

[b] All two-way/three-way interactions are included in the subgroup variable imputation model.

[c] Method proposed in Chapter 2, Section 2.2.2.

Table 4.5: Simulation study results for estimating marginal causal parameters in an observational setting with an incomplete subgroup variable using Chapter 3 simulation study parameters (see Section 3.5.1). The number of imputed datasets is $K = 5$ for each simulated dataset.

Imputation	β_1				$\beta_1 + \beta_3$			
Model	EBias	ASE	ESE	ECP	EBias	ASE	ESE	ECP
				$P(R=1)=0.8$				
YW, YZ_1, WZ_1	-0.0013	0.1588	0.1570	0.9526	0.0033	0.1494	0.1502	0.9478
All two-way [a]	-0.0058	0.1589	0.1588	0.9494	-0.0032	0.1494	0.1483	0.9500
All three-way [a]	-0.0037	0.1590	0.1570	0.9528	-0.0001	0.1495	0.1500	0.9484
Weighted EM [b]	-0.0008	0.1759	0.1763	0.9470	0.0033	0.1692	0.1714	0.9460
				$P(R=1)=0.6$				
YW, YZ_1, WZ_1	0.0015	0.1746	0.1716	0.9538	-0.0004	0.1631	0.1569	0.9576
All two-way [a]	-0.0011	0.1752	0.1743	0.9478	-0.0028	0.1635	0.1643	0.9468
All three-way [a]	0.0015	0.1759	0.1740	0.9496	-0.0047	0.1640	0.1621	0.9510
Weighted EM [b]	0.0001	0.1746	0.1743	0.9504	0.0020	0.1633	0.1628	0.9508

Abbreviations: EBias Empirical bias, ASE asymptotic standard error, ESE empirical standard error, ECP empirical coverage probability; EM expectation-maximization

[a] All two-way/three-way interactions are included in the subgroup variable imputation model.

[b] Doubly weighted EM-type algorithm method presented in Chapter 3, Section 3.4.2.

4.5.2 Discussion of Simulation Studies

In Tables 4.1 and 4.2, we see that the complete case model for estimating $\boldsymbol{\vartheta}$ is appropriate and results in consistent estimates for ϑ_1 and $\vartheta_1+\vartheta_3$ as expected. However, there is a loss of efficiency in comparison to the multiple imputation estimates. The imputation models with main effects only, and with YZ_1 or WZ_1 as the only interaction term, result in inconsistent estimates of ϑ_1 and $\vartheta_1 + \vartheta_3$. This is interesting, as the complete case analysis results in consistent estimates, but the models with an incorrect imputation model do not. Therefore specifying the correct imputation model is important. We showed in Section 4.1.2 that the two-way interaction terms that are necessary in the imputation model in order to estimate the true relationship between all of the variables are YW, YZ_1 and WZ_1. However, we do see that in an imputation model where YW is the only interaction term, consistent estimates are obtained. This may be due to the relatively weak relationship between Y and Z_1 ($\vartheta_4 = \log(1.25)$). Including all two-way, or even all three-way, interaction terms in the imputation model results in consistent estimates and does not result in a loss of efficiency.

In Tables 4.3 and 4.4, we see that the complete case method which ignores the missing data and includes one weight for confounding results in slightly biased estimates of β_1 and $\beta_1+\beta_3$. The imputation model for S which includes main effects only performs worse than the complete case method with only one weight for confounding in terms of consistency, which is interesting. This highlights the importance of choosing an appropriate imputation model. The imputation model with YW as the only interaction term performs well under the parameter settings. The multiple imputation method is quite a bit more efficient than the doubly weighted estimating equation method presented in Chapter 2.

The results in Table 4.5 are similar to the results in Tables 4.3 and 4.4. In terms of a comparison between the weighted EM-type algorithm method introduced in Chapter 3 and the multiple imputation method, we see that the efficiency is very similar and both methods result in consistent estimates for β_1 and $\beta_1 + \beta_3$ when the percentage of missing data is high (40%). However, the weighted multiple imputation model is more efficient when the percentage of missing data is lower (20%).

4.6 Summary

In this chapter, we explore a method for selecting an imputation model for a missing covariate/subgroup variable in the presence of interaction terms in the response (substantive) model. We review existing methods for imputing missing values to create K 'full' datasets. We also review existing methods for estimating conditional regression parameters using multiply imputed data in the randomized and observational settings. We do the same for estimating marginal regression parameters in randomized and observational settings, where a weight for confounding is included in an observational setting. A method for estimating the variance in the weighted multiple imputation method is given, as well as suggestions for inference and confidence interval calculations.

In simulation studies, we find that correct specification of the imputation model is important in both conditional parameter estimation and marginal parameter estimation. The weighted multiple imputation method is more efficient than the doubly weighted estimating equation approach from Chapter 2. Also, the weighted multiple imputation model is comparable to the doubly weighted EM algorithm method introduced in Chapter 3 in terms of efficiency when the percentage of missing data is high, and the weighted multiple imputation model is more efficient when the percentage of missing data is low.

Chapter 5

Incidence of Thrombotic Events in the Treatment of Metastatic Colorectal Cancer

In this chapter we apply the methods presented in Chapters 2, 3 and 4 in a causal analysis of the effect of bevacizumab plus chemotherapy versus chemotherapy alone on the incidence of thrombotic events in a cohort of patients with metastatic colorectal cancer treated between 2006 and 2011.

5.1 Background, Exclusions and Data Summary

Interest here lies in comparing the risk of a thrombotic event (TE) in metastatic colorectal patients who receive either a combination of bevacizumab (BV, Avastin®) and chemotherapy (FOLFIRI), or FOLFIRI alone. The causal link between BV plus FOLFIRI and a TE during the course of therapy is of interest within two subgroups: those who are treated as a second line of treatment, and those who are treated as a first line of treatment. Data are from a cohort study of patients from the Juravinski Cancer Centre in Ontario, Canada between 2004 and 2011. Data were obtained from a registry of patients diagnosed

Table 5.1: Year of start of treatment for study participants treated with BV plus FOLFIRI or FOLFIRI alone between 2004 and 2011.

Year of start of therapy	BV plus FOLFIRI (n = 261)	FOLFIRI (n = 189)
2004	0	1
2005	0	9
2006	3	38
2007	4	47
2008	47	27
2009	79	20
2010	67	28
2011	39	19
Missing	22	0

with metastatic colorectal cancer who were referred to the Juravinski Cancer Centre from facilities within the centre's catchment area, as well as patients who were referred from other cancer centres. In September 2006, BV became a publicly funded drug, therefore we restrict attention to patients who started treatment between 2006 and 2011, since all patients who started treatment before 2006 were given the control therapy. See Table 5.1 for information on the distribution of treatment assignment by year, and Table 5.2 for exclusion criteria. The analyses performed in this chapter are intended as an exploratory application of the methods proposed in Chapters 2, 3 and 4.

Al-Shamsi et al. (2015) give the reasons for not receiving BV in addition to FOLFIRI for the control patients, which include risk of perforation due to a large tumor, post-surgical wound complication, and prior deep vein thrombosis (DVT) and/or pulmonary embolus (PE). A small percentage of patients treated with BV in addition to FOLFIRI have a history of DVT/PE. In 34% of control patients, there is no documented reason for not receiving BV in addition to FOLFIRI.

See Table 5.3 for a summary of baseline characterstics for study participants. Overall we see that BV plus FOLFIRI patients are healthier. The median age is lower in the BV

Table 5.2: Exclusions for patients treated with BV plus FOLFIRI or FOLFIRI alone.

Exclusion criteria	BV plus FOLFIRI (n = 261)	FOLFIRI (n = 189)
Missing date of start of therapy, n	22	0
Year of start of therapy < 2006, n	0	10
Patients included in the analysis, n	239	179

plus FOLFIRI group (61 vs. 64 years). BV plus FOLFIRI patients are less likely to have hypertension (33.9% vs. 41.9%) or diabetes (8.8% vs. 22.3%). BV plus FOLFIRI patients are less likely to have rectal/rectosigmoid cancer (29.3% vs. 43.6%) and more likely to have colon cancer (69.9% vs. 56.4%).

In Table 5.4 we summarize the occurrence of a TE during the treatment period for study participants. A patient is considered as having a TE if the event occurs during the course of treatment, or within 30 days after discontinuation. For patients who did not have a TE, the follow-up time is the difference between the start of therapy, and the end of therapy, plus 30 days. The incidence of a TE is higher in the BV plus FOLFIRI group compared to the group receiving FOLFIRI alone (16.3% vs. 12.3%). In Al-Shamsi et al. (2015), where control patients who began treatment in 2004 and 2005 (before BV was introduced) were included, the incidence of TE was lower in the BV plus FOLFIRI group (14.9% vs. 15.9%). One patient in the FOLFIRI without BV treatment group experienced a TE at the start of treatment (day 0).

5.2 Variable Selection and Notation

We define the following variables to reflect the set-up of Chapters 2 through 4. The binary response variable is denoted by Y, where $Y_i = 1$ if patient i experiences a TE during the course of treatment or during the 30 days after treatment is discontinued, and $Y_i = 0$ otherwise. The treatment variable is denoted by W in this observational setting, where

Table 5.3: Baseline characteristics for 239 colorectal cancer patients treated with BV plus FOLFIRI, and 179 colorectal cancer patients treated with FOLFIRI alone.

Baseline characteristic	BV plus FOLFIRI (n = 239)	FOLFIRI (n = 179)
Male, n (%) [a]	149/234 (63.7%)	112/179 (62.6%)
Age (years), median (range) [a]	61 (24, 83)	64 (37, 92)
BMI (kg/m^2), median (range)	26 (17, 52)	27 (18, 46)
Diabetes, n (%)	21 (8.8%)	40 (22.3%)
Hypertension, n (%)	81 (33.9%)	75 (41.9%)
Smoking, n (%)	27 (11.3%)	24 (13.4%)
Type of cancer, n (%)		
Colon	167 (69.9%)	101 (56.4%)
Rectal or Rectosigmoid	70 (29.3%)	78 (43.6%)
Site of metasteses, n (%)		
Liver	119 (49.8%)	134 (74.9%)
Lung	63 (26.4%)	89 (49.7%)
Second line of treatment [b]	17/221 (7.7%)	114/178 (64.0%)
History of DVT/PE	8 (3.3%)	31 (17.3%)
History of MI	26 (10.9%)	24 (13.4%)

Abbreviatons: BV bevacizumab; BMI body mass index; DVT deep vein thrombosis; PE pulmonary embolus; MI myocardial infarction

[a] Sex is missing in 5 patients and age is missing in 1 patient (all BV plus FOLFIRI).

[b] Line of treatment is missing in 19 patients, 18 BV plus FOLFIRI, 1 FOLFIRI.

$W_i = 1$ if patient i receives BV plus FOLFIRI, and $W_i = 0$ otherwise. The subgroup variable denoted by S indicates line of treatment, where $S_i = 1$ if patient i receives second line of treatment, and $S_i = 0$ if patient i receives first line of treatment. The subgroup variable of interest is incomplete for 5% of patients included in the study (Table 5.3). The possible confounder, denoted by Z_{1i}, indicates whether patient i has a history of DVT/PE; we consider this as a confounder because it is both a risk factor for a TE and is associated with treatment selection in this patient sample (see Table 5.3). The auxiliary variable Z_{2i}

Table 5.4: Frequency and location of thrombotic events by treatment group (BV plus FOLFIRI and FOLFIRI).

	BV plus FOLFIRI (n = 239)	FOLFIRI (n = 179)
Thrombotic event [a], n (%)	39 (16.3%)	22 (12.3%)
Thrombotic location [b], n (%)		
Venous upper limb	5 (12.8%)	8 (36.4%)
Venous lower limb	8 (20.5%)	4 (18.2%)
Pulmonary embolus	20 (51.3%)	9 (40.9%)
Other	6 (15.4%)	1 (4.5%)
Diagnosed as [b], n (%)		
Symptomatic	25 (64.1%)	14 (63.6%)
Incidental	14 (35.9%)	8 (36.4%)
Time (days), median (range)		
Treatment start to stop, all patients	211 (6, 1261)	170 (0, 1398)
Treatment start to thrombotic event	96 (6, 893)	110 (0, 289)

[a] Thrombotic event is defined as a thrombotic event that occurs during the course of treatment or within 30 days of discontinuation.

[b] The denominator for the percentages is the total number of thrombotic events for each treatment group.

indicates whether patient i has malignant colon cancer and if $Z_{2i} = 0$, patient i has a malignant rectal/rectosigmoid cancer. Note that out of 268 patients with malignant colon cancer, one patient has rectal/rectosigmoid cancer as well, and for this patient, we set $Z_{2i} = 1$. We let R_i denote the missing data indicator where $R_i = 1$ if we have observed the subgroup variable for patient i, and 0 otherwise.

Figure 5.1 contains a causal DAG which summarizes the conceptualized causal relationships between the variables. We make the following assumptions, in alignment with

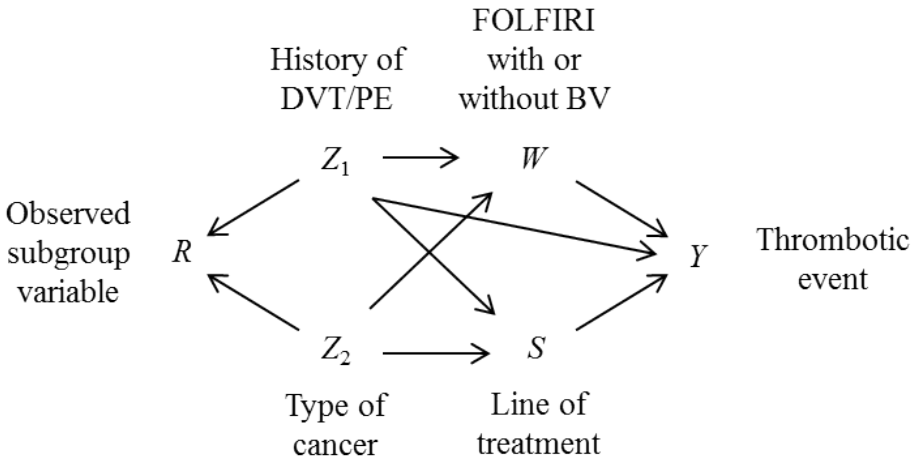

Figure 5.1: Simple causal DAG for treatment (BV plus FOLFIRI) versus control (FOLFIRI), denoted by W. The binary response variable Y denotes whether a patient has a TE during the course of treatment or within 30 days of discontinuation. Potential confounder Z_1 denotes whether a patient has a history of DVT/PE, and auxiliary variable Z_2 denotes the indicator variable for malignant colon cancer. The important, and possibly incomplete, subgroup variable is denoted by S which represents the line of treatment (second or first). Variable R indicates whether S is observed.

those of Chapters 2 and 4.

B.0	$R_i \perp (Y_i, W_i, S_i) \mid (Z_{1i}, Z_{2i})$	
B.1	$Y_i \perp Z_{2i} \mid (W_i, S_i, Z_{1i})$	
B.2	$W_i \perp S_i \mid (Z_{1i}, Z_{2i})$	(observational setting)
B.3	$Z_{1i} \perp Z_{2i}$.	

Note that we make the assumption that S and W are conditionally independent, and we explore whether this assumption is appropriate given the data. In Chapter 3, we further assume that Z_1 and S are conditionally independent and we explore whether this assumption is appropriate as well.

To investigate whether the potential confounding variable Z_1 and auxiliary variable Z_2 are appropriately chosen based on the study data, we model the propensity score for

treatment selection, see Table 5.5. An analysis of variables that may be associated with treatment selection are included in the full model (Model T1) which contains all potential confounding and auxiliary variables as well as the subgroup variable, and a reduced model (Model T2) where only Z_1 (history of DVT/PE) and Z_2 (malignant colon cancer) are included. We see that Z_1 and Z_2 are both significantly associated with treatment selection, and that other variables are also significantly associated with treatment selection, including diabetes, liver metasteses and the subgroup variable. The assumption that S and W are conditionally independent is not appropriate, however we restrict the propensity score model to two important covariates to reflect the models proposed in our derivation of methods in Chapters 2, 3 and 4. (See Appendix B for results of a simulation study investigating the effect of a subgroup variable that is also a confounding variable, where the subgroup variable is omitted from the propensity score model.) The propensity score from Model T2 is estimated by fitting the following logistic regression model

$$\pi_i(\boldsymbol{\xi}_1) = P(W_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_1) = \text{expit}(\xi_{10} + \xi_{11}Z_{1i} + \xi_{12}Z_{2i}) \ .$$

To investigate whether variables Z_1 (history of DVT and/or PE) and Z_2 (malignant colon cancer) are associated with subgroup variable S (line of treatment), we fit a logistic regression model with S as the binary outcome variable and Z_1 and Z_2 as the independent variables (Model S2, Table 5.6). Model S2 is the following logistic regression model:

$$\pi_i(\boldsymbol{\xi}_2) = P(S_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_2) = \text{expit}(\xi_{20} + \xi_{21}Z_{1i} + \xi_{22}Z_{2i}) \ .$$

We also fit a larger logistic regression model to identify any other variables measured before the start of therapy that are associated with S (Model S1, Table 5.6). Variables Z_1 and Z_2 are both significantly associated with S, therefore the assumption that we make in Chapter 3 which is that Z_1 and S are independent, is not appropriate. Other variables that are associated with line of therapy are gender, age, diabetes and liver metasteses.

To investigate whether variables Z_1 and Z_2 are associated with R (the indicator for whether S is observed), we fit a logistic regression model with R as the binary outcome variable and Z_1 and Z_2 as the covariates (Model R2, Table 5.7). We also fit a larger logistic regression model to identify any other variables measured before the start of therapy that

Table 5.5: Estimates from full (Model T1) and reduced (Model T2) logistic regression models for the propensity score. All independent variables listed are included in Model T1, and only Z_1 and Z_2 are included in Model T2.

		Model T1		Model T2	
Covariate	Comparison	Est. (SE)	p	Est. (SE)	p
History of DVT/PE (Z_1)	Yes vs. no	-1.66 (0.52)	<0.01	-1.85 (0.41)	< 0.01
Colon cancer (Z_2) [a]	Yes vs. no	0.72 (0.29)	0.01	0.64 (0.21)	< 0.01
Gender	F vs. M	-0.004 (0.29)	0.99		
Age (years)		-0.03 (0.01)	0.07		
Body mass index (kg/m^2)		0.03 (0.03)	0.30		
Diabetes	Yes vs. no	-0.91 (0.39)	0.02		
Hypertension	Yes vs. no	-0.07 (0.31)	0.81		
Smoking	Yes vs. no	-0.21 (0.41)	0.61		
Liver metasteses	Yes vs. no	-1.05 (0.29)	<0.01		
History of MI	Yes vs. no	-0.32 (0.41)	0.44		
Line of treatment (S)	2$^{\text{nd}}$ vs. 1$^{\text{st}}$	-2.94 (0.32)	<0.01		

Abbreviatons: Est. log odds ratio estimate; SE; standard error; DVT deep vein thrombosis; PE pulmonary embolus; MI myocardial infarction

[a] Comparing patients with malignant colon cancer to those with malignant rectal cancer.

are associated with R (Model R1, Table 5.7). Model R2 is the following logistic regression model:

$$\pi_i(\boldsymbol{\rho}) = P(R_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\rho}) = \text{expit}(\rho_0 + \rho_1 Z_{1i} + \rho_2 Z_{2i}) \ .$$

From Table 5.7, we see that Z_1 is not significantly associated with the missing data process, however Z_2 is significantly associated with R at the 0.10 significance level (p=0.07).

Tables 5.8 and 5.9 are 2×2 tables displaying the number of TEs by treatment group (BV plus FOLFIRI vs. FOLFIRI) for each subgroup (second or first line of treatment). For those whose treatment is the second line, the probability of a TE is 0.18 versus 0.11 comparing BV plus FOLFIRI to FOLFIRI alone. For those whose treatment is the first line of treatment, the probability of a TE is the same in each treatment group (0.16).

Table 5.6: Results of full and reduced logistic regression models to investigate the relationship between the subgroup variable and other baseline variables. In Model S1, all variables listed are covariates; in Model S2, only Z_1 and Z_2 are included as covariates.

		Model S1		Model S2	
Covariates	Comparison	Est. (SE)	p	Est. (SE)	p
History of DVT/PE (Z_1)	Yes vs. no	1.08 (0.36)	< 0.01	1.19 (0.35)	< 0.01
Colon cancer (Z_2) [a]		-0.68 (0.24)	< 0.01	-0.50 (0.22)	0.03
Gender	F vs. M	0.53 (0.24)	0.03		
Age (years)		0.04 (0.01)	< 0.01		
Body mass index (kg/m^2)		0.02 (0.02)	0.46		
Diabetes	Yes vs. no	0.70 (0.31)	0.03		
Hypertension	Yes vs. no	-0.09 (0.26)	0.74		
Smoking	Yes vs. no	0.53 (0.35)	0.13		
Liver metasteses	Yes vs. no	0.94 (0.25)	< 0.01		
History of MI	Yes vs. no	-0.35 (0.37)	0.34		

Abbreviatons: Est. log odds ratio estimate; SE standard error; DVT deep vein thrombosis; PE pulmonary embolus; MI myocardial infarction

[a] Comparing patients with malignant colon cancer to those with malignant rectal cancer.

Table 5.7: Results of full and reduced logistic regression models to investigate the relationship between R (indicator of missing data) and baseline variables. In Model R1, all variables listed are covariates; in Model R2, only Z_1 and Z_2 are included as covariates.

		Model R1		Model R2	
Covariates	Comparison	Est. (SE)	p	Est. (SE)	p
History of DVT/PE (Z_1)	Yes vs. no	0.41 (1.08)	0.70	0.67 (1.04)	0.52
Colon cancer (Z_2) [a]		-1.64 (0.77)	0.03	-1.14 (0.64)	0.07
Gender	F vs. M	0.11 (0.52)	0.83		
Age (years)		0.003 (0.02)	0.90		
Body mass index (kg/m^2)		-0.05 (0.04)	0.29		
Diabetes	Yes vs. no	1.01 (1.09)	0.35		
Hypertension	Yes vs. no	0.08 (0.56)	0.89		
Smoking	Yes vs. no	0.15 (0.80)	0.85		
Liver metasteses	Yes vs. no	0.93 (0.51)	0.07		
History of MI	Yes vs. no	0.84 (1.07)	0.44		

Abbreviatons: Est. log odds ratio estimate; SE standard error; DVT deep vein thrombosis; PE pulmonary embolus; MI myocardial infarction

[a] Comparing patients with malignant colon cancer to those with malignant rectal cancer.

Table 5.8: 2×2 table comparing the percentage of patients with a TE in second line of treatment patients ($S_i = 1$).

	Thrombotic event	
Second line of treatment	No	Yes
BV plus FOLFIRI	14 (82%)	3 (18%)
FOLFIRI	102 (89%)	12 (11%)

Table 5.9: 2×2 table comparing the percentage of patients with a TE for first line of treatment patients ($S_i = 0$).

	Thrombotic event	
First line of treatment	No	Yes
BV plus FOLFIRI	171 (84%)	33 (16%)
FOLFIRI	54 (84%)	10 (16%)

5.3 Conditional Response Model

In Table 5.10, we explore the relationship between the response, TE during the course of therapy or within 30 days after discontinuation, and variables measured at baseline in separate logistic regression models each with one independent variable. In unadjusted analyses, treatment (BV plus FOLFIRI vs. FOLFIRI alone) is not significantly associated with TE during therapy, however the odds ratio is 1.39 (p=0.25), indicating that the risk of a TE may be higher in the BV plus FOLFIRI group.

We also report the results of an adjusted logistic regression analysis in Table 5.11 which is similar to the conditional response model fitted in Al-Shamsi et al. (2015); here, we use data from the cohort of patients who started therapy between 2006 and 2011. In this table, we see that the odds of a TE is higher for the BV plus FOLFIRI group, whereas in the original analysis of patients who started therapy between 2004 and 2011, the odds are lower for the BV plus FOLFIRI group. In both analyses, the treatment effect is not statistically significantly different from zero.

Interest lies in estimating the marginal causal effect of BV plus FOLFIRI on the risk of a TE during the course of therapy, for two subgroups: (i) first line of therapy and (ii) second line of therapy. However, we begin our exploration of the response model by fitting the full conditional response model using complete cases in the following model

$$P(Y = 1|W, S, Z_1) = \text{expit}(\vartheta_0 + \vartheta_1 W + \vartheta_2 S + \vartheta_3 WS + \vartheta_4 Z_1) \ , \tag{5.1}$$

where the variables are defined in Section 5.2. Note that in our proposed conditional response model (e.g. equation (2.7)), we have assumed that the main effect of Z_1 on the response is null, however we have relaxed that assumption here and we are including a main effect for Z_1 and omitting the interaction effect with the treatment variable because we do not wish to estimate the treatment effect for different levels of Z_1.

In an ignorable missingness setting, where all variables that are associated with the response are included in the response model, the full conditional response model results in consistent estimates of the conditional causal parameters. This may not be a reasonable assumption, however, since we have restricted attention to only three independent variables: treatment W (BV plus FOLFIRI versus FOLFIRI alone), subgroup variable S (line

Table 5.10: Results of logistic regression models to investigate the relationship between a TE, treatment and baseline covariates. In each regression analysis, one covariate is included as an independent variable and TE during the course of treatment or within 30 days of discontinuation is the dependent variable.

Covariate	Comparison	OR (95% CI)	p
Treatment (W)	BV + FOLFIRI vs. FOLFIRI	1.39 (0.79, 2.44)	0.25
Line of treatment (S) [a]	2^{nd} vs. 1^{st}	0.68 (0.36, 1.27)	0.22
History of DVT/PE (Z_1)	Yes vs. no	0.65 (0.22, 1.89)	0.42
Colon cancer (Z_2) [b]		1.08 (0.61, 1.91)	0.80
Gender	F vs. M	0.70 (0.39, 1.27)	0.24
Age (years)		0.99 (0.97, 1.01)	0.44
Body mass index (kg/m^2)		1.08 (1.04, 1.14)	<0.01
Diabetes	Yes vs. no	0.87 (0.39, 1.92)	0.72
Hypertension	Yes vs. no	1.52 (0.88, 2.63)	0.14
Smoking	Yes vs. no	0.46 (0.16, 1.33)	0.15
Liver metasteses	Yes vs. no	1.09 (0.62, 1.91)	0.76
Past history of MI	Yes vs. no	0.78 (0.32, 1.91)	0.58

Abbreviatons: OR odds ratio; DVT deep vein thrombosis;
PE pulmonary embolus; MI myocardial infarction

[a] Line of treatment is missing in 5% of patients.

[b] Comparing patients with colon cancer to rectal cancer.

of treatment) and potential confounding variable Z_1 (history of DVT/PE). The estimated odds ratios for both levels of the subgroup variable in a complete case conditional response model analysis are given in Table 5.12. We see that the treatment effect is not statistically significant for either of the levels of the subgroup variable in this setting, and the test for interaction is not significant (p-value=0.51).

Table 5.11: Results of a conditional logistic regression model to investigate the relationship between the response, treatment and baseline covariates. This response model is similar to the one in Al-Shamsi et al. (2015), but restricted to patients who started therapy between 2006 and 2011; all covariates listed are included in the adjusted model.

Covariates	Comparison	OR (95% CI)	p
Age (years)		0.99 (0.96, 1.02)	0.41
Body mass index (kg/m^2)		1.08 (1.03, 1.14)	< 0.01
Gender	F vs. M	0.78 (0.42, 1.44)	0.42
Colon cancer (Z_2) [a]		1.16 (0.63, 2.11)	0.63
Metastatic disease	Yes vs. no	1.75 (0.86, 3.56)	0.13
Risk factors for VTE	Yes vs. no	0.62 (0.22, 1.74)	0.36
Risk factors for ATE	Yes vs. no	1.02 (0.56, 1.88)	0.94
Treatment (W)	BV + FOLFIRI vs. FOLFIRI	1.39 (0.74, 2.58)	0.30

Abbreviatons: OR odds ratio; CI confidence interval;
VTE venous thromboembolism; ATE arterial thromboembolism
[a] Comparing patients with malignant colon cancer to rectal cancer.

Table 5.12: Results of fitting a complete case logistic regression model to investigate the relationship between the response, treatment, subgroup variable and potential confounder. See equation (5.1) for the logistic regression model, and Section 5.2 for details on variable selection and notation.

Independent variable	OR (95% CI)
Treatment effect, for $S = 0$ ($\exp(\vartheta_1)$)	1.02 (0.47, 2.21)
Treatment effect, for $S = 1$ ($\exp(\vartheta_1 + \vartheta_3)$) [a]	1.74 (0.43, 7.02)

Abbreviatons: OR odds ratio; CI confidence interval
[a] Interaction p-value = 0.51.

5.4 Marginal Causal Effect Estimation

Interest lies in estimating the marginal causal effect of BV plus FOLFIRI on the risk of a TE during treatment for two subgroups: (i) those whose treatment was a second line of treatment and (ii) those whose treatment was a first line of treatment. We use the following logistic regression model to relate the treatment and subgroup variables to the response:

$$P(Y = 1|X, S; \boldsymbol{\beta}) = \text{expit}(\beta_0 + \beta_1 X + \beta_2 S + \beta_3 XS) \ , \tag{5.2}$$

where X denotes the treatment variable in a randomized setting. See Section 5.2 for details on variable selection and definitions of other variables. Since we are working with observational data with an incomplete subgroup variable, we do not assume that the missingness is ignorable when we condition on treatment and the subgroup variable alone; also treatment selection is highly correlated with baseline characteristics that may be associated with the response therefore confounding may be an issue in this study. To account for the non-ignorably missing data and confounding, we use the approaches introduced in Chapters 2, 3 and 4 to estimate the marginal causal effects.

We begin by using the doubly weighted estimating equation approach introduced in Chapter 2 (see equation (2.18)). We compare the doubly weighted estimating equation approach to the following possibly misspecified models: complete case analysis without weights (equation 2.25); complete case analysis with a weight for confounding, with the addition of R_i in the numerator of the weight to ensure that only complete cases are included in the model (equation (2.11)); and complete case analysis with a weight for missingness (equation (2.15)).

In an application of the doubly weighted estimating equation method, the median (range) of the product of the double inverse probability weights is 1.99 (1.18, 7.15). See Table 5.13 for results; we see that in using the doubly weighted estimating equation approach, there may be a subgroup effect, where patients receiving treatment as a second line of therapy may have a higher odds ratio of a TE (1.82, 95% CI 0.45 to 7.39) compared to those receiving treatment as a first line of therapy (1.15, 95% CI 0.50 to 2.61). We note however that the interaction effect is not statistically significant, and the causal effect

Table 5.13: Results of fitting a complete case estimating function with or without weights in an application of the methods proposed in Chapter 2 to estimate marginal causal parameters in a logistic regression setting. See Section 5.2 for notation and equation (5.2) for the model of interest.

Method	OR (95% CI [a])	OR (95% CI [b])
Complete case unweighted method		
Treatment effect, for $S = 0$	1.04 (0.48, 2.25)	1.04 (0.49, 2.64)
Treatment effect, for $S = 1$	1.82 (0.46, 7.26)	1.82 (0.45, 6.67)
Complete case, IPW for confounding		
Treatment effect, for $S = 0$	1.14 (0.44, 2.98)	1.14 (0.52, 3.07)
Treatment effect, for $S = 1$	1.89 (0.39, 9.09)	1.89 (0.46, 6.91)
Complete case, IPW for missingness		
Treatment effect, for $S = 0$	1.05 (0.49, 2.27)	1.05 (0.50, 2.67)
Treatment effect, for $S = 1$	1.75 (0.44, 6.97)	1.75 (0.43, 6.40)
Complete case, doubly weighted estimating equation [c]		
Treatment effect, for $S = 0$	1.15 (0.50, 2.61)	1.15 (0.52, 3.08)
Treatment effect, for $S = 1$	1.82 (0.45, 7.39)	1.82 (0.44, 6.61)

Abbreviatons: OR odds ratio; CI confidence interval; IPW inverse probability weight

[a] Confidence interval derived using methods proposed in Section 2.3.2.

[b] The 2.5^{th} and 97.5^{th} percentiles from 5000 bootstrap samples.

[c] Interaction p-value = 0.41.

is not statistically significant in either subgroup. We also report the confidence intervals (CIs) obtained from bootstrapping and we find that the bootstrap CIs are comparable to those obtained using the standard error derivation in Section 2.3.2. In the estimation of bootstrap CIs, for approximately 5% of the bootstrap samples, there is not enough data to obtain an estimate of the interaction effect between the treatment and subgroup variables. We omit these bootstrap sample estimates to obtain the 2.5^{th} and 97.5^{th} percentiles of the estimates.

See Table 5.14 for results of an EM-type approach with double inverse probability

Table 5.14: Results of applying the doubly weighted EM-type algorithm introduced in Chapter 3 to the BV plus FOLFIRI data to estimate marginal causal parameters in a logistic regression setting. See Section 5.2 for notation and equation (5.2) for the model of interest.

	OR (95% CI [a])
Doubly weighted EM-type algorithm	
Treatment effect, for $S = 0$ ($\exp(\beta_1)$)	1.32 (0.51, 3.39)
Treatment effect, for $S = 1$ ($\exp(\beta_1 + \beta_3)$)	2.66 (0.10, 68.29)

Abbreviatons: OR odds ratio; CI confidence interval; EM expectation-maximization

[a] Confidence interval calculated using the method proposed in Section 3.4.4.

The interaction p-value is 0.70.

weights for missingness and confounding. The median (range) of the product of the inverse probability weights for confounding and missing data is 1.99 (1.18, 118.0). The effect estimates are similar in comparison to the doubly weighted estimating equation approach, although the confidence intervals are much wider. We note that there is a violation of the assumption that S and Z_1 are independent (see Table 5.6), therefore the results from the application of the weighted EM-type approach are not as useful as the results from the other marginal models.

See Table 5.15 for results of an application of the multiple imputation approach with an inverse probability weight for confounding. Following the imputation model recommendations given in Section 4.1.2, three interaction terms are included in the imputation model for S: YW, YZ_1 and WZ_1, and the number of imputations is $K = 5$. The effect estimates are very similar to those obtained in an application of the doubly weighted estimating equation method.

Table 5.15: Results of applying the weighted multiple imputation method introduced in Chapter 4 to the BV plus FOLFIRI data to estimate marginal causal parameters in a logistic regression setting. See Section 5.2 for notation and equation (5.2) for the model of interest.

	OR (95% CI [a])
Weighted multiple imputation	
Treatment effect, for $S = 0$ (β_1)	1.16 (0.43, 3.12)
Treatment effect, for $S = 1$ ($\beta_1 + \beta_3$)	1.71 (0.36, 8.24)

Abbreviatons: OR odds ratio; CI confidence interval

[a] Confidence interval calculated using the method proposed in Section 4.4.2.

The interaction p-value is 0.67.

5.5 Additional Estimates of Marginal Causal Effects

Here we perform an additional analysis which is not restricted to variables Y, W, S, Z_1, Z_2, and R. In this analysis, all variables that are associated with treatment selection are included in the propensity score model, and all variables that are associated with missing data are included in the missing data model. The marginal response model of interest is unchanged (see equation (5.2)); see Section 5.2 for details on notation. In an application of the method proposed in Chapter 2, a complete case doubly weighted estimating equation approach is used to obtain an estimate of the marginal causal effects, and confidence intervals are estimated using the 2.5^{th} and 97.5^{th} percentiles of 5000 bootstrap samples.

To select the variables included in the propensity score model and the missing data model, we use a p-value cut-off of 0.10. For the treatment selection model, we include the following variables as covariates (see Model T1 in Table 5.5): history of DVT/PE (Z_1), colon cancer (Z_2), age, diabetes, and liver metasteses. Although the subgroup variable is also significantly associated with treatment selection, we omit this variable from the propensity score model. (See Appendix B for a discussion of misspecified propensity score models where the subgroup variable of interest is a confounder, but is omitted from the

propensity score model.) For the missing data model, we include the following variables as covariates (see Model R1 in Table 5.7): colon cancer (Z_2) and liver metasteses.

The estimate of the odds ratio of a TE comparing those treated with BV plus FOLFIRI to FOLFIRI alone, in the first line of treatment group, is 1.07 (95% CI 0.48 to 2.92). In the second line of treatment group, the estimated odds ratio is 1.39 (95% CI 0.27 to 6.59).

5.6 Discussion and Conclusions

In an application of the methods proposed in this thesis, we find that the risk of a TE is the same between treatment groups (BV plus FOLFIRI compared to FOLFIRI alone), and the subgroup effect (line of treatment) is not statistically significant. In an application of the doubly weighted estimating equation approach introduced in Chapter 2, the estimate of the marginal causal odds ratio in the first line of treatment group is 1.15 (95% confidence interval (CI) 0.50 to 2.61), and the odds ratio in the second line of treatment group is 1.82 (95% CI 0.45 to 7.39). We remark that while these estimates are not statistically significantly different, the point estimates convey quite different messages. The estimates are similar when the methods introduced in Chapters 3 and 4 are applied, although the confidence intervals resulting from an application of the EM-type algorithm method are wider in comparison.

We note that this study is likely underpowered to detect a clinically significant subgroup effect. See Appendix A for results of simulation studies designed to estimate the power to detect a significant subgroup effect using the methods proposed in this thesis.

Chapter 6

Discussion, Conclusions and Future Work

6.1 Summary and Discussion

In a clinical setting, investigators often wish to estimate marginal causal parameters that are estimable in randomized controlled trials, using observational data. Missing data and confounding arise frequently in observational studies; in this thesis, we investigate the estimation of marginal causal odds ratios in the observational setting where an important subgroup variable is incomplete. In Chapter 1, we review statistical methods for causal inference, which includes a brief overview of counterfactual notation. We then review methods for incomplete data.

In Chapter 2, we propose a method for estimating marginal causal odds ratios in an observational setting, for each level of a subgroup variable, using a doubly inverse probability weighted estimating equation. The inverse probability weights account for both confounding and missing data, as the subgroup variable is incomplete. Prior to introducing this method, we review existing methods for causal analysis where an inverse probability weight for confounding is incorporated. We then review existing inverse probability weighted methods for adjusting for non-ignorably missing data. In the doubly weighted method,

we show that the asymptotic standard error estimation is relatively straightforward, and involves accounting for the extra variability introduced by estimating the parameters in the inverse probability weights. A method for hypothesis testing is given, and simulation studies show that this method is straightforward to implement and that the estimators are consistent.

In implementing the doubly weighted estimating equation approach, knowledge of the model to generate the incomplete subgroup variable is not necessary; only knowledge about the missing data process and the appropriateness of the no unmeasured confounders assumption is required. The weight models must be properly specified and omission of one or both of the weights can lead to biased estimates. We recommend using a more liberal approach to building a model for the missing data process as well as including as many potential confounding variables as is necessary in the propensity score model. Although variance estimation is relatively straightforward when there are only one or two covariates included in the weight models, bootstrapping to obtain confidence intervals may be useful when the number of covariates in the inverse probability weight models is large.

The doubly weighted estimating equation approach involves complete case analysis, where data from subjects with an incomplete subgroup variable is omitted, which may impact efficiency. In Chapter 3, we review an existing method for handling missing covariate data in conditional causal models, which involves an EM algorithm where all available data is included in the analysis. We extend this method to the setting where marginal causal parameters are of interest and this involves an EM-type algorithm approach with the addition of two inverse probability weights - one for non-ignorably missing data, and one for confounding. To implement this method, one must have knowledge about the subgroup variable model, the missing data process and whether the no unmeasured confounders assumption is appropriate. We derive the asymptotic properties of the estimators, and a method for hypothesis testing is proposed. In simulation studies, we see that the weighted EM-type algorithm can be used to obtain consistent estimates of causal parameters, however, estimation of the variance of the estimators is not straightforward.

In Chapter 4, we review existing methods for multiple imputation in the setting of a missing covariate. We discuss the selection of an appropriate imputation model given the true conditional distribution of the incomplete subgroup variable. We investigate the

use of multiply imputed data with an imputed subgroup variable, in weighted estimating equations with one weight for confounding. Standard methods to combine estimates from multiply imputed datasets are used; the extra variability introduced by estimating the parameters in the model for the weight for confounding is incorporated into standard methods for variance estimation in a multiple imputation setting. Simulation studies show that this method is straightforward to implement, and consistent estimates of the parameters of interest are obtained. However, care must be taken when choosing an imputation model for the missing covariate, and an imputation model that characterizes the model for the missing covariate as adequately as possible is desirable.

In terms of a comparison between the multiple imputation method introduced in Chapter 4 and the doubly inverse probability weighted method introduced in Chapter 2, in simulation studies, the estimators of the multiple imputation method tend to be more efficient. In the multiple imputation setting, knowledge about the incomplete subgroup variable model is required for imputation whereas knowledge about the missing data process is required for the doubly weighted estimating equation approach. In an observational data setting, it may not be feasible to collect data on the variables that are associated with the missing data process, and studies are not often designed to answer such questions, which may make the doubly inverse probability weighted method more difficult to implement in practice. In terms of selecting an imputation model for the incomplete subgroup variable, we see in simulation studies that exact knowledge of the model for imputing the subgroup variable is not necessary, and a more liberal approach to variable inclusion and interaction effect inclusion does not affect consistency or efficiency of the estimates of the marginal causal effects. In both methods, knowledge about the appropriateness of the propensity score model and the no unmeasured confounders assumption is required.

In comparing the weighted EM-type algorithm method introduced in Chapter 3, and the multiple imputation method of Chapter 4, we need to use a more restrictive approach to estimating the conditional distribution of the subgroup variable in the weighted EM-type method. This is because estimation of the subgroup model is tied into the marginal causal model in the EM-type algorithm method, whereas the multiple imputation model used to approximate the true conditional distribution of the subgroup variable need not be restricted by the marginal response model. Existing software to facilitate imputation

makes the weighted multiple imputation method preferable in practice.

In Appendix A, we summarize the results of simulation studies designed to estimate the power to detect a significant subgroup effect under different parameter settings and sample sizes. We find that a large sample size (n $\geq$ 2000) is required have adequate power to detect a relatively large subgroup effect ($\beta_3 > 0.5$). The three methods are comparable in terms of power to detect a subgroup effect.

In Chapter 5, we apply the methods proposed in Chapters 2, 3 and 4 to investigate the causal link between a novel treatment for metastatic colorectal cancer, and thrombotic events. We find that there is no significant difference in the risk of a thrombotic event comparing patients who receive the new treatment, bevacizumab (BV) plus chemotherapy (FOLFIRI), to patients who receive FOLFIRI alone. The subgroup of interest is line of treatment (second or first); the interaction effect is not statistically significant, although the data indicate that perhaps the odds ratio of a thrombotic event is higher in those whose treatment was a second line therapy. Data from 418 patients are included in this study, and we note that the study is likely underpowered to detect a significant and clinically meaningful treatment effect and/or subgroup effect. In the marginal models described in Chapters 2, 3 and 4, we make the assumption that the subgroup variable is not also a confounder. This assumption is likely not appropriate in the application to the BV plus FOLFIRI data; see Appendix B for results of simulation studies designed to estimate the degree of bias introduced from a misspecified propensity score model where an important confounder is omitted.

6.2 Future Work

This section is an outline of future research related to this thesis.

6.2.1 Extension to Other Response Models

In this thesis, we focus on estimation of the marginal causal odds ratio. With a binary response variable, risk differences and risk ratios may be of interest. The methods presented

in this thesis can be explored for estimating other types of summary statistics. We note that log linear models are collapsible, and therefore the marginal relative risk is estimable using a conditional response model, and a weight for confounding and/or missing data may not be necessary, depending on the missing data process (Zou, 2004).

When estimation of the causal hazard ratio is of interest, the methods presented in Chapter 2 may be useful, as the hazard ratio is subject to non-collapsibility and therefore conditioning on confounding variables and covariates that are associated with the missing data process in the response model may not be desirable. Hernán et al. (2000) introduce weighted Cox models for adjustment for time-dependent confounding. These methods can be modified to account for missingness and selection bias introduced when an important subgroup variable is incomplete, and marginal causal effects are of interest.

6.2.2 Stabilized Weights

The inverse probability weights used in Chapters 2, 3 and 4 can be stabilized by adding the product of the marginal probability of treatment and the marginal probability of missing data in the numerator of the weight. In the longitudinal data setting, the stabilized weight for confounding at time t is

$$sw_i(t) = \prod_{k=0}^{t} \frac{f[A(k)|\bar{A}(k-1)]}{f[A(k)|\bar{A}(k-1), \bar{L}(k)]} \tag{6.1}$$

where $A(k)$ is the observed treatment at visit k, $\bar{A}(k-1)$ is the observed treatment history up to and including visit $k-1$, $L(k)$ is a vector of time-varying covariates and $\bar{L}(k)$ contains the covariate history up to and including visit k (Hernán et al., 2002). This can be simplified to the setting with only one follow-up visit, and extended to the setting where a weight for missing data is also incorporated. One may wish to investigate the properties of the following weighted estimating equation with stabilized weights:

$$\widetilde{\overline{\mathbf{U}}}_i(\boldsymbol{\beta};\boldsymbol{\psi}) \;=\; \sum_{l=0}^{1} \frac{R_i P(R_i=1)}{\pi_i(\boldsymbol{\rho})} \frac{I(W_i=l)P(W_i=l)}{\pi_i(\boldsymbol{\xi}_1)^l(1-\pi_i(\boldsymbol{\xi}_1))^{1-l}} \; \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\big(Y_i - \mu_i(\boldsymbol{\beta})\big) \;, \tag{6.2}$$

where the parameters and $\mathbf{D}_i(\boldsymbol{\beta})$ and $V_i(\boldsymbol{\beta})$ are defined in Section 2.2.2.

Because the use of stabilized weights may result in estimators with more desirable properties, it is of interest to investigate the sample size requirements for detecting a significant treatment effect and/or a significant subgroup effect with the use of stabilized weights.

6.2.3 Extension to Missing Response Data

In this thesis, we have discussed methods to deal with one incomplete covariate. The multiple imputation method investigated in Chapter 2 can be extended to the setting with more than one incomplete covariate, and/or an incomplete response variable. Suppose that, in addition to missing subgroup data, the response of interest is also incomplete, and let $\pi_i(\boldsymbol{\alpha}) = P(R_i^y = 1|\mathcal{D}_i;\boldsymbol{\alpha})$ where R_i^y indicates whether response variable Y_i is observed and $\mathcal{D}_i$ is some combination of the data measured at baseline. It is of interest to investigate the properties of the following weighted estimating equation:

$$\mathbf{U}_i(\boldsymbol{\beta};\boldsymbol{\rho},\boldsymbol{\alpha},\boldsymbol{\xi}_1) \;=\; \sum_{l=0}^{1} \frac{R_i}{\pi_i(\boldsymbol{\rho})} \frac{R_i^y}{\pi_i(\boldsymbol{\alpha})} \frac{I(W_i = l)}{\pi_i(\boldsymbol{\xi}_1)^l(1-\pi_i(\boldsymbol{\xi}_1))^{1-l}} \; \mathbf{D}_i(\boldsymbol{\beta})[V_i(\boldsymbol{\beta})]^{-1}\big(Y_i - \mu_i(\boldsymbol{\beta})\big) \;,$$

where $\boldsymbol{\beta}$, $\boldsymbol{\xi}_1$, $\boldsymbol{\rho}$ and $\mathbf{D}_i(\boldsymbol{\beta})$ and $V_i(\boldsymbol{\beta})$ are defined in Section 2.2.2.

In an extension of the multiple imputation method proposed in Chapter 4, an investigation of an appropriate model for imputing an incomplete response variable is of interest. Fully conditional specification (FCS) multiple imputation methods can be easily implemented through the use of standard statistical software, and there are many options for the imputation model (e.g. logistic regression, linear regression, predictive mean matching, etc.) (van Buuren and Groothuis-Oudshoorn, 2011). As well, in FCS multiple imputation, the missing data need not have a monotone missing data pattern to implement the imputation, which allows for flexibility.

6.2.4 Extension to Longitudinal Data

Interest may lie in estimating marginal causal odds ratios in the longitudinal data setting; the doubly weighted estimating function method described in Chapter 2 and multiple

imputation with a weight for confounding investigated in Chapter 4 can be extended to the longitudinal data setting.

Chen et al. (2010) introduce a method for estimation of conditional causal parameters using inverse probability weighting for missing data in a longitudinal setting; covariate and response data may be incomplete at each follow-up visit. We propose an extension of these methods to estimate marginal causal parameters in an observational setting where confounding may be present.

Let Y_{ij} denote a binary response variable, X_{ij} denote a binary treatment variable in a randomized setting, and S_{ij} denote a binary subgroup variable, for subject i in a random sample of n individuals, $i = 1, ..., n$ with measurement at the jth assessment/follow-up visit. In an observational setting, we let W_{ij} denote the treatment. Let $\mathbf{Z}_{1ij}$ denote a vector of variables that are directly associated with response, and $\mathbf{Z}_{2ij}$ denote a vector of variables that do not have direct effects on the response.

Suppose the model for the response at assessment j is given by

$$\mu_{ij}(\boldsymbol{\vartheta}) = P(Y_{ij} = 1|X_{ij}, S_{ij}, \mathbf{Z}_{ij}; \boldsymbol{\vartheta}) = \text{expit}(\vartheta_0 + \vartheta_x X_{ij} + \vartheta_s S_{ij} + \vartheta_{xs} X_{ij} S_{ij} + \mathbf{Z}_{ij}\boldsymbol{\vartheta}_z)$$

for $j = 1, ..., J$. Suppose we are interested in fitting the marginal response model

$$\mu_{ij}(\boldsymbol{\beta}) = \text{expit}(\beta_0 + \beta_x X_{ij} + \beta_s S_{ij} + \beta_{xs} X_{ij} S_{ij})$$

where

$$\begin{aligned} \mu_{ij}(\boldsymbol{\beta}) &= P(Y_{ij} = 1|X_{ij}, S_{ij}; \boldsymbol{\beta}) \\ &= \sum_{z=0}^{1} P(Y_{ij} = 1|X_{ij}, S_{ij}, Z_{ij} = z; \boldsymbol{\vartheta}) P(Z_{ij} = z|S_{ij}) \end{aligned}$$

with $\mathbf{Z}_{ij} = Z_{ij}$ for convenience. We make the assumption that $X_{ij} \perp S_{ij}, Z_{ij}$ in a randomized setting.

Let R^y_{ij} indicate whether the response variable is measured at the jth assessment for subject i, and R^s_{ij} denote whether the subgroup variable is measured at the jth assessment. Let $\pi^{sy}_{ijk} = P(R^s_{ij} = 1, R^s_{ik} = 1, R^y_{ik} = 1|\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i})$, where $\mathbf{Y}_i, \mathbf{W}_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}$ represent

the vectors of variables measured at each assessment (some components may be incomplete). Chen et al. (2010) propose a weighted estimating equation with weight matrix $\boldsymbol{\Delta}_i = [w_{ijk}]_{J\times J}$, where $w_{ijk} = I(R^s_{ij} = 1, R^s_{ik} = 1, R^y_{ik} = 1)/\pi^{sy}_{ijk}$.

In an observational setting, it is of interest to investigate the addition of an inverse probability weight for confounding. Let

$$\pi^w_{ij} = \prod_{k=0}^{j}\left[\sum_{l=0}^{1}\frac{I(W_{ik} = l)}{P(W_{ik} = l|\bar{\mathbf{W}}_{ik}, \bar{\mathbf{Z}}_{1ik}, \bar{\mathbf{Z}}_{2ik})}\right],$$

where $\bar{\mathbf{W}}_{ik} = \{W_{i1}, ..., W_{i,k-1}\}$, $\bar{\mathbf{Z}}_{1ik} = \{Z_{1i1}, ..., Z_{1ik}\}$ and $\bar{\mathbf{Z}}_{2ik} = \{Z_{2i1}, ..., Z_{2ik}\}$. We propose an augmentation of weight matrix $\boldsymbol{\Delta}_{ijk}$ to include π^w_{ij}.

References

H. Aguinis, J. C. Beaty, R. J. Boik, and C. A. Pierce. Effect size and power in assessing moderating effects of categorical variables using multiple regression: a 30-year review. *Journal of Applied Psychology*, 90(1):94–107, 2005. doi: 10.1037/0021-9010.90.1.94.

H. O. Al-Shamsi, A. A. Farsi, M. Anjum, H. Shen, K. Zbuk, R. J. Cook, L. Linkins, and P. Major. Thrombotic events in metastatic colorectal cancer patients treated with leucovorin, fluorouracil and irinotican (FOLFIRI) plus bevacizumab. *J Gastrointest Oncol*, 6(3):274–279, 2015. doi: 10.3978/j.issn.2078-6891.2015.025.

P. D. Allison. Multiple imputation for missing data: a cautionary tale. *Sociological Methods & Research*, 28(3):301–309, 2000.

D. G. Altman and J. N. S. Matthews. Interaction 1: heterogeneity of effects. *BMJ*, 313: 486, 1996.

S. F. Assman, S. J. Pocock, L. E. Enos, and L. E. Kasten. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, 355:1064–69, 2000.

P. C. Austin. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, 29:661–677, 2009. doi: 10.1177/0272989X09341755.

P. C. Austin and M. D. Escobar. Bayesian modeling of missing data in clinical research. *Computational Statistics & Data Analysis*, 49:821–836, 2005.

P. C. Austin, P. Grootendorst, and G. M. Anderson. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine*, 26:734–753, 2007a.

P. C. Austin, P. Grootendorst, S. T. Normand, and G. M. Anderson. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in Medicine*, 26:754–768, 2007b.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

J. Barnard and X. Meng. Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*, 8:17–36, 1999.

K. Bhaskaran and L. Smeeth. What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology*, 43(4):1336–1339, 2014. doi: 10.1093/ije/dyu080.

N. Black. Why we need observational studies to evaluate the effectiveness of health care. *BMJ*, 312:1215–1218, 1996.

J. M. Bland and D. G. Altman. Statistics notes: the odds ratio. *BMJ*, 320(7247):1468, 2000.

S. T. Brookes, E. Whitely, M. Egger, G. Davey Smith, P. A. Mulheran, and T. J. Peters. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *Journal of Clinical Epidemiology*, 57:229–236, 2004. doi: 10.1016/j.jclinepi.2003.08.009.

B. A. Brumback, M. A. Hernán, S. J. P. A. Haneuse, and J. M. Robins. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*, 23:749–767, 2004. doi: 10.1002/sim.1657.

J. R. Carpenter and M. G. Kenward. *Multiple Imputation and its Application*. John Wiley & Sons, New York, 2013.

J. R. Carpenter, M. G. Kenward, and S. Vansteelandt. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J. R. Statist. Soc. A*, 169: 571–584, 2006.

J. R. Carpenter, M. G. Kenward, and I. R. White. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research*, 16:259–275, 2007.

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

S. R. Cole and M. A. Hernán. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168:656–664, 2008.

L. M. Collins, J. L. Schafer, and C. Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351, 2001.

R. J. Cook, K.-A. Lee, M. Cuerden, and C. A. Cotton. Inverse probability weighted estimating equations for randomized trials in transfusion medicine. *Statistics in Medicine*, 32:4380–4399, 2013.

C. A. Cotton. *Inference for Treatments Targeting Control of an Intermediate Measure*. PhD thesis, University of Washington, 2009.

B. J. Crowe, I. A. Lipkovich, and O. Wang. Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. *Pharmaceutical Statistics*, 9:269–279, 2010.

R. H. Dehejia and S. Wahba. Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84:151–161, 2002.

C. Eulenburg, A. Suling, P. Neuser, A. Reuss, U. Canzler, T. Fehm, A. Luyten, M. Hellriegel, L. Woelber, and S. Mahner. Propensity scoring after multiple imputation in a

retrospective study on adjuvant radiation therapy in lymph-node positive vulvar cancer. *PLoS ONE*, 11(11), 2016. doi: 10.1371/journal.pone.0165705.

A. X. Garg, A. Kurz, D. I. Sessler, M. Cuerden, A. Robinson, M. Mrkobrada, C. R. Parikh, R. Mizera, P. M. Jones, and M. Tiboni. Perioperative aspirin and clonidine and risk of acute kidney injury: a randomized clinical trial. *JAMA*, 312(21):2254–2264, 2014. doi: 10.1001/jama.2014.15284.

S. Greenland. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*, 125:761–768, 1987.

S. Greenland. Randomization, statistics, and causal inference. *Epidemiology*, 1(6):421–429, 1990.

S. Greenland and B. Brumback. An overview of relations among causal modelling methods. *International Journal of Epidemiology*, 31:1030–1037, 2002.

S. Greenland and H. Morgenstern. Ecological bias, confounding, and effect modification. *International Journal of Epidemiology*, 18(1):269–274, 1989.

S. Greenland and J. M. Robins. Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15:413–418, 1986.

S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14:29–46, 1999.

B. Hasselman. *nleqslv: Solve Systems of Nonlinear Equations*, 2017. URL https://CRAN.R-project.org/package=nleqslv. R package version 3.3.1.

M. A. Hernán and J. M. Robins. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*, 60:578–586, 2006. doi: 10.1136/jech.2004.029496.

M. A. Hernán, B. A. Brumback, and J. M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11: 561–570, 2000.

M. A. Hernán, B. A. Brumback, and J. M. Robins. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine*, 21:1689–1709, 2002.

J. Hill. Reducing bias in treatment effect estimation in observational studies suffering from missing data. Columbia University Institute for Social and Economic Research and Policy (ISERP) Working Paper, 2004.

H. I. Hurwitz, L. Fehrenbacher, J. D. Hainsworth, W. Heim, J. Berlin, E. Holmgren, J. Hambleton, W. F. Novotny, and F. Kabbinavar. Bevacizumab in combination with fluorouracil and leucovorin: An active regimen for first-line metastatic colorectal cancer. *Journal of Clinical Oncology*, 23(15):3502–3508, 2005.

J. G. Ibrahim, S. R. Lipsitz, and M. Chen. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61:173–190, 1999.

J. G. Ibrahim, M. Chen, S. R. Lipsitz, and A. H. Herring. Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association*, 100:332–347, 2005.

M. P. Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91:222–230, 1996.

C. Leyrat, S. R. Seaman, I. A. White, I. Douglas, L. Smeeth, J. Kim, M. Resche-Rigon, J. R. Carpenter, and E. J. Williamson. Propensity score analysis with partially observed covariates: how should multiple imputation be used? *Statistical Methods in Medical Research*, 2017. doi: 10.1177/0962280217713032.

T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society*, 44:226–233, 1982.

T. M. Loux, C. Drake, and J. Smith-Gagen. A comparison of marginal odds ratio estimators. *Statistical Methods in Medical Research*, 26(1):155–175, 2017.

J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23: 2937–2960, 2004. doi: 10.1002/sim.1903.

T. Martinussen and S. Vansteelandt. On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Analysis*, 19:279–296, 2013.

C. D. Mazer, R. P. Whitlock, D. A. Fergusson, E. Belley-Cote, K. Connolly, B. Khanykin, A. J. Gregory, É. Médicis, F. M. Carrier, and S. McGuinness. Six-month outcomes after restrictive or liberal transfusion for cardiac surgery. *N Engl J Med*, 379:1224–1233, 2018.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

M. McIsaac and R.J. Cook. Statistical methods for incomplete data: some results on model misspecification. *Statistical Methods in Medical Research*, 26(1):248–267, 2017. doi: 10.1177/0962280214544251.

X. Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9:538–558, 1994.

O. S. Miettinen and E. F. Cook. Confounding: essence and detection. *American Journal of Epidemiology*, 114:593–603, 1981.

R. Mitra and J. P. Reiter. A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical Methods in Medical Research*, 25:188–204, 2016.

E. E. M. Moodie, J. A. C. Delaney, G. Lefebvre, and R. W. Platt. Missing confounding data in marginal structural models: a comparison of inverse probability weighting and multiple imputation. *The International Journal of Biostatistics*, 4(1), 2008.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994. doi: 10.1016/S1573-4412(05)80005-4.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:465–480, 1923.

M.C. Paik and W. Tsai. On using the cox proportional hazards model with missing covariates. *Biometrika*, 84(3):579–593, 1997.

P.N. Papanikolaou, G.D. Christidi, and J.P.A. Ioannidis. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *Canadian Medical Association Journal*, 174(5):635–641, 2006.

J. Pearl. Causality: models, reasoning, and inference. *Economic Theory*, 19:675–685, 2003.

J. Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6, 2010.

B. B. L. Penning de Vries and R. H. H. Groenwold. Comments on propensity score matching following multiple imputation. *Statistical Methods in Medical Research*, 25(6):3066–3068, 2016.

M. L. Petersen, K. Porter, S. Gruber, Y. Wang, and M. J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *The Berkeley Electronic Press*, 2010.

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

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

C. L. Raison, R. E. Rutherford, B. J. Woolwine, C. Shuo, P. Schettler, D. F. Drake, E. Haroon, and A. H. Miller. A randomized controlled trial of the tumor necrosis factor antagonist infliximab for treatment-resistant depression. *JAMA Psychiatry*, 70:31–41, 2013.

V. Ranpura, S. Hapani, and S. Wu. Treatment-related mortality with bevacizumab in cancer patients: a meta-analysis. *JAMA*, 305:487–497, 2011.

P. H. Rezvan, K. J. Lee, and J. A. Simpson. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*, 15(30), 2015.

J. M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran and D. Berry, editors, *Statistical Models in Epidemiology: The Environment and Clinical Trials*, pages 95–134. New York: Springer-Verlag, 2000.

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

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

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

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

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

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

P. M. Rothwell. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*, 365:176–186, 2005.

A. Rotnitzky and J. M. Robins. Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scandinavian Journal of Statistics*, 22(3): 323–333, 1995.

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

D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–92, 1976.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.

D. B. Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

D. B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489, 1996.

M. W. Saif and R. Mehra. Incidence and management of bevacizumab-related toxicities in colorectal cancer. *Expert Opin Drug Saf*, 5:553–566, 2006.

L. B. Saltz, S. Clarke, E. Diaz-Rubio, W. Scheithauer, A. Figer, R. Wong, S. Koski, M. Lichinitser, T. Yang, and F. Rivera. Bevacizumab in combination with oxaliplatin-based chemotherapy as first-line therapy in metastatic colorectal cancer: A randomized phase III study. *Journal of Clinical Oncology*, 26(12):2013–2019, 2008.

J. L. Schafer. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8: 3–15, 1999.

J. L. Schafer. Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, 57(1):19–35, 2003.

K. F. Schulz and D. A. Grimes. Multiplicity in randomised trials II: subgroup and interim analyses. *Lancet*, 365:1657–61, 2005.

K. F. Schulz, D. G. Altman, and D. Moher. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, 8(18), 2010.

S. Seaman and I. White. Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communication in Statistics - Theory and Methods*, 43(16): 3499–3515, 2014.

S. R. Seaman and I. R. White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22:278–95, 2013.

J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338(b2393), 2009. doi: 10.1136/bmj.b2393.

A. C. Tsai, S. D. Weiser, M. L. Petersen, K. Ragland, M. B. Kushel, and D. R. Bangsberg. A marginal structural model to estimate the causal effect of antidepressant medication treatment on viral suppression among homeless and marginally housed persons living with HIV. *Arch Gen Psychiatry*, 67(12):1282–1290, 2010. doi: 10.1001/archgenpsychiatry.2010.160.

W. Vach. *Logistic Regression with Missing Values in the Covariates*. Springer, New York, 1994.

W. Vach and M. Blettner. 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*, 134:895–907, 1991.

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. URL http://www.jstatsoft.org/v45/i03/.

W. M. van der Wal and R. B. Geskus. ipw: An R Package for Inverse Probability Weighting. *Journal of Statistical Software*, 43(13):1–23, 2011. URL http://www.jstatsoft.org/v43/i13/.

T. J. VanderWeele. On the distinction between interaction and effect modification. *Epidemiology*, 20(6):863–871, 2009.

A. J. Viera. Odds ratios and risk ratios: what's the difference and why does it matter? *Southern Medical Journal*, 101(7), 2008. doi: 10.1097/SMJ.0b013e31817a7ee4.

E. von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. Vandenbroucke. The strengthening the reporting of observational studies in epidemiology

(STROBE) statement: guidelines for reporting observational studies. *Epidemiology*, 18 (6):800–804, 2007.

S. Yusuf, J. Wittes, J. Probstfield, and H. A. Tyroler. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*, 266: 93–98, 1991.

G. Zou. A modified Poisson regression approach to prospective studies with binary data. *Am J Epi*, 159:702–706, 2004.

APPENDICES

Appendix A

Empirical Power Calculations

A.1 Marginal Causal Effects

Here we present the results of simulation studies designed to investigate the empirical power to detect a significant subgroup effect using the marginal causal model methods presented in Chapters 2, 3 and 4. We use the parameter settings given in Section 2.4.1 for the observational setting with non-ignorable missingness, unless otherwise specified. To investigate the power of the methods introduced, we vary the value of β_3 which is the regression coefficient associated with the interaction effect WS. For the application of multiple imputation method proposed in Section 4.4.2, all two-way interaction terms are included in the imputation model for the incomplete subgroup variable. The number of imputed datasets is $K = 5$.

For the weighted EM-type algorithm method, the parameter specifications are given in Chapter 3, Section 3.5.1, unless otherwise specified.

We consider two sample sizes: $n = 500$ and $n = 2000$. Although these are the sample sizes for the entire sample of subjects, when a complete case analysis is used, the sample size is smaller. We vary the percentage of missing data to be 20% and 40%.

For each simulated dataset, a subgroup effect is considered significant if the 2-sided p-value for a test of the null hypothesis that the interaction term $\beta_3 = 0$ is less than 0.05

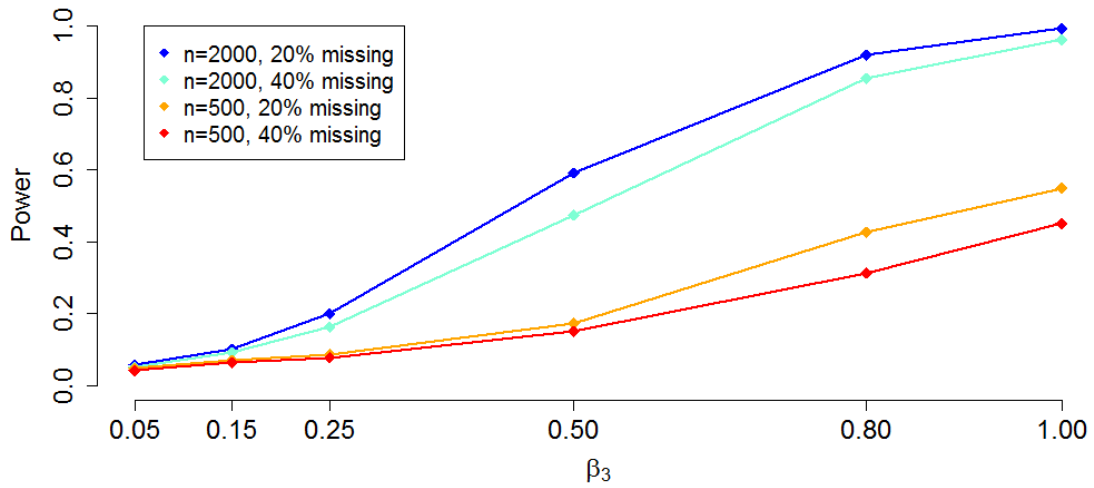

Figure A.1: Empirical power to detect a significant subgroup effect in a marginal causal model using the doubly weighted estimating equation method (Section 2.2.2); $\alpha = 0.05$.

(i.e. $\alpha = 0.05$). Each simulation study contains 1000 independent simulated datasets. The seed value for generating 1000 independent datasets is recorded for reproducibility.

A.1.1 Discussion

In Figure A.1, we see that the empirical power to detect a significant interaction effect using the doubly weighted estimating equation approach introducted in Chapter 2 increases as the subgroup effect parameter (β_3) increases, as expected. As well, the empirical power is higher when the percentage of missing data is lower, also as expected. For a sample size of 2000, the power to detect a significant subgroup effect is >80% when $\beta_3 \geq 0.80$, for either percentage of missingness (20% or 40%). The power to detect a significant subgroup effect is $< 80\%$ when the sample size is 500 and for $0.0 < \beta_3 \leq 1.0$.

The doubly weighted EM-type algorithm method and the weighted multiple imputation method are very similar to the doubly weighted estimating equation approach in terms of power. See Figures A.2 and A.3.

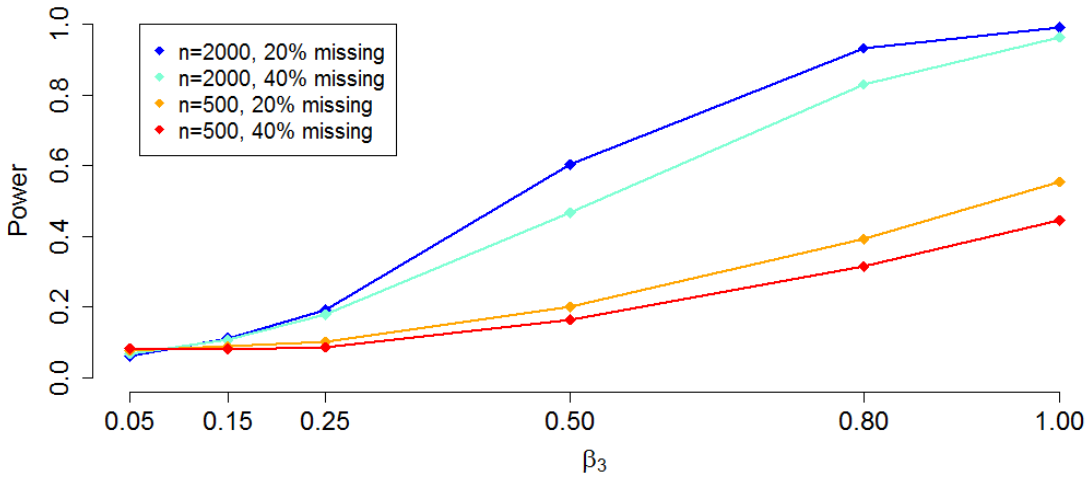

Figure A.2: Empirical power to detect a significant subgroup effect in a marginal causal model using the doubly weighted EM-type algorithm method (Section 3.4.2); $\alpha = 0.05$.

A.2 Conditional Causal Effects

Next, we investigate the empirical power to detect a significant subgroup effect in a complete case conditional regression model without weights. The parameter settings for $\boldsymbol{\xi}_1$, $\boldsymbol{\xi}_2$, ζ_1, ζ_2 and $\boldsymbol{\rho}$ in the observational setting with non-ignorable missing data (when we do not include Z_1 in the response model) are given in Chapter 2. The conditional regression parameters $\boldsymbol{\vartheta}$ from equation (2.2) are set to

$$P(Y_i = 1|W_i, S_i, Z_{1i}) = \text{expit}(\log(1) + \log(0.75)W_i + \log(1.2)S_i + \vartheta_3 W_i S_i + \log(1.25)W_i Z_{1i}) \ ,$$

where ϑ_3, the subgroup effect coefficient, is allowed to vary. Note that in the conditional response model, the missingness is ignorable since we condition on Z_1. Also, the confounding effect is adjusted for when we include Z_1 in the response model. Therefore a complete case unweighted analysis is appropriate here.

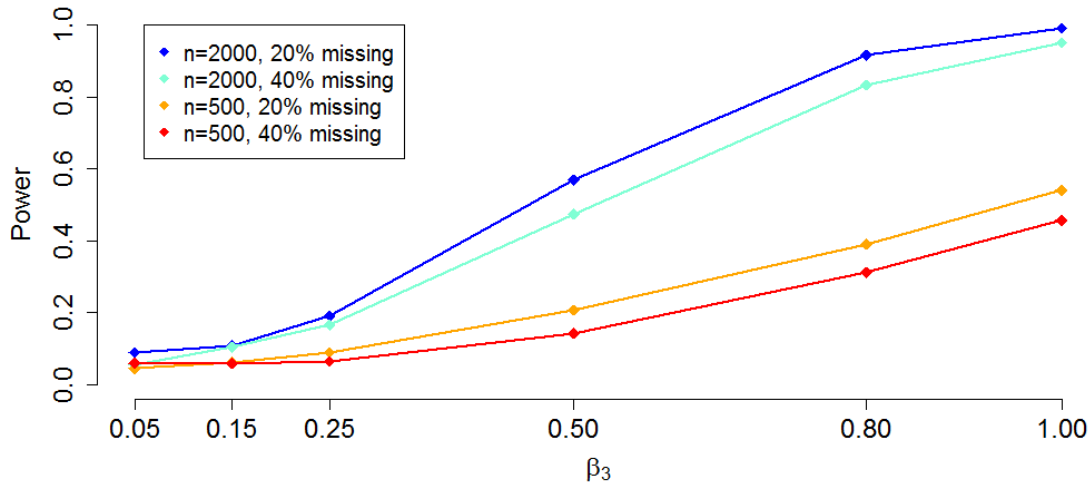

Figure A.3: Empirical power to detect a significant subgroup effect in a marginal causal model using the weighted multiple imputation method (Section 4.4.2); $\alpha = 0.05$.

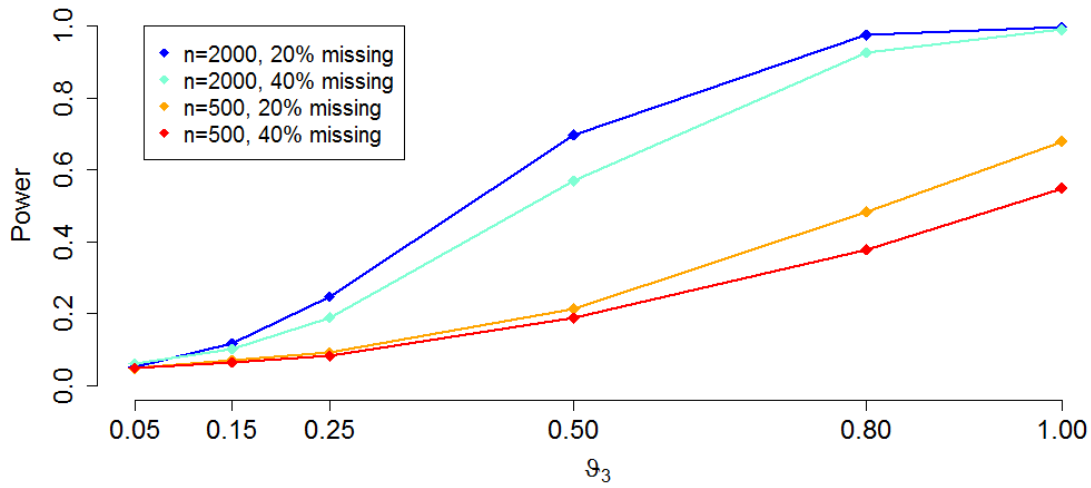

Figure A.4: Empirical power to detect a significant subgroup effect in a conditional causal model using complete case analysis; $\alpha = 0.05$.

A.2.1 Discussion

The power to detect a significant subgroup effect is higher in the conditional causal model compared to the marginal causal models. This is likely due to the extra variability introduced in the estimation of the weights, and the extra variability introduced in the imputation of the missing subgroup variable.

Overall, the power estimates found in these simulation studies are consistent with other results (Aguinis et al., 2005; Brookes et al., 2004). In our studies, given the chosen parameters, a large sample size ($n \geq 2000$) is required have adequate power to detect a relatively large subgroup effect ($\beta_3 > 0.5$).

Appendix B

Simulation Study to Explore the Effect of a Subgroup Variable That Is Also a Confounder

Here we perform a simulation study to investigate the estimation of marginal causal effects in the setting where the subgroup variable is also a confounder.

Suppose the true propensity score model is

$$P(W_i = 1|Z_{1i}, Z_{2i}, S_i; \boldsymbol{\xi}_1^\dagger) \quad = \quad \text{expit}(\xi_{10}^\dagger + \xi_{11}^\dagger Z_{1i} + \xi_{12}^\dagger Z_{2i} + \xi_{13}^\dagger S_i) \ ,$$

but we omit the incomplete subgroup variable from the propensity score model and fit the following model using the full dataset

$$P(W_i = 1|Z_{1i}, Z_{2i}; \boldsymbol{\xi}_1) \quad = \quad \text{expit}(\xi_{10} + \xi_{11} Z_{1i} + \xi_{12} Z_{2i}) \ .$$

In our simulation study, we use the parameter settings for $\boldsymbol{\beta}$, ϑ_4, $\boldsymbol{\xi}_2$, ζ_1, ζ_2 and $\boldsymbol{\rho}$ for the observational setting with non-ignorable missing data given in Section 2.4.1. The parameters for the true propensity score model are

$$P(W_i = 1|Z_{1i}, Z_{2i}, S_i; \boldsymbol{\xi}_1^\dagger) \quad = \quad \text{expit}(\text{logit}(0.2) + \log(4.0) Z_{1i} + \log(4.0) Z_{2i} + \log(2.0) S_i) \ .$$

Table B.1: Empirical bias and efficiency of estimated marginal regression coefficients, using the doubly inverse probability weighted estimating equation approach, the doubly weighted EM-type algorithm, and weighted multiple imputation.

	β_1				$\beta_1 + \beta_3$			
	EBias	ASE	ESE	ECP	EBias	ASE	ESE	ECP
Doubly weighted estimating equation method								
20% missing	0.0037	0.1640	0.1650	0.9512	0.0080	0.1661	0.1675	0.9462
40% missing	0.0020	0.1896	0.1882	0.9540	0.0113	0.1922	0.1939	0.9460
Doubly weighted EM-type algorithm								
20% missing	0.0311	0.1760	0.1769	0.9458	0.0148	0.1804	0.1830	0.9470
40% missing	0.0296	0.1724	0.1715	0.9488	0.0129	0.1715	0.1724	0.9464
Weighted multiple imputation method								
20% missing	0.0025	0.1563	0.1554	0.9482	-0.0022	0.1590	0.1624	0.9426
40% missing	0.0016	0.1719	0.1713	0.9466	-0.0034	0.1750	0.1744	0.9488

Abbreviations: EBias Empirical bias, ASE asymptotic standard error, ESE empirical standard error, ECP empirical coverage probability

In the simulation studies where the doubly weighted EM-type algorithm is used, the parameter settings are given in Section 3.5.1, with the exception of propensity score model parameters $\boldsymbol{\xi}_1^{\dagger}$ which are given above.

For each simulation study, the number of subjects per dataset is 2000 and the number of simulated datasets is 5000. The simulated datasets are independent, and the seed value for the first simulated dataset is the same across the simulation studies.

B.1 Discussion

In Table B.1 we see that consistency is affected in the doubly weighted estimating equation method when we omit the subgroup variable from the propensity score model which is the basis for the weight for confounding, but the degree of bias is relatively low. In the EM-

type algorithm where the weight for confounding is misspecified, the bias is quite large in comparison. This is likely because the conditional distribution for the subgroup variable is also misspecified. Consistency is not affected in the weighted multiple imputation method.

In the complete case doubly weighted estimating equation approach, one may wish to estimate the propensity score including the subgroup variable as a covariate using complete cases only. If the propensity score model is properly specified, the parameters are independent from the parameters of the missing data model, and the missingness does not depend on the treatment variable, a complete case analysis is a valid approach. Weights can therefore be estimated for each subject in the complete cases dataset, and the doubly weighted estimating equation method can be implemented.

In the multiple imputation approach, the subgroup variable can be imputed, and the propensity score can be estimated within each imputed dataset by correctly specifying the propensity score model (i.e. including S). Standard methods to combine the estimates from each imputed dataset can be used, while incorporating the inverse probability weight for confounding in the variance estimation (see Section 4.4.2) (Crowe et al., 2010; Qu and Lipkovich, 2009; Seaman and White, 2014).