Fractional Imputation for Ordinal and Mixed-type Responses with Missing Observations

by

Xichen She

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

Waterloo, Ontario, Canada, 2017

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	NAME: Christian Léger Title: Professor (Université de Montréal)
Supervisor(s)	NAME: Changbao Wu Title: Professor
Internal Member	NAME: Joel Dubin Title: Associate Professor
Internal Member	NAME: Peisong Han Title: Assistant Professor
Internal-external Member	NAME: Liping Fu Title: Professor (Civil and Environmental Engineering)

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

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

Abstract

This thesis addresses two essential aspects of large scale public-use data files involving ordinal and mixed-type responses with missing observations: **(i)** the creation of single complete data sets with imputation for missing values; and **(ii)** the statistical analysis of imputed data sets by public data users with different objectives. Large scale data sets are typically collected by statistical agencies, research institutes or commercial organizations and missing observations are a common feature. Our research focuses on scenarios where one ordinal response or several mixed-type responses are part of the data sets and are subject to missingness. We develop a sequential regression fractional imputation procedure to create single complete data sets which provide valid and efficient statistical analysis for commonly encountered inferential problems by public data users.

Ordinal variables are widely collected and analyzed in many scientific fields. They share some common tools with discrete data analysis but have much richer structure to explore as compared to general categorical variables. More importantly, statistical methods developed for ordinal variables can be readily extended to cover categorical data. In this thesis, we present the sequential regression fractional imputation strategy through three major research projects, starting from ordinal variables and extending to mixed-type responses. The proposed method takes into account unique features of ordinal responses and is theoretically sound and practically appealing.

The first project considers a simple scenario where there is only one ordinal response with missing values. We provide detailed steps for the proposed imputation procedure and develop asymptotic properties of subsequent estimators derived under a general setting. We discuss in great detail three inferential problems of practical importance: **(1)** estimation of category probabilities; **(2)** regression analysis using all available covariates; and **(3)** regression analysis involving a subset of all the covariates. For each problem, the proposed procedure is compared with existing alternative methods in terms of validity and efficiency of the analysis. Finite sample performances are demonstrated through simulation studies.

The second research project extends the proposed procedure to more complex scenarios where multiple variables of mixed types, including continuous, ordered and unordered categorical variables, all contain missing observations. We outline the key steps for the sequential regression fractional imputation procedure under general

settings and present asymptotic results on statistical analysis through two specific inferential problems: **(1)** test of independence for two ordinal responses via association measures; and **(2)** regression of an ordinal response on continuous covariates where both the response and the covariates are subject to missingness. Simulation studies reveal that our proposed procedure provides superior results as compared to existing methods.

In the third research project, we study the robustness of the estimators for marginal population quantities by incorporating missing data mechanisms into the proposed procedure. Two cases are considered: one of a univariate ordinal response with missing values and the other of longitudinal ordinal responses with monotone missingness. We show the power of the proposed procedure through an application to a causal inference problem in a point-treatment study. The double robustness property of the estimators for marginal population quantities using the fractionally imputed data sets against misspecification of the imputation models as well as the response probability models is confirmed through results from simulation studies.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Changbao Wu for his patience and support in overcoming obstacles I have been facing through my PhD studies. It is under his guidance that I was able to finish writing this thesis and learned how to be an effective researcher and how to stay motivated and enthusiastic through whatever one might be facing.

I would also like to thank the rest of my thesis examining committee: Dr. Joel Dubin, Dr. Peisong Han, Dr. Liping Fu and Dr. Christian Léger for their invaluable comments and insightful questions, which gave me incentives to widen and deepen the current research from different perspectives.

My special thanks go to my parents and my girlfriend Jie Shen for their continuous support and encouragement throughout the past four years.

Lastly, my research receives financial support from various sources, including the research assistantship from the Canadian Statistical Sciences Institute (CANSSI), the Natural Sciences and Engineering Research Council of Canada (NSERC) grant to Prof. Changbao Wu, the International Doctoral Student Award and the Doctoral Thesis Completion Award from University of Waterloo and teaching assistantships from the Department of Statistics and Actuarial Science. I am very grateful to these funding sources which make my dream of pursuing PhD studies at University of Waterloo a reality.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Statistical Analysis with Missing Data

The problem of missing data is pervasive in many scientific fields. Its presence hinders researchers' ability to draw reliable conclusions from the data and thus statistical techniques to deal with incomplete observations are essential. Statistical analysis with missing data faces two distinct scenarios. It could be investigating a data set of small or moderate size collected for specific scientific purposes and the analysis is carried out by specific researchers who have full access to the data set and are equipped with a profound knowledge of statistics. It has become increasingly common, however, that data sets are collected by a large research team or a statistical agency and contain missing values for multiple variables. The researchers handling missing data only serve as data suppliers who create one or several complete data sets with missing values properly treated and then make them available for public use. The processed data sets can be accessed by multiple users for subsequent analyses with different research objectives. The main focus of this thesis is to address the missing data problem in the second scenario. Discussions on handling missing data for in-house use as in the first scenario can be found in Little and Rubin (2002).

In general, methods for handling incomplete data for public use can be readily applied to a specific problem, but tailor-made methods to deal with missing data from particular studies are not always suitable for creating public use data files because of two fundamental requirements for the public use files: **(i)** ease of implementation of

subsequent analyses, and **(ii)** validity of subsequent inferences with various objectives and partially available information. For the first requirement, the data users usually "have access only to complete-data software and possesses limited knowledge of specific reasons and models for nonresponse" (Rubin 1996), therefore, it is critical that the analyses by the users can be carried out through explicit steps which only involve standard complete-data analysis, or at most, with some minor and easy-to-implement modifications. For the second requirement, to provide valid inferential results is a primary task for any statistical methods, but it is a particularly challenging one for constructing public use files. The data files are accessible by multiple users, who may have different scientific interests and may choose different approaches to inferences. It is, therefore, necessary that the data files are created with a wide range of possible subsequent analyses taken into consideration. The restrictions on the access to and usage of complete information by the data users make the task even more difficult. This could happen when, for example, the file creators use supplementary information such as administrative records to construct the data set but this information is removed due to confidentiality concerns when the data are disseminated to users. See Rubin (1987), Raghunathan et al. (2003) and Reiter (2008).

There exists extensive literature on handling missing data for general purpose estimation. Assumptions or estimating techniques may vary, but most of these methods are based on three strategies: **(1)** to ignore, **(2)** to re-weight and **(3)** to impute. The first strategy is to simply ignore missing observations and to analyze the observed responses only. This is also known as available-case analysis. When the missing data are "ignorable" (Molenberghs and Kenward 2007), likelihood-based analyses of the available cases provide valid results. For public use data files, in order to supply to the data users "normal" datasets without missing values, a more aggressive method, called complete-case analysis (CCA) or listwise deletion, is widely adopted in practice. The CCA method deletes observations with missing values for at least one response and applies standard complete-data analyses to the remaining fully observed cases. When the fully observed units are not representative of the original sample, the CCA approach generally leads to biased inferences, but if the missing rate is low, the adverse impact of deleting incomplete observations is negligible. As we will show in Chapter 3, for some particular analyses, CCA is not only valid but also efficient.

The second strategy adjusts the weights of observed units to compensate for ignoring incomplete observations and is often called the inverse probability weighting

(IPW) or propensity score adjusting method, as termed by Rosenbaum and Rubin (1983). It shares the same spirit with the well-known Horvitz-Thompson estimator (Horvitz and Thompson 1952) used in survey sampling under unequal probability sampling designs. Discussions on the IPW methods can be found in Kim and Riddles (2012); Seaman and White (2013) and references therein. With properly chosen weights, the IPW methods usually correct the potential bias induced by CCA, but one major drawback of IPW methods is the lack of efficiency, since they fail to take full advantage of information contained in the incomplete observations. To improve the efficiency of IPW estimators, Robins et al. (1994) proposed the class of augmented IPW (AIPW) estimators, which is further discussed in Robins and Rotnitzky (1995). See Chapter 2 for a detailed introduction. Many other estimators stemming from the AIPW method with favorable properties such as multiple robustness and maximum efficiency have been developed in recent years; see for example, Tan (2010), Tang and Qin (2012), Han and Wang (2013), among others. However, AIPW is not well suited for constructing public use data, because it requires the data users to devise different augmented terms for different subsequent analyses and extra efforts to solve the augmented equations. Nevertheless, the idea of incorporating models for both the data generating process and the missing data mechanism can be borrowed to improve the robustness of subsequent estimators derived from public use files. See Chapter 5 for details.

The third strategy on imputation for missing values has attracted tremendous amount of attention from researchers in the past 30 years. It is extremely appealing for the creation of public use files because by filling in missing values with plausible predictions, imputation usually results in a synthetic "complete" data set which can be conveniently investigated by data users (Brick and Kalton 1996). Early attempts of imputation mainly focus on single imputation techniques such as regression imputation and hot deck imputation, for which a missing response is replaced by a single imputed one, leading to a single complete data set which resembles the original one. Two major drawbacks hinder the wide use of single imputation, with the first being the lack of efficiency and second being the absence of good variance estimation techniques. To tackle these two problems, Rubin (1978) proposed the multiple imputation (MI) method and Rubin (1987) further addressed this topic. MI generates multiple sets of imputed values for missing responses and creates several copies of complete data sets. This slightly increases the burden of manipulating files on the end-users,

but in exchange, it reduces the extra variation, termed imputation variance, induced by the imputation procedure and more importantly, it captures the uncertainty in generating an imputed value and provides a simple way of estimating variances of estimators based on the imputed data sets which involves repeating standard analyses with the multiple copies of complete data sets separately and combining the results through an intuitive formula, known as the Rubin's combining rule. See Section 2.2.3. The MI was originally developed for creating public use data files (Rubin 1987, 1996), but it has seen widespread applications in various fields, see, for example, Lavori et al. (1995), Van Buuren et al. (1999), Raghunathan et al. (2003) and Zhao et al. (2015).

However, multiple imputation is not the ultimate solution to the problem of public use file creation. The MI was motivated under the Bayesian framework, but the validity of variance estimators obtained from the combining rule is controversial from a frequentist's perspective. Many authors discussed cases where the combining rule failed to yield sensible variance estimators. See Meng (1994), Robins and Wang (2000), Nielsen (2003), Kim et al. (2006) and Yang and Kim (2016*b*) among others. It turns out that extra conditions are required for the combining rule to be justifiable. Meng (1994) proposed a sufficient condition called "congeniality" for a multiple imputation method to be "proper" (Rubin 1996), that is, the variance estimators produced by the combining rule correctly estimate the variances of subsequent estimators derived from the imputed data sets. The congeniality condition, however, imposes constraints on both the imputation procedure and the complete-data analyses carried out by the end-user, which is very restrictive for general purpose estimation. Even when this condition does hold, Wang and Robins (1998) and Nielsen (2003) both showed that Rubin's variance estimator is weakly unbiased rather than consistent, for finite number of imputations, in the sense that the variance estimator converges to some non-degenerate distribution with the mean equal to the true variance. This can result in longer-than-desirable confidence intervals in some cases.

Fractional imputation (FI) has recently surged as an attractive alternative to multiple imputation for handling incomplete data set for general purpose estimation, with its idea dating back to Kalton and Kish (1984). Under fractional imputation, an incomplete observation is replaced by a cluster of imputed units, each assigned with a fractional weight to recover the distributional structure of the missing responses, resulting in a single but enlarged data file. From a practical point of view, this is more appealing than the MI which requires the storage and manipulation of multiple data

sets. The FI approach can also effectively reduce the imputation variance (Kalton and Kish 1984; Fay 1996) and provide valid and efficient inferences without the "congeniality" condition. There have been increased research activities on the topic since the paper of Kim and Fuller (2004). Yang and Kim (2016*a*) presented an excellent review on the recent development of fractional imputation.

1.2 Analysis of Ordinal Responses

Ordinal responses are categorical variables with an intrinsic order among categories but without quantitative measurements on the scales. Ordinal data are routinely collected and analyzed in many scientific fields, such as psychological and behavioural sciences, public health and medical studies, and business and management. Examples of ordinal responses include variables measuring performance (poor, average, excellent), attitude (disagree, neutral, agree), severity of disease (mild, moderate, severe), and many others. Ordinal variables are sometimes observed directly, such as responses to survey questions on 3-point or 5-point Likert scales, and sometimes can also be derived based on values of other observed variables, especially for measuring level of performance and severity of diseases. A well-known example is the stage of obesity defined by the body mass index (BMI) (Zhao et al. 2015), which is derived based on one's body weight and height.

Statistical analyses of ordinal responses are closely related to methodologies developed for binary and categorical data. They share some common tools, but ordinal data have much richer structure to explore as compared to general categorical variables owing to the ranking of categories. Past 30 years have seen major advances in literature to address the ordinal nature of the data. More importantly, statistical methods developed for ordinal variables can be readily extended to cover categorical data by simply disregarding the order. Agresti (2013) contains an excellent coverage on categorical data analysis and Agresti (2010) provides a comprehensive account on analysis of ordinal responses.

There are two fundamental inference problems in statistics: **(i)** estimation of the mean response; and **(ii)** regression against covariates. The mean response represents the overall population average of the variable and is the main summary characteristic of the population. The treatment effect in health and medical studies and the

effectiveness of an intervention in social sciences is usually measured by the difference of two population means. Regression analysis has been one of the pillars of modern statistics. Its primary objective is to establish the relationship between the conditional expectation of the response variable given a set of covariates and the covariates themselves, and to further identify significant factors which affect the response variable. Dependence between the response variable and the set of covariates established through regression analysis also serves as a crucial step for causal inference. Rao (2009) is a classic reference on linear regression analysis and Fan and Gijbels (1996) contains stimulating materials on nonparametric regression methods. An important development on regression analysis with categorical and discrete response variables is the emergence of quasi-likelihood theory and generalized linear modelling techniques (McCullagh and Nelder 1983). For longitudinal and clustered responses and other multivariate responses, the generalized estimating equation (GEE) methodology (Liang and Zeger 1986) has become the most powerful tool for semiparametric regression analysis.

For ordinal data, the population distribution is characterized by the probabilities of the ordinal categories, which can be considered as mean responses of indicator vectors; see Section 2.4.2 for details. Estimation of those probabilities would be of interest for many applications; see Section 5.2 for examples. When multiple ordinal responses are under consideration, the contingency table analysis is an extended effort to characterize the population which goes beyond marginal category probabilities of each ordinal variable and investigates the interrelation between responses. In this thesis, we mainly focus on non-model-based association measures as studied in Kendall (1945), Goodman and Kruskal (1954), Somers (1962) and Lang (2008). Other methods to measure the association based on log-linear models, which are described in Chapter 6 of Agresti (2010), are not treated here.

For regression analysis of ordinal responses, the ordinal nature of the data did not receive appropriate treatment until McCullagh (1980) proposed the popular proportional odds model based on the cumulative probabilities. McCullagh and Nelder (1983) interpreted the latent variable motivation behind the proportional odds model and suggested several extensions to the original model. Peterson and Harrell Jr (1990) proposed the partial proportional odds model to allow for a different odds ratio at each level. For the analysis of discrete survival time data, models based on continuation ratios are preferred, see, for example, Tutz (1991), Cole and Ananth (2001) and

Tutz and Binder (2004). For multivariate ordinal responses, the GEE is commonly adopted when the primary interest lies in the marginal dependence of the responses on the covariates. See Lipsitz et al. (1994), Heagerty and Zeger (1996), Parsons et al. (2006) and Touloumis et al. (2013). If the correlation between responses is also of interest, likelihood-based methods can be used. The joint distribution of multivariate ordinal responses can be specified either by a set of models for the marginal expectations and association measures or a set of sequential regression models for each ordinal response with other responses as covariates. For the first approach, see Dale (1986), Molenberghs and Lesaffre (1994) and Ekholm et al. (2003); for the second, see Lindsey et al. (1997) and Müller and Czado (2005).

1.3 Contributions and Outline of the Thesis

It is apparent that there exists a rich literature on statistical analysis of missing data and of ordinal responses as two separate topics. However, little attention has been given to the analysis of ordinal responses with missing observations, especially in public-use data files intended for general-purpose estimation. This thesis addresses two essential aspects of large scale public-use data files involving ordinal and mixed-type incomplete responses: **(i)** the creation of single complete data sets with imputation for missing values; and **(ii)** the statistical analysis of imputed data sets by data users with different objectives.

We present a fractional imputation strategy based on sequential regression modeling, starting from data sets with a single ordinal variable subject to missingness and extending to those containing multiple mixed-type incomplete responses. With the single fractionally imputed "complete" data set, we demonstrate that users are able to conduct a variety of valid inferences with existing complete-data softwares plus minor extra efforts to incorporate the fractional weights. We further improve the proposed approach by adding protection for the estimators of marginal quantities against model misspecification. This can be extremely beneficial to problems such as estimating average treatment effects, where marginal distributions are of primary interest. The proposed method takes into account unique features of ordinal responses and are theoretically sound and practically appealing.

In Chapter 2, we first set the stage for our discussions by introducing some key

notation and assumptions, and then provide a brief review of methods for handling missing data and for analyzing ordinal responses in the current literature.

In Chapter 3, we consider a simple scenario where there is only one ordinal response with missing values. We provide detailed steps for the proposed fractional imputation procedure and develop asymptotic properties of estimators derived under a general setting. We discuss in great detail three inferential problems of practical importance: **(1)** estimation of category probabilities; **(2)** regression analysis using all available covariates; and **(3)** regression analysis involving a subset of all the covariates. For each problem, the proposed procedure is compared with existing alternatives in terms of validity and efficiency of the analysis. Finite sample performances are demonstrated through simulation studies.

Chapter 4 extends the proposed procedure to more complex scenarios where multiple variables of mixed types, including continuous, ordered and unordered categorical variables, all contain missing observations. We outline the key steps for the sequential regression fractional imputation procedure under general settings and present asymptotic results on statistical analysis through two specific inferential problems: **(1)** test of independence for two ordinal responses via association measures; and **(2)** regression of an ordinal response on continuous covariates where both the response and the covariates are subject to missingness. Simulation studies reveal that our proposed procedure provides superior results as compared to existing methods.

In Chapter 5, we study the robustness of the estimators of marginal quantities by incorporating missing data mechanisms into the proposed imputation procedure. Two cases are considered: one of a univariate ordinal response with missing values and the other of longitudinal ordinal responses with monotone missingness. We show the power of the improved procedure through an application to a causal inference problem in a point-treatment study (Robins et al. 2000). The double robustness property of the estimators of the marginal probabilities using the fractionally imputed data sets against misspecification of the imputation models as well as the response probability models is confirmed through results from simulation studies.

We conclude this thesis in Chapter 6 with a detailed discussion on the application of the proposed method to data sets collected from complex surveys along with several interesting topics worthy of further exploration in the future.

Chapter 2

Missing Data Methods and Ordinal Response Analysis

2.1 Basic Settings

In this section, we first describe the general problem we attempt to tackle and set up notation that will be used throughout the thesis. The original incomplete data set is denoted by $\mathcal{O} = \big\{(\boldsymbol{r}_i, \boldsymbol{y}_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\big\}$ which is an independent and identically distributed (*i.i.d.*) sample of size n of $(\boldsymbol{R}, \boldsymbol{Y}, \boldsymbol{X})$, where $\boldsymbol{X}$ is a p-dimensional vector of fully-observed baseline variables and $\boldsymbol{Y} = (Y_1, \ldots, Y_T)'$ consists of mixed-type variables with missing values in observations, which may include continuous, unordered categorical and ordinal components, and $\boldsymbol{R}$ is the corresponding indicator vector recording the availability of the components of $\boldsymbol{Y}$ such that

$$R_t = \begin{cases} 1, & \text{if } Y_t \text{ is observed,} \\ 0, & \text{if } Y_t \text{ is missing.} \end{cases} \qquad \text{for } t = 1, \ldots, T\,. \tag{2.1}$$

Let $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{mis}$ be the observed and missing components of $\boldsymbol{Y}$, respectively. Throughout this thesis, we use uppercase letters to represent random variables and lowercase letters for their realizations indexed by i. For example, $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,T})$ is the ith observation of $\boldsymbol{Y}$. Our aim is to create a complete version of the data set $\mathcal{O}$ to facilitate general purpose inferences conducted by different data users. To investigate the performance of estimators derived from the created data set, we consider

parameters of interest $\boldsymbol{\theta}$ defined in a fairly flexible way by an unbiased estimating function $\boldsymbol{U}(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta})$ (Godambe 1991) such that:

$$E\Big\{\boldsymbol{U}(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta}_0)\Big\} = \mathbf{0}\,, \tag{2.2}$$

for some $\boldsymbol{\theta}_0$ in the parameter space. This class covers a wide range of important parameters that may be of interest in practical studies, including the marginal means, regression coefficients, association measures, etc. Detailed discussions and examples are presented in the following chapters.

In the absence of missing data, $\boldsymbol{\theta}$ can be consistently estimated by solving the following sample-based estimating equations,

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta})\,. \tag{2.3}$$

Let $\hat{\boldsymbol{\theta}}_n$ be the solution to (2.3). It belongs to the classic m-estimator family (Newey and McFadden 1994; Tsiatis 2006).The following theorem summarizes the asymptotic behaviour of $\hat{\boldsymbol{\theta}}_n$:

Theorem 2.1. *Let* $\big\{(\boldsymbol{y}_i, \boldsymbol{x}_i)\ i = 1, \ldots, n\big\}$ *be an* i.i.d. *sample from some joint distribution* P *and* $\boldsymbol{U}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta})$ *be a fixed vector-valued function with* $\boldsymbol{\theta}$ *taking values in parameter space* Θ*. Denote*

$$\Psi(\boldsymbol{\theta}) = E\big[\boldsymbol{U}(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta})\big] \quad \textit{and} \quad \Psi_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta})\,.$$

Under the regularity conditions given in Section 2.5, the sequence of estimators $\hat{\boldsymbol{\theta}}_n$ *satisfying* $\Psi_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$ *converges in probability to* $\boldsymbol{\theta}_0$*, which satisfies* $\Psi(\boldsymbol{\theta}_0) = \mathbf{0}$*. Furthermore,*

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = -\Big[\mathring{\Psi}(\boldsymbol{\theta}_0)\Big]^{-1} n^{-1} \sum_{i=1}^{n} \boldsymbol{U}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}_0) + o_p\big(n^{-1/2}\big)\,, \tag{2.4}$$

where $\mathring{\Psi}(\boldsymbol{\theta}) = E\big[\partial \boldsymbol{U}(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}\big]$.

Proof of Theorem 2.1 can be found in Tsiatis (2006).

2.2 Missing Data Methods

2.2.1 Missing Mechanisms and Patterns

Missing data are a common feature of large data sets. The mechanisms underlying the occurrence of missing data can be classified into three types by the dependence of response probabilities on the observed information (Little and Rubin 2002). We say the data are missing-completely-at-random (MCAR) if the distribution of the response indicator $\boldsymbol{R}$ is independent of all the other variables, whether observed or not, that is,

$$P(\boldsymbol{R} \mid \boldsymbol{Y}, \boldsymbol{X}) = P(\boldsymbol{R}) \,.$$

The data are called missing-at-random (MAR) if the response probabilities only depend on observed information, that is,

$$P(\boldsymbol{R} \mid \boldsymbol{Y}, \boldsymbol{X}) = P(\boldsymbol{R} \mid \boldsymbol{Y}_{obs}, \boldsymbol{X}) \,,$$

where $\boldsymbol{Y}_{obs}$ consists of observed components of $\boldsymbol{Y}$. In some cases, it is also plausible to consider a scenario which falls in between MCAR and MAR, known as covariate-dependent-missing (CDM) (Little 1995), when the response probabilities only depend on the fully-observed baseline covariates but not the partially observed variables, that is,

$$P(\boldsymbol{R} \mid \boldsymbol{Y}, \boldsymbol{X}) = P(\boldsymbol{R} \mid \boldsymbol{X}) \,.$$

Note that in the univariate case where only one response is subject to missingness, CDM is equivalent to MAR. If the data do not satisfy either of the above assumptions, they are called missing-not-at-random (MNAR). In practice, the MNAR is probably the most realistic assumption to impose, because the probability of observing a variable often relies, more or less, on the potential value of that variable, especially for variables of sensitive nature. Unfortunately, it usually requires more complicated model assumptions and additional information to identify model parameters. Even for a given MNAR model, the model is not fully verifiable from the available data, just like the MAR assumption. Molenberghs and Kenward (2007) showed that for every MNAR model, we can always construct an MAR counterpart that achieves exactly the same fit to the observed data. For these reasons, arguments throughout the thesis are established under the MAR assumption, which is common practice adopted by

many other studies in the literature. If the data are indeed MNAR, the analysis under MAR still serves as an anchor for the sensitivity analysis suggested by Molenberghs and Kenward (2007).

When there exists multiple variables with missing observations, we can alternatively classify missing data into two major patterns: monotone missingness and intermittent missingness. The data are said to follow a monotone missing pattern if there exists a permutation $\mathcal{P}(\cdot)$ of $\{1, \ldots, T\}$ such that $R_{\mathcal{P}(t_1)} = 0$ implies $R_{\mathcal{P}(t_2)} = 0$ for any $\mathcal{P}(t_2) > \mathcal{P}(t_1)$, that is, after some proper reordering, if one response is missing then all the following responses in the reordered sequence are not observed. Monotone missingness can often be found in longitudinal studies, where each individual is repeatedly measured over a period of time. Once a subject drops out of the study at a certain stage, he/she usually will never return, and the monotone missing data are thereby sometimes called "data with dropouts" in these studies. If the data are missing arbitrarily and we can not observe a clear pattern even after permutation, we say the data are intermittently missing. Table 2.1 shows two simple examples of monotone and intermittent missing data.

Table 2.1: An Illustration of Different Missing Patterns

Monotone Missng					
Y_5	Y_4	Y_3	Y_2	Y_1	X
×	×	×	×	×	×
·	·	×	×	×	×
·	·	×	×	×	×
·	·	·	×	×	×
·	·	·	·	×	×
·	·	·	·	×	×
·	·	·	·	·	×

Intermittent Missing					
Y_5	Y_4	Y_3	Y_2	Y_1	X
×	·	×	×	·	×
·	×	·	×	·	×
·	·	×	·	×	×
·	×	·	·	×	×
×	×	×	×	·	×
×	·	·	·	·	×
·	×	×	×	×	×

2.2.2 Complete-case Analysis and Inverse Probability Weighting

Complete-case analysis (CCA), or listwise deletion, simply ignores the observations with missing values and only keeps the fully observed units. Under our settings, the subsequent estimator of $\boldsymbol{\theta}$ using data sets created by CCA is the solution to

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \delta_i \boldsymbol{U}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) \,, \tag{2.5}$$

where $\delta_i = \boldsymbol{I}(\boldsymbol{r}_i = \mathbf{1})$ is the complete case indicator and $\mathbf{1} = (1, \ldots, 1)'$ is a T-dimensional vector of 1.

Inverse probability weighting (IPW) is a technique attempting to correct the potential bias of estimators based on CCA by assigning an inverse-probability weight to each complete observation. Under our settings, the full response probabilities, also called the propensity scores by Rosenbaum and Rubin (1983), used to construct the weights are given by

$$\pi_i = P(\delta_i = 1 \mid \boldsymbol{y}_i, \boldsymbol{x}_i) = P(\boldsymbol{r}_i = \mathbf{1} \mid \boldsymbol{y}_i, \boldsymbol{x}_i) \quad \text{for } i = 1, \ldots, n \,. \tag{2.6}$$

An IPW estimator of $\boldsymbol{\theta}$ then follows by solving

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \delta_i \pi_i^{-1} \boldsymbol{U}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) \,. \tag{2.7}$$

In some cases, π_i's are known, for example, the design weights in survey sampling, but in general missing data problems, proper models for the missing data process (MDP) are required to estimate the π_i's. We will elaborate on the modelling in the following chapters. Let $\hat{\pi}_i$'s be the estimated response probabilities from the MDP models, then the IPW estimator $\hat{\boldsymbol{\theta}}_{ipw}$ of $\boldsymbol{\theta}$ is the solution to

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \delta_i \hat{\pi}_i^{-1} \boldsymbol{U}(\boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) \,. \tag{2.8}$$

To take full advantage of available information and to improve efficiency, Robins et al. (1994) proposed a class of augmented IPW (AIPW) estimators. In the simple

univariate case, an AIPW estimator can be obtained by solving

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \left[\delta_i \hat{\pi}_i^{-1} \boldsymbol{U}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) + (1 - \delta_i \hat{\pi}_i^{-1}) h(\boldsymbol{x}_i) \right], \tag{2.9}$$

where $h(\cdot)$ is an arbitrary vector-valued function of the same dimension as $\boldsymbol{U}(y, \boldsymbol{x}; \boldsymbol{\theta})$. The resulting estimator, denoted by $\hat{\boldsymbol{\theta}}_{aipw}(h)$ to emphasize its dependence on the choice of h, is consistent for any function h. For properly chosen h, the AIPW estimators can be more efficient then the original IPW estimators. In this particular case, the optimal h leading to the most efficient estimator is $E\{\boldsymbol{U}(Y, \boldsymbol{X}; \boldsymbol{\theta}) \mid \boldsymbol{X} = \boldsymbol{x}\}$.Tsiatis (2006) provides a comprehensive discussion on AIPW estimators under more general settings. To construct appropriate augmented terms, it requires modelling the distribution of $(\boldsymbol{Y}, \boldsymbol{X})$, that is, the data generating process (DGP). Scharfstein et al. (1999) showed that the optimal AIPW estimator is doubly robust in the sense that it is consistent if either the MDP model or the DGP model is correct but not necessarily both. We will further investigate this topic in Chapter 5.

2.2.3 Multiple Imputation

For the missing components $\boldsymbol{y}_{mis}$ in each observation, multiple imputation generates M imputed values, denoted by $\boldsymbol{y}_{mis}^{(l)}$, $l = 1, \ldots, M$, and creates M complete data sets by filling in missing values with these imputed responses. Each complete data set is then analyzed using standard methods as if they were fully-observed. Under our settings, an estimator $\hat{\boldsymbol{\theta}}_{mi}^{(l)}$ is obtained by solving

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}(\boldsymbol{y}_i^{(l)}, \boldsymbol{x}_i; \boldsymbol{\theta}), \qquad l = 1, \ldots, M, \tag{2.10}$$

where $\boldsymbol{y}_i^{(l)} = (\boldsymbol{y}_{i,obs}, \boldsymbol{y}_{i,mis}^{(l)})$ is the imputed vector of $\boldsymbol{y}_i$ in the lth data set. A final point estimator is obtained by taking the average of the estimators obtained separately from the M imputed data sets and is given by

$$\hat{\boldsymbol{\theta}}_{mi} = M^{-1} \sum_{l=1}^{M} \hat{\boldsymbol{\theta}}_{mi}^{(l)}. \tag{2.11}$$

To conduct further inferences, Rubin proposed a simple and convenient combining rule to estimate the variance of $\hat{\boldsymbol{\theta}}_{mi}$. Let $\hat{\boldsymbol{V}}^{(l)}$ be the variance estimator from the lth imputed data set using standard analysis. Rubin suggested the variance of the final estimator be approximated by

$$\hat{\boldsymbol{\Sigma}}_{Rubin} = M^{-1} \sum_{l=1}^{M} \hat{\boldsymbol{V}}^{(l)} + (1 + M^{-1})(M-1)^{-1} \sum_{l=1}^{M} (\hat{\boldsymbol{\theta}}_{mi}^{(l)} - \hat{\boldsymbol{\theta}}_{mi})^{\otimes 2}, \tag{2.12}$$

where $\boldsymbol{A}^{\otimes 2}$ denotes $\boldsymbol{A}\boldsymbol{A}'$. The first term on the right hand side is the average of estimated variances from M complete data sets and the second term corresponds to the variance inflation caused by imputation.

The most important step of MI is to generate $\boldsymbol{y}_{mis}^{(l)}$, $l = 1, \ldots, M$. There are two distinct approaches. For the first approach, the values $\boldsymbol{y}_{mis}^{(l)}$ are drawn from $f(\boldsymbol{y}_{mis} \mid \boldsymbol{y}_{obs}, \boldsymbol{x}; \hat{\boldsymbol{\eta}}^{(l)})$ for $l = 1, \ldots, M$, where $f(\boldsymbol{y}_{mis} \mid \boldsymbol{y}_{obs}, \boldsymbol{x}; \boldsymbol{\eta})$ denotes the conditional distribution of $\boldsymbol{y}_{mis}$ given all the observed information parameterized by $\boldsymbol{\eta}$, which is available from the DGP model and $\hat{\boldsymbol{\eta}}^{(l)}$'s are independent samples from the Bayesian posterior distribution of $\boldsymbol{\eta}$ derived from the observed data, the DGP model and a proper prior. The estimator $\hat{\boldsymbol{\theta}}_{mi}^{a}$ given by (2.11) is named the "type-A" estimator by Rubin (1987). For the second approach, the values $\boldsymbol{y}_{mis}^{(l)}$ are an *i.i.d.* sample from a fixed distribution $f(\boldsymbol{y}_{mis} \mid \boldsymbol{y}_{obs}, \boldsymbol{x}; \hat{\boldsymbol{\eta}})$, where $\hat{\boldsymbol{\eta}}$ is a preliminary consistent estimator of $\boldsymbol{\eta}$ from the observed data, often taken as the maximum observed likelihood estimator (MOLE). The resulting estimator $\hat{\boldsymbol{\theta}}_{mi}^{b}$ is referred to as the "type-B" estimator.

Wang and Robins (1998) and Robins and Wang (2000) investigated the asymptotic properties of both types of estimators and advocated the use of "type-B" estimators if "one's concern is with estimation efficiency", because the "type-B" estimator is strictly more efficient than the "type-A" estimator for finite number of imputation and the difference can be significant under some circumstances. One drawback, however, of the "type-B" estimator is the absence of a computationally convenient variance estimator. Both Rubin (1987) and Wang and Robins (1998) noted that unlike the "type-A" estimator whose variance can be easily estimated by the combing rule, $\hat{\boldsymbol{\Sigma}}_{Rubin}$ in (2.12) does not correctly estimate the variance of a "type-B" estimator. Robins and Wang (2000) proposed a consistent variance estimator based on the derived asymptotic variance, but their approach requires not only the data creator to supply additional information and to disclose details of the imputation procedure, but also the user to

conduct extra computation that is not directly available from existing softwares. This is not desirable for public use data files.

In Chapter 3, we will compare estimators based on the proposed method with the "type-B" MI estimators regarding efficiency and show that our proposed fractional imputation estimators have even smaller asymptotic variance than the already more efficient "type-B" estimators for finite number of imputation. Moreover, the asymptotic variance of subsequent estimators can be consistently estimated directly from the imputed data set with existing softwares.

2.3 Contingency Table Analysis of Bivariate Ordinal Responses

For simplicity, we confine our discussion to two-way contingency tables. Let Y_1 and Y_2 be the two ordinal variables of interest on a J- and K-level scale, respectively. In the absence of missing values, observations can be cross-classified into a $J \times K$ table of cell counts, denoted by n_{jk} for the cell in the jth row and kth column, based on the response values. For a fixed sample size n, the cell counts of the contingency table follow a multinomial distribution. We denote the probability of the bivariate ordinal responses falling into the cell in the jth row and kth column by

$$\pi_{jk} = P(Y_1 = j, Y_2 = k), \quad j = 1, \ldots, J, \quad k = 1, \ldots, K\,.$$

Let $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{1K}, \ldots, \pi_{J1}, \ldots, \pi_{JK})'$ be the vector of all cell probabilities. We have $\sum_{j=1}^{J}\sum_{k=1}^{K}\pi_{jk} = 1$. The marginal distributions of the responses are of basic interest and are denoted by $\boldsymbol{\pi}_1 = (\pi_{1+}, \ldots, \pi_{J+})'$ and $\boldsymbol{\pi}_2 = (\pi_{+1}, \ldots, \pi_{+K})'$, where $\pi_{j+} = \sum_{k=1}^{K}\pi_{jk}$ and $\pi_{+k} = \sum_{j=1}^{J}\pi_{jk}$. The dependence between the two ordinal responses, however, is often the main focus for the analysis of bivariate data. In such cases measures of association are of primary concern. A simple example is the conditional distribution of Y_1 given Y_2 at level k:

$$\boldsymbol{\pi}_{1|k} = (\pi_{1|k}, \ldots, \pi_{J|k})', \quad k = 1, \ldots, K\,,$$

where $\pi_{j|k} = P(Y_1 = j \mid Y_2 = k) = \pi_{jk}/\pi_{+k}$. A more popular example of measuring association is through different types of ordinal odds ratios, including the local (θ^L_{jk}),

the cumulative (θ^C_{jk}) and the global (θ^G_{jk}) odds ratios, defined respectively as

$$\theta^L_{jk} = \frac{\pi_{jk}\pi_{j+1,k+1}}{\pi_{j,k+1}\pi_{j+1,k}}, \qquad \theta^C_{jk} = \frac{\left(\sum_{b\le k}\pi_{jb}\right)\left(\sum_{b>k}\pi_{j+1,b}\right)}{\left(\sum_{b>k}\pi_{jb}\right)\left(\sum_{b\le k}\pi_{j+1,b}\right)}$$
$$\text{and} \quad \theta^G_{jk} = \frac{\left(\sum_{a\le j}\sum_{b\le k}\pi_{ab}\right)\left(\sum_{a>j}\sum_{b>k}\pi_{ab}\right)}{\left(\sum_{a\le j}\sum_{b>k}\pi_{ab}\right)\left(\sum_{a>j}\sum_{b\le k}\pi_{ab}\right)}$$

for $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Note that both θ^C_{jk} and θ^G_{jk} have incorporated the ordinality of the responses in the definition and are only well-defined for ordinal variables, not nominal ones.

It is sometimes more appealing to characterize the association between two ordinal variables by a single summary index rather than a set of odds ratios. Several such measures have been proposed based on the probabilities of concordance and discordance. Two ordinal observations (y_{i1}, y_{i2}) and (y_{m1}, y_{m2}) are concordant if the subject ranking higher on Y_1 also ranks higher on Y_2; while they are discordant if the one ranking higher on Y_1 ranks lower on Y_2. Goodman and Kruskal (1954) proposed to use the parameter *gamma* defined as

$$\gamma = \left(\prod\nolimits_c - \prod\nolimits_d\right) / \left(\prod\nolimits_c + \prod\nolimits_d\right), \tag{2.13}$$

where $\prod_c = 2\sum_{j<a}\sum_{k<b}\pi_{jk}\pi_{ab}$ and $\prod_d = 2\sum_{j<a}\sum_{k>b}\pi_{jk}\pi_{ab}$, corresponding to the probabilities of concordance and discordance for two randomly selected observations. The value of γ ranges from -1 to 1. When $|\gamma| = 1$, there is a monotone relationship between Y_1 and Y_2, but not necessarily strictly monotone. For example, $\gamma = 1$ indicates that if $y_{i1} < y_{m1}$ then $y_{i2} \le y_{m2}$. When Y_1 and Y_2 are independent, we have $\gamma = 0$, but the reverse statement is not true. Other examples of association measures include Kendall's *Tau-b* (Kendall 1945) and Somers' d (Somers 1962), both having the same numerator $\prod_c - \prod_d$. The plug-in estimator of $\prod_c - \prod_d$ is given by $C - D$, where $C = 2\sum_{j<a}\sum_{k<b}\hat{\pi}_{jk}\hat{\pi}_{ab}$ and $D = 2\sum_{j<a}\sum_{k>b}\hat{\pi}_{jk}\hat{\pi}_{ab}$ and $\hat{\pi}_{jk} = n_{jk}/n$. Simon (1978) showed that any estimated measures based on $C - D$ are equivalent in terms of efficacy for testing independence. The Wald-type test statistic for independence is given by

$$z = \left(C - D\right) / \hat{\sigma}_{\text{C-D}}, \tag{2.14}$$

where $\hat{\sigma}_{\text{C-D}}$ can be the *nonnull standard error* of $C-D$ or the *null standard error* using the relations $\pi_{rj} = \pi_{r+}\pi_{+j}$ under independence. Agresti (2010) recommended to use the latter one and claimed that the test statistic with *null standard error* converges to normal distribution faster under the null hypothesis. Since ordinal responses are a special type of categorical data, the Pearson χ^2 test is also applicable. However, the latter is designed for a general alternative and may not have good power for testing a trend, which is of primary interest for ordinal responses. On the contrary, the z statistic given in (2.14) is very natural for alternative hypotheses such as $\prod_c > \prod_d$ or $\prod_c < \prod_d$, corresponding to a positive and negative trend.

2.4 Regression Analysis of Ordinal Responses

2.4.1 Model Formulation

We consider regression models for an ordinal element Y_t of $\boldsymbol{Y}$ with J_t ordinal levels against $\boldsymbol{X}$ and assume Y_t is fully observed for all individuals. Throughout the section, the subscription "t" is suppressed for simplicity of notation. The data set and relevant variables are denoted by $\big\{(y_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\big\}$ and $(Y, \boldsymbol{X})$.

A regression model can be specified in the following general form:

$$G^{-1}\big[\omega_j(\boldsymbol{X})\big] = \alpha_j - \boldsymbol{\beta}_j'\boldsymbol{X}, \qquad j = 1, \ldots, J-1, \tag{2.15}$$

where G^{-1} is a link function, $(\alpha_j, \boldsymbol{\beta}_j')'$ are the intercept and coefficients of level j and $\omega_j(\boldsymbol{X})$'s are a one-to-one transformation of the category probabilities $P(Y = j \mid \boldsymbol{X})$. Clearly, the choice of $\omega_j(\boldsymbol{X})$ is of critical importance. It should reflect the ordinal nature of the response and have an intuitive interpretation at the same time. Two commonly adopted quantities are $\omega_j(\boldsymbol{X}) = P(Y \le j \mid \boldsymbol{X})$ as in *cumulative link models* and $\omega_j(\boldsymbol{X}) = P(Y = j \mid Y \ge j, \boldsymbol{X})$ as in *continuation-ratio link models*. For link functions, the *logit*, *probit* and *c-log-log* functions are all possible options. When $\omega_j(\boldsymbol{X})$ is chosen as the cumulative probability of level j, it is often sensible and preferred to assume a common effect of $\boldsymbol{X}$ on each level, that is

$$G^{-1}\big[\gamma_j(\boldsymbol{X})\big] = \alpha_j - \boldsymbol{\beta}'\boldsymbol{X}, \qquad j = 1, \ldots, J-1, \tag{2.16}$$

with $\alpha_1 \leq \cdots \leq \alpha_{J-1}$ and $\gamma_j(\boldsymbol{X}) = P(Y \leq j \mid \boldsymbol{X})$. Model (2.16) with the *logit* link is often referred to as the *proportional odds model*, which is one of the most popular models for ordinal responses in practice. As we shall see in Section 2.4.3, the *cumulative link models* of the form (2.16) are motivated by an underlying process which discretizes a latent variable with a set of thresholds and hence are perfectly suited for ordinal responses derived by categorizing continuous variables. A big advantage of having the same slope for all categories is that it preserves the order structure of the cumulative probabilities. In a general form of (2.15), the curves of cumulative probabilities at different levels may overlap which leads to negative category probabilities. This will never happen in (2.16) since the curves are parallel to each other with shifts determined by α_j's. However, the common effect assumption should not be taken for granted. There exist methods developed to test the validity of this assumption, see, for example, Brant (1990), Peterson and Harrell Jr (1990) and Kim (2003). In cases where the common effect assumption is not appropriate, the more general form (2.15) is necessary. Discussions on general models with different effects for each level include Peterson and Harrell Jr (1990), Cox (1995) and Cole et al. (2004). The *continuation-ratio link models* are more appropriate for ordinal responses on development scales which are determined by a sequential process, for example, the survival times in medical studies. They do not have the limitation of the *cumulative link models* and always provide valid probabilities even without the common effect assumption.

2.4.2 Parameter Estimation

Let $\boldsymbol{\eta}$ be the parameters in the regression model (2.15) and they can be estimated by maximum likelihood method. Note that results in this section are derived more for a theoretical need to estimate parameters in different models under a unified framework. In practical problems, more computationally convenient parameter estimation is possible. We first derive the cumulative probabilities $\gamma_j(\boldsymbol{X};\boldsymbol{\eta}) = P(Y \leq j \mid \boldsymbol{X})$ from the models. This is straightforward for *cumulative link models* and also not difficult for *continuation-ratio link models* because continuation ratios are a one-to-one transformation of category probabilities. It can be shown that, for *continuation-ratio link models* of form (2.15),

$$\gamma_j(\boldsymbol{X};\boldsymbol{\eta}) = 1 - (1 - G_1) \cdots (1 - G_j)\,, \tag{2.17}$$

where $G_j = G(\alpha_j - \boldsymbol{\beta}_j' \boldsymbol{X})$. We then define the cumulative indicator vector for the ordinal response as $\boldsymbol{Z} = (Z_1, \ldots, Z_{J-1})'$, where $Z_j = \boldsymbol{I}(Y \leq j)$ for $j = 1, \ldots, J-1$ and $\boldsymbol{I}(\cdot)$ is the indicator function. The realizations are denoted by $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{i(J-1)})'$ for $i = 1, \ldots, n$. Let $z_{i0} = 0$ and $z_{iJ} = 1$ for all i. Let $\boldsymbol{\gamma}_i = E(\boldsymbol{Z} \mid \boldsymbol{x}_i) = (\gamma_{i1}, \ldots, \gamma_{i(J-1)})'$, where $\gamma_{ij} = E(Z_j \mid \boldsymbol{x}_i) = P(Y \leq j \mid \boldsymbol{x}_i) = \gamma_j(\boldsymbol{x}_i; \boldsymbol{\eta})$. It follows that $P(Y = j \mid \boldsymbol{x}_i) = \gamma_{ij} - \gamma_{i(j-1)}$ and the likelihood function is given by

$$L(\boldsymbol{\eta}) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} \Big[\gamma_{ij} - \gamma_{i(j-1)} \Big]^{z_{ij} - z_{i(j-1)}} \right\},$$

where $\gamma_{i0} = 0$ and $\gamma_{iJ} = 1$ for all i. The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\eta}}$ is the solution to the score equations given by

$$\boldsymbol{0} = \frac{\partial}{\partial \boldsymbol{\eta}} \log L(\boldsymbol{\eta}) = \sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{B}_i (\boldsymbol{z}_i - \boldsymbol{\gamma}_i), \tag{2.18}$$

where $\boldsymbol{D}_i = \partial \boldsymbol{\gamma}_i / \partial \boldsymbol{\eta}$ is of dimension $(J-1) \times (J-1+p)$, p is the dimension of $\boldsymbol{X}$, $\boldsymbol{B}_i = \boldsymbol{V}_i^{-1}$, and $\boldsymbol{V}_i$ is the $(J-1) \times (J-1)$ variance-covariance matrix of $\boldsymbol{z}_i$ with the (jk)th entry given by $\gamma_{ij}(1 - \gamma_{ik})$. We use the notation $\boldsymbol{D}(\boldsymbol{x}_i; \boldsymbol{\eta})$, $\boldsymbol{B}(\boldsymbol{x}_i; \boldsymbol{\eta})$ and $\boldsymbol{\gamma}(\boldsymbol{x}_i; \boldsymbol{\eta})$ to emphasize the dependence of the terms $\boldsymbol{D}_i$, $\boldsymbol{B}_i$ and $\boldsymbol{\gamma}_i$ on $\boldsymbol{x}_i$ and $\boldsymbol{\eta}$. From (2.18), the MLE $\hat{\boldsymbol{\eta}}$ is also the inverse variance weighted least square estimator. The parameter $\boldsymbol{\eta}$ can be defined through the following unbiased estimating function:

$$\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\eta}) = \boldsymbol{D}(\boldsymbol{x}; \boldsymbol{\eta}) \boldsymbol{B}(\boldsymbol{x}; \boldsymbol{\eta}) \Big[\boldsymbol{z} - \boldsymbol{\gamma}(\boldsymbol{x}; \boldsymbol{\eta}) \Big]. \tag{2.19}$$

2.4.3 Latent Variable Interpretation

Ordinal responses can be viewed as manifestations of unobservable latent variables. Consider first the *cumulative link models* of the form (2.16). Let L be a continuous latent variable associated with Y and we assume L depends linearly on $\boldsymbol{X}$, that is,

$$L = \beta_0 + \boldsymbol{\beta}' \boldsymbol{X} + \epsilon, \tag{2.20}$$

where ϵ is an error term with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$. Imagine L is coarsened to Y by a set of *cut-points* $\{\alpha_1, \ldots, \alpha_{J-1}\}$ in the following way:

$$\alpha_{j-1} < L \leq \alpha_j \quad \Longleftrightarrow \quad Y = j\,, \;\; j = 1, \ldots, J\,, \tag{2.21}$$

where $\alpha_0 = -\infty$ and $\alpha_J = \infty$. Note that for different values of (β_0, σ^2), we can shift or rescale α_j's to produce the same Y. Therefore, to make parameters identifiable, we restrict $\beta_0 = 0$ and $\sigma = 1$. It follows that

$$\omega_j(\boldsymbol{X}) = P(Y \leq j \mid \boldsymbol{X}) = P(L \leq \alpha_j \mid \boldsymbol{X}) = P(\epsilon \leq \alpha_j - \boldsymbol{\beta}'\boldsymbol{X}) = G(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x})\,,$$

where G is the cumulative density function (*c.d.f.*) of ϵ. We obtain model (2.16) by taking the inverse of G on both sides. The underlying process for the *continuation-ratio link models* involves a sequence of latent variables $\{L_1, \ldots, L_{R-1}\}$ which depend linearly on $\boldsymbol{X}$:

$$L_j = \boldsymbol{\beta}_j'\boldsymbol{X} + \epsilon_j\,, \qquad j = 1, \ldots, J-1, \tag{2.22}$$

where ϵ_j's are independent with mean 0 and *c.d.f.* G. The *cut-points* α_j's are associated with L_j's. Starting from L_1, if $L_1 \leq \alpha_1$, then $Y = 1$, otherwise move on to L_2. If $L_2 \leq \alpha_2$, then $Y = 2$, otherwise move on to L_3 and so on. In general, if $L_j \leq \alpha_j$, then $Y = j$, otherwise move on to L_{j+1}. It follows directly from this process that

$$\omega_j(\boldsymbol{X}) = P(Y = j \mid Y \geq j, \boldsymbol{X}) = P(L_j \leq \alpha_j \mid \boldsymbol{X}) = G(\alpha_j - \boldsymbol{\beta}_j'\boldsymbol{X})\,. \tag{2.23}$$

2.5 Regularity Conditions of Theorem 2.1

We require the following regularity conditions for the proof of the consistency of $\hat{\boldsymbol{\theta}}_n$:

S1. The parameter space Θ is compact;

S2. $\Psi(\boldsymbol{\theta}) = \mathbf{0}$ has a unique root;

S3. $\boldsymbol{U}(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta})$ is continuous at each $\boldsymbol{\theta} \in \Theta$ with probability one;

S4. There exists $\boldsymbol{H}(\boldsymbol{Y}, \boldsymbol{X})$ such that $|\boldsymbol{U}(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta})| \leq \boldsymbol{H}(\boldsymbol{Y}, \boldsymbol{X})$ for all $\boldsymbol{\theta}$, and $E[\boldsymbol{H}(\boldsymbol{Y}, \boldsymbol{X})] < \infty$.

In addition to S1-S4, the following conditions are also required to derive the equation (2.4):

S5. $\boldsymbol{\theta}_0$ is an interior point of Θ;

S6. $\boldsymbol{U}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$ for every $(\boldsymbol{y}, \boldsymbol{x})$;

S7. The second-order partial derivatives of $\boldsymbol{U}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta})$ satisfy

$$\left| \frac{\partial^2 \boldsymbol{U}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right| \leq \boldsymbol{U}_0(\boldsymbol{y}, \boldsymbol{x})$$

for some integrable function $\boldsymbol{U}_0(\boldsymbol{y}, \boldsymbol{x})$ for every $\boldsymbol{\theta}$ in a neighbourhood of $\boldsymbol{\theta}_0$;

S8. $\mathring{\Psi}(\boldsymbol{\theta}_0) = E\big[\partial \boldsymbol{U}(\boldsymbol{Y}, \boldsymbol{X}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}\big]\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ exists and is non-singular.

Chapter 3

Fractional Imputation for Univariate Incomplete Ordinal Variable

In this chapter, we consider a simple case where only one ordinal variable is subject to missingness. Under the notation introduced in Section 2.1, we have $T = 1$ and Y_1 is an ordinal response variable on a J_1-level scale. For simplicity of notation, we suppress the subscription "1" throughout this chapter. The data set under consideration is thus denoted by $\mathcal{O} = \big\{(r_i, y_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\big\}$. Our interest lies in creating one or several complete data sets for public use and investigating the validity and efficiency of subsequent estimators based on the synthetic data sets.

3.1 Existing Methods

3.1.1 Complete-case Analysis

In the single missing variable case, the complete case indicator δ_i is the same as the item response indicator r_i. Therefore, the CCA estimator $\hat{\boldsymbol{\theta}}_{cc}$ of the general parameter of interest defined in (2.2) is the solution to the following estimating equations:

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} r_i \boldsymbol{U}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta})\,. \tag{3.1}$$

Noting that

$$E\big[R\boldsymbol{U}(Y,\boldsymbol{X};\boldsymbol{\theta}_0)\big] = E\Big\{\pi(\boldsymbol{X})E\big[\boldsymbol{U}(Y,\boldsymbol{X};\boldsymbol{\theta}_0)\mid\boldsymbol{X}\big]\Big\},$$

we have, if $\pi(\boldsymbol{X}) = P(R=1\mid\boldsymbol{X})$ is a constant, that is, if the data is MCAR, the estimating function in (3.1) is unbiased and $\hat{\boldsymbol{\theta}}_{cc}$ is thereby consistent. If the data is strictly MAR, no clear conclusion can be drawn without knowing the form of $\boldsymbol{U}(y,\boldsymbol{x};\boldsymbol{\theta})$.

3.1.2 Inverse Probability Weighting

We focus on cases where the response probabilities $\pi(\boldsymbol{X}) = P(R=1\mid\boldsymbol{X})$ are unknown and require to be estimated. This can be done by imposing a parametric model $\pi(\boldsymbol{X}) = \pi(\boldsymbol{X};\boldsymbol{\phi})$ on the missing data process. Because $\big\{(r_i,\boldsymbol{x}_i),\ i=1,\ldots,n\big\}$ are fully observed, the parameter $\boldsymbol{\phi}$ can be estimated by $\hat{\boldsymbol{\phi}}$ using the maximum likelihood method. A common choice for $\pi(\boldsymbol{X};\boldsymbol{\phi})$ is the logistic regression model

$$\log\Big[\frac{\pi(\boldsymbol{X};\boldsymbol{\phi})}{1-\pi(\boldsymbol{X};\boldsymbol{\phi})}\Big] = a(\boldsymbol{X};\boldsymbol{\phi}), \tag{3.2}$$

where $a(\boldsymbol{X};\boldsymbol{\phi})$ belongs to a known family of functions parameterized by $\boldsymbol{\phi}$. The estimator $\hat{\boldsymbol{\phi}}$ is the solution to the score equations

$$\mathbf{0} = n^{-1}\sum_{i=1}^{n}\boldsymbol{T}(r_i,\boldsymbol{x}_i;\boldsymbol{\phi}), \tag{3.3}$$

where

$$\boldsymbol{T}(r,\boldsymbol{x};\boldsymbol{\phi}) = \frac{r-\pi(\boldsymbol{x};\boldsymbol{\phi})}{\pi(\boldsymbol{x};\boldsymbol{\phi})[1-\pi(\boldsymbol{x};\boldsymbol{\phi})]}\frac{\partial\pi(\boldsymbol{x};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}.$$

Estimation of parameter $\boldsymbol{\theta}$ then involves inverse probability weighting of the complete cases with the estimated response probabilities $\pi(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}})$. Specifically, an IPW estimator $\hat{\boldsymbol{\theta}}_{ipw}$ is obtained by solving:

$$\mathbf{0} = n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_{ipw}(r_i,y_i,\boldsymbol{x}_i;\boldsymbol{\theta},\hat{\boldsymbol{\phi}}), \tag{3.4}$$

where

$$\boldsymbol{U}_{ipw}(r, y, \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) = r\pi^{-1}(\boldsymbol{x}; \boldsymbol{\phi})\boldsymbol{U}(y, \boldsymbol{x}; \boldsymbol{\theta}). \tag{3.5}$$

Large sample properties of $\hat{\boldsymbol{\theta}}_{ipw}$ are well established, see for example, Rotnitzky and Robins (1997) and Rotnitzky et al. (1998), and are summarized as follows:

Theorem 3.1. *Given the notation and assumptions above, under the regularity conditions given in Section 3.6, $\hat{\boldsymbol{\theta}}_{ipw}$ is a consistent estimator of $\boldsymbol{\theta}$ and satisfies:*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{ipw} - \boldsymbol{\theta}_0) \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{ipw}), \tag{3.6}$$

where

$$\boldsymbol{\Sigma}_{ipw} = \boldsymbol{\Gamma}\left\{Var\big(\boldsymbol{U}_{ipw}\big) - \boldsymbol{A}\Big[Var\big(\boldsymbol{T}\big)\Big]^{-1}\boldsymbol{A}'\right\}\boldsymbol{\Gamma}',$$

and $\boldsymbol{\Gamma} = \big[-E\big(\partial\boldsymbol{U}/\partial\boldsymbol{\theta}'\big)\big]^{-1}$, $\boldsymbol{A} = -E\big(\partial\boldsymbol{U}_{ipw}/\partial\boldsymbol{\phi}'\big)$, all evaluated at the true parameter values $\boldsymbol{\theta}_0$, $\boldsymbol{\phi}_0$ and $\boldsymbol{U}$, $\boldsymbol{T}$ and $\boldsymbol{U}_{ipw}$ are short forms of the estimating functions defined in (2.2), (3.3) and (3.4).

3.1.3 Multiple Imputation

To characterize the data generating process, we assume the response Y depends on $\boldsymbol{X}$ through the model (2.15) and denote the conditional probability mass function by $f(y \mid \boldsymbol{X}; \boldsymbol{\eta})$. Let $\hat{\boldsymbol{\eta}}^{cc}$ be the complete-case estimator of $\boldsymbol{\eta}$ obtained by solving the estimating equations

$$\boldsymbol{0} = \sum_{i=1}^{n} r_i \, \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\eta}), \tag{3.7}$$

where $\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\eta})$ is the score function defined in (2.19). It is shown in Section 3.3.2 that $\hat{\boldsymbol{\eta}}^{cc}$ is a consistent estimator of $\boldsymbol{\eta}$. Let $\hat{\boldsymbol{\gamma}}_i = (\hat{\gamma}_{i1}, \ldots, \hat{\gamma}_{i(J-1)})' = \boldsymbol{\gamma}(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc})$ and $\hat{\gamma}_{i0} = 0$, $\hat{\gamma}_{iJ} = 1$ for all i. The conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}_i$ can be estimated by $f(y \mid \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc})$, which can be re-expressed as

$$P(Y = j \mid \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc}) = \hat{\gamma}_{ij} - \hat{\gamma}_{i(j-1)}, \quad j = 1, \ldots, J.$$

For unit i with $r_i = 0$, generate an imputed value $\tilde{y}_{il}$ from $f(y \mid \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc})$. The resulting imputed data set can be represented by $\{(r_i y_i + (1 - r_i)\tilde{y}_{il}, \boldsymbol{x}_i), i = 1, \ldots, n\}$. Do this independently for $l = 1, \ldots, M$ to create M imputed data files.

Each imputed data set is analyzed as if they are complete. Specifically, for the lth data file, an estimator $\hat{\boldsymbol{\theta}}^{(l)}$ is obtained by solving

$$\mathbf{0} = n^{-1}\sum_{i=1}^{n}\Big\{r_i\boldsymbol{U}(y_i,\boldsymbol{x}_i;\boldsymbol{\theta}) + (1-r_i)\boldsymbol{U}(\tilde{y}_{il},\boldsymbol{x}_i;\boldsymbol{\theta})\Big\}\,. \tag{3.8}$$

Finally, we take average of all these estimators and estimate $\boldsymbol{\theta}$ with

$$\hat{\boldsymbol{\theta}}_{mi} = M^{-1}\sum_{l=1}^{M}\hat{\boldsymbol{\theta}}^{(l)}. \tag{3.9}$$

The following theorem summarizes the asymptotic properties of $\hat{\boldsymbol{\theta}}_{mi}$, first proved by Robins and Wang (2000). We present the details of the proof in Section 3.6.

Theorem 3.2. *Let*

$$\boldsymbol{U}_{mi}(r,y,\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\eta}) = r\boldsymbol{U}(y,\boldsymbol{x},\boldsymbol{\theta}) + (1-r)M^{-1}\sum_{l=1}^{M}\boldsymbol{U}(\tilde{y}_l,\boldsymbol{x};\boldsymbol{\theta})\,, \tag{3.10}$$

where the $\tilde{y}_l$'s are independent draws from $f(y\mid\boldsymbol{x};\boldsymbol{\eta})$. Under the regularity conditions given in Section 3.6, $\hat{\boldsymbol{\theta}}_{mi}$ is a consistent estimator of $\boldsymbol{\theta}$ for any $M>0$. Furthermore,

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{mi}-\boldsymbol{\theta}_0) \sim \boldsymbol{N}(\mathbf{0},\boldsymbol{\Sigma}_{mi})\,, \tag{3.11}$$

where

$$\boldsymbol{\Sigma}_{mi} = \boldsymbol{\tau}Var\Big\{\boldsymbol{U}_{mi} + \boldsymbol{\kappa}I_{obs}^{-1}R\boldsymbol{S}\Big\}\boldsymbol{\tau}'\,,$$

and $\boldsymbol{\tau} = \big[-E\big(\partial\boldsymbol{U}/\partial\boldsymbol{\theta}'\big)\big]^{-1}$, $\boldsymbol{\kappa} = E\big[(1-R)\boldsymbol{U}\boldsymbol{S}'\big]$, $I_{obs} = -E\big[R\partial\boldsymbol{S}/\partial\boldsymbol{\eta}'\big]$, all evaluated at $\boldsymbol{\theta}_0$, $\boldsymbol{\eta}_0$ and $\boldsymbol{U}$, $\boldsymbol{U}_{mi}$, $\boldsymbol{S}$ are shortened expression for the functions defined in (2.2), (3.10) and (2.19).

3.2 Fully Efficient Fractional Imputation

3.2.1 Fractional Imputation Procedure

Let $\hat{\boldsymbol{\theta}}_{fi,e}$ be the solution to the following estimating equations:

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \left\{ r_i \boldsymbol{U}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) + (1 - r_i) M^{-1} \sum_{l=1}^{M} \boldsymbol{U}(\tilde{y}_{il}, \boldsymbol{x}_i; \boldsymbol{\theta}) \right\}, \tag{3.12}$$

then from the proof of Theorem 3.2, it can be shown that $\hat{\boldsymbol{\theta}}_{fi,e}$ is asymptotically equivalent to $\hat{\boldsymbol{\theta}}_{mi}$ in the sense that

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{fi,e} - \hat{\boldsymbol{\theta}}_{mi}) = o_p(1)\,.$$

See Section 3.6 for details. The estimator $\hat{\boldsymbol{\theta}}_{fi,e}$ is proposed by Fay (1996) and is known as the fractional imputation estimator with equal weights. From (3.12), we clearly see that each unit $(y_i, \boldsymbol{x}_i)$ with $r_i = 0$ is imputed by a cluster $\{(\tilde{y}_{il}, \boldsymbol{x}_i),\, l = 1, \ldots, M\}$ with members of the cluster receiving a fractional weight $w_{il} = 1/M$. The fractional weights should satisfy $\sum_{l=1}^{M} w_{il} = 1$, but they are not necessarily equal. Compared with multiple imputation, fractional imputation provides an extra degree of freedom so that we are able to impute the missing observations not only by choosing plausible values but also by assigning proper fractional weights.

Data files with missing ordinal responses are ideally suited for fractional imputation. Let $f(y \mid \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc})$ be the imputation model established in Section 3.1.3. Note that the variable Y takes J ordinal levels. Instead of drawing random samples from $f(y \mid \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc})$, a more efficient way of recovering the distributional structure of the missing y_i is to take all possible levels as imputed values and then to assign appropriate fractional weights. Specifically, we create a single complete data set by the following steps: if $r_i = 1$, unit i stays unchanged; otherwise, we replicate the unit J times and fill in J deterministic imputed values $\tilde{y}_{ij} = j$ for $j = 1, \ldots, J$, with the fractional weights given by

$$w_{ij} = P(Y = j \mid \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc}) = \hat{\gamma}_{ij} - \hat{\gamma}_{i(j-1)}\,, \quad j = 1, \ldots, J\,. \tag{3.13}$$

We will use the notation $w_{ij} = w_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc})$ to emphasize the dependance of w_{ij} on j,

Table 3.1: A Simple Example of a Fractionally Imputed Data Set with $J = 3$ and $n = 4$ (the column e_i^* indicates which original observation the unit corresponds to)

e_i^*	r_i^*	y_i^*	x_{i1}^*	x_{i2}^*	x_{i3}^*	i	w_i^*
1	1	y_1	x_{11}	x_{12}	x_{13}	1	1
2	0	1	x_{21}	x_{22}	x_{23}	2	w_{21}
2	0	2	x_{21}	x_{22}	x_{23}	3	w_{22}
2	0	3	x_{21}	x_{22}	x_{23}	4	w_{23}
3	1	y_3	x_{31}	x_{32}	x_{33}	5	1
4	0	1	x_{41}	x_{42}	x_{43}	6	w_{41}
4	0	2	x_{41}	x_{42}	x_{43}	7	w_{42}
4	0	3	x_{41}	x_{42}	x_{43}	8	w_{43}

$\boldsymbol{x}_i$ and $\hat{\boldsymbol{\eta}}^{cc}$. It is apparent that $\sum_{j=1}^{J} w_{ij} = 1$. The proposed fractional imputation procedure is fully efficient, in the sense that it does not introduce additional variations. Let the resulting data set be $\mathcal{O}^* = \left\{(r_i{}^*, y_i{}^*, \boldsymbol{x}_i{}^*, w_i{}^*),\ i = 1, \ldots, n^*\right\}$, where n^* is the size of the imputed data set. It is understood that for units with $r_i{}^* = 1$, $(y_i{}^*, \boldsymbol{x}_i{}^*)$ are the actually observed values with $w_i{}^* = 1$, while for units with $r_i{}^* = 0$, $(y_i{}^*, \boldsymbol{x}_i{}^*)$ comprises an imputed value for the missing Y and a duplicated value for the observed $\boldsymbol{X}$ with $w_i{}^*$ being the fractional weight. The column $\left\{r_i^*,\ i = 1, \ldots, n^*\right\}$ is only needed when we construct the complete data set and can be hidden for confidentiality concerns when the data is released for public use. Table 3.1 shows a simple example of a fractionally imputed data set, where the original data set involves an ordinal response with $J = 3$ levels and has $n = 4$ observations with missing responses in the second and fourth observation.

3.2.2 Point Estimation

The enlarged data file can be analyzed by standard tools with minor modifications to incorporate the weights. In fact, most existing packages allow users to specify a weight for each observation. The subsequent estimator $\hat{\boldsymbol{\theta}}_{fi}$ of $\boldsymbol{\theta}$ based on the fractionally

imputed data set can be obtained by solving:

$$\mathbf{0} = \Big\{ \sum_{i=1}^{n^*} {w_i}^* \Big\}^{-1} \sum_{i=1}^{n^*} {w_i}^* \boldsymbol{U}({y_i}^*, {\boldsymbol{x}_i}^*; \boldsymbol{\theta}) \,. \tag{3.14}$$

Or equivalently in the form of the "truly observed" data:

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \left\{ r_i \boldsymbol{U}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) + (1 - r_i) \sum_{j=1}^{J} w_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc}) \boldsymbol{U}(j, \boldsymbol{x}_i; \boldsymbol{\theta}) \right\} . \tag{3.15}$$

From the above definition, $\hat{\boldsymbol{\theta}}_{fi}$ is a "two-step" estimator based on $\hat{\boldsymbol{\eta}}^{cc}$. The asymptotic properties of $\hat{\boldsymbol{\theta}}_{fi}$ can be derived by viewing $(\hat{\boldsymbol{\theta}}_{fi}, \hat{\boldsymbol{\eta}}^{cc})$ as the solution to the following joint estimating equations:

$$\begin{aligned} \mathbf{0} &= n^{-1} \sum_{i=1}^{n} \boldsymbol{U}_{fi}(r_i, y_i, \boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\eta}) \,, \\ \mathbf{0} &= n^{-1} \sum_{i=1}^{n} r_i \, \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\eta}) \,, \end{aligned} \tag{3.16}$$

where

$$\boldsymbol{U}_{fi}(r, y, \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\eta}) = r \boldsymbol{U}(y, \boldsymbol{x}; \boldsymbol{\theta}) + (1 - r) \sum_{j=1}^{J} w_j(\boldsymbol{x}; \boldsymbol{\eta}) \boldsymbol{U}(r, \boldsymbol{x}; \boldsymbol{\theta}) \,,$$

and $\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\eta})$ is defined in (2.19). We show that both functions in the above estimating equations are unbiased. By the MAR assumption and the definition of $\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\eta})$

$$\begin{aligned} E\big[R\boldsymbol{S}(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\eta}_0)\big] &= E\Big\{ E(R \mid \boldsymbol{X}) E\big[\boldsymbol{S}(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\eta}_0) \mid \boldsymbol{X}\big] \Big\} \\ &= \mathbf{0} \,, \end{aligned}$$

and

$$\begin{aligned} E\big[\boldsymbol{U}_{fi}(R, Y, \boldsymbol{X}; \boldsymbol{\theta}, \boldsymbol{\eta}_0)\big] &= E\Big\{ E\big[\boldsymbol{U}_{fi}(R, Y, \boldsymbol{X}; \boldsymbol{\theta}, \boldsymbol{\eta}_0) \mid \boldsymbol{X}\big] \Big\} \\ &= E\Big\{ E(R \mid \boldsymbol{X}) E\big[\boldsymbol{U}(Y, \boldsymbol{X}; \boldsymbol{\theta}) \mid \boldsymbol{X}\big] \\ &\quad + \big[1 - E(R \mid \boldsymbol{X})\big] E\big[\boldsymbol{U}(Y, \boldsymbol{X}; \boldsymbol{\theta}) \mid \boldsymbol{X}\big] \Big\} \\ &= E\big[\boldsymbol{U}(Y, \boldsymbol{X}; \boldsymbol{\theta})\big] \,, \end{aligned} \tag{3.17}$$

therefore, $E\big[\boldsymbol{U}_{fi}(R, Y, \boldsymbol{X}; \boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\big] = \mathbf{0}$. By applying Theorem 2.1 to the joint estimating equations, the asymptotic properties of $\hat{\boldsymbol{\theta}}_{fi}$ are derived as follows. See Section 3.6 for the details of the proof.

Theorem 3.3. *Under the regularity conditions given in Section 3.6, the fractional imputation estimator $\hat{\boldsymbol{\theta}}_{fi}$ is consistent and satisfies*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{fi} - \boldsymbol{\theta}_0) \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\Sigma}_{fi})\,, \tag{3.18}$$

where

$$\boldsymbol{\Sigma}_{fi} = \boldsymbol{\tau} Var\Big\{\boldsymbol{U}_{fi} + \boldsymbol{\kappa} I_{obs}^{-1} R \boldsymbol{S}\Big\}\boldsymbol{\tau}'\,,$$

evaluated at $(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)$, $\boldsymbol{\tau}$, $\boldsymbol{\kappa}$ and I_{obs} are defined in Theorem 3.2.

To compare the efficiency of estimators based on fractional imputation and multiple imputation, we note that the middle term of $\boldsymbol{\Sigma}_{mi}$ in Theorem 3.2 can be decomposed by conditioning on the observed variable $(R, Y_{obs}, \boldsymbol{X})$:

$$\begin{aligned}
&Var\Big\{\boldsymbol{U}_{mi} + \boldsymbol{\kappa} I_{obs}^{-1} R \boldsymbol{S}\Big\} \\
&= Var\Big\{E\big[\boldsymbol{U}_{mi} + \boldsymbol{\kappa} I_{obs}^{-1} R \boldsymbol{S} \mid R, Y_{obs}, \boldsymbol{X}\big]\Big\} + E\Big\{Var\big[\boldsymbol{U}_{mi} + \boldsymbol{\kappa} I_{obs}^{-1} R \boldsymbol{S} \mid R, Y_{obs}, \boldsymbol{X}\big]\Big\} \\
&= Var\Big\{E\big[\boldsymbol{U}_{mi} \mid R, Y_{obs}, \boldsymbol{X}\big] + \boldsymbol{\kappa} I_{obs}^{-1} R \boldsymbol{S}\Big\} + E\Big\{Var\big[\boldsymbol{U}_{mi} \mid R, Y_{obs}, \boldsymbol{X}\big]\Big\},
\end{aligned}$$

where the second equation holds because $\boldsymbol{\kappa} I_{obs}^{-1} R \boldsymbol{S}$ is a function of $(R, Y_{obs}, \boldsymbol{X})$. Let $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ denote the first and second term in the above decomposition. When evaluated at $(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)$, it is easy to check that

$$E\big[\boldsymbol{U}_{mi} \mid R, Y_{obs}, \boldsymbol{X}\big] = \boldsymbol{U}_{fi}(R, Y, \boldsymbol{X}; \boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\,,$$

and it follows that $\boldsymbol{V}_1 = Var\Big\{\boldsymbol{U}_{fi} + \boldsymbol{\kappa} I_{obs}^{-1} R \boldsymbol{S}\Big\}$ is the middle term of $\boldsymbol{\Sigma}_{fi}$. A further look into $\boldsymbol{V}_2$ reveals that

$$\boldsymbol{V}_2 = M^{-1} E\Big\{Var\big[(1 - R)\boldsymbol{U} \mid R, Y_{obs}, \boldsymbol{X}\big]\Big\},$$

which is positive-definite for finite M and vanishes as $M \to \infty$.

Proposition 3.2.1. *Estimators based on the proposed fractional imputation is more efficient than the type-B multiple imputation estimators for any given M. When the*

number of imputations M goes to infinity, they are asymptotically equivalent.

The relation between the MI estimator $\hat{\boldsymbol{\theta}}_{mi}$ and the FI estimator $\hat{\boldsymbol{\theta}}_{fi}$ can also be established as follows. Note that $\hat{\boldsymbol{\theta}}_{mi}$ is asymptotically equivalent to $\hat{\boldsymbol{\theta}}_{fi,e}$ defined in (3.12). The second term in (3.12) is given by $(1-r_i)M^{-1}\sum_{l=1}^{k}\boldsymbol{U}(\tilde{y}_{il},\boldsymbol{x}_i;\boldsymbol{\theta})$, where the $\tilde{y}_{il}$'s are random samples drawn from $f(y \mid \boldsymbol{x}_i;\hat{\boldsymbol{\eta}}^{cc})$. It can be viewed as a sample mean of M observations and hence converges in probability to $(1-r_i)E[\boldsymbol{U}(Y,\boldsymbol{x}_i;\boldsymbol{\theta}) \mid \boldsymbol{x}_i;\hat{\boldsymbol{\eta}}^{cc}] = \sum_{j=1}^{J} w_j(\boldsymbol{x}_i;\hat{\boldsymbol{\eta}}^{cc})\boldsymbol{U}(j,\boldsymbol{x}_i;\boldsymbol{\theta})$ as $M \to \infty$. In other words, our proposed FI estimator corresponds to the MI estimator with $M = \infty$.

3.2.3 Variance Estimation

Following the idea in Robins and Wang (2000), we can estimate the variance of $\hat{\boldsymbol{\theta}}_{fi}$ by expanding the $\boldsymbol{\Sigma}_{fi}$ in Theorem 3.3 and then substituting sample moments based on the imputed data set for population moments. Note that

$$Var\Big\{\boldsymbol{U}_{fi}+\boldsymbol{\kappa}I_{obs}^{-1}R\boldsymbol{S}\Big\} = E\Big[\boldsymbol{U}_{fi}^{\otimes 2}\Big] + \boldsymbol{\kappa}I_{obs}^{-1}\boldsymbol{\kappa}' + E\Big[\boldsymbol{U}_{fi}R\boldsymbol{S}'\Big]I_{obs}^{-1}\boldsymbol{\kappa}' + \boldsymbol{\kappa}I_{obs}^{-1}E\Big[R\boldsymbol{S}\boldsymbol{U}_{fi}'\Big].$$

A consistent variance estimator for $\hat{\boldsymbol{\theta}}_{fi}$ is given by

$$\hat{\boldsymbol{V}}_A = n^{-1}\hat{\boldsymbol{\tau}}\Big\{\hat{\boldsymbol{\Sigma}}_1 + \hat{\boldsymbol{\kappa}}\hat{I}_{obs}^{-1}\hat{\boldsymbol{\kappa}}' + \hat{\boldsymbol{\Sigma}}_2\hat{I}_{obs}^{-1}\hat{\boldsymbol{\kappa}}' + \hat{\boldsymbol{\kappa}}\hat{I}_{obs}^{-1}\hat{\boldsymbol{\Sigma}}_2'\Big\}\hat{\boldsymbol{\tau}}',$$

where

$$\hat{\boldsymbol{\tau}} = \Big\{-n^{-1}\sum_{i=1}^{n^*} w_i{}^*\frac{\partial}{\partial\boldsymbol{\theta}'}\boldsymbol{U}(y_i{}^*,\boldsymbol{x}_i{}^*;\hat{\boldsymbol{\theta}}_{fi})\Big\}^{-1},$$

$$\hat{\boldsymbol{\kappa}} = n^{-1}\sum_{i=1}^{n^*}(1-r_i{}^*)w_i{}^*\boldsymbol{U}(y_i{}^*,\boldsymbol{x}_i{}^*;\hat{\boldsymbol{\theta}}_{fi})\boldsymbol{S}'(\boldsymbol{z}_i{}^*,\boldsymbol{x}_i{}^*;\hat{\boldsymbol{\eta}}^{cc}),$$

$$\hat{\boldsymbol{\Sigma}}_1 = n^{-1}\sum_{i=1}^{n^*}\big[w_i{}^*\boldsymbol{U}(y_i{}^*,\boldsymbol{x}_i{}^*;\hat{\boldsymbol{\theta}}_{fi})\big]^{\otimes 2}, \quad \hat{I}_{obs} = -n^{-1}\sum_{i=1}^{n^*} w_i{}^*r_i^*\frac{\partial}{\partial\boldsymbol{\eta}'}\boldsymbol{S}(\boldsymbol{z}_i{}^*,\boldsymbol{x}_i{}^*;\hat{\boldsymbol{\eta}}^{cc}),$$

$$\hat{\boldsymbol{\Sigma}}_2 = n^{-1}\sum_{i=1}^{n^*} w_i{}^*r_i^*\boldsymbol{U}(y_i{}^*,\boldsymbol{x}_i{}^*;\hat{\boldsymbol{\theta}}_{fi})\boldsymbol{S}'(\boldsymbol{z}_i{}^*,\boldsymbol{x}_i{}^*;\hat{\boldsymbol{\eta}}^{cc}),$$

and $\boldsymbol{z}_i^*$'s are the cumulative indicator vectors for $y_i{}^*$ in the imputed data set, similarly to $\boldsymbol{z}_i$ defined in Section 2.4.2. The $\hat{\boldsymbol{V}}_A$ is known as the linearization variance

estimators for imputed data sets (Kim and Rao 2009).

The linearization approach, however, is usually not appropriate for public use data files, because it typically requires full access to the information used for imputation, including the response indicators r_i which are often suppressed for confidentiality considerations, and all the covariates $\boldsymbol{x}_i$, even if the subsequent analysis only involves part of them (see Section 3.3.3). In addition, the linearization variance estimators cannot be conveniently calculated from the imputed data set with existing softwares.

In such cases, the resampling methods (Rao and Shao 1992; Efron 1994) become an attractive alternative, especially combined with the unique feature of the proposed fractional imputation, namely, each incomplete observation is imputed by the same set of units whether it is in the original data set or the resampled ones, and the resampling variation is reflected through the weights only. Consequently, variance estimation can be done through the use of additional columns of replication weights. We complete our proposed method by introducing the procedure for creating these replication weights, using the bootstrap method as an example. For ease of illustration, we add an index $e_i = i$ to the ith observation in the original data set, for $i = 1, \ldots, n$. These indices are either unchanged or replicated following the proposed procedure and are denoted by $\left\{e_i^*,\, i = 1, \ldots, n^*\right\}$ in the imputed data set (see Table 3.1).

1. Draw a bootstrap sample $\mathcal{O}^{(b)} = \left\{(e_i^{(b)}, r_i^{(b)}, y_i^{(b)}, \boldsymbol{x}_i^{(b)}),\, i = 1, \ldots, n\right\}$ **WITH** replacement from the original data, keeping all the missing values.

2. Treat the bootstrap sample as a real data set and apply the proposed fractional imputation procedure, that is, re-estimate the parameter $\boldsymbol{\eta}$ with observed units in $\mathcal{O}^{(b)}$, then calculate the fractional weights of imputed units with the updated estimates of $\boldsymbol{\eta}$. Let the resulting data set be denoted by $\mathcal{O}^{*(b)} = \left\{(e_i^{*(b)}, r_i^{*(b)}, y_i^{*(b)}, \boldsymbol{x}_i^{*(b)}, w_i^{*(b)}),\, i = 1, \ldots, n^{*(b)}\right\}$.

3. Re-express $\mathcal{O}^{*(b)}$ with units in $\mathcal{O}^*$ and a new set of weights $\left\{\tilde{w}_i^{*(b)},\, i = 1, \ldots, n^*\right\}$, where $\tilde{w}_i^{*(b)} = 0$, if the original observation indicated by e_i^* is not selected in $\mathcal{O}^{(b)}$; and $\tilde{w}_i^{*(b)} = \sum_{l \in S(e_i^*)} w_l^{*(b)}$ otherwise, where $S(e_i^*) = \left\{l \mid l \in \{1, \ldots, n^{*(b)}\},\, e_l^{*(b)} = e_i^* \text{ and } y_l^{*(b)} = y_i^*\right\}$.

4. Repeat STEP 1 - 3 B times and obtain B columns of replication weights $\left\{\tilde{w}_i^{*(b)},\, i = 1, \ldots, n^*\right\}$ for $b = 1, \ldots, B$.

5. Finally, the fractionally imputed data set with replication weights is given by

$$\bar{\mathcal{O}} = \left\{ (y_i{}^*, \boldsymbol{x}_i{}^*, w_i{}^*, \tilde{w}_i^{*(1)}, \ldots, \tilde{w}_i^{*(B)}),\ i = 1, \ldots, n^* \right\}.$$

Sometimes, the data set $\bar{\mathcal{O}}$ is only partially released for public access with some variables suppressed, usually for confidentiality considerations.

Note that STEP 3 is not applicable to multiple imputation, because for each bootstrap sample, the estimated imputation parameter $\hat{\boldsymbol{\eta}}^{cc}$ (or the posterior distribution if the Bayesian approach is used) has changed and the imputed values are re-generated and are typically different from the imputed values drawn based on the original sample.

With the provided data set, the data users first apply the desired complete-data analysis to units in $\bar{\mathcal{O}}$ weighted by $w_i{}^*$ to obtain a point estimator $\hat{\boldsymbol{\theta}}$, then simply repeat the same standard analysis B times with the replication weights $\tilde{w}_i^{*(b)}$ and obtain an estimate $\hat{\boldsymbol{\theta}}^{(b)}$ for $b = 1, \ldots, B$. The variance of $\hat{\boldsymbol{\theta}}$ can be estimated by

$$\hat{\boldsymbol{V}}_B = B^{-1} \sum_{b=1}^{B} (\hat{\boldsymbol{\theta}}^{(b)} - \bar{\boldsymbol{\theta}}), \tag{3.19}$$

where $\bar{\boldsymbol{\theta}} = B^{-1} \sum_{b=1}^{B} \hat{\boldsymbol{\theta}}^{(b)}$. Rigorous proof of the consistency of $\hat{\boldsymbol{V}}_B$ requires careful examination of complicated conditions, which is beyond the scope of this thesis. In Section 3.6, we sketch the key idea underlying the proof the validity of the bootstrap variance estimator without going too deep into the technical details. Replication weights have been widely used in survey sampling to facilitate inferences under complex designs and many softwares already have the ability to incorporate them automatically.

3.3 Subsequent Analyses by the Data Users

3.3.1 Estimation of the Category Probabilities

Let $\boldsymbol{p} = (p_1, \ldots, p_{J-1})'$ where $p_j = P(Y \leq j)$ are the unconditional cumulative probabilities. Let $p_0 = 0$ and $p_J = 1$. Estimation of the category probabilities, i.e.,

$P(Y = j)$ for $j = 1, \ldots, J$, is equivalent to the estimation of $\boldsymbol{p}$ since $P(Y = j) = p_j - p_{j-1}$. With the definition of the cumulative indicator vector $\boldsymbol{Z}$ from Section 2.4.2, we have $E(\boldsymbol{Z}) = \boldsymbol{p}$. In other words, we are interested in estimating the mean response of the indicator vector. Let $\boldsymbol{p}_0$ denote the true value of $\boldsymbol{p}$. In this case, the estimating function defining the parameter of interest is

$$\boldsymbol{U}_p(\boldsymbol{z}; \boldsymbol{p}) = \boldsymbol{z} - \boldsymbol{p}\,. \tag{3.20}$$

The CCA estimator

The CCA estimator $\hat{\boldsymbol{p}}^{cc}$ of $\boldsymbol{p}$ is defined as the solution to

$$\boldsymbol{0} = n^{-1} \sum_{i=1}^{n} r_i \boldsymbol{U}_p(\boldsymbol{z}_i; \boldsymbol{p}) = n^{-1} \sum_{i=1}^{n} r_i (\boldsymbol{z}_i - \boldsymbol{p})\,,$$

and is given by

$$\hat{\boldsymbol{p}}^{cc} = \frac{\sum_{i=1}^{n} r_i \boldsymbol{z}_i}{\sum_{i=1}^{n} r_i}\,.$$

Noting that

$$E\big[R(\boldsymbol{Z} - \boldsymbol{p}_0)\big] = E\big\{\pi(\boldsymbol{X})[E(\boldsymbol{Z} \mid \boldsymbol{X}) - \boldsymbol{p}_0]\big\}\,,$$

we have $E[R\boldsymbol{U}_p(\boldsymbol{Z}; \boldsymbol{p}_0)] \neq \boldsymbol{0}$ unless $\pi(\boldsymbol{X})$ is a constant or $E(\boldsymbol{Z} \mid \boldsymbol{X}) = E(\boldsymbol{Z})$. In other words, the CCA estimator $\hat{\boldsymbol{p}}^{cc}$ is not consistent for $\boldsymbol{p}$ under the MAR assumption unless the responses are MCAR or the response variable is independent of the covariates.

The IPW estimator

The IPW estimator $\hat{\boldsymbol{p}}^{ipw}$ of $\boldsymbol{p}$ is defined as the solution to

$$\boldsymbol{0} = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}_{ipw}(r_i, \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{p}, \hat{\boldsymbol{\phi}})\,, \tag{3.21}$$

where $\boldsymbol{U}_{ipw}(r, \boldsymbol{z}, \boldsymbol{x}; \boldsymbol{p}, \boldsymbol{\phi}) = r\pi^{-1}(\boldsymbol{x}; \boldsymbol{\phi})\boldsymbol{U}_p(\boldsymbol{z}; \boldsymbol{p})$ and $\hat{\boldsymbol{\phi}}$ is the solution to (3.3). The estimator is given by

$$\hat{\boldsymbol{p}}^{ipw} = \left[\sum_{i=1}^{n} \frac{r_i}{\pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})}\right]^{-1} \left[\sum_{i=1}^{n} \frac{r_i \boldsymbol{z}_i}{\pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})}\right]\,.$$

Asymptotic results for $\hat{\boldsymbol{p}}^{ipw}$ follow directly from Theorem 3.1.

The MI estimator

Let $\{(r_i y_i + (1-r_i)\tilde{y}_{il}, \boldsymbol{x}_i), i = 1, \ldots, n\}$, $l = 1, \ldots, M$ be the M imputed data files, where $\tilde{y}_{il}$ is generated from $f(y \mid \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc})$ as described in Section 3.1.3. Let $\boldsymbol{z}_i$ and $\tilde{\boldsymbol{z}}_{il}$ be the cumulative indicators for y_i and $\tilde{y}_{il}$, respectively. The MI estimator of $\boldsymbol{p}$ is computed as

$$\hat{\boldsymbol{p}}^{mi} = M^{-1} \sum_{l=1}^{M} \hat{\boldsymbol{p}}^{l} \,,$$

where $\hat{\boldsymbol{p}}^{l}$ is the solution to the estimating equations

$$\boldsymbol{0} = n^{-1} \sum_{i=1}^{n} \Big\{ r_i \boldsymbol{U}_p(\boldsymbol{z}_i; \boldsymbol{p}) + (1-r_i) \boldsymbol{U}_p(\tilde{\boldsymbol{z}}_{il}; \boldsymbol{p}) \Big\} \,. \tag{3.22}$$

The MI estimator can be alternatively written as

$$\hat{\boldsymbol{p}}^{mi} = n^{-1} \sum_{i=1}^{n} \left[r_i \boldsymbol{z}_i + (1-r_i) M^{-1} \sum_{l=1}^{M} \tilde{\boldsymbol{z}}_{il} \right] \,,$$

which is exactly the same as the fractional imputation estimator with equal weights defined in (3.12).

The FI estimator

The FI estimator $\hat{\boldsymbol{p}}^{fi}$ of $\boldsymbol{p}$ based on the proposed procedure is the solution to

$$\boldsymbol{0} = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}_{fi}(r_i, \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{p}, \hat{\boldsymbol{\eta}}^{cc}) \,, \tag{3.23}$$

where $\boldsymbol{U}_{fi}(r, \boldsymbol{z}, \boldsymbol{x}; \boldsymbol{p}, \boldsymbol{\eta}) = r \boldsymbol{U}_p(\boldsymbol{z}; \boldsymbol{p}) + (1-r) \sum_{j=1}^{J} w_j(\boldsymbol{x}; \boldsymbol{\eta}) \boldsymbol{U}_p(\boldsymbol{c}_j; \boldsymbol{p})$ and $\boldsymbol{c}_j$ is the cumulative indicator of level j for $j = 1, \ldots, J$. The estimator can be written as

$$\hat{\boldsymbol{p}}^{fi} = n^{-1} \sum_{i=1}^{n} \left[r_i \boldsymbol{z}_i + (1-r_i) \sum_{j=1}^{J} w_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc}) \boldsymbol{c}_j \right] .$$

Proposition 3.3.1. *Suppose that the response probability model (3.2) and the imputation model (2.15) are correctly specified and assume that the responses are missing-*

at-random. For estimating the mean responses,

(1) The CCA estimator is not consistent unless the responses are missing completely at random (MCAR) or the response variable is independent of all the covariates, whereas the IPW estimator, the MI estimator and the FI estimator are all consistent.

(2) The FI estimator is equivalent to the MI estimator with $M = +\infty$ and hence is more efficient than the MI estimator for a finite M.

There is no clear-cut comparison in efficiency between the IPW estimator and the FI estimator, since the two estimators involve two different models: the MDP model and the DGP model. Our limited simulation results presented in Section 3.5 seem to indicate that the FI estimator has better efficiency.

3.3.2 Regression Analysis: The First Scenario

We now turn our attention to regression analysis where the objective of the data user is to establish associations between the response variable Y and a set of covariates $\boldsymbol{V}$. In this section we consider the first scenario where $\boldsymbol{V} = \boldsymbol{X}$, i.e., all covariates used for imputation are included in the subsequent regression analysis. This is often the case when no sensitive or confidential information is involved in the imputation process so that the whole data set is accessible to the data users and the user is interested in the relations between the response and all the covariates, for example, when one conducts initial exploration of the data file to have an overview of the dependence structure.

Assume the data user imposes a model of form (2.15) on the response against $\boldsymbol{X}$, often called *the analysis model* and we denote the parameters in the model by $\boldsymbol{\theta}$ to distinguish them from the $\boldsymbol{\eta}$ in the model used by the data file creator (See Section 3.1.3), which is known as *the imputation model*. For these two models to be "compatible", in the current scenario, they are essentially the same, that is, they share a common score function $\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\eta})$ defined in (2.19). We will elaborate on the "model compatibility" in Section 3.4. Let $\boldsymbol{\eta}_0$ and $\boldsymbol{\theta}_0$ be the true values of the parameters. Following the above discussion, $\boldsymbol{\eta}_0 = \boldsymbol{\theta}_0$ and the estimating function

defining the parameters of interest is

$$\boldsymbol{U}^{(1)}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta}), \tag{3.24}$$

where $\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta})$ is defined in (2.19).

The CCA estimator

The CCA estimator $\hat{\boldsymbol{\theta}}^{cc}$ of $\boldsymbol{\theta}$ is the solution to

$$n^{-1} \sum_{i=1}^{n} r_i \, \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \boldsymbol{0}\,, \tag{3.25}$$

where

$$\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{D}(\boldsymbol{x}; \boldsymbol{\theta}) \boldsymbol{B}(\boldsymbol{x}; \boldsymbol{\theta}) \big\{ \boldsymbol{z} - \boldsymbol{\gamma}(\boldsymbol{x}; \boldsymbol{\theta}) \big\}\,.$$

Note that $\hat{\boldsymbol{\theta}}^{cc}$ is the same as the $\hat{\boldsymbol{\eta}}^{cc}$ in Section 3.1.3, which is used for multiple and fractional imputation. It turns out that the CCA estimator for the regression coefficients is a valid and efficient estimator under the MAR assumption. In fact, if the separability condition discussed in Molenberghs and Kenward (2007) holds, $\hat{\boldsymbol{\theta}}^{cc}$ is the maximum observed likelihood estimator. The consistency of the estimator follows from

$$E\Big[R\boldsymbol{S}(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\theta}_0)\Big] = E\Big[\pi(\boldsymbol{X}) \boldsymbol{D}(\boldsymbol{X}; \boldsymbol{\theta}_0) \boldsymbol{B}(\boldsymbol{X}; \boldsymbol{\theta}_0) \big\{ E(\boldsymbol{Z} \mid \boldsymbol{X}) - \boldsymbol{\gamma}(\boldsymbol{X}; \boldsymbol{\theta}_0) \big\}\Big] = \boldsymbol{0}\,.$$

Asymptotic variance can be derived by Theorem 2.1. Efficiency comparisons between $\hat{\boldsymbol{\theta}}^{cc}$ and other alternative estimators are given in Proposition 3.3.2.

The IPW estimator

Under the MAR assumption, it is mandatory to use all available covariates $\boldsymbol{X}$ for the response probability model such as (3.2). With $\hat{\boldsymbol{\phi}}$ obtained by solving (3.3) the IPW estimator $\hat{\boldsymbol{\theta}}^{ipw}$ of $\boldsymbol{\theta}$ is the solution to

$$\boldsymbol{0} = \sum_{i=1}^{n} \boldsymbol{U}^{(1)}_{ipw}(r_i, \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\theta}, \hat{\boldsymbol{\phi}})\,, \tag{3.26}$$

where

$$\boldsymbol{U}^{(1)}_{ipw}(r, \boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{r}{\pi(\boldsymbol{x}; \boldsymbol{\phi})} \boldsymbol{U}^{(1)}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta}) = \frac{r}{\pi(\boldsymbol{x}; \boldsymbol{\phi})} \boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta})\,,$$

The IPW estimator $\hat{\boldsymbol{\theta}}^{ipw}$ is still consistent, however, it is less efficient than the CCA estimator as shown in Proposition 3.3.2.

The MI estimator

The multiple imputation estimator of $\boldsymbol{\theta}$ with M imputed data sets is computed as

$$\hat{\boldsymbol{\theta}}^{mi} = M^{-1} \sum_{l=1}^{M} \hat{\boldsymbol{\theta}}_l,$$

where $\hat{\boldsymbol{\theta}}_l$ is the solution to

$$n^{-1} \sum_{i=1}^{n} \Big\{ r_i \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) + (1 - r_i) \boldsymbol{S}(\tilde{\boldsymbol{z}}_{il}, \boldsymbol{x}_i; \boldsymbol{\theta}) \Big\} = \mathbf{0}, \tag{3.27}$$

and $\tilde{\boldsymbol{z}}_{il}$'s are the derived cumulative indicators for the random draws $\tilde{y}_{il}$'s from $f(y \mid \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc})$. The asymptotic variance presented in Proposition 3.3.2 is based on Theorem 3.2 and properties of the score function $\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta})$.

The FI estimator

The FI estimator $\hat{\boldsymbol{\theta}}^{fi}$ of $\boldsymbol{\theta}$ is the solution to

$$n^{-1} \sum_{i=1}^{n} \boldsymbol{U}_{fi}^{(1)}(r_i, \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}^{cc}) = \mathbf{0}\,, \tag{3.28}$$

where

$$\boldsymbol{U}_{fi}^{(1)}(r, \boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\eta}) = r \boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\theta}) + (1 - r) \sum_{j=1}^{J} w_j(\boldsymbol{x}; \boldsymbol{\eta}) \boldsymbol{S}(\boldsymbol{c}_j, \boldsymbol{x}; \boldsymbol{\theta})\,,$$

and $w_j(\boldsymbol{x}; \boldsymbol{\eta})$ is defined after equation (3.13). The asymptotic variance can be derived from Theorem 3.3. Alternatively, noting that

$$\begin{aligned} \sum_{j=1}^{J} w_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc}) \boldsymbol{S}(\boldsymbol{c}_j, \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}^{cc}) &= \boldsymbol{D}(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}^{cc}) \boldsymbol{B}(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}^{cc}) \Big\{ \sum_{j=1}^{J} w_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc}) \boldsymbol{c}_j - \boldsymbol{\gamma}(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}^{cc}) \Big\} \\ &= \mathbf{0}, \end{aligned}$$

since $\hat{\boldsymbol{\eta}}^{cc} = \hat{\boldsymbol{\theta}}^{cc}$, we have

$$n^{-1}\sum_{i=1}^{n} \boldsymbol{U}_{fi}^{(1)}(r_i, \boldsymbol{z}_i, \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}^{cc}, \hat{\boldsymbol{\eta}}^{cc}) = n^{-1}\sum_{i=1}^{n} r_i \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}^{cc}) = \mathbf{0},$$

which implies that the estimator based on the fractionally imputed data set is the same as the CCA estimator.

Proposition 3.3.2. *Suppose that the response probability model (3.2) is correctly specified and the responses are missing-at-random. Suppose also that the imputation procedure and the main analysis are based on the same correct regression model. For estimating the regression coefficients $\boldsymbol{\theta}$ in the analysis model,*

(1) The CCA estimator, the IPW estimator, the MI estimator and the FI estimator are all consistent.

(2) The CCA estimator and the FI estimator are equivalent and hence are equally efficient. Both are generally more efficient than the IPW estimator and the MI estimator with a finite M.

(3) When $M \to \infty$, the MI estimator becomes equivalent to the CCA estimator and the FI estimator. The IPW estimator and the CCA estimator are equivalent under MCAR.

(4) The asymptotic variances (AV) of the CCA, IPW, MI and FI estimators are given respectively by

$$\begin{aligned}
AV\big(\hat{\boldsymbol{\theta}}^{cc}\big) &= AV\big(\hat{\boldsymbol{\theta}}^{fi}\big) = n^{-1} I_{obs}^{-1}\,, \\
AV\big(\hat{\boldsymbol{\theta}}^{ipw}\big) &= n^{-1} I_{com}^{-1} E\Big[\pi^{-1}(\boldsymbol{X}; \boldsymbol{\phi}_0)\boldsymbol{S}(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\theta}_0)^{\otimes 2}\Big] I_{com}^{-1}\,, \\
AV\big(\hat{\boldsymbol{\theta}}^{mi}\big) &= n^{-1}\left\{ I_{obs}^{-1} + M^{-1} I_{com}^{-1}\Big[I_{com} - I_{obs}\Big] I_{com}^{-1}\right\}\,,
\end{aligned}$$

where $I_{com} = E\big[-\partial \boldsymbol{S}(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}'\big]$ and $I_{obs} = E\big[R\partial \boldsymbol{S}(\boldsymbol{Z}, \boldsymbol{X}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}'\big]$, both evaluated at $\boldsymbol{\theta}_0$, are often respectively referred to as the complete and the observed information matrix.

Proof. Results on consistency follow from discussions above. The asymptotic variance formulas can be derived by applying theorems under the general settings. For

efficiency comparisons, we first note that, by the alternative representation of the information matrices,

$$
\begin{aligned}
I_{com} = E\big[\boldsymbol{S}(\boldsymbol{\theta}_0)^{\otimes 2}\big] &= Cov\Big\{\pi^{1/2}(\boldsymbol{X};\boldsymbol{\phi}_0)\boldsymbol{S}(\boldsymbol{\theta}_0), \pi^{-1/2}(\boldsymbol{X};\boldsymbol{\phi}_0)\boldsymbol{S}(\boldsymbol{\theta}_0)\Big\}\,, \\
I_{obs} = E\big[R\boldsymbol{S}(\boldsymbol{\theta}_0)^{\otimes 2}\big] &= Var\Big\{\pi^{1/2}(\boldsymbol{X};\boldsymbol{\phi}_0)\boldsymbol{S}(\boldsymbol{\theta}_0)\Big\}\,, \\
E\big[\pi^{-1}(\boldsymbol{X};\boldsymbol{\phi}_0)\boldsymbol{S}(\boldsymbol{\theta}_0)^{\otimes 2}\big] &= Var\Big\{\pi^{-1/2}(\boldsymbol{X};\boldsymbol{\phi}_0)\boldsymbol{S}(\boldsymbol{\theta}_0)\Big\}\,,
\end{aligned}
$$

where $\boldsymbol{S}(\boldsymbol{\theta}_0)$ is the abbreviated expression for $\boldsymbol{S}(\boldsymbol{z},\boldsymbol{x};\boldsymbol{\theta}_0)$. By applying the matrix form of the Cauchy-Schwarz Inequality (Tripathi 1999), we have

$$
I_{obs} \geq I_{com}\Big\{E\big[\pi^{-1}(\boldsymbol{X};\boldsymbol{\phi}_0)\boldsymbol{S}(\boldsymbol{\theta}_0)^{\otimes 2}\big]\Big\}^{-1} I_{com}\,,
$$

The equation holds if and only if $\pi(\boldsymbol{X};\boldsymbol{\phi}_0)$ is a constant, i.e., the responses are MCAR. It follows that the CCA estimator is more efficient than the IPW estimator under general MAR. The MI estimator is obviously less efficient than the CCA and the FI estimators, since the second term in $AV\big(\hat{\boldsymbol{\theta}}^{mi}\big)$ is positive definite for any finite M. □

3.3.3 Regression Analysis: The Second Scenario

We now consider a practically important scenario in regression analysis, when the data user only includes in the analysis model a subset of covariates used for imputation. This could be the case, for instance, when the user has a specific scientific objective which requires exploration on how the responses are associated with specific covariates, or when the data file creator uses some confidential information in the imputation procedure and that information is concealed thereafter and hence inaccessible to the user in the public data file.

Let $\boldsymbol{X} = (\boldsymbol{V}', \boldsymbol{S}')'$, where $\boldsymbol{X}$ are the covariates in the imputation model and $\boldsymbol{V}$ are the covariates in the analysis model. Both models are assumed to follow the general form (2.15). To avoid any confusion on notation, we let $\boldsymbol{\theta}$ be the parameters in the analysis model and $\boldsymbol{D}$, $\boldsymbol{B}$, $\boldsymbol{\gamma}$ and $\boldsymbol{S}$ be the corresponding functions defined in (2.18) and (2.19). The parameters in the imputation model are denoted by $\boldsymbol{\eta}$ and the corresponding functions by $\boldsymbol{D}^*$, $\boldsymbol{B}^*$, $\boldsymbol{\gamma}^*$ and $\boldsymbol{S}^*$. The focus is on statistical inferences for $\boldsymbol{\theta}$. Let $\boldsymbol{\theta}_0$ be the true value of the parameters $\boldsymbol{\theta}$. In this case, the estimating

function defining the parameters of interest is

$$\boldsymbol{U}^{(2)}(\boldsymbol{z}, \boldsymbol{v}; \boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{z}, \boldsymbol{v}; \boldsymbol{\theta}) \,. \tag{3.29}$$

The CCA estimator

The CCA estimator $\hat{\boldsymbol{\theta}}^{cc}$ is the solution to the estimating equations

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} r_i \, \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{v}_i; \boldsymbol{\theta}) \,,$$

where

$$\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{v}; \boldsymbol{\theta}) = \boldsymbol{D}(\boldsymbol{v}; \boldsymbol{\theta}) \boldsymbol{B}(\boldsymbol{v}; \boldsymbol{\theta}) \big\{ \boldsymbol{z} - \boldsymbol{\gamma}(\boldsymbol{v}; \boldsymbol{\theta}) \big\} \,, \tag{3.30}$$

with the forms of $\boldsymbol{D}(\boldsymbol{v}; \boldsymbol{\theta})$ and $\boldsymbol{B}(\boldsymbol{v}; \boldsymbol{\theta})$ similarly specified in (2.18). To see the validity of the CCA estimator, we note that, by the MAR assumption, the nonresponse indicator variable R and the cumulative indicator variable $\boldsymbol{Z}$ derived from the response are conditionally independent given ALL the $\boldsymbol{X}$ variables. It follows that

$$E\Big[R \, \boldsymbol{S}(\boldsymbol{Z}, \boldsymbol{V}; \boldsymbol{\theta}_0) \Big] = E\Big[\pi(\boldsymbol{X}) \boldsymbol{D}(\boldsymbol{V}; \boldsymbol{\theta}_0) \boldsymbol{B}(\boldsymbol{V}; \boldsymbol{\theta}_0) \big\{ E\big(\boldsymbol{Z} \mid \boldsymbol{X}\big) - E\big(\boldsymbol{Z} \mid \boldsymbol{V}\big) \big\} \Big] \,.$$

In general, $E[R \, \boldsymbol{S}(\boldsymbol{Z}, \boldsymbol{V}; \boldsymbol{\theta}_0)] \neq \mathbf{0}$ unless the response variable and the not-in-the-model covariates $\boldsymbol{S}$ are independent given $\boldsymbol{X}$. In the latter case we have $E\big(\boldsymbol{Z} \mid \boldsymbol{X}\big) = E\big(\boldsymbol{Z} \mid \boldsymbol{V}\big)$ and $E[R\boldsymbol{S}(\boldsymbol{Z}, \boldsymbol{V}; \boldsymbol{\theta}_0)] = \mathbf{0}$. In fact, if this is the case, the problem reduces to the first scenario. Otherwise, the CCA estimator for the regression coefficients is invalid under the current setting.

The IPW estimator

The IPW estimator $\hat{\boldsymbol{\theta}}^{ipw}$ of $\boldsymbol{\theta}$ is the solution to

$$\mathbf{0} = \sum_{i=1}^{n} \boldsymbol{U}^{(2)}_{psa}(r_i, \boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{v}_i; \boldsymbol{\theta}, \hat{\boldsymbol{\phi}}) \,, \tag{3.31}$$

where

$$\boldsymbol{U}^{(2)}_{psa}(r, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{v}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{r}{\pi(\boldsymbol{x}; \boldsymbol{\phi})} \boldsymbol{S}(\boldsymbol{z}, \boldsymbol{v}; \boldsymbol{\theta}) \,.$$

The estimate $\hat{\boldsymbol{\phi}}$ is the solution to the estimating equations specified by (3.3). The data

file creator uses all the covariates $\boldsymbol{X}$ when creating the inverse-probability weights as required by the MAR assumption. Asymptotic results are derived from Theorem 3.1.

The MI estimator

The MI estimator $\hat{\boldsymbol{\theta}}^{mi}$ is computed as

$$\hat{\boldsymbol{\theta}}^{mi} = M^{-1} \sum_{l=1}^{M} \hat{\boldsymbol{\theta}}_l \,,$$

where $\hat{\boldsymbol{\theta}}_l$ is the solution to

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \Big\{ r_i \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{v}_i; \boldsymbol{\theta}) + (1 - r_i) \boldsymbol{S}(\tilde{\boldsymbol{z}}_{il}, \boldsymbol{v}_i; \boldsymbol{\theta}) \Big\} \,, \tag{3.32}$$

and $\tilde{\boldsymbol{z}}_{il}$'s are the derived cumulative indicators for the random draws $\tilde{y}_{il}$'s from $f(y \mid \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc})$. Here $\hat{\boldsymbol{\eta}}^{cc}$ solves the following estimating equations based on the imputation model

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} r_i \boldsymbol{S}^*(\boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\eta}) \,. \tag{3.33}$$

The FI estimator

The FI estimator $\hat{\boldsymbol{\theta}}^{fi}$ is defined as the solution to

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}_{fi}^{(2)}(r_i, \boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{v}_i; \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}^{cc}) \,, \tag{3.34}$$

where

$$\boldsymbol{U}_{fi}^{(2)}(r, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{v}; \boldsymbol{\theta}, \boldsymbol{\eta}) = r \boldsymbol{S}(\boldsymbol{z}, \boldsymbol{v}; \boldsymbol{\theta}) + (1 - r) \sum_{j=1}^{J} w_j(\boldsymbol{x}; \boldsymbol{\eta}) \boldsymbol{S}(\boldsymbol{c}_j, \boldsymbol{v}; \boldsymbol{\theta}) \,,$$

with $\hat{\boldsymbol{\eta}}^{cc}$ obtained from (3.33) and

$$w_j(\boldsymbol{x}; \boldsymbol{\eta}) = \gamma_j^*(\boldsymbol{x}; \boldsymbol{\eta}) - \gamma_{j-1}^*(\boldsymbol{x}; \boldsymbol{\eta})$$

being the fractional weights based on the imputation model involving $\boldsymbol{X}$.

Proposition 3.3.3. *Suppose that the response probability model and the imputation model are built using all available covariates, and assume that the responses are missing at random. For estimating the regression coefficients* $\boldsymbol{\theta}$ *in the analysis model involving a set of selected covariates,*

(1) The CCA estimator is inconsistent unless the response variable is independent of covariates not included in the analysis model given those in the model.

(2) The IPW estimator, the MI estimator and the FI estimator are all consistent under the assumed response probability model and the imputation model.

(3) The FI estimator is equivalent to the MI estimator when $M = +\infty$ *and hence is more efficient than the MI estimator for a fixed* M *.*

3.4 A Discussion on Model Compatibility

When the subsequent analysis conducted by the data user is model-free, for example, the estimation of the category probabilities or nonparametric regression, the validity of the analysis only relies on the correct specification of the imputation model. However, when the data user also posits an analysis model on the data, the "compatibility" issue would arise. Specifically, the analysis model and imputation model have to be correct simultaneously, which cannot always be taken as granted because they are imposed by two disconnected entities but are internally related. We give an example of compatible analysis and imputation models under the second scenario of regression analysis.

Suppose the data file creator assumes a *cumulative link model* of the form (2.16) with the *probit* link function for response Y against $\boldsymbol{X}$. Then by the latent variable interpretation discussed in Section 2.4.3, there exits a latent variable L such that

$$L \mid \boldsymbol{X} \sim N(0, \boldsymbol{\beta}'\boldsymbol{X}) \,.$$

Noting that $\boldsymbol{X} = (\boldsymbol{V}', \boldsymbol{S}')'$, we re-write the above expression as

$$L \mid \boldsymbol{V}, \boldsymbol{S} \sim N(0, \boldsymbol{\beta}_1'\boldsymbol{V} + \boldsymbol{\beta}_2'\boldsymbol{S}) \,,$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are components of $\boldsymbol{\beta}$ corresponding to $\boldsymbol{V}$ and $\boldsymbol{S}$. We further assume that $\boldsymbol{S}$ depends linearly on $\boldsymbol{V}$ through the model

$$\boldsymbol{S} \mid \boldsymbol{V} \sim \boldsymbol{N}(\boldsymbol{\nu}_0 + \boldsymbol{\nu}_1' \boldsymbol{V}, \boldsymbol{\Sigma}) .$$

It can be shown that

$$L \mid \boldsymbol{V} \sim N(\boldsymbol{\beta}_2' \boldsymbol{\nu}_0 + \boldsymbol{\beta}_1' \boldsymbol{V} + \boldsymbol{\beta}_2' \boldsymbol{\nu}_1' \boldsymbol{V}, 1 + \boldsymbol{\beta}_2' \boldsymbol{\Sigma} \boldsymbol{\beta}_2) .$$

This implies that if the data user also imposes a *cumulation link model* defined in (2.16) with the *probit* link function on Y against $\boldsymbol{V}$, the analysis model and the imputation model are compatible.

Compatibility between the imputation procedure and the subsequent analyses is a common issue for all imputation-based approaches, including multiple imputation and fractional imputation. For data creators, it is preferable to build a flexible imputation model involving all available information and possible interaction terms. Model selection tools can be used if necessary. For data users, it is essential to perform some preliminary goodness-of-fit analysis with the imputed data set to choose a plausible analysis model that fits the data well. As an additional note, the "congeniality" condition proposed by Meng (1994) for Rubin's variance estimator to be valid is much stronger than the "compatibility" we discussed here. In the first place, it also requires the analysis model to be compatible with the imputation model, and moreover, it places restrictions on the estimation techniques used to fit the analysis model.

3.5 Simulation Studies

We conduct three simulation studies to evaluate the finite sample performance of the estimators, corresponding to the three inferential problems we discussed in Section 3.3.

In the first study, we consider an ordinal response variable Y with three categories and two covariates: a continuous $X_1 \sim \text{Exp}(1)$ and a discrete $X_2 \sim \text{Bernoulli}(0.5)$. The ordinal response Y follows the *cumulative link model* of the form (2.16) and the response probability follows a logistic model. The model parameters $\boldsymbol{\phi}$ in the response probability model are chosen so that the average response rate $E(R)$ is at a designated

Table 3.2: Absolute Relative Bias (%) and Root Mean Squared Error ($\times 10^{-2}$) of Different Estimators of $p_1 = P(Y \leq 1)$

RR	SS		COMP	CCA	IPW	MI(1)	MI(5)	MI(10)	FI
85%	500	ARB	0.1	12.7	0.0	0.1	0.1	0.1	0.1
		RMSE	(1.9)	(3.7)	(2.2)	(2.3)	(2.2)	(2.2)	(2.2)
	200	ARB	0.3	12.4	0.2	0.5	0.5	0.4	0.4
		RMSE	(3.0)	(4.5)	(3.3)	(3.5)	(3.4)	(3.3)	(3.3)
5%	500	ARB	—	7.2	0.1	0.1	0.2	0.2	0.2
		RMSE	—	(2.9)	(2.3)	(2.4)	(2.3)	(2.2)	(2.2)
	200	ARB	—	7.6	0.3	0.5	0.4	0.4	0.4
		RMSE	—	(4.0)	(3.5)	(3.6)	(3.5)	(3.4)	(3.4)
50%	500	ARB	—	17.8	0.4	0.3	0.1	0.1	0.2
		RMSE	—	(5.4)	(2.7)	(2.5)	(2.4)	(2.4)	(2.4)
	200	ARB	—	17.5	0.7	0.4	0.3	0.3	0.2
		RMSE	—	(6.4)	(4.1)	(4.0)	(3.7)	(3.7)	(3.6)

level. For the MI estimator, we include results for $M = 1, 5$ and 10. The simulated absolute relative bias (ARB, in %) and the root mean square error (RMSE, multiplied by 10^2) for three levels of $E(R)$ at 85%, 75% and 50% and two different sample sizes n at 500 and 200, based on 2000 simulation samples, are reported in Table 3.2 for estimating the first component of $\boldsymbol{p}$. The table also includes results on the full sample estimator denoted by "COMP" with no missing values. The simulation results provide empirical evidence on the theoretical development in Sections 3.3. In particular, the CCA estimator is inconsistent, with larger bias corresponding to further departures from the MCAR assumption. The IPW estimator and the MI estimator seem to all perform well, and the FI estimator performs the best among all of them. The second part of the study involves estimating asymptotic variances of consistent estimators and constructing confidence intervals. For the IPW and MI method, results from Theorem 3.1 and Theorem 3.2 are used to estimate the asymptotic variances; for the proposed FI method, both the linearization approach and the resampling approach

Table 3.3: Absolute Relative Bias (%) of Variance Estimators for $p_1 = P(Y \leq 1)$

RR	SS	IPW	MI(1)	MI(5)	MI(10)	FI(L)	FI(R)
85%	500	2.1	2.5	2.0	2.4	1.9	2.1
	200	3.0	2.0	2.9	3.4	3.6	3.6
75%	500	1.7	5.0	2.1	2.5	2.9	3.1
	200	2.6	4.2	3.3	3.4	3.6	3.5
50%	500	8.7	0.6	1.7	1.9	1.7	2.0
	200	5.9	2.0	4.9	4.5	4.3	4.1

are considered, denoted by "FI(L)" and "FI(R)", respectively. For the resampling approach, the bootstrap method is used with 100 bootstrap replications for each simulated sample. Table 3.3 shows the ARB (in %) of different variance estimators and Table 3.4 contains the coverage probability (CP) and average length (AL) of 95% confidence intervals constructed by different methods. All variance estimators have a reasonably small ARB (less than 10%) and are thus consistent. The IPW, the MI and the FI all produce confidence intervals with a CP close to the nominal 95%; the proposed FI method using either the linearization or resampling variance estimators has the shortest AL among others.

For the second study, the settings are identical to those in the first one, but the focus is on the regression coefficients $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \beta_1, \beta_2)'$. Table 3.5 presents the simulation results of different estimators of the parameter β_1 under three levels of nonresponse rate and two different sample sizes. The gold standard full sample estimator "COMP" is also included for comparison. All the absolute relative biases are smaller than 4% except for cases where $E(R) = 50\%$ and $n = 200$. The IPW estimator and the MI estimator with $M = 1$ have the largest biases. The CCA estimator and the FI estimator have almost identical performances, which is in line with the theoretical results. Table 3.6 and Table 3.7 report results for the variance estimators and confidence intervals. For the CCA, the IPW and the MI, variance estimators are based on results in Proposition 3.3.2. We observe that the resampling variance estimators for the FI method have relatively large ARBs for the small sample

Table 3.4: Coverage Probability (%) and Average Length ($\times 10^{-2}$) of 95% Confidence Intervals for $p_1 = P(Y \leq 1)$

RR	SS		IPW	MI(1)	MI(5)	MI(10)	FI(L)	FI(R)
85%	500	CP	95.5	95.1	95.3	95.3	95.2	94.8
		AL	(8.4)	(8.8)	(8.4)	(8.4)	(8.3)	(8.3)
	200	CP	94.3	94.3	94.8	94.7	94.7	94.4
		AL	(13.2)	(13.9)	(13.3)	(13.3)	(13.2)	(13.2)
75%	500	CP	95.7	94.7	95.0	95.0	94.7	94.6
		AL	(8.8)	(9.2)	(8.7)	(8.7)	(8.6)	(8.6)
	200	CP	94.6	94.5	94.8	94.9	95.2	94.7
		AL	(13.9)	(14.5)	(13.8)	(13.7)	(13.6)	(13.6)
50%	500	CP	94.3	94.2	95.0	94.8	94.8	94.5
		AL	(9.9)	(9.9)	(9.3)	(9.3)	(9.2)	(9.1)
	200	CP	94.4	94.2	94.6	94.8	94.9	95.0
		AL	(15.3)	(15.7)	(14.8)	(14.7)	(14.6)	(14.5)

size ($n = 200$), but the biases are still within acceptable range (less than 15%). All four methods have good coverage probabilities with the FI method using linearization variance estimators having the shortest average length.

In the third study, in addition to the two covariates (X_1, X_2), we now include a third covariate X_3 which depends on the other two covariates: $X_3 = 0.5 - X_1 + X_2 + \epsilon$, where $\epsilon \sim N(0, 2)$. Both the imputation model and the analysis model are assumed to follow (2.16) with the *probit* link. The imputation model and the propensity scores involve all three covariates, while the analysis model involves only the first two covariates and has four regression coefficients: $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \beta_1, \beta_2)'$. The ordinal response Y is generated from the analysis model. We simulated a trivial case where Y is generated independent of ϵ. Table 3.8 shows the ARB and RMSE of the CCA estimator for β_1 in this trivial case and it is apparent that the CCA method is consistent which confirms our conclusion that this problem reduces to the

Table 3.5: Absolute Relative Bias (%) and Root Mean Squared Error ($\times 10^{-2}$) of Different Estimators of β_1 (First Scenario)

RR	SS		COMP	CCA	IPW	MI(1)	MI(5)	MI(10)	FI
85%	500	ARB	1.2	1.4	1.4	1.6	1.4	1.4	1.4
		RMSE	(14.8)	(16.2)	(16.3)	(17.2)	(16.4)	(16.3)	(16.2)
	200	ARB	1.9	2.3	2.3	2.6	2.3	2.3	2.3
		RMSE	(22.8)	(25.5)	(25.7)	(27.2)	(25.9)	(25.7)	(25.5)
75%	500	ARB	—	1.5	1.6	1.8	1.6	1.6	1.6
		RMSE	—	(17.5)	(17.6)	(19.2)	(17.9)	(17.6)	(17.5)
	200	ARB	—	2.6	2.7	3.4	2.8	2.7	2.6
		RMSE	—	(28.0)	(28.3)	(30.8)	(28.7)	(28.3)	(28.0)
50%	500	ARB	—	1.8	2.5	2.3	1.9	1.8	1.8
		RMSE	—	(22.0)	(23.9)	(24.4)	(22.5)	(22.2)	(22.0)
	200	ARB	—	3.7	5.4	5.0	3.9	3.9	3.7
		RMSE	—	(36.1)	(38.9)	(40.4)	(37.2)	(36.8)	(36.1)

one in the second study if the ordinal response and the not-in-the-model covariate are conditionally independent. In the non-trivial case where Y depends on ϵ, results on β_1 under three nonresponse rates and two sample sizes based on 2000 simulated samples are presented in Table 3.9. Major observations can be summarized as follows: (i) The CCA estimator has non-negligible biases when $E(R) = 0.75$ and huge biases when $E(R) = 0.50$; (ii) The IPW estimator does not perform well for $n = 200$ or $E(R) = 0.50$; (iii) The MI estimator performs well for $M = 5$ and 10; (iv) The FI estimator performs the best among all methods. Table 3.10 contains the ARBs of variance estimators and Table 3.11 covers the results for the 95% confidence interval. We observe that the IPW and MI with $M = 1$, although theoretically justified, do not perform well in practice with lower-than-normal CP and long AL, especially when $E(R) = 0.5$. On the other hand, the MI with $M = 5$ and 10 and the proposed FI with both the linearization and bootstrap variance estimators have satisfactory performances.

Table 3.6: Absolute Relative Bias (%) of Different Variance Estimators for β_1 (First Scenario)

RR	SS	CCA	IPW	MI(1)	MI(5)	MI(10)	FI(L)	FI(R)
85%	500	3.9	4.2	2.5	3.1	3.7	3.9	0.1
	200	1.2	0.7	1.4	0.4	0.6	1.2	13.4
75%	500	3.9	4.0	4.1	4.1	3.5	3.9	0.4
	200	1.9	2.3	0.2	2.4	1.7	1.9	12.0
50%	500	5.2	3.2	3.5	4.5	4.9	5.2	0.7
	200	5.2	11.2	2.2	5.6	5.8	5.2	14.0

Table 3.7: Coverage Probability (%) and Average Length of 95% Confidence Intervals for β_1 (First Scenario)

RR	SS		CCA	IPW	MI(1)	MI(5)	MI(10)	FI(L)	FI(R)
85%	500	CP	95.0	94.9	95.7	95.1	95.2	95.0	95.5
		AL	(0.61)	(0.62)	(0.65)	(0.62)	(0.62)	(0.61)	(0.62)
	200	CP	95.7	95.7	95.4	96.0	96.1	95.7	96.3
		AL	(0.98)	(0.99)	(1.05)	(0.99)	(0.99)	(0.98)	(1.03)
75%	500	CP	95.6	95.2	94.8	95.2	95.6	95.6	94.9
		AL	(0.66)	(0.66)	(0.72)	(0.67)	(0.67)	(0.66)	(0.67)
	200	CP	95.7	95.5	96.0	95.0	95.7	95.7	96.2
		AL	(1.06)	(1.07)	(1.17)	(1.08)	(1.07)	(1.06)	(1.12)
50%	500	CP	95.3	95.8	95.1	95.3	94.9	95.3	95.1
		AL	(0.82)	(0.93)	(0.92)	(0.84)	(0.93)	(0.82)	(0.84)
	200	CP	95.7	97.1	95.8	95.7	95.5	95.7	96.2
		AL	(1.34)	(1.53)	(1.50)	(1.37)	(1.35)	(1.34)	(1.45)

Table 3.8: Absolute Relative Bias (%) and Root Mean Squared Error ($\times 10^{-2}$) of the CCA Estimator of β_1 when Y and X_3 are Conditionally Independent

	85%		75%		50%	
	$n = 200$	$n = 500$	$n = 200$	$n = 500$	$n = 200$	$n = 500$
ARB	3.3	1.5	3.6	1.5	4.6	2.1
RMSE	(27.1)	(15.8)	(29.3)	(17.2)	(41.9)	(24.3)

Table 3.9: Absolute Relative Bias (%) and Root Mean Squared Error ($\times 10^{-2}$) of Different Estimators of β_1 (Second Scenario).

RR	SS		COMP	CCA	IPW	MI(1)	MI(5)	MI(10)	FI
85%	500	ARB	0.9	4.8	1.9	1.1	1.0	1.0	1.0
		RMSE	(14.6)	(19.3)	(19.7)	(17.2)	(16.3)	(16.2)	(16.2)
	200	ARB	2.2	4.0	2.3	2.8	2.5	2.4	2.4
		RMSE	(23.5)	(31.1)	(25.7)	(28.1)	(27.1)	(26.9)	(26.6)
75%	500	ARB	—	9.2	3.4	1.4	1.1	1.1	1.1
		RMSE	—	(26.1)	(24.6)	(19.2)	(17.9)	(17.8)	(17.6)
	200	ARB	—	11.4	6.5	3.4	2.8	2.6	2.6
		RMSE	—	(37.8)	(38.2)	(31.4)	(29.4)	(29.1)	(28.8)
50%	500	ARB	—	23.5	13.1	2.2	1.8	1.8	1.7
		RMSE	—	(54.3)	(51.2)	(26.6)	(25.0)	(24.8)	(24.6)
	200	ARB	—	27.0	21.4	4.8	3.8	3.7	3.7
		RMSE	—	(70.1)	(74.3)	(44.5)	(41.7)	(41.1)	(40.7)

Table 3.10: Absolute Relative Bias (%) of Different Variance Estimators for β_1 (Second Scenario).

RR	SS	IPW	MI(1)	MI(5)	MI(10)	FI(L)	FI(R)
85%	500	8.3	3.0	3.2	3.7	4.7	0.9
	200	0.04	6.1	6.3	6.2	6.1	5.3
75%	500	12.5	4.1	3.6	4.4	3.8	0.0
	200	4.2	3.5	1.8	2.0	1.5	8.7
50%	500	14.5	1.3	0.9	0.8	0.2	1.0
	200	7.2	3.1	5.0	6.0	7.2	15.3

Table 3.11: Coverage Probability (%) and Average Length of 95% Confidence Intervals for β_1 (Second Scenario).

RR	SS		IPW	MI(1)	MI(5)	MI(10)	FI(L)	FI(R)
85%	500	CP	93.9	93.6	93.9	94.1	94.2	94.3
		AL	(0.72)	(0.65)	(0.62)	(0.62)	(0.61)	(0.62)
	200	CP	95.6	93.2	93.4	93.7	94.0	95.2
		AL	(1.17)	(1.03)	(1.00)	(0.99)	(0.98)	(1.04)
75%	500	CP	92.9	93.5	93.8	94.0	94.7	94.7
		AL	(0.86)	(0.72)	(0.68)	(0.67)	(0.67)	(0.68)
	200	CP	93.9	92.0	93.6	94.2	94.4	95.0
		AL	(1.36)	(1.15)	(1.10)	(1.10)	(1.09)	(1.14)
50%	500	CP	85.7	92.9	94.6	95.0	95.0	94.8
		AL	(1.56)	(1.00)	(0.96)	(0.95)	(0.95)	(0.95)
	200	CP	87.9	92.8	94.6	94.6	95.4	95.7
		AL	(2.25)	(1.65)	(1.61)	(1.60)	(1.60)	(1.65)

3.6 Regularity Conditions and Proofs

3.6.1 Regularity Conditions of Theorem 3.1

Define the joint parameter $\tilde{\boldsymbol{\theta}}_p = (\boldsymbol{\theta}', \boldsymbol{\phi}')'$. Let Θ_p be the joint parameter space of $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{U}}_p(r, y, \boldsymbol{x}; \tilde{\boldsymbol{\theta}}_p) = (\boldsymbol{U}_{ipw}(r, y, \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\phi})', \boldsymbol{T}(r, \boldsymbol{x}; \boldsymbol{\phi})')'$. The following conditions are required for the proof of Theorem 3.1.

S1. The parameter space Θ_p is compact with $(\boldsymbol{\theta}_0', \boldsymbol{\phi}_0')'$ being an interior point;

S2. $E\big[\tilde{\boldsymbol{U}}_p(R, Y, \boldsymbol{X}; \tilde{\boldsymbol{\theta}}_p)\big] = \boldsymbol{0}$ has a unique root;

S3. $\tilde{\boldsymbol{U}}_p(R, Y, \boldsymbol{X}; \tilde{\boldsymbol{\theta}}_p)$ is continuous at each $\tilde{\boldsymbol{\theta}}_p \in \Theta_p$ with probability one;

S4. There exists $\boldsymbol{H}_p(R, Y, \boldsymbol{X})$ such that $|\boldsymbol{U}_p(R, Y, \boldsymbol{X}; \tilde{\boldsymbol{\theta}}_p)| \leq \boldsymbol{H}_p(R, Y, \boldsymbol{X})$ for all $\tilde{\boldsymbol{\theta}}_p$, and $E[\boldsymbol{H}_p(R, Y, \boldsymbol{X})] < \infty$.

S5. $\tilde{\boldsymbol{U}}_p(r, y, \boldsymbol{x}; \tilde{\boldsymbol{\theta}}_p)$ is twice continuously differentiable with respect to $\tilde{\boldsymbol{\theta}}_p$ for every $(r, y, \boldsymbol{x})$;

S6. The second-order partial derivatives of $\tilde{\boldsymbol{U}}_p(r, y, \boldsymbol{x}; \tilde{\boldsymbol{\theta}}_p)$ satisfy

$$\left| \frac{\partial^2 \tilde{\boldsymbol{U}}_p(r, y, \boldsymbol{x}; \tilde{\boldsymbol{\theta}}_p)}{\partial \tilde{\theta}_i \partial \tilde{\theta}_j} \right| \leq \tilde{\boldsymbol{U}}_{p0}(r, y, \boldsymbol{x})$$

for some integrable function $\tilde{\boldsymbol{U}}_{p0}(r, y, \boldsymbol{x})$ for every $\tilde{\boldsymbol{\theta}}_p$ in a neighbourhood of $(\boldsymbol{\theta}_0', \boldsymbol{\phi}_0')'$;

S7. $E\big[\|\tilde{\boldsymbol{U}}_p(R, Y, \boldsymbol{X}; \boldsymbol{\theta}_0, \boldsymbol{\phi}_0)\|^2\big] < \infty$;

S8. $E\big[\partial \tilde{\boldsymbol{U}}_p(R, Y, \boldsymbol{X}; \tilde{\boldsymbol{\theta}}_p)/\partial \tilde{\boldsymbol{\theta}}_p\big]\big|_{\tilde{\boldsymbol{\theta}}_p = (\boldsymbol{\theta}_0', \boldsymbol{\phi}_0')'}$ exists and is non-singular.

3.6.2 Regularity Conditions and Proof of Theorem 3.2

Let $\boldsymbol{S}_{obs}(r, y, \boldsymbol{x}; \boldsymbol{\eta}) = r\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\eta})$, where $\boldsymbol{z}$ is the cumulative indicator vector of y. We assume that $\boldsymbol{S}_{obs}(r, y, \boldsymbol{x}; \boldsymbol{\eta})$ satisfies conditions S1-S8 in Section 2.5. Therefore, by

Theorem 2.1, we have

$$\hat{\boldsymbol{\eta}}^{cc} - \boldsymbol{\eta}_0 = n^{-1} \sum_{i=1}^{n} I_{obs}^{-1} r_i \boldsymbol{S}_i(\boldsymbol{\eta}_0) + o_p(n^{-1/2}) \,, \tag{3.35}$$

where $\boldsymbol{\eta}_0$ is the unique root of $E\big[\boldsymbol{S}_{obs}(\boldsymbol{\eta})\big] = \mathbf{0}$ and is the true parameter value when the imputation model is correct. Let

$$\boldsymbol{U}_i^{(l)}(\boldsymbol{\theta}, \boldsymbol{\eta}) = r_i \boldsymbol{U}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) + (1 - r_i) \boldsymbol{U}(\tilde{y}_{il}, \boldsymbol{x}_i; \boldsymbol{\theta}),$$

where $\tilde{y}_{il}$ is drawn from $f(y \mid \boldsymbol{x}_i; \boldsymbol{\eta})$, then $\hat{\boldsymbol{\theta}}^{(l)}$ is the solution to

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}_i^{(l)}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}^{cc}) \,. \tag{3.36}$$

Suppose that the imputation model is correct. The imputed values are then asymptotically drawn from the true data generating distribution, and thus $\hat{\boldsymbol{\theta}}^{(l)}$ converges in probability to $\boldsymbol{\theta}_0$. To derive the asymptotic variance, we assume that the following conditions also hold:

S1. Let $\Lambda_i(\boldsymbol{\theta}, \boldsymbol{\eta}) = r_i \boldsymbol{U}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) + (1 - r_i) E\big[\boldsymbol{U}(\tilde{Y}, \boldsymbol{x}_i; \boldsymbol{\theta}) \mid \boldsymbol{x}_i; \boldsymbol{\eta}\big]$, where the expectation is taken over $\tilde{Y}$ based on $f(y \mid \boldsymbol{x}_i; \boldsymbol{\eta})$, and $\Lambda(\boldsymbol{\theta}, \boldsymbol{\eta}) = E\big[\Lambda_i(\boldsymbol{\theta}, \boldsymbol{\eta})\big]$, where the expectation is taken over $(R, Y_{obs}, \boldsymbol{X})$. Assume $\partial \Lambda(\boldsymbol{\theta}, \boldsymbol{\eta}) / \partial \boldsymbol{\theta}'$ and $\partial \Lambda(\boldsymbol{\theta}, \boldsymbol{\eta}) / \partial \boldsymbol{\eta}'$ exist and are continuous in $(\boldsymbol{\theta}, \boldsymbol{\eta})$.

S2. Let $\mathcal{L}_{n,\boldsymbol{p}}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = n^{1/2} |\boldsymbol{W}_n(\boldsymbol{\theta}, \boldsymbol{\eta}_1) - \boldsymbol{W}_n(\boldsymbol{\theta}, \boldsymbol{\eta}_2)|$ where

$$\boldsymbol{W}_n(\boldsymbol{\theta}, \boldsymbol{\eta}) = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}_i^{(l)}(\boldsymbol{\theta}, \boldsymbol{\eta}) - \Lambda(\boldsymbol{\theta}, \boldsymbol{\eta}) \,.$$

There exists a positive d such that for any $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ in a neighbourhood of $\boldsymbol{\eta}_0$, $\sup_{|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2| < d} \mathcal{L}_{n,\boldsymbol{p}}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \to 0$ uniformly in $\boldsymbol{\theta}$ as $n \to \infty$.

S3. $E\big[\|\boldsymbol{U}_i^{(l)}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\|^2\big] < \infty$ and $E\big[\|r_i \boldsymbol{S}_i(\boldsymbol{\eta}_0)\|^2\big] < \infty$.

Starting from (3.36), we have

$$
\begin{aligned}
\mathbf{0} &= n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i^{(l)}(\hat{\boldsymbol{\theta}}^{(l)},\hat{\boldsymbol{\eta}}^{cc}) \\
&= n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i^{(l)}(\hat{\boldsymbol{\theta}}^{(l)},\hat{\boldsymbol{\eta}}^{cc}) - \Lambda(\hat{\boldsymbol{\theta}}^{(l)},\hat{\boldsymbol{\eta}}^{cc}) - \Big[n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i^{(l)}(\hat{\boldsymbol{\theta}}^{(l)},\boldsymbol{\eta}_0) - \Lambda(\hat{\boldsymbol{\theta}}^{(l)},\boldsymbol{\eta}_0)\Big] \\
&+ \Lambda(\hat{\boldsymbol{\theta}}^{(l)},\hat{\boldsymbol{\eta}}^{cc}) + n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i^{(l)}(\hat{\boldsymbol{\theta}}^{(l)},\boldsymbol{\eta}_0) - \Lambda(\hat{\boldsymbol{\theta}}^{(l)},\boldsymbol{\eta}_0) \\
&= o_p(n^{-1/2}) + \Lambda(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) + \frac{\partial\Lambda}{\partial\boldsymbol{\theta}'}(\boldsymbol{\theta}^*,\boldsymbol{\eta}_0)(\hat{\boldsymbol{\theta}}^{(l)}-\boldsymbol{\theta}_0) + \frac{\partial\Lambda}{\partial\boldsymbol{\eta}}(\boldsymbol{\theta}_0,\boldsymbol{\eta}^*)(\hat{\boldsymbol{\eta}}^{cc}-\boldsymbol{\eta}_0) \\
&+ n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i^{(l)}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) + \big[n^{-1}\sum_{i=1}^{n}\frac{\partial\boldsymbol{U}_i^{(l)}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}^*,\boldsymbol{\eta}_0)\big](\hat{\boldsymbol{\theta}}^{(l)}-\boldsymbol{\theta}_0) \\
&- \Lambda(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) - \frac{\partial\Lambda}{\partial\boldsymbol{\theta}'}(\boldsymbol{\theta}^*,\boldsymbol{\eta}_0)(\hat{\boldsymbol{\theta}}^{(l)}-\boldsymbol{\theta}_0)
\end{aligned}
$$

where $\boldsymbol{\theta}^* \to \boldsymbol{\theta}_0$ and $\boldsymbol{\eta}^* \to \boldsymbol{\eta}_0$. Noting that, when $\boldsymbol{\eta} = \boldsymbol{\eta}_0$, $\tilde{y}_{il}$'s are samples from $f(y \mid \boldsymbol{x}_i;\boldsymbol{\eta}_0)$ just like the original responses y_i, hence

$$
n^{-1}\sum_{i=1}^{n}\frac{\partial\boldsymbol{U}_i^{(l)}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}^*,\boldsymbol{\eta}_0) \to E\left[\frac{\partial\boldsymbol{U}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right].
$$

From the above discussion:

$$
\Big\{-E\big[\frac{\partial\boldsymbol{U}}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\big]\Big\}(\hat{\boldsymbol{\theta}}^{(l)}-\boldsymbol{\theta}_0) = n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i^{(l)}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) + \frac{\partial\Lambda}{\partial\boldsymbol{\eta}}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)(\hat{\boldsymbol{\eta}}^{cc}-\boldsymbol{\eta}_0) + o_p(n^{-1/2}).
$$

By substituting (3.35), we have

$$
\hat{\boldsymbol{\theta}}^{(l)} - \boldsymbol{\theta}_0 = \boldsymbol{\tau}n^{-1}\sum_{i=1}^{n}\left\{\boldsymbol{U}_i^{(l)}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) + \frac{\partial\Lambda}{\partial\boldsymbol{\eta}}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)I_{obs}^{-1}r_i\boldsymbol{S}_i(\boldsymbol{\eta}_0)\right\} + o_p(n^{-1/2}). \quad (3.37)
$$

Now we consider $\partial\Lambda(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)/\partial\boldsymbol{\eta}$. Note that only the second term in $\Lambda_i(\boldsymbol{\theta},\boldsymbol{\eta})$ involves $\boldsymbol{\eta}$, so by the definition of $\Lambda(\boldsymbol{\theta},\boldsymbol{\eta})$, we have,

$$\begin{aligned}
\frac{\partial\Lambda}{\partial\boldsymbol{\eta}}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) &= \int(1-r)\frac{\partial}{\partial\boldsymbol{\eta}}\int \boldsymbol{U}(\tilde{y},\boldsymbol{x};\boldsymbol{\theta}_0)f(\tilde{y}\mid\boldsymbol{x};\boldsymbol{\eta})d\tilde{y}\,dF(r,y_{obs},\boldsymbol{x}) \\
&= \int(1-r)\int \boldsymbol{U}(\tilde{y},\boldsymbol{x};\boldsymbol{\theta}_0)\frac{\partial}{\partial\boldsymbol{\eta}}f(\tilde{y}\mid\boldsymbol{x};\boldsymbol{\eta})d\tilde{y}\,dF(r,y_{obs},\boldsymbol{x}) \\
&= \int(1-r)\int \boldsymbol{U}(\tilde{y},\boldsymbol{x};\boldsymbol{\theta}_0)(1-r)\boldsymbol{S}(\boldsymbol{\eta}_0)f(\tilde{y}\mid\boldsymbol{x};\boldsymbol{\eta})d\tilde{y}\,dF(r,y_{obs},\boldsymbol{x}) \\
&= \int\int(1-r)\boldsymbol{U}(\tilde{y},\boldsymbol{x};\boldsymbol{\theta}_0)\boldsymbol{S}(\boldsymbol{\eta}_0)f(\tilde{y}\mid\boldsymbol{x};\boldsymbol{\eta})d\tilde{y}\,dF(r,y_{obs},\boldsymbol{x}) \\
&= E\Big[(1-R)\boldsymbol{U}(\boldsymbol{\theta}_0)\boldsymbol{S}(\boldsymbol{\eta}_0)\Big] \\
&= \boldsymbol{\kappa}\,.
\end{aligned}$$

Since (3.37) holds for all l, it is not difficult to obtain:

$$\hat{\boldsymbol{\theta}}_{mi}-\boldsymbol{\theta}_0=\boldsymbol{\tau}n^{-1}\sum_{i=1}^{n}\left\{\boldsymbol{U}_{i,mi}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)+\boldsymbol{\kappa}I_{obs}^{-1}r_i\boldsymbol{S}_i(\boldsymbol{\eta}_0)\right\}+o_p(n^{-1/2})\,.$$

where $\boldsymbol{U}_{i,mi}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)=r_i\boldsymbol{U}(y_i,\boldsymbol{x}_i;\boldsymbol{\theta}_0)+(1-r_i)M^{-1}\sum_{l=1}^{M}\boldsymbol{U}(\tilde{y}_{il},\boldsymbol{x}_i;\boldsymbol{\theta}_0)$ and $\tilde{y}_{il}$ are drawn from $f(y\mid\boldsymbol{x}_i;\boldsymbol{\eta}_0)$. Theorem 3.2 follows by the Central Limit Theorem. The asymptotic properties of $\hat{\boldsymbol{\theta}}_{fi,e}$ can be derived following similar arguments and it is easy to see $\hat{\boldsymbol{\theta}}_{fi,e}$ and $\hat{\boldsymbol{\theta}}_{mi}$ are asymptotically equivalent.

3.6.3 Regularity Conditions and Proof of Theorem 3.3

Let $\tilde{\boldsymbol{\theta}}_f=(\boldsymbol{\theta}',\boldsymbol{\eta}')'$ and $\tilde{\boldsymbol{U}}_f(r,y,\boldsymbol{x};\tilde{\boldsymbol{\theta}}_f)=(\boldsymbol{U}_{fi}(r,y,\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\eta})',r\boldsymbol{S}(\boldsymbol{z},\boldsymbol{x};\boldsymbol{\eta})')$. Note that unlike the $\boldsymbol{U}^{(l)}(\boldsymbol{\theta},\boldsymbol{\eta})$ in Theorem 3.2 which depends on $\boldsymbol{\eta}$ implicitly through random draws from $f(y\mid\boldsymbol{x}_i;\boldsymbol{\eta})$, $\boldsymbol{U}_{fi}(r,y,\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\eta})$ is an explicit function of $(\boldsymbol{\theta}',\boldsymbol{\eta}')'$. We assume that the regularity conditions S1-S8 of Theorem 3.1 hold for $\tilde{\boldsymbol{U}}_f(r,y,\boldsymbol{x};\tilde{\boldsymbol{\theta}}_f)$ and $\tilde{\boldsymbol{\theta}}_f$. Then by Theorem 2.1, we have

$$\begin{pmatrix}\hat{\boldsymbol{\theta}}^{fi}-\boldsymbol{\theta}_0\\ \hat{\boldsymbol{\eta}}^{cc}-\boldsymbol{\eta}_0\end{pmatrix}=-\Big[\mathring{\Psi}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\Big]^{-1}\begin{pmatrix}n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_{fi,i}(\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\\ n^{-1}\sum_{i=1}^{n}r_i\boldsymbol{S}_i(\boldsymbol{\eta}_0)\end{pmatrix}+o_p(n^{-1/2})\,,\qquad(3.38)$$

where $\boldsymbol{U}_{fi,i}$ and $\boldsymbol{S}_i$ are the functions $\boldsymbol{U}_{fi}$ and $\boldsymbol{S}$ evaluated at the ith observation and

$$\mathring{\Psi}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0) = \begin{pmatrix} E[\partial \boldsymbol{U}_{fi}/\partial \boldsymbol{\theta}]' & E[\partial \boldsymbol{U}_{fi}/\partial \boldsymbol{\eta}]' \\ \mathbf{0} & E[\partial r \boldsymbol{S}/\partial \boldsymbol{\eta}]' \end{pmatrix} \Bigg|_{\boldsymbol{\theta}_0, \boldsymbol{\eta}_0}.$$

It is apparent that $E[\partial \boldsymbol{U}_{fi}/\partial \boldsymbol{\theta}]'\big|_{\boldsymbol{\theta}_0,\boldsymbol{\eta}_0} = -\boldsymbol{\tau}^{-1}$ and $E[\partial r \boldsymbol{S}/\partial \boldsymbol{\eta}]'\big|_{\boldsymbol{\eta}_0} = -I_{obs}$ given in Theorem 3.2. Noting that $\boldsymbol{U}_{fi}$ depends on $\boldsymbol{\eta}$ only through $w_j = w_j(\boldsymbol{x}; \boldsymbol{\eta})$, we have

$$\begin{aligned} E\Big[\partial \boldsymbol{U}_{fi}/\partial \boldsymbol{\eta}\Big]'_{\boldsymbol{\theta}_0,\boldsymbol{\eta}_0} &= E\left\{(1-r)\sum_{j=1}^{J}\big[\partial w_j(\boldsymbol{x};\boldsymbol{\eta}_0)/\partial \boldsymbol{\eta}\big]\boldsymbol{U}(j,\boldsymbol{x};\boldsymbol{\theta}_0)'\right\}' \\ &= E\left\{(1-r)\sum_{j=1}^{J}\boldsymbol{U}(j,\boldsymbol{x};\boldsymbol{\theta}_0)w_j(\boldsymbol{x};\boldsymbol{\eta}_0)\big[\partial \log w_j(\boldsymbol{x};\boldsymbol{\eta}_0)/\partial \boldsymbol{\eta}\big]'\right\} \\ &= E\left\{(1-r)\sum_{j=1}^{J} w_j(\boldsymbol{x};\boldsymbol{\eta}_0)\boldsymbol{U}(j,\boldsymbol{x};\boldsymbol{\theta}_0)\boldsymbol{S}(\boldsymbol{c}_j,\boldsymbol{x};\boldsymbol{\eta}_0)'\right\} \\ &= \boldsymbol{\kappa}\,, \end{aligned}$$

where the third equality holds by the definition of $\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\eta})$ in (2.19). By the inverse formula for block matrix (Henderson and Searle 1981), we have

$$-\Big[\mathring{\Psi}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_0)\Big]^{-1} = \begin{pmatrix} \boldsymbol{\tau} & \boldsymbol{\tau}\boldsymbol{\kappa} I_{obs}^{-1} \\ \mathbf{0} & I_{obs}^{-1} \end{pmatrix},$$

and the asymptotic variance formula in Theorem 3.3 follows from (3.38) and the Central Limit Theorem.

3.6.4 Proof of the Validity of Bootstrap Variance Estimators

First, it is apparent that when $B \to \infty$, $\hat{\boldsymbol{V}}_B$ consistently estimates the variance of $\hat{\boldsymbol{\theta}}^{(b)}$, so it suffices to show that $\hat{\boldsymbol{\theta}}^{(b)}$ has the same asymptotic variance as $\hat{\boldsymbol{\theta}}^{fi}$. Let $\tilde{\boldsymbol{\theta}}_f = (\boldsymbol{\theta}', \boldsymbol{\eta}')'$ and $\tilde{\boldsymbol{U}}_f(r, y, \boldsymbol{x}; \tilde{\boldsymbol{\theta}}_f) = (\boldsymbol{U}_{fi}(r, y, \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\eta})', r\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\eta})')$ be the joint parameter vector and joint estimating function defined in Section 3.6.3. Consider an infinite path of the response variables $\mathcal{P} = \big\{(r_1, y_1, \boldsymbol{x}_1), \ldots, (r_n, y_n, \boldsymbol{x}_n), \ldots\big\}$. By the

Table 3.12: The Triangular Array Formed by Bootstrap Samples

$$
\begin{array}{llll}
(R^*_{11}, Y^*_{11}, \boldsymbol{X}^*_{11}) & & & \\
(R^*_{21}, Y^*_{21}, \boldsymbol{X}^*_{21}) & (R^*_{22}, Y^*_{22}, \boldsymbol{X}^*_{22}) & & \\
\vdots & \vdots & & \\
(R^*_{n1}, Y^*_{n1}, \boldsymbol{X}^*_{n1}) & (R^*_{n2}, Y^*_{n2}, \boldsymbol{X}^*_{n2}) & \cdots & (R^*_{nn}, Y^*_{nn}, \boldsymbol{X}^*_{nn}) \\
\vdots & \vdots & & \vdots
\end{array}
$$

Strong Law of Large Numbers, the following two conditions hold for almost all paths.

$$
\begin{aligned}
&\text{(i).}\ n^{-1}\sum_{i=1}^{n} \mathring{\tilde{\boldsymbol{U}}}_f(r_i, y_i, \boldsymbol{x}_i; \hat{\tilde{\boldsymbol{\theta}}}_f) \longrightarrow E\Big[\mathring{\tilde{\boldsymbol{U}}}_f(R, Y, \boldsymbol{X}; \tilde{\boldsymbol{\theta}}_{f0})\Big], \\
&\text{(ii).}\ n^{-1}\sum_{i=1}^{n} \tilde{\boldsymbol{U}}_f^{\otimes 2}(r_i, y_i, \boldsymbol{x}_i; \hat{\tilde{\boldsymbol{\theta}}}_f) \longrightarrow E\Big[\tilde{\boldsymbol{U}}_f^{\otimes 2}(R, Y, \boldsymbol{X}; \tilde{\boldsymbol{\theta}}_{f0})\Big],
\end{aligned}
$$

where $\hat{\tilde{\boldsymbol{\theta}}}_f = (\hat{\boldsymbol{\theta}}', \hat{\boldsymbol{\eta}}^{cc'})'$, $\tilde{\boldsymbol{\theta}}_{f0} = (\boldsymbol{\theta}_0', \boldsymbol{\eta}_0')'$ and $\mathring{\tilde{\boldsymbol{U}}}_f = \partial \tilde{\boldsymbol{U}}_f / \partial \tilde{\boldsymbol{\theta}}_f$. Conditional on one such path, the bootstrap samples form a triangular array as shown in Table 3.12, where the nth row consists of n *i.i.d.* samples from the empirical distribution of the first n points in the path, i.e., $\{(r_1, y_1, \boldsymbol{x}_1), \ldots, (r_n, y_n, \boldsymbol{x}_n)\}$. It then follows that

$$
E\Big\{\tilde{\boldsymbol{U}}_f(R^*_{ni}, Y^*_{ni}, \boldsymbol{X}^*_{ni}; \hat{\tilde{\boldsymbol{\theta}}}_f) \mid \mathcal{P}\Big\} = n^{-1}\sum_{i=1}^{n} \tilde{\boldsymbol{U}}_f(r_i, y_i, \boldsymbol{x}_i; \hat{\tilde{\boldsymbol{\theta}}}_f) = \boldsymbol{0}.
$$

The bootstrap estimator $\hat{\tilde{\boldsymbol{\theta}}}_f^{(b)}$ solves the estimating equations

$$
\boldsymbol{0} = n^{-1}\sum_{i=1}^{n} \tilde{\boldsymbol{U}}_f(R^*_{ni}, Y^*_{ni}, \boldsymbol{X}^*_{ni}; \tilde{\boldsymbol{\theta}}_f).
$$

We now expand $n^{-1}\sum_{i=1}^{n} \tilde{\boldsymbol{U}}_f(R^*_{ni}, Y^*_{ni}, \boldsymbol{X}^*_{ni}; \hat{\tilde{\boldsymbol{\theta}}}_f^{(b)})$ around $\hat{\tilde{\boldsymbol{\theta}}}_f$ and have

$$
\boldsymbol{0} = n^{-1}\sum_{i=1}^{n} \tilde{\boldsymbol{U}}_f(R^*_{ni}, Y^*_{ni}, \boldsymbol{X}^*_{ni}; \hat{\tilde{\boldsymbol{\theta}}}_f) + n^{-1}\sum_{i=1}^{n} \mathring{\tilde{\boldsymbol{U}}}_f(R^*_{ni}, Y^*_{ni}, \boldsymbol{X}^*_{ni}; \hat{\tilde{\boldsymbol{\theta}}}_f)(\hat{\tilde{\boldsymbol{\theta}}}_f^{(b)} - \hat{\tilde{\boldsymbol{\theta}}}_f) + o_p(n^{-1/2}).
$$

It is easy to see that

$$\hat{\tilde{\boldsymbol{\theta}}}_f^{(b)} - \hat{\tilde{\boldsymbol{\theta}}}_f = \left\{ n^{-1} \sum_{i=1}^{n} \mathring{\tilde{\boldsymbol{U}}}_f(R_{ni}^*, Y_{ni}^*, \boldsymbol{X}_{ni}^*; \hat{\tilde{\boldsymbol{\theta}}}_f) \right\}^{-1} \left\{ n^{-1} \sum_{i=1}^{n} \tilde{\boldsymbol{U}}_f(R_{ni}^*, Y_{ni}^*, \boldsymbol{X}_{ni}^*; \hat{\tilde{\boldsymbol{\theta}}}_f) \right\} + o_p(n^{-1/2}). \tag{3.39}$$

Note that

$$E\Big\{ \mathring{\tilde{\boldsymbol{U}}}_f(R_{ni}^*, Y_{ni}^*, \boldsymbol{X}_{ni}^*; \hat{\tilde{\boldsymbol{\theta}}}_f) \mid \mathcal{P} \Big\} = n^{-1} \sum_{i=1}^{n} \mathring{\tilde{\boldsymbol{U}}}_f(r_i, y_i, \boldsymbol{x}_i; \hat{\tilde{\boldsymbol{\theta}}}_f).$$

By the Weak Law of Large Numbers for the triangular arrays and condition (i), we show

$$n^{-1} \sum_{i=1}^{n} \mathring{\tilde{\boldsymbol{U}}}_f(R_{ni}^*, Y_{ni}^*, \boldsymbol{X}_{ni}^*; \hat{\tilde{\boldsymbol{\theta}}}_f) \xrightarrow{p} E\Big[\mathring{\tilde{\boldsymbol{U}}}_f(R, Y, \boldsymbol{X}; \tilde{\boldsymbol{\theta}}_{f0}) \Big]. \tag{3.40}$$

Also noting that

$$E\Big\{ \tilde{\boldsymbol{U}}_f(R_{ni}^*, Y_{ni}^*, \boldsymbol{X}_{ni}^*; \hat{\tilde{\boldsymbol{\theta}}}_f) \mid \mathcal{P} \Big\} = n^{-1} \sum_{i=1}^{n} \tilde{\boldsymbol{U}}_f(r_i, y_i, \boldsymbol{x}_i; \hat{\tilde{\boldsymbol{\theta}}}_f) = \boldsymbol{0}\,,$$

$$Var\Big\{ \tilde{\boldsymbol{U}}_f(R_{ni}^*, Y_{ni}^*, \boldsymbol{X}_{ni}^*; \hat{\tilde{\boldsymbol{\theta}}}_f) \mid \mathcal{P} \Big\} = n^{-1} \sum_{i=1}^{n} \tilde{\boldsymbol{U}}_f^{\otimes 2}(r_i, y_i, \boldsymbol{x}_i; \hat{\tilde{\boldsymbol{\theta}}}_f)\,,$$

by the Central Limit Theorem for triangular arrays, we have

$$n^{1/2} \times \frac{n^{-1} \sum_{i=1}^{n} \tilde{\boldsymbol{U}}_f(R_{ni}^*, Y_{ni}^*, \boldsymbol{X}_{ni}^*; \hat{\tilde{\boldsymbol{\theta}}}_f)}{\sqrt{E\Big[\mathring{\tilde{\boldsymbol{U}}}_f(R, Y, \boldsymbol{X}; \tilde{\boldsymbol{\theta}}_{f0}) \Big]}} \times \frac{\sqrt{E\Big[\mathring{\tilde{\boldsymbol{U}}}_f(R, Y, \boldsymbol{X}; \tilde{\boldsymbol{\theta}}_{f0}) \Big]}}{\sqrt{n^{-1} \sum_{i=1}^{n} \tilde{\boldsymbol{U}}_f^{\otimes 2}(r_i, y_i, \boldsymbol{x}_i; \hat{\tilde{\boldsymbol{\theta}}}_f)}} \xrightarrow{d} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{3.41}$$

By equations (3.39), (3.40), (3.41) and the Slutsky's Theorem, it follows that

$$n^{1/2}(\hat{\tilde{\boldsymbol{\theta}}}_f^{(b)} - \hat{\tilde{\boldsymbol{\theta}}}_f) \xrightarrow{d} \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{fi})\,,$$

where $\boldsymbol{\Sigma}_{fi}$ is defined in Theorem 3.3. In other words, we have shown conditional on almost all paths of the response variables, the bootstrap estimator has the same asymptotic distribution as the original fractional imputation estimator. If the uniform integrability condition (Serfling 1980) is also satisfied, the convergence in distribution

implies the convergence in the second moment, therefore the variance estimator $\hat{\boldsymbol{V}}_B$ is consistent.

Chapter 4

Fractional Imputation for Multivariate Incomplete Mixed-type Variables

In this chapter, we extend our study to the general setting described in Section 2.1, where multiple variables of mixed types in the data set are subject to missingness. We will focus on two types of variables: ordinal and continuous, but the proposed method can be easily adapted for unordered categorical variables.

For multivariate incomplete data, Schafer (1997) presented a unified framework to conduct multiple imputation based on the joint modelling of incomplete variables. However, when different types of variables are involved, a full Bayesian model for joint multiple imputation would be problematic. The sequential regression multiple imputation (SRMI) method, proposed by Raghunathan et al. (2001) and also known as multiple imputation with chained equations (MICE) (Buuren and Groothuis-Oudshoorn 2011), is a flexible and practical procedure for generating multiple imputed data sets, and the method is intended to handle different types of variables. In Section 4.1.2, we elaborate on key steps to implement the method. White et al. (2011) provided an excellent overview on SRMI. One of the major drawbacks of SRMI, however, is the lack of theoretical justifications. The general procedure resembles the Markov Chain Monte Carlo technique but the explicit relationship between the two and theoretical properties of the method have yet to be developed (Kenward and Carpenter 2007). The popularity of SRMI in practical applications rests largely on empirical studies

rather than theoretical arguments (White et al. 2011).

4.1 Existing Methods

4.1.1 Complete-case Analysis and Inverse Probability Weighting

As in the univariate case, the CCA creates a complete data set by deleting observations with missing values for at least one variable. The subsequent estimator of $\boldsymbol{\theta}$ based on the remaining fully observed units is defined in (2.5).

The IPW method attaches an additional column of inverse probability weights to the data set created by CCA, attempting to correct the potential bias. This requires estimating the full response probabilities

$$\pi_i = P(\delta_i = 1 \mid \boldsymbol{y}_i, \boldsymbol{x}_i) = P(\boldsymbol{r}_i = \mathbf{1} \mid \boldsymbol{y}_i, \boldsymbol{x}_i).$$

In the multivariate case, this is not easy without further assumptions on the missing data process, since the probabilities depend on partially observed $\boldsymbol{y}_i$. One such assumption is to consider covariate-dependent missingness (Little 1995), under which the distribution of the response indicator vector $\boldsymbol{R}$ only depends on fully-observed covariates. It then follows that

$$\pi_i = P(\delta_i = 1 \mid \boldsymbol{y}_i, \boldsymbol{x}_i) = P(\boldsymbol{r}_i = \mathbf{1} \mid \boldsymbol{y}_i, \boldsymbol{x}_i) = P(\boldsymbol{r}_i = \mathbf{1} \mid \boldsymbol{x}_i),$$

and so we can specify a parametric form $\pi(\boldsymbol{x}_i; \boldsymbol{\phi})$ such as a logistic model given in (3.2), for π_i. The parameters $\boldsymbol{\phi}$ can be estimated by maximum likelihood method using data $\big\{(\delta_i, \boldsymbol{x}_i),\, i = 1, \ldots, n\big\}$. If the data set is from a longitudinal study and is missing monotonically, a set of conditional models can be used to estimate the full response probabilities. See Section 5.3 for details. Robins and Gill (1997) proposed a class of Markov random monotone missing (RMM) models to describe the underlying physical mechanism that generates intermittently missing responses. But the estimation of RMM models is computationally heavy, which limits its use when a considerable amount of variables are missing.

4.1.2 Sequential Regression Multiple Imputation

The sequential regression multiple imputation procedure can be carried out through the following steps:

(1) Specify a regression model for Y_1 and the fully observed covariates $\boldsymbol{X}$ along with a prior distribution for model parameters. Draw imputed values of Y_1 from the posterior predictive distribution for missing observations of Y_1.

(2) Repeat the process by regressing the next variable Y_t with missing values on all the previously imputed $Y_{t-1}, \ldots, Y_1$ (imputed values are treated as if they were observed) and fully observed covariates, then drawing imputed values for missing observations of Y_t from the posterior predictive distribution, for $t = 2, \ldots, T$, until we obtain a complete data set.

(3) Specify a regression model for Y_1 with $Y_2, \ldots, Y_T$ and $\boldsymbol{X}$ as covariates along with a prior distribution for model parameters. Update the imputed values for Y_1 with draws from the posterior predictive distribution based on the assumed model and the "complete" data set obtained in the previous step, treating the imputed values for $Y_2, \ldots, Y_T$ as if they were observed.

(4) Repeat Step (3) by regressing Y_t on $Y_1, \ldots, Y_{t-1}, Y_{t+1}, \ldots, Y_T$ and $\boldsymbol{X}$ and updating imputed values for Y_t with draws from the new posterior predictive distribution, for $t = 2, \ldots, T$, until we update all the imputed values for $Y_1, \ldots, Y_T$.

(5) Repeat Step (3) and (4) for a pre-specified number of times or until certain stability criterion is met to obtain the first imputed data set.

(6) Repeat Steps (1)-(5) to obtain multiple imputed data sets.

Subsequent analysis based on the multiple imputed data sets can be carried out following the general procedure for multiple imputation discussed in Section 2.2.3.

When the ordinal components in $\boldsymbol{Y}$ serve as predictors in the regression models required for the SRMI method, there are two possible approaches: the first is to ignore the ordinality and use dummy variables; the second is to assign proper scores to each level and treat them as regular discrete numeric variables. For most applications, the dummy variable approach is preferable (Royston et al. 2009).

4.2 Sequential Regression Fractional Imputation

In this section, we propose a fractional imputation procedure to create a single enlarged complete data file when $\boldsymbol{Y}$ has both ordinal and continuous components. Let $\mathcal{D}$ and $\mathcal{C}$ be the index sets of the ordinal and continuous components of $\boldsymbol{Y}$, respectively. To better explain the procedure, we add the indices $e_i = i$ to the original data set, so the available data set is $\mathcal{O} = \big\{(e_i, \boldsymbol{r}_i, \boldsymbol{y}_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\big\}$.

We first impose a sequence of regression models on each component of $\boldsymbol{Y}$ given all the previous components and the fully observed covariates $\boldsymbol{X}$:

$$Y_1 \mid \boldsymbol{X} \sim f(y_1 \mid \boldsymbol{x}; \boldsymbol{\eta}_1), \quad Y_t \mid Y_{t-1}, \ldots, Y_1, \boldsymbol{X} \sim f(y_t \mid y_{t-1}, \ldots, y_1, \boldsymbol{x}; \boldsymbol{\eta}_t), \tag{4.1}$$

for $t = 2, \ldots, T$, where $f(\cdot)$ denotes the density function if Y_t is continuous and is the mass function if it is ordinal. Here $\boldsymbol{\eta}_t$'s are the corresponding model parameters. The forms of these models are chosen by the type of the response. For ordinal components, Y_t can take the general form of (2.15), while for continuous components, a normal linear or non-linear model on a suitable scale can be postulated. As noted above, when ordinal variables serve as predictors, we use dummy variables. The fractionally imputed data set, created by the procedure we will discuss later in this section, can be used to assess the goodness of fit of these sequential regression models and thus helps the data creator to choose sensible forms for these models.

We further posit another set of marginal models involving the continuous responses and the fully-observed covariates $\boldsymbol{X}$,

$$\boldsymbol{Y}_t \mid \boldsymbol{X} \sim h(y_t \mid \boldsymbol{x}; \boldsymbol{\psi}_t), \quad \text{for } t \in \mathcal{C}, \tag{4.2}$$

parameterized by $\boldsymbol{\psi}_t$ with $h(\cdot)$ being the density functions. These marginal models are then fitted with the observed data $\big\{(y_{it}, \boldsymbol{x}_i),\ i \in \{i : r_{it} = 1\}\big\}$. Let $\hat{\psi}_t$ be the resulting estimators. It is important to note that we do not require these marginal models to be correctly specified, and $\hat{\boldsymbol{\psi}}_t$ is not necessarily a valid estimator. As we will show later, we generate imputed values as random draws from these distributions, but they need not to be the "true" generating process underlying the original data, because we will calibrate the imputed values with fractional weights. We choose to use the marginal models in (4.2) and $\hat{\psi}_t$ so that the imputed values will have the same range as the original response and appear to be plausible. The single fractionally imputed

data set is then created in two stages.

Stage One: Create imputed values for the mixed-type responses

We have different imputation strategies for ordinal and continuous variables, so it is better to consider them separately by groups. Missing values are imputed for one response at a time and the way in which these values are generated depends on the type (continuous or ordinal) of the variable and the nonresponse status (observed or missing) of previously considered variables. The detailed procedures are as follows:

1. Reorder the response variables as $Y_{Q(1)}, \ldots, Y_{Q(T)}$ where $Q(\cdot)$ is a permutation of $1, \ldots, T$, such that the first T_1 variables are continuous and the remaining are ordinal. For example, consider Y_1, Y_2, Y_3, where Y_2 is continuous and Y_1, Y_3 are ordinal, then the reordered variables are Y_2, Y_1, Y_3.

2. Start from the first continuous response $Y_{Q(1)}$. Units with $Y_{Q(1)}$ observed remain unchanged. If the ith observation has $Y_{Q(1)}$ missing, the whole unit is replicated M times and the missing values of $Y_{Q(1)}$ are imputed by random draws $\tilde{y}_{i1}, \ldots, \tilde{y}_{iM}$ from $h(y_{Q(1)} \mid \boldsymbol{x}_i; \hat{\boldsymbol{\psi}}_{Q(1)})$, where M is a pre-specified positive integer. Move on to $Y_{Q(2)}$ when all missing values of $Y_{Q(1)}$ are imputed.

3. Consider the second continuous response $Y_{Q(2)}$ based on the enlarged data file from last step. Units with $Y_{Q(2)}$ observed remain unchanged. If the ith observation has $Y_{Q(2)}$ missing and all previous variables, in this case $Y_{Q(1)}$, observed, then replicate the whole unit M times and impute the missing values of $Y_{Q(2)}$ by random draws $\tilde{y}_{i1}, \ldots, \tilde{y}_{iM}$ from $h(y_{Q(2)} \mid \boldsymbol{x}_i; \hat{\boldsymbol{\psi}}_{Q(2)})$. Otherwise, draw a single point $\tilde{y}_i$ from $h(y_{Q(2)} \mid \boldsymbol{x}_i; \hat{\boldsymbol{\psi}}_{Q(2)})$ as the imputed value for $Y_{Q(2)}$.

4. Repeat STEP 3 for the remaining continuous variables $Y_{Q(3)}, \ldots, Y_{Q(T_1)}$, until we obtain a data set with complete observations for all continuous response variables.

5. Based on the data set from last step, consider the first ordinal response $Y_{Q(T_1+1)}$ on a $J_{Q(T_1+1)}$-level scale. Units with $Y_{Q(T_1+1)}$ observed remain unchanged. Observations with missing values for $Y_{Q(T_1+1)}$ are replicated $J_{Q(T_1+1)}$ times and the missing values are filled in with $1, \ldots, J_{Q(T_1+1)}$.

6. Repeat STEP 5 for all remaining ordinal variables $Y_{Q(T_1+2)}, \ldots, Y_{Q(T)}$ until we obtain a complete data set.

Let $\mathcal{O}^* = \left\{(e_i^*, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*),\ i = 1, \ldots, n^*\right\}$ be the resulting complete data set. Each observation in the original data $\mathcal{O}$ with missing values for at least one response variable is replaced by a cluster of observations with imputed values in $\mathcal{O}^*$, while the fully-observed observations in $\mathcal{O}$ remain the same in $\mathcal{O}^*$. The value of e_i^* indicates the index of the observation in the original data set that corresponds to the ith unit in $\mathcal{O}^*$. The columns $\left\{(e_i^*, \boldsymbol{r}_i^*),\ i = 1, \ldots, n^*\right\}$ facilitate the calculation of fractional weights in the second stage and can be removed before the release of the file for confidentiality considerations.

Stage Two: Calculate fractional weights

Each observation in the imputed data set is accompanied by a weight w_i^*, which can be calculated iteratively by the following procedures:

1. Choose initial values $\left\{\boldsymbol{\eta}_1^{(0)}, \ldots, \boldsymbol{\eta}_T^{(0)}\right\}$ for parameters in model (4.1).

2. Define the intermediate function

$$\begin{aligned} & g(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T) \\ = & \frac{f(y_1 \mid \boldsymbol{x}; \boldsymbol{\eta}_1) f(y_2 \mid y_1, \boldsymbol{x}; \boldsymbol{\eta}_2) \cdots f(y_T \mid y_{T-1}, \ldots, y_1, \boldsymbol{x}; \boldsymbol{\eta}_T)}{h(y_{Q(1)} \mid \boldsymbol{x}; \hat{\boldsymbol{\psi}}_{Q(1)})^{I(r_{Q(1)}=0)} \cdots h(y_{Q(T_1)} \mid \boldsymbol{x}; \hat{\boldsymbol{\psi}}_{Q(T_1)})^{I(r_{Q(T_1)}=0)}}, \end{aligned} \tag{4.3}$$

   and the general weight function

$$W(i_0, \boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T) = \frac{g(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T)}{\sum_{l \in S(i_0)} g(\boldsymbol{r}_l^*, \boldsymbol{y}_l^*, \boldsymbol{x}_l^*; \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T)}, \tag{4.4}$$

   where $S(i_0) = \left\{l \mid l \in \{1, \ldots, n^*\} \text{ and } e_l^* = i_0\right\}$.

3. Calculate the initial weights:

$$w_i^{*(0)} = W(e_i^*, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*; \boldsymbol{\eta}_1^{(0)}, \ldots, \boldsymbol{\eta}_T^{(0)}), \quad i = 1, \ldots, n^*.$$

4. Apply the maximum likelihood method to fit the models in (4.1) using the imputed data set $\mathcal{O}^*$ with the weights $w_i^{*(0)}$ for the first iteration or the weights

$w_i^{*(1)}$ from STEP 5 for subsequent iterations and obtain updated estimates $\boldsymbol{\eta}_1^{(1)}, \ldots, \boldsymbol{\eta}_T^{(1)}$.

5. Update the fractional weights as

$$w_i^{*(1)} = W(e_i^*, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*; \boldsymbol{\eta}_1^{(1)}, \ldots, \boldsymbol{\eta}_T^{(1)}), \quad i = 1, \ldots, n^*.$$

6. Repeat STEP 4 and STEP 5 until the fractional weights converge. Denote the final converged weights by $\boldsymbol{w}^* = (w_1^*, \ldots, w_{n^*}^*)$.

Note that $S(i_0)$ is the index set of observations in $\mathcal{O}^*$ that correspond to the i_0th observation in $\mathcal{O}$. If the original i_0th observation has missing values for at least one of the response variables, $S(i_0)$ indicates the cluster of imputed observations for that incomplete original observation. From the definition, it is easy to see that

$$\sum_{i \in S(i_0)} W(i_0, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*; \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T) = 1,$$

the imputed observations receive fractional weights adding up to 1 for any i_0. If the original i_0th observation is fully observed, then $S(i_0)$ contains exactly one point, which is the original observation itself and hence the weight assigned to a fully observed unit is 1.

Replication weights can be obtained following the same procedure as discussed in Section 3.2.3. When applying the above procedure to the replication sample, we skip *Stage One* and directly use the imputed values in the data set $\mathcal{O}^*$, and then follow steps in *Stage Two* to re-calculate the weights based on the replication sample. Let the resulting complete data set with replication weights be denoted by

$$\bar{\mathcal{O}} = \left\{ (\boldsymbol{y}_i^*, \boldsymbol{x}_i^*, w_i^*, w_i^{*(1)}, \ldots, w_i^{*(B)}),\ i = 1, \ldots, n^* \right\}$$

The subsequent analysis is then carried out based on $\bar{\mathcal{O}}$. A fractional imputation estimator $\hat{\boldsymbol{\theta}}_{fi}$ of $\boldsymbol{\theta}$ is obtained by solving

$$\left\{ \sum_{i=1}^{n^*} w_i^* \right\}^{-1} \sum_{i=1}^{n^*} w_i^* \boldsymbol{U}(\boldsymbol{y}_i^*, \boldsymbol{x}_i^*; \boldsymbol{\theta}) = \boldsymbol{0}. \tag{4.5}$$

In the following two sections, we investigate the properties of $\hat{\boldsymbol{\theta}}_{fi}$ in two specific inferential problems.

4.3 Analysis of Incomplete Bivariate Ordinal Responses

Suppose the data set under consideration has two ordinal response variables subject to missingness and is denoted by $\mathcal{O} = \big\{(\boldsymbol{r}_i, \boldsymbol{y}_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\big\}$, where $\boldsymbol{y}_i = (y_{i1}, y_{i2})$ are the two responses on J- and K-level scales, respectively, and both are partially observed, with $\boldsymbol{r}_i = (r_{i1}, r_{i2})$ being the corresponding response indicators: $r_{it} = 1$ if y_{it} is observed and $r_{it} = 0$ otherwise, $t = 1, 2$. Units in the sample can be partitioned into four groups, depending on the missing pattern of the responses:

$$\begin{aligned} \mathcal{R} &= \big\{i:\ \delta_{i1} = 1, \delta_{i2} = 1\big\}, \quad & \mathcal{P}_1 &= \big\{i:\ \delta_{i1} = 1, \delta_{i2} = 0\big\}, \\ \mathcal{P}_2 &= \big\{i:\ \delta_{i1} = 0, \delta_{i2} = 1\big\}, \quad & \mathcal{M} &= \big\{i:\ \delta_{i1} = 0, \delta_{i2} = 0\big\}. \end{aligned}$$

The research interest lies in estimating the marginal distributions and measuring the association of the bivariate ordinal responses as discussed in Section 2.3. In this section, the probability mass function of a discrete random variable is denoted by $f(\cdot)$.

4.3.1 Analysis Based on Fractional Imputation

To implement the general fractional imputation procedure proposed above for the current case, we impose two regression models on the data as in (4.1). To be more specific, we consider:

$$\begin{aligned} G_1^{-1}\big[\gamma_{1j}(\boldsymbol{X})\big] &= \alpha_{1j} - \boldsymbol{\beta}_1'\boldsymbol{X}, \\ G_2^{-1}\big[\gamma_{2k}(Y_1, \boldsymbol{X})\big] &= \alpha_{2k} - \boldsymbol{\beta}_2'\boldsymbol{X} - \sum_{j=2}^{J} \nu_j \boldsymbol{I}(Y_1 = j), \end{aligned} \tag{4.6}$$

where $\gamma_{1j}(\boldsymbol{X}) = P(Y_1 \le j \mid \boldsymbol{X})$ and $\gamma_{2k}(Y_1, \boldsymbol{X}) = P(Y_2 \le k \mid Y_1, \boldsymbol{X})$ are the cumulative probabilities given the covariates, and G_1 and G_2 are link functions. Let

$\boldsymbol{\eta}_1 = (\alpha_{11}, ..., \alpha_{1J}, \boldsymbol{\beta}_1')'$ and $\boldsymbol{\eta}_2 = (\alpha_{21}, ..., \alpha_{2K}, \boldsymbol{\beta}_2', \nu_2, ..., \nu_J)'$ be the parameters in the models (4.6). Both models belong to the *cumulative link model* family of the form (2.16), but our proposed method can be easily adapted to more complex parametric forms and other ordinal regression models of form (2.15), such as the *continuation-ratio link models*. The dummy variable approach is used to incorporate ordinal predictors in the second model.

A practical question regarding (4.6) is on which response variable to be used for the first model. The decision could be based on results from two preliminary model fittings for each response variable using complete-case analysis and choose the better fitted model. Another important factor to consider is the amount of observed units for each variable. Modelling the response with a larger proportion of observed values provides a more accurate starting point. Let Y_1 be the response variable chosen for the first model.

The conditional distributions required in (4.1) are fully determined by models (4.6):

$$\begin{aligned} f(j \mid \boldsymbol{x}; \boldsymbol{\eta}_1) &= P(Y_1 = j \mid \boldsymbol{x}; \boldsymbol{\eta}_1) = P(Y_1 \leq j \mid \boldsymbol{x}; \boldsymbol{\eta}) - P(Y_1 \leq j-1 \mid \boldsymbol{x}; \boldsymbol{\eta}_1), \\ f(k \mid j, \boldsymbol{x}; \boldsymbol{\eta}_2) &= P(Y_2 = k \mid Y_1 = j, \boldsymbol{x}; \boldsymbol{\eta}_2) \\ &= P(Y_2 \leq k \mid Y_1 = j, \boldsymbol{x}; \boldsymbol{\eta}_2) - P(Y_2 \leq k-1 \mid Y_1 = j, \boldsymbol{x}; \boldsymbol{\eta}_2). \end{aligned}$$

Following the two stages in Section 4.2, we create a single weighted complete data file denoted by $\mathcal{O}^* = \left\{(e_i^*, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*, w_i^*),\ i = 1, \ldots, n^*\right\}$. The initial values $\boldsymbol{\eta}_1^{(0)}, \boldsymbol{\eta}_2^{(0)}$ in Step 1 of *Stage Two* could be the estimates obtained by the available-case analysis method for the models in (4.6). More specifically, we can fit the model for Y_1 with data from $\mathcal{R}$ and $\mathcal{P}_1$, and fit the model for Y_2 with data from $\mathcal{R}$ alone and use the resulting estimates as $\boldsymbol{\eta}_1^{(0)}, \boldsymbol{\eta}_2^{(0)}$. A practical issue is that when the size of group $\mathcal{R}$ is too small, the transitional model may not be numerically identifiable. Should that be the case, we take initial values of ν_j in the second model as 0 and estimate the remaining parameters in $\boldsymbol{\eta}_2$ with data from $\mathcal{R}$ and $\mathcal{P}_2$.

Table 4.1 shows the structure of the imputed data set for a toy example with $n = 4$ observations, one for each of the four groups $\mathcal{R}$, $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{M}$. The bivariate ordinal response variables each has two levels ($J = K = 2$) and there are three auxiliary variables. The imputed data set is an enlarged data file with the same number of variables as the initial sample and a total number of $n^* = n_r + Kn_{p1} + Jn_{p2} + JKn_m$

observations, where n_r, n_{p1}, n_{p2} and n_m are the size of group $\mathcal{R}$, $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{M}$, respectively. For the simple example shown in Table 4.1 we have $n_r = n_{p1} = n_{p2} = n_m = 1$, $J = K = 2$ and $n^* = 9$.

Table 4.1: A Simple Example of Fractionally Imputed Data Set with $J = K = 2$ and $n = 4$

e_i^*	r_{i1}	r_{i2}	y_{i1}	y_{i2}	x_{i1}	x_{i2}	x_{i3}	i	w_i^*
1	1	1	y_{11}	y_{12}	x_{11}	x_{12}	x_{13}	1	w_1^*
2	1	0	y_{21}	1	x_{21}	x_{22}	x_{23}	2	w_2^*
2	1	0	y_{21}	2	x_{21}	x_{22}	x_{23}	3	w_3^*
3	0	1	1	y_{32}	x_{31}	x_{32}	x_{33}	4	w_4^*
3	0	1	2	y_{32}	x_{31}	x_{32}	x_{33}	5	w_5^*
4	0	0	1	1	x_{41}	x_{42}	x_{43}	6	w_6^*
4	0	0	1	2	x_{41}	x_{42}	x_{43}	7	w_7^*
4	0	0	2	1	x_{41}	x_{42}	x_{43}	8	w_8^*
4	0	0	2	2	x_{41}	x_{42}	x_{43}	9	w_9^*

Note that both the marginal distributions and association measures rely on the cell probabilities $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{1K}, \ldots, \pi_{J1}, \ldots, \pi_{JK})'$, which are defined by the estimating functions $\boldsymbol{U}(y_1, y_2; \boldsymbol{\pi}) = (U_{11}, \ldots, U_{1K}, \ldots, U_{J1}, \ldots, U_{JK})'$ where

$$U_{jk} = \boldsymbol{I}(y_1 = j, y_2 = k) - \pi_{jk}, \quad \text{for } j = 1, \ldots, J, \quad k = 1, \ldots, K.$$

From the general formula (4.5), an estimator $\hat{\boldsymbol{\pi}}^{fi}$ based on the fractionally imputed data set is the solution to

$$\left\{ \sum_{i=1}^{n^*} w_i^* \right\}^{-1} \sum_{i=1}^{n^*} w_i^* \boldsymbol{U}(y_{i1}^*, y_{i2}^*; \boldsymbol{\pi}) = \mathbf{0}, \tag{4.7}$$

or equivalently, $\hat{\boldsymbol{\pi}}^{fi} = (\hat{\pi}_{11}^{fi}, \ldots, \hat{\pi}_{1K}^{fi}, \ldots, \hat{\pi}_{J1}^{fi}, \ldots, \hat{\pi}_{JK}^{fi})'$, where

$$\hat{\pi}_{rj}^{fi} = \sum_{i=1}^{n^*} w_i^* \boldsymbol{I}(y_{i1}^* = j, y_{i2}^* = k) / \sum_{i=1}^{n^*} w_i^* \,. \tag{4.8}$$

The marginal probabilities π_{j+} of Y_1 can be similarly estimated by

$$\hat{\pi}_{j+}^{fi} = \sum_{i=1}^{n^*} w_i^* \boldsymbol{I}(y_{i1}^* = j) / \sum_{i=1}^{n^*} w_i^* \,. \tag{4.9}$$

The association parameter γ can be estimated by

$$\hat{\gamma}^{fi} = \big(C^{fi} - D^{fi}\big) / \big(C^{fi} + D^{fi}\big), \tag{4.10}$$

where $C^{fi} = 2\sum_{j<a}\sum_{k<b}\hat{\pi}_{jk}^{fi}\hat{\pi}_{ab}^{fi}$ and $D^{fi} = 2\sum_{j<a}\sum_{k>b}\hat{\pi}_{jk}^{fi}\hat{\pi}_{ab}^{fi}$. In general, any parameters in the form of $\boldsymbol{g}(\boldsymbol{\pi})$ for a differentiable function $\boldsymbol{g}(\cdot)$ can be estimated by $\boldsymbol{g}(\hat{\boldsymbol{\pi}}^{fi})$.

Other analyses, such as fitting regression models involving Y_1 and a subset of covariates shown in Chapter 3, can be carried out in similar ways as in (4.7) by solving weighted estimating equations with the fractionally imputed data set. In the following sections, we first address the convergence issue of the fractional weights calculated from the iterative procedure introduced in Section 4.2 and then derive the asymptotic properties for the fractional imputation estimators given in (4.8), (4.9) and (4.10).

4.3.2 Convergence of the Fractional Weights

We now demonstrate that the weights from the proposed fractional imputation procedure do converge to a set of stable values. We start by scrutinizing the fractional weights received by observations in $\mathcal{O}^*$ that correspond to the i_0th observation in the original data. If $i_0 \in \mathcal{R}$, this observation remains the same in $\mathcal{O}^*$ and receives weight 1. If $i_0 \in \mathcal{P}_1$, from STEP 5 of *Stage One*, it is replicated K time in the imputed data set with missing values of Y_2 filled in with $1, \ldots, K$. The set $S(i_0)$ defined in (4.4) consists of these K units. By the general weight function (4.4), for any unit $(\boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*)$ with $i \in S(i_0)$,

$$\begin{aligned} W(i_0, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) &= \frac{f(y_{i_0 1} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1) f(y_{i2}^* \mid y_{i_0 1}, \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_2)}{\sum_{l \in S(i_0)} f(y_{i_0 1} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1) f(y_{l2}^* \mid y_{i_0 1}, \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_2)} \\ &= f(y_{i2}^* \mid y_{i_0 1}, \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_2). \end{aligned} \tag{4.11}$$

Similarly, it can be shown that, if $i_0 \in \mathcal{P}_2$,

$$W(i_0, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = f(y_{i1}^* \mid y_{i_0 2}, \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2), \tag{4.12}$$

and if $i_0 \in \mathcal{M}$,

$$W(i_0, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = f(y_{i1}^*, y_{i2}^* \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2). \tag{4.13}$$

In what follows, we reveal the link between the fractional imputation procedure and the EM algorithm (Dempster et al. 1977) used in the likelihood approach to estimating parameters in (4.6). This finding leads to our statement on the convergence of the iterative procedure for computing the fractional weights.

The likelihood function of the observed data is given by

$$\begin{aligned} L_{obs} &= \prod_{i=1}^{n} \int f(r_{i1}, r_{i2}, y_{i1}, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) d\mu(\boldsymbol{y}_{i,mis}) \\ &= \prod_{i=1}^{n} \int f(r_{i1}, r_{i2} \mid \boldsymbol{x}_i, y_{i1}, y_{i2}) f(y_{i1}, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) d\mu(\boldsymbol{y}_{i,mis}) \,, \end{aligned}$$

where $\boldsymbol{y}_{i,mis}$ is the missing part of the bivariate responses. Under the MAR assumption, response indicators and the missing responses are conditionally independent given the observed responses and covariates, i.e., $f(r_1, r_2 \mid \boldsymbol{x}, y_1, y_2) = f(r_1, r_2 \mid \boldsymbol{x}, \boldsymbol{y}_{obs})$, which does not involve $\boldsymbol{y}_{mis}$ and hence can be taken to the outside of the integral. We can re-write L_{obs} into two parts as

$$L_{obs} = \prod_{i=1}^{n} f(r_{i1}, r_{i2} \mid \boldsymbol{x}_i, \boldsymbol{y}_{i,obs}) \prod_{i=1}^{n} \int f(y_{i1}, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) d\mu(\boldsymbol{y}_{i,mis}) \,.$$

We assume the separability condition (Molenberghs and Kenward 2007) holds when considering likelihood-based approaches hereafter, so only the second part in L_{obs} involving parameters $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ is of interest. Noting that Y_1, Y_2 are discrete variables, the integrals can be written as summations over all possible values. By considering

the four groups of sampled units separately, we can re-write L_{obs} as

$$L_{obs} \propto \prod_{i\in\mathcal{R}} f(y_{i1}, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \times \prod_{i\in\mathcal{P}_1} \Big[\sum_{y_2=1}^{K} f(y_{i1}, y_2 \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)\Big] \times \prod_{i\in\mathcal{P}_2} \Big[\sum_{y_1=1}^{J} f(y_1, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)\Big], \tag{4.14}$$

where $f(y_1, y_2 \mid \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = f(y_1 \mid \boldsymbol{x}; \boldsymbol{\eta}_1) f(y_2 \mid y_1, \boldsymbol{x}; \boldsymbol{\eta}_2)$, which can be obtained from (4.6). The term involving group $\mathcal{M}$ vanishes because the double summation of the joint probability mass function equals 1.

It follows from (4.14) that the log-likelihood function is given by

$$\begin{aligned} l_{obs} \;=\; & \sum_{i\in\mathcal{R}} \log\big[f(y_{i1} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1)\big] + \sum_{i\in\mathcal{R}} \log\big[f(y_{i2} \mid y_{i1}, \boldsymbol{x}_i; \boldsymbol{\eta}_2)\big] + \\ & \sum_{i\in\mathcal{P}_1} \log\big[f(y_{i1} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1)\big] + \sum_{i\in\mathcal{P}_2} \log\Big[\sum_{y_1=1}^{J} f(y_1 \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1) f(y_{i2} \mid y_1, \boldsymbol{x}_i; \boldsymbol{\eta}_2)\Big]. \end{aligned} \tag{4.15}$$

By taking derivatives of l_{obs} with respect to $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ and setting them equal to zeros, we obtain the set of score functions as

$$\begin{aligned} \mathbf{0} &= \sum_{i\in\mathcal{R},\mathcal{P}_1} \boldsymbol{S}_1(y_{i1}, \boldsymbol{x}_i; \boldsymbol{\eta}_1) + \sum_{i\in\mathcal{P}_2} \sum_{y_1=1}^{J} \boldsymbol{S}_1(y_1, \boldsymbol{x}_i; \boldsymbol{\eta}_1) f(y_1 \mid y_{i2}, \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2), \\ \mathbf{0} &= \sum_{i\in\mathcal{R}} \boldsymbol{S}_2(y_{i2}, y_{i1}, \boldsymbol{x}_i; \boldsymbol{\eta}_2) + \sum_{i\in\mathcal{P}_2} \sum_{y_1=1}^{J} \boldsymbol{S}_2(y_{i2}, y_1, \boldsymbol{x}_i; \boldsymbol{\eta}_2) f(y_1 \mid y_{i2}, \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2), \end{aligned} \tag{4.16}$$

where

$$\begin{aligned} \boldsymbol{S}_1(y_1, \boldsymbol{x}; \boldsymbol{\eta}_1) &= \frac{\partial}{\partial \boldsymbol{\eta}_1'} \log\big[f(y_1 \mid \boldsymbol{x}; \boldsymbol{\eta}_1)\big], \\ \boldsymbol{S}_2(y_2, y_1, \boldsymbol{x}; \boldsymbol{\eta}_2) &= \frac{\partial}{\partial \boldsymbol{\eta}_2'} \log\big[f(y_2 \mid y_1, \boldsymbol{x}; \boldsymbol{\eta}_2)\big] \end{aligned}$$

are the score functions of $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ when the two models in (4.6) are fitted separately

with complete data and

$$f(y_1 \mid y_2, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \frac{f(y_1 \mid \boldsymbol{x}; \boldsymbol{\eta}_1) f(y_2 \mid y_1, \boldsymbol{x}; \boldsymbol{\eta}_2)}{\sum_{y_1=1}^{J} f(y_1 \mid \boldsymbol{x}; \boldsymbol{\eta}_1) f(y_2 \mid y_1, \boldsymbol{x}; \boldsymbol{\eta}_2)} \tag{4.17}$$

is the derived conditional probability mass function of Y_1 given Y_2 and $\boldsymbol{X}$.

It is difficult to solve the score equations (4.16) directly. An alternative approach is to apply the EM algorithm to find the maximum likelihood estimators of $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2$.

E-step Calculate

$$Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}) = E\Big\{ \sum_{i=1}^{n} \log\big[f(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \big] \mid \boldsymbol{y}_{obs}, \boldsymbol{r}, \boldsymbol{x}; \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)} \Big\},$$

where $\boldsymbol{y}_{obs}$ denotes the observed part of $\boldsymbol{y}$. Following the same partition used for L_{obs}, we can re-write $Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})$ into four terms:

$$\begin{aligned} Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}) = & \sum_{i \in \mathcal{R}} \log\big[f(y_{i1}, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \big] \\ & + \sum_{i \in \mathcal{P}_1} \sum_{y_2=1}^{K} \log\big[f(y_{i1}, y_2 \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \big] f(y_2 \mid y_{i1}, \boldsymbol{x}_i; \boldsymbol{\eta}_2^{(t)}) \\ & + \sum_{i \in \mathcal{P}_2} \sum_{y_1=1}^{J} \log\big[f(y_1, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \big] f(y_1 \mid y_{i2}, \boldsymbol{x}_i; \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}) \\ & + \sum_{i \in \mathcal{M}} \sum_{y_1=1}^{J} \sum_{y_2=1}^{K} \log\big[f(y_1, y_2 \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \big] f(y_1, y_2 \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}) \,. \end{aligned} \tag{4.18}$$

M-step Obtain $\boldsymbol{\eta}_1^{(t+1)}$ and $\boldsymbol{\eta}_2^{(t+1)}$ which maximize $Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})$ with respect to $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$. Note that $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ in $Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})$ are separable. This leads to simpler forms of score functions. For example, the score equations for $\boldsymbol{\eta}_1$ are given

by

$$\begin{aligned}\mathbf{0} &= \sum_{i\in\mathcal{R}} \boldsymbol{S}_1(y_{i1}, \boldsymbol{x}_i; \boldsymbol{\eta}_1) + \sum_{i\in\mathcal{P}_1}\sum_{y_2=1}^{K} f(y_2 \mid y_{i1}, \boldsymbol{x}_i; \boldsymbol{\eta}_2^{(t)})\boldsymbol{S}_1(y_{i1}, \boldsymbol{x}_i; \boldsymbol{\eta}_1) \\ &+ \sum_{i\in\mathcal{P}_2}\sum_{y_1=1}^{J} f(y_1 \mid y_{i2}, \boldsymbol{x}_i; \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})\boldsymbol{S}_1(y_1, \boldsymbol{x}_i; \boldsymbol{\eta}_1) \\ &+ \sum_{i\in\mathcal{M}}\sum_{y_1=1}^{J}\sum_{y_2=1}^{K} f(y_1, y_2 \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})\boldsymbol{S}_1(y_1, \boldsymbol{x}_i; \boldsymbol{\eta}_1)\,. \end{aligned} \tag{4.19}$$

It is important for our following arguments to note that (4.19) are the same as the score equations obtained by fitting the first model in (4.6) with the imputed data set weighted by $\boldsymbol{w}^{*(t)} = (w_1^{*(t)}, ..., w_{n^*}^{*(t)})$, where $w_i^{*(t)} = W(e_i^*, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*; \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})$. Same results can also be shown for $\boldsymbol{\eta}_2$. In other words, our proposed joint fractional imputation procedures have the same spirit as the EM algorithm.

The convergence properties of the EM algorithm were studied by Wu (1983). In our case, $Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})$ is continuous with respect to $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}$, and hence the EM sequence $\left\{\boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}\right\}$ converges to a stationary point $(\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2)$ which is the solution to the score equations (4.16). We summarize the above discussions with the following theorem.

Theorem 4.1. *The fractional weights* $\left\{\boldsymbol{w}^{*(t)}\right\}$ *defined in the proposed fractional imputation procedure converge to a stable set of values denoted by* $\boldsymbol{w}^*$ *as* $t \to \infty$*, and the* i*th element of* $\boldsymbol{w}^*$ *is given by*

$$w_i^* = W(e_i^*, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*; \hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2),$$

where $(\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2)$ *is the solution to the score equations (4.16).*

4.3.3 Asymptotic Properties of Fractional Imputation Estimators

We begin with the estimator $\hat{\boldsymbol{\pi}}^{fi} = (\hat{\pi}_{11}^{fi}, ..., \hat{\pi}_{1K}^{fi}, ..., \hat{\pi}_{J1}^{fi}, ..., \hat{\pi}_{JK}^{fi})'$ of the vector $\boldsymbol{\pi}$ of joint cell probabilities, where $\hat{\pi}_{rj}^{fi}$ is given in (4.8). Note that $\hat{\pi}_{rj}^{fi}$ is a weighted sum of indicator functions of "non-independent" observations in the imputed data file. To

investigate the asymptotic behaviour of $\hat{\pi}_{rj}^{fi}$, it is essential to write it in the form of the original sample.

For the fractionally imputed data set, the i_0th observation in the original sample with one or both missing responses corresponds to a "cluster of observations" indexed by $S(i_0)$ in the imputed file. Consider the case $i_0 \in \mathcal{P}_1$, units in $S(i_0)$ have the same values for $(Y_1, \boldsymbol{X})$ equal to $(y_{i_01}, \boldsymbol{x}_{i_0})$ and the missing values of Y_2 are filled with $1, \ldots, K$, therefore, we have

$$\sum_{i \in S(i_0)} w_i^* \boldsymbol{I}(y_{i1}^* = j, y_{i2}^* = k) = \sum_{i \in S(i_0)} w_i^* \boldsymbol{I}(y_{i_01}^* = j, y_{i2}^* = k)$$

and at most one term on the right hand side is non-zero, which is $w_{i_0k}\boldsymbol{I}(y_{i_01} = j)$ where $w_{i_0k} = W(i_0, \boldsymbol{r}_{i_0}, (y_{i_01}, k), \boldsymbol{x}_{i_0}; \hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2)$. Similar arguments can be made for observations from other groups. Define the estimating function for π_{rj} as

$$\begin{aligned} U_{jk}^{fi}(e, \boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \pi_{jk}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \;=\; & r_1 r_2 \boldsymbol{I}(y_1 = j, y_2 = k) \\ & + r_1(1 - r_2) W(e, \boldsymbol{r}, (y_1, k), \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \boldsymbol{I}(y_1 = j) \\ & + (1 - r_1) r_2 W(e, \boldsymbol{r}, (j, y_2), \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \boldsymbol{I}(y_2 = k) \\ & + (1 - r_1)(1 - r_2) W(e, \boldsymbol{r}, (j, k), \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) - \pi_{rj} \,. \end{aligned} \tag{4.20}$$

It can be seen that $\hat{\pi}_{rj}^{fi}$ given in (4.8) is the same as the solution to the estimating equation

$$0 = n^{-1} \sum_{i=1}^{n} U_{jk}^{fi}(e_i, \boldsymbol{r}_i, \boldsymbol{y}_i, \boldsymbol{x}_i; \pi_{jk}, \hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2), \tag{4.21}$$

which depends on preliminary estimators of $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$. This two-step estimator $\hat{\pi}_{rj}$ can be more conveniently handled as a component of solutions to an extended system of estimating equations. Let

$$\begin{aligned} \boldsymbol{S}_{obs}^{(1)}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) &= r_1 \boldsymbol{S}_1(y_1, \boldsymbol{x}; \boldsymbol{\eta}_1) + (1 - r_1) r_2 E\big[\boldsymbol{S}_1(y_1, \boldsymbol{x}; \boldsymbol{\eta}) \mid y_2, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2\big], \\ \boldsymbol{S}_{obs}^{(2)}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) &= r_1 r_2 \boldsymbol{S}_2(y_2, y_1, \boldsymbol{x}; \boldsymbol{\eta}_2) \\ &\quad + (1 - r_1) r_2 E\big[\boldsymbol{S}_2(y_2, y_1, \boldsymbol{x}; \boldsymbol{\eta}_2) \mid y_2, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2\big] \,. \end{aligned} \tag{4.22}$$

The estimators $(\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2)$ are initially defined as the solution to the score equations

(4.16) and can be re-written as the solution to

$$\mathbf{0} \;=\; \sum_{i=1}^{n} \boldsymbol{S}_{obs}^{(1)}(\boldsymbol{r}_i, \boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2), \quad \mathbf{0} \;=\; \sum_{i=1}^{n} \boldsymbol{S}_{obs}^{(2)}(\boldsymbol{r}_i, \boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2). \tag{4.23}$$

Let $\boldsymbol{U}^{fi}(\boldsymbol{\pi}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = (U_{11}^{fi}, ..., U_{1K}^{fi}, ..., U_{J1}^{fi}, ..., U_{JK}^{fi})'$, $\boldsymbol{S}_{obs}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = (\boldsymbol{S}_{obs}^{(1)'}, \boldsymbol{S}_{obs}^{(2)'})'$ and $\boldsymbol{S}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = (\boldsymbol{S}_1', \boldsymbol{S}_2')'$, where U_{jk}^{fi}, $\boldsymbol{S}_{obs}^{(1)}$, $\boldsymbol{S}_{obs}^{(2)}$, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are short forms of functions defined in (4.20), (4.16) and (4.23). The following theorem summarizes the asymptotic properties of $\hat{\boldsymbol{\pi}}^{fi}$. Proofs are outlined in Section 4.5.

Theorem 4.2. *Let $\boldsymbol{\pi}_0$, $\boldsymbol{\eta}_{10}$ and $\boldsymbol{\eta}_{20}$ be the true values of $\boldsymbol{\pi}$, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$. Under the regularity conditions specified in Section 4.5, $\hat{\boldsymbol{\pi}}^{fi}$ with elements given by (4.8) is a consistent estimator of $\boldsymbol{\pi}_0$. Furthermore,*

$$n^{1/2}(\hat{\boldsymbol{\pi}}^{fi} - \boldsymbol{\pi}_0) \;\sim\; \boldsymbol{N}\Big(\mathbf{0}, Var\big[\boldsymbol{U}^{fi}(\boldsymbol{\pi}_0, \boldsymbol{\eta}_{10}, \boldsymbol{\eta}_{20}) + \boldsymbol{\kappa}\boldsymbol{I}_{obs}^{-1}\boldsymbol{S}_{obs}(\boldsymbol{\eta}_{10}, \boldsymbol{\eta}_{20})\big]\Big),$$

where "$\sim$" represents "is asymptotically distributed as",

$$\boldsymbol{I}_{obs} = \big(E\big[-\partial \boldsymbol{S}_{obs}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)/\partial \boldsymbol{\eta}_1\big], E\big[-\partial \boldsymbol{S}_{obs}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)/\partial \boldsymbol{\eta}_2\big]\big),$$

evaluated at the true values of the parameters, $\boldsymbol{\kappa} = (\kappa_{11}', ..., \kappa_{1K}', ..., \kappa_{J1}', ...\kappa_{JK}')'$, and

$$\kappa_{jk} = E\Big\{\boldsymbol{I}(y_1 = j, y_2 = k)\big[\boldsymbol{S}((j,k), \boldsymbol{x}; \boldsymbol{\eta}_{10}, \boldsymbol{\eta}_{20}) - \boldsymbol{S}_{obs}(\boldsymbol{r}, (j,k), \boldsymbol{x}; \boldsymbol{\eta}_{10}, \boldsymbol{\eta}_{20})\big]'\Big\}.$$

Corollary 4.2.1. *Let $\boldsymbol{g}(\boldsymbol{\pi})$ be a differentiable function of $\boldsymbol{\pi}$, either scalar or vector valued. Denote the asymptotic variance of $n^{1/2}(\hat{\boldsymbol{\pi}}^{fi} - \boldsymbol{\pi}_0)$ given in Theorem 4.2 as $\boldsymbol{\Sigma}^{fi}$. Then $\boldsymbol{g}(\hat{\boldsymbol{\pi}}^{fi})$ is a consistent estimator of $\boldsymbol{g}(\boldsymbol{\pi})$ and*

$$n^{1/2}\Big[\boldsymbol{g}(\hat{\boldsymbol{\pi}}^{fi}) - \boldsymbol{g}(\boldsymbol{\pi}_0)\Big] \;\sim\; \boldsymbol{N}\big(\mathbf{0},\; \boldsymbol{\Gamma}\boldsymbol{\Sigma}^{fi}\boldsymbol{\Gamma}'\big),$$

where $\boldsymbol{\Gamma} = \partial \boldsymbol{g}(\boldsymbol{\pi})/\partial \boldsymbol{\pi}$ and is evaluated at $\boldsymbol{\pi}_0$.

The corollary follows directly from the Continuous Mapping Theorem and the Delta method. The marginal probabilities and association measures are all special cases with different $\boldsymbol{g}(\cdot)$. For example, the marginal probabilities of y_1 can be written as $\boldsymbol{\pi}_1 = \boldsymbol{C}\boldsymbol{\pi}$, where $\boldsymbol{C} = \text{diag}(\mathbf{1}', \ldots, \mathbf{1}')$ is a $J \times (JK)$ block diagonal matrix and $\mathbf{1} = (1, \ldots, 1)'$ with length K. It follows that $\hat{\boldsymbol{\pi}}_1^{fi}$ with elements given in (4.9) is

asymptotically normal with mean being the true value of $\boldsymbol{\pi}_1$ and variance-covariance matrix $\boldsymbol{C}\boldsymbol{\Sigma}^{fi}\boldsymbol{C}'$. For the association parameter γ, we have

$$\boldsymbol{\Gamma} = \frac{\partial \gamma}{\partial \boldsymbol{\pi}}\bigg|_{\boldsymbol{\pi}_0} = \frac{(1-\gamma_0)^2}{2\prod_{c0}} \frac{\partial \prod_c}{\partial \boldsymbol{\pi}}\bigg|_{\boldsymbol{\pi}_0} - \frac{(1+\gamma_0)^2}{2\prod_{d0}} \frac{\partial \prod_d}{\partial \boldsymbol{\pi}}\bigg|_{\boldsymbol{\pi}_0},$$

where γ_0, $\prod_{c0}$ and $\prod_{d0}$ are true values of the terms defined in (2.13).

Variance estimation can be done through the linearization method, which replaces unknown population quantities in asymptotic variances given in Corollary 4.2.1 with weighted sample estimates from the imputed data set. For example, the quantity κ_{jk} defined in Theorem 4.2 can be estimated by $\hat{\kappa}_{jk}$, which is computed as

$$\left\{\sum_{m=1}^{n^*} w_i^*\right\}^{-1} \sum_{i=1}^{n^*} w_i^* \boldsymbol{I}(y_{i1}^* = j, y_{i2}^* = k)\Big[\boldsymbol{S}((j,k), \boldsymbol{x}_i^*; \hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2) - \boldsymbol{S}_{obs}(\boldsymbol{r}_i^*, (j,k), \boldsymbol{x}_i^*; \hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2)\Big].$$

For public use data files, however, the resampling methods we mentioned in Section 4.2 are always preferred.

4.3.4 Simulation Studies

We report results from simulation studies on the finite sample performance of the estimators based on the fractionally imputed data file, with comparisons to existing methods. We consider (Y_1, Y_2), each with three categories, and two covariates: a continuous variable X_1 generated from Exp(1) and a discrete variable X_2 following Bernoulli(0.5). The responses (Y_1, Y_2) follow the models given in (4.6). In order to apply the IPW method, we simulate the response indicators under the CDM assumption and the full-response probability follows a logistic regression model. The SRMI technique is implemented using the package MICE (Buuren and Groothuis-Oudshoorn 2011) with default non-informative priors for the imputation procedure.

The missing data process is carefully chosen such that the proportions of units in the four groups $\mathcal{R}$, $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{M}$ are controlled to have desirable patterns to mimic two real world scenarios. The first scenario has the majority of the sample fully observed, with proportions being $(50\%, 20\%, 20\%, 10\%)$ for the four groups. For the second scenario, only one of the two responses is observed for the majority of sampled units, with the proportions being $(20\%, 30\%, 40\%, 10\%)$. The simulation

Table 4.2: Absolute Relative Bias (%) and Mean Squared Error ($\times 10^{-4}$) of Estimators of $\pi_{+1} = P(Y_2 = 1)$

RP	n		COMP	CCA	IPW	SRMI5	FI
5221	200	ARB	0.2	16.5	0.04	1.0	0.03
		MSE	(8.9)	(30.8)	(12.1)	(11.8)	(11.6)
	500	ARB	0.33	17.1	0.3	0.2	0.3
		MSE	(3.6)	(23.4)	(5.2)	(5.0)	(4.9)
2341	200	ARB	—	57.2	0.008	1.5	0.1
		MSE	—	(251.5)	(12.9)	(12.5)	(12.2)
	500	ARB	—	56.6	0.2	0.2	0.3
		MSE	—	(208.8)	(5.1)	(5.1)	(4.9)

studies consist of three parts: (i) Point estimators; (ii) Variance estimators; and (iii) Tests of independence.

Table 4.2 presents results from the first part of the simulation on Absolute Relative Bias (ARB, in %) and Mean Squared Error (MSE, multiplied by 10^4) of different estimators of the first element π_{+1} of the marginal probabilities of Y_2 under the two response patterns (RP, indicated by 5221 and 2341) and two sample sizes $n = 200$ and $n = 500$. The complete sample estimator without any missing values is denoted by COMP and is listed as the gold-standard reference; the estimator from complete-case analysis is denoted as CCA; the inverse probability weighting estimator is indicated by IPW; the SRMI method with 5 imputed data sets is denoted by SRMI5. Our proposed fractional imputation estimator is denoted by FI. Simulation results for the association measure γ are summarized in Table 4.3.

The simulation results show clearly that the CCA estimator is not consistent for either the marginal probability π_{+1} or the association measure γ. The other three methods IPW, SRMI and FI provide comparable results for estimating marginal probabilities with negligible biases. However, for the estimation of γ, the IPW estimator is far less efficient than the two imputation-based estimators. The SRMI estimator

Table 4.3: Absolute Relative Bias (%) and Mean Squared Error ($\times 10^{-3}$) of Estimators of γ

RP	n		COMP	CCA	IPW	SRMI5	FI
5221	200	ARB	0.5	8.7	0.9	1.5	0.04
		MSE	(7.3)	(15.0)	(25.4)	(13.2)	(12.7)
	500	ARB	0.04	9.1	0.4	1.1	0.1
		MSE	(2.8)	(6.9)	(13.1)	(5.3)	(5.0)
2341	200	ARB	—	10.9	5.0	10.6	0.5
		MSE	—	(42.3)	(87.8)	(27.9)	(25.1)
	500	ARB	—	12.5	3.9	7.8	0.03
		MSE	—	(17.2)	(56.8)	(11.3)	(9.5)

is close to the proposed FI estimator under the first response pattern but has unreasonably large biases under the second scenario where there are only 20% of the sampled units having both responses observed. Our proposed FI estimator performs well for all cases and is uniformly better than the alternative methods considered in the simulation.

The second part of the simulation is on variance estimation. For the SRMI method, the variance estimator uses Rubin's combining rule; for the FI method, two versions of variance estimators are considered: the linearization method (FIL) and the bootstrap method (FIB). Table 4.4 reports the Absolute Relative Bias (ARB, in %) of the variances estimators for the two parameters π_{+1} and γ. For estimating π_{+1}, all variance estimators have acceptable ARB. For estimating γ, the variance estimator of the SRMI estimator has large negative biases, which implies that the variance estimator based on Rubin's combining rule underestimates the true variance. Both the linearization and the bootstrap variance estimators for the FI method are consistent.

The third part of the simulation is on a test of independence between the two ordinal responses. We use the Wald-type test statistic given in (2.14) based on a particular pair of point and variance estimators with significant level at 0.05. By

Table 4.4: Absolute Relative Bias (%) of Variance Estimators for π_{+1} and γ

		π_{+1}			γ		
RP	n	SRMI5	FIL	FIB	SRMI5	FIL	FIB
5221	200	3.0	5.7	3.7	−16.9	5.1	3.3
	500	8.0	1.3	2.2	−16.0	4.1	3.6
2341	200	4.4	4.0	3.7	−24.2	1.4	4.5
	500	7.5	2.1	1.3	−26.4	2.3	2.3

tuning the parameters in (4.6), we simulate the power of tests for a series of cases where the true value of the association measure γ increases from 0 to 1, departing gradually from the null hypothesis of independence.

The power of a test is computed as the simulated rejection probability under the given scenario. Plots of the power function for four scenarios are shown in Figure 4.1 to Figure 4.4, corresponding to two sample sizes ($n = 200, 500$) and two missing patterns (5221 and 2341). Each plot shows the power functions of three different tests: `SRFI_non`, `SRFI_nul` and SRMI. The first test uses the regular linearization variance estimator without considering the null hypothesis; the second test uses the linearization variance estimator under the null hypothesis (i.e., $\pi_{rj} = \pi_{r+}\pi_{+j}$); the third test uses the regular point and variance estimators for the SRMI method. Test results for fractional imputation estimators using bootstrap variance estimators are very similar to the ones using linearization variance estimators and not reported here to save space. The horizontal line in each figure represents the nominal value 0.05 for the level of the test.

There are three major observations from the power functions displayed in Figure 4.1 to Figure 4.4: (i) The test based on the SRMI method has type I errors bigger than the nominal value 0.05, and it becomes more pronounced when the sample size is small or the proportion of units in $\mathcal{R}$ is small. (ii) The type I errors for the two FI-based tests are very close to the nominal value and both tests have similar power. (iii) The response patterns have significant impact on the power of the tests, with the pattern 5221 producing more powerful tests than the pattern 2341. The first

observation is in line with the results on underestimation of variance for the SRMI method. The second observation shows that there is no significant advantage of using the variance estimator under the null hypothesis. The last observation is in agreement with common sense since data with the pattern 5221 provide more information on the association between the two response variables than the other pattern.

Figure 4.1: Power Function with $n = 200$ and Pattern 5221

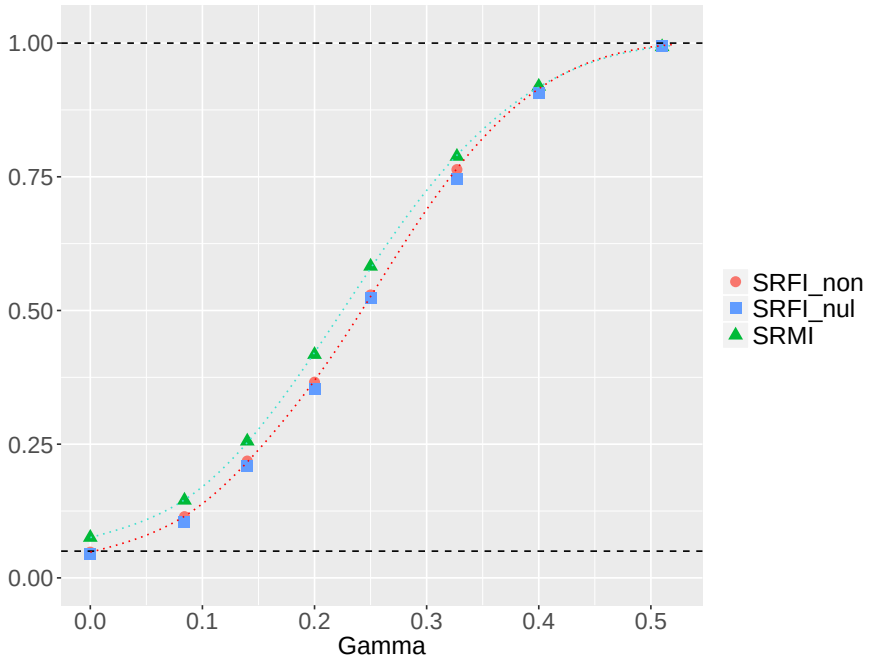


Figure 4.2: Power Function with $n = 500$ and Pattern 5221

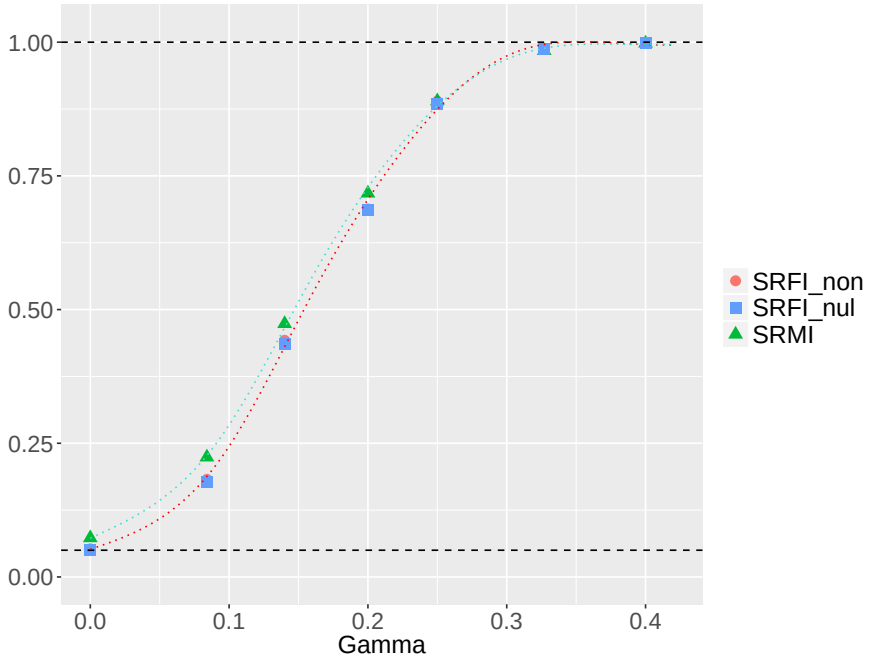


Figure 4.3: Power Function with $n = 200$ and Pattern 2341

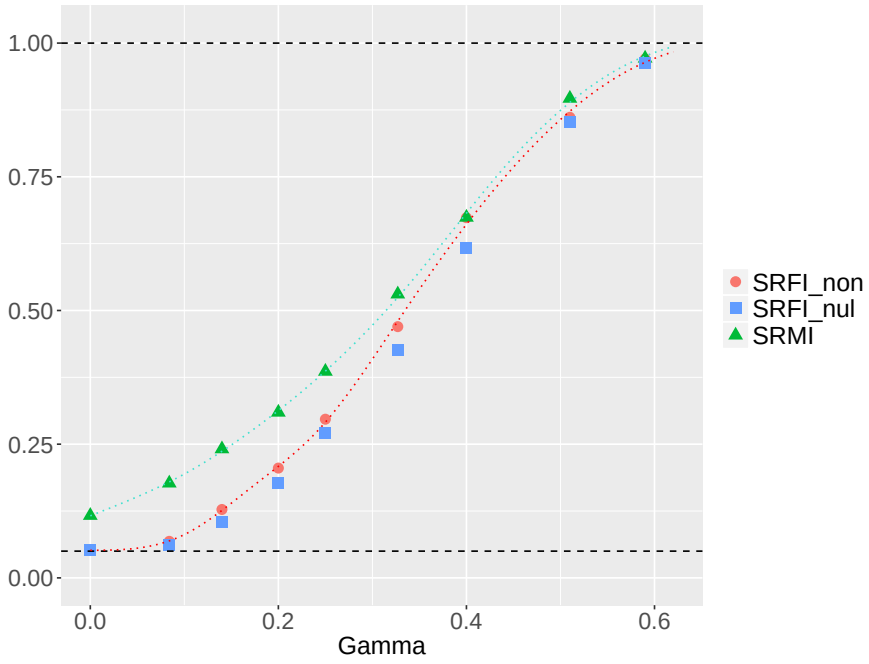
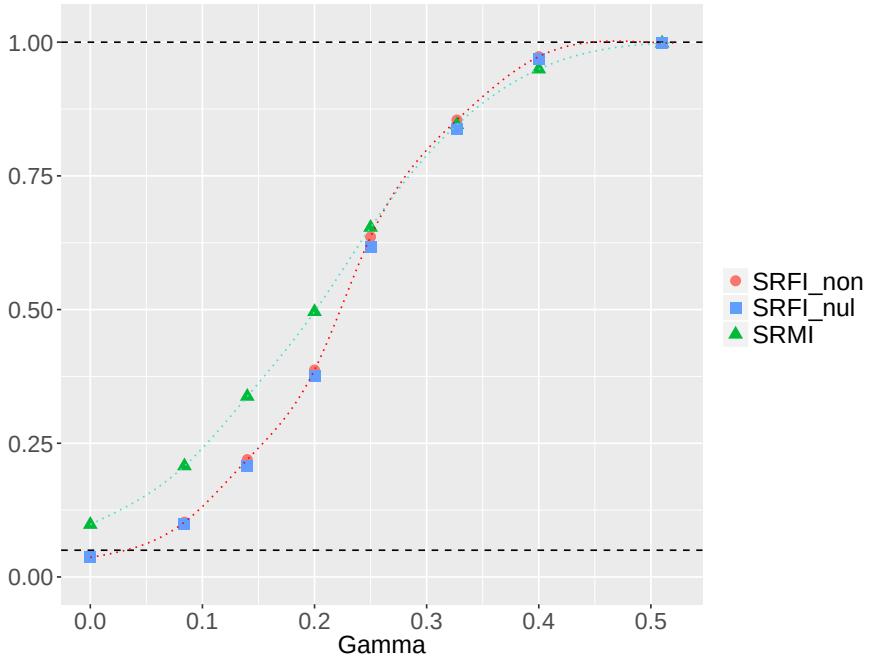


Figure 4.4: Power Function with $n = 500$ and Pattern 2341



4.4 Regression Analysis with Responses and Covariates Both Missing

In this section, we consider another practically important application of the proposed method to the estimation of regression coefficients in an ordinal regression model with both the ordinal response and one of the continuous covariates subject to missingness. The latter case involves imputing a continuous variable and an ordinal variable simultaneously, and hence the demonstration of convergence of the fractional weights and asymptotic properties differs from that in the previous section. We will mainly focus on the differences and cover the common parts briefly.

To be consistent with notation in previous sections, let Y_2 be the ordinal response variable on a J-level scale, Y_1 be the continuous covariates with missing observations and $\boldsymbol{X}$ be the remaining fully-observed covariates. The available data set is $\mathcal{O} = \{(\boldsymbol{r}_i, \boldsymbol{y}_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\}$ and is still partitioned into four groups defined in Section 4.3. An *analysis model* of form (2.15) is postulated by the data user for Y_2 with Y_1 and $\boldsymbol{X}$ as covariates. The model parameters of interest $\boldsymbol{\theta}$ are defined by the estimating function

$$\boldsymbol{U}(y_1, y_2, \boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{z}_2, (y_1, \boldsymbol{x}')'; \boldsymbol{\theta}), \tag{4.24}$$

where $\boldsymbol{z}_2$ is the cumulative indicator vector of y_2 and the function $\boldsymbol{S}$ is defined in (2.19). In this section, $f(\cdot)$ denotes the probability mass function if the random variable is discrete and the probability density function if the random variable is continuous.

4.4.1 Analysis Based on Fractional Imputation

Following the general fractional imputation procedure, we impose two regression models sequentially on Y_1 and Y_2:

$$\begin{aligned} Y_1 &= \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{X} + \epsilon, \\ G^{-1}\big[\gamma_{2j}(Y_1, \boldsymbol{X})\big] &= \alpha_{2j} - \boldsymbol{\beta}_{21}' \boldsymbol{X} - \beta_{22} Y_1, \end{aligned} \tag{4.25}$$

where $\epsilon \sim N(0, \sigma^2)$ and $\gamma_{2j}(Y_1, \boldsymbol{X}) = P(Y_2 \le j \mid Y_1, \boldsymbol{X})$. Let $\boldsymbol{\eta}_1 = (\beta_0, \boldsymbol{\beta}_1', \sigma^2)'$ and $\boldsymbol{\eta}_2 = (\alpha_{21}, \ldots, \alpha_{2J}, \boldsymbol{\beta}_{21}', \beta_{22})'$ be the parameters of the two models in (4.25). The first

model belongs to the *normal linear model* family. More complex forms such as those with non-linear predictors can be assumed if necessary. Note that for the *analysis model* and the second model in (4.25) to be compatible, they actually take the same form. Nevertheless, we still use $\boldsymbol{\theta}$ and $\boldsymbol{\eta}_2$ to denote the parameters in these two models to avoid confusion. See Section 3.4 for more discussion on model compatibility. The conditional distributions required in (4.1) are readily available from (4.25). For the continuous Y_1, the marginal distribution $h(y_1 \mid \boldsymbol{X}; \boldsymbol{\psi}_1)$ required in (4.2) is also given by the first model in (4.25). The parameter $\boldsymbol{\psi}_1$ is estimated by $\hat{\boldsymbol{\psi}}_1 = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1', \hat{\sigma}^2)'$, where $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1')'$ is the solution to

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} r_{i1}(1, \boldsymbol{x}_i')'(y_{i1} - \beta_0 - \boldsymbol{\beta}_1' \boldsymbol{x}_i)\,,$$

and $\hat{\sigma}^2 = \sum_{i=1}^n r_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}_1' \boldsymbol{x}_i)^2 / (\sum_{i=1}^n r_{i1} - p - 1)$. We create a single complete data file $\mathcal{O}^* = \left\{(e_i^*, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*),\ i = 1, \ldots, n^*\right\}$ with weights $\left\{w_i^*,\ i = 1, \ldots, n^*\right\}$ following the procedure in Section 4.2. An estimator $\hat{\boldsymbol{\theta}}^{fi}$ can be obtained by solving weighted estimating equations with the imputed data set:

$$\mathbf{0} = \left\{ \sum_{i=1}^{n^*} w_i^* \right\}^{-1} \sum_{i=1}^{n^*} w_i^* \boldsymbol{U}(y_{i1}^*, y_{i2}^*, \boldsymbol{x}_i^*; \boldsymbol{\theta}), \tag{4.26}$$

where $\boldsymbol{U}(y_1, y_2, \boldsymbol{x}; \boldsymbol{\theta})$ is defined in (4.24).

4.4.2 Convergence of the Fractional Weights

Following the same arguments as in Section 4.3.2, the likelihood function of the observed data is written as

$$\begin{aligned} L_{obs} \ &\propto \ \prod_{i=1}^{n} \int f(y_{i1}, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) d\mu(\boldsymbol{y}_{i,mis}) \\ &= \ \prod_{i \in \mathcal{R}} f(y_{i1}, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \times \prod_{i \in \mathcal{P}_1} \Big[\sum_{y_2=1}^{J} f(y_{i1}, y_2 \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \Big] \\ &\times \prod_{i \in \mathcal{P}_2} \int_{\mathcal{D}_1} f(y_1, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) dy_1, \end{aligned} \tag{4.27}$$

where $\mathcal{D}_1$ is the domain of Y_1. The log-likelihood function is then given by

$$\begin{aligned} l_{obs} &= \sum_{i\in\mathcal{R}} \log f(y_{i1}, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1) + \sum_{i\in\mathcal{P}_1} \log\Big[\sum_{y_2=1}^{J} f(y_{i1}, y_2 \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)\Big] \\ &\quad + \sum_{i\in\mathcal{P}_2} \log\Big[\int_{\mathcal{D}_1} f(y_1, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) dy_1\Big], \end{aligned} \tag{4.28}$$

which can be maximized by the EM algorithm. However, unlike the case in Section 4.3, the *Expectation* step of the algorithm requires the calculation of an integral for every iteration and is different from the proposed fractional imputation procedure. So the convergence results for the EM algorithm cannot be applied directly to the proposed procedure.

To demonstrate the convergence of the fractional weights, we define a function $l^*_{obs}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ as follows, which approximates the log-likelihood function using the imputed data set,

$$\begin{aligned} l^*_{obs}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) =& \sum_{i\in\mathcal{R}} \log\big[f(y_{i1}, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1)\big] + \sum_{i\in\mathcal{P}_1} \log\Big[\sum_{y_2=1}^{J} f(y_{i1}, y_2 \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)\Big] \\ &+ \sum_{i\in\mathcal{P}_2} \log\left[M^{-1} \sum_{l\in S(i)} \frac{f(y^*_{l1}, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{h(y^*_{l1} \mid \boldsymbol{x}_i; \boldsymbol{\psi}_1)}\right] \\ &+ \sum_{i\in\mathcal{M}} \log\left[M^{-1} \sum_{l\in S(i)} \frac{f(y^*_{l1}, y^*_{l2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{h(y^*_{l1} \mid \boldsymbol{x}_i; \boldsymbol{\psi}_1)}\right]. \end{aligned} \tag{4.29}$$

Noting that for $i \in \mathcal{P}_2$, the set $S(i)$ consists of M replications of the ith observation with missing values of Y_1 replaced by random draws $y^*_{l1} \sim h(y_1 \mid \boldsymbol{x}_i; \boldsymbol{\psi}_1)$, we have

$$\begin{aligned} M^{-1} \sum_{l\in S(i)} \frac{f(y^*_{l1}, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{h(y^*_{l1} \mid \boldsymbol{x}_i; \boldsymbol{\psi}_1)} &\to \int_{\mathcal{D}_1} \frac{f(y_1, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{h(y_1 \mid \boldsymbol{x}_i; \boldsymbol{\psi}_1)} h(y_1 \mid \boldsymbol{x}_i; \boldsymbol{\psi}_1) dy_1 \\ &= \int_{\mathcal{D}_1} f(y_1, y_{i2} \mid \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) dy_1, \end{aligned}$$

when M goes to infinity. Similarly, we can show the last term in (4.29) converges to 0. Therefore, $l^*_{obs}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \to l_{obs}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ in probability when $M \to \infty$.

From STEP 4 in *Stage Two*, the proposed procedure updates the parameters by

maximizing the "weighted log-likelihood" obtained by treating the imputed data set as if it were observed. Specifically, we find $(\boldsymbol{\eta}_1^{(t+1)}, \boldsymbol{\eta}_2^{(t+1)})$ that maximizes

$$Q^*(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}) = \sum_{i=1}^{n^*} w_i^{*(t)} \log f(y_{i1}^*, y_{i2}^* \mid \boldsymbol{x}_i^*; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2),$$

where $w_i^{*(t)} = W(e_i^*, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*; \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})$ are the fractional weights obtained from last step. It then follows that

$$Q^*(\boldsymbol{\eta}_1^{(t+1)}, \boldsymbol{\eta}_2^{(t+1)} \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}) \geq Q^*(\boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)} \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}), \tag{4.30}$$

since $(\boldsymbol{\eta}_1^{(t+1)}, \boldsymbol{\eta}_2^{(t+1)})$ is the maximum point. It follows from the Jensen's inequality and the definition of $W(e, \boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ that (4.30) implies

$$l_{obs}^*(\boldsymbol{\eta}_1^{(t+1)}, \boldsymbol{\eta}_2^{(t+1)}) \geq l_{obs}^*(\boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}), \tag{4.31}$$

that is, the proposed procedure increases the approximated log-likelihood $l_{obs}^*(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ monotonically. See Section 4.5 for details on the derivation. By the discussion in Wu (1983), we have the following results:

Theorem 4.3. *The fractional weights* $\{\boldsymbol{w}^{*(t)}\}$ *defined in the proposed fractional imputation procedure converge to a stable set of values denoted by* $\boldsymbol{w}^*$ *as* $t \to \infty$ *for any fixed* M*. When* $M \to \infty$*, the* i*th element of* $\boldsymbol{w}^*$ *is given by*

$$w_i^* = W(e_i^*, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*; \hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2),$$

where $(\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2)$ *maximizes the log-likelihood function in (4.28).*

4.4.3 Asymptotic Properties of Fractional Imputation Estimators

In this section, we investigate the asymptotic distribution of $\hat{\boldsymbol{\theta}}^{fi}$ defined in (4.26), under the assumption that $M \to \infty$. Let

$$\begin{aligned}\boldsymbol{U}^{fi}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = r_1 r_2 \boldsymbol{U}(y_1, y_2, \boldsymbol{x}; \boldsymbol{\theta}) \\ + r_1(1-r_2) E\Big[\boldsymbol{U}(y_1, y_2, \boldsymbol{x}; \boldsymbol{\theta}) \mid y_1, \boldsymbol{x}; \boldsymbol{\eta}_2\Big] \\ + (1-r_1) r_2 E\Big[\boldsymbol{U}(y_1, y_2, \boldsymbol{x}; \boldsymbol{\theta}) \mid y_2, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2\Big] \\ + (1-r_1)(1-r_2) E\Big[\boldsymbol{U}(y_1, y_2, \boldsymbol{x}; \boldsymbol{\theta}) \mid \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2\Big]. \qquad (4.32)\end{aligned}$$

It can be checked that $\hat{\boldsymbol{\theta}}^{fi}$ is asymptotically equivalent to the solution to

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}^{fi}(\boldsymbol{r}_i, \boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\theta}, \hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2), \qquad (4.33)$$

and the maximum likelihood estimator $(\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2)$ can be obtained by solving estimating equations

$$\mathbf{0} = \sum_{i=1}^{n} \boldsymbol{S}_{obs}^{(1)}(\boldsymbol{r}_i, \boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2), \quad \mathbf{0} = \sum_{i=1}^{n} \boldsymbol{S}_{obs}^{(2)}(\boldsymbol{r}_i, \boldsymbol{y}_i, \boldsymbol{x}_i; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2). \qquad (4.34)$$

with $\boldsymbol{S}_{obs}^{(1)}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$, $\boldsymbol{S}_{obs}^{(2)}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$, $\boldsymbol{S}(y_1, \boldsymbol{x}; \boldsymbol{\eta}_1)$ and $\boldsymbol{S}(y_2, y_1, \boldsymbol{x}; \boldsymbol{\eta}_2)$ defined by the same expressions as in (4.22) and (4.16). See Section 4.5 for details. Let $\boldsymbol{S}_{obs}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = (\boldsymbol{S}_{obs}^{(1)'}, \boldsymbol{S}_{obs}^{(2)'})'$ and $\boldsymbol{S}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = (\boldsymbol{S}_1', \boldsymbol{S}_2')'$, where $\boldsymbol{S}_{obs}^{(1)}$, $\boldsymbol{S}_{obs}^{(2)}$, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are short forms of functions given above. We present the asymptotic properties of $\hat{\boldsymbol{\theta}}^{fi}$ when $M \to \infty$ as follows:

Theorem 4.4. *Let $\boldsymbol{\theta}_0$, $\boldsymbol{\eta}_{10}$ and $\boldsymbol{\eta}_{20}$ be the true values of $\boldsymbol{\theta}$, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$. Under the regularity conditions specified in Section 4.5, $\hat{\boldsymbol{\theta}}^{fi}$ given by (4.26) is a consistent estimator of $\boldsymbol{\theta}$. Furthermore,*

$$n^{1/2}(\hat{\boldsymbol{\theta}}^{fi} - \boldsymbol{\theta}_0) \sim \boldsymbol{N}\Big(\mathbf{0}, \boldsymbol{\tau} Var\big[\boldsymbol{U}^{fi}(\boldsymbol{\theta}_0, \boldsymbol{\eta}_{10}, \boldsymbol{\eta}_{20}) + \boldsymbol{\kappa} \boldsymbol{I}_{obs}^{-1} \boldsymbol{S}_{obs}(\boldsymbol{\eta}_{10}, \boldsymbol{\eta}_{20})\big] \boldsymbol{\tau}'\Big),$$

where "$\sim$" represents "is asymptotically distributed as", $\boldsymbol{\tau} = \left\{-E(\partial \boldsymbol{U}/\partial \boldsymbol{\theta}')\right\}^{-1}$,

$$\boldsymbol{I}_{obs} = \left(E\left[-\partial \boldsymbol{S}_{obs}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)/\partial \boldsymbol{\eta}_1'\right], E\left[-\partial \boldsymbol{S}_{obs}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)/\partial \boldsymbol{\eta}_2'\right]\right),$$

both evaluated at the true values of the parameters and

$$\kappa = E\Big\{\boldsymbol{U}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}_0)\left[\boldsymbol{S}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_{10}, \boldsymbol{\eta}_{20}) - \boldsymbol{S}_{obs}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}_{10}, \boldsymbol{\eta}_{20})\right]'\Big\}.$$

Variance estimation methods discussed Section 4.3.3 are all applicable to the current case and are not repeated here.

4.4.4 Simulation Studies

For the case of ordinal regression with both the response and a covariate missing, a simple simulation study is conducted to demonstrate the finite sample performance of the estimators based on the fractionally imputed data file. We consider a single ordinal response Y_2 with three categories and two covariates: a fully observed X generated from Exp(1) and an incomplete Y_1 following N(1, 1). The covariate Y_1 depends linearly on $\boldsymbol{X}$. The response Y_2 follow the second model given in (4.25) and the same model is used by the data user for subsequent analysis. Missingness is simulated following the randomized monotone missing mechanism proposed by Robins and Gill (1997). The proportions of units in the four groups $\mathcal{R}$, $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{M}$ are approximately 40%, 30%, 20% and 10%, respectively. Two sample sizes $n = 200$ and $n = 500$ are considered, each replicated 2000 times.

Table 4.5 presents results for different estimators of the marginal probability of the first category of Y_2, denoted by π_1, and the regression coefficient β_{21} of X in the *analysis model*. The same abbreviations are used for different approaches as in Section 4.3.4. For SRMI, $M = 10$ imputed data sets are created and Rubin's combining rule is applied for variance estimation. For the proposed FI, we draw $M = 10$ samples for each missing value of Y_1 and implement the jackknife resampling method to estimate the variance. The absolute relative bias (ARBv) of these two variance estimators is also listed in Table 4.5.

From the results, we observe that the CCA method is severely biased; both imputation-based methods perform well when estimating the marginal probability,

Table 4.5: Absolute Relative Bias (%) and Mean Squared Error ($\times 10^{-4}$ for π_1 and $\times 10^{-2}$ for β_{21}) of Different Estimators of π_1 and β_{21} and Absolute Relative Bias (%) of Variance Estimators

		π_1			β_{21}		
n	Methods	ARB	MSE	ARBv	ARB	MSE	ARBv
200	COMP	0.1	12.0	—	3.7	3.0	—
	CCA	60.5	896.9	—	8.8	9.5	—
	SRMI	0.3	16.9	2.2	7.5	5.9	19.9
	FI	0.2	16.3	0.6	1.9	6.7	5.7
500	COMP	0.0	4.9	—	1.3	1.1	—
	CCA	60.5	891.7	—	2.8	2.9	—
	SRMI	0.1	6.9	0.8	9.8	2.8	31.2
	FI	0.2	6.5	2.9	2.5	2.3	1.7

but for estimating the regression coefficient, the SRMI has unreasonably large bias in both the point estimator and the variance estimator. Our proposed FI method still provides satisfactory results as expected. As an additional note, the FI method is justified under the assumption that $M \to \infty$, but at least from this limited simulation study, the number of imputations required for each missing continuous variable does not have to be very large.

4.5 Regularity Conditions and Proofs

4.5.1 Regularity Conditions and Proof of Theorem 4.2

We only consider the asymptotic properties of $\hat{\pi}_{jk}^{fi}$ and the extension to $\hat{\boldsymbol{\pi}}^{fi}$ is straightforward. Let $\tilde{\boldsymbol{\theta}}_g = (\pi_{jk}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ and $\tilde{\boldsymbol{U}}_g(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \tilde{\boldsymbol{\theta}}_g) = (U_{jk}^{fi'}, \boldsymbol{S}'_{obs})'$, where U_{jk}^{fi} and $\boldsymbol{S}_{obs}$ are defined after equation (4.23). We assume that conditions S1-S8 for Theorem 3.1 hold for $\tilde{\boldsymbol{U}}_g(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \tilde{\boldsymbol{\theta}})$ and $\tilde{\boldsymbol{\theta}}_g$. Following similar arguments to those in Section 3.6.3, we can derive the asymptotic distribution of $\hat{\pi}_{jk}^{fi}$. The key step is to find the expression for $E\big[\partial U_{rj}^{fi}/\partial \boldsymbol{\eta}\big]$, where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2')'$ with true value denoted by $\boldsymbol{\eta}_0$.

We note that U_{jk}^{fi} depends on $\boldsymbol{\eta}$ through W defined in (4.4). For the second term of U_{jk}^{fi} given in (4.20), we have

$$\begin{aligned} \partial W(e,(1,0),(y_1,k),\boldsymbol{x};\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)/\partial\boldsymbol{\eta} &= (\boldsymbol{0}, \partial f(k \mid y_1, \boldsymbol{x}; \boldsymbol{\eta}_2)/\partial \boldsymbol{\eta}_2) \\ &= (\boldsymbol{0}, \boldsymbol{S}_2(k, y_1, \boldsymbol{x}; \boldsymbol{\eta}_2) f(k \mid y_1, \boldsymbol{x}; \boldsymbol{\eta}_2))\,. \end{aligned}$$

For the third term of U_{jk}^{fi} given in (4.20), we have

$$\begin{aligned} & \partial W(e,(0,1),(j,y_2),\boldsymbol{x};\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)/\partial\boldsymbol{\eta} \\ =\ & \partial f(j \mid y_2, \boldsymbol{x}; \boldsymbol{\eta})/\partial \boldsymbol{\eta} \\ =\ & f(j \mid y_2, \boldsymbol{x}; \boldsymbol{\eta})\Big\{\partial \log\big[f(j \mid y_2, \boldsymbol{x}; \boldsymbol{\eta})\big]/\partial \boldsymbol{\eta}\Big\} \\ =\ & f(j \mid y_2, \boldsymbol{x}; \boldsymbol{\eta}) \begin{pmatrix} \boldsymbol{S}_1(j, \boldsymbol{x}; \boldsymbol{\eta}_1) - E[\boldsymbol{S}_1(y_1, \boldsymbol{x}; \boldsymbol{\eta}_1) \mid y_2, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2] \\ \boldsymbol{S}_2(y_2, j, \boldsymbol{x}; \boldsymbol{\eta}_2) - E[\boldsymbol{S}_2(y_2, y_1, \boldsymbol{x}; \boldsymbol{\eta}_2) \mid y_2, \boldsymbol{x}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2] \end{pmatrix}', \end{aligned}$$

where the last equation uses the conditional probability mass function of y_1 given y_2 in (4.17). Similarly, for the fourth term of U_{rj}^{fi}, we have

$$\begin{aligned} \partial W(e,(0,0),(j,k),\boldsymbol{x};\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)/\partial\boldsymbol{\eta} &= \partial f(j,k \mid \boldsymbol{x}; \boldsymbol{\eta})/\partial \boldsymbol{\eta} \\ &= f(j,k \mid \boldsymbol{x}; \boldsymbol{\eta})(\boldsymbol{S}'_1(j, \boldsymbol{x}; \boldsymbol{\eta}_1), \boldsymbol{S}'_2(k, j, \boldsymbol{x}; \boldsymbol{\eta}_2)). \end{aligned}$$

After some tedious but straightforward algebra, it can be shown that

$$\begin{aligned} E\big[\partial U_{jk}^{fi}/\partial\boldsymbol{\eta}\big]\big|_{\boldsymbol{\eta}_0} &= E\Big\{f(j,k \mid \boldsymbol{x};\boldsymbol{\eta}_0)\big[\boldsymbol{S}((j,k),\boldsymbol{x};\boldsymbol{\eta}_0) - \boldsymbol{S}_{obs}(\boldsymbol{r},(j,k),\boldsymbol{x};\boldsymbol{\eta}_0)\big]'\Big\} \\ &= E\Big\{\boldsymbol{I}(y_1=j,y_2=k)\big[\boldsymbol{S}((j,k),\boldsymbol{x};\boldsymbol{\eta}_0) - \boldsymbol{S}_{obs}(\boldsymbol{r},(j,k),\boldsymbol{x};\boldsymbol{\eta}_0)\big]'\Big\} \\ &= \kappa_{jk}\,, \end{aligned}$$

where $\boldsymbol{S}$ and $\boldsymbol{S}_{obs}$ are defined in Theorem 4.2.

4.5.2 Proof of the Monotonicity of l_{obs}^* in the Fractional Imputation Procedure

We examine the weights defined by (4.4) for imputed observations corresponding to the i_0th observation in the original data. If $i_0 \in \mathcal{R}$, the original observation is not imputed and receives weight 1 in the imputed file. If $i_0 \in \mathcal{P}_1$, missing values of the ordinal response Y_2 are imputed, as in (4.11),

$$W(i_0,\boldsymbol{r}_i^*,\boldsymbol{y}_i^*,\boldsymbol{x}_i^*;\boldsymbol{\eta}_1,\boldsymbol{\eta}_2) = \frac{f(y_{i_01},y_{i2}^* \mid \boldsymbol{x}_{i_0};\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)}{\sum_{l\in S(i_0)} f(y_{i_01},y_{l2}^* \mid \boldsymbol{x}_{i_0};\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)},$$

for any unit $(\boldsymbol{r}_i^*,\boldsymbol{y}_i^*,\boldsymbol{x}_i^*)$ with $i \in S(i_0)$. If $i_0 \in \mathcal{P}_2$, the original observation is replicated M times with missing values of Y_1 replaced by independent draws from $h(y_1 \mid \boldsymbol{x}_{i_0};\boldsymbol{\psi}_1)$. For any unit $(\boldsymbol{r}_i^*,\boldsymbol{y}_i^*,\boldsymbol{x}_i^*)$ with $i \in S(i_0)$,

$$\begin{aligned} W(i_0,\boldsymbol{r}_i^*,\boldsymbol{y}_i^*,\boldsymbol{x}_i^*;\boldsymbol{\eta}_1,\boldsymbol{\eta}_2) &= \frac{f(y_{i1}^*,y_{i_02} \mid \boldsymbol{x}_{i_0};\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)/h(y_{i1}^* \mid \boldsymbol{x}_{i_0};\boldsymbol{\psi}_1)}{\sum_{l\in S(i_0)} f(y_{l1}^*,y_{i_02} \mid \boldsymbol{x}_{i_0};\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)/h(y_{l1}^* \mid \boldsymbol{x}_{i_0};\boldsymbol{\psi}_1)} \\ &= C_{i_0}(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)\frac{f(y_{i1}^*,y_{i_02} \mid \boldsymbol{x}_{i_0};\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)}{h(y_{i1}^* \mid \boldsymbol{x}_{i_0};\boldsymbol{\psi}_1)}, \end{aligned}$$

where $C_{i_0}(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2) = \Big\{\sum_{l\in S(i_0)} f(y_{l1}^*,y_{i_02} \mid \boldsymbol{x}_{i_0};\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)/h(y_{l1}^* \mid \boldsymbol{x}_{i_0};\boldsymbol{\psi}_1)\Big\}^{-1}$ is the normalizing constant. Similarly, if $i_0 \in \mathcal{M}$,

$$W(i_0,\boldsymbol{r}_i^*,\boldsymbol{y}_i^*,\boldsymbol{x}_i^*;\boldsymbol{\eta}_1,\boldsymbol{\eta}_2) = D_{i_0}(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)\frac{f(y_{i1}^*,y_{i2}^* \mid \boldsymbol{x}_{i_0};\boldsymbol{\eta}_1,\boldsymbol{\eta}_2)}{h(y_{i1}^* \mid \boldsymbol{x}_{i_0};\boldsymbol{\psi}_1)},$$

for $(\boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*)$ with $i \in S(i_0)$ with $D_{i_0}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \left\{ \sum_{l \in S(i_0)} f(y_{l1}^*, y_{l2}^* \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) / h(y_{l1}^* \mid \boldsymbol{x}_{i_0}; \boldsymbol{\psi}_1) \right\}^{-1}$ being the normalizing constant. So by the definition,

$$
\begin{aligned}
Q^*(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}) =& \sum_{i=1}^{n^*} w_i^{*(t)} \log f(y_{i1}^*, y_{i2}^* \mid \boldsymbol{x}_i^*; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \\
=& \sum_{i_0 \in \mathcal{R}} \log f(y_{i_0 1}, y_{i_0 2} \mid \boldsymbol{x}_{i0}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \\
&+ \sum_{i_0 \in \mathcal{P}_1} \sum_{i \in S(i_0)} w_i^{*(t)} \log f(y_{i_0 1}, y_{i2}^* \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \\
&+ \sum_{i_0 \in \mathcal{P}_2} \sum_{i \in S(i_0)} w_i^{*(t)} \log f(y_{i1}^*, y_{i_0 2} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \\
&+ \sum_{i_0 \in \mathcal{M}} \sum_{i \in S(i_0)} w_i^{*(t)} \log f(y_{i1}^*, y_{i2}^* \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2),
\end{aligned} \tag{4.35}
$$

which is written in four parts corresponding to the four groups in the original data. Similarly, $Q^*(\boldsymbol{\eta}_1^{(t+1)}, \boldsymbol{\eta}_2^{(t+1)} \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}) - Q^*(\boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)} \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})$ also consists of four terms. We take the third term as an example to illustrate the idea of the proof. The complete proof involves applying the same arguments to other terms. The third term of $Q^*(\boldsymbol{\eta}_1^{(t+1)}, \boldsymbol{\eta}_2^{(t+1)} \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}) - Q^*(\boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)} \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})$ is

$$
\begin{aligned}
&\sum_{i_0 \in \mathcal{P}_2} \sum_{i \in S(i_0)} w_i^{*(t)} \log \frac{f(y_{i1}^*, y_{i_0 2} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1^{(t+1)}, \boldsymbol{\eta}_2^{(t+1)})}{f(y_{i1}^*, y_{i_0 2} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})} \\
\leq& \sum_{i_0 \in \mathcal{P}_2} \sum_{i \in S(i_0)} \log \left\{ w_i^{*(t)} \frac{f(y_{i1}^*, y_{i_0 2} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1^{(t+1)}, \boldsymbol{\eta}_2^{(t+1)})}{f(y_{i1}^*, y_{i_0 2} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})} \right\} \\
=& \sum_{i_0 \in \mathcal{P}_2} \sum_{i \in S(i_0)} \log \left\{ C_{i_0}(\boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}) \frac{f(y_{i1}^*, y_{i_0 2} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1^{(t+1)}, \boldsymbol{\eta}_2^{(t+1)})}{h(y_{i1}^* \mid \boldsymbol{x}_{i_0}; \boldsymbol{\psi}_1)} \right\},
\end{aligned}
$$

where the first step holds by the fact that $\sum_{i \in S(i_0)} w_i^{*(t)} = 1$ and the Jensen's Inequality. It is easy to check that the term on the right hand side of the equation is the third term in $l_{obs}^*(\boldsymbol{\eta}_1^{(t+1)}, \boldsymbol{\eta}_2^{(t+1)}) - l_{obs}^*(\boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)})$. Therefore, we showed that

$$
\begin{aligned}
&Q^*(\boldsymbol{\eta}_1^{(t+1)}, \boldsymbol{\eta}_2^{(t+1)} \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}) - Q^*(\boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)} \mid \boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}) \\
\leq& l_{obs}^*(\boldsymbol{\eta}_1^{(t+1)}, \boldsymbol{\eta}_2^{(t+1)}) - l_{obs}^*(\boldsymbol{\eta}_1^{(t)}, \boldsymbol{\eta}_2^{(t)}),
\end{aligned}
$$

that is, the iterative procedure in the fractional imputation monotonically increases $l^*_{obs}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$.

4.5.3 Regularity Conditions and Proof of Theorem 4.4

The estimator $\hat{\boldsymbol{\theta}}^{fi}$ is the solution to (4.26). As in (4.35), we can re-write the right hand side of (4.26) into four terms:

$$\begin{aligned}\mathbf{0} =& n^{-1}\Bigg\{\sum_{i_0\in\mathcal{R}} \boldsymbol{U}(y_{i_0 1}, y_{i_0 2}, \boldsymbol{x}_{i_0}; \boldsymbol{\theta}) + \sum_{i_0\in\mathcal{P}_1}\sum_{i\in S(i_0)} w_i^* \boldsymbol{U}(y_{i_0 1}, y^*_{i2}, \boldsymbol{x}_{i_0}; \boldsymbol{\theta}) \\ &+ \sum_{i_0\in\mathcal{P}_2}\sum_{i\in S(i_0)} w_i^* \boldsymbol{U}(y^*_{i1}, y_{i_0 2}, \boldsymbol{x}_{i_0}; \boldsymbol{\theta}) + \sum_{i_0\in\mathcal{M}}\sum_{i\in S(i_0)} w_i^* \boldsymbol{U}(y^*_{i1}, y^*_{i2}, \boldsymbol{x}_{i_0}; \boldsymbol{\theta})\Bigg\}.\end{aligned}$$

We still take the third term as an example and for $i_0 \in \mathcal{P}_2$, we have

$$\begin{aligned}&\sum_{i\in S(i_0)} w_i^* \boldsymbol{U}(y^*_{i1}, y_{i_0 2}, \boldsymbol{x}_{i_0}; \boldsymbol{\theta}) \\ =& MC_{i_0}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) M^{-1} \sum_{i\in S(i_0)} \frac{f(y^*_{i1}, y_{i_0 2} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{h(y^*_{i1} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\psi}_1)} \boldsymbol{U}(y^*_{i1}, y_{i_0 2}, \boldsymbol{x}_{i_0}; \boldsymbol{\theta}).\end{aligned}$$

Noting that, when $M \to \infty$,

$$MC_{i_0}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \xrightarrow{p} \Big\{f(y_{i_0 2} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)\Big\}^{-1}$$

and

$$\begin{aligned}&M^{-1} \sum_{i\in S(i_0)} \frac{f(y^*_{i1}, y_{i_0 2} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2)}{h(y^*_{i1} \mid \boldsymbol{x}_{i_0}; \boldsymbol{\psi}_1)} \boldsymbol{U}(y^*_{i1}, y_{i_0 2}, \boldsymbol{x}_{i_0}; \boldsymbol{\theta}) \\ \xrightarrow{p}& \int_{\mathcal{D}_1} f(y_1, y_{i_0 2} \mid \boldsymbol{x}_{i0}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \boldsymbol{U}(y_1, y_{i_0 2}, \boldsymbol{x}_{i_0}; \boldsymbol{\theta}) dy_1,\end{aligned}$$

by putting two pieces together, we have

$$\sum_{i\in S(i_0)} w_i^* \boldsymbol{U}(y^*_{i1}, y_{i_0 2}, \boldsymbol{x}_{i_0}; \boldsymbol{\theta}) \xrightarrow{p} E\big[\boldsymbol{U}(y_1, y_{i_0 2}, \boldsymbol{x}_{i_0}; \boldsymbol{\theta}) \mid \boldsymbol{y}_{i_0 2}, \boldsymbol{x}_{i_0}; \boldsymbol{\eta}_1, \boldsymbol{\eta}_2\big].$$

Similar results can be shown for other terms. Also when $M \to \infty$, by Theorem 4.3,

$$w_i^* = W(e_i^*, \boldsymbol{r}_i^*, \boldsymbol{y}_i^*, \boldsymbol{x}_i^*; \hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2),$$

and it thus follows that $\hat{\boldsymbol{\theta}}^{fi}$ is equivalent to the solution to (4.33).

Let $\boldsymbol{\eta}_e = (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2')$, $\tilde{\boldsymbol{\theta}}_e = (\boldsymbol{\theta}', \boldsymbol{\eta}_e')'$ and $\tilde{\boldsymbol{U}}_e(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \tilde{\boldsymbol{\theta}}_e) = (\boldsymbol{U}^{fi'}(\tilde{\boldsymbol{\theta}}_e), \boldsymbol{S}_{obs}'(\boldsymbol{\eta}_e))'$, where $\boldsymbol{U}^{fi'}(\tilde{\boldsymbol{\theta}}_e)$ and $\boldsymbol{S}_{obs}'(\boldsymbol{\eta}_e)$ are defined in (4.32) and after (4.34), respectively. We assume that conditions S1-S8 in Theorem 3.1 hold for $\tilde{\boldsymbol{U}}_e(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \tilde{\boldsymbol{\theta}})$ and $\tilde{\boldsymbol{\theta}}_e$. Theorem 4.4 follows directly from similar arguments to those in Section 3.6.3. The only step missing is to find an expression for $E\big[\partial \boldsymbol{U}^{fi}/\partial \boldsymbol{\eta}'\big]$. This can be done by representing $\boldsymbol{U}^{fi}$ as $E\big[\boldsymbol{U}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}) \mid \boldsymbol{r}, \boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\eta}\big]$ and we have

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\eta}'} \boldsymbol{U}^{fi}(\boldsymbol{\theta}, \boldsymbol{\eta}) \\
=&\frac{\partial}{\partial \boldsymbol{\eta}'} \int \boldsymbol{U}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}) f(\boldsymbol{y}_{mis} \mid \boldsymbol{r}, \boldsymbol{y}_{obs}, \boldsymbol{x}; \boldsymbol{\eta}) d\mu(\boldsymbol{y}_{mis}) \\
=&\int \boldsymbol{U}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\eta}'} f(\boldsymbol{y}_{mis} \mid \boldsymbol{r}, \boldsymbol{y}_{obs}, \boldsymbol{x}; \boldsymbol{\eta}) d\mu(\boldsymbol{y}_{mis}) \\
=&\int \boldsymbol{U}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}) \big\{ \frac{\partial}{\partial \boldsymbol{\eta}'} \log f(\boldsymbol{y}_{mis} \mid \boldsymbol{r}, \boldsymbol{y}_{obs}, \boldsymbol{x}; \boldsymbol{\eta}) \big\} f(\boldsymbol{y}_{mis} \mid \boldsymbol{r}, \boldsymbol{y}_{obs}, \boldsymbol{x}; \boldsymbol{\eta}) d\mu(\boldsymbol{y}_{mis}) \\
=&\int \boldsymbol{U}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}) \big[\boldsymbol{S}(\boldsymbol{\eta}) - \boldsymbol{S}_{obs}(\boldsymbol{\eta}) \big] f(\boldsymbol{y}_{mis} \mid \boldsymbol{r}, \boldsymbol{y}_{obs}, \boldsymbol{x}; \boldsymbol{\eta}) d\mu(\boldsymbol{y}_{mis}),
\end{aligned}$$

where $\boldsymbol{y}_{mis}$ and $\boldsymbol{y}_{obs}$ are the missing and observed components of $\boldsymbol{y}$. Therefore,

$$\begin{aligned}
E\Big[\frac{\partial}{\partial \boldsymbol{\eta}'} \boldsymbol{U}^{fi}(\boldsymbol{\theta}, \boldsymbol{\eta}) \Big] &= \int \frac{\partial}{\partial \boldsymbol{\eta}'} \boldsymbol{U}^{fi}(\boldsymbol{\theta}, \boldsymbol{\eta}) f(\boldsymbol{r}, \boldsymbol{y}_{obs}, \boldsymbol{x}) d\boldsymbol{r} d\boldsymbol{y} d\boldsymbol{x} \\
&= E\left\{ \boldsymbol{U}(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\theta}) \big[\boldsymbol{S}(\boldsymbol{\eta}) - \boldsymbol{S}_{obs}(\boldsymbol{\eta}) \big] \right\}.
\end{aligned}$$

Chapter 5

Doubly Robust Fractional Imputation

As with most imputation-based approaches, the validity of subsequent analyses based on the fractionally imputed data set rests on the correct specification of models for the data generating process. In most cases, these models are reliable, because the data creators are equipped with an in-depth knowledge of statistical modelling and have access to extra information which contributes to revealing the structure of the responses. Nevertheless, it is always desirable if we could improve the robustness of estimation, at least for some important inferential problems.

Scharfstein et al. (1999) first pointed out the "doubly robust" property of the AIPW estimator, that is, it is consistent if either the model for the data generating process (DGP) or the model for the missing data process (MDP) is correctly specified. Bang and Robins (2005) extended the method for constructing doubly robust estimators to the longitudinal case through an alternative "regression representation" of the AIPW estimator. The key idea underlying double robustness is to incorporate both the DGP model and the MDP model simultaneously. In this chapter, we integrate the MDP model into the proposed fractional imputation procedure by imposing additional "calibration constraints" when estimating parameters in the DGP model and show that the estimators of marginal population quantities are doubly robust against model misspecification.

5.1 Univariate Ordinal Responses with Missing observations

5.1.1 Basic Settings and Motivation

We begin with a simple case where only one ordinal response is subject to missingness with notation and assumptions given in Chapter 3. The available data set is denoted by $\mathcal{O} = \big\{(r_i, y_i, \boldsymbol{x}_i),\, i = 1, \ldots, n\big\}$. We mainly focus on the subsequent analysis of estimating the unconditional cumulative probabilities $\boldsymbol{p}$ defined in Section 3.3.1. But it is important to note that the following discussion is also true for estimating a general marginal parameter $\boldsymbol{\theta} = E[\boldsymbol{g}(Y)]$, where $\boldsymbol{g}(\cdot)$ is a scalar- or vector-valued function, because $\boldsymbol{\theta} = \sum_{j=1}^{J} \boldsymbol{g}(j) P(Y = j) = \sum_{j=1}^{J} \boldsymbol{g}(j)(p_j - p_{j-1})$.

First, we postulate a *cumulative link model* of the form (2.16) for the response Y with $\boldsymbol{X}$ as covariates

$$G^{-1}\big[\gamma_j(\boldsymbol{X})\big] = \alpha_j - \boldsymbol{\beta}'\boldsymbol{X}, \qquad j = 1, \ldots, J-1, \tag{5.1}$$

where $\gamma_j(\boldsymbol{X}) = P(Y \le j \mid \boldsymbol{X})$ and G^{-1} is a link function. Let $\boldsymbol{\eta} = (\alpha_1, \ldots, \alpha_{J-1}, \boldsymbol{\beta}')'$ be the parameters in this DGP model. Note that our discussion can be readily extended to cover other ordinal regression models introduced in Section 2.4. At the same time, we characterize the MDP with a parametric model $\pi(\boldsymbol{X}; \boldsymbol{\phi})$ for the response probability

$$\pi(\boldsymbol{X}) = P(R = 1 \mid \boldsymbol{X})\,.$$

A common choice for $\pi(\boldsymbol{X}; \boldsymbol{\phi})$ is the logistic regression model given in (3.2). The parameters $\boldsymbol{\phi}$ can be estimated by maximum likelihood method based on the data $\big\{(r_i, \boldsymbol{x}_i),\, i = 1. \ldots, n\big\}$ and let $\hat{\boldsymbol{\phi}}$ be the resulting estimator.

Suppose that the ith observation has the response missing. A general fractional imputation procedure replicates this unit M times and fill in M imputed values $\tilde{y}_{ij}$ for $j = 1, \ldots, M$, each assigned a fractional weight w_{ij}. For the procedure we proposed in Chapter 3, $M = J$ and $\tilde{y}_{ij} = j$ with

$$w_{ij} = \gamma_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc}) - \gamma_{j-1}(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}^{cc}) \tag{5.2}$$

and $\hat{\boldsymbol{\eta}}^{cc}$ defined in (3.1). Based on the imputed data set, the fractional imputation

estimator $\hat{\boldsymbol{p}}^{fi}$ is given by

$$\hat{\boldsymbol{p}}^{fi} = n^{-1} \sum_{i=1}^{n} \left[r_i \boldsymbol{z}_i + (1 - r_i) \sum_{j=1}^{J} w_{ij} \boldsymbol{c}_j \right] , \tag{5.3}$$

where $\boldsymbol{z}_i$ and $\boldsymbol{c}_j$ are the cumulative indicator vectors for y_i and level j. When model (5.1) is correct, we have shown in Section 3.3.1 that $\hat{\boldsymbol{p}}^{fi}$ is a consistent and efficient estimator. However, when model (5.1) fails, the second term in (5.3) cannot recover the true distributional structure of the missing responses and thus introduces bias. To quantify this bias, we compare $\hat{\boldsymbol{p}}^{fi}$ with the unobservable sample mean $\bar{\boldsymbol{p}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{z}_i$, called the "strong robust estimator" by Kang and Schafer (2007), since it is free of model assumptions and always consistent:

$$\begin{aligned} \bar{\boldsymbol{p}} - \hat{\boldsymbol{p}}^{fi} &= n^{-1} \sum_{i=1}^{n} (1 - r_i) \big[\boldsymbol{z}_i - \sum_{j=1}^{J} w_{ij} \boldsymbol{c}_j \big] \\ &= n^{-1} \sum_{i=1}^{n} \big[1 - r_i \pi^{-1}(\boldsymbol{x}_i) \big] \big[\boldsymbol{z}_i - \sum_{j=1}^{J} w_{ij} \boldsymbol{c}_j \big] \\ &\quad + n^{-1} \sum_{i=1}^{n} r_i \big[\pi^{-1}(\boldsymbol{x}_i) - 1 \big] \big[\boldsymbol{z}_i - \sum_{j=1}^{J} w_{ij} \boldsymbol{c}_j \big], \end{aligned} \tag{5.4}$$

where $\pi(\boldsymbol{x}_i)$ is the response probability of the ith observation. Note that even when the DGP model does not hold, the fractional weights w_{ij} usually still converge to some values w_{ij}^* which only depends on $\boldsymbol{x}_i$, therefore, under the MAR assumption, the first term in (5.4) converges to 0 considering the fact that $E\big[1 - R\pi^{-1}(\boldsymbol{X}) \mid \boldsymbol{X}\big] = 0$. The second term involves only the observed values and thus can be actually calculated from the data assuming the response probabilities are known. If we use the same set of imputed values, but choose a proper set of fractional weights w_{ij} such that the second term in (5.4) vanishes, then the resulting estimator will remain consistent even when (5.1) is incorrect. In practice, the response probabilities are usually unknown and need to be estimated through a correctly specified model $\pi(\boldsymbol{X}; \boldsymbol{\phi})$.

A further look into the second term in (5.4) reveals an interesting interpretation that this term essentially estimates the potential bias introduced by imputing missing values incorrectly. To see this, suppose that all responses, whether observed or missing, are replaced by $\sum_{j=1}^{J} w_{ij} \boldsymbol{c}_j$. The actual bias of the observed group can be

directly calculated as

$$\Delta_R = \sum_{i=1}^{n} r_i \big[\boldsymbol{z}_i - \sum_{j=1}^{J} w_{ij} \boldsymbol{c}_j \big].$$

Based on the idea of inverse probability weighting, we can estimate the total bias with

$$\Delta_T = \sum_{i=1}^{n} r_i \pi^{-1}(\boldsymbol{x}_i) \big[\boldsymbol{z}_i - \sum_{j=1}^{J} w_{ij} \boldsymbol{c}_j \big].$$

The bias of the missing group is given by

$$\Delta_M = \Delta_T - \Delta_R = \sum_{i=1}^{n} r_i \big[\pi^{-1}(\boldsymbol{x}_i) - 1 \big] \big[\boldsymbol{z}_i - \sum_{j=1}^{J} w_{ij} \boldsymbol{c}_j \big].$$

5.1.2 Doubly Robust Fractional Imputation

Motivated by the above observation, we propose three different choices of fractional weights, with which the imputed data set produces valid results for various subsequent analyses as in Chapter 3 when the DGP model holds, and still provides consistent estimators of the marginal cumulative probabilities when the DGP model fails but a correct model for the MDP is available.

The first approach shares the same idea as in Bang and Robins (2005) and considers an extended version of (5.1) with the estimated inverse response probabilities $\pi^{-1}(\boldsymbol{X}; \hat{\boldsymbol{\phi}})$ as a covariate:

$$G^{-1}\big[\tilde{\gamma}_j(\boldsymbol{X})\big] = \alpha_j - \boldsymbol{\beta}'\boldsymbol{X} - \nu_j \pi^{-1}(\boldsymbol{X}; \hat{\boldsymbol{\phi}}), \qquad j = 1, \ldots, J-1. \tag{5.5}$$

Let $\boldsymbol{\eta}_{(a)} = (\alpha_1, \ldots, \alpha_{J-1}, \boldsymbol{\beta}', \nu_1, \ldots, \nu_{J-1})$ be the parameters in the extended model. We obtain an estimator $\hat{\boldsymbol{\eta}}_{(a)}$ by solving the following estimating equations:

$$\mathbf{0} = \sum_{i=1}^{n} r_i \left\{ \frac{\partial}{\partial \boldsymbol{\eta}_{(a)}} G^{-1}\big[\tilde{\boldsymbol{\gamma}}(\boldsymbol{x}_i; \boldsymbol{\eta}_{(a)}, \hat{\boldsymbol{\phi}})\big] \right\} \big[\boldsymbol{z}_i - \tilde{\boldsymbol{\gamma}}(\boldsymbol{x}_i; \boldsymbol{\eta}_{(a)}, \hat{\boldsymbol{\phi}}) \big], \tag{5.6}$$

where $\tilde{\boldsymbol{\gamma}}(\boldsymbol{x}_i; \boldsymbol{\eta}_{(a)}, \hat{\boldsymbol{\phi}}) = (\tilde{\gamma}_1(\boldsymbol{x}_i; \boldsymbol{\eta}_{(a)}, \hat{\boldsymbol{\phi}}), \ldots, \tilde{\gamma}_{J-1}(\boldsymbol{x}_i; \boldsymbol{\eta}_{(a)}, \hat{\boldsymbol{\phi}}))'$. The new fractional weights $w_{ij}^{(a)}$ are given by

$$w_{ij}^{(a)} = \tilde{w}_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_{(a)}, \hat{\boldsymbol{\phi}}) = \tilde{\gamma}_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_{(a)}, \hat{\boldsymbol{\phi}}) - \tilde{\gamma}_{j-1}(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_{(a)}, \hat{\boldsymbol{\phi}}). \tag{5.7}$$

As Bang and Robins (2005) pointed out when the model belongs to the *generalized linear models* family with the canonical link function, then $\hat{\boldsymbol{\eta}}_{(a)}$ is essentially the maximum likelihood estimator. But in our case, the assumed *cumulative link models* do not have the canonical link and thus the above estimating equations (5.6) have a different form from the score equations in (2.18). This is also true for the *continuation-ratio link models*.

A practical issue arises for the first approach when model (5.1) is imposed. The new covariate with varying coefficients invalidates the common slope assumption of the original model and breaks the order structure of the cumulative probabilities. One challenge faced by the extended model is the possible overlap of curves for different cumulative probabilities, which leads to negative estimated category probabilities, in our case, negative fractional weights. An ad hoc workaround for this issue is to set all negative weights to 0 and re-allocate the fractional weights within the same cluster proportional to their original weights. Specifically, suppose that the ith observation has a missing response and is imputed by the cluster $\big\{(1, \boldsymbol{x}_i), \ldots, (J, \boldsymbol{x}_i)\big\}$ with fractional weights $(w_{i1}^{(a)}, \ldots, w_{iJ}^{(a)})$ given by (5.7). If $w_{ij_0}^{(a)} < 0$, the fractional weights are adjusted such that

$$\breve{w}_{ij_0}^{(a)} = 0, \qquad \breve{w}_{ij}^{(a)} = \frac{w_{ij}^{(a)}}{\sum_{j \neq j_0} w_{ij}^{(a)}} \quad \text{for } j \neq j_0. \tag{5.8}$$

The adjustment may have certain unknown impact on the resulting estimator, but our simulation studies seem to suggest that the occurrence rate of negative weights is low and the impact is ignorable. Note that this is not a problem if *continuation-ratio link models* are used, because they always produce legitimate probabilities whether the common slope assumption holds or not.

The second approach still uses the original model (5.1) but estimates the parameters by solving the following weighted estimating equations:

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} r_i \big[\pi^{-1}(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}}) - 1\big] \left\{ \frac{\partial}{\partial \boldsymbol{\eta}} G^{-1}\big[\boldsymbol{\gamma}(\boldsymbol{x}_i; \boldsymbol{\eta})\big] \right\} \big[\boldsymbol{z}_i - \boldsymbol{\gamma}(\boldsymbol{x}_i; \boldsymbol{\eta})\big], \tag{5.9}$$

where $\boldsymbol{\gamma}(\boldsymbol{x}_i; \boldsymbol{\eta}) = (\gamma_1(\boldsymbol{x}_i; \boldsymbol{\eta}), \ldots, \gamma_{J-1}(\boldsymbol{x}_i; \boldsymbol{\eta}))'$. Kang and Schafer (2007) mentioned a similar idea under a simpler case where a continuous response follows a linear regression model. Let $\hat{\boldsymbol{\eta}}_{(b)}$ be the solution to (5.9). Note that $\hat{\boldsymbol{\eta}}_{(b)}$ can be viewed as a

weighted least square estimator of $\boldsymbol{\eta}$ based on the observed units with weights given by

$$\left[\pi^{-1}(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}})-1\right]\begin{pmatrix}\mathring{G}^{-}\left[\gamma_1(\boldsymbol{x}_i;\boldsymbol{\eta})\right] & & \\ & \ddots & \\ & & \mathring{G}^{-}\left[\gamma_{J-1}(\boldsymbol{x}_i;\boldsymbol{\eta})\right]\end{pmatrix},$$

where $\mathring{G}^{-}(x)$ is the derivative of $G^{-1}(x)$. It then follows the new fractional weigths:

$$w_{ij}^{(b)}=w_j(\boldsymbol{x}_i;\hat{\boldsymbol{\eta}}_{(b)})=\gamma_j(\boldsymbol{x}_i;\hat{\boldsymbol{\eta}}_{(b)})-\gamma_{j-1}(\boldsymbol{x}_i;\hat{\boldsymbol{\eta}}_{(b)})\,. \tag{5.10}$$

The third approach also rests on the original model. However, instead of re-weighting the score equations, we estimate $\boldsymbol{\eta}$ by introducing auxiliary $(J-1)$-dimensional parameters $\boldsymbol{\lambda}$ in the estimating equation. Specifically, we solve

$$\begin{aligned}\mathbf{0}&=n^{-1}\sum_{i=1}^{n}r_i\boldsymbol{S}(\boldsymbol{z}_i,\boldsymbol{x}_i;\boldsymbol{\eta})-\boldsymbol{\lambda}'n^{-1}\sum_{i=1}^{n}r_i\left[\pi^{-1}(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}})-1\right]\frac{\partial}{\partial\boldsymbol{\eta}}\boldsymbol{\gamma}(\boldsymbol{x}_i;\boldsymbol{\eta})\,,\\ \mathbf{0}&=n^{-1}\sum_{i=1}^{n}r_i\left[\pi^{-1}(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}})-1\right]\left[\boldsymbol{z}_i-\boldsymbol{\gamma}(\boldsymbol{x}_i;\boldsymbol{\eta})\right],\end{aligned} \tag{5.11}$$

where $\boldsymbol{S}(\boldsymbol{z},\boldsymbol{x};\boldsymbol{\eta})$ is the score function defined in (2.19). Let the solution be denoted by $(\hat{\boldsymbol{\eta}}'_{(c)},\hat{\boldsymbol{\lambda}}')'$. Note that $\hat{\boldsymbol{\eta}}_{(c)}$ can be alternatively defined through the following constrained optimization problem:

$$\text{argmax}_{\boldsymbol{\eta}\in H_n}\ n^{-1}\sum_{i=1}^{n}l_{obs}(r_i,y_i,\boldsymbol{x}_i;\boldsymbol{\eta})\,, \tag{5.12}$$

where

$$H_n=\left\{\boldsymbol{\eta}:n^{-1}\sum_{i=1}^{n}r_i\left[\pi^{-1}(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}})-1\right][\boldsymbol{z}_i-\boldsymbol{\gamma}(\boldsymbol{x}_i;\boldsymbol{\eta})]=\mathbf{0}\right\}$$

and $l_{obs}(r,y,\boldsymbol{x};\boldsymbol{\eta})=\sum_{j=1}^{J}r_i(z_j-z_{j-1})\log\left[\gamma_j(\boldsymbol{x};\boldsymbol{\eta})-\gamma_{j-1}(\boldsymbol{x};\boldsymbol{\eta})\right]$ is the observed log-likelihood function. We apply the Lagrange multiplier method and define the Lagrange function as

$$\mathcal{L}(\boldsymbol{\eta},\boldsymbol{\lambda})=n^{-1}\sum_{i=1}^{n}l_{obs}(r_i,y_i,\boldsymbol{x}_i;\boldsymbol{\eta})+\boldsymbol{\lambda}'n^{-1}\sum_{i=1}^{n}r_i\left[\pi^{-1}(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}})-1\right][\boldsymbol{z}_i-\boldsymbol{\gamma}(\boldsymbol{x}_i;\boldsymbol{\eta})].$$

Equations (5.11) then follow by setting the derivatives of $\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ to zero. Therefore, $\hat{\boldsymbol{\eta}}_{(c)}$ can be treated as the maximum observed likelihood estimator calibrated with the response probabilities to correct the potential bias of estimators of marginal cumulative probabilities. The optimal points $(\hat{\boldsymbol{\eta}}'_{(c)}, \hat{\boldsymbol{\lambda}}')'$ can be found by existing constrained optimization routines such sequential quadratic programming (Boggs and Tolle 1995). The fractional weights are constructed in the same way as the second approach with $\boldsymbol{\eta}$ estimated by $\hat{\boldsymbol{\eta}}_{(c)}$:

$$w_{ij}^{(c)} = w_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_{(c)}) = \gamma_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_{(c)}) - \gamma_{j-1}(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_{(c)}) \,. \tag{5.13}$$

The following theorem formally presents the "doubly robust" property of the estimators of marginal cumulative probabilities derived from the fractionally imputed data sets which are created through the proposed approaches.

Theorem 5.1. *Suppose that a single complete data set is created by replicating each missing observation J times and filling in J imputed values $\tilde{y}_{ij} = j$, for $j = 1, \ldots, J$. With the fractional weights $w_{ij}^{(a)}$, $w_{ij}^{(b)}$ and $w_{ij}^{(c)}$ defined in (5.7), (5.10) and (5.13), the estimators of marginal cumulative probabilities given in (5.3) are consistent under the regularity conditions in Section 5.5, if either the DGP model (5.1) or the MDP model $\pi(\boldsymbol{X}; \boldsymbol{\phi})$ is correctly specified.*

Detailed proof of Theorem 5.1 is given in Section 5.5. Here we provide a sketch of key ideas used in the derivation. First consider the case when model (5.1) is correct. For the first approach, in the extended model with $\pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})$ replaced by its probability limit, the true values of ν_j are zero and hence the added covariate will have little impact on the fractional weights. For the second approach, the estimating function in (5.9) with $\pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})$ replaced by its probability limit is unbiased evaluated at $\boldsymbol{\eta}_0$. For the last approach, the constraints are approximately met at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$, so based on the same idea as in the Lagrange multiplier test (Breusch and Pagan 1980), the estimator under the constraints is close to the unconstrained estimator. When the model for $\pi(\boldsymbol{X}; \boldsymbol{\phi})$ is correct, it can be easily checked that the second term in (5.4) with response probabilities estimated by $\pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})$ equals zero for all the three approaches.

5.2 Causal Effects of a Point Treatment with Ordinal Outcomes

In this section, we apply the fractional imputation methods proposed in Section 5.1 to observational studies concerning the causal effects of a dichotomous treatment measured by an ordinal outcome, by using a "complete" data set where both potential outcomes of each individual are either observed or imputed.

5.2.1 Basic Settings

Suppose that the data set from an observational study is given by $\{(r_i, y_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\}$, which is an *i.i.d.* sample of variables $(R, Y, \boldsymbol{X})$, where R denotes the level of a dichotomous treatment (with level 0 and 1) received by the subject, Y is an outcome on a J-level ordinal scale measured at the end of the study and $\boldsymbol{X}$ consists of a set of confounding variables. For a random subject, we envisage that there exist two potential outcomes $Y^{(0)}$ and $Y^{(1)}$, with $Y^{(r)}$ denoting the outcome at treatment level r for $r = 0, 1$. For subject i, the actually observed outcome y_i is a realization of the potential outcome $Y^{(r_i)}$ corresponding to the treatment level received by this subject. This is known as the consistency theorem in the causal inference literature (Pearl 2010). Without loss of generality, we assume that the first n_1 subjects receive treatment at level 1, that is, $r_i = 1$ for $i = 1, \ldots, n_1$ and the rest receive level 0. The fundamental problem of causal inference is to construct and estimate a quantity, known as the average treatment effect (ATE), which measures the difference between $Y^{(1)}$ and $Y^{(0)}$.

Unlike in randomized experiments, the treatment in observational studies is not assigned randomly among subjects, but often associated with some pretreatment variables. For example, a doctor may suggest that a patient have a surgery if he/she is young, and get medical treatment if he/she is in old age. In most cases, the age also has an impact on the potential outcomes, therefore the potential outcomes $Y^{(r)}$ are usually associated with the actual treatment R. We assume that the association can be fully explained by measured variables $\boldsymbol{X}$ and that there are no unmeasured confounders. This implies that

$$Y^{(r)} \perp R \mid \boldsymbol{X} \quad \text{for } r = 0, 1\,.$$

In other words, the potential outcomes are conditionally independent of the actual treatment given the confounding variables $\boldsymbol{X}$. For example, suppose there are two patients with same values for $\boldsymbol{X}$, but one receives treatment at level 0, while the other receives level 1. We assume that the patient receiving level 0 would have the same outcome as the other, if he/she had received level 1. It is important to note that $Y^{(r)} \perp R \mid \boldsymbol{X}$ is different from $Y \perp R \mid \boldsymbol{X}$. If the treatment has a causal effect on the outcome, then the observed outcome Y is associated with the actual treatment R conditional on $\boldsymbol{X}$. The confounders $\boldsymbol{X}$ are usually chosen based on background knowledge of the investigative team (Robins 2001). However, in the absence of randomization, there is no guarantee that all confounders are included in $\boldsymbol{X}$ and uncontrolled confounding gives rise to biased effect estimates. In that case, special techniques are necessary to adjust for these unmeasured confounders, see, for example, Stürmer et al. (2005), Johnston et al. (2008), VanderWeele and Arah (2011).

Under the potential outcome framework, the causal inference problem can be viewed from a missing data perspective (Westreich et al. 2015). For any observation with $r_i = 1$, the observed outcome is $y_i = y_i^{(1)}$ and the counterpart $y_i^{(0)}$ is unobservable, thus missing. The same thing can be said of units with $r_i = 0$. Therefore, the causal inference involves comparing two responses both subject to missingness and never observed simultaneously. Table 5.1 illustrates a typical data set from a point-treatment study viewed as a data set with missing observations.

Table 5.1: An Observational Data Set in a Point-treatment Study from a Missing Data Perspective

i	$Y^{(1)}$	$Y^{(0)}$	R	X_1	X_2	X_3
1	y_1	$*$	1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_1	y_{n_1}	$*$	1	$x_{n_1,1}$	$x_{n_1,2}$	$x_{n_1,3}$
n_1+1	$*$	y_{n_1+1}	0	$x_{n_1+1,1}$	$x_{n_1+1,2}$	$x_{n_1+1,3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$*$	y_n	0	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$

5.2.2 Doubly Robust Causal Effects Estimation

A well-defined ATE is of crucial importance in measuring the causal effects, but the construction is not straightforward for ordinal outcomes. The risk difference defined by $E[Y^{(1)}] - E[Y^{(0)}]$ for continuous responses is not suitable in this case, because the values of ordinal variables only reflect comparative orders rather than quantitative scales. The risk/odds ratio used for binary responses is not directly applicable either, because an ordinal response usually has more than two categories. We consider two commonly adopted approaches for ordinal outcomes. The first approach assigns a score s_j to each level j from 1 to J and thus transforms the ordinal responses to scalar variables $U^{(r)} = \sum_{j=1}^{J} s_j \boldsymbol{I}(Y^{(r)} = j)$ for $r = 0, 1$. The ATE is then defined as the risk difference of the two scores: $E[U^{(1)}] - E[U^{(0)}]$. The second approach chooses a reference level j_0 based on which the ordinal variable collapses into a dichotomous variable $V^{(r)} = \boldsymbol{I}(Y^{(r)} \geq j_0)$ for $r = 0, 1$. We can then define the risk ratio (RR) or odds ratio (OR) by

$$RR = \frac{P(V^{(1)} = 1)}{P(V^{(0)} = 1)}, \qquad OR = \frac{P(V^{(1)} = 1)P(V^{(0)} = 0)}{P(V^{(1)} = 0)P(V^{(0)} = 1)}. \tag{5.14}$$

Both methods involve subjective assessment from the investigators through the choices of the scores s_j and the reference level j_0. Even for the same data set, different researchers may have different choices depending on their specific needs. One attractive property of fractional imputation is that the procedure is independent of the choices of ATE. It simply creates a "complete" data set where both potential outcomes are available. With this data set, different data users then choose appropriate scores and reference levels according to their needs and obtain estimates of ATE through straightforward calculation. Furthermore, the ATEs defined above are functions of the marginal probabilities of the potential outcomes. By assigning fractional weights proposed in Section 5.1, the "complete" data set provides doubly robust estimators of ATE.

We first impose two models on $Y^{(1)}$ and $Y^{(0)}$ against $\boldsymbol{X}$ separately:

$$\begin{aligned} G_1^{-1}\left[\gamma_j^{(1)}(\boldsymbol{X})\right] &= \alpha_j^{(1)} + \boldsymbol{\beta}^{(1)\prime}\boldsymbol{X}, \\ G_0^{-1}\left[\gamma_j^{(0)}(\boldsymbol{X})\right] &= \alpha_j^{(0)} + \boldsymbol{\beta}^{(0)\prime}\boldsymbol{X}, \quad j = 1, \ldots, J-1, \end{aligned} \tag{5.15}$$

where $\gamma_j^{(r)} = P(Y^{(r)} \leq j \mid \boldsymbol{X})$ for $r = 0, 1$. Let $\boldsymbol{\eta}^{(1)} = (\alpha_1^{(1)}, \ldots, \alpha_{J-1}^{(1)}, \boldsymbol{\beta}^{(1)'})'$ and $\boldsymbol{\eta}^{(0)} = (\alpha_1^{(0)}, \ldots, \alpha_{J-1}^{(0)}, \boldsymbol{\beta}^{(0)'})'$ be the parameters in these models. At the same time, we model the treatment assignment process by positing a model $\pi(\boldsymbol{X}; \boldsymbol{\phi})$ for $P(R = 1 \mid \boldsymbol{X})$, for example, in a logistic regression form:

$$\log \left[\frac{\pi(\boldsymbol{X}; \boldsymbol{\phi})}{1 - \pi(\boldsymbol{X}; \boldsymbol{\phi})} \right] = a(\boldsymbol{X}; \boldsymbol{\phi}), \tag{5.16}$$

where $a(\boldsymbol{X}; \boldsymbol{\phi})$ is specified up to some unknown parameters $\boldsymbol{\phi}$. If model (5.16) holds, a consistent estimator $\hat{\boldsymbol{\phi}}$ of $\boldsymbol{\phi}$ is given by maximizing the likelihood function based on $\{(t_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$. This is the basis for the inverse-probabilities-of-treatment weighting method proposed by Robins et al. (2000).

To create the "complete" data set, we begin with $Y^{(1)}$ which is treated as a response variable with missing observations and R is the response indicator. Each incomplete observation $i = n_1, \ldots, n$ is replicated J times with missing values filled in by $\tilde{y}_{ij}^{(1)} = 1, \ldots, J$. The corresponding fractional weights can be taken as $\breve{w}_{ij}^{(a)}$, $w_{ij}^{(b)}$ and $w_{ij}^{(c)}$ defined in (5.8), (5.10) and (5.13). We repeat the procedure for $Y^{(0)}$. Missing values of $Y^{(0)}$ in observations $i = 1, \ldots, n_1$ are imputed in the same way as $Y^{(1)}$ and the fractional weights $\breve{w}_{ij}^{(a)}$, $w_{ij}^{(b)}$ and $w_{ij}^{(c)}$ are calculated by substituting $1 - r_i$ and $1 - \pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})$ for r_i and $\pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})$. Finally, we obtain a single data set of size Jn with $Y^{(1)}$ and $Y^{(0)}$ fully observed for all subjects, denoted by $\{(r_i^*, y_i^{*(1)}, y_i^{*(0)}, \boldsymbol{x}_i^*), i = 1, \ldots, Jn\}$ with fractional weights $\{\breve{w}_i^{*(a)}, i = 1, \ldots, Jn\}$, $\{w_i^{*(b)}, i = 1, \ldots, Jn\}$ and $\{w_i^{*(c)}, i = 1, \ldots, Jn\}$ corresponding to the three approaches proposed in Section 5.1.

The subsequent analysis is straightforward. For example, we choose j_0 as the reference level and are interested in estimating the risk ratio $\boldsymbol{\theta} = RR$ defined in (5.14). An estimator based on the "complete" data set is given by

$$\hat{\boldsymbol{\theta}} = \sum_{i=1}^{Jn} w_i^* \boldsymbol{I}(y_i^{*(1)} \geq j_o) / \sum_{i=1}^{Jn} w_i^* \boldsymbol{I}(y_i^{*(0)} \geq j_o), \tag{5.17}$$

where w_i^* can take values $\breve{w}_i^{*(a)}$, $w_i^{*(b)}$ and $w_i^{*(c)}$. By Theorem 5.1, it is apparent that $\hat{\boldsymbol{\theta}}$ is doubly robust in the sense that it is consistent if either model (5.15) or model (5.16) is correctly specified, but not necessarily both.

5.3 Longitudinal Ordinal Responses with Monotone Missingness

5.3.1 Basic Settings

In this section, we extend the discussion in Section 5.1 to multivariate cases and consider the data set $\mathcal{O} = \{(\boldsymbol{r}_i, \boldsymbol{y}_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\}$ consisting of *i.i.d.* samples from $(\boldsymbol{R}, \boldsymbol{Y}, \boldsymbol{X})$, where $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$ is a T-dimensional response vector from a longitudinal study with each element Y_t being an ordinal variable with J_t categories and subject to missingness. We assume that the data follow a monotone missing pattern (see Section 2.2.1). The MAR assumption in this case implies that

$$P(R_t = 1 \mid \bar{\boldsymbol{R}}_{t-1}, \boldsymbol{Y}, \boldsymbol{X}) = P(R_t = 1 \mid \bar{\boldsymbol{R}}_{t-1}, \bar{\boldsymbol{Y}}_{t-1}, \boldsymbol{X}), \tag{5.18}$$

for $t = 1, \ldots, T$, where $\bar{\boldsymbol{R}}_{t-1} = (R_{t-1}, \ldots, R_1)$ and $\bar{\boldsymbol{Y}}_{t-1} = (Y_{t-1}, \ldots, Y_1)$ for $t \geq 2$ and $\bar{\boldsymbol{R}}_0$, $\bar{\boldsymbol{Y}}_0$ are equal to 1. In other words, the probability of observing Y_t is fully determined by the history $(\bar{\boldsymbol{R}}_{t-1}, \bar{\boldsymbol{Y}}_{t-1})$ and baseline covariates. Our interest lies in creating a single imputed data set with fractional weights to facilitate various subsequent analyses.

The fractional imputation procedure proposed in Chapter 4 is readily applicable to this problem. Under current settings, we are able to simplify the process and improve the robustness of estimators of marginal cumulative probabilities. As in Chapter 4, we impose a sequence of conditional models on the responses:

$$G_t^{-1}\Big[\gamma_{t,j}(\bar{\boldsymbol{Y}}_{t-1}, \boldsymbol{X})\Big] = \alpha_{t,j} + b_t(\bar{\boldsymbol{Y}}_{t-1}, \boldsymbol{X}; \boldsymbol{\beta}_t), \quad j = 1, \ldots, J_t - 1, \tag{5.19}$$

for $t = 1, \ldots, T$, where G_t is the link function, $\gamma_{t,j}(\bar{\boldsymbol{Y}}_{t-1}, \boldsymbol{X}) = P(Y_t \leq j \mid \bar{\boldsymbol{Y}}_{t-1}, \boldsymbol{X})$ and $b_t(\cdot)$ is a pre-specified function parameterized by $\boldsymbol{\beta}_t$. The *cut-points* $\alpha_{t,j}$'s satisfy $\alpha_{t,1} \leq \cdots \leq \alpha_{t,J_t-1}$. Let $\boldsymbol{\eta}_t = (\alpha_{t,1}, \ldots, \alpha_{t,J_t-1}, \boldsymbol{\beta}_t')'$ be the parameters in the model for Y_t. In addition, we also postulate a set of models based on (5.18) to characterize the missing data process:

$$\log\left[\frac{\lambda_t(\bar{\boldsymbol{Y}}_{t-1}, \boldsymbol{X}; \boldsymbol{\phi}_t)}{1 - \lambda_t(\bar{\boldsymbol{Y}}_{t-1}, \boldsymbol{X}; \boldsymbol{\phi}_t)}\right] = a_t(\bar{\boldsymbol{Y}}_{t-1}, \boldsymbol{X}; \boldsymbol{\phi}_t), \tag{5.20}$$

for $t = 1, \ldots, T$, where $\lambda_t(\bar{\boldsymbol{Y}}_{t-1}, \boldsymbol{X}; \boldsymbol{\phi}_t) = P(R_t = 1 \mid \bar{\boldsymbol{R}}_{t-1} = \mathbf{1}, \bar{\boldsymbol{Y}}_{t-1}, \boldsymbol{X})$ and $a_t(\cdot)$ is a function given up to the unknown parameters $\boldsymbol{\phi}_t$. Let $\hat{\boldsymbol{\phi}}_t$ be the maximum likelihood estimators of $\boldsymbol{\phi}$ based on the units with $\bar{\boldsymbol{R}}_{t-1} = \mathbf{1}$. Since the data are missing monotonically, Y_t is observed only if $\bar{\boldsymbol{Y}}_{t-1}$ are all observed, therefore the response probabilities can be expressed as a product of conditional response probabilities:

$$\pi_t(\boldsymbol{X}, \bar{\boldsymbol{Y}}_{t-1}; \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_t) = P(\bar{\boldsymbol{R}}_t = 1 \mid \boldsymbol{X}, \bar{\boldsymbol{Y}}_{t-1}) = \prod_{l=1}^{t} \lambda_l(\bar{\boldsymbol{Y}}_{l-1}, \boldsymbol{X}; \boldsymbol{\phi}_l) \tag{5.21}$$

and can be estimated by substituting $\hat{\boldsymbol{\phi}}_l$ for $\boldsymbol{\phi}_l$ for $l = 1, \ldots, t$.

5.3.2 Doubly Robust Fractional Imputation

Let $f(\cdot)$ denote the probability mass functions, then from (5.19) it is not difficult to obtain $f(y_t \mid \boldsymbol{x}, \bar{y}_{t-1}; \boldsymbol{\eta}_t)$ for $t = 1, \ldots, T$. As in Chapter 4, the observed log-likelihood function is given by:

$$\begin{aligned}
l_{obs} &= \sum_{i=1}^{n} \sum_{t=1}^{T} r_{i,t}(1 - r_{i,t+1}) \log f(y_{i,1}, \ldots, y_{i,t} \mid \boldsymbol{x}_i) \\
&= \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{l=1}^{t} r_{i,t}(1 - r_{i,t+1}) \log f(y_{i,l} \mid \boldsymbol{x}_i, \bar{y}_{i,l-1}; \boldsymbol{\eta}_l), \\
&= \sum_{i=1}^{n} \sum_{l=1}^{T} \sum_{t=l}^{T} r_{i,t}(1 - r_{i,t+1}) \log f(y_{i,l} \mid \boldsymbol{x}_i, \bar{y}_{i,l-1}; \boldsymbol{\eta}_l) \\
&= \sum_{i=1}^{n} \sum_{l=1}^{T} r_{i,l} \log f(y_{i,l} \mid \boldsymbol{x}_i, \bar{y}_{i,l-1}; \boldsymbol{\eta}_l),
\end{aligned} \tag{5.22}$$

where it is understood that $r_{i,T+1} = 0$ for all i. This implies that when the models in (5.19) are correct, the parameters $\boldsymbol{\eta}_t$ can be consistently estimated by fitting the model for Y_t alone with observations satisfying $r_{i,t} = 1$. The iterative procedure in Chapter 4 can be simplified by imputing missing values and calculating fractional weights simultaneously. Let $\mathcal{O}^{(0)} = \mathcal{O}$ be the original data set and $\boldsymbol{w}^{(0)} = \mathbf{1}$ be the initial weights. We impute the missing values for each response sequentially from Y_1 to Y_T. Let $\mathcal{O}^{(t)}$ be the data set after Y_t is fully imputed, and let $\boldsymbol{w}^{(t)}$ be the corresponding weights. The third approach in Section 5.1 can be extended to the

longitudinal case. For $t = 1, \ldots, T$, we update the data set $\mathcal{O}^{(t-1)}$ to $\mathcal{O}^{(t)}$ and weights $\boldsymbol{w}^{(t-1)}$ to $\boldsymbol{w}^{(t)}$ through the following steps:

1. For observations with missing values for Y_t in $\mathcal{O}^{(t-1)}$, i.e., $\{i : r_{i,t} = 0\}$, replicate each missing unit J_t times and fill in imputed values $\tilde{y}_{i,tj} = j$, for $j = 1, \ldots, J_t$.

2. Fit the model in (5.19) for Y_t with observed units, that is, $\{i : r_{i,t} = 1\}$ and obtain an estimator $\hat{\boldsymbol{\eta}}_t$ by solving

$$\text{argmax}_{\boldsymbol{\eta}_t \in H_{t,n}} \, n^{-1} \sum_{i=1}^{n} r_{i,t} \log f(y_{i,t} \mid \boldsymbol{x}_i, \bar{\boldsymbol{y}}_{i,t-1}; \boldsymbol{\eta}_t), \tag{5.23}$$

   where

$$H_{t,n} = \left\{ \boldsymbol{\eta}_t : n^{-1} \sum_{i=1}^{n} A_{i,t} = \mathbf{0} \right\},$$

   and

$$\begin{aligned} A_{i,t} = {} & r_{i,t}(\hat{\pi}_{i,t}^{-1} - 1) \left[\boldsymbol{z}_{i,t} - \boldsymbol{\gamma}_t(\bar{\boldsymbol{y}}_{i,t-1}, \boldsymbol{x}_i; \boldsymbol{\eta}_t) \right] \\ & + \sum_{l=1}^{t-1} r_{i,l}(\hat{\pi}_{i,l}^{-1} - 1) \left[\hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{y}}_{i,l}, \boldsymbol{x}_i) - \hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{y}}_{i,l-1}, \boldsymbol{x}_i) \right], \end{aligned}$$

   with $\boldsymbol{Z}_t$ being the cumulative indicator vector of Y_t, $\boldsymbol{\gamma}_t = (\gamma_{t,1}, \ldots, \gamma_{t,J_t-1})$ and $\hat{\pi}_{i,t} = \pi_t(\boldsymbol{x}_i, \bar{\boldsymbol{y}}_{i,t-1}; \hat{\boldsymbol{\phi}}_1, \ldots, \hat{\boldsymbol{\phi}}_t)$ estimated from (5.21). Here $\hat{E}(\cdot)$ denotes the expectation based on the assumed sequential regression models (5.19). For example, $\hat{E}(\boldsymbol{z}_t \mid \bar{\boldsymbol{y}}_{i,t-1}, \boldsymbol{x}_i) = \boldsymbol{\gamma}_t(\bar{\boldsymbol{y}}_{i,t-1}, \boldsymbol{x}_i; \boldsymbol{\eta}_t)$ and

$$\hat{E}(\boldsymbol{z}_t \mid \bar{\boldsymbol{y}}_{i,t-2}, \boldsymbol{x}_i) = \sum_{j=1}^{J_{t-1}} \boldsymbol{\gamma}_t((j, \bar{\boldsymbol{y}}_{i,t-2}), \boldsymbol{x}_i; \boldsymbol{\eta}_t)(\hat{\gamma}_{i,t-1,j} - \hat{\gamma}_{i,t-1,j-1}),$$

   where $\hat{\gamma}_{i,t-1,j} = \gamma_{t,j}(\bar{\boldsymbol{y}}_{i,t-2}, \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_{t-1})$ and $\hat{\boldsymbol{\eta}}_{t-1}$ is obtained when we create $\mathcal{O}^{(t-1)}$.

3. For the observed units with $r_{i,t} = 1$, the weights remain unchanged, while for a unit i with $r_{i,t} = 0$, the original weight $w_i^{(t-1)}$ is split among the cluster of imputed values with the jth imputed value $\tilde{y}_{i,tj}$ receiving a fractional weight:

$$w_{ij}^{(t)} = w_i^{(t-1)} \left[\gamma_{t,j}(\boldsymbol{x}_i, \bar{\boldsymbol{y}}_{i,t-1}; \hat{\boldsymbol{\eta}}_t) - \gamma_{t,j-1}(\boldsymbol{x}_i, \bar{\boldsymbol{y}}_{i,t-1}; \hat{\boldsymbol{\eta}}_t) \right]. \tag{5.24}$$

Consequently, we obtain a new data set $\mathcal{O}^{(t)}$ with weights $\boldsymbol{w}^{(t)}$, where all missing values of Y_t are imputed.

Note that the calculation of $\hat{\pi}_{i,t}$ and $\gamma_{t,j}(\boldsymbol{x}_i, \bar{\boldsymbol{y}}_{i,t-1}; \hat{\boldsymbol{\eta}}_t)$ in STEP 2 and 3 is possible because $\mathcal{O}^{(t-1)}$ has no missing value in $\bar{\boldsymbol{Y}}_{t-1}$. The following theorem presents the double robustness of estimators of marginal cumulative probabilities based on the fractionally imputed data set:

Theorem 5.2. *Based on the final data set $\mathcal{O}^{(T)}$ with $\boldsymbol{w}^{(T)}$, the estimator of marginal cumulative probabilities of Y_t given by*

$$\hat{\boldsymbol{p}}_t^{fi} = \Big\{ \sum_{i \in \mathcal{O}^{(T)}} w_i^{(T)} \Big\}^{-1} \Big\{ \sum_{i \in \mathcal{O}^{(T)}} w_i^{(T)} (\boldsymbol{I}(y_{i,t} \leq 1), \ldots, \boldsymbol{I}(y_{i,t} \leq J_t - 1)) \Big\} \tag{5.25}$$

for $t = 1, \ldots T$, is consistent if either the DGP models (5.19) or the MDP models (5.20) are correct, but not necessarily both.

Proof of the Theorem is presented in Section 5.5. The proposed procedure doubly protects the marginal estimators for all responses against model misspecification, but it is also possible to protect a specific response Y_{t_0} by carrying out the above STEP 2 only when we create $\mathcal{O}^{(t_0)}$ and following the regular fractional imputation procedure when imputing other responses.

We use a special case with $T = 2$ to intuitively illustrate the rationale underlying the constraints in STEP 2. For Y_1, it is easy to check that the constraint is equivalent to that imposed by (5.12) in the univariate case. For Y_2,

$$\begin{aligned} A_{i,2} = {} & r_{i,2}(\hat{\pi}_{i,2}^{-1} - 1) \left[\boldsymbol{z}_{i,2} - \boldsymbol{\gamma}_2(y_{i,1}, \boldsymbol{x}_i; \boldsymbol{\eta}_2) \right] \\ & + r_{i,1}(\hat{\pi}_{i,1}^{-1} - 1) \left[\boldsymbol{\gamma}_2(y_{i,1}, \boldsymbol{x}_i; \boldsymbol{\eta}_2) - \hat{E}(\boldsymbol{Z}_2 \mid \boldsymbol{x}_i) \right], \end{aligned}$$

where $\hat{E}(\boldsymbol{Z}_2 \mid \boldsymbol{x}_i)$ estimates the cumulative probabilities $(P(Y_2 \leq 1 \mid \boldsymbol{x}_i), \ldots, P(Y_2 \leq J_2 - 1 \mid \boldsymbol{x}_i))'$ based on the DGP models. Similar to (5.4), the first term of $\sum_{i=1}^n A_{i,2}$ approximates the potential bias of imputation if every unit with $r_{i,2} = 0$ is imputed in the same way as those with $r_{i,2} = 0$ and $r_{i,1} = 1$. However, in the actual imputation procedure, the units with both Y_1 and Y_2 missing are imputed differently and the actual contribution of such a unit i to the final estimator $\hat{\boldsymbol{p}}_2^{fi}$ is $\hat{E}(\boldsymbol{Z}_2 \mid \boldsymbol{x}_i)$. Therefore, we add a term $\sum_{i=1}^n (1 - r_{i,1}) \left[\boldsymbol{\gamma}_2(y_{i,1}, \boldsymbol{x}_i; \boldsymbol{\eta}_2) - \hat{E}(\boldsymbol{Z}_2 \mid \boldsymbol{x}_i) \right]$ to adjust for the difference.

But this term is not directly observable, so we replace it with an IPW estimator, which is essentially the second term of $\sum_{i=1}^{n} A_{i,2}$. Consequently, $\sum_{i=1}^{n} A_{i,2}$ quantifies the potential bias of estimating $\boldsymbol{p}_2$ based on the actual imputed data set.

5.3.3 Discussion

Since the models in (5.19) are fitted sequentially, there exist naive extensions of the three approaches in Section 5.1 to the longitudinal case. In STEP 2 of the above procedure, when fitting the model for Y_t with observed units, we simply consider Y_t as the single response with missing values and all previous variables $\bar{\boldsymbol{Y}}_{t-1}, \boldsymbol{X}$ are treated as fully observed covariates, although some values of these variables are in fact imputed. We then apply the three approaches in the univariate case to the subset of $\mathcal{O}^{(t-1)}$ involving $(Y_t, \bar{\boldsymbol{Y}}_{t-1}, \boldsymbol{X})$ by substituting $r_{i,t}$ for r_i, $\hat{\pi}_{i,t}$ for $\pi(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})$, $\boldsymbol{z}_{i,t}$ for $\boldsymbol{z}_i$ and $\boldsymbol{\gamma}_t(\bar{\boldsymbol{y}}_{i,t-1}, \boldsymbol{x}_i; \boldsymbol{\eta}_t)$ for $\boldsymbol{\gamma}(\boldsymbol{x}_i; \boldsymbol{\eta})$ and obtain the fractional weights $\breve{w}_{i,tj}^{(a)}$, $w_{i,tj}^{(b)}$ and $w_{i,tj}^{(c)}$ defined in (5.7), (5.10) and (5.13) for the imputed values $\tilde{y}_{i,tj}$. For example, to extend the second approach, we estimate $\boldsymbol{\eta}_t$ in STEP 2 by solving

$$\mathbf{0} = n^{-1} \sum_{i=1}^{n} r_{i,t} \big[\hat{\pi}_{i,t}^{-1} - 1\big] \left\{ \frac{\partial}{\partial \boldsymbol{\eta}_t} G_t^{-1} \big[\boldsymbol{\gamma}_t(\bar{\boldsymbol{y}}_{i,t-1}, \boldsymbol{x}_i; \boldsymbol{\eta}_t)\big] \right\} \big[\boldsymbol{z}_{i,t} - \boldsymbol{\gamma}_t(\bar{\boldsymbol{y}}_{i,t-1}, \boldsymbol{x}_i; \boldsymbol{\eta}_t)\big].$$

In STEP 3, the weights of imputed values are updated to $w_{ij}^{(t,b)} = w_i^{(t-1,b)} w_{i,tj}^{(b)}$ where $w_i^{(t-1,b)}$ is the original weight associated with unit i in $\mathcal{O}^{(t-1)}$ and

$$w_{i,tj}^{(b)} = \gamma_{t,j}(\bar{\boldsymbol{y}}_{i,t-1}, \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_{t(b)}) - \gamma_{t,j-1}(\bar{\boldsymbol{y}}_{i,t-1}, \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_{t(b)}).$$

The theoretical justification of these approaches is not straightforward (See the proof of Theorem 5.2 in Section 5.5 for details), though the simulation results we present in Section 5.4 seem to suggest these methods also lead to doubly robust estimators of marginal cumulative probabilities.

When the data have an intermittent missingness pattern, the naive methods are apparently not applicable, while extending the proposed method based on calibration constraints is theoretically possible. However, two practical issues prevent the implementation of the double robust fractional imputation with intermittently missing data. The first issue is that modelling the MDP is challenging. Robins and Gill (1997) discussed a class of models for non-monotone missing processes, but fitting

these models is computationally cumbersome. Even if we can estimate the response probabilities under some strong assumptions such as the covariate dependent missingness, it is almost impossible to solve the constrained optimization problem numerically with existing algorithms, because in the intermittent case, both the target function and the constraints involve parameters from all the sequential regression models.

5.4 Simulation Studies

Numerical simulations are conducted to examine the finite sample properties of different estimators under various choices of models. The details of models we impose, either true or false, are listed in Table 5.2 for the three problems we consider in this chapter, where "DP" refers to models for the data generating process, "MP" to models for the missing data process and "TP" to models for treatment assignment. The fully observed baseline covariates X_1 and X_2, also called the measured confounders in the causal inference literature, are generated from Exp(1) and N(0.5, 1), respectively. All the ordinal responses we simulate have three ordered categories and are modelled by *cumulative link models* of the form (2.16) with *probit* link function. The sample size is taken as $n = 200$ and $n = 500$, each replicated 2000 times.

Tables 5.3, 5.4 and 5.5 report the results of the univariate missing data problem, the point-treatment causal effect problem and the longitudinal data with dropouts, respectively. For each estimator, the absolute relative bias (ARB) and mean squared error (MSE) are reported. We use P**a**R**b** to indicate the set of models applied, where

$$\mathbf{a} = \boldsymbol{I}(\text{missing data/treatment assignment models are correct}),$$
$$\mathbf{b} = \boldsymbol{I}(\text{data generating models are correct}).$$

For example, P**1**R**0** indicates the case where the missing data/treatment assignment models are correct, but the data generating models are misspecified.

For the univariate missing data problem, each row of Table 5.3 contains the results of one method for estimating the first category probability of the response $\pi_1 = P(Y = 1)$. The three rows denoted by "DFIa", "DFIb" and "DFIc" correspond to the three doubly robust fractional imputation methods in Section 5.1. For the first approach, results based on adjusted fractional weights are reported. For comparison, we also

Table 5.2: Details of the Model Specifications for Simulations

1. Univariate response with missing observations		
DP	True	$\text{probit}\big[\gamma_j(\boldsymbol{X};\boldsymbol{\eta})\big] = \alpha_j + (X_1, X_2, X_1X_2)\boldsymbol{\beta}$, $(\alpha_1, \alpha_2) = (-1, 1)$, $\boldsymbol{\beta} = (2, 1, -4)'$,
	False	$\text{probit}\big[\gamma_j(\boldsymbol{X};\boldsymbol{\eta})\big] = \alpha_j + (X_1, X_2)\boldsymbol{\beta}$;
MP	True	$\text{logit}\big[\pi(\boldsymbol{X};\boldsymbol{\phi})\big] = (1, X_1, X_2, X_1X_2)\boldsymbol{\phi}$, $\boldsymbol{\phi} = (-1, 1, 1, -1)'$,
	False	$\text{logit}\big[\pi(\boldsymbol{X};\boldsymbol{\phi})\big] = (1, X_1, X_2)\boldsymbol{\phi}$;
2. Causal inference in a point-treatment study		
DP	True	$\text{probit}\big[\gamma_j(\boldsymbol{X};\boldsymbol{\eta}^{(1)})\big] = \alpha_j^{(1)} + (X_1, X_2, X_1X_2)\boldsymbol{\beta}^{(1)}$, $\text{probit}\big[\gamma_j(\boldsymbol{X};\boldsymbol{\eta}^{(0)})\big] = \alpha_j^{(0)} + (X_1, X_2, X_1X_2)\boldsymbol{\beta}^{(0)}$, $(\alpha_1^{(1)}, \alpha_2^{(1)}) = (0.5, 1.2)$, $(\alpha_1^{(0)}, \alpha_2^{(0)}) = (-1, 1)$, $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} = (2, 1, -4)'$,
	False	$\text{probit}\big[\gamma_j(\boldsymbol{X};\boldsymbol{\eta}^{(1)})\big] = \alpha_j^{(1)} + (X_1, X_2)\boldsymbol{\beta}^{(1)}$, $\text{probit}\big[\gamma_j(\boldsymbol{X};\boldsymbol{\eta}^{(0)})\big] = \alpha_j^{(0)} + (X_1, X_2)\boldsymbol{\beta}^{(0)}$;
TP	True	$\text{logit}\big[\pi(\boldsymbol{X};\boldsymbol{\phi})\big] = (1, X_1, X_2, X_1X_2)\boldsymbol{\phi}$, $\boldsymbol{\phi} = (-1, 1, 1, -1)'$,
	False	$\text{logit}\big[\pi(\boldsymbol{X};\boldsymbol{\phi})\big] = (1, X_1, X_2)\boldsymbol{\phi}$;
3. Longitudinal data with monotone missingness		
DP	True	$\text{probit}\big[\gamma_{1,j}(\boldsymbol{X};\boldsymbol{\eta}_1)\big] = \alpha_{1,j} + (X_1, X_2, X_1X_2)\boldsymbol{\beta}_1$, $\text{probit}\big[\gamma_{2,j}(\boldsymbol{X}, Y_1;\boldsymbol{\eta}_2)\big] = \alpha_{2,j} + (X_1, X_2, I(Y_1 = 2), I(Y_1 = 3))\boldsymbol{\beta}_2$, $(\alpha_{1,1}, \alpha_{1,2}) = (1, 2.5)$, $\boldsymbol{\beta}_1 = (1, 2, 4)'$, $(\alpha_{2,1}, \alpha_{2,2}) = (1, 3)$, $\boldsymbol{\beta}_2 = (1, 2, 2, -2)'$,
	False	$\text{probit}\big[\gamma_{1,j}(\boldsymbol{X};\boldsymbol{\eta}_1)\big] = \alpha_{1,j} + (X_1, X_2)\boldsymbol{\beta}_1$, $\text{probit}\big[\gamma_{2,j}(\boldsymbol{X}, Y_1;\boldsymbol{\eta}_2)\big] = \alpha_{2,j} + (X_1, X_2, I(Y_1 = 2), I(Y_1 = 3))\boldsymbol{\beta}_2$;
MP	True	$\text{logit}\big[\lambda_1(\boldsymbol{X};\boldsymbol{\phi}_1)\big] = (1, X_1, X_2, X_1X_2)\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_1 = (-1, 2, -1, 4)'$, $\text{logit}\big[\lambda_2(\boldsymbol{X}, Y_1;\boldsymbol{\phi}_2)\big] = (1, X_1, X_2, I(Y_1 = 2), I(Y_1 = 3))\boldsymbol{\phi}_2$, $\boldsymbol{\phi}_2 = (1.5, -1, 1, 1, 1.5)'$,
	False	$\text{logit}\big[\lambda_1(\boldsymbol{X};\boldsymbol{\phi}_1)\big] = (1, X_1, X_2)\boldsymbol{\phi}_1$, $\text{logit}\big[\lambda_2(\boldsymbol{X}, Y_1;\boldsymbol{\phi}_2)\big] = (1, X_1, X_2, I(Y_1 = 2), I(Y_1 = 3))\boldsymbol{\phi}_2$.

considered the complete case analysis (CCA), the inverse probability weighting (IPW) and the regular fractional imputation (FI) proposed in Chapter 3. Results from the true complete data set (COMP) are included as a gold standard as well.

The CCA approach is clearly not valid. The IPW and FI produce consistent estimators when corresponding models are correctly specified, but when the models are misspecified, they both lead to severely biased inferences. For the three proposed estimators, we observe that: (i) they are all consistent if either or both models are true; (ii) the impact of adjustments on the first approach is not significant; (iii) none of them is still valid when both models are wrong. Another interesting observation is that when the MP model is correct, the doubly robust estimators are more efficient than the IPW estimator even when the DP model is wrong. It seems that the incorrect DP model still manages to extract some useful information about the response structure. On the other hand, when the DP model is correct, the proposed estimators lose some efficiency compared to the FI method, but the loss is not significant.

For the point-treatment causal effect problem, we chose level 2 as a reference level in the simulation and estimated the risk ratio given by (5.14). Table 5.4 includes results for the crude risk ratio (CRR) estimator, the inverse-probability-of-treatment weighted estimator (IPTW), the regular fractional imputation (FI) estimator and the doubly robust fractional imputation estimators. Because of the presence of confounders, the CRR is clearly not consistent. The IPTW and FI methods are only valid under correct model specification and results for the proposed estimators confirm the double robustness property.

For the longitudinal data with monotone missingness, we simulated two responses Y_1 and Y_2, with approximately 40% of the subjects dropping out at the first stage and only having baseline covariates observed, 10% of them dropping out during the second stage and having Y_1 observed, the rest 50% finishing the study and having both Y_1 and Y_2 observed. We were interested in the category probabilities of the final response Y_2. Table 5.5 presents the results for estimating the third category probability of Y_2: $\pi_{2,3} = P(Y_2 = 3)$. The proposed doubly robust estimator is denoted by "DFI" and the three naive extensions discussed in Section 5.3 are denoted by "NAIa", "NAIb" and "NAIc". The double robustness of "DFI" is obvious. From the results of this particular simulation study, the three naive extensions seem to provide the same protection for the marginal probabilities, except that the first approach has larger biases when $n = 200$, possibly due to the issue of negative weights. The performance

Table 5.3: Absolute Relative Bias (%) and Mean Squared Error ($\times 10^{-4}$) of Different Estimators of $\pi_1 = P(Y = 1)$

		P**1**R**1**		P**1**R**0**		P**0**R**1**		P**0**R**0**	
n	Methods	ARB	MSE	ARB	MSE	ARB	MSE	ARB	MSE
200	COMP	0.3	(8.9)	0.3	(8.9)	0.3	(8.9)	0.3	(8.9)
	CCA	10.6	(22.6)	10.6	(22.6)	10.6	(22.6)	10.6	(22.6)
	IPW	1.7	(19.1)	1.7	(19.1)	17.2	(30.6)	17.2	(30.6)
	FI	0.1	(13.2)	13.9	(24.5)	0.1	(13.2)	13.9	(24.5)
	DFIa	1.0	(15.4)	2.9	(17.2)	0.2	(14.0)	14.1	(24.0)
	DFIb	0.2	(14.2)	1.2	(15.7)	0.0	(13.4)	18.2	(32.4)
	DFIc	0.2	(14.2)	1.7	(16.2)	0.1	(13.3)	18.7	(33.7)
500	COMP	0.2	(3.6)	0.2	(3.6)	0.2	(3.6)	0.2	(3.6)
	CCA	10.1	(13.1)	10.1	(13.1)	10.1	(13.1)	10.1	(13.1)
	IPW	0.7	(8.3)	0.7	(8.3)	16.9	(22.5)	16.9	(22.5)
	FI	0.1	(5.4)	13.8	(16.7)	0.1	(5.4)	13.8	(16.7)
	DFIa	0.6	(6.3)	0.9	(7.5)	0.0	(5.7)	14.8	(17.8)
	DFIb	0.2	(5.9)	0.3	(6.8)	0.1	(5.5)	18.1	(24.9)
	DFIc	0.1	(5.9)	0.7	(7.0)	0.1	(5.5)	18.6	(25.8)

of other methods is similar to that in the univariate case.

Table 5.4: Absolute Relative Bias (%) and Mean Squared Error ($\times 10^{-3}$) of Different Estimators of the Risk Ratio

		P**1**R**1**		P**1**R**0**		P**0**R**1**		P**0**R**0**	
n	Methods	ARB	MSE	ARB	MSE	ARB	MSE	ARB	MSE
200	COMP	0.1	(1.4)	0.1	(1.4)	0.1	(1.4)	0.1	(1.4)
	CRR	13.4	(13.9)	13.4	(13.9)	13.4	(13.9)	13.4	(13.9)
	IPTW	1.9	(6.7)	1.9	(6.7)	20.4	(23.9)	20.4	(23.9)
	FI	0.1	(3.7)	18.9	(20.9)	0.1	(3.7)	18.9	(20.9)
	DFIa	1.8	(4.6)	0.6	(4.2)	0.5	(4.0)	18.4	(21.0)
	DFIb	0.1	(4.0)	2.0	(5.0)	0.2	(3.8)	20.0	(22.9)
	DFIc	0.1	(4.0)	1.9	(5.0)	0.1	(3.8)	20.9	(24.8)
500	COMP	0.1	(0.6)	0.1	(0.6)	0.1	(0.6)	0.1	(0.6)
	CRR	12.6	(9.6)	12.6	(9.6)	12.6	(9.6)	12.6	(9.6)
	IPTW	0.9	(3.3)	0.9	(3.3)	19.8	(19.4)	19.8	(19.4)
	FI	0.0	(1.4)	18.4	(16.9)	0.0	(1.4)	18.4	(16.9)
	DFIa	0.8	(1.7)	0.3	(1.7)	0.0	(1.5)	18.3	(17.3)
	DFIb	0.0	(1.5)	0.7	(2.0)	0.0	(1.4)	19.3	(18.6)
	DFIc	0.0	(1.5)	0.7	(2.0)	0.0	(1.4)	20.1	(20.0)

Table 5.5: Absolute Relative Bias (%) and Mean Squared Error ($\times 10^{-4}$) of Different Estimators of $\pi_{2,3} = P(Y_2 = 3)$

		P**1**R**1**		P**1**R**0**		P**0**R**1**		P**0**R**0**	
n	Methods	ARB	MSE	ARB	MSE	ARB	MSE	ARB	MSE
200	COMP	0.3	(8.0)	0.3	(8.0)	0.3	(8.0)	0.3	(8.0)
	CCA	20.6	(34.5)	20.6	(34.5)	20.6	(34.5)	20.6	(34.5)
	IPW	1.1	(28.6)	1.1	(28.6)	9.1	(18.2)	9.1	(18.2)
	FI	1.0	(12.6)	8.0	(12.5)	1.0	(12.6)	8.0	(12.5)
	DFI	0.8	(16.6)	0.4	(16.4)	0.7	(13.4)	6.3	(13.1)
	NAIa	3.6	(19.8)	7.0	(21.9)	3.3	(18.2)	1.7	(17.2)
	NAIb	0.7	(16.3)	0.5	(16.3)	0.5	(13.5)	5.0	(12.7)
	NAIc	0.8	(16.1)	0.5	(15.8)	0.7	(13.3)	6.1	(12.8)
500	COMP	0.2	(3.1)	0.2	(3.1)	0.2	(3.1)	0.2	(3.1)
	CCA	19.5	(21.5)	19.5	(21.5)	19.5	(21.5)	19.5	(21.5)
	IPW	0.4	(12.6)	0.4	(12.6)	9.9	(10.0)	9.9	(10.0)
	FI	0.1	(4.7)	8.7	(6.8)	0.1	(4.7)	8.7	(6.8)
	DFI	0.3	(6.6)	0.5	(6.7)	0.0	(5.2)	7.2	(6.5)
	NAIa	1.0	(7.5)	3.3	(8.5)	1.0	(7.2)	3.4	(7.3)
	NAIb	0.3	(6.4)	0.7	(6.5)	0.1	(5.2)	5.4	(5.7)
	NAIc	0.3	(6.4)	0.6	(6.5)	0.0	(5.2)	6.6	(6.2)

5.5 Regularity Conditions and Proofs

5.5.1 Regularity Conditions and Proof of Theorem 5.1

Let

$$\boldsymbol{U}_{(a)}(r, y, \boldsymbol{x}; \boldsymbol{\eta}_{(a)}, \boldsymbol{\phi}) = r \left\{ \frac{\partial}{\partial \boldsymbol{\eta}_{(a)}} G^{-1}\left[\tilde{\boldsymbol{\gamma}}(\boldsymbol{x}; \boldsymbol{\eta}_{(a)}, \boldsymbol{\phi})\right] \right\} \left[\boldsymbol{z} - \tilde{\boldsymbol{\gamma}}(\boldsymbol{x}; \boldsymbol{\eta}_{(a)}, \boldsymbol{\phi})\right]$$

be the estimating function in (5.6) and $\boldsymbol{T}(r, \boldsymbol{x}; \boldsymbol{\phi})$ be the estimating function in (3.3). Let also $\tilde{\boldsymbol{\theta}}_{(a)} = (\boldsymbol{\eta}'_{(a)}, \boldsymbol{\phi}')'$ and $\tilde{\boldsymbol{U}}_{(a)}(r, y, \boldsymbol{x}; \tilde{\boldsymbol{\theta}}_{(a)}) = (\boldsymbol{U}'_{(a)}(r, y, \boldsymbol{x}; \boldsymbol{\eta}_{(a)}, \boldsymbol{\phi}), \boldsymbol{T}'(r, \boldsymbol{x}; \boldsymbol{\phi}))'$. Assume that $\tilde{\boldsymbol{U}}_{(a)}(r, y, \boldsymbol{x}; \tilde{\boldsymbol{\theta}}_{(a)})$ satisfies condition S1-S4 in Theorem 2.1 and the unique root is denoted by $\tilde{\boldsymbol{\theta}}^*_{(a)} = (\boldsymbol{\eta}^{*'}_{(a)}, \boldsymbol{\phi}^{*'})'$. Note that the root is not necessarily $(\boldsymbol{\eta}'_{(a)0}, \boldsymbol{\phi}'_0)'$ because models in (5.1) and (3.2) can be misspecified.

When the DGP model (5.1) is correct, it can be easily checked that $\boldsymbol{\eta}^*_{(a)} = \boldsymbol{\eta}_{(a)0} = (\alpha_{1,0}, \ldots, \alpha_{J-1,0}, \boldsymbol{\beta}'_0, 0, \ldots, 0)'$ where $(\alpha_{1,0}, \ldots, \alpha_{J-1,0}, \boldsymbol{\beta}'_0)$ are the true values of parameters in (5.1). The $\hat{\boldsymbol{p}}^{fi}$ can be treated as the solution to

$$\boldsymbol{0} = n^{-1} \sum_{i=1}^{n} \boldsymbol{U}_{imp}(r_i, y_i, \boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_{(a)}, \hat{\boldsymbol{\phi}}),$$

where

$$\boldsymbol{U}_{imp}(r, y, \boldsymbol{x}; \boldsymbol{\eta}_{(a)}, \boldsymbol{\phi}) = r(\boldsymbol{z} - \boldsymbol{p}) + (1 - r) \sum_{j=1}^{J} w_j^{(a)} (\boldsymbol{c}_j - \boldsymbol{p}).$$

Therefore, $(\boldsymbol{\eta}'_{(a)0}, \boldsymbol{\phi}^{*'}, \boldsymbol{p}'_0)'$ is the unique root of $E(\tilde{\boldsymbol{U}}'_{(a)}, \boldsymbol{U}'_{imp})' = \boldsymbol{0}$, and thus the consistency of $\hat{\boldsymbol{p}}^{fi}$ follows.

When the MDP model (3.2) is correct, it is apparent that $\boldsymbol{\phi}^* = \boldsymbol{\phi}_0$, the true parameter values in the model (3.2). From (5.6), note that

$$\frac{\partial}{\partial(\alpha_1, \ldots, \alpha_{J-1})} G^{-1}\left[\tilde{\boldsymbol{\gamma}}(\boldsymbol{x}_i; \boldsymbol{\eta}_{(a)}, \hat{\boldsymbol{\phi}})\right] = \mathbb{I},$$

where $\mathbb{I}$ is the identity matrix of dimension $J - 1$ and

$$\frac{\partial}{\partial(\nu_1, \ldots, \nu_{J-1})} G^{-1}\left[\tilde{\boldsymbol{\gamma}}(\boldsymbol{x}_i; \boldsymbol{\eta}_{(a)}, \hat{\boldsymbol{\phi}})\right] = -\pi^{-1}(\boldsymbol{x}_i; \hat{\boldsymbol{\phi}})\mathbb{I},$$

so we have

$$n^{-1}\sum_{i=1}^{n} r_i\big[\pi^{-1}(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}})-1\big]\big[\boldsymbol{z}_i-\sum_{j=1}^{J} w_{ij}^{(a)}\boldsymbol{c}_j\big]=\mathbf{0},$$

the second term in (5.4) with the response probability estimated by $\pi(\boldsymbol{x}_i;\hat{\boldsymbol{\phi}})$ is equal to zero. Let

$$\boldsymbol{U}_{rp}(r,y,\boldsymbol{x};\boldsymbol{\eta}_{(a)},\boldsymbol{\phi},P)=\big[1-r\pi^{-1}(\boldsymbol{x};\boldsymbol{\phi})\big]\big[\boldsymbol{z}-\sum_{j=1}^{J} w_j^{(a)}(\boldsymbol{x};\boldsymbol{\eta}_{(a)})\boldsymbol{c}_j\big]-P,$$

then the first term in (5.4) can be considered as the solution $\hat{P}$ to

$$\mathbf{0}=n^{-1}\sum_{i=1}^{n}\boldsymbol{U}_{rp}(r_i,y_i,\boldsymbol{x}_i;\hat{\boldsymbol{\eta}}_{(a)},\hat{\boldsymbol{\phi}},P).$$

It is obvious that $(\boldsymbol{\eta}_{(a)}^{*\prime},\boldsymbol{\phi}_0',\mathbf{0}')'$ is the solution to $E\big[(\tilde{\boldsymbol{U}}_{(a)}',\boldsymbol{U}_{rp}')\big]=\mathbf{0}$, and hence $\hat{P}\xrightarrow{p}\mathbf{0}$. It then follows that the bias in (5.4) converges to $\mathbf{0}$ in probability and $\hat{\boldsymbol{p}}^{fi}$ is consistent.

The double robustness of $\hat{\boldsymbol{p}}^{fi}$ based on the second can be shown similarly. For the third approach, let

$$\boldsymbol{U}_{(c1)}(r,y,\boldsymbol{x};\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\lambda})=r\boldsymbol{S}(\boldsymbol{z},\boldsymbol{x};\boldsymbol{\eta})-\boldsymbol{\lambda}'r\big[\pi^{-1}(\boldsymbol{x};\boldsymbol{\phi})-1\big]\frac{\partial}{\partial\boldsymbol{\eta}}\boldsymbol{\gamma}(\boldsymbol{x};\boldsymbol{\eta}),$$

and

$$\boldsymbol{U}_{(c2)}(r,y,\boldsymbol{x};\boldsymbol{\eta},\boldsymbol{\phi})=r\big[\pi^{-1}(\boldsymbol{x};\boldsymbol{\phi})-1\big]\big[\boldsymbol{z}-\boldsymbol{\gamma}(\boldsymbol{x};\boldsymbol{\eta})\big].$$

We note that $(\boldsymbol{\eta}_0,\boldsymbol{\phi}^*,\mathbf{0})$ is the root of $E\big[(\boldsymbol{U}_{(c1)}',\boldsymbol{U}_{(c2)}',\boldsymbol{T}')\big]=\mathbf{0}$, when the DGP model is correct. The rest of the proof is the same as that for the first approach.

5.5.2 Regularity Conditions and Proof of Theorem 5.2

It is important to note that the estimator of marginal cumulative probabilities of Y_t based on the final dataset $\mathcal{O}^{(T)}$ is the same as the one based on $\mathcal{O}^{(t)}$ which has all missing values of Y_t imputed. This follows immediately from the fractional imputation procedure in Section 5.3, because in the subsequent steps, units in $\mathcal{O}^{(t)}$ are either kept unchanged or replicated with weights split among the replicated observations.

In either way, the estimator of marginal cumulative probabilities of Y_t is unaffected. It then suffices to consider the estimator based on $\mathcal{O}^{(t)}$.

We first consider the case when the DGP models (5.19) are correct. Let the score function of $\boldsymbol{\eta}_t$ be denoted by $\boldsymbol{S}_t(y_t, \bar{\boldsymbol{y}}_{t-1}, \boldsymbol{x}; \boldsymbol{\eta}_t)$. By the Lagrange multiplier method, the constrained optimization problem (5.23) can be considered as solving the following equations:

$$\begin{aligned}
\mathbf{0} &= n^{-1} \sum_{i=1}^{n} \left\{ r_{i,t} \boldsymbol{S}_t(y_{i,t}, \bar{\boldsymbol{y}}_{i,t-1}, \boldsymbol{x}_i; \boldsymbol{\eta}_t) - \boldsymbol{\lambda}_t' \frac{\partial}{\partial \boldsymbol{\eta}_t} A_t(\bar{\boldsymbol{r}}_{i,t}, \bar{\boldsymbol{y}}_{i,t}, \boldsymbol{x}_i; \boldsymbol{\eta}_t, \hat{\bar{\boldsymbol{\eta}}}_{t-1}, \hat{\boldsymbol{\phi}}) \right\}, \\
\mathbf{0} &= n^{-1} \sum_{i=1}^{n} A_t(\bar{\boldsymbol{r}}_{i,t}, \bar{\boldsymbol{y}}_{i,t}, \boldsymbol{x}_i; \boldsymbol{\eta}_t, \hat{\bar{\boldsymbol{\eta}}}_{t-1}, \hat{\boldsymbol{\phi}}),
\end{aligned}$$

where $\hat{\bar{\boldsymbol{\eta}}}_{t-1} = (\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_{t-1})$ and

$$\begin{aligned}
A_t(\bar{\boldsymbol{r}}_t, \bar{\boldsymbol{y}}_t, \boldsymbol{x}; \boldsymbol{\eta}_t, \bar{\boldsymbol{\eta}}_{t-1}, \boldsymbol{\phi}) =& r_t(\pi_t^{-1} - 1) \left[\boldsymbol{z}_t - \boldsymbol{\gamma}_t(\bar{\boldsymbol{y}}_{t-1}, \boldsymbol{x}; \boldsymbol{\eta}_t) \right] \\
&+ \sum_{l=1}^{t-1} r_l(\pi_l^{-1} - 1) \left[\hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{y}}_l, \boldsymbol{x}) - \hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{y}}_{l-1}, \boldsymbol{x}) \right].
\end{aligned}$$

As in Section 5.5.1, we assume that conditions S1-S4 of Theorem 2.1 hold for the joint estimating functions. Note that, by the MAR assumption

$$\begin{aligned}
& E\left\{ R_l(\pi_l^{-1} - 1) \big[\hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{Y}}_l, \boldsymbol{X}) - \hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{Y}}_{l-1}, \boldsymbol{X}) \big] \right\} \\
=& E\left\{ E\big[R_l(\pi_l^{-1} - 1) \mid \bar{\boldsymbol{Y}}_{l-1}, \boldsymbol{X} \big] \Big[E\big[\hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{Y}}_l, \boldsymbol{X}) \mid \bar{\boldsymbol{Y}}_{l-1}, \boldsymbol{X} \big] - \hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{Y}}_{l-1}, \boldsymbol{X}) \Big] \right\}
\end{aligned}$$

and when models (5.19) are correct

$$E\big[\hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{Y}}_l, \boldsymbol{X}) \mid \bar{\boldsymbol{Y}}_{l-1}, \boldsymbol{X} \big] = \hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{Y}}_{l-1}, \boldsymbol{X}),$$

evaluated at the true parameter values $\bar{\boldsymbol{\eta}}_{t0} = (\boldsymbol{\eta}_{10}, \ldots, \boldsymbol{\eta}_{t0})$. Therefore, we have

$$E\big[A_t(\bar{\boldsymbol{R}}_t, \bar{\boldsymbol{Y}}_t, \boldsymbol{X}; \bar{\boldsymbol{\eta}}_{t0}, \boldsymbol{\phi}^*) \big] = \mathbf{0}.$$

Then by arguments in the proof of Theorem 5.1, $\hat{\boldsymbol{p}}_t^{fi}$ is consistent.

If the MDP models (5.20) are correct, we have

$$\begin{aligned}
\bar{\boldsymbol{p}}_t - \hat{\boldsymbol{p}}_t^{fi} =& \sum_{l=1}^{t} r_{i,l-1}(1 - r_{i,l})\big[\boldsymbol{z}_{i,t} - \hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{y}}_{i,l-1}, \boldsymbol{x}_i)\big] \\
=& (1 - r_{i,t})\big[\boldsymbol{z}_{i,t} - \boldsymbol{\gamma}_t(\bar{\boldsymbol{y}}_{i,t-1}, \boldsymbol{x}_i; \boldsymbol{\eta}_t)\big] \\
&+ \sum_{l=1}^{t-1}(1 - r_{i,l})\left[\hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{y}}_{i,l}, \boldsymbol{x}_i) - \hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{y}}_{i,l-1}, \boldsymbol{x}_i)\right] \\
=& A_{i,t} + (1 - r_{i,t}\hat{\pi}_{i,t}^{-1})\big[\boldsymbol{z}_{i,t} - \boldsymbol{\gamma}_t(\bar{\boldsymbol{y}}_{i,t-1}, \boldsymbol{x}_i; \boldsymbol{\eta}_t)\big] \\
&+ \sum_{l=1}^{t-1}(1 - r_{i,l}\hat{\pi}_{i,l}^{-1})\left[\hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{y}}_{i,l}, \boldsymbol{x}_i) - \hat{E}(\boldsymbol{Z}_t \mid \bar{\boldsymbol{y}}_{i,l-1}, \boldsymbol{x}_i)\right].
\end{aligned}$$

Note that

$$E\big[1 - R_l\pi_l^{-1} \mid \bar{\boldsymbol{Y}}_{l-1}, \boldsymbol{X}\big] = \boldsymbol{0} \quad \text{for } l = 1, \ldots, t,$$

evaluated at the true parameter values $\boldsymbol{\phi}_0$. By arguments in Section 5.5.1, $\bar{\boldsymbol{p}}_t - \hat{\boldsymbol{p}}_t^{fi}$ converges to $\boldsymbol{0}$ and thus $\hat{\boldsymbol{p}}_t^{fi}$ is consistent.

For the naive extensions, the consistency under the correct DGP models is easy to show. When the DGP models are misspecified, it can be checked that

$$\sum_{i=1}^{n} r_{i,t}(\pi_{i,t}^{-1} - 1)\left[\boldsymbol{z}_{i,t} - \boldsymbol{\gamma}_t(\bar{\boldsymbol{y}}_{i,t-1}, \boldsymbol{x}_i; \boldsymbol{\eta}_t)\right] = \boldsymbol{0},$$

for $\boldsymbol{\eta}_t = \hat{\boldsymbol{\eta}}_{t(a)}$, $\hat{\boldsymbol{\eta}}_{t(b)}$ and $\hat{\boldsymbol{\eta}}_{t(c)}$, the estimates corresponding to the three approaches, which implies that the naive extensions only eliminate the bias if all units with $r_{i,t} = 0$ are imputed in the same way as those with $r_{i,t} = 0$ and $r_{i,t-1} = 1$ and fail to adjust for the true bias of the actual imputation procedure.

Chapter 6

Discussion and Future Work

In previous chapters, we proposed a fractional imputation procedure based on sequential regression modelling and addressed the important problem in missing data literature on the creation of a single complete data set to facilitate various subsequent analyses by multiple users when variables of mixed types are subject to missingness. In this chapter, we conclude the thesis with a discussion on the integration of the proposed method with another main research topic, the analysis of complex survey data. We also point out some other directions that deserve further consideration in the future.

6.1 Fractional Imputation for Complex Survey Data

Statistical agencies are one of the most important sources of public use data files and these data sets are often collected through carefully designed complex surveys. The proposed fractional imputation procedure can be easily adapted to integrate the design features of the complex survey data by combining the fractional weights and the design weights. We take the simple case in Chapter 3 as an example to demonstrate the idea of incorporating surveys weights into the proposed procedure.

We first consider the problem of parameter estimation in the context of survey sampling without missing responses. Let $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, N\}$ be a finite population, from which a sample is drawn according to some sampling design. Let $\boldsymbol{I}_i$ be selection indicator for the ith unit, such that $\boldsymbol{I}_i = 1$ if the unit is selected in

the sample and $\boldsymbol{I}_i = 0$ otherwise. The first-order inclusion probabilities q_i are defined as $P(\boldsymbol{I}_i = 1)$ for $i = 1, \ldots, N$, which are given by the sampling design. Let $\{(d_i, y_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\}$ denote the sample of size n selected from the finite population, where $d_i = q_i^{-1}$ are called the basic design weights.

In typical survey sampling problems, interest lies on estimating finite population quantities, such as the population mean $\mu = N^{-1} \sum_{i=1}^{N} y_i$. In such cases, the most popular way of inferences with survey data is the "design-based" approach. Here the values $(y_i, \boldsymbol{x}_i)$ are treated as nonrandom and the selection process, reflected by indicator variables $\boldsymbol{I}_i$, is the only source of randomization. For public use data files, however, it is not unusual that the user is interested in investigating widely applicable models underlying the finite population. The imputation of missing responses also requires fitting models for the data generating process. To make inferences on the parameters in these models with survey data, a different approach, sometimes called the "joint-randomization" framework, is more appropriate. Under this framework, the finite population is assumed to be *i.i.d.* realizations of variables $(Y, \boldsymbol{X})$ following a distribution $F(y, \boldsymbol{x})$, often known as the *superpopulation* distribution. The sample can thus be thought of as "a second phase of sampling" from the *superpopulation* (Godambe and Thompson 1986; Binder and Roberts 2003). When evaluating estimators based on the sample, we take two sources of randomization into account, the first generating the finite population from the *superpopulation* distribution and the second selecting samples from the finite population.

Let $\boldsymbol{\theta}$ be a parameter of the *superpopulation* distribution defined by an unbiased estimating function $\boldsymbol{U}(y, \boldsymbol{x}; \boldsymbol{\theta})$ such that

$$E_\xi \big[\boldsymbol{U}(Y, \boldsymbol{X}; \boldsymbol{\theta})\big] = \mathbf{0},$$

for some $\boldsymbol{\theta}_0$, where the expectation $E_\xi(\cdot)$ is taken with respect to the *superpopulation* distribution. The interest lies in making inferences on $\boldsymbol{\theta}$. Because the finite population is treated as an *i.i.d.* sample of $(Y, \boldsymbol{X})$, by Theorem 2.1, we are able to estimate $\boldsymbol{\theta}$ by solving

$$\mathbf{0} = N^{-1} \sum_{i=1}^{N} \boldsymbol{U}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}), \tag{6.1}$$

if the whole finite population is surveyed. In reality, however, inferences have to be based on the sample rather than the finite population. A natural idea motivated by

the design-based approach is to estimate the right hand side of (6.1), which can be viewed as a population mean, using the Horvitz-Thompson (HT) estimator

$$N^{-1}\sum_{i=1}^{n} d_i \boldsymbol{U}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}),$$

which is unbiased with respect to the sampling design. It follows that an estimator $\hat{\boldsymbol{\theta}}_n$ based on the sample can be obtained as the solution to

$$\mathbf{0} = \sum_{i=1}^{n} d_i \boldsymbol{U}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}). \tag{6.2}$$

Carrillo et al. (2010) proved the consistency of $\hat{\boldsymbol{\theta}}_n$ jointly under the *superpopulation* distribution and the sampling design for generalized estimating equations.

Our main focus is on a sample with missing responses denoted by $\{(d_i, r_i, y_i, \boldsymbol{x}_i),\ i = 1, \ldots, n\}$ where r_i is the response indicator of the ith sampled unit, such that $r_i = 1$ if y_i is observed and $r_i = 0$ otherwise. To estimate $\boldsymbol{\theta}$, a direct application of (6.2) is impossible, since y_i is not always observed. As discussed in Shao and Steel (1999) and Kim and Rao (2009), we envisage that the response indicators also exist in the finite population, i.e., the finite population consists of units $\{(r_i, y_i, \boldsymbol{x}_i),\ i = 1, \ldots, N\}$, which are *i.i.d.* realizations of the *superpopulation* variables $(R, Y, \boldsymbol{X})$. We assume that the data are missing at random at the finite population level in the sense that

$$P(R = 1 \mid Y, \boldsymbol{X}) = P(R = 1 \mid \boldsymbol{X}).$$

We further assume that the response Y depends on $\boldsymbol{X}$ through a model $f(y \mid \boldsymbol{x}; \boldsymbol{\eta})$ parameterized by $\boldsymbol{\eta}$. For ordinal responses, the model could take the form of (2.16). By applying the proposed procedure to the finite population, we can estimate $\boldsymbol{\theta}$ by solving

$$\mathbf{0} = N^{-1}\sum_{i=1}^{N}\left\{ r_i \boldsymbol{U}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) + (1 - r_i)\sum_{j=1}^{J} w_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_p)\boldsymbol{U}(j, \boldsymbol{x}; \boldsymbol{\theta})\right\}, \tag{6.3}$$

where $w_j(\boldsymbol{x};\boldsymbol{\eta})$ is defined in (3.13) and $\hat{\boldsymbol{\eta}}_p$ is the solution to

$$\mathbf{0} = N^{-1}\sum_{i=1}^{N} r_i \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\eta}), \tag{6.4}$$

with $\boldsymbol{S}(\boldsymbol{z}, \boldsymbol{x}; \boldsymbol{\eta})$ defined in (2.19) and $\boldsymbol{z}_i$ being the cumulative indicator vector of y_i. We then substitute the corresponding HT estimators for all population quantities in (6.3) and (6.4) and a sample-based estimator $\hat{\boldsymbol{\theta}}^{fi}$ is given by solving

$$\mathbf{0} = N^{-1}\sum_{i=1}^{n}\left\{ r_i d_i \boldsymbol{U}(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) + (1-r_i)\sum_{j=1}^{J} d_i w_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_s)\boldsymbol{U}(j, \boldsymbol{x}; \boldsymbol{\theta})\right\}, \tag{6.5}$$

where $\hat{\boldsymbol{\eta}}_s$ solves

$$\mathbf{0} = N^{-1}\sum_{i=1}^{n} r_i d_i \boldsymbol{S}(\boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\eta}). \tag{6.6}$$

The factor N^{-1} in (6.5) and (6.6) is not required for computational purposes.

Therefore, a modified fractional imputation procedure to incorporate survey weights involves the following steps: (i) fit the imputation model with observed units weighted by the survey weights d_i as shown in (6.6); (ii) replicate the units with missing values J times including the survey weights and fill in imputed values $\tilde{y}_{ij} = j$ for $j = 1, \ldots, J$; (iii) assign a fractional weight to the jth imputed observation:

$$w_{ij} = w_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_s) = \gamma_j(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_s) - \gamma_{j-1}(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}_s), \quad \text{for } j = 1, \ldots, J,$$

so that the final weight of the jth imputed observation is $d_i w_{ij}$. The resampling approach to variance estimation can be straightforwardly adapted by integrating the survey weights of each replication sample. Subsequent analyses can be carried out by treating the fractionally imputed data set as a normal complete data set with survey weights and by solving equations similar to (6.2).

The above discussions can be easily extended to the general case introduced in Chapter 4 by two modified stages: (i) replicate the survey weights when imputing the incomplete observations in *Stage One* and (ii) weight the estimating equations with survey weights when fitting the imputation models iteratively in *Stage Two*.

6.2 Future Work

The current work of this thesis can be continued or extended in several directions:

(1) Applications to real data sets. Although we have shown the power of the proposed methods through extensive simulation studies, identification of real world problems and applications of our proposed methods to those problems are of great interest. We are currently investigating several possibilities in real survey data and for causal inference.

(2) Unconventional problems for subsequent analyses. We discussed a fairly flexible class of parameters defined by unbiased estimating equations, but there exist other inferential problems the users may be interested in, for example, the model-based contingency table analysis for bivariate ordinal responses. Some of them belong the above-mentioned class and some do not. It is interesting to investigate the performance of these analyses based on the fractionally imputed data set.

(3) Applications of the doubly robust fractional imputation method to longitudinal causal inference problems. As for the univariate case, the doubly robust method we developed for longitudinal missing data can be applied to longitudinal causal inference based on marginal structural models (Robins et al. 2000).

(4) Derivations of the asymptotic properties of estimators based on the fractionally imputed survey data set. We sketched the idea of applying the proposed method to survey data, yet the asymptotic properties of subsequent estimators need to be rigorously derived under the joint-randomization framework.

(5) More flexible imputation models. A set of correct imputation models is a crucial part of the proposed method. The use of semi- or non-parametric modelling techniques would greatly improve the robustness of the proposed method, but at the same time they also bring about much heavier computational burden, especially for data files with a large set of variables. Noting that our method is based on sequential regression modelling, a possible solution would be to use semi- or non-parametric models only for variables about which we do not have sufficient information.

(6) Double robustness for general inferential problems. An alternative way to improve robustness is to incorporate the MDP models into the imputation procedure as we did in Chapter 5. We only considered the estimation of marginal mean responses, and an interesting question would be whether the same idea could be applied to doubly protect other parameters, such the regression coefficients.

References

Agresti, A. (2010), *Analysis of ordinal categorical data*, Wiley Series in Probability and Statistics, 2 edn, John Wiley & Sons. 5, 6, 18

Agresti, A. (2013), *Categorical data analysis*, Wiley Series in Probability and Statistics, 3 edn, John Wiley & Sons. 5

Bang, H., and Robins, J. M. (2005), "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61(4), 962–973. 96, 99, 100

Binder, D. A., and Roberts, G. R. (2003), "Design-based and model-based methods for estimating model parameters," in *Analysis of survey data*, eds. R. L. Chamber, and C. J. Skinner John Wiley & Sons. 123

Boggs, P. T., and Tolle, J. W. (1995), "Sequential quadratic programming," *Acta Numerica*, 4, 1–51. 102

Brant, R. (1990), "Assessing proportionality in the proportional odds model for ordinal logistic regression," *Biometrics*, 46(4), 1171–1178. 19

Breusch, T. S., and Pagan, A. R. (1980), "The Lagrange multiplier test and its applications to model specification in econometrics," *The Review of Economic Studies*, 47(1), 239–253. 102

Brick, J. M., and Kalton, G. (1996), "Handling missing data in survey research," *Statistical methods in medical research*, 5(3), 215–238. 3

Buuren, S., and Groothuis-Oudshoorn, K. (2011), "mice: Multivariate imputation by chained equations in R," *Journal of statistical software*, 45(3). 60, 77

Carrillo, I. A., Chen, J., and Wu, C. (2010), "The pseudo-GEE approach to the analysis of longitudinal surveys," *Canadian Journal of Statistics*, 38(4), 540–554. 124

Cole, S. R., Allison, P. D., and Ananth, C. V. (2004), "Estimation of cumulative odds ratios," *Annals of Epidemiology*, 14(3), 172–178. 19

Cole, S. R., and Ananth, C. V. (2001), "Regression models for unconstrained, partially or fully constrained continuation odds ratios," *International Journal of Epidemiology*, 30(6), 1379–1382. 6

Cox, C. (1995), "Location?scale cumulative odds models for ordinal data: A generalized non-linear model approach," *Statistics in Medicine*, 14(11), 1191–1203. 19

Dale, J. R. (1986), "Global cross-ratio models for bivariate, discrete, ordered responses," *Biometrics*, 42(4), 909–917. 7

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38. 71

Efron, B. (1994), "Missing data, imputation, and the bootstrap," *Journal of the American Statistical Association*, 89(426), 463–475. 32

Ekholm, A., Jokinen, J., McDonald, J. W., and Smith, P. W. (2003), "Joint regression and association modeling of longitudinal ordinal data," *Biometrics*, 59(4), 795–803. 7

Fan, J., and Gijbels, I. (1996), *Local polynomial modelling and its applications*, Chapman and Hall/CRC Monographs on Statistics and Applied Probability, Chapman and Hall/CRC. 6

Fay, R. E. (1996), "Alternative paradigms for the analysis of imputed survey data," *Journal of the American Statistical Association*, 91(434), 490–498. 5, 27

Godambe, V. P. (1991), *Estimating Functions*, Oxford Statistical Science Series, Clarendon Press. 10

Godambe, V. P., and Thompson, M. E. (1986), "Parameters of superpopulation and survey population: their relationships and estimation," *International Statistical Review/Revue Internationale de Statistique*, 54(4), 127–138. 123

Goodman, L. A., and Kruskal, W. H. (1954), "Measures of association for cross classifications," *Journal of the American Statistical Association*, 49(268), 732–764. 6, 17

Han, P., and Wang, L. (2013), "Estimation with missing data: beyond double robustness," *Biometrika*, 100(2), 417–430. 3

Heagerty, P. J., and Zeger, S. L. (1996), "Marginal regression models for clustered ordinal measurements," *Journal of the American Statistical Association*, 91(435), 1024–1036. 7

Henderson, H. V., and Searle, S. R. (1981), "On deriving the inverse of a sum of matrices," *Siam Review*, 23(1), 53–60. 56

Horvitz, D. G., and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, 47(260), 663–685. 3

Johnston, K., Gustafson, P., Levy, A., and Grootendorst, P. (2008), "Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research," *Statistics in medicine*, 27(9), 1539–1556. 104

Kalton, G., and Kish, L. (1984), "Some efficient random imputation methods," *Communications in Statistics-Theory and Methods*, 13(16), 1919–1939. 4, 5

Kang, J. D., and Schafer, J. L. (2007), "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical Science*, 22(4), 523–539. 98, 100

Kendall, M. G. (1945), "The treatment of ties in ranking problems," *Biometrika*, 33(3), 239–251. 6, 17

Kenward, M. G., and Carpenter, J. (2007), "Multiple imputation: current perspectives," *Statistical Methods in Medical Research*, 16(3), 199–218. 60

Kim, J.-H. (2003), "Assessing practical significance of the proportional odds assumption," *Statistics & Probability Letters*, 65(3), 233–239. 19

Kim, J. K., Brick, J. M., Fuller, W. A., and Kalton, G. (2006), "On the bias of the multiple-imputation variance estimator in survey sampling," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3), 509–521. 4

Kim, J. K., and Fuller, W. A. (2004), "Fractional hot deck imputation," *Biometrika*, 91(3), 559–578. 5

Kim, J. K., and Rao, J. N. K. (2009), "A unified approach to linearization variance estimation from survey data after imputation for item nonresponse," *Biometrika*, 96(4), 917–932. 32, 124

Kim, J. K., and Riddles, M. K. (2012), "Some Theory for Propensity-score-adjustment Estimators in Survey Sampling," *Survey Methodology*, 38, 157–65. 3

Lang, J. B. (2008), "Score and profile likelihood confidence intervals for contingency table parameters," *Statistics in Medicine*, 27(28), 5975–5990. 6

Lavori, P. W., Dawson, R., and Shera, D. (1995), "A multiple imputation strategy for clinical trials with truncation of patient data," *Statistics in Medicine*, 14(17), 1913–1925. 4

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 73(1), 13–22. 6

Lindsey, J., Jones, B., and Ebbutt, A. (1997), "Simple models for repeated ordinal responses with an application to a seasonal rhinitis clinical trial," *Statistics in Medicine*, 16(24), 2873–2882. 7

Lipsitz, S. R., Kim, K., and Zhao, L. (1994), "Analysis of repeated categorical data using generalized estimating equations," *Statistics in Medicine*, 13(11), 1149–1163. 7

Little, R. J. (1995), "Modeling the drop-out mechanism in repeated-measures studies," *Journal of the American Statistical Association*, 90(431), 1112–1121. 11, 61

Little, R. J., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data*, Wiley Series in Probability and Statistics, Wiley. 1, 11

McCullagh, P. (1980), "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2), 109–142. 6

McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models*, Chapman and Hall/CRC Monographs on Statistics and Applied Probability, Chapman and Hall/CRC. 6

Meng, X. L. (1994), "Multiple-imputation inferences with uncongenial sources of input," *Statistical Science*, 9(4), 538–558. 4, 44

Molenberghs, G., and Kenward, M. (2007), *Missing data in clinical studies*, Statistics in Practice, John Wiley & Sons. 2, 11, 12, 37, 71

Molenberghs, G., and Lesaffre, E. (1994), "Marginal modeling of correlated ordinal data using a multivariate Plackett distribution," *Journal of the American Statistical Association*, 89(426), 633–644. 7

Müller, G., and Czado, C. (2005), "An autoregressive ordered probit model with application to high-frequency financial data," *Journal of Computational and Graphical Statistics*, 14(2), 320–338. 7

Newey, W. K., and McFadden, D. (1994), "Large sample estimation and hypothesis testing," in *Handbook of Econometrics* North Holland. 10

Nielsen, S. F. (2003), "Proper and improper multiple imputation," *International Statistical Review*, 71(3), 593–607. 4

Parsons, N. R., Edmondson, R. N., and Gilmour, S. G. (2006), "A generalized estimating equation method for fitting autocorrelated ordinal score data with an application in horticultural research," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 55(4), 507–524. 7

Pearl, J. (2010), "On the consistency rule in causal inference: axiom, definition, assumption, or theorem," *Epidemiology*, 21(6), 872–875. 103

Peterson, B., and Harrell Jr, F. E. (1990), "Partial proportional odds models for ordinal response variables," *Applied Statistics*, 39(2), 205–217. 6, 19

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. (2001), "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey Methodology*, 27(1), 85–96. 60

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003), "Multiple imputation for statistical disclosure limitation," *Journal of Official Statistics*, 19(1), 1–16. 2, 4

Rao, C. R. (2009), *Linear statistical inference and its applications*, Wiley Series in Probability and Statistics, 2 edn, John Wiley & Sons. 6

Rao, J. N. K., and Shao, J. (1992), "Jackknife variance estimation with survey data under hot deck imputation," *Biometrika*, 79(4), 811–822. 32

Reiter, J. P. (2008), "Multiple imputation when records used for imputation are not used or disseminated for analysis," *Biometrika*, 95(4), 933–946. 2

Robins, J. M. (2001), "Data, design, and background knowledge in etiologic inference," *Epidemiology*, 12(3), 313–320. 104

Robins, J. M., and Gill, R. D. (1997), "Non-response models for the analysis of non-monotone ignorable missing data," *Statistics in Medicine*, 16(1), 39–56. 61, 89, 111

Robins, J. M., Hernan, M. A., and Brumback, B. (2000), "Marginal structural models and causal inference in epidemiology," *Epidemiology*, 11(5), 550–560. 8, 106, 126

Robins, J. M., and Rotnitzky, A. (1995), "Semiparametric efficiency in multivariate regression models with missing data," *Journal of the American Statistical Association*, 90(429), 122–129. 3

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89(427), 846–866. 3, 13

Robins, J. M., and Wang, N. (2000), "Inference for imputation estimators," *Biometrika*, 87(1), 113–124. 4, 15, 26, 31

Rosenbaum, P. R., and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70(1), 41–55. 3, 13

Rotnitzky, A., and Robins, J. (1997), "Analysis of semi-parametric regression models with non-ignorable non-response," *Statistics in medicine*, 16, 81–102. 25

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998), "Semiparametric regression for repeated outcomes with nonignorable nonresponse," *Journal of the American Statistical Association*, 93(444), 1321–1339. 25

Royston, P. et al. (2009), "Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables," *Stata Journal*, 9(3), 466. 62

Rubin, D. B. (1978), "Multiple imputations in sample surveys-A phenomenological Bayesian approach to nonresponse," in *Proceedings of the Section on Survey Research Methods*. 3

Rubin, D. B. (1987), *Multiple imputation for nonresponse in surveys*, Wiley Series in Probability and Statistics, Wiley. 2, 3, 4, 15

Rubin, D. B. (1996), "Multiple imputation after 18+ years," *Journal of the American Statistical Association*, 91(434), 473–489. 2, 4

Schafer, J. L. (1997), *Analysis of incomplete multivariate data*, Chapman and Hall/CRC Monographs on Statistics and Applied Probability, Chapman and Hall/CRC. 60

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), "Adjusting for nonignorable drop-out using semiparametric nonresponse models," *Journal of the American Statistical Association*, 94(448), 1096–1120. 14, 96

Seaman, S. R., and White, I. R. (2013), "Review of inverse probability weighting for dealing with missing data," *Statistical Methods in Medical Research*, 22(3), 278–295. 3

Serfling, R. J. (1980), *Approximation theorems of mathematical statistics*, Wiley Series in Probability and Statistics, John Wiley & Sons. 58

Shao, J., and Steel, P. (1999), "Variance estimation for survey data with composite imputation and nonnegligible sampling fractions," *Journal of the American Statistical Association*, 94(445), 254–265. 124

Simon, G. A. (1978), "Efficacies of measures of association for ordinal contingency tables," *Journal of the American Statistical Association*, 73(363), 545–551. 17

Somers, R. H. (1962), "A new asymmetric measure of association for ordinal variables," *American Sociological Review*, 27(6), 799–811. 6, 17

Stürmer, T., Schneeweiss, S., Avorn, J., and Glynn, R. J. (2005), "Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration," *American journal of epidemiology*, 162(3), 279–289. 104

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

Tang, C. Y., and Qin, Y. (2012), "An efficient empirical likelihood approach for estimating equations with missing data," *Biometrika*, 99(4), 1001–1007. 3

Touloumis, A., Agresti, A., and Kateri, M. (2013), "GEE for multinomial responses using a local odds ratios parameterization," *Biometrics*, 69(3), 633–640. 7

Tripathi, G. (1999), "A matrix extension of the Cauchy-Schwarz inequality," *Economics Letters*, 63(1), 1–3. 40

Tsiatis, A. (2006), *Semiparametric theory and missing data*, Springer Series in Statistics, Springer. 10, 14

Tutz, G. (1991), "Sequential models in categorical regression," *Computational Statistics and Data Analysis*, 11(3), 275–295. 6

Tutz, G., and Binder, H. (2004), "Flexible modelling of discrete failure time including time-varying smooth effects," *Statistics in Medicine*, 23(15), 2445–2461. 7

Van Buuren, S., Boshuizen, H. C., Knook, D. L. et al. (1999), "Multiple imputation of missing blood pressure covariates in survival analysis," *Statistics in Medicine*, 18(6), 681–694. 4

VanderWeele, T. J., and Arah, O. A. (2011), "Unmeasured confounding for general outcomes, treatments, and confounders: Bias formulas for sensitivity analysis," *Epidemiology*, 22(1), 42. 104

Wang, N., and Robins, J. M. (1998), "Large-sample theory for parametric multiple imputation procedures," *Biometrika*, 85(4), 935–948. 4, 15

Westreich, D., Edwards, J. K., Cole, S. R., Platt, R. W., Mumford, S. L., and Schisterman, E. F. (2015), "Imputation approaches for potential outcomes in causal inference," *International Journal of Epidemiology*, 44(5), 1731–1737. 104

White, I. R., Royston, P., and Wood, A. M. (2011), "Multiple imputation using chained equations: issues and guidance for practice," *Statistics in Medicine*, 30(4), 377–399. 60, 61

Wu, C. F. J. (1983), "On the convergence properties of the EM algorithm," *The Annals of Statistics*, 11(1), 95–103. 74, 87

Yang, S., and Kim, J. K. (2016*a*), "Fractional imputation in survey sampling: a comparative review," *Statistical Science*, 31(3), 415–432. 5

Yang, S., and Kim, J. K. (2016*b*), "A note on multiple imputation for method of moments estimation," *Biometrika*, 103(1), 244–251. 4

Zhao, J., Cook, R. J., and Wu, C. (2015), "Multiple imputation for the analysis of incomplete compound variables," *Canadian Journal of Statistics*, 43(2), 240–264. 4, 5