Analysis of Survey Data

With Non-Ignorable Missing covariates

BY

FIMA LANRA FREDRIK GERARLD LANGI
M.D., Universitas Sam Ratulangi, Manado, Indonesia, 1999
M.Med.Stats., University of Newcastle, Newcastle, Australia, 2006

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Public Health Sciences (Biostatistics)
in the Graduate College of the
University of Illinois at Chicago, 2017

Chicago, Illinois

Defense Committee:

Hua Yun Chen, Chair and Advisor
Sally Freels, Epidemiology and Biostatistics
Hakan Demirtas, Epidemiology and Biostatistics
Li Liu, Epidemiology and Biostatistics
Timothy Johnson, Survey Research Laboratory

I thank Lord GOD the Almighty,

through His Son, Jesus Christ, My Beautiful Savior,

for this wonderful life and journey

This thesis is dedicated to:

Ellen, my lovely wife

Natasha and Matthew, my blessed children

Mega, my comforting mother

In Loving Memory of:

Fhillie Bert Langi (December 27, 1947 - August 20, 2007), my beloved father

ACKNOWLEDGMENTS

I would like to express my gratitude to the thesis committee: Prof. Dr. Hua Yun Chen (chairman and thesis advisor), Prof. Dr. Sally Freels, Prof. Dr. Hakan Demirtas, Prof. Dr. Li Liu, and Prof. Dr. Timothy Johnson; thank you for all the guidance and the unwavering support that had made this journey possible. Prof. Dr. Hua Yun Chen was my academic advisor as well, and had guided me over many years I was in the Ph.D. program. I am greatly indebted to Mr. George E. Manning II, Prof. (Emeritus) Dr. Edwin Chen, and Prof. Dr. Fabricio E. Balcazar for the generous help, assistance, wisdom, and encouragement throughout the whole stages of my thesis writing. Mr. George E. Manning II's helps were instrumental for my "survival" during the tumultuous times I encountered while still trying to finish the Ph.D. program. It is also important to acknowledge the U.S. Department of State, the government of the Republic of Indonesia, and the Institute of International Agency (IIE), which had enabled me to undertake this Ph.D. program with the provision of the Fulbright Presidential scholarship.

FLFG

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASE	Asymptotic standard error
CR	(Augmentation with) the constants removed
EM	Expectation-maximization (algorithm)
GLM	Generalized linear model
IDHS	Indonesia Demographic and Health Survey
MAE	Median absolute error
MAR	Missing at random
MCAR	Missing completely at random
MICE	Multiple imputation by chained equations
MDGs	Millennium Development Goals
MLE	Maximum likelihood estimate
MNAR	Missing not at random
MSE	Mean squared error
NB	Negative-binomial (distribution)
URTI	Upper respiratory tract infection
ZIP	Zero-inflated Poisson.

SUMMARY

Missing data are common in survey sampling. Such loss creates a spectrum of inferential problems, depending on the type of missingness. In this thesis, a method to analyze survey data with potentially non-ignorable covariates is proposed. The approach is particularly developed to address the limitations in current routines of the standard statistical packages when, simultaneously, the model of interest has a mixture of categorical and continuous missing covariates, the analysis needs to incorporate the sampling design under different assumptions about its functional form, and there is a demand for manageable computation time in practical sense.

Essentially, the proposed method is a modification of the Expectation-Maximization (EM) algorithm, particularly the E-step, where the missing elements of data are first augmented, before the estimation continues using the conditional probability of the missing variables given the observed data as the weight. The algorithm proceeds as a full likelihood procedure if the sampling probability function is known for all observations, but it becomes a quasi-likelihood approach when the quantity of survey weight is instead the only available information about sample selection.

There are three classes of survey data considered during the development, which include those of which none (Case 1), all (Case 2), or some (Case 3) of the covariates are observable outside the samples. Two situations are further defined on each of them, that is, whether the functional form of sample selection is known (Situation 1) or unknown (Situation 2). Given its construction, the proposed method, termed the augmentation assisted EM algorithm or

simply the augmentation method, retains the desirable properties of the maximum likelihood estimates, while flexible enough to handle both continuous and categorical missing covariates, and can adapt the use of survey weight to improve inference.

The simulation studies indicates that the proposed method performs reliably well across all classes of survey data. In terms of unbiasedness, it is competitive with and may occasionally outperform the multiple imputation by chained equations (MICE), a well-known technique in multiple imputation. Efficiency of its estimates are also comparable to MICE. In the real data application using the dataset from the Indonesia Demographic and Health Survey of 2012, the proposed method successfully estimates the demographic, health, and birth-related factors associated with the infant mortality. Most importantly, it is able to improve the results of complete case analyses by both correcting the magnitude of effect size and increasing the power of analysis to detect the variable significance.

CHAPTER 1

INTRODUCTION

Data with missing values are common in surveys and observational studies. The loss may occur in various stages. For instance, the data may not be collected for the selected households because they are in geographically challenging areas, or the residents are not at home, have moved and cannot be located. At individual level, information of certain variables may be missing because the eligible person is unable, unavailable, or refuse to respond. In the surveys with interviewer-administered questionnaire, data loss can be experienced when the interviewer mistakenly passes the question or fail to record the answer. It is also possible that the data were collected but are then lost due to error during data entry. Incompleteness of information can also be a natural consequence of the instrument or research design. Many studies use questionnaires where not all the questions or items in it are applicable to the subjects of interest. The sampling selection itself may be regarded as a part of design that causes missing data.

Missing data create a spectrum of inferential problems. They could be limited to a compromised precision or power of the study, which might not be serious, particularly when it has been anticipated in the design. However, if the missing fraction is substantial and not properly anticipated, then the effects on inference may no longer be trivial. In fact, even with a relatively small number of incomplete observations there is already a risk of inaccurate or biased estimates when the mechanism of which the values missing of a variable is related to the variable itself.

The implication of this type of missing data is that there can be systematic differences between the subjects with and without missing values.

Ideally, variables with missing values not by study design are remedied through retrieval of the information from the original sources or subjects. This is of course impractical in most settings. Furthermore, the data analysts may not be part of the former research team and do not have access to the data managers, let alone the study subjects. Such situation has prompted the development of techniques for dealing with incomplete data. Complete-case analysis is the most traditional approach, which remains as an option for modeling missing data because of its availability in all standard statistical software, and the fact that it produces unbiased parameter estimates for the class of missing-data mechanism called missing completely at random (MCAR)(1). The more advanced strategies are generally fall in one of the following: multiple imputation, weighted estimating equations, Bayesian, and maximum likelihood methods(2; 3). Earlier works of these methodologies are polarized on data with missing variables at random (MAR). Their use in data with missing variables not at random (MNAR), also termed non-ignorable missingness, remains an active research area in the missing data literature.

Development of the proposed method in this dissertation is motivated by the data from the Indonesia Demographic and Health Survey (IDHS) of 2012. This survey is a complex multistage probability sampling study designed to provide a number of demographic statistics related to population health. These include fertility and childhood mortality rates, level of contraceptive knowledge and practice, key child health indicators (level of immunization, prevalence and treatment of diarrhea and several childhood diseases, and child feeding practices), coverage

of maternity care services, men's involvement in reproductive health, data on awareness of AIDS/STI, and determinants of maternal and child health (4). Sample sizes were calculated with the objective of providing reliable estimates at national, provincial, and urban/rural levels. The primary sampling units were the census blocks formed during the 2010 population census. At the first stage of sampling, the census blocks were stratified into provinces and urban/rural areas. Allocation for each stratum was not proportional to the population size; a minimum of 43 census blocks was imposed for individual province. There were in total 1,840 census blocks allocated for the 33 provinces: 874 in urban areas and 966 in rural areas. Complete listing and mapping of households was then conducted in the selected census blocks. The second stage sampling comprises a systematic selection of the households from each block, for a total of 21,850 households in the urban areas and 24,150 in the rural areas, or an average of 25 households per census block survey wide. All women age 15-49 and never-married men age 15-24 on the selected households were eligible for interview. In addition, eight of the 25 households were subsequently sampled in a systematic fashion to find currently married men age 15-49 eligible for interview. The survey used separate questionnaires for household, woman's, currently married man's, and never-married man's interviews. Overall, the IDHS 2012 had a very high response rate. Of 44,302 households which occupied a house (there were originally 46,024 households selected, but not all houses occupied), 43,852 (99%) were able to complete the interview. Within the interviewed households, 45,607 of 47,533 eligible women (96%) completed their interview. The response rate for the interview of currently married men was also high at 92%. In general, the rural areas had a higher response rate than the urban areas.

Childhood mortality data were collected through the complete birth history of live births among the eligible women in the selected households. Of 45,607 women who provided data during the study, 32,129 had ever given birth (15,262 had the delivery within the five years before the survey). The information asked includes child's birth order, gender, birth year and month, survival status, and if relevant, age at death (in days for children died in the first month of life, in months for those died before the second birthday, and in years for children died at later ages). For live births in the preceding five years, the study recorded the data on maternity: previous and succeeding birth intervals, antenatal care, any complication during pregnancy and at birth, delivery attendant, child's birth size and weight, place of birth, maternal age when the child was born, postnatal care, and breastfeeding. Further investigation was conducted on children who were born in the last five years and still alive at the time of survey to obtain health histories, such as the immunization status, the presence and treatment of major childhood illnesses (diarrhea, acute respiratory infection, and fever) in the last two weeks, and child's nutrition. There are 83,650 live births observed in the 2012 IDHS, where 18,021 of them preceded the survey by five years or less. The data are accessible in the public domain www.dhsprogram.com. This website, in particular, provides separate datasets for all live births and those that took place in five years before the survey. They are respectively named the birth and children datasets. The variable names of these two sets of data are identical. Nevertheless, the birth dataset in general does not contain the information related to maternity and child's health histories. IDHS provides the sampling weight for each sample in the datasets.

TABLE I
MISSING OBSERVATIONS IN SELECTED VARIABLES OF THE INDONESIA DEMOGRAPHIC AND HEALTH SURVEY (IDHS) OF 2012 DATASETS

Variable	n_{obs}	n_{mis} (%)
Household Level (n = 43,852)		
Major source of drinking water	43,830	22 (0.1)
Type of toilet facility in the household	43,821	31 (0.1)
Frequency household members smoke inside the house	43,821	31 (0.1)
Any member has electricity	43,793	59 (0.1)
Any member has a car	43,769	83 (0.2)
Time taken to the source for drinking water	43,320	532 (1.2)
Presence of water at hand washing place	33,434	10,418 (23.8)
Location of source for water	24,678	19,174 (43.7)
Woman Level (n = 45,607)		
Ever had a terminated pregnancy	45,602	5 (0.0)
Age of the respondent at first birth	32,129	13,478 (29.6)
Ever had complications during pregnancy	15,206	30,401 (66.7)
Birth in the Preceding 10 Years (n = 36,484)		
Mother has any postpartum problem	36,463	21 (0.1)
Preceding birth interval	23,909	12,575 (34.5)
Birth in the Preceding 5 Years (n = 18,021)		
Birth assistance: health professional	17,886	135 (0.7)
Place of delivery	17,855	166 (0.9)
Delivery by Caesarean section	17,847	174 (1.0)
Size of child as reported subjectively by the mother	17,293	728 (4.0)
Problem at time of birth: prolonged labor	15,227	2,794 (15.5)
Problem at time of birth: vaginal bleeding	15,210	2,811 (15.6)
Problem at time of birth: convulsion	15,209	2,812 (15.6)
Prenatal care: health professional	15,203	2,818 (15.6)
Problem at time of birth: fever and foul smelling vaginal discharge	15,183	2,838 (15.7)
Problem at time of birth: water broke > 6 hrs before delivery	15,167	2,854 (15.8)
Birth weight (kg)	15,124	2,897 (16.1)
Total antenatal visits during pregnancy	15,121	2,900 (16.1)

Continued on next page

TABLE I (Continued)
MISSING OBSERVATIONS IN SELECTED VARIABLES OF THE INDONESIA DEMOGRAPHIC AND HEALTH SURVEY (IDHS) OF 2012 DATASETS

Variable	n_{obs}	n_{mis} (%)
Postnatal check within 2 months	15,075	2,946 (16.3)
Timing of first antenatal check (month)	14,475	3,546 (19.7)
Place baby first checked	9,762	8,259 (45.8)
Postnatal attendant	9,723	8,298 (46.0)
Timing of first postnatal check	9,438	8,583 (47.6)
Birth in the Preceding 5 Years and Alive (n = 17,367)		
Received POLIO 1	17,288	79 (0.5)
Received BCG	17,285	82 (0.5)
Received MEASLES	17,243	124 (0.7)
Had cough in last 2 weeks	17,230	137 (0.8)
Received DPT 1	17,221	146 (0.8)
Had fever in last 2 weeks	17,211	156 (0.9)
Had diarrhea in last 2 weeks	17,207	160 (0.9)
Ever had vaccination	12,166	5,201 (29.9)
Had cough with short, rapid breaths in last 2 weeks	5,935	11,432 (65.8)

Survival outcome of a child is fully recorded in the IDHS 2012 data. A total of 654 (3.6 percent) children reportedly died. There are a number of variables in the IDHS 2012 data that may be related to the survival or death of the children, and thus are potential covariates for regression models. They are a mixture of categorical and continuous variables, and include those about birth history, child's health, and household characteristics. Some of these potential variables, however, are subject to missing values with a varying degree. Table I shows a few of them based on their level or nature of data collection. The missingness is broad spectrum, from much less than 1% to more than 60%. Figure 1 presents the pattern of the missing variables at household level. Not surprisingly, the patterns are arbitrary. It is fair to consider that the missingness of these potential covariates might not be ignorable. Intuitively, for instance, a missing response on the child's vaccination status could be determined by some of the household characteristics. Information on postnatal care, as another example, may also be lost due to some other variables regarding birth histories, maternal education, and household characteristics. These are just educated guess until we undertake a formal investigation. They, however, emphasize the need to not only deal with the missing variables when analyzing the survival or death status of children using a survey dataset as that of the IDHS 2012, but also to take into account the possibility that such missingness is not ignorable.

Another problem in survey data analysis, particularly for the datasets of large scale or national surveys, is in accommodating the sampling design. It is very common that such surveys are implemented through a complex survey scheme, which is exactly the case of the IDHS 2012. If the functional form of sample selection is known and such information is available for all

target individuals, then the solution is quite straightforward; the analysis may be accomplished in a traditional model based fashion. This is, of course, difficult to expect in a real data scenario. However, many surveys provide the sample weight of all samples. As a consequence, the standard statistical software packages such as Stata, SAS, and R are increasingly equipped themselves with routines for survey analysis that allow the application of sample weight into the models. The area where they appear to require further improvement is in the analysis of survey data of which the model of interest is hampered by non-ignorable missing variables, particularly when they are a mixture of categorical and continuous variables.

The aforementioned backgrounds provide the rationale for this thesis research. In particular, I am interested in developing a method that may address the lack of statistical routines in the standard packages for handling situations where the model of interest has a mixture of categorical and continuous missing covariates, and the analysis needs to incorporate the sampling design under different assumptions about its functional form. An important consideration is given to an approach that is computationally manageable for practical application. The organization of this thesis is as follows. Chapter 2 reviews missing data literature and the approaches for modeling data from surveys. The idea is to present the variety of existing methods, the problems they addressed, and the areas yet explored and thus eligible for further improvement or refinement. A brief discussion about the models for count data closes Chapter 2. Development of the proposed methods is the core of Chapter 3. Three classes of survey data are defined based on whether the missing covariates are all, partially, or not at all observable among the non-samples. Implementation of the proposed method on each class of survey data is simulated

in Chapter 4. Chapter 5 then demonstrates its application on a real dataset. Finally, Chapter 6 concludes this thesis.

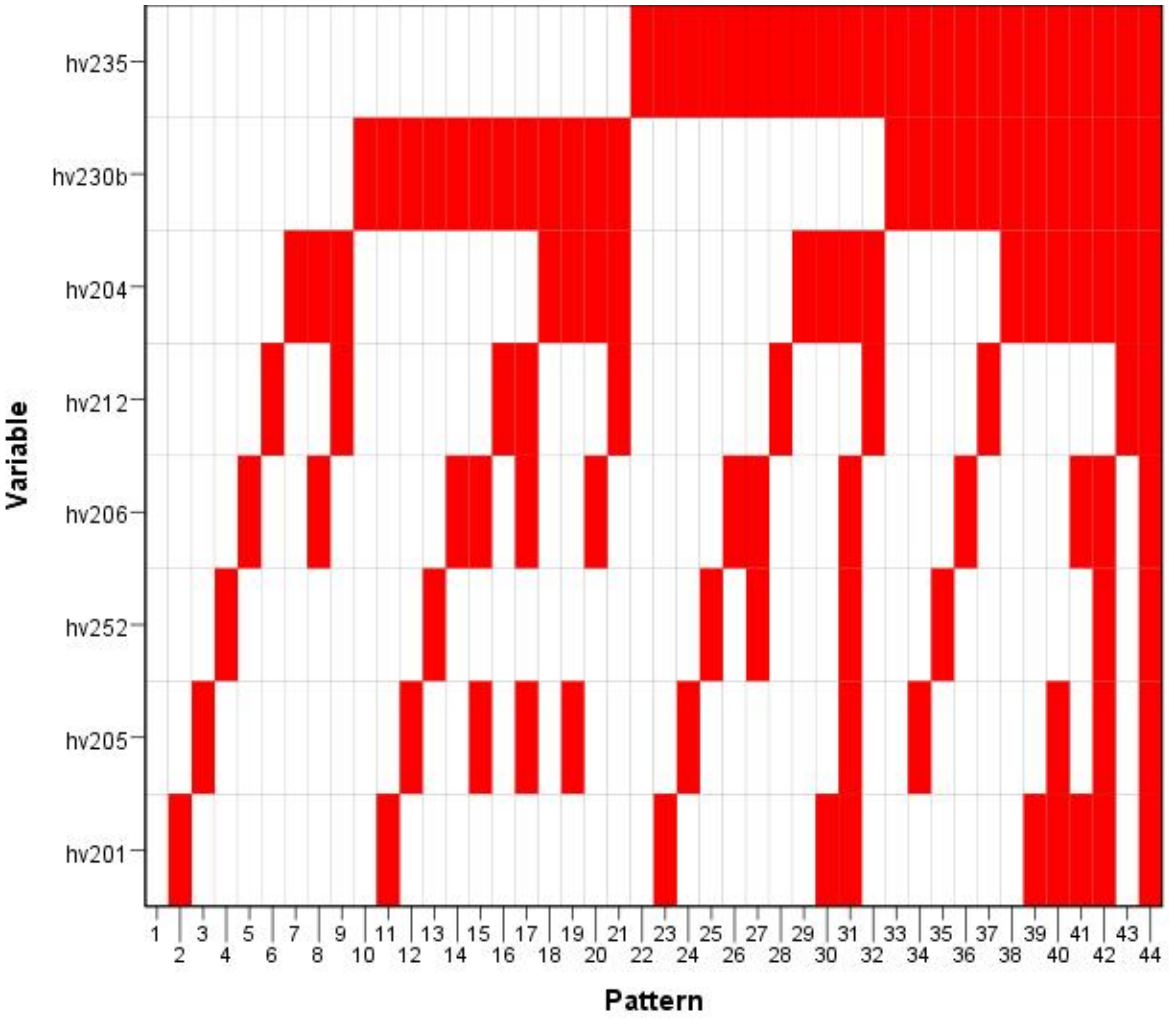

Figure 1. Patterns of Missing Variables in the Household Level Data, the IDHS of 2012 (n = 43,852)

CHAPTER 2

LITERATURE REVIEW

We are often interested with the relationship of a vector of a random variable $\mathbf{Y}$ with a matrix of covariates $\mathbf{X}$ in the population. Suppose that in the population

$$\mathbf{Y}, \mathbf{X} \sim f(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\beta}) f(\mathbf{x}; \boldsymbol{\alpha})$$

holds, where the lower cases indicate the realization of the upper cases, and $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are vectors of model parameters. Clearly, $\boldsymbol{\beta}$ are the parameters of interest and $\boldsymbol{\alpha}$ are nuisance parameters. We thus consider a population $\mathscr{F}$ with N subjects. It may not always be feasible to census $\mathscr{F}$. Instead, we take a sample of size n and perform the estimation of $(Y_i, \mathbf{X}_i)$ relationship, $i = 1, \cdots, N$. Let us denote the event where subject i is selected during the sampling as $I_i = 1$, and otherwise $I_i = 0$. Accordingly, $n = \sum_{i=1}^{N} I_i$. In addition, let $R_{ik} = 1$; $k = 1, \cdots, K$ indicates that the covariate X_k is observed on subject i, and $R_{ik} = 0$ indicates X_{ik} is missing. This implies $\mathbf{X}_i = (\mathbf{X}_{i,obs}, \mathbf{X}_{i,mis})$, where $\mathbf{X}_{i,obs}$ represent the observed part of $\mathbf{X}_i$, and $\mathbf{X}_{i,mis}$ the missing part. Throughout this chapter bold letters such as $\mathbf{Y}$ and $\mathbf{X}$ denote a vector or a matrix, depending on the context. Unless otherwise stated, subjects are considered independent and Y_i is always observed within the samples.

There are two technical parts that have to be dealt with in the analysis of survey data with missing covariates. The first is how to handle the incomplete observations, which come not

only from the covariates with missing values but also the outcome as a consequence of survey selection. The second part is about incorporation of the sampling mechanism into parameters estimation. Various approaches on each of the issues have appeared in statistics literature. I will review a few of them in the following sections. To facilitate the discussion, I fix the assumption of sampling mechanism to be a simple random when I address the statistical methods for missing data. This assumption is changed when the presentation switches to the role of sampling weights in the analysis of survey data. For the purpose of continuation, however, I maintain the notation I develop throughout the chapter even as I refer to the reviewed studies. Accordingly, the presented expressions are in general not identical to the original papers, but are so in spirit.

The rest of this chapter is organized as follows. Section 1 contains the discussion about estimation methods for data with incomplete observations from a simple random sampling. A concise introduction of missing data classification will open this section. The methods that will be reviewed include complete-case analysis, multiple imputation, Bayesian methods, weighted estimating equations, and maximum likelihood procedures. Assumption about the sample selection is then relaxed in Section 2, as I discuss the importance of sampling weights in survey data. I also present the settings where sampling probability is ignorable or informative, followed by the review of approaches for incorporating sampling weights in the inference process. In Section 3, I present the models for count data. Information of infant mortality in survey data is often available as counts, thus explains the need of this section.

2.1 Taxonomy of Missing Data Mechanisms

It is important to classify the types of missing-data mechanism, because different mechanisms require different methods to handle. Rubin (1) and Little and Rubin (5) described three different mechanisms of missing data based on the conditional distribution of $\mathbf{R}_i$ given Y_i and $\mathbf{X}_i$. They are: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Data are MCAR if the missing data probability $\Pr\{\mathbf{r}_i \mid y_i, \mathbf{x}_i\} = \Pr\{\mathbf{r}_i\}$ for all y_i and all $\mathbf{x}_i$. That is, MCAR is the case where the distribution of $\mathbf{R}_i$ does not depend on the values of either Y_i or $\mathbf{X}_i$; the observed data are simply a random sample of all the data. On the other hand, data are MAR if $\Pr\{\mathbf{r}_i \mid y_i, \mathbf{x}_i\} = \Pr\{\mathbf{r}_i \mid y_i, \mathbf{x}_{i,\text{obs}}\}$ for all $\mathbf{x}_{i,\text{mis}}$. One may note that MCAR is a special case of MAR, just like the situations where $\Pr\{\mathbf{r}_i \mid y_i, \mathbf{x}_i\} = \Pr\{\mathbf{r}_i \mid y_i\}$ for all $\mathbf{x}_i$, or when $\Pr\{\mathbf{r}_i \mid y_i, \mathbf{x}_i\} = \Pr\{\mathbf{r}_i \mid \mathbf{x}_{i,\text{obs}}\}$ for all y_i and $\mathbf{x}_{i,\text{mis}}$. Lastly, data are regarded MNAR if the distribution of $\mathbf{R}_i$ depend on $\{\mathbf{D}_i : \mathbf{D}_i \subseteq (Y_i, \mathbf{X}_i), \mathbf{X}_{i,\text{mis}} \in \mathbf{D}_i\}$.

Another common classification of data with incomplete observations is based on whether the law governing missingness can be ignored or not in statistical inference. Using the chapter example, if data are MCAR or MAR then likelihood-based inference of the parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ of the joint distribution $(Y_i, \mathbf{X}_i)$ does not require incorporation of the model for $\mathbf{R}_i$. The asymptotic theory guarantees that the parameter estimates will still be unbiased, provided that certain regularity conditions are met. This advantage is unfortunately not shared by data with MNAR. Any analysis involving data with this type of missingness has to account for the missing-data mechanism in order to avoid bias. Hence, MNAR is non-ignorable missing data. This thesis

uses the terms MNAR and non-ignorable missing data interchangeably to refer to the cases where the model for missingness has to be incorporated in parameters estimation.

Most but not all MCAR can be tested. However, MAR and MNAR are in general hardly verifiable. The treatment of missing data in practice is thus a delicate balance between science and art. Historically, MCAR is an attractive assumption as it allows the use of simpler approaches for statistical inferences, in particular complete-case analysis. Yet with the accumulated knowledge of missing data techniques and ever-increasing computing power nowadays, there is a tendency in the statistical communities to use it as the last instead of first resort. On the other hand, MAR is a more realistic assumption than MCAR. Its use is greatly facilitated by the common practice of major statistical software packages to set MAR as the default setting for missing-data procedures. The safest assumption is of course MNAR. However, its implementation is plagued by the issue of parameters identifiability.

Little (3), Glynn and Laird (6), Little and Rubin (5), Ibrahim, Chen, Lipsitz and Herring (2), Chen, Ibrahim, Chen, and Senchaudhuri (7), and White and Carlin (8) discusses the use of complete-case analysis in data with missing covariates. In this approach, only subjects with complete information on Y_i and all $\mathbf{X}_i$ are used in the modeling. Those with any missing value are discarded. It is a primitive method for the purpose, yet has the appeals of availability in any standard statistical package, simplicity of implementation, and most importantly, ability to produce valid inference if the missingness is MCAR, or MAR where the missing-data mechanism

depends only on the regressors. Following the examples from Chen et al. (7), let us write the complete-case analysis as a conditional inference given $\mathbf{R}_i = \mathbf{1} = (1, 1, \cdots, 1)'$. For MCAR,

$$\begin{aligned} f(y_i, x_i \mid r_i = 1; \alpha, \beta, \gamma) &= \frac{f(r_i = 1; \alpha) f(y_i \mid x_i; \beta) f(x_i; \gamma)}{f(r_i = 1; \alpha) \int\int f(y_i \mid x_i; \beta) f(x_i; \gamma) dy_i dx_i} \\ &= f(y_i \mid x_i; \beta) f(x_i; \gamma), \end{aligned}$$

provided that α, β, and γ are distinct. When data are MAR and the missingness mechanism depends only on $\mathbf{X}$,

$$\begin{aligned} f(y_i, x_i \mid r_i = 1; \alpha, \beta, \gamma) &= \frac{f(r_i = 1 \mid x_i; \alpha) f(y_i \mid x_i; \beta) f(x_i; \gamma)}{\int f(r_i = 1 \mid x_i; \alpha) \int f(y_i \mid x_i; \beta) f(x_i; \gamma) dy_i dx_i} \\ &= f(y_i \mid x_i; \beta) \frac{f(r_i = 1 \mid x_i; \alpha) f(x_i; \gamma)}{\int f(r_i = 1 \mid x_i; \alpha) f(x_i; \gamma) dx_i}, \end{aligned} \tag{2.1}$$

which implies that the estimates for β are unbiased or asymptotically consistent under certain regularity conditions. This property of complete-case analysis is not shared by any other method, thus reserves its importance in missing data analysis.

Complete-case analysis, however, is a biased procedure for other cases of MAR and for MNAR. In addition, it wastes the subjects with incomplete information that may lead to a substantial reduction in power. This risk increases as the number of X's becomes greater, since even a sparse pattern of missing X's can result in a number of incomplete cases that is not trivial. Little (3) suggested that the way to avoid such a risk is either dropping X's with large proportion of missing values or opting for other methods that incorporates subjects with incomplete information. Despite its limitations, complete-case analysis is frequently used as a valuable benchmark for other missing data procedures.

2.2 Analysis of Data with Missing Covariates

2.2.1 Multiple Imputation

Rubin (1; 9; 10; 11) first proposed multiple imputations in the late 1970s, and ever since then, it has become the most popular technique for dealing with missing data. There is a plethora of literature discussing its underlying theory and general examples, including those of Rubin and Schenker (12) and Rubin (1; 9; 11). Implementation of multiple imputation for data with missing covariates was explored in the studies by Little (3), Van Buuren and Groothuis-Oudshoorn (13), Raghunatan, Lepkowski, Van Hoewyk, and Solenberger (14), Rubin (11), Ibrahim, Chen, Lipsitz, and Herring (2), and by Little and Rubin (5). The idea of multiple imputation is to create $M > 1$ "complete" datasets by imputing the missing data, then analyze each $m = 1, \cdots, M$ dataset as if they were complete data. The estimates $\hat{\theta}' = (\hat{\alpha}', \hat{\beta}', \hat{\gamma}')$ are simply $\frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$. Rubin(1; 9; 11) showed that the variance estimate

$$\hat{V}[\hat{\theta}] = \hat{V}_W + \left(1 + \frac{1}{M}\right)\hat{V}_B,$$

where $\hat{V}_W = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_m[\hat{\theta}]$ and $\hat{V}_B = \sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta})(\hat{\theta}_m - \hat{\theta})'$. Large-sample inference for θ in multiple imputation is based on setting $(\hat{\theta} - \theta)(\hat{V}[\hat{\theta}])^{-1} \sim t$ with degrees of freedom $(M-1)(1 + \frac{M}{M+1}\hat{V}_W(\hat{V}_B)^{-1})$.

There are thus two models to deal with in multiple imputation: imputation model and analysis model. The latter is performed after each missing values has been filled. Rubin (11), Nielsen (15), and Little and Rubin (5) differentiated proper to improper imputation when

obtaining $\hat{\theta}$ on the m-th replication. They argued that improper imputation would result in biased or inconsistent estimators in large-sample inference. A typical example for improper imputation is the use of $\hat{\theta}$ from complete-case analysis. Proper imputation, in contrast, is based on Bayesian predictive distribution that integrates out parameters. It thus requires specification of priors for θ to get the posterior distribution $\varphi(\theta \mid y, x_{obs}, r)$. The predictive distribution of the missing values, assuming non-ignorable missingness, is

$$f(x_{i,mis} \mid y_i, x_{i,obs}, r_i) \propto \int f(x_{i,mis} \mid y_i, x_{i,obs}, r_i; \theta)\varphi(\theta \mid y, x_{obs}, r)d\theta. \tag{2.2}$$

The pathway to obtain $\varphi(\theta \mid y, x_{obs}, r)$ is presented in the next subsection about Bayesian methods. Of course it is clear, multiple imputation is a Bayesian approach in spirit. Rubin (11) and Little and Rubin (5) showed that proper imputation yields asymptotically valid inferences with the estimators have nice large-sample properties. At the downside, proper imputation in general requires intensive computation.

Various approaches for performing multiple imputation have appeared in the literature. Among them, the multiple imputation by chained equations (MICE) due to van Buuren and Groothuis-Oudshoorn (16; 17; 18; 19; 20) is one of the most adopted methods in recent years and it increasingly becomes the mainstream imputation approach. The general idea of MICE is to handle incomplete multivariate data by fully conditional specification (FCS) (19; 20). In particular, it iterates imputations over a series of conditional densities, one for each variable with missing values. After imputations completed, MICE proceeds through steps common to any

approach in this class of missing data procedure, which include simultaneous analysis of each imputed dataset using the same statistical method, and pooling of the results to estimate the quantities of scientific interest. Van Buuren and Groothuis-Oudshoorn (16) show that FCS has been implemented to certain extent in the previous studies (21; 20; 14; 22; 23; 24; 25; 19) using different names, such as stochastic relaxation, sequential regressions, ordered pseudo-Gibbs sampler, partially incompatible Monte Carlo Markov Chain, and iterated univariate imputation. The primary attraction of MICE is its easy application and tremendous flexibility, especially when compared to imputation by joint modeling (26; 27). However, MICE is vulnerable to incompatible model specification. Consequently, it is difficult to find the theory to support its use. One solution for this problem is proposed by Chen and colleagues (28).

Regardless of how it is performed, multiple imputation remains a powerful and extremely general method. It also has the wealth of literature in various applications, a luxury not always shared by the competing methods. Moreover, it is available in virtually every software package that carries features for missing data management. If the potential user or analyst is different from the person who would do imputation, such as that in database construction, then multiple imputation is almost the prescribed method as it allows end users to analyze the multiply imputed data using ordinary complete-data procedures. This by no means implies that one should use multiple imputation in all missing data problems. Like any statistical procedure, it can only be applied after some careful thoughts. Classical multiple imputation, for instance, may pose challenges when it comes to data with substantial collection of variables having incomplete observations. As listed by van Buuren and Groothuis-Oudshoorn (13), the challenges include the

selection of sensible imputation regressors particularly when the dataset is large, dealing with imputation cascade, treatment of variables with different measures (continuous and categories), accounting for the sampling design, handling optimal imputation model that may be non-linear, avoiding imputation-order hazard in data where the order might be meaningful, and addressing incompatibility issues between the imputation and analysis models. The traditional solution for multivariate data imputation is to assume a multivariate normal distribution (27). Schafer (26) then obtained the algorithm for multivariate continuous, categorical, and mixed data. Later, van Buuren and Groothuis-Oudshoorn (16; 17; 18; 19; 20) introduced MICE to address most of the problems they listed for multivariate data imputation. Yet the potential conflict between the imputation model and the analysis model continues to be a major obstacle (29). For MICE, the imputation models may themselves be incompatible. In addition, the variance of proper imputation is not consistent in conventional sense. To get full efficiency in multiple imputation, one has to impute and analyze an infinite replicates of datasets; something that will never be achievable. Another caveat is that the nature of imputation process makes this procedure produces different results for one dataset. Allison (30) also noted that when compared to maximum likelihood approach, multiple imputation entails far more decisions, which include the choice of iterative algorithm, imputation model for each incomplete variable, number of data replications, total iterations, prior distribution, incorporation of interactions and non-linearities, and the methods for multivariate testing. If computation is feasible, likelihood approach is clearly a better alternative than multiple imputation.

2.2.2 Bayesian Methods

Earlier application of Bayesian methods on missing data centered on incomplete dependent variables (31; 32; 9; 33; 34; 35). Its implementation in data with missing covariates was presented in Little (3), Ibrahim, Chen, Lipsitz, and Herring (2), Huang, Chen, and Ibrahim (36), and Mason, Richardson, Plewis, and Best (37). A 'full' Bayesian approach requires the specification of priors on all the parameters and the conditional distributions of missing covariates given the observed values and the parameters. Inference for the parameters then proceeds through sampling from their posterior distribution, for instance, with Gibbs sampler (38; 39; 40; 41; 2; 36). As a concrete example, let us note that by re-expressing θ in Equation 2.2 as $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ and assuming the model $f(\mathbf{r}_i \mid y_i, \mathbf{x}_i; \boldsymbol{\gamma})$ for $\mathbf{R}_i$, then $\vartheta(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{x}_{obs}, \mathbf{r})$ is proportional to

$$\left\{ \prod_{i=1}^{n} \int f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_i; \boldsymbol{\alpha}) f(\mathbf{r}_i \mid y_i, \mathbf{x}_i; \boldsymbol{\gamma}) d\mathbf{x}_{i,mis} \right\} \varphi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}).$$

One can set $f(\mathbf{x}_i; \boldsymbol{\alpha}) = f(\mathbf{x}_{i,mis} \mid \mathbf{x}_{i,obs}; \boldsymbol{\alpha})$ such as that of Ibrahim, Chen, Lipsitz, and Herring (2), or keep it as a 'prior' covariates model as discussed in Mason, Richardson, Plewis, and Best (37). The joint prior distribution $\varphi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ can be proper or improper, but the latter should be avoided if feasible. Once the prior and covariate models are set, a sample from the posterior distribution via Gibbs sampler is obtained by drawing the samples consecutively from $f(\mathbf{x}_{i,mis} \mid y_i, \mathbf{x}_{i,obs}, \mathbf{r}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, $\varphi(\boldsymbol{\alpha} \mid, y_i, \mathbf{x}_i, \mathbf{r}_i; \boldsymbol{\beta}, \boldsymbol{\gamma})$, $\varphi(\boldsymbol{\beta} \mid y_i, \mathbf{x}_i, \mathbf{r}_i; \boldsymbol{\alpha}, \boldsymbol{\gamma})$, and $\varphi(\boldsymbol{\gamma} \mid y_i, \mathbf{x}_i, \mathbf{r}_i; \boldsymbol{\alpha}, \boldsymbol{\beta})$. Hence, compared to the non-missing data situations that involve sampling from the posterior distribution of each parameter, Bayesian methods with missing covariates

using Gibbs sampler require only an additional step of sampling from the predictive distribution of the missing values.

Bayesian methods are powerful and general tools for dealing with missing covariates. Any additional problem, for instance, non-ignorable instead of ignorable missingness mechanism, can be accommodated without the need of new techniques for inference. However, implementation of this class of missing-data methods requires determination of priors, which in certain cases is far from straightforward. The trick is to find a prior distribution that is computationally convenient while keeping the results not radically change with a different set of priors. Huang, Chen, and Ibrahim (42) showed that in certain settings of non-ignorable missing covariates, improper uniform priors may lead to a joint posterior that will always be improper. Ibrahim, Chen, Lipsitz, and Herring (2) suggested informative prior using historical data for MAR covariates in generalized linear models (GLMs). Such information may not be available in surveys. Another potential problem in practice may occur when one deals with large data having substantial missing covariates while the marginal posterior distribution of parameters has a non-explicit expression (3). In this case, complexity of the likelihood function requires approximation either by numerical integration or stochastic simulation (such as Monte Carlo Markov Chain or MCMC sampling), and it brings the same issue of ensuring that the computation is manageable.

2.2.3 Weighted Estimating Equations

A class of semiparametric approaches for handling missing data based on inverse-probability weighted estimating equations was proposed by Robins, Rotnitzky, and Zhao (43; 44). The methods were generalized by Robins and Ritov (45). Robins and Rotnitzky (46), Bang and

Robins (47), Lunceford and Davidian (48), Kang and Schafer (49), and Chen and Zhou (50) then developed 'doubly-robust' weighted estimating equations for missing-data analysis, which only required either the missingness mechanism or the score vector of the response model to be correctly specified. Robins, Rotnitzky, and Zhao (51), Rotnitzky and Robins (52), Robins and Rotnitzky (44; 46), Zhao, Lipsitz, and Lew (53), Rotnitzky, Robins, and Scharfstein (54), Lipsitz, Ibrahim, and Zao (55), Scharfstein, Rotnitzky, and Robins (56), Robins, Rotnitzky, and Scharfstein (57), Parzen, Lipsitz, Ibrahim, and Lipshultz (58), Scharfstein and Irizarry (59), Herring and Ibrahim (60), and Seaman and White (61) explored the use of the weighted estimating equations for missing covariates.

In principle, weighted estimating equations use the inverse of the probability of missingness to weight the score vector of the model relating the response Y to explanatory variables $\mathbf{X}$. Let us assume here that there is only one missing covariate pattern in the data. Hence, the missing indicator becomes a scalar R_i. Its realization $r_i = 1$ means $\mathbf{x}_i$ are fully observed, and $r_i = 0$ refers to the situation where the elements of $\mathbf{x}_i$ with potential missing values are all not observed. A natural choice of the distribution of R_i, letting it to be dependent on both Y_i and $\mathbf{X}_i$, is Bernoulli with $\pi_i = f(r_i = 1 \mid y_i, \mathbf{x}_i; \boldsymbol{\gamma})$. To obtain $\hat{\boldsymbol{\beta}}$ in weighted estimating equations is to solve $\mathbf{U}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, where assuming π_i is known, the simplest form of these methods is (43)

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{r_i}{\pi_i} \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) \right\}$$

based on the previous specification of the conditional distribution of Y_i. For generalized linear models (GLMs) (62; 2), the term inside the curly brackets is equal to $\mathbf{d}w_i^{-1}(y_i - \mu_i)$, where $\mathbf{d} = \partial\mu_i/\partial\boldsymbol{\beta}$, $\mu_i = E[y_i \mid \mathbf{x}_i; \boldsymbol{\beta}]$, and $w_i = V[y_i \mid \mathbf{x}_i; \boldsymbol{\beta}]$. The expression for doubly-robust weighted estimating equations have several alternatives (49), including in particular (43; 51; 52)

$$\mathbf{U}_{DR}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\frac{r_i}{\pi_i} \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) \right\} + (1 - \frac{r_i}{\pi_i}) \left\{ \mathscr{E} \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) \Big| y_i, \mathbf{x}_{i,obs}; \boldsymbol{\alpha}, \boldsymbol{\beta} \right] \right\} \right],$$

where the expectation here is with respect to the missing covariates given the observed data and the parameters. Recall that we assume in this paragraph only one pattern of missing data (r_i is a scalar). A consistent variance estimate of $\widehat{\boldsymbol{\beta}}$ in weighted estimating equations is normally obtained through a robust sandwich estimator, for instance, as shown in Ibrahim, Chen, Lipsitz, and Herring (2).

More relaxed settings (that is, not requiring full specification of the complete-data likelihood) and asymptotic consistency of the estimators when the missingness mechanism is correctly specified make weighted estimating estimations an attractive approach for missing-data analysis. However, weighted estimating estimations are notorious for being less efficient than likelihood-based methods when the distributional assumptions of the likelihood are satisfied. The methods are also prone to bias for finite samples. Ibrahim, Chen, Lipsitz, and Herring (2) found that in small samples, doubly-robust weighted estimating equations yielded large finite-sample variances. Parzen, Lipsitz, Ibrahim, and Lipshultz (58) warned the need of a sufficient amount of missing data for obtaining the estimates of π_i with acceptable precision. Kang and

Schafer (49) noted that methods with inverse probabilities as weights were sensitive to misspecification of the model of missingness mechanism when some π_i are small, irrespective of whether the methods were doubly robust or not. They also showed that when both response model and missingness-mechanism model were incorrectly specified, inverse-probability weighting methods were biased procedures. The requirement for specifying and estimating the model of missing-data mechanism is practically difficult, unless the covariates $\mathbf{X}$ are either always observed or always missing (58).

2.2.4 Maximum Likelihood

Missing data analysis using maximum likelihood has the longest history among all procedures for this purpose, and yet continues to be an active area of research. Afifi and Elashoff (63) traced the studies back to Wilks (64). Anderson (65) introduced the concept of likelihood factorization for certain patterns of missing data to obtain maximum likelihood solutions in closed forms. Dempster, Laird, and Rubin (66) proposed the Expectation-Maximization (EM) algorithm for deriving maximum likelihood estimates of data with missing values based on the complete-data likelihood. Gourieroux and Montfort (67) used Anderson's technique for linear regression with missing covariates. Little and Schluchter (68) suggested an approach using EM algorithm for mixed categorical and continuous variables with missing data and applied it to linear and logistic regressions. Schluchter and Jackson (69) presented EM algorithm for survival analysis with missing categorical covariates. Ibrahim (70), Lipsitz and Ibrahim (71), Ibrahim, Lipsitz, and Chen (72), Lipsitz, Ibrahim, and Zhao (55), Lipsitz, Ibrahim, Chen, and Peterson (73), and Ibrahim, Lipsitz, and Horton (74) discussed EM algorithm in GLMs with

missing covariates. Chen and Fienberg (75), Fuchs (76), and Vach (77) also addressed about missing covariates in GLMs. On the other hand, Pepe and Fleming (78), Reilly and Pepe (79), Lawless, Kalbfleisch, and Wild (80), and Tang, Little, and Raghunathan (81) explored the use of quasi-likelihood methods in nonlinear regression with missing covariates.

Factorization of the data likelihood forms the basis of maximum likelihood methods on virtually all studies in the recent years. The trick is to get a closed form of the observed data likelihood in the presence of missing data. Schafer (26) and Little and Rubin (5) discussed at length the missing data settings where the observed data likelihood could be analyzed using conventional complete-data procedures. They showed that these settings involved alternative reparameterizations of certain models with ignorable missingness mechanism and monotone patterns of missing data. Ibrahim, Chen, Lipsitz, and Herring (2) argued, however, that the standard complete-data techniques for models like GLMs with nonmonotone pattern of missingness would be difficult, because it might require approximations of high-dimensional intergrals. An attractive and popular alternative for general patterns of missing data is the expectation-maximization (EM) algorithm by Dempster, Laird, and Rubin (66). This iterative optimization procedure works by augmenting incomplete data to accommodate unobserved or latent variables. The term "data augmentation" itself was introduced by Tanner and Wong (82; 83) when they proposed the idea of inserting a sampling or imputation step prior to optimization in the EM algorithm. In contrast to Dempster, Laird, and Rubin, data augmentation of Tanner and Wong is a stochastic algorithm, since they optimized the posterior density instead of only the likelihood function. Gelfand and Smith (38) suggested several approaches for the sampling step

of the data augmentation algorithm, including the use of Gibbs sampler by Geman and Geman (84). Wei and Tanner (85) then formalized the implementation of stochastic sampling to assist the EM algorithm. They named their method the Monte Carlo EM algorithm. Ibrahim (70) utilized the idea of Wei and Tanner to devise a general method for estimation of data with missing covariates in GLMs. Gilks and Wild (39) and Gilks, Best, and Tan (41) introduced the application of adaptive rejection sampling for the Gibbs sampler, which was later used by Ibrahim, Lipsitz, and Chen (72), and Ibrahim, Chen, Lipsitz, and Herring (2) to handle the continuous covariates with incomplete observations of their proposed models.

Let us briefly explore the maximum likelihood methods, in particular, those proposed by Ibrahim and his colleagues in Ibrahim (70), Ibrahim, Lipsitz, and Chen (72), and Ibrahim, Chen, Lipsitz, and Herring (2). Suppose that in n observations drawn through a simple random sampling from a population, the response Y are completely observed and the covariates $\mathbf{X}$ are partially missing where the missingness is dependent on the values of both Y and $\mathbf{X}$. Data from each subject $i = 1, \cdots, n$ are assumed to be independent. $\boldsymbol{\beta}$ are the parameters of interest, while $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ act as nuisance parameters. All parameters are assumed to be distinct. The joint distribution of Y, $\mathbf{X}$, and the missing data indicators $\mathbf{R}$ for subject i in the complete data is thus

$$f(y_i, \mathbf{x}_i, \mathbf{r}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = f(\mathbf{x}_i; \boldsymbol{\alpha}) f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{r}_i \mid y_i, \mathbf{x}_i; \boldsymbol{\gamma}).$$

It is not trivial to model the marginal density of $\mathbf{X}$, as they in general can consist of continuous and categorical variables with missing data. Little and Schluchter (68) used a multinomial model for the categorical variables and then applied a multivariate normal model for the con-

ditional distribution of the continuous variables given the observed pattern of the categorical variables. Ibrahim, Lipsitz, and Chen (72), however, argued that within parametric approaches the indexing parameters for $\mathbf{X}$ would be nuisance and, therefore, it would be critical to reduce the number of these nuisance parameters. Inferences in data with large fraction of missing values and substantial amount of nuisance parameters would be computationally burdensome and inefficient. As an alternative strategy, Lipsitz and Ibrahim (71) proposed a product of one-dimensional conditional distributions to model the joint distribution of $\mathbf{X}$. Ibrahim, Lipsitz, and Chen (72) then suggested to model the missing data mechanism in a similar fashion. They listed several advantages of this strategy, including reduction of nuisance parameters of the missing data mechanism, increased flexibility in model specification, good approximation to the 'standard' log-linear model, provision of a natural way for incorporating the dependency of missingness of a covariate on the missingness of another missing covariate, and facilitation of the E-step of the EM algorithm by the construction of scheme for efficient sampling from the conditional distribution of missing covariates given the observed data. Ibrahim, Lipsitz, and Chen (72) also noted that the proposed joint distribution of the missing data mechanism would be log-concave in $\boldsymbol{\gamma}$, because each of the one-dimensional conditional model was a logistic regression and, hence, r_{ik} being log-concave in $\boldsymbol{\gamma}$. Log-concavity helps reduce the burden of computation of the EM algorithm.

The observed-data likelihood with the presence of nonignorable missing covariates becomes

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{c=1}^{C} \left\{ \int f(y_i, \mathbf{x}_i, \mathbf{r}_i; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) d\mathbf{x}_{i,\text{mis}} \right\}^{I_{\{\mathbf{r}_i = c\}}}$$

where $\theta' = (\alpha', \beta', \gamma')$, and $I_{\{r_i=c\}} = 1$ if the missing data for subject i conforms to pattern $c = 1, \cdots, C$ and equals 0 otherwise. The integral sign becomes summation for the missing categorical covariates. The Q function in the E-step of the EM algorithm (66; 72) for subject i and $r_i = c$ can be written as

$$\begin{aligned} Q_i(\theta \mid \theta^{(t)}) =& \mathscr{E}\Big[\ell(\alpha; x_i) \mid y_i, x_{i,obs}, r_i = c; \theta^{(t)}\Big] + \\ & \mathscr{E}\Big[\ell(\beta; y_i \mid x_i) \mid y_i, x_{i,obs}, r_i = c; \theta^{(t)}\Big] + \\ & \mathscr{E}\Big[\ell(\gamma; r_i \mid y_i, x_i) \mid y_i, x_{i,obs}, r_i = c; \theta^{(t)}\Big], \end{aligned}$$

$\ell(.)$ being the log-likelihood function.

In terms of the asymptotic variance of $\hat{\theta}$, Ibrahim (70) recommended to use of the observed information matrix by Louis (86), where

$$\begin{aligned} \mathscr{I}(\hat{\theta}) =& -\sum_{i=1}^{n}\sum_{c=1}^{C} I_{\{r_i=c\}} \mathscr{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta; y_i, x_i, r_i)\middle| y_i, x_{i,obs}, r_i; \hat{\theta}\right] \\ & -\sum_{i=1}^{n}\sum_{c=1}^{C} I_{\{r_i=c\}} \mathscr{E}\left[\left\{\frac{\partial}{\partial\theta}\ell(\theta; y_i, x_i, r_i)\right\}\left\{\frac{\partial}{\partial\theta}\ell(\theta; y_i, x_i, r_i)\right\}'\middle| y_i, x_{i,obs}, r_i; \hat{\theta}\right] \\ & +\Bigg\{\sum_{i=1}^{n}\sum_{c=1}^{C} I_{\{r_i=c\}} \mathscr{E}\left[\frac{\partial}{\partial\theta}\ell(\theta; y_i, x_i, r_i)\middle| y_i, x_{i,obs}, r_i; \hat{\theta}\right] \\ & \mathscr{E}\left[\frac{\partial}{\partial\theta}\ell(\theta; y_i, x_i, r_i)\middle| y_i, x_{i,obs}, r_i; \hat{\theta}\right]'\Bigg\}. \end{aligned} \tag{2.3}$$

The quantities for $\hat{\theta}$ can be obtained from the estimates at convergence.

Ibrahim et al.'s approach, however, can easily run to estimability issue when the model for missing data mechanism (and also for the missing covariates) becomes too large. Ibrahim,

Lipsitz, and Chen (72) acknowledged this problem and warned the importance of proper choice for functions of $(Y, \mathbf{X})$ to be included in $\mathbf{R} \mid Y, \mathbf{X}$, and cautious selection of interaction terms or other higher order terms that would be added on the model. They were yet to provide either solution or an ad hoc workaround that would ensure the model identifiable. Another concern for the joint modeling via a product of one-dimensional conditional distributions is the requirement of a particular ordering for the variables with incomplete observations (87; 13; 16). The real data complexity may turn out to be against such a condition. Furthermore, the number of possible orderings grows substantially as the dimension of partially observed variables increases. Bartlett et al. (87) had noted that very few applied statisticians used Ibrahim et al.'s method thus far, and they hypothesized it to be attributable to this ordering problem. One additional issue, though minor, is that application of the approach was primarily studied under a simple random draw assumption. General sampling schemes, such as those applied in survey studies, may require some adjustments that have not been addressed in the previous works. For instance, the outcome is clearly subject to missing data due to sampling and thus the management of incomplete observations has to be extended to the outcome as well. Despite such problems, Ibrahim et al.'s approach is fairly straightforward to implement. If the models are correctly specified and meet certain regularity conditions, the estimates possess many attractive limiting properties of the maximum likelihood estimators, such as consistency and efficiency.

2.3 Sampling Design in Inferences using Survey Data

2.3.1 Spectrum of Opinions on the Role of Sampling Weights

Sampling is a special case of missing data. In particular, the values of some units or subjects are missing by design. A frequently challenging question in the analysis of sampling data is whether the study design should be accounted for. Sampling weights are routinely included in data from major surveys. The samples are weighted to compensate for unequal inclusion probabilities with respect to the target population. While their role in descriptive population quantities (such as means, proportions, totals) is generally accepted, implementation of sampling weights in analytical inference is a subject of controversy and confusion among theorists and practitioners (88; 89). Pfeffermann (89) discussed in depth the differing positions of statisticians on this issue. He proposed the classification into survey or design-based, model-based, and robust model-based statisticians.

At one extreme, survey statisticians tend to consider it impractical to fit models that appropriately approximate the true population models. This is due to the large heterogeneity of populations in practice and the complexity of designs that were often used to draw samples. They consequently focus their inferences on finding the finite population quantities $\boldsymbol{\theta}_{\mathscr{F}}$ for the model parameters. They also incorporate the sampling weights into every survey analysis. The advantages of this approach are the interpretability of the estimates and the robustness of the relationship between the survey variables. That is, the estimated quantities do not lose their meaning despite the models fail. Consistency of the estimates using such an approach is emphasized at the perspective of study design. Classical examples for this class of inferences

include the works by Kish and Frankel (90), Jonrup and Rennermalm (91), and Shah, Holt, and Folsom (92). Further discussion on the theories and applications can be found in Little (88), Skinner, Holt, and Smith (93), and Fuller (94).

On the other extreme, model-based statisticians refer to the group who is faithful to the philosophical purpose of inference of approximating the true parameters θ of the population. The models developed during the inference process are in such a way that they would apply to populations more general than the fixed population giving rise to the sample. Statisticians of this group concentrate on the correct specification of the models and the conditions that ensure the models hold. The estimates of the model parameters are hypothetical in nature, meaning that they are the quantities that would be observed had the model holds and the conditions satisfied. From this perspective, it becomes irrelevant to incorporate sampling weights (95; 96). If the model holds, the weights have no role but complicating the inference. An obvious strength of this approach is the optimal properties of the model parameter estimates. Unbiasedness of the estimates is the ultimate goal. Consistency is judged with respect to the assumed population model. Casella and Berger (97), Lehmann (98), and Kasprzyk, Duncan, Kalton, and Singh (99) provide detailed discussion on the theoretical and empirical results of model-based inferences.

The third group of inferential approach for survey data, according to Pfeffermann (89), is represented by theorists and practitioners who equally weigh the model unbiasedness and design consistency. They consider the model parameters as the ultimate parameters of interest, but at the same time they are also concerned about preserving the robustness of the inference. In order to achieve that, a model is first postulated under the condition $U(Y, X; \theta) = 0$, where $U(.)$ is a

real valued function. The solution $\hat{\theta} = T(Y, X)$ is the corresponding finite population quantities for θ under the estimation rule $R[(Y, X) \rightarrow \theta]$. In contrast to the model-based statisticians who focus on finding an optimal estimators under the model, however, the interest here is on obtaining the estimators based on sample, $t(n)$, that are design-consistent for the population estimator $T(N)$. That is, $\lim_{n\rightarrow\infty, N\rightarrow\infty} \Pr\{|t(n) - T(N)| \geq \epsilon\} = 0$, $\epsilon > 0$. The advantages of this approach are two sided. First, if the model is correctly specified, then as the population size increases t will converge to θ. Second, if the model is unfortunately misspecified, any finite population values such as T and their corresponding sample quantities t are still real entity that have actual interpretation. See for example Binder (100), Godambe and Thompson (101), Little (88), Kreiger and Pfeffermann (102), and Breckling, Chambers, Dorfman, et al. (103).

2.3.2 Conditions Requiring Inclusion of Sampling Design

Let Z_i, $i = 1, \cdots, N$, be a vector of design variables for subject i that may include information such as cluster or stratum indicators, other grouping variables, and quantitative characteristics such as measures of size of sampling units. It is reasonable to expand the model developed in the beginning of this chapter such that it becomes $f(y_i, x_i, z_i; \beta, \alpha, \psi) = f(y_i \mid x_i, z_i; \beta) f(x_i; \alpha) f(z_i; \psi)$, under the assumption that $x_i \perp z_i$. In contrast to the settings on the section about missing data, X_i are allowed to be fully observed for all i. We assume that each i is independent. Suppose that a sample $\mathscr{S}$ of size n is drawn from the population $\mathscr{F}$ under the sampling scheme $p_i \equiv \Pr\{i \in \mathscr{S}\}$. The objective is to estimate β. Let I_i be the sampling indi-

cator such that $I_i = 1$ for $i \in \mathscr{S}$ and $I_i = 0$ otherwise. The joint distribution of $(I_i, Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ is

$$\left[p_i f(y_i, x_i, z_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\psi})\right]^{I_i} \left[\int (1 - p_i) f(y_i, x_i, z_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\psi}) dy_i\right]^{1-I_i} \tag{2.4}$$

where $\{\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\psi}, \boldsymbol{\delta}\}$ are assumed to be distinct parameters.

We consider two important situations as discussed in Rubin (1; 9), Little (104), Sugden and Smith (105), Kish (106), Pfeffermann (89), and Chambers, Dorfman, and Wang (107): First,

$$p_i = \Pr\{\mathbf{I} \mid \mathbf{Y}, \mathbf{Z}\} \tag{2.5}$$

and second,

$$\begin{gathered} p_i = \Pr\{\mathbf{I}|\mathbf{Z}\}, \text{ or alternatively,} \\ \Pr\{\mathbf{Y}_{\mathscr{S}}|\mathbf{Y}_{\bar{\mathscr{S}}}, \mathbf{X}, \mathbf{Z}, \mathbf{I}\} = \Pr\{\mathbf{Y}_{\mathscr{S}}|\mathbf{Y}_{\bar{\mathscr{S}}}, \mathbf{X}, \mathbf{Z}\} \end{gathered} \tag{2.6}$$

where $\mathbf{I} = (I_1, \cdots, I_N)'$, $\mathbf{Y} = (Y_1, \cdots, Y_N)'$, $\mathbf{Z} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_N)'$, $\mathbf{Y}_{\mathscr{S}} = \{(i, Y_i) : i \in \mathscr{S}\}$, and $\bar{\mathscr{S}}$ represents the units that are not in $\mathscr{S}$. In the first situation, the sampling design is 'informative'. The unobserved values of Y among the non-sampled subjects, in Rubin's terminology (1), are MNAR. An unbiased inference on $\boldsymbol{\beta}$ thus requires the inclusion of sampling design. By contrast, the probability of a subject being selected into $\mathscr{S}$ for the second situation leads to MAR Y_i within the non-sampled subjects. Survey statisticians refer to this as 'ignorable' sampling scheme (89; 107). As $(1 - p_i)$ becomes outside the integration, the inference on $\boldsymbol{\beta}$ proceeds safely without the need for inclusion of the sampling design. We note, however, that the

independence of Y_i and I_i in the second situation are conditional on $\mathbf{Z}_i$, the design variables. In other words, the sampling design is ignorable for estimating $\boldsymbol{\beta}$ if the model include all the design variables. A more theoretical treatment of the conditions under which the sampling design is ignorable or informative is provided in Rubin (1), Little (104), and Sugden and Smith (105).

It is generally a complicated task, unfortunately, to satisfy the ignorability conditions in data from complex surveys. The number of design variables often follows the multistage nature of samples selection. On the one hand, it becomes less likely for the analyst who is not the sampler to know the values of all the design variables for all subjects in the population. On the other hand, even when the design information is fully known and available, the variables may be too many to handle. Alexander (108) reminds the formidable consequences in the derivation, fit, and validation of the models when one include all the relevant design variables. Thus in either way, the idea of incorporating all the design variables into the model seems rather ambitious than practical.

The impracticality of a complete inclusion of the design information within the models brings another problem. It has been shown that such an incomplete incorporation may seriously impact the inferences (93). Similar effects have also been demonstrated in the misspecification of the distribution for the survey variables given the design variables. There are indeed cases where the use of partial design information is sufficient to avoid biased inferences. Sugden and Smith (105) define the conditions for which ignorability holds in these cases. They first propose a set of adequate summary of $\mathbf{Z}_i$ for any $i \in \mathscr{S}$, namely $\mathbf{D}_{\mathscr{S}} = \{(i, D(\mathbf{Z}_i)) : i \in \mathscr{S}\}$ with realization $\mathbf{d}_{\mathscr{S}}$,

such that the relation $\mathbf{I} \perp \mathbf{Z} \mid \mathbf{D}_{\mathscr{S}} = \mathbf{d}_{\mathscr{S}}$ holds for all $\mathtt{i} \in \mathscr{S}$. Note that $\mathbf{I}$ is a vector and $\mathbf{Z}$ is a matrix. The design information $\mathbf{D}_{\mathscr{S}}$ may be obtained from any known quantities or functions of $\mathbf{Z}$, knowledge of the sampling design, and if they are available, the values of the sampling probability. Accordingly, the forms of $\mathbf{D}_{\mathscr{S}}$ are dependent on the information available to the analyst. If it consists of a single variable then $\mathbf{D}_{\mathscr{S}}$ is a vector, otherwise a matrix. The above condition allows the conditional distribution of $\mathrm{I}_{\mathtt{i}} \mid \mathbf{Z}_{\mathtt{i}}$ on the second situation we discussed previously to be replaced with $\mathrm{I}_{\mathtt{i}} \mid \mathrm{D}(\mathbf{Z}_{\mathtt{i}})$ for all $\mathtt{i} \in \mathscr{S}$. If such condition is not satisfied, the authors suggested that the design may still be ignorable for inferences on $\boldsymbol{\beta}$ if for all $\boldsymbol{\beta}$ and all $\mathtt{i} \in \mathscr{S}$, $\mathbf{Y} \perp \mathbf{Z} \mid \mathbf{D}_{\mathscr{S}}; \boldsymbol{\beta}$. All formulation in their paper are based on the assumption that the analyst of survey data is not the sampler, and thus the discussion does not cover the non-sampled units.

A replacement of the relevant design variables with a proxy variable containing partial design information gives rise to the need of testing whether the design is indeed ignorable given the available design information. Pfeffermann (89) explores such tests and concludes that they may either give inconclusive results or indicate that the conditions for ignorability are not met. In the end, he concludes that sampling weights should be used for model based inferences as they can protect against bias due to informative designs. He also adds that the sampling weights can play a critical role in protecting against model specification.

2.3.3 Strategies to Incorporate Sampling Design in Inferences

It becomes evident from the preceding discussion that incorporation of the sampling design in the inference process may be the best alternative for the analysts of survey data who them-

selves are not the samplers. The key argument is the practical difficulties of verifying whether the ignorability of sampling design is satisfied given the limitation of information within the samples. Fortunately, we have a rich literature addressing the strategies to include sampling designs in the inferences. I use some of them in the following description. It includes the strategies frequently implemented in practice, which are categorized into: modifications at estimators, model, and estimating functions levels.

2.3.3.1 Modifications at Estimators Level

There are certain cases where the model-based estimators have explicit expressions. For example, in a linear regression model $f(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta})$ where $\mathbf{Y}$ and $\boldsymbol{\beta}$ are vectors, and $\mathbf{X}$ is a matrix, $\hat{\boldsymbol{\beta}} = \sum_{i=1}^{N}(\mathbf{x}_i\mathbf{x}_i')^{-1}\sum_{i=1}^{N}\mathbf{x}_i y_i$ is the least square estimator for $\boldsymbol{\beta}$. It is also $\hat{\boldsymbol{\beta}}_{\mathscr{F}}$ since it produces the census estimates or the finite population quantities for the population $\mathscr{F}$ (compare this to the definition from Binder (100), Godambe and Thompson (101), Little (88), and Pfeffermann (89). One can modify this estimator using the method by Horvitz and Thompson (109), such that $\hat{\boldsymbol{\beta}}^{w}_{\mathscr{S}} = \sum_{i=1}^{N}(w_i\mathbf{x}_i\mathbf{x}_i')^{-1}\sum_{i=1}^{N}w_i\mathbf{x}_i y_i$, which is also equals to$\sum_{i=1}^{n}(1/p_i)(\mathbf{x}_i\mathbf{x}_i')^{-1}\sum_{i=1}^{n}(1/p_i)\mathbf{x}_i y_i$, provides the weighted sample quantities for $\mathscr{S}$. The theory of linear models has shown that $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$. In addition, as n and N become sufficiently large, the probability that $\hat{\boldsymbol{\beta}}^{w}_{\mathscr{S}}$ and $\boldsymbol{\beta}_{\mathscr{F}}$ are different subsides, thus the design consistency. Pfeffermann (89) indicates that this class of approach for incorporating sampling design in general produces design-consistent estimators for the finite population quantities. Further examples of this strategy are available in Kish and Frankel (90), Shah, Holt, and Folson (92), Nathan and Holt (110), Dumouchel and Duncan (111), and Fuller (94).

The drawbacks of this otherwise convenient strategy are the fact that not every estimator has an explicit expression, and also the possibility of more than one design-consistent estimators for the same estimands. The later problem has been discussed, for instance, in Pfeffermann and Holmes (112), Little (88), and Pfeffermann (89). Suggested treatments have been proposed in Nathan and Holt (110), Pfeffermann and Holmes (112) and Rao, Kovar and Mantel (113).

2.3.3.2 Modifications at Model Level

Another strategy to account for the sampling design in the inference process is making it part of the model. This idea can be accomplished in several ways, such as by taking the inclusion probability as a covariate, weighting the distribution with the inclusion probabilities, or modifying the model parameters to include the sampling design.

Suppose that the inclusion probability p_i is available for all N units and it is deemed too complicated to model $\mathbf{Y}$ given $\mathbf{Z}$. Additionally, there is a vector $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\mathbf{Z})$ such that $\Pr\{\mathbf{I}|\mathbf{Z}\} = \Pr\{\mathbf{I}|\boldsymbol{\alpha}\}$ for all i, where $\mathbf{I}$ is the vector of sampling indicator as it was defined previously, and $\mathbf{Z}$ is the matrix of design variables. Then following Rubin (114), $\boldsymbol{\alpha}$ is an 'adequate' summary of $\mathbf{Z}$. Rubin indicates two consequences of this property. First, there cannot be any adequate summary of $\mathbf{Z}$ that is coarser than $\mathbf{p} = (p_1, \cdots, p_N)$, because $p_i = \Pr\{I_i = 1|\mathbf{z}_i\} = \Pr\{I_i = 1|\boldsymbol{\alpha}_i\}$ and thus $\mathbf{p}$ must be a function of $\boldsymbol{\alpha}$, and $\boldsymbol{\alpha}$ cannot be coarser than $\mathbf{p}$. Second, by Bayes theorem $\Pr\{\mathbf{Y}_{\mathscr{S}}|\mathbf{Y}_{\bar{\mathscr{S}}}, \boldsymbol{\alpha}, \mathbf{I}\} = \Pr\{\mathbf{Y}_{\mathscr{S}}|\mathbf{Y}_{\bar{\mathscr{S}}}, \boldsymbol{\alpha}\}$. Though Rubin admits that $\mathbf{p}$ is the coarsest possible adequate summary of $\mathbf{Z}$, even too coarse, he argues that it gives advantages in specifying realistic models for $\Pr\{\mathbf{Y}|\boldsymbol{\alpha}\}$; the coarser the summary $\boldsymbol{\alpha}$, the simpler in general the task of specifying the models. Of course, the specification can be expanded to $\Pr\{\mathbf{Y}|\mathbf{X}, \boldsymbol{\alpha}\}$ as the model developed

in this chapter. Other than the coarseness of $\mathbf{p}$, another problem with this approach is the requirement to have information of the inclusion probability for all $i \in N$. Sugden and Smith (105) provide remedies for these limitations, which have been discussed in previous section. In essence, they define the conditions of which $\mathbf{D}_{\mathscr{S}}$ can be an adequate summary of $\mathbf{Z}$ for any $i \in \mathscr{S}$. It is also important to note that the models with the inclusion probabilities as covariates do not necessarily produce design-consistent estimators (89).

Model-level modifications can also be implemented when the available information is limited to the n sampled units. In this case, one needs to specify (115; 116)

$$f(y_i, x_i | I_i = 1; \delta, \beta, \alpha) = \frac{f(I_i = 1 | y_i, x_i; \delta) f(y_i, x_i; \beta, \alpha)}{\iint f(I_i = 1 | y_i, x_i; \delta) f(y_i, x_i; \beta, \alpha) dy_i dx_i},$$

where δ, β, and α are assumed to be distinct vector parameters. Notice that the integration is effectively only for $\mathbf{Y}_{\mathscr{S}}$ and $\mathbf{X}_{\mathscr{S}}$. Specification of $f(I_i = 1 | y_i, x_i; \delta)$ may be performed empirically using the sample inclusion probabilities and the observed measurements. It is suggested that δ are either empirically estimated or fixed based on external estimates (89). Examples of the use of this method are provided in Patil and Rao (115), and Krieger and Pfeffermann (102). As it was with the models with the inclusion probabilities as covariates, the resulting estimators of the models with weighted distribution are not necessarily design consistent for their finite population quantities (89).

Little (88) suggests the methods to take account for sampling design in the case of simple stratified sampling through a modification of the model parameters. The idea is based

on Bayesian inference about the variables given the observed data using the posterior predictive distribution. To be more concrete, one defines $f(\mathbf{y},\mathbf{x}|\mathbf{y}_{\mathscr{S}},\mathbf{x}_{\mathscr{S}}) = \prod_{j=1}^{J}\int\prod_{i=1}^{N_j} f(y_{ij},\mathbf{x}_{ij};\theta_j)$ $\varphi(\theta_j;\mathbf{y}_{\mathscr{S}},\mathbf{x}_{\mathscr{S}})d\theta_j$, where i and j respectively index units and strata, J is the number of strata, and N_j the number of population units in stratum j. The function $f(y_{ij}\mathbf{x}_{ij};\theta_j)$ is equal to $f(y_{ij}|\mathbf{x}_{ij};\theta_{j1})f(\mathbf{x}_{ij};\theta_{j2})\varphi(\theta_j)$. Let $\theta_{j2} = (\upsilon_j,\boldsymbol{\nu}_j)$, where $\upsilon_j = (\upsilon_{j1},\cdots,\upsilon_{jK})$ are the location parameters for stratum $\mathbf{X}_{ij}$, and $\boldsymbol{\nu}_j$ the dispersion or shape parameters. For fixed stratum-effect models, $\varphi(\upsilon_j) \propto C$, a constant, while for random stratum-effect models $\varphi(\upsilon_j) = \iint\prod_{j=1}^{J}$ $\varphi(\upsilon_j|\upsilon,\omega)d\upsilon d\omega$. Little (88) shows that the estimators using this approach are design-consistent for their counterpart finite population quantities and unbiased for θ. Development of his models in the referred paper, however, focuses on the linear regression coefficients. There is also no extension to arbitrary sampling design. The simulations show that in small samples, the inferences can be very sensitive to the assumption of the priors. Nevertheless, there have been studies of the same spirit conducted by DuMouchel and Duncan (111) and Alexander (108), though their estimators do not necessarily design consistent.

2.3.3.3 Modifications at Estimating Functions Level

Godambe and Thompson (101) introduce a general method of obtaining design-consistent estimators by modification at the estimating functions level. To fix the idea, we consider again the vector $\mathbf{Y}$ and the matrix $\mathbf{X}$ as defined previously and assume that they are generated from

$f(y, x; \theta) = \xi(\theta)$. Let us denote the estimating function for the unknown parameters θ based on the finite population values of Y and X as $g = g(Y, X; \theta)$. To be model unbiased,

$$\mathscr{E}_\xi[g(Y, X; \theta(\xi))] = 0, \tag{2.7}$$

where $\mathscr{E}_\xi$ denotes expectation under the model $\xi(\theta)$. In addition, the estimating function $g^* = g^*(Y, X; \theta)$ is also optimal if, under appropriate regularity conditions, it minimizes among unbiased g

$$\left\{\mathscr{E}_\xi[\partial g/\partial\theta]^{-1}\mathscr{E}_\xi[g g']\mathscr{E}_\xi[\partial g/\partial\theta]^{-1}\right\}_{\theta=\theta(\xi)}. \tag{2.8}$$

If g^* is optimal, then $g^* = 0$ is the optimal estimating equation and the solution for θ is the optimal estimates. Godambe and Thompson (101) indicate that for the finite population $\mathscr{F}$ with N units, the solution $\theta_{\mathscr{F}}(\xi)$ of $g^*(Y, X; \theta_{\mathscr{F}}(\xi)) = 0$ can be both 'estimates' of θ when all N components of (Y, X) are known, and 'parameters' for $\mathscr{F}$ when not all components are known.

Optimality of g^* may include the cases where it is of linear form. Suppose that all (Y_i, X_i), $i = 1, \cdots, N$, are independent, then there exists a real-valued estimating function in the form of (101)

$$\sum_{i=1}^{N} u(y_i, x_i; \theta) \tag{2.9}$$

where $\mathscr{E}_\xi[U(Y_i, \mathbf{X}_i; \theta(\xi))] = 0$. Alternatively, the estimating function may be linear in $\mathbf{U} = (\mathbf{U}_1, \cdots, \mathbf{U}_N)$, $\mathbf{U}_i = U(y_i, \mathbf{x}_i; \theta)$, such that

$$\sum_{i=1}^{N} U(y_i, \mathbf{x}_i; \theta)\varphi_i(\theta), \tag{2.10}$$

$\varphi_i(\theta)$ being arbitrary real functions of θ. Since $\mathscr{E}_\xi[\mathbf{U}_i] = 0$, both Equation 2.9 and Equation 2.10 are model unbiased. If $\mathbf{g}^*$ minimizes Equation 2.8 in the form of Equation 2.9 for $\mathbf{g} = \mathbf{g}^*$, then $\mathbf{g}^*$ is referred to as optimal; if it is in the form of Equation 2.10, $\mathbf{g}^*$ is termed 'linearly optimal'. Godambe and Thompson (101) provide the sufficient condition for $\mathbf{g}^*$ to be linearly optimal. The importance of linearly optimal $\mathbf{g}^*$ comes in place when we consider the sample $\mathscr{S}$ instead of the finite population $\mathscr{F}$. Let the sample $\mathscr{S}$ of size n be drawn from the finite population $\mathscr{F}$ based on a sampling design $p_i = f(I_i = 1|z_i; \delta)$, where as before I_i is the sampling indicator, Z_i either a scalar or a vector of design variables that in general may include Y_i and/or $\mathbf{X}_i$, and δ can be a scalar or a vector of either fixed or unknown design parameters. For example, a sample selection using probability proportional to size approach with Z as the size variable, has $p_i = nz_i/N\bar{z}$. We denote $\rho(\delta) = \mathbf{p} = (p_1, \cdots, p_N)'$, and $(\mathbf{Y}_\mathscr{S}, \mathbf{X}_\mathscr{S}) = \{(i, Y_i, \mathbf{X}_i) : i \in \mathscr{S}\}$. The function $h(\mathbf{Y}_\mathscr{S}, \mathbf{X}_\mathscr{S}; \theta)$ is design unbiased if

$$\mathscr{E}_\rho[h(\mathbf{Y}_\mathscr{S}, \mathbf{X}_\mathscr{S}; \theta)] = \sum_{i=1}^{N} U(y_i, \mathbf{x}_i; \theta), \tag{2.11}$$

where $\mathscr{E}_\rho$ represents expectation with respect to the design $\rho(\delta)$. An optimal choice for $\mathbf{h} = h(\mathbf{Y}_\mathscr{S}, \mathbf{X}_\mathscr{S}; \theta)$ is $\mathbf{h}^* = h^*(\mathbf{Y}_\mathscr{S}, \mathbf{X}_\mathscr{S}; \theta)$, which minimizes

$$\left\{ \mathscr{E}_\xi \mathscr{E}_\rho [\partial \mathbf{h}/\partial\theta]^{-1} \mathscr{E}_\xi \mathscr{E}_\rho [\mathbf{h}\mathbf{h}'] \mathscr{E}_\xi \mathscr{E}_\rho [\partial \mathbf{h}/\partial\theta]^{-1} \right\}_{\theta=\theta(\xi)}. \tag{2.12}$$

The solution for θ from $\mathbf{h}^*$, denoted $\hat{\theta}_\mathscr{S}$, is thus optimal for $\theta_\mathscr{F}$, which itself is the solution for θ from $\mathbf{g}^*$. Godambe and Thompson (101) demonstrate that

$$\mathbf{h}^* = \sum_{i=1}^{n} \frac{U(y_i, \mathbf{x}_i; \theta)}{p_i} = \sum_{i=1}^{N} w_i U(y_i, \mathbf{x}_i; \theta), \tag{2.13}$$

where $w_i = I_i/p_i$. We note here that under certain regularity conditions, $U(y_i, \mathbf{x}_i; \theta) = \frac{\partial}{\partial\theta} \log f(y_i, \mathbf{x}_i; \theta)$, the score function. Consequently, Equation 2.13 becomes a pseudo likelihood approach. Pfeffermann (89) indicates that for some regularity conditions the estimating function $\mathbf{h}^*$ meets the condition

$$\mathscr{E}_\xi \mathscr{E}_\rho \left[\left[\mathbf{h}^* - \frac{\partial}{\partial\theta} \log f(y_i, \mathbf{x}_i; \theta) \right] \left[\mathbf{h}^* - \frac{\partial}{\partial\theta} \log f(y_i, \mathbf{x}_i; \theta) \right]' \right] \leq$$
$$\mathscr{E}_\xi \mathscr{E}_\rho \left[\left[\mathbf{h} - \frac{\partial}{\partial\theta} \log f(y_i, \mathbf{x}_i; \theta) \right] \left[\mathbf{h} - \frac{\partial}{\partial\theta} \log f(y_i, \mathbf{x}_i; \theta) \right]' \right],$$

which theoretically justifies the application of the pseudo likelihood approach in cases where it produces the optimal estimating equation. The pseudo likelihood approach has been studied by many, for examples, Binder (100), Chambless and Boyle (117), Fuller (118; 94), Kreiger and Pfeffermann (102), Breckling, Chambers, Dorfman, et al. (103), and Chambers, Dorfman, and

Wang (107). Recently, Skinner and Mason (119) propose $w_i q_i U(y_i, x_i; \theta)$ for Equation 2.13, where $q_i = q(x_i)$ being an arbitrary function. This newer approach has the appeal of improving estimation efficiency.

2.4 Models for Count Data

2.4.1 Traditional Approaches

Suppose that the data $(Y_i, \mathbf{X}_i)$, $i = 1, \cdots, N$ are independent, where $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iP})'$. Let us assume that the data are collected through simple random sampling and there is no missing values among the samples. In the spirit of the generalized linear model (GLM), due to Nelder and Wedderburn (120), let

$$f(y_i \mid x_i; \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right], \tag{2.14}$$

where functions $a_i(.)$, $b(.)$ and $c(.)$ are known. The interest is on estimating the canonical parameter θ_i, such that the dispersion parameter ϕ is considered a nuisance. Setting θ_i as $\log \lambda_i = x_i'\beta$, $a_i(\phi)$ as unity, $b(\theta_i) = e^{x_i'\beta}$, and $c(y_i, \phi) = -\log(y_i!)$, one obtains the conditional distribution of Y_i as Poisson with parameter $\lambda_i = e^{\eta_i} > 0$. The log-likelihood function is accordingly

$$\ell(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{n} [y_i x_i'\boldsymbol{\beta} - e^{x_i'\beta} - \log(y_i!)], \tag{2.15}$$

$\mathbf{Y} = (Y_1, \cdots, Y_N)'$, $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_N)'$. A standard approach to solve the GLM estimating equations for Equation 2.15 is the iteratively weighted least squares (IWLS) algorithm. In

particular, one creates $\mathbf{Z}$ with elements $Z_i = \hat{\eta}_i + \frac{Y_i - \hat{\lambda}_i}{\hat{\lambda}_i}$, $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$, such that $\mathscr{E}[\mathbf{Z}] = \boldsymbol{\eta}$ and $\mathscr{V}[\mathbf{Z}] = \text{diag}\{\hat{\lambda}_i\}$, and regresses $\mathbf{Z}$ on $\mathbf{X}$ using IWLS with weight $\mathbf{W} = \{\mathscr{V}[\mathbf{Z}]\}^{-1}$.

2.4.2 Overdispersion and Excess Zeros Model

The conventional Poisson model may require adjustment because of the restriction $\mathscr{E}[Y_i \mid \mathbf{X}_i] = \mathscr{E}[Y_i \mid \mathbf{X}_i] = \lambda_i$. Count data in practice might violate this equidispersion property, often in the form of much larger conditional variance than the mean (overdispersion). A simple remedy is the quasi-likelihood method (121). To take advantage of it, one introduces a dispersion parameter ϕ and sets $\mathscr{V}[Y_i \mid \mathbf{X}_i] = \phi\lambda_i$, then revises the weight matrix of IWLS. The estimates are typically similar to Poisson GLM, but the standard errors tend to be larger. A caveat of the approach is that one no longer assumes any distribution for the data. If in fact the counts are Poisson distributed, the quasi-likelihood method can give less efficient estimates then the maximum likelihood solution.

A more involved but distributional solution for circumventing overdispersion is the application of the mixed (also called multilevel or hierarchical) models. In this context, the models can be expressed, for instance, as

$$f(y_i \mid \mathbf{x}_i; \lambda_i) = \int f(y_i \mid \mathbf{x}_i; \lambda_i)\varphi(\lambda_i; \boldsymbol{\eta})d\eta_i, \quad \text{or} \tag{2.16a}$$

$$= \int f(y_i \mid \mathbf{x}_i; \lambda_i, \epsilon_i)\varphi(\epsilon_i)d\epsilon_i \tag{2.16b}$$

where ϵ_i may be a vector. Thus, given $\boldsymbol{\eta}$ or ϵ_i, the observed data follow a certain distribution, but the random term is not observed. For many instances, the conditional distribution of y_i

in Equation 2.16a or Equation 2.16b is set to be Poisson, but the random term is specified to follow another reasonable distribution, for instance, Gamma, Inverse-Gaussian, or Log-Normal (62; 122; 123). An interesting result is obtained for the Poisson-Gamma model. One writes

$$Y_i \mid \mathbf{X}_i \sim \text{Poisson}(\lambda_i^*);$$

$$\lambda_i^* \sim \text{Gamma}(\text{shape} = \kappa_i, \text{scale} = \nu_i); \kappa_i > 0, \nu_i > 0$$

such that

$$f(y_i \mid \mathbf{X}_i; \lambda_i) = \frac{1}{y!\Gamma(\kappa)\nu_i^\kappa} \int \lambda_i^{*(y_i+\kappa-1)} e^{-\lambda_i^*(1+\frac{1}{\nu_i})} d\lambda_i^*$$

Setting $E(\lambda_i^*) = \lambda_i$ and $\kappa_i = \kappa$, which implies $\nu_i = \lambda_i/\kappa$, leads to a familiar expression

$$\frac{\Gamma(y_i + \kappa)}{y_i!\Gamma(\kappa)} \left(\frac{\lambda_i}{\lambda_i + \kappa}\right)^{y_i} \left(\frac{\kappa}{\lambda_i + \kappa}\right)^{\kappa} \tag{2.17}$$

since it is a Negative-Binomial (NB) model with parameters κ and $\pi_i = \frac{\lambda_i}{\lambda_i+\kappa}$. Note that this result can also be achieved through explicitly setting $\lambda_i^* = \lambda_i \epsilon_i$, and assume that ϵ_i is Gamma distributed with $\mathscr{E}[\epsilon_i] = 1$ and $\mathscr{V}[\epsilon_i] = 1/\nu$. The NB model is a well-known extension of the Poisson model to address overdispersed count data (62). Accordingly, while $\mathscr{E}[Y_i \mid \mathbf{X}_i] = \lambda_i$, $\mathscr{V}[Y_i \mid \mathbf{X}_i] = \lambda_i + \frac{\lambda_i^2}{\kappa}$. The ML estimates (MLE) of the NB regression can be solved using the numerical procedures for GLMs.

Overdispersed count data might be caused by or exist with excess number of zero observations. To deal with excessive zero counts, Lambert (124) suggests a zero-inflated Poisson (ZIP) regression, where Y is considered to come from two distributions

$$\begin{aligned} Y_i &\sim 0 && \text{with probability } \xi_i \\ Y_i &\sim \text{Poisson}(\lambda_i) && \text{with probability } 1-\xi_i. \end{aligned}$$

And accordingly,

$$\begin{aligned} P(Y_i = 0) &= \xi_i + (1-\xi_i)e^{-\lambda_i} \\ P(Y_i > 0) &= (1-\xi_i)e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!}. \end{aligned}$$

The relationship between Y_i and the covariates $\mathbf{X}_i = (\mathbf{b}_i', \mathbf{c}_i')'$ is specified using the GLM canonical links, such that $\lambda_i = e^{\mathbf{X}_i'\boldsymbol{\beta}_i}$ and $\xi_i = (1+e^{-\mathbf{c}_i'\boldsymbol{\gamma}_i})^{-1}$. Suppose that $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}$ are related, then Lambert suggests the parameterization $\xi_i = (1+\lambda_i^{\zeta})^{-1}$, where $\zeta \in \mathbb{R}$. Solutions for the ZIP MLEs are derived, for instance, using the EM algorithm, due to Dempster, Laird, and Rubin (66).

2.4.3 Nested Data Structure

Count data in many applications are nested. For example, the data for children may be nested within mothers or households, which in turn also a part of larger clusters. Hierarchical structure typically introduces correlation among the observations within the same cluster. The most widely used methods to handle correlated counts due to nested structure are the mixed

models. Expanding Equation 2.16b to accommodate n_i counts on the i-th cluster in the sample and assuming conditional independence across clusters, one has

$$f(y_i \mid \mathbf{x}_i; \boldsymbol{\lambda}_i) = \int \prod_{j=1}^{n_i} f(y_{ij} \mid \mathbf{x}_{ij}; \lambda_{ij}, \epsilon_i) \varphi(\epsilon_i) d\epsilon_i, \tag{2.18}$$

where in this case $\mathbf{x}_i = (\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{in_i})'$.

An example of setting for Equation 2.18 is the Poisson-Log Normal model. In particular, let $\theta_{ij} = \lambda_{ij} h_i$, where $\lambda_{ij} = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})$ and $h_i = \exp(\epsilon_i)$, and set the distribution of ϵ_i as Gaussian with mean $\mu = -\sigma^2/2$ and variance σ^2. Accordingly,

$$f(y_i \mid \mathbf{x}_i; \boldsymbol{\lambda}_i) = \int_{-\infty}^{\infty} \left(\prod_{j=1}^{n_i} \frac{e^{-\theta_{ij}} (\theta_{ij})^{y_{ij}}}{y_{ij}!} \right) \frac{\exp(-\epsilon_i^2/2\sigma^2)}{\sqrt{2\pi}} d\epsilon_i.$$

CHAPTER 3

METHODOLOGY

3.1 Notation

A finite population $\mathscr{F}$ with N subjects is assumed to be the setting of a research interested in estimation of the characteristic Y based on its relationship with $\mathbf{X} = (X_1, \cdots, X_p)$. The subset $\mathscr{S}$ of $\mathscr{F}$ is then drawn under the sampling selection probability $p_i = \Pr\{i \in \mathscr{S}\}$, $i = 1, \cdots, N$ where each i is considered independent, and the observed $(Y_i, \mathbf{X}_i)$ are measured among the subjects in $\mathscr{S}$. Let $I_i = 1$ denotes $i \in \mathscr{S}$ and $I_i = 0$ otherwise, such that the sample size $n = \sum_{i=1}^{N} I_i$. Except for the sampling indicator I that is always in upper case, throughout the chapter the small cases represent the realization of the upper cases unless otherwise stated. Both the density and probability mass are denoted with $f(.)$ for those including the variables and $\varphi(.)$ for the functions only showing the parameters. Unknown parameters are expressed in Greek letters. The notations $\mathscr{E}[.]$ and $\mathscr{V}[.]$ are reserved for expectation and variance. An explanation may accompany bold letters to indicate as to whether they refer to a vector or a matrix.

The review of available studies in the previous chapter have covered the most influential procedures for missing data management and several theoretical approaches for incorporating sampling design into the parameters estimation. We are now ready to develop the models of interest, which are focused around survey data with missing covariates, and obtain the possible

solutions for parameters estimation in such presence of incomplete observations. Both survey selection and partially observed covariates are missing data problems, but it is important to recognize them as separate entities because the former is governed by sampling design while the latter is likely caused by non-responses or entry errors.

This chapter is structured as the following. Sections 2 and 3 introduce the models for survey sampling without and with missing covariates in certain practical situations. The assumptions about sampling design and missing data mechanisms are considered along with their consequences in the model likelihood. There are in particular three major classes of survey data with non-ignorable missing covariates that become the focus of model development: first, data with covariates observable only among samples; second, those of which covariates information available in both samples and non-samples; and third, the hybrid class where some covariates are observable on all subjects regardless of sampling status, but the rest are only available among samples. In Section 4, the parameters estimation is formulated. The discussion includes general strategies to computation and the form of variance estimators. Section 5 provides details of computation algorithms.

3.2 Survey Sampling without Missing Covariates

Let the joint distribution between the random variables Y_i and $\mathbf{X}_i$ be $f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_i; \boldsymbol{\alpha})$. We further assume that it is reasonable to model the covariates vector $\mathbf{X}_i$ as a product of one-dimensional conditional distributions. Specifically,

$$f(\mathbf{x}_i; \boldsymbol{\alpha}) = \left\{ \prod_{p=2}^{P} f(x_{ip} \mid \mathbf{x}_{ip}^{-}; \alpha_p) \right\} f(x_{i1}; \alpha_1), \tag{3.1}$$

where $\mathbf{x}_{ip}^{-} = (x_{i(p-1)}, x_{i(p-2)}, \ldots, x_{i1})$. In terms of the sampling model p_i, two situations are considered:

Situation 1 There is enough evidence to suggest $p_i = f(I_i = 1|\mathbf{x}_i^{(D)}; \boldsymbol{\delta})$, where $\mathbf{x}_i^{(D)}$ are part of $\mathbf{X}_i$ that constitute the design variables. Hence $\mathbf{X}_i = (\mathbf{x}_i^{(D)}, \mathbf{x}_i^{(D-)})$, where superscript D− denoting complement of D.

Situation 2 The functional form of p_i is not known for all i, but the quantities for p_i are available for $i \in \mathscr{S}$.

One of the following cases may happen in the survey sampling for estimating the characteristics of Y based on $(Y, \mathbf{X})$ relationship.

Case 1. Only $\mathbf{X}_i$ for those i with $I_i = 1$ (or, $i \in \mathscr{S}$) are observable.

By treating the sampling as a missing data problem, this case corresponds to missing both outcome and covariates. The likelihood function $f(\mathbf{I}, \mathbf{Y}, \mathbf{X}; \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})$, assuming each i independent, is equals to

$$\prod_{i=1}^{N} \left\{ p_i f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_i; \boldsymbol{\alpha}) \right\}^{I_i} \left\{ \int\int (1 - p_i) f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_i; \boldsymbol{\alpha}) dy_i d\mathbf{x}_i \right\}^{1-I_i}, \tag{3.2}$$

which leads to the log-likelihood score for $\boldsymbol{\beta}$

$$\begin{aligned} \sum_{i=1}^{N} \Bigg\{ & I_i \frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) \; + \\ & (1 - I_i) \frac{\int\int (1 - p_i) \frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_i; \boldsymbol{\alpha}) dy_i d\mathbf{x}_i}{\int\int (1 - p_i) f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_i; \boldsymbol{\alpha}) dy_i d\mathbf{x}_i} \Bigg\}. \end{aligned} \tag{3.3}$$

Under Situation 1 of p_i, however, Equation 3.2 reduces to

$$\prod_{i=1}^{N} \left\{ p_i f(y_i \mid x_i; \beta) f(x_i; \alpha) \right\}^{I_i} \left\{ \int (1 - p_i) f(x_i; \alpha) dx_i \right\}^{1-I_i}$$

and thus the second part of the log-likelihood score for β in Equation 3.3 is a zero function. This is not necessarily the case for Situation 2. When the functional form of p_i is not known for i such that $I_i = 0$, the second part of Equation 3.3 cannot be computed. To circumvent this problem, an alternative approach such as weighting can be used. In particular, the second part of Equation 3.3 is approximated by

$$\sum_{i=1}^{N} \frac{I_i(1 - p_i)}{p_i} \frac{\partial}{\partial \beta} \log f(y_i \mid x_i; \beta),$$

which leads to the simple weighting form

$$\sum_{i=1}^{N} \frac{I_i}{p_i} \frac{\partial}{\partial \beta} \log f(y_i \mid x_i; \beta). \tag{3.4}$$

Case 2. $X_i, i = 1, \cdots, N$ are all observable.

This case corresponds to missing outcome. The likelihood function is different from that of Case 1 in terms of the expression for $I_i = 0$. In particular, the likelihood becomes

$$\prod_{i=1}^{N} \left\{ p_i f(y_i \mid x_i; \beta) f(x_i; \alpha) \right\}^{I_i} \left\{ f(x_i; \alpha) \int (1 - p_i) f(y_i \mid x_i; \beta) dy_i \right\}^{1-I_i},$$

Similarly with Case 1, nevertheless, the log-likelihood score for $\boldsymbol{\beta}$ in Case 2 Situation 1 is only computed for samples ($I_i = 1$) as the score is zero for non-samples ($I_i = 0$). For Situation 2, the log-likelihood score for $\boldsymbol{\beta}$ can be approximated using the weighting approach of Equation 3.4.

Case 3. Some of the components of $\mathbf{X}$, namely $\mathbf{X}_{i1}$, are in Case 1, while the rest, $\mathbf{X}_{i2}$, are in case 2.

In the missing-data perspective, one has missing outcome and partial covariates if the data from survey sampling turn out to be this case. Hence, the likelihood function is

$$\prod_{i=1}^{N} \left\{ p_i f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_i; \boldsymbol{\alpha}) \right\}^{I_i} \left\{ \int\int (1 - p_i) f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_i; \boldsymbol{\alpha}) dy_i d\mathbf{x}_{i1} \right\}^{1-I_i}.$$

The log-likelihood score for $\boldsymbol{\beta}$ is thus zero for $I_i = 0$ under Situation 1, and under Situation 2 it can be computed using the form of Equation 3.4.

3.3 Survey Sampling with Missing Covariates

Suppose that aside from the missing data problem due to the probability sampling, $K \leq P$ elements of $\mathbf{X}$ are also subject to missing values. Let $\mathbf{X}_{i,\text{obs}}$ represents the observed part of $\mathbf{X}_i$ and $\mathbf{X}_{i,\text{mis}}$ the missing part. Subscript $k = 1, \cdots, K$ indexes the missing data indicator R_{ik} in $\mathbf{R}_i$. We will also use the subscript "(p)" for $\mathbf{R}_i$ to conveniently indicate the pth element of $\mathbf{X}_i$ it refers to. Thus, $R_{i(p)} = 1$ denotes X_{ip} is observed, and $R_{i(p)} = 0$ indicates X_{ip} is missing. To accommodate the missing data pattern, let $I_{\{\mathbf{r}_i = c\}} = 1$ denotes the missing data for subject i conforms to pattern $c \in \{1, \cdots, C\}$ and $I_{\{\mathbf{r}_i = c\}} = 0$ otherwise. We consider a functional form for

the missing data mechanism that resembles the joint model of covariates in Equation 3.1. This leads to

$$f(\mathbf{r}_i \mid y_i, \mathbf{x}_i; \boldsymbol{\gamma}) = \left\{ \prod_{k=2}^{K} f(r_{ik} \mid \mathbf{r}_{ik}^{-}, y_i, \mathbf{x}_i; \boldsymbol{\gamma}_k) \right\} f(r_{i1} \mid y_i, \mathbf{x}_i; \boldsymbol{\gamma}_1), \tag{3.5}$$

$\mathbf{r}_{ik}^{-} = (r_{i(k-1)}, r_{i(k-2)}, \ldots, r_{i1})$. Manifestation of these additional assumptions to the previous likelihood functions is described below.

3.3.1 Case 1: Covariates Observed only among Samples

We suppose now that $\mathbf{X}_i$ are only observed among $i \in \mathscr{S}$ and $\mathbf{X}_{i\in\mathscr{S}} = (\mathbf{X}_{i,\text{obs}}, \mathbf{X}_{i,\text{miss}})'$, where the missing data mechanisms are assumed to be non-ignorable. The likelihood function as expressed by Equation 3.2 has to be modified for accommodating the missing data mechanisms. With each i is independent to each other, $\mathscr{L}(\boldsymbol{\theta}) = f(\mathbf{I}, \mathbf{R}, \mathbf{Y}, \mathbf{X}; \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ is equals to

$$\begin{aligned} &\prod_{i=1}^{N} \left\{ \prod_{c=1}^{C} \left[\int p_i f(\mathbf{r}_i \mid y_i, \mathbf{x}_i; \boldsymbol{\gamma}) f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_i; \boldsymbol{\alpha}) d\mathbf{x}_{i,\text{mis}} \right]^{I_{\{r_i=c\}}} \right\}^{I_i} \times \\ &\quad \left\{ \iint (1 - p_i) f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_i; \boldsymbol{\alpha}) dy_i d\mathbf{x}_i \right\}^{1-I_i}, \end{aligned} \tag{3.6}$$

which, under the assumption that all the parameters are distinct, gives rise to the following log-likelihood score function for $\boldsymbol{\beta}$ under Situation 1

$$\begin{aligned} &\sum_{i=1}^{N} I_i \sum_{c=1}^{C} I_{\{r_i=c\}} \int \frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_{i,\text{miss}} \mid I_i = 1, \mathbf{r}_i, y_i, \mathbf{x}_{i,\text{obs}}) d\mathbf{x}_{i,\text{obs}} \\ &= \sum_{i=1}^{n} \sum_{c=1}^{C} I_{\{r_i=c\}} \mathscr{E} \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) \mid \mathbf{r}_i, y_i, \mathbf{x}_{i,\text{obs}} \right]. \end{aligned} \tag{3.7}$$

When the functional form of p_i is not known (Situation 2), the log-likelihood score function is weighted by the survey weights I_i/p_i such that it becomes

$$\sum_{i=1}^{N} \frac{I_i}{p_i} \sum_{c=1}^{C} I_{\{r_i=c\}} \mathscr{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) \mid I_i = 1, \boldsymbol{r}_i, y_i, \boldsymbol{x}_{i,\text{obs}}\right]. \tag{3.8}$$

3.3.2 Case 2: Covariates Observed on Both Samples and Non-Samples

Let $\mathbf{X}_i$ are observable for all $i = 1, \cdots, N$, but also subject to missing values. The likelihood function is therefore

$$\begin{aligned} \prod_{i=1}^{N} \Bigg\{ \prod_{c=1}^{C} \left[\int p_i f(\boldsymbol{r}_i \mid y_i, \boldsymbol{x}_i; \boldsymbol{\gamma}) f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) f(\boldsymbol{x}_i; \boldsymbol{\alpha}) d\boldsymbol{x}_{i,\text{mis}} \right]^{I_{\{r_i=c\}}} \Bigg\}^{I_i} \times \\ \left\{ \prod_{c=1}^{C} \left[\iint (1-p_i) f(\boldsymbol{r}_i \mid y_i, \boldsymbol{x}_i; \boldsymbol{\gamma}) f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) f(\boldsymbol{x}_i; \boldsymbol{\alpha}) dy_i d\boldsymbol{x}_{i,\text{mis}} \right]^{I_{\{r_i=c\}}} \right\}^{1-I_i}. \end{aligned} \tag{3.9}$$

The structure of this likelihood requires that both samples ($I_i = 1$) and non-samples ($I_i = 0$) contribute the log-likelihood score for $\boldsymbol{\beta}$. Under Situation 1, $\frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\theta})$, $\boldsymbol{\theta}' = (\boldsymbol{\delta}', \boldsymbol{\gamma}', \boldsymbol{\beta}', \boldsymbol{\alpha}')$, is equal to

$$\begin{aligned} \sum_{i=1}^{N} \Bigg[\Bigg\{ I_i \sum_{c=1}^{C} I_{\{r_i=c\}} \mathscr{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) \mid I_I = 1, \boldsymbol{r}_i, y_i, \boldsymbol{x}_{i,\text{obs}}\right] \Bigg\} \\ + \Bigg\{ (1-I_i) \sum_{c=1}^{C} I_{\{r_i=c\}} \mathscr{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) \mid I_i = 0, \boldsymbol{r}_i, \boldsymbol{x}_{i,\text{obs}}\right] \Bigg\} \Bigg], \end{aligned} \tag{3.10}$$

and under Situation 2, it is Equation 3.8.

3.3.3 Case 3: Covariates are A Mixture of Case 1 and Case 2

We consider, on the other hand, the case of which the components $\mathbf{X}_i^{(1)}$ of $\mathbf{X}$ are in Case 1, while the rest $\mathbf{X}_i^{(2)}$ are in Case 2. In addition, some or all $\mathbf{X}_i$ are subject to missing values. The likelihood function is thus

$$\begin{aligned}\prod_{i=1}^{N}\Bigg\{\prod_{c=1}^{C}\left[\int p_i f(\mathbf{r}_i \mid y_i, \mathbf{x}_i; \boldsymbol{\gamma}) f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_i; \boldsymbol{\alpha}) d\mathbf{x}_{i,\text{mis}}\right]^{I_{\{\mathbf{r}_i=c\}}}\Bigg\}^{I_i} \times \\ \left\{\prod_{c=1}^{C}\left[\iiint (1-p_i) f(\mathbf{r}_i \mid y_i, \mathbf{x}_i; \boldsymbol{\gamma}) f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) f(\mathbf{x}_i; \boldsymbol{\alpha}) dy_i d\mathbf{x}_i^{(1)} d\mathbf{x}_{i,\text{miss}}^{(2)}\right]^{I_{\{\mathbf{r}_i=c\}}}\right\}^{1-I_i}.\end{aligned} \tag{3.11}$$

The log-likelihood score for $\boldsymbol{\beta}$ under Situation 1 looks similar to Equation 3.10, except that the second term is

$$(1-I_i)\sum_{c=1}^{C} I_{\{\mathbf{r}_i=c\}} \mathscr{E}\left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) \mid I_i = 0, \mathbf{r}_i, \mathbf{x}_{i,\text{obs}}^{(2)}\right]. \tag{3.12}$$

The score function for Situation 2 is again Equation 3.8.

3.4 Parameters Estimation

It is obvious from Equation 3.7, Equation 3.8, Equation 3.10, and Equation 3.12 that solution for the parameters of interest $\boldsymbol{\beta}$ in the presence of non-ignorable missing covariates has to be derived through modeling the joint distribution $f(\mathbf{I}, \mathbf{R}, \mathbf{Y}, \mathbf{X}; \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})$ either in the form of Equation 3.6, Equation 3.9, or Equation 3.11. I obtain in this section the maximum likelihood estimators (MLEs) for $\boldsymbol{\theta}' = (\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha})'$. The solutions do not in general have a closed form, and thus, one needs to apply iterative methods. Assuming certain regularity conditions hold, I use the EM algorithm (66) in a fashion nearly analogous to Ibrahim et al. (2; 70; 72), Lipsitz et al. (71; 73), and Wei and Tanner (85) to find the solution for $\boldsymbol{\theta}$.

Let $\mathbf{Z} \equiv (\mathbf{I}, \mathbf{R}, \mathbf{Y}, \mathbf{X})$ with realization $\boldsymbol{z}$. Hence, $\boldsymbol{z}_i = (I_i, \mathbf{r}_i', y_i, \boldsymbol{x}_i')'$. We will also use $\boldsymbol{z}_{i;I_i\in\{0,1\}}$ to denote $\boldsymbol{z}_i$ with $I_i \in \{0, 1\}$, and $\boldsymbol{z}_{i,obs;I_i=1} = (I_i = 1, \mathbf{r}_i', y_i, \boldsymbol{x}_{i,obs}')'$. Suppose that the functional form of p_i is $f(I_i = 1|\boldsymbol{x}_i^{(D)}; \boldsymbol{\delta})$ (Situation 1), then the E-step of the EM algorithm computes $Q(\theta \mid \theta^{(t)})$, t denoting iteration, which equals to

$$
\begin{aligned}
\text{Case 1:}\quad & \textstyle\sum_{i=1}^{N} \Big\{ I_i \sum_{c=1}^{C} I_{\{r_i=c\}} \mathscr{E}\Big[\ell(\theta; \boldsymbol{z}_{i;I_i=1}) \mid \boldsymbol{z}_{i,obs;I_i=1}; \theta^{(t)}\Big] \\
& + \ (1 - I_i) \mathscr{E}\Big[\ell(\boldsymbol{\delta}, \boldsymbol{\alpha}; I_i = 0, \boldsymbol{x}_i) \mid I_i = 0; \theta^{(t)}\Big] \Big\} \\
\text{Case 2:}\quad & \textstyle\sum_{i=1}^{N} \Big\{ I_i \sum_{c=1}^{C} I_{\{r_i=c\}} \mathscr{E}\Big[\ell(\theta; \boldsymbol{z}_{i;I_i=1}) \mid \boldsymbol{z}_{i,obs;I_i=1}; \theta^{(t)}\Big] \\
& + \ (1 - I_i) \textstyle\sum_{c=1}^{C} I_{\{r_i=c\}} \mathscr{E}\Big[\ell(\theta; \boldsymbol{z}_{i;I_i=0}) \mid I_i = 0, \mathbf{r}_i, \boldsymbol{x}_{i,obs}; \theta^{(t)}\Big] \Big\} \\
\text{Case 3:}\quad & \textstyle\sum_{i=1}^{N} \Big\{ I_i \sum_{c=1}^{C} I_{\{r_i=c\}} \mathscr{E}\Big[\ell(\theta; \boldsymbol{z}_{i;I_i=1}) \mid \boldsymbol{z}_{i,obs;I_i=1}; \theta^{(t)}\Big] \\
& + \ (1 - I_i) \textstyle\sum_{c=1}^{C} I_{\{r_i=c\}} \mathscr{E}\Big[\ell(\theta; \boldsymbol{z}_{i;I_i=0}) \mid I_i = 0, \mathbf{r}_i, \boldsymbol{x}_{i,obs}^{(2)}; \theta^{(t)}\Big] \Big\}.
\end{aligned}
\tag{3.13}
$$

Assuming $\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ are distinct,

$$
\ell(\theta; \boldsymbol{z}_i) = \ell(\boldsymbol{\delta}; I_i, \boldsymbol{x}_i^{(D)}) + \ell(\boldsymbol{\gamma}; \mathbf{r}_i, y_i, \boldsymbol{x}_i) + \ell(\boldsymbol{\beta}; y_i, \boldsymbol{x}_i) + \ell(\boldsymbol{\alpha}; \boldsymbol{x}_i), \tag{3.14}
$$

where the adjustment for less parameters like $\ell(\boldsymbol{\delta}, \boldsymbol{\alpha}; I_i = 0, \boldsymbol{x}_i)$ in Case 1 is straightforward. We further denote the components of $\mathbf{Z}$ that are subject to missing values as $\mathbf{V} = (\mathbf{Y}, \mathbf{X})$ with realization $\boldsymbol{\nu}$. Accordingly, $\boldsymbol{\nu}_{i,obs}$ denote the observed part of $\boldsymbol{\nu}$ in individual i, and $\boldsymbol{\nu}_{i,mis}$ the

missing part. Expectation of (Equation 3.14) with respect to the conditional distribution of the missing variables, that is, the Q function of (Equation 3.13), has the following general form

$$\int \ell(\boldsymbol{\theta}; \boldsymbol{z}_i) f(\boldsymbol{\nu}_{i,mis} \mid \boldsymbol{z}_{i,obs}; \boldsymbol{\theta}) d\boldsymbol{\nu}_{i,mis} = \int \ell(\boldsymbol{\theta}; \boldsymbol{z}_i) \frac{f(\boldsymbol{z}_i; \boldsymbol{\theta})}{\int f(\boldsymbol{z}_i; \boldsymbol{\theta}) d\boldsymbol{\nu}_{i,mis}} d\boldsymbol{\nu}_{i,mis}, \tag{3.15}$$

$\boldsymbol{z}_i = (I_i, \boldsymbol{r}_i', \boldsymbol{\nu}_i')'$, and $\boldsymbol{z}_{i,obs} = (I_i, \boldsymbol{r}_i', \boldsymbol{\nu}_{i,obs}')'$. To perform the integration in Equation 3.15 when $\boldsymbol{\nu}_{i,mis}$ are all continuous, Tanner and Wong (82), followed by Wei and Tanner (85), and Gelfand and Smith (38) suggested the use of Monte Carlo based sampling. Theoretical details and proof of convergence for this approach is provided in Tanner and Wong's article. For the same situation in GLMs where $\boldsymbol{\nu}_{i,mis} \equiv \boldsymbol{x}_{i,mis}$ (that is, only the covariates that have incomplete observations), Ibrahim and Weisberg (125) utilized a Gaussian quadrature. Ibrahim (70) also introduced what he called the EM algorithm by the method of weight to handle Equation 3.15 in GLMs of which the elements of $\boldsymbol{\nu}_{i,mis}$ consist only of categorical $\boldsymbol{x}_{mis}$. His approaches involved data augmentation using all possible values of the variables with missingness. Extension of Ibrahim's method to GLMs with mixed categorical and continuous $\boldsymbol{x}_{i,mis}$ is discussed by Lipsitz et al. (71; 73) and Ibrahim et al. (72; 2).

Maximization of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ in the M-step can be performed using, for instance, the Newton-Raphson technique, where

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} - \left\{ \left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} Q(\boldsymbol{\theta}) \right]^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta}) \right\}_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(t)}}. \tag{3.16}$$

The E-step and M-step are then repeated until convergence. Further details of how I approach the EM algorithm computation are described in the next section.

Implementation of the EM algorithm when the functional form of p_i is not known (Situation 2) follows the same principal as it is under Situation 1. The basic difference is that the computation of Q function only entails the data where $I_i = 1$, that is, $z_{i;I_i=1}$ in the current notation.

3.5 Variance Formula

I derive the estimated covariance matrix for $\hat{\theta}$ under Situation 1 through the Louis' method (86) shown in Equation 2.3. This method has been recommended in several studies of missing covariates where the parameters are estimated within ML framework (70; 72; 74; 2). Suppose that $\hat{\theta}$ is the estimate of θ at the convergence of the EM algorithm. Let $\ell(\theta; z_i)$ be the simplified form of the log-likelihood of fully observed data for Case 1, 2, and 3 under Situation 1, and $z_{i,obs}$ the relevant observed parts of z_i. Application of the Louis' method involves the computation of the information matrix for $\hat{\theta}$, where

$$
\begin{aligned}
&\mathcal{I}(\hat{\theta}) = \\
&-\sum_{i=1}^{N} \mathcal{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\ell(\theta; z_i)\middle| z_{i,obs}; \hat{\theta}\right] \\
&-\sum_{i=1}^{N} \mathcal{E}\left[\left\{\frac{\partial}{\partial\theta}\ell(\theta; z_i)\right\}\left\{\frac{\partial}{\partial\theta}\ell(\theta; z_i)\right\}'\middle| z_{i,obs}; \hat{\theta}\right] \\
&+\left\{\sum_{i=1}^{N} \mathcal{E}\left[\frac{\partial}{\partial\theta}\ell(\theta; z_i)\middle| z_{i,obs}; \hat{\theta}\right]\mathcal{E}\left[\frac{\partial}{\partial\theta}\ell(\theta; z_i)\middle| z_{i,obs}; \hat{\theta}\right]'\right\}.
\end{aligned}
\tag{3.17}
$$

An adjustment of Equation 3.17 is needed under Situation 2. To see this, let $\boldsymbol{\tau} = (\boldsymbol{\gamma}', \boldsymbol{\beta}', \boldsymbol{\alpha}')'$, that is, $\boldsymbol{\theta}$ minus the parameters for the sampling design $\boldsymbol{\delta}$ shown for Situation 1. Suppose that $\boldsymbol{\tau}^w$ is the solution of

$$\sum_{i=1}^{N} \frac{I_i}{p_i} \frac{\partial}{\partial \boldsymbol{\tau}} \ell(\boldsymbol{\tau}; z_i) = 0. \tag{3.18}$$

Godambe and Thompson (101) showed that $\boldsymbol{\tau}^w$ is consistent for $\boldsymbol{\tau}$ if $\mathscr{E}\left[\frac{I_i}{p_i} \frac{\partial}{\partial \boldsymbol{\tau}} \ell(\boldsymbol{\tau}; z_i)\right] = 0$, where the expectation is with respect to both the sampling design and the model. Under certain regularity conditions, the expansion of Equation 3.18 to the second term of Taylor series leads to

$$\sqrt{N}(\boldsymbol{\tau}^w - \boldsymbol{\tau}) = \left[\frac{1}{N} \sum_{i=1}^{N} \frac{I_i}{p_i} \frac{\partial^2}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}'} \ell(\boldsymbol{\tau}; z_i)\right]^{-1} \left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{I_i}{p_i} \frac{\partial}{\partial \boldsymbol{\tau}} \ell(\boldsymbol{\tau}; z_i)\right].$$

Therefore, assuming the finite population correction (FPC) $= \sqrt{\frac{N-n}{N-1}}$ can be ignored, the asymptotic covariance of $\boldsymbol{\tau}^w$ can be consistently estimated by

$$\hat{J}^{-1} \hat{V} \hat{J}^{-1}, \tag{3.19}$$

nothing that $\hat{J} = \sum_{i=1}^{N} \frac{I_i}{p_i} \frac{\partial^2}{\partial \boldsymbol{\tau} \partial \boldsymbol{\tau}'} \ell(\boldsymbol{\tau}; z_i)$, setting $\boldsymbol{\tau} = \hat{\boldsymbol{\tau}}$, and $\hat{V}$ is derived from setting $\boldsymbol{\tau} = \hat{\boldsymbol{\tau}}$ for V where

$$
\begin{aligned}
V &\equiv \mathscr{V}\left[\sum_{i=1}^{N} \frac{I_i}{p_i}\frac{\partial}{\partial \boldsymbol{\tau}}\ell(\boldsymbol{\tau}; z_i)\right] \\
&= \sum_{i=1}^{N} \mathscr{V}\left[\frac{I_i}{p_i}\frac{\partial}{\partial \boldsymbol{\tau}}\ell(\boldsymbol{\tau}; z_i)\right], \text{ due to independence assumption} \\
&= \sum_{i=1}^{N} \left\{\mathscr{E}_{\boldsymbol{\tau}}\left[\mathscr{V}_{\delta}\left[\frac{I_i}{p_i}\frac{\partial}{\partial \boldsymbol{\tau}}\ell(\boldsymbol{\tau}; z_i)\right]\right] + \mathscr{V}_{\boldsymbol{\tau}}\left[\mathscr{E}_{\delta}\left[\frac{I_i}{p_i}\frac{\partial}{\partial \boldsymbol{\tau}}\ell(\boldsymbol{\tau}; z_i)\right]\right]\right\} \\
&= \sum_{i=1}^{N} \left\{\mathscr{E}_{\boldsymbol{\tau}}\left[\frac{p_i(1-p_i)}{p_i^2}\left[\frac{\partial}{\partial \boldsymbol{\tau}}\ell(\boldsymbol{\tau}; z_i)\right]\left[\frac{\partial}{\partial \boldsymbol{\tau}}\ell(\boldsymbol{\tau}; z_i)\right]'\right] + \mathscr{V}_{\boldsymbol{\tau}}\left[\frac{\partial}{\partial \boldsymbol{\tau}}\ell(\boldsymbol{\tau}; z_i)\right]\right\} \\
&= \sum_{i=1}^{N} \left\{\frac{1}{p_i}\mathscr{E}_{\boldsymbol{\tau}}\left[\left[\frac{\partial}{\partial \boldsymbol{\tau}}\ell(\boldsymbol{\tau}; z_i)\right]\left[\frac{\partial}{\partial \boldsymbol{\tau}}\ell(\boldsymbol{\tau}; z_i)\right]'\right]\right\},
\end{aligned}
\tag{3.20}
$$

where $\mathscr{E}_{\boldsymbol{\tau}}$ and $\mathscr{V}_{\boldsymbol{\tau}}$ denote the expectation and variance, respectively, with respect to the model $f(z_i; \boldsymbol{\tau})$, and $\mathscr{E}_{\delta}$ and $\mathscr{V}_{\delta}$ respectively denote the expectation and variance with respect to the sampling design. Note that with the presence of missing values, the Hessian matrix $\left[\frac{\partial}{\partial \boldsymbol{\tau}}\ell(\boldsymbol{\tau}; z_i)\right]\left[\frac{\partial}{\partial \boldsymbol{\tau}}\ell(\boldsymbol{\tau}; z_i)\right]'$ becomes the Louis information matrix (similar to Equation 3.17, except for the components δ). Also, $\sum_{i=1}^{N} \frac{1}{p_i}$ in the survey sample is estimated by $\sum_{i=1}^{N} \frac{I_i}{p_i^2}$, as $\mathscr{E}_{\delta}[I_i] = p_i$.

Equation 3.20, however, assumes that there is neither stratification nor clustering. In what follows, I provide the variance estimator accommodating a complex survey design. The general form of the covariance estimate is still Equation 3.19, and for all sampling designs $\hat{J} = \sum_{i=1}^{N} \frac{I_i}{p_i}\frac{\partial^2}{\partial \boldsymbol{\tau}\partial \boldsymbol{\tau}'}\ell(\boldsymbol{\tau}; z_i)$. The difference is in the expression for $\hat{V}$. Suppose that the samples

are drawn from H strata, where each stratum $h = 1, \cdots, H$ has J_h clusters. At observation i of cluster $j =, 1, \cdots, J_h$ in stratum h, let

$$\mathbf{U}_{ijh} = \frac{I_{ijh}}{p_{ijh}} \mathscr{E}_{\boldsymbol{\tau}} \left[\frac{\partial}{\partial \boldsymbol{\tau}} \ell(\boldsymbol{\tau}; \boldsymbol{z}_{ijh}) \right],$$

and at cluster j

$$\mathbf{D}_{jh} = \sum_{i=1}^{n_{j_h}} \mathbf{U}_{ijh} - \frac{\sum_{i=1}^{n_{j_h}} \sum_{j=1}^{J_h} \mathbf{U}_{ijh}}{J_h}.$$

Therefore, the expression for $\hat{V}$ can be obtained by setting $\boldsymbol{\tau} = \hat{\boldsymbol{\tau}}$ for $V_{(CS)}$ where

$$V_{(CS)} = \sum_{h=1}^{H} \frac{\sum_{j=1}^{J_h} \mathbf{D}_{jh} \mathbf{D}'_{jh}}{J_h - 1}, \tag{3.21}$$

and CS stands for "complex survey".

3.6 Computation Algorithm

I now devise the algorithm for computing the parameter estimates. This computation extends Ibrahim et al.'s (72; 2) and Lipsitz et al.'s (71; 73) approach for handling $\mathbf{X}_{\text{mis}}$ to the settings where the missing variables $\boldsymbol{V}$ may include both $\boldsymbol{Y}$ and $\mathbf{X}$. Such approach has an appeal of allowing certain flexibility in missing data management when the incomplete multivariate dataset is a mixture of continuous and categorical variables. In particular, one specifies the joint distribution as a product of univariate densities. The goal is to reduce possible nuisance parameters (70; 72; 2), and also by breaking the problem of joint modeling into a series of univariate models, it enables the functional specification of each variable to proceed based on

the corresponding type of the variable (continuous, categorical, ordered categorical) (87). Since $f(\boldsymbol{\nu}_{i,mis} \mid \boldsymbol{z}_{i,obs}; \theta^{(t)})$ is basically the weight of $\ell(\theta; \boldsymbol{z}_i)$ in Equation 3.15, Ibrahim named this approach the EM algorithm by the method of weight (70). However, his proposal is actually an extension of data augmentation technique by Tanner and Wong (82). The principal difference is that he modified Tanner and Wong's weight $1/m$, m is the number of imputed values for $\boldsymbol{\nu}_{i,mis}$, with $f(\boldsymbol{\nu}_{i,mis} \mid \boldsymbol{z}_{i,obs}; \theta^{(t)})$, and use the exact values of the incomplete variables to augment data.

Let us keep all notations we already have from the early parts of this chapter. The proposed algorithm then proceeds as follows:

1. Start with initial values $\theta^{(0)}$.

   At iteration t:

2. Augment data using:

   – All possible or most likely values if categorical or finite discrete

   – Gauss-Hermite quadrature nodes if continuous

3. E-step: Estimate $Q(\theta|\theta^{(t)})$ by

$$\sum_{i=1}^{N} \sum_{j=1}^{J} \ell(\theta; \boldsymbol{z}_i) f(\boldsymbol{\nu}_{i,mis(j)} \mid \boldsymbol{z}_{i,obs}; \theta^{(t)}) \tag{3.22}$$

if all $\boldsymbol{\nu}_i = (\boldsymbol{\nu}_{i,obs}, \boldsymbol{\nu}_{i,mis})$ are categorical or finite discrete, where j indexes the pattern of $\boldsymbol{\nu}_{i,mis}$; or, by

$$\frac{1}{\sqrt{\pi}} \sum_{i=1}^{N} \sum_{q_1=1}^{Q} \cdots \sum_{q_H=1}^{Q} \ell(\theta; \boldsymbol{z}_{iq}^*) f(a_{iq_1}, \cdots, a_{iq_H} \mid \boldsymbol{z}_{i,obs}; \theta^{(t)}) w_{iq_1} \cdots w_{iq_H}, \tag{3.23}$$

if all $\boldsymbol{\nu}_i$ are continuous, where $\boldsymbol{\nu}_{i,mis}$ is in a transformed form $\boldsymbol{\nu}_i^*$, and $\boldsymbol{z}_{iq}^* = (\boldsymbol{z}_{i,obs}', \boldsymbol{\nu}_{iq}^{*\prime})'$, $\boldsymbol{\nu}_{iq}^* = (\nu_{iq_1}^*, \cdots, \nu_{iq_H}^*)'$, $\nu_{iq_h}^* = \mu_{\nu_{ih}} + \sqrt{2}\sigma_{\nu_{ih}} a_q$, $h = 1, \cdots, H$, $\mu_{\nu_{ih}}$ and $\sigma_{\nu_{ih}}$ are the mode and scale for ν_{ih}, $w_{iq_h} = w(a_{iq_h})$; a_{iq_h} and $w(a_{iq_h})$ are the abscissas and weights of the Hermite polynomial; or by

$$\frac{1}{\sqrt{\pi}} \sum_{i=1}^{N} \sum_{j=1}^{J} \sum_{q_1=1}^{Q} \cdots \sum_{q_H=1}^{Q} \ell(\theta; \boldsymbol{z}_{iq}^{**}) f(\boldsymbol{\nu}_{i,mis(j)}^{(cat)}, a_{iq_1}, \cdots, a_{iq_H} \mid \boldsymbol{z}_{i,obs}; \theta^{(t)}) w_{iq_1} \cdots w_{iq_H}, \tag{3.24}$$

$\boldsymbol{z}_{iq}^{**} = (\boldsymbol{z}_{i,obs}', \boldsymbol{\nu}_{i(j)}^{(cat)\prime}, \boldsymbol{\nu}_{iq}^{(cont)*\prime})'$, $\boldsymbol{\nu}_{iq}^{(cont)*} = (\nu_{iq_1}^*, \cdots, \nu_{iq_H}^*)'$, if $\boldsymbol{\nu}_i$ are a mixture of continuous and categorical (or finite discrete) variables.

4. M-step: Maximize $Q(\theta|\theta^{(t)})$ to obtain $\widehat{\theta}^{(t+1)}$ using Equation 3.16.

5. Repeat Steps 2 to 4 until convergence.

There are a few notes to complement the application of this computation algorithm. For categorical or finite discrete variables, ideally one uses all possible values. That may, however, pose a challenge in cases where the missing variable is assumed to follow a Poisson distribution, as the possible values run from 0 to ∞. One may then restrict the augmentation to those values around the mode and ignore others of which the empirical probability is acceptably low. My

observation during the simulation studies indicates that such a decision lead to fairly satisfying results with a good trade off between computational cost and accuracy of the estimated parameters. It is recommended that one takes all integers within the range of values served as the investigator's choice of lower and upper bounds to be used for augmentation. Another note is regarding continuous variables. An adaptive or non-adaptive Gaussian quadrature may be used. For the latter, a choice between $5-10$ nodes appear to provide acceptable accuracy in my simulations. The other option for imputing incomplete continuous variables is to use a Gibbs sampler or any Monte Carlo based sampling (71; 73; 72; 2). In general, their application demands a more expensive computation than Gaussian quadrature because of the need to assure independent random draws from the joint distribution. However, if it is computationally feasible to use Gibbs sampler for sampling $\boldsymbol{\nu}_{\mathrm{i,mis}}$ from their conditional distribution given $\boldsymbol{z}_{\mathrm{i,obs}}$ — that is, not only for continuous $\boldsymbol{x}_{\mathrm{i,mis}}$ as suggested in Ibrahim et al. (72; 2) and Lipsitz et al. (71; 73) — then that will add substantial flexibility to this algorithm.

CHAPTER 4

SIMULATION STUDIES

Survey dataset with missing variables is unique with respect to the structure of missingness. It represents the settings where there are both sample and variable selections. That is, the elements of data matrix are missing row wise and column wise. In this chapter, I simulate several datasets having a certain scenario of sample selection and missing variables as described in Chapter 3 to evaluate the performance of the proposed method. The term "observable" is used extensively throughout the chapter to refer to the conditions where an element in data matrix is not subject to row wise missingness. That is, the presence or absence of the element is independent of sample selection. It may at the same time be subject to column wise missingness or variable selection. To emphasize the focus of the proposed method development, I attach these conditions into the variables; thus, the variable is either observable or not observable (sample selection issue), missing or not missing (variable selection issue).

The simulation studies are presented in the following order: Section 1 provides the general setup of all simulations, which include the generation of variables of interest, sample selection, and missing covariate mechanisms; Section 2 describes the computation of parameter estimates based on the proposed and the competing methods, and how they will be compared; Section 3 depicts the simulations for Case 2 of the Chapter 3 formulation, that is, survey data with non-ignorable missing covariates of which the covariates are observable on both samples and non-samples; Section 4 entails the simulation studies of survey data having one part of the covariates

observed only among samples, while the other part observable regardless of sampling status but subject to non-ignorable missingness (Case 3); Section 5 demonstrates the performance of the proposed method for the case of survey data with non-ignorable missing covariates that are observed strictly among samples (Case 1); And finally, Section 6 discusses the overall findings and the limitations of the proposed method and simulation setting.

4.1 Simulation Setup and Notation

We consider a hypothetical population $\mathscr{F}$ having $N < \infty$ individuals at certain time point. Data for the outcome Y and the covariates of interest $X_2, \cdots, X_8$ are generated using the models demonstrated in Table II. Y is thought as a health-related outcome of non-negative discrete nature. Another variable, X_1, is also created to be one of the dominant sources of sampling variation. Overall, Table II shows that the covariates $X_1, \cdots, X_8$ are a mixture of categorical and continuous variables. The conditional distribution of $Y, X_5, \cdots, X_8$ is in the class of the generalized linear models (GLM), with a link function that represents the most natural choice in health and medical studies. These include log-link for Poisson, identity-link for Gaussian, and logit-link for binary dependent variables. The parameters of $X_1, \cdots, X_5$ are set to be $\mu_{(X_1)} = 50$, $\sigma^2_{(X_1)} = 25$, $\pi_{(X_2)} = 0.5$, $\lambda_{(X_3)} = 2$, $\mu_{(X_4)} = 5$, $\sigma^2_{(X_4)} = 1$, $\boldsymbol{\alpha} = (\alpha_0, \cdots, \alpha_4)' = (-1, 1, -1, 1)'$, and $\sigma^2_{(X_5)} = 1$. The GLM parameters of Y, X_6, X_7, X_8, however, are determined in a less straightforward way. Their quantities are derived by first fixing the expected values $\lambda_{(Y)} = 0.35$, $\pi_{(X_6)} = 0.7$, $\pi_{(X_7)} = 0.7$, and $\pi_{(X_8)} = 0.4$. Then, with brute force, it is found that $\boldsymbol{\beta} = 0.1 \times (1, -1, -1, -1, -1, -1, -1, 1)'$, $\boldsymbol{\gamma} = 0.1412139 \times (1, -1, -1, 1, 1)'$, $\boldsymbol{\iota} = (-0.25, 0.25, 0.2, 0.25, -0.25, 0.15)'$, and $\boldsymbol{\kappa} = (-0.25, 0.1 \times (-1, -1, 1, -1, -1, -1))'$, respec-

tively, produce an empirical average for Y, X_6, X_7, X_8 that is reasonably close to their expected value. One advantage of setting the categorical GLM parameters this way is the avoidance of unwanted quantities, like a proportion close to 0 or 1 in the binomial models. I use the same technique later to create the sampling fraction and missing data proportion that meet the simulation interests.

Selection of the sample set $\mathscr{S}$ from $\mathscr{F}$is simulated conditional on X_1, X_2, and X_8 under a binomial GLM model, that is, for observation $i = 1, \cdots, N$,

$$\text{logit } f(I_i = 1 | x_{i1}, x_{i2}, x_{i5}; \delta) = \delta_0 + \delta_1 X_{i1} + \delta_2 X_{i2} + \delta_3 X_{i5}, \tag{4.1}$$

where $I_i = 1$ indicates observation $i \in \mathscr{S}$, while $I_i = 0$ is otherwise. Thus the sample size $n = \sum_{i=1}^{N} I_i$. One may notice that the covariates of I_i consist of both completely and incompletely observed variables (see Table II). Each subject i is treated as independent to one another. In simulating Case 1 of the Chapter 3 formulation, $\pi_{(I|X_1,X_2,X_5)} \equiv \mathscr{E}\left[\mathbf{I}|\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_5; \boldsymbol{\delta}\right]$ is set to be 0.01, while for Case 2 and Case 3, it is 0.7. The latter sampling fraction is chosen primarily for computational convenience. That of Case 1, however, is generated to reasonably represents a traditional survey, where the sampling fraction is much smaller than the actual population size. Applying a similar technique to that for obtaining the GLM parameters of Y, X_6, X_7, and X_8, I choose $\delta = (-0.1, -0.1, 0.1, 0.2)'$ in the Case 1 simulation, and $(-0.5, 0.1, -1, -1)'$ in the Case 2 and Case 3 simulations. Two situations are then devised for the Case 1 simulation following the model development in the Methodology chapter: First, the true sampling model is known

and all information to compute the conditional probability are available. Second, the functional form of $f(I_i = 1|x_{i1}, x_{i2}, x_{i5}; \boldsymbol{\delta})$ as shown in Equation 4.1 is not known for all i, but its quantity p_i is available among the samples. In the studies of Case 2 and Case 3, the functional form of sample selection is always assumed known. There is no need to replicate the second situation of sampling information in the simulation of Case 2 and Case 3, since the log-likelihood score for the parameters of interest $\boldsymbol{\beta}$ is identical to that of Case 1, which is Equation 3.8.

The covariates $X_5, \cdots, X_8$ are then subjected to non-ignorable missing data. Let $\mathbf{R} = (\mathbf{R}_5, \cdots, \mathbf{R}_8)$ indicate their missingness, where $\mathbf{R}_k$ is a missing indicator vector for the corresponding $\mathbf{X}_k$, $k = 5, \cdots, 8$, $R_{ik} = 1$ represents the situation when X_{ik} are observed for individual i, and $R_{ik} = 0$ otherwise. $\mathbf{R}$ are generated for each i based on the following relational forms

$$
\begin{aligned}
R_{i5} &\sim \text{Bernoulli}\Big(\pi_{i(R_5|X_2,\cdots,X_8,Y)} = \text{expit}(\zeta_0 + \zeta_1 X_{i2} + \cdots + \zeta_7 X_{i8} + \zeta_8 Y_i)\Big) \\
R_{i6} &\sim \text{Bernoulli}\Big(\pi_{i(R_6|X_2,\cdots,X_8,Y,R_5)} = \text{expit}(\nu_0 + \nu_1 X_{i2} + \cdots + \nu_7 X_{i8} + \nu_8 Y_i + \nu_9 R_{i5})\Big) \\
R_{i7} &\sim \text{Bernoulli}\Big(\pi_{i(R_7|X_2,\cdots,X_8,Y,R_5,R_6)} = \text{expit}(\upsilon_0 + \upsilon_1 X_{i2} + \cdots + \upsilon_7 X_{i8} + \upsilon_8 Y_i + \\
&\qquad \upsilon_9 R_{i5} + \upsilon_{10} R_{i6})\Big) \\
R_{i8} &\sim \text{Bernoulli}\Big(\pi_{i(R_8|X_2,\cdots,X_8,Y,R_5,R_6,R_7)} = \text{expit}(\omega_0 + \omega_1 X_{i2} + \cdots + \omega_7 X_{i8} + \omega_8 Y_i + \\
&\qquad \omega_9 R_{i5} + \omega_{10} R_{i6} + \omega_{11} R_{i7})\Big).
\end{aligned}
\tag{4.2}
$$

To produce the missing proportions displayed in Table II, $\pi_{(R_k)}$ are fixed accordingly, such that the GLM parameters $\boldsymbol{\zeta}' = (\zeta_0, \cdots, \zeta_8)$, $\boldsymbol{\nu}' = (\nu_0, \cdots, \nu_9)$, $\boldsymbol{\upsilon}' = (\upsilon_0, \cdots, \upsilon_{10})$, and $\boldsymbol{\omega}' = (\omega_0, \cdots, \omega_{11})$ are $0.5245125 \times (-1, 1, -1, 1, -1, -1, 1, 1, 1)$, $0.177842 \times (1, 1, 1, 1, 1, -1, 1, 1, 1, 1)$,

$0.1400954 \times (-1, 1, 1, 1, 1, -1, -1, 1, 1, 1, 1)$, and $0.1 \times (-1, -1, -1, 1, 1, -1, 1, 1, -1, -1, 1, 1)$, respectively.

As it is already obvious, the notations follow those of the previous chapters, where small cases represent the realization of upper cases, except for the sampling indicator I that is always in upper case. Also, the density and probability mass are both denoted with $f(.)$ and the unknown parameters in Greek letters. $\mathscr{E}[.]$ and $\mathscr{V}[.]$ are respectively representing expectation and variance. Depending on the context, bold letters may indicate a vector or matrix.

4.2 Likelihood and Computation

Let $f(I_i, \mathbf{R}_i, Y_i, \mathbf{X}_i; \theta)$ be the joint distribution for observation i of the simulated data, where $\mathbf{R}_i' = (R_{i5}, \cdots, R_{i8})$, $\mathbf{X}_i' = (X_{i1}, \cdots, X_{i8})$, $\theta' = (\theta_X', \beta', \theta_R', \delta')$, $\theta_X' = (\mu_{(X_1)}, \sigma^2_{(X_1)}, \pi_{(X_2)}, \lambda_{(X_3)}, \mu_{(X_4)}, \sigma^2_{(X_4)}, \alpha', \gamma', \iota', \kappa')$, and $\theta_R' = (\zeta', \nu', \upsilon', \omega')$. Following Table II, Equation 4.1, and Equation 4.2, for the full dataset $f(I_i, \mathbf{R}_i, Y_i, \mathbf{X}_i; \theta)$ equals to Equation 4.3. Under the independence of i, the likelihood $\mathscr{L}(\theta) = \prod_{i=1}^{N} \mathscr{L}_i(\theta) = \prod_{i=1}^{N} f(I_i, \mathbf{R}_i, Y_i, \mathbf{X}_i; \theta)$. And based on Section 3.3 of Chapter 3, the likelihood becomes Equation 3.6 for Case 1, Equation 3.9 for Case 2, and Equation 3.11 for Case 3, with an important exception that the notation for α and γ in those expressions is replaced here with θ_X', and θ_R', respectively. If the relation expressed by Equation 4.1 is assumed unknown, a full likelihood estimation is not feasible. However, when additionally the quantities p_i is available for all i, a quasi-likelihood approach using survey weight such as Equation 3.8 can be computed. I generate this situation, which is termed Situation 2 in Chapter 3, during the simulation of Case 1.

$$\left(Y_i!X_{i3}!(2\pi)^{\frac{3}{2}}\sigma_{(X_1)}\sigma_{(X_4)}\sigma_{(X_5|X_2,\cdots,X_4)}\right)^{-1}\exp\left\{-\frac{1}{2}\left[\frac{(X_{i1}-\mu_{(X_1)})^2}{\sigma^2_{(X_1)}}+\frac{(X_{i4}-\mu_{(X_4)})^2}{\sigma^2_{(X_4)}}\right.\right.$$

$$\left.+\frac{(X_{i5}-(\alpha_0+\alpha_1X_2+\cdots+\alpha_3X_4))^2}{\sigma^2_{(X_5|X_2,\cdots,X_4)}}\right]+X_{i2}\log(\pi_{(X_2)})+(1-X_{i2})\log(1-\pi_{(X_2)})$$

$$+X_{i6}(\gamma_0+\gamma_1X_{i2}+\cdots+\gamma_4X_{i5})+X_{i7}(\iota_0+\iota_1X_{i2}+\cdots+\iota_5X_{i6})+X_{i8}(\kappa_0+\kappa_1X_{i2}+\cdots+\kappa_6X_{i7})$$

$$-\log(1+\exp(\gamma_0+\gamma_1X_{i2}+\cdots+\gamma_4X_{i5}))-\log(1+\exp(\iota_0+\iota_1X_{i2}+\cdots+\iota_5X_{i6}))$$

$$-\log(1+\exp(\kappa_0+\kappa_1X_{i2}+\cdots+\kappa_6X_{i7}))+Y_i(\beta_0+\beta_1X_{i2}+\cdots+\beta_7X_{i8})$$

$$-\exp(\beta_0+\beta_1X_{i2}+\cdots+\beta_7X_{i8})+R_{i5}(\zeta_0+\zeta_1X_{i2}+\cdots+\zeta_7X_{i8}+\zeta_8Y_i)$$

$$+R_{i6}(\nu_0+\nu_1X_{i2}+\cdots+\nu_7X_{i8}+\nu_8Y_i+\nu_9R_{i5})+R_{i7}(\upsilon_0+\upsilon_1X_{i2}+\cdots+\upsilon_7X_{i8}$$

$$+\upsilon_8Y_i+\upsilon_9R_{i5}+\upsilon_{10}R_{i6})+R_{i8}(\omega_0+\omega_1X_{i2}+\cdots+\omega_7X_{i8}+\omega_8Y_i+\omega_9R_{i5}+\omega_{10}R_{i6}+\omega_{11}R_{i7})$$

$$-\log(1+\exp(\zeta_0+\zeta_1X_{i2}+\cdots+\zeta_7X_{i8}+\zeta_8Y_i))-\log(\nu_0+\nu_1X_{i2}+\cdots+\nu_7X_{i8}+\nu_8Y_i+\nu_9R_{i5}))$$

$$-\log(\upsilon_0+\upsilon_1X_{i2}+\cdots+\upsilon_7X_{i8}+\upsilon_8Y_i+\upsilon_9R_{i5}+\upsilon_{10}R_{i6}))$$

$$-\log(\omega_0+\omega_1X_{i2}+\cdots+\omega_7X_{i8}+\omega_8Y_i+\omega_9R_{i5}+\omega_{10}R_{i6}+\omega_{11}R_{i7}))$$

$$\left.+I_i(\delta_0+\delta_1X_{i1}+\delta_2X_{i2}+\delta_3X_{i5})-\log(1+\exp(\delta_0+\delta_1X_{i1}+\delta_2X_{i2}+\delta_3X_{i5}))\right\}. \tag{4.3}$$

The parameters are estimated through

$$\sum_{i=1}^{n}\sum_{c=1}^{C}\frac{1}{p_i}I_{\{r_i=c\}}\mathscr{E}\left[\frac{\partial}{\partial\theta}\log f(I_i,\mathbf{R}_i,Y_i,\mathbf{X}_i;\theta)\mid I_i=1,\mathbf{r}_i,y_i,\mathbf{x}_{i,obs}\right]. \tag{4.4}$$

where C is the total patterns of missing data in the ith observation.

TABLE II
SETUP OF SIMULATION VARIABLES

Variable	Type	Distribution	Parameters	Missing[a,b]	Non-Samples Observable		
					Case 1[d]	Case 2[e]	Case 3[f]
X_1	Continuous	Normal	$\mu_{(X_1)}, \sigma^2_{(X_1)}$	0%	No	Yes	Yes
X_2	Categorical	Binary	$\pi_{(X_2)}$	0%	No	Yes	Yes
X_3	Categorical	Poisson	$\lambda_{(X_3)}$	0%	No	Yes	Yes
X_4	Continuous	Normal	$\mu_{(X_4)}, \sigma^2_{(X_4)}$	0%	No	Yes	Yes
X_5	Continuous	Normal	$\mu_{(X_5\|X_2,\cdots,X_4)} = \alpha_0 + \alpha_1 X_2 + \cdots + \alpha_3 X_4,\ \sigma^2_{(X_5\|X_2,\cdots,X_4)}$	40%	No	Yes	Yes
X_6	Categorical	Binary	$\pi_{(X_6\|X_2,\cdots,X_5)} = \text{expit}(\gamma_0 + \gamma_1 X_2 + \cdots + \gamma_4 X_5)$	10%	No	Yes	Yes
X_7	Categorical	Binary	$\pi_{(X_7\|X_2,\cdots,X_6)} = \text{expit}(\iota_0 + \iota_1 X_2 + \cdots + \iota_5 X_6)$	20%	No	Yes	No[c]
X_8	Categorical	Binary	$\pi_{(X_8\|X_2,\cdots,X_7)} = \text{expit}(\kappa_0 + \kappa_1 X_2 + \cdots + \kappa_6 X_7)$	35%	No	Yes	No[c]
Y	Categorical	Poisson	$\lambda_{(Y\|X_2,\cdots,X_8)} = \exp(\beta_0 + \beta_1 X_2 + \cdots + \beta_7 X_8)$	30%(0%[b])	No	No	No

[a] Approximate proportion, except for 0%, which is an exact value.
[b] For Case 1: missing among samples.
[c] The actual missing proportion for X_7 and X_8 is larger in Case 3 due to this.
[d] Case 1: survey data with non-ignorable missing covariates, data observable only among samples.
[e] Case 2: survey data with non-ignorable missing covariates, all covariates observable on samples and non-samples.
[f] Case 3: survey data with non-ignorable missing covariates, some covariates observable on samples and non-samples.

The likelihood-based estimation of the parameters follows the augmentation assisted EM algorithm outlined in Section 3.6 of Chapter 3. I refer to it as the augmentation method from this point forward. To augment data row with any missing element, I use the two possible values for the binary variables, and five nodes of Gauss-Hermite quadrature with their corresponding weights for the continuous variable. Preliminary trials with similarly simulated data indicate that five nodes of Gauss-Hermite quadrature keeps the computation time manageable for an acceptable level of accuracy. If the missing element is the count variable Y, I augment the observation with Y_{mis} =(0, 1, 2, 3). Given the expected value of Y as I created it, that is, $\lambda_{(Y)} = 0.35$, the probability mass function of Y equals any value larger than 3 is about zero at 3 decimals. The function $Q(\theta \mid \theta^{(t)})$ at iteration t is Equation 3.13, with notational adjustment, when the sampling form is assumed known; it is Equation 4.4 for the survey weighted analysis, with $\theta = \theta^{(t)}$. The algorithm estimates $Q(\theta \mid \theta^{(t)})$ in the E-step via Equation 3.22, Equation 3.23, and Equation 3.24, respectively, for missing variables of categorical, continuous, and mixture natures. In the M-step, it maximizes θ using Equation 3.16. For the convergence criterion, I opt for $\left\|\theta_{XYI}^{(t+1)} - \theta_{XYI}^{(t)}\right\|^2 < 10^{-3}$, where θ_{XYI} are the parameters of $I, Y, X_1, \cdots, X_8$ in θ. Preliminary trials show that the differences in terms of δ, β, and θ_X, that is, the parameters of $I, Y, X_1, \cdots, X_8$ between the above criterion and using the whole estimates $\left\|\theta^{(t+1)} - \theta^{(t)}\right\|^2 < 10^{-3}$ are as small as two decimals or less, but slightly unstable for θ_R. However, the required iteration for convergence in the latter criterion is greater by a factor of about 10^2 or more. As a comparison, Stubbendick and Ibrahim (126) who also applied Ibrahim et al.'s approach (70; 72; 127) used only the parameters of interest $\left\|\beta^{(t+5)} - \beta^{(t)}\right\|^2 < 10^{-2}$ for convergence.

Another computation of the parameter estimates is performed through a slight modification of the augmentation method. Here, I modify the E-step by removing one or some components of the weight $f(\boldsymbol{\nu}_{i,mis} \mid \boldsymbol{z}_{i,obs}; \theta)$ in Equation 3.15 which relate to constant values in the missing part of data. For instance, the value of $R_{i8} = 0$ for all i where X_{i8} is missing. Hence, assuming Equation 4.1 is known and only X_8 that is missing for the ith observation, the numerator of the weight $f(x_{i8} \mid I_i, \boldsymbol{r}_i, y_i, x_{i1}, \cdots, x_{i7}; \theta^{(t)})$ at the jth augmentation of i is

$$\begin{aligned}
&f(I_i \mid x_{i1}, x_{i2}, x_{i5(j)}; \delta^{(t)}) \times f(r_{i7} \mid r_{i6}, r_{i5}, y_i, \boldsymbol{x}_{i(2-8)(j)}; \upsilon^{(t)}) \times \\
&f(r_{i6} \mid r_{i5}, y_i, \boldsymbol{x}_{i(2-8)(j)}; \nu^{(t)}) \times f(r_{i5} \mid y_i, \boldsymbol{x}_{i(2-8)(j)}; \zeta^{(t)}) \times f(y_i \mid \boldsymbol{x}_{i(2-8)(j)}; \boldsymbol{\beta}^{(t)}) \times \\
&f(x_{i8(j)} \mid x_{i2}, \cdots, x_{i7}; \kappa^{(t)}) \times f(x_{i7} \mid x_{i2}, \cdots, x_{i6}; \iota^{(t)}) \times \\
&f(x_{i6} \mid x_{i2}, \cdots, x_{i5}; \boldsymbol{\gamma}^{(t)}) \times f(x_{i5} \mid x_{i2}, \cdots, x_{i4}; \boldsymbol{\alpha}^{(t)})
\end{aligned} \tag{4.5}$$

where $\boldsymbol{x}_{i(2-8)(j)} = x_{i2}, \cdots, x_{i7}, x_{i8(j)}$. That is, the term $f(r_{i8} \mid r_{i7}, r_{i6}, x_{i5}, y_i, \boldsymbol{x}_{i(2-8)(j)}; \boldsymbol{\omega}^{(t)})$ disappears from the weight. Accordingly, the computation for $X_{i8,mis}$, and in general for each of $\boldsymbol{V}_{i,mis} \in \{\boldsymbol{X}_{i,mis}, Y_{i,mis}\}$, works under MAR assumption. The algorithm, however, still uses the joint distribution containing the missing data mechanisms, which differentiates it from an approach assuming ignorable missingness. A justification for this modification is that constant values add no information to the estimation. In fact, when the constant values are quite dominant, the estimated parameters can be very unstable because of the potentially flat likelihood. Thus, the modified version of the proposed method is created to circumvent the identifiability problem. Moreover, it gives a computational advantage in the speed of convergence. This ad hoc approach is not new; the multiple imputation by chained equations (MICE) (13; 16), for

instance, does the same thing. I termed such computational modification the augmentation-constant removed (CR) method, as a reference of what it does. The comparison with the proposed method is clearly aimed at evaluating the trade off between a slight violation of non-ignorable assumption and the accuracy of the estimates.

I compare the estimated parameters from the proposed method and its computational variant with those based on the full case and the complete case analyses. Since standard statistical packages still default to the complete case method when encountering data with missingness, the comparison has a non-trivial meaning. Besides, it can serve as a sensitivity analysis concerning a more complicated model of missing data, given that the underlying assumption of a complete case analysis is MCAR. The comparison with the full case analysis involves all parameters. It only, however, covers the parameters of the outcome and covariate models with the complete case analysis. This should be reasonable since in practice it is unlikely we model the missing data mechanism when we resort to a complete case analysis. Furthermore, fitting a regression model for the missing data mechanism when $\{X_i : X_i \in \mathscr{S}\}$ hardly results in convergence, as $R_{ik} = 1$ for all $\{i, k : i \in \mathscr{S}, k = 5, \cdots, 8\}$.

Multiple imputation is arguably the most prescribed approach for missing data in the recent literature. I thus additionally compute the estimates by multiply imputing the simulated datasets. In particular, I use MICE, which is also known as the fully conditional specifications (FCS) method (13; 16). MICE consists of imputation and analysis steps. The first step is initialized by a simple random draw from the incomplete variables to replace the missing values. The sequence of imputation can be arbitrary. At iteration $s = 1, \cdots, S$, the dataset is mul-

tiplied into J replicates, and for each $j = 1, \cdots, J$, successive draws are implemented through Gibbs sampler. At the analysis step, MICE proceeds by analyzing each imputed datasets using the model of interest. Then, the estimates are obtained using Rubin's rule (11).

There are N = 100,000 observations on each simulated data of Case 1, and N = 1,000 observations on those of Case 2 and Case 3. Every simulation run is replicated M = 1,000 times. Information on the estimation error, defined as $(\theta - \hat{\theta}_m)$, $m = 1, \cdots, M$, where the hat differentiates the estimated from the true θ, is collected within each run in terms of observed, squared, and absolute values. Thereafter, the quantity of the empirical bias $\sum_{m=1}^{M}(\theta - \hat{\theta}_m)/M$, mean squared error (MSE) $\sum_{m=1}^{M}(\theta - \hat{\theta}_m)^2/M$, and median absolute error (MAE), are computed and reported. The asymptotic standard error (ASE) for each estimated parameter is also calculated. For the proposed method and its alternative, the ASE is derived from inverting Equation 3.17 if the functional form of sample selection is assumed known, or through Equation 3.20 otherwise. All analyses are conducted using the R statistical software version 3.3.3. I use the R package `mice` for application of MICE, leaving everything at its default settings. This includes an imputation based on predictive mean matching for numeric variables and logistic regression for binary variables. Count variable Y is considered as numeric in `mice`. By default, `mice` uses a total of five for both data replicates and maximum iteration during the imputation step. Convergence in `mice` is assessed visually, as suggested by its creators (16). Finally, I also implement the `survey` library in R for the survey weighted analyses of the full and complete cases in the Case 1 Situation 2 simulations. To my best knowledge, `mice` does not yet imple-

ment survey weighting in its programming. Hence, I exclude MICE in the evaluation for Case 1 Situation 2 to keep a fair comparison with the other methods.

4.3 Simulation of Case 2 Survey

It is likely, though not typical, that in some survey settings the covariates are observable for all elements of the population. One instance would be a survey of a particular outcome among a small, well-defined community, where cost or some other concern prohibits the use of census. The outcome can be mortality or morbidity of a recent outbreak, and other current status of individuals, which is normally not known prior to the survey. On the other hand, covariates information that includes demographic and socioeconomic characteristics, as well as risk factors, may have been parts of the regular registration in the local administration offices or clinics. Certainly, they are subject to missing values for various of reasons, including those associated with the missingness itself. Such class of survey cases, which is classified as Case 2 in the previous chapter formulation, thus represents the mildest setting of survey data with non-ignorable missing covariates because some information of the non-samples is not subject to sample selection (that is, "observable").

Overall, the augmentation method needs a median of 79 (first quartile $Q_1 = 37$, third quartile $Q_3 = 194$) iterations to achieve convergence in the Case 2 simulation. Not surprisingly, the augmentation-CR method requires fewer iterations, with the median 16 ($(Q_1, Q_3) = (14, 17)$). The actual missing proportion of $Y, X_5, \cdots, X_8$ is nicely close to the simulation setup, where the rounded average $\bar{N}_{Y_{\text{mis}}} = 304$ (30%), $\bar{N}_{X_{5,\text{mis}}} = 410$ (41%), $\bar{N}_{X_{6,\text{mis}}} = 107$ (11%), $\bar{N}_{X_{7,\text{mis}}} = 207$ (21%), and $\bar{N}_{X_{8,\text{mis}}} = 377$ (38%). Note that I use the notation N since the

numbers involve all observations. Based on the setup, $\bar{n} = \bar{N}_{Y_{obs}} = 696$ is the average sample size in the simulation. To analyze the relationship between Y and $X_2, \cdots, X_8$ utilizing the complete cases, each simulated data only has an average of about 181 observations. This is a loss of approximately 82% of data.

Table III and Figure 2 show the simulation results for the parameters of interest $\boldsymbol{\beta}$. The augmentation method appears to have a smaller bias than the other methods except the full case analysis, which serves as a reference. MICE performs fine, but it tends to have a larger absolute bias than the augmentation method. Their estimated $\boldsymbol{\beta}$ are, however, about equally efficient. In a stark contrast, the complete case analysis generally performs the worst, having a bias than can be substantial and extremely lacking in efficiency.

Table IV compares the results for the parameters of the missing covariate models across the competing techniques. With the full case analysis as an exception, performance of the augmentation method remains well in comparison to the others. Its modification, the augmentation-CR method seems to be either as good as the augmentation method, or somewhere between the augmentation method and MICE. Efficiency wise, however, there is barely any gain the proposed method has over MICE. As a side note, the current R package `mice` to the best knowledge does not include an estimate of the linear regression residual; hence, $\widehat{\sigma^2}_{(X_5)}$ is not computed for MICE in all simulations. On the other hand, R produces such estimate in both `lm` and `glm` environments, but without an estimated standard error. Accordingly, the table cell for the asymptotic standard error (ASE) of $\sigma^2_{(X_5)}$ is also empty for the analyses utilizing those environments, which include the full and complete case analyses.

There are several things worth mentioning with regard to the parameter estimates of the missing data mechanism in this Case 2 simulation (Table V). First of all, the proposed method still in general provides the least biased estimates among the three competing approaches using the incomplete data (MICE, the augmentation-CR, and the augmentation methods). Its estimated parameters also tend to be the closest ones to those of a full case analysis. Such argument, however, does not necessarily apply to the GLM parameter representing the partial relationship between the missing data indicator and its corresponding missing covariate. For instance, all simulation measures (bias, MSE, MAE, and ASE) for ω_7, the parameter of the (R_8, X_8) relationship adjusted for the effect of other variables in the R_8 model, are worse in the augmentation method than either MICE or the augmentation-CR method. A fairly similar problem is noted for ν_5 and υ_6, which are the parameters relating (R_6, X_6) and (R_7, X_7), respectively, even though the bias of the augmentation method for υ_6 is relatively smaller than both MICE and the augmentation-CR method. The root of this problem is intuitively clear. All X_6, X_7, X_8 are designed simulation wise as correlated binary variables, with a missing proportion on each ranging between 10% and 35%. Accordingly, the risk of considerable constant values in the augmented data is high during the estimation of their missing data mechanism, which leads to a relatively flat likelihood and identifiability issue. Such problem does not present in MICE and the modified, augmentation-CR method, because their algorithm avoids dealing with constant values. Nevertheless, if the missing covariate is continuously distributed such as X_5, then the issue seems to be slightly resolved. Table V shows that ζ_4, the GLM parameter

for the (R_5, X_5) relationship, is fairly well estimated by the augmentation method. In fact, its estimate appears to be slightly better than MICE.

All missing data methods, including the complete case analysis, demonstrate comparable results when it comes to the estimates of the sample selection mechanism (Table VI). Such finding should not come as a surprise, given that the affected variable Y is by simulation design not part of the model for I. That is, I and Y are conditionally independent and thus the missingness is MAR. In such situation, it is well known that even standard statistical methods are able to produce valid estimates (see Equation 2.1).

4.4 Simulation of Case 3 Survey

An extension of the survey settings illustrated in the previous section is when the covariates set is only partially observable among the non-samples. Take for instance, current income of household head, months of pregnancy and the accumulated antenatal care, number of ongoing postnatal visits following a recent delivery, any medical problem in the previous two weeks, and other health or socioeconomic status that is very dynamic in nature. It is understandable if they are not part of the population record, and only available through specific purpose surveys. On the other hand, the more static demographic variables, such as household wealth index, total children at the end of the year, the primary source of drinking water, the average doctor visits for certain time point, and other similar examples, may be maintained by the local authorities or health providers for the whole population, although they are probably missing for some individuals or households. These are just to emphasize that such case of survey data is possible, and can be realistic for a small, well-defined population.

TABLE III
SIMULATION RESULTS FOR THE PARAMETERS OF INTEREST, CASE 2 (M = 1,000 REPLICATIONS; N = 1,000 OBSERVATIONS)

	Full Case			Complete Case			MICE			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE(ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
$\beta_0 =$ 0.10	-0.02	0.09	0.20 (0.29)	0.20	0.42	0.45 (0.65)	-0.01	0.13	0.24(0.36)	-0.01	0.11	0.23 (0.34)	-0.04	0.12	0.24 (0.34)
β_1 = -0.10	0.00	0.01	0.08 (0.12)	-0.06	0.06	0.16 (0.25)	-0.04	0.02	0.10(0.15)	-0.05	0.02	0.09 (0.14)	-0.01	0.02	0.09 (0.14)
β_2 = -0.10	0.00	0.00	0.04 (0.06)	0.04	0.02	0.10 (0.14)	0.03	0.01	0.06(0.09)	0.04	0.01	0.06 (0.08)	0.01	0.01	0.06 (0.08)
β_3 = -0.10	0.00	0.01	0.05 (0.07)	-0.05	0.03	0.11 (0.16)	-0.03	0.01	0.07(0.10)	-0.04	0.01	0.07 (0.09)	-0.01	0.01	0.06 (0.09)
β_4 = -0.10	0.00	0.00	0.03 (0.05)	0.05	0.02	0.09 (0.12)	0.04	0.01	0.06(0.08)	0.04	0.01	0.06 (0.07)	0.01	0.01	0.05 (0.07)
β_5 = -0.10	0.00	0.01	0.07 (0.11)	0.06	0.06	0.16 (0.23)	-0.01	0.02	0.09(0.14)	-0.01	0.02	0.09 (0.13)	0.00	0.02	0.09 (0.13)
β_6 = -0.10	0.00	0.01	0.08 (0.12)	-0.04	0.07	0.19 (0.27)	0.02	0.03	0.11(0.17)	0.02	0.03	0.11 (0.15)	0.01	0.03	0.11 (0.15)
$\beta_7 =$ 0.10	0.00	0.01	0.07 (0.10)	-0.05	0.05	0.15 (0.22)	0.01	0.03	0.12(0.17)	0.01	0.03	0.11 (0.16)	0.01	0.03	0.11 (0.16)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error.

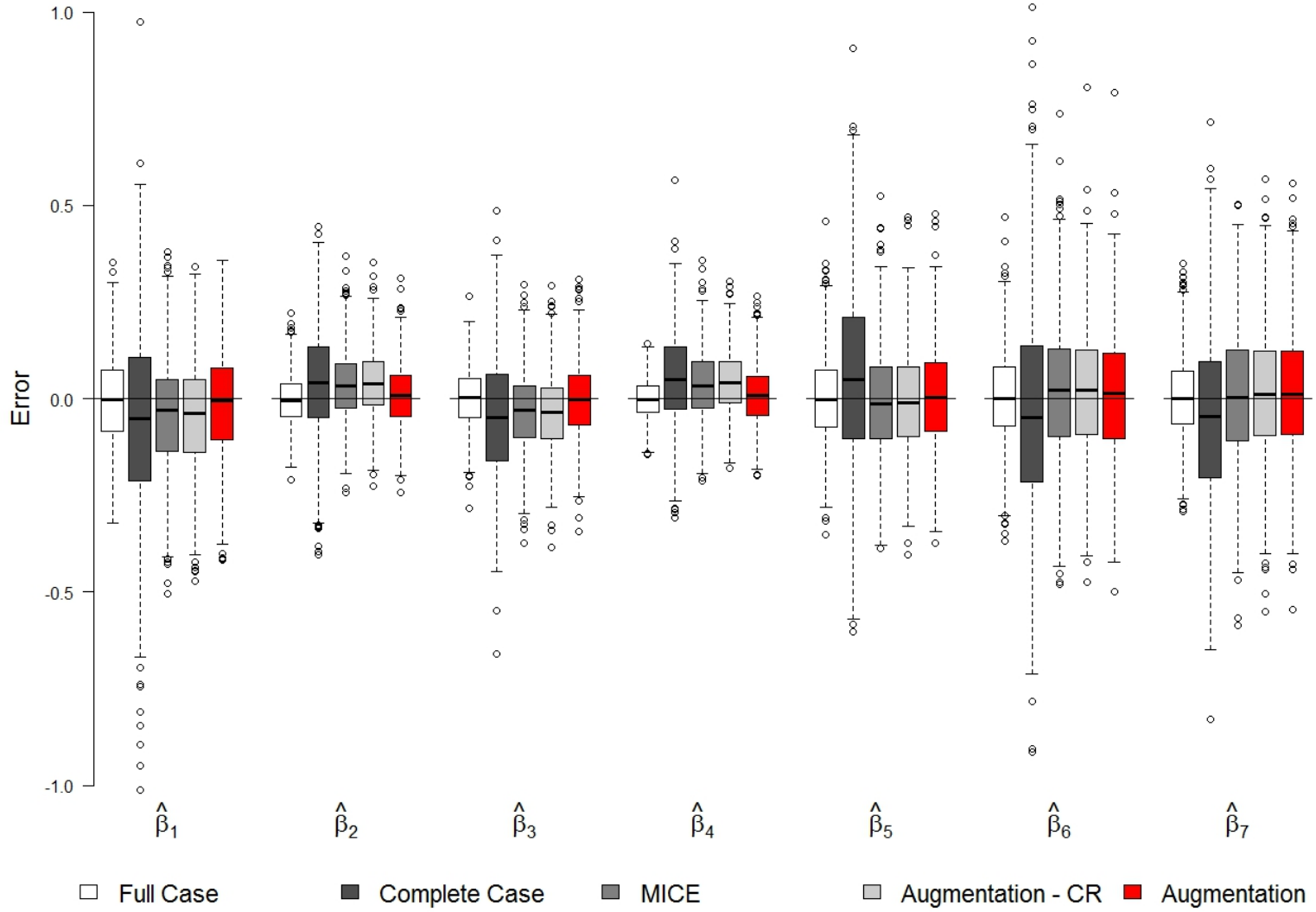

Figure 2. Distribution of Errors in the Parameters of Interest on Case 2 Simulation

TABLE IV
SIMULATION RESULTS FOR THE COVARIATES PARAMETERS, CASE 2 (M = 1,000 REPLICATIONS; N = 1,000 OBSERVATIONS)

	Full Case			Complete Case			MICE			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
α_0 = -1.00	0.00	0.03	0.12 (0.17)	-0.19	0.08	0.21 (0.21)	-0.12	0.06	0.17 (0.21)	-0.16	0.07	0.18 (0.20)	-0.04	0.05	0.15 (0.20)
α_1 = 1.00	0.00	0.00	0.04 (0.06)	0.00	0.01	0.05 (0.08)	0.01	0.01	0.05 (0.08)	0.01	0.01	0.05 (0.07)	0.01	0.01	0.05 (0.07)
α_2 = -1.00	0.00	0.00	0.02 (0.02)	0.01	0.00	0.02 (0.03)	0.00	0.00	0.02 (0.03)	0.00	0.00	0.02 (0.03)	0.00	0.00	0.02 (0.03)
α_3 = 1.00	0.00	0.00	0.02 (0.03)	-0.01	0.00	0.03 (0.04)	-0.01	0.00	0.03 (0.04)	0.00	0.00	0.03 (0.04)	0.00	0.00	0.03 (0.04)
$\sigma^2_{X_5}$ = 1.00	0.00	0.00	0.03 (.)	-0.06	0.01	0.07 (.)	.	.	. (.)	-0.07	0.01	0.07 (0.05)	-0.07	0.01	0.07 (0.06)
γ_0 = 0.14	-0.02	0.14	0.26 (0.38)	-0.30	0.37	0.41 (0.52)	-0.10	0.19	0.29 (0.42)	-0.10	0.18	0.28 (0.42)	-0.13	0.22	0.32 (0.41)
γ_1 = -0.14	0.00	0.02	0.10 (0.16)	0.05	0.05	0.15 (0.21)	0.05	0.03	0.13 (0.18)	0.04	0.03	0.13 (0.18)	0.02	0.03	0.12 (0.17)
γ_2 = -0.14	0.00	0.01	0.06 (0.09)	-0.05	0.02	0.09 (0.12)	-0.04	0.01	0.08 (0.11)	-0.04	0.01	0.08 (0.11)	0.00	0.01	0.07 (0.10)
γ_3 = 0.14	0.01	0.01	0.07 (0.10)	0.06	0.02	0.10 (0.14)	0.05	0.02	0.09 (0.12)	0.05	0.02	0.09 (0.12)	0.02	0.02	0.08 (0.12)
γ_4 = 0.14	0.00	0.01	0.05 (0.07)	-0.06	0.01	0.08 (0.10)	-0.04	0.01	0.08 (0.10)	-0.04	0.01	0.07 (0.09)	-0.01	0.01	0.06 (0.09)
ι_0 = -0.25	0.00	0.16	0.26 (0.41)	0.19	0.50	0.47 (0.67)	-0.01	0.23	0.30 (0.48)	-0.03	0.21	0.30 (0.47)	-0.02	0.30	0.35 (0.46)
ι_1 = 0.25	0.00	0.03	0.11 (0.16)	-0.04	0.07	0.18 (0.27)	-0.03	0.05	0.14 (0.20)	-0.03	0.04	0.13 (0.20)	-0.01	0.04	0.13 (0.19)
ι_2 = 0.20	0.00	0.01	0.06 (0.09)	0.06	0.03	0.12 (0.16)	0.04	0.02	0.09 (0.13)	0.04	0.02	0.10 (0.13)	0.01	0.02	0.09 (0.12)
ι_3 = 0.25	0.00	0.01	0.07 (0.10)	-0.05	0.03	0.12 (0.17)	-0.04	0.02	0.10 (0.14)	-0.04	0.02	0.10 (0.14)	-0.01	0.02	0.09 (0.13)
ι_4 = -0.25	0.00	0.01	0.05 (0.07)	0.05	0.02	0.10 (0.12)	0.04	0.02	0.09 (0.12)	0.04	0.01	0.08 (0.11)	0.02	0.01	0.07 (0.11)
ι_5 = 0.15	0.00	0.03	0.11 (0.16)	0.05	0.07	0.18 (0.25)	-0.02	0.04	0.13 (0.20)	-0.02	0.04	0.13 (0.19)	-0.01	0.04	0.14 (0.19)
κ_0 = -0.25	-0.01	0.15	0.26 (0.37)	0.40	0.75	0.56 (0.75)	0.08	0.27	0.35 (0.51)	0.07	0.26	0.34 (0.48)	0.15	0.29	0.36 (0.48)
κ_1 = -0.10	0.00	0.02	0.09 (0.15)	-0.06	0.09	0.19 (0.29)	-0.05	0.04	0.13 (0.21)	-0.04	0.04	0.13 (0.20)	0.00	0.04	0.12 (0.19)
κ_2 = -0.10	0.00	0.01	0.05 (0.08)	0.05	0.03	0.12 (0.16)	0.05	0.02	0.09 (0.13)	0.04	0.02	0.09 (0.12)	0.02	0.01	0.08 (0.12)
κ_3 = 0.10	0.00	0.01	0.06 (0.09)	-0.06	0.04	0.13 (0.18)	-0.05	0.02	0.10 (0.14)	-0.04	0.02	0.10 (0.13)	-0.02	0.02	0.09 (0.13)
κ_4 = -0.10	0.00	0.00	0.05 (0.07)	0.05	0.02	0.10 (0.13)	0.05	0.02	0.09 (0.12)	0.04	0.01	0.07 (0.11)	0.00	0.01	0.07 (0.10)
κ_5 = -0.10	-0.01	0.02	0.10 (0.14)	0.06	0.08	0.20 (0.27)	-0.01	0.04	0.13 (0.20)	-0.01	0.04	0.13 (0.19)	0.01	0.04	0.12 (0.19)
κ_6 = -0.10	0.00	0.02	0.10 (0.15)	-0.06	0.11	0.22 (0.30)	0.02	0.05	0.15 (0.21)	0.01	0.05	0.14 (0.21)	0.00	0.05	0.14 (0.20)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error; { . }, not estimated.

TABLE V
SIMULATION RESULTS FOR THE PARAMETERS OF MISSING DATA MECHANISM, CASE 2 (M = 1,000 REPLICATIONS; N = 1,000 OBSERVATIONS)

	Full Case			MICE			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
ζ_0 = -0.52	-0.01	0.18	0.29 (0.41)	0.33	0.30	0.39 (0.43)	0.29	0.27	0.37 (0.42)	0.03	0.21	0.31 (0.43)
ζ_1 = 0.52	0.01	0.02	0.11 (0.16)	-0.43	0.21	0.43 (0.17)	-0.40	0.18	0.40 (0.17)	-0.11	0.05	0.15 (0.18)
ζ_2 = -0.52	0.00	0.01	0.06 (0.09)	0.42	0.18	0.43 (0.11)	0.39	0.16	0.39 (0.11)	0.11	0.03	0.13 (0.12)
ζ_3 = 0.52	0.00	0.01	0.07 (0.10)	-0.43	0.19	0.43 (0.12)	-0.40	0.16	0.40 (0.12)	-0.11	0.04	0.12 (0.13)
ζ_4 = -0.52	0.00	0.01	0.05 (0.07)	0.44	0.19	0.44 (0.10)	0.40	0.16	0.40 (0.10)	0.11	0.03	0.12 (0.11)
ζ_5 = -0.52	0.00	0.03	0.11 (0.16)	-0.04	0.03	0.12 (0.17)	-0.04	0.03	0.12 (0.16)	-0.01	0.03	0.12 (0.17)
ζ_6 = 0.52	0.01	0.03	0.11 (0.16)	0.10	0.05	0.15 (0.18)	0.08	0.04	0.14 (0.17)	0.03	0.04	0.13 (0.18)
ζ_7 = 0.52	0.01	0.02	0.10 (0.14)	0.03	0.04	0.13 (0.19)	0.03	0.03	0.12 (0.18)	0.02	0.03	0.12 (0.18)
ζ_8 = 0.52	0.01	0.02	0.09 (0.12)	0.02	0.02	0.11 (0.16)	0.03	0.02	0.10 (0.15)	0.02	0.02	0.10 (0.15)
ν_0 = 0.18	0.01	0.39	0.43 (0.61)	-0.03	0.43	0.45 (0.66)	-0.05	0.41	0.42 (0.73)	-0.14	0.61	0.51 (0.84)
ν_1 = 0.18	0.01	0.06	0.16 (0.24)	0.01	0.07	0.18 (0.26)	0.01	0.07	0.18 (0.26)	0.02	0.07	0.18 (0.26)
ν_2 = 0.18	0.00	0.02	0.09 (0.13)	0.02	0.03	0.11 (0.17)	0.01	0.03	0.11 (0.16)	0.02	0.03	0.11 (0.16)
ν_3 = 0.18	0.01	0.02	0.11 (0.15)	-0.01	0.04	0.13 (0.18)	0.00	0.03	0.12 (0.18)	-0.01	0.04	0.12 (0.18)
ν_4 = 0.18	0.00	0.01	0.07 (0.11)	0.01	0.02	0.10 (0.15)	0.00	0.02	0.09 (0.14)	0.00	0.02	0.09 (0.14)
ν_5 = -0.18	-0.03	0.05	0.15 (0.24)	0.17	0.04	0.17 (0.35)	0.18	0.04	0.19 (0.65)	0.35	0.82	0.39 (0.81)
ν_6 = 0.18	-0.01	0.06	0.17 (0.24)	0.00	0.09	0.19 (0.28)	-0.01	0.08	0.18 (0.28)	-0.02	0.08	0.19 (0.28)
ν_7 = 0.18	0.02	0.05	0.15 (0.22)	0.02	0.09	0.20 (0.30)	0.03	0.09	0.19 (0.29)	0.02	0.09	0.20 (0.29)
ν_8 = 0.18	0.01	0.03	0.12 (0.18)	0.01	0.04	0.14 (0.21)	0.02	0.04	0.13 (0.21)	0.02	0.04	0.14 (0.21)
ν_9 = 0.18	0.00	0.05	0.16 (0.23)	-0.05	0.05	0.15 (0.23)	-0.05	0.05	0.16 (0.23)	0.03	0.06	0.17 (0.24)
υ_0 = -0.14	-0.01	0.27	0.36 (0.50)	-0.02	0.28	0.35 (0.53)	-0.03	0.27	0.34 (0.54)	0.08	0.56	0.40 (1.06)
υ_1 = 0.14	0.00	0.03	0.12 (0.18)	0.00	0.04	0.13 (0.20)	0.00	0.04	0.13 (0.19)	0.00	0.04	0.13 (0.20)
υ_2 = 0.14	0.01	0.01	0.07 (0.10)	0.00	0.02	0.08 (0.13)	0.00	0.01	0.08 (0.12)	0.00	0.02	0.08 (0.12)
υ_3 = 0.14	0.00	0.01	0.08 (0.12)	0.00	0.02	0.09 (0.14)	0.00	0.02	0.09 (0.13)	0.00	0.02	0.09 (0.14)
υ_4 = 0.14	0.00	0.01	0.05 (0.08)	0.01	0.01	0.07 (0.11)	0.01	0.01	0.07 (0.11)	0.01	0.01	0.08 (0.11)
υ_5 = -0.14	-0.01	0.03	0.12 (0.18)	-0.01	0.04	0.14 (0.20)	-0.01	0.04	0.13 (0.19)	-0.02	0.04	0.14 (0.20)
υ_6 = -0.14	-0.01	0.03	0.12 (0.19)	0.15	0.03	0.15 (0.28)	0.15	0.03	0.15 (0.39)	0.04	0.96	0.45 (1.00)
υ_7 = 0.14	0.01	0.03	0.12 (0.17)	0.01	0.05	0.15 (0.23)	0.01	0.05	0.15 (0.22)	0.01	0.05	0.14 (0.22)
υ_8 = 0.14	0.01	0.02	0.09 (0.14)	0.01	0.02	0.10 (0.16)	0.01	0.02	0.10 (0.15)	0.01	0.03	0.11 (0.16)

Continued on next page

TABLE V (Continued)
SIMULATION RESULTS FOR THE PARAMETERS OF MISSING DATA MECHANISM, CASE 2 (M = 1,000 REPLICATIONS; N = 1,000 OBSERVATIONS)

	Full Case			MICE			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
υ_9 = 0.14	0.00	0.03	0.12 (0.17)	-0.08	0.04	0.13 (0.17)	-0.07	0.04	0.13 (0.18)	-0.03	0.04	0.14 (0.18)
υ_{10} = 0.14	-0.01	0.06	0.16 (0.25)	-0.01	0.06	0.17 (0.25)	-0.01	0.06	0.16 (0.25)	0.00	0.06	0.17 (0.26)
ω_0 = -0.10	0.03	0.19	0.28 (0.44)	0.10	0.21	0.31 (0.46)	0.10	0.20	0.30 (0.45)	0.14	0.22	0.31 (0.45)
ω_1 = -0.10	0.00	0.02	0.11 (0.15)	-0.01	0.03	0.11 (0.16)	-0.01	0.03	0.11 (0.16)	-0.02	0.03	0.11 (0.16)
ω_2 = -0.10	0.00	0.01	0.05 (0.08)	0.01	0.01	0.07 (0.10)	0.00	0.01	0.06 (0.10)	0.00	0.01	0.07 (0.10)
ω_3 = 0.10	-0.01	0.01	0.06 (0.10)	-0.01	0.01	0.07 (0.11)	-0.01	0.01	0.07 (0.11)	0.00	0.01	0.07 (0.11)
ω_4 = 0.10	0.00	0.00	0.05 (0.07)	0.01	0.01	0.06 (0.09)	0.00	0.01	0.05 (0.09)	0.00	0.01	0.05 (0.09)
ω_5 = -0.10	0.00	0.02	0.10 (0.15)	0.01	0.03	0.11 (0.16)	0.00	0.03	0.11 (0.16)	0.00	0.03	0.11 (0.16)
ω_6 = 0.10	0.00	0.02	0.11 (0.16)	-0.01	0.03	0.12 (0.18)	-0.01	0.03	0.12 (0.18)	-0.02	0.03	0.12 (0.18)
ω_7 = 0.10	0.00	0.02	0.09 (0.14)	-0.10	0.01	0.10 (0.21)	-0.10	0.01	0.10 (0.22)	-0.31	0.21	0.35 (0.23)
ω_8 = -0.10	0.00	0.01	0.08 (0.11)	0.00	0.02	0.09 (0.13)	0.00	0.02	0.09 (0.13)	0.00	0.02	0.09 (0.13)
ω_9 = -0.10	0.00	0.02	0.10 (0.15)	-0.03	0.02	0.10 (0.15)	-0.03	0.02	0.10 (0.15)	0.03	0.02	0.11 (0.15)
ω_{10} = 0.10	0.00	0.05	0.15 (0.22)	0.01	0.05	0.14 (0.22)	0.01	0.05	0.15 (0.22)	0.03	0.05	0.15 (0.22)
ω_{11} = 0.10	0.00	0.03	0.12 (0.17)	0.00	0.03	0.12 (0.17)	0.00	0.03	0.12 (0.17)	0.01	0.03	0.12 (0.17)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error.

TABLE VI
SIMULATION RESULTS FOR THE PARAMETERS OF SAMPLE SELECTION MECHANISM, CASE 2 (M = 1,000 REPLICATIONS; N = 1,000 OBSERVATIONS)

	Full Case			Complete Case			MICE			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
δ_0 = -0.50	-0.02	0.89	0.63 (0.92)	0.01	1.68	0.89 (1.24)	-0.08	0.91	0.65 (0.95)	-0.15	0.96	0.69 (0.95)	-0.05	0.96	0.68 (0.95)
δ_1 = 0.10	0.00	0.00	0.01 (0.02)	0.00	0.00	0.02 (0.03)	0.00	0.00	0.01 (0.02)	0.00	0.00	0.01 (0.02)	0.00	0.00	0.01 (0.02)
δ_2 = -1.00	-0.01	0.03	0.12 (0.19)	-0.02	0.07	0.17 (0.25)	0.03	0.04	0.13 (0.19)	0.00	0.04	0.13 (0.19)	0.00	0.04	0.13 (0.19)
δ_3 = -1.00	-0.01	0.01	0.05 (0.07)	-0.02	0.01	0.07 (0.10)	0.03	0.01	0.06 (0.08)	-0.01	0.01	0.06 (0.08)	-0.02	0.01	0.06 (0.08)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error.

TABLE VII
SIMULATION RESULTS FOR THE PARAMETERS OF INTEREST, CASE 3 (M = 1,000 REPLICATIONS; N = 1,000 OBSERVATIONS)

	Full Case			Complete Case			MICE			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
β_0 = 0.10	-0.02	0.08	0.20 (0.29)	0.20	0.48	0.48 (0.65)	0.01	0.12	0.23 (0.36)	-0.03	0.11	0.23 (0.34)
β_1 = -0.10	0.00	0.01	0.07 (0.12)	-0.06	0.07	0.17 (0.26)	-0.04	0.02	0.09 (0.15)	-0.01	0.02	0.09 (0.14)
β_2 = -0.10	0.00	0.00	0.04 (0.06)	0.04	0.02	0.09 (0.14)	0.04	0.01	0.06 (0.09)	0.01	0.01	0.05 (0.08)
β_3 = -0.10	0.00	0.01	0.05 (0.07)	-0.05	0.03	0.12 (0.16)	-0.04	0.01	0.07 (0.10)	-0.01	0.01	0.06 (0.09)
β_4 = -0.10	0.00	0.00	0.04 (0.05)	0.05	0.02	0.09 (0.12)	0.04	0.01	0.07 (0.08)	0.01	0.01	0.05 (0.07)
β_5 = -0.10	0.00	0.01	0.07 (0.11)	0.06	0.06	0.16 (0.24)	-0.01	0.02	0.09 (0.14)	0.01	0.02	0.09 (0.13)
β_6 = -0.10	0.00	0.01	0.08 (0.12)	-0.03	0.08	0.19 (0.27)	0.01	0.03	0.11 (0.17)	0.01	0.03	0.11 (0.15)
β_7 = 0.10	-0.01	0.01	0.07 (0.10)	-0.07	0.05	0.15 (0.23)	0.00	0.03	0.11 (0.17)	0.00	0.02	0.11 (0.16)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error.

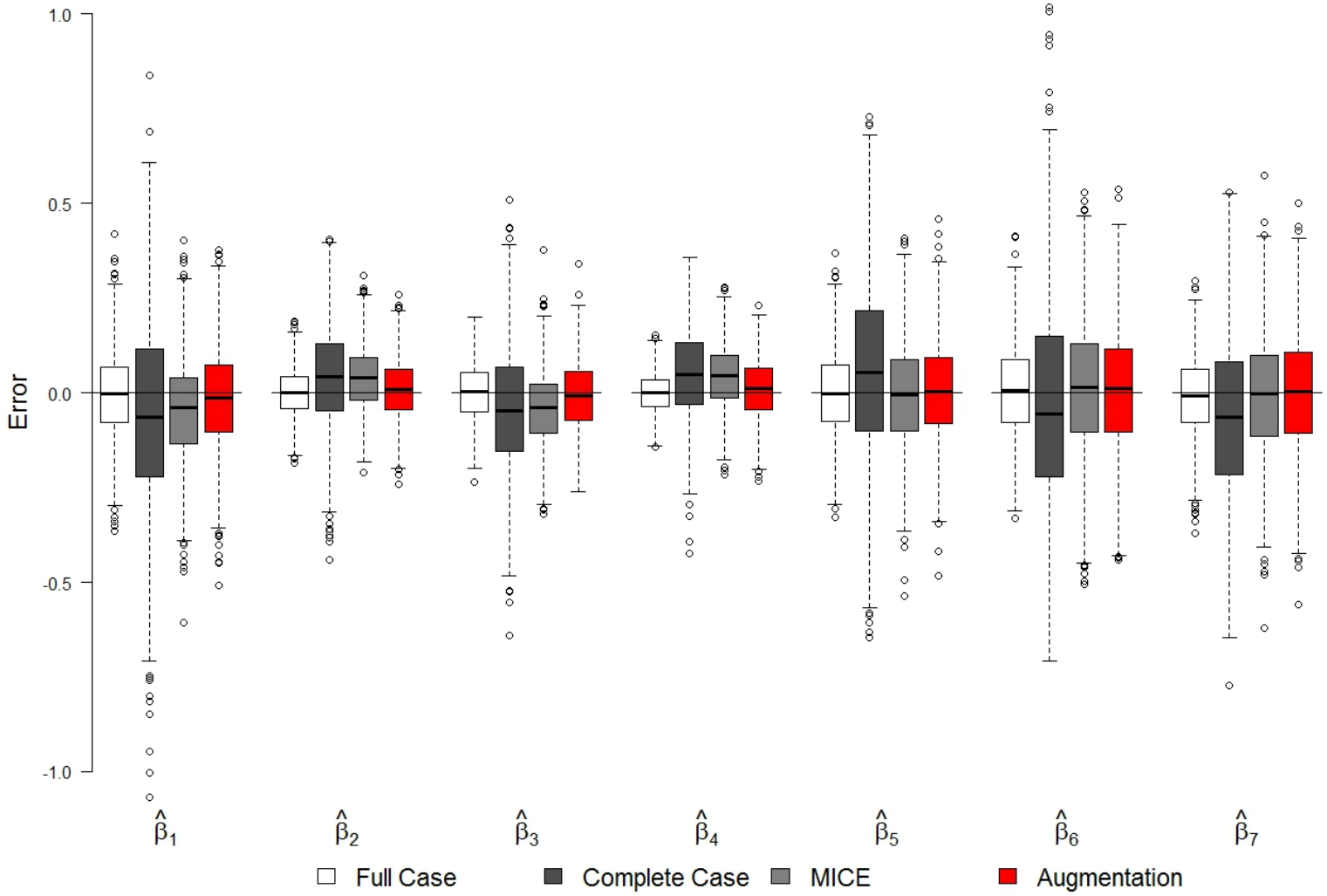

Figure 3. Distribution of Errors in the Parameters of Interest on Case 3 Simulation

TABLE VIII
SIMULATION RESULTS FOR THE COVARIATES PARAMETERS, CASE 3 (M = 1,000 REPLICATIONS; N = 1,000 OBSERVATIONS)

	Full Case			Complete Case			MICE			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
α_0 = -1.00	0.00	0.03	0.12 (0.17)	-0.20	0.09	0.22 (0.21)	-0.15	0.07	0.18 (0.22)	-0.05	0.05	0.14 (0.20)
α_1 = 1.00	0.00	0.00	0.04 (0.06)	-0.01	0.01	0.05 (0.08)	0.00	0.01	0.06 (0.08)	0.00	0.01	0.05 (0.07)
α_2 = -1.00	0.00	0.00	0.02 (0.02)	0.01	0.00	0.02 (0.03)	0.00	0.00	0.02 (0.03)	0.00	0.00	0.02 (0.03)
α_3 = 1.00	0.00	0.00	0.02 (0.03)	0.00	0.00	0.03 (0.04)	-0.01	0.00	0.03 (0.04)	0.00	0.00	0.03 (0.04)
$\sigma^2_{X_5}$ = 1.00	0.00	0.00	0.03 (.)	-0.07	0.01	0.07 (.)	.	.	. (.)	-0.07	0.01	0.07 (0.06)
γ_0 = 0.14	-0.02	0.14	0.26 (0.38)	-0.33	0.40	0.42 (0.52)	-0.10	0.19	0.28 (0.42)	-0.12	0.22	0.32 (0.41)
γ_1 = -0.14	-0.01	0.03	0.11 (0.16)	0.06	0.05	0.14 (0.21)	0.05	0.04	0.13 (0.18)	0.02	0.03	0.12 (0.17)
γ_2 = -0.14	0.00	0.01	0.05 (0.09)	-0.06	0.02	0.09 (0.12)	-0.05	0.02	0.08 (0.11)	-0.01	0.01	0.07 (0.10)
γ_3 = 0.14	0.01	0.01	0.07 (0.10)	0.07	0.03	0.11 (0.14)	0.06	0.02	0.10 (0.12)	0.03	0.02	0.08 (0.12)
γ_4 = 0.14	0.00	0.01	0.05 (0.07)	-0.06	0.01	0.08 (0.10)	-0.05	0.01	0.08 (0.10)	-0.01	0.01	0.07 (0.09)
ι_0 = -0.25	-0.01	0.16	0.27 (0.41)	0.11	0.69	0.54 (0.83)	-0.02	0.39	0.43 (0.62)	-0.04	0.44	0.44 (0.57)
ι_1 = 0.25	0.00	0.03	0.11 (0.16)	-0.03	0.11	0.21 (0.33)	-0.04	0.07	0.18 (0.26)	-0.01	0.06	0.17 (0.24)
ι_2 = 0.20	0.01	0.01	0.06 (0.09)	0.07	0.04	0.14 (0.19)	0.04	0.03	0.12 (0.17)	0.01	0.03	0.11 (0.15)
ι_3 = 0.25	0.00	0.01	0.07 (0.10)	-0.04	0.05	0.15 (0.21)	-0.04	0.04	0.13 (0.19)	-0.01	0.03	0.12 (0.17)
ι_4 = -0.25	0.00	0.01	0.05 (0.07)	0.05	0.03	0.11 (0.16)	0.05	0.03	0.11 (0.16)	0.01	0.02	0.10 (0.14)
ι_5 = 0.15	0.00	0.03	0.11 (0.16)	0.07	0.10	0.21 (0.31)	-0.02	0.06	0.16 (0.25)	-0.01	0.05	0.16 (0.23)
κ_0 = -0.25	-0.01	0.14	0.25 (0.37)	0.36	0.99	0.70 (0.92)	0.06	0.41	0.41 (0.63)	0.16	0.42	0.41 (0.58)
κ_1 = -0.10	0.00	0.02	0.10 (0.15)	-0.08	0.13	0.23 (0.35)	-0.06	0.07	0.17 (0.25)	-0.02	0.06	0.15 (0.23)
κ_2 = -0.10	0.00	0.01	0.06 (0.08)	0.06	0.04	0.14 (0.19)	0.05	0.03	0.11 (0.16)	0.02	0.02	0.10 (0.14)
κ_3 = 0.10	0.00	0.01	0.06 (0.09)	-0.06	0.06	0.16 (0.23)	-0.05	0.03	0.12 (0.18)	-0.02	0.03	0.11 (0.16)
κ_4 = -0.10	0.00	0.00	0.04 (0.07)	0.05	0.03	0.12 (0.17)	0.05	0.02	0.10 (0.15)	0.00	0.02	0.09 (0.13)
κ_5 = -0.10	0.00	0.02	0.09 (0.14)	0.08	0.11	0.22 (0.33)	0.00	0.06	0.15 (0.25)	0.02	0.06	0.15 (0.23)
κ_6 = -0.10	0.00	0.02	0.10 (0.15)	-0.06	0.16	0.27 (0.39)	0.03	0.08	0.19 (0.29)	0.00	0.07	0.17 (0.26)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error.

TABLE IX
SIMULATION RESULTS FOR THE PARAMETERS OF MISSING DATA MECHANISM, CASE 3 (M = 1,000 REPLICATIONS; N = 1,000 OBSERVATIONS)

	Full Case			MICE			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
ζ_0 = -0.52	0.00	0.16	0.25 (0.41)	0.39	0.33	0.41 (0.45)	0.04	0.20	0.28 (0.44)
ζ_1 = 0.52	0.01	0.02	0.10 (0.16)	-0.47	0.24	0.47 (0.18)	-0.11	0.06	0.15 (0.18)
ζ_2 = -0.52	-0.01	0.01	0.06 (0.09)	0.46	0.22	0.46 (0.12)	0.11	0.04	0.12 (0.12)
ζ_3 = 0.52	0.01	0.01	0.07 (0.10)	-0.47	0.23	0.47 (0.13)	-0.11	0.04	0.13 (0.13)
ζ_4 = -0.52	-0.01	0.01	0.05 (0.07)	0.48	0.23	0.48 (0.11)	0.12	0.04	0.12 (0.11)
ζ_5 = -0.52	-0.02	0.02	0.11 (0.16)	-0.05	0.03	0.12 (0.17)	-0.02	0.03	0.12 (0.17)
ζ_6 = 0.52	0.01	0.02	0.11 (0.16)	0.11	0.07	0.18 (0.24)	0.04	0.06	0.16 (0.22)
ζ_7 = 0.52	0.00	0.02	0.11 (0.14)	0.02	0.06	0.17 (0.24)	0.02	0.05	0.15 (0.22)
ζ_8 = 0.52	0.01	0.02	0.09 (0.12)	0.02	0.02	0.10 (0.15)	0.02	0.02	0.10 (0.15)
ν_0 = 0.18	0.04	0.39	0.41 (0.61)	0.03	0.44	0.42 (0.68)	-0.08	0.64	0.50 (0.83)
ν_1 = 0.18	0.00	0.06	0.16 (0.24)	0.01	0.07	0.17 (0.26)	0.01	0.07	0.17 (0.26)
ν_2 = 0.18	0.01	0.02	0.09 (0.13)	0.02	0.03	0.11 (0.17)	0.03	0.03	0.10 (0.17)
ν_3 = 0.18	-0.01	0.02	0.10 (0.15)	-0.01	0.03	0.12 (0.19)	-0.02	0.03	0.12 (0.18)
ν_4 = 0.18	0.00	0.01	0.07 (0.11)	0.00	0.02	0.10 (0.15)	0.00	0.02	0.10 (0.15)
ν_5 = -0.18	-0.01	0.06	0.17 (0.24)	0.18	0.04	0.18 (0.35)	0.33	0.88	0.40 (0.82)
ν_6 = 0.18	-0.01	0.05	0.16 (0.24)	-0.01	0.13	0.25 (0.35)	-0.03	0.12	0.24 (0.33)
ν_7 = 0.18	0.00	0.05	0.15 (0.22)	0.02	0.13	0.23 (0.37)	0.02	0.12	0.23 (0.34)
ν_8 = 0.18	0.01	0.03	0.12 (0.18)	0.02	0.04	0.14 (0.21)	0.02	0.05	0.14 (0.21)
ν_9 = 0.18	-0.01	0.05	0.16 (0.23)	-0.07	0.06	0.16 (0.23)	0.01	0.07	0.17 (0.24)
υ_0 = -0.14	0.01	0.25	0.35 (0.50)	0.00	0.41	0.41 (0.64)	0.14	0.79	0.50 (1.81)
υ_1 = 0.14	0.01	0.03	0.13 (0.18)	0.00	0.05	0.16 (0.23)	0.00	0.05	0.15 (0.23)
υ_2 = 0.14	0.00	0.01	0.07 (0.10)	-0.01	0.02	0.09 (0.15)	-0.01	0.02	0.10 (0.14)
υ_3 = 0.14	0.01	0.01	0.08 (0.12)	0.00	0.03	0.11 (0.16)	0.00	0.03	0.11 (0.16)
υ_4 = 0.14	0.00	0.01	0.05 (0.08)	0.00	0.02	0.09 (0.14)	0.01	0.02	0.09 (0.13)
υ_5 = -0.14	0.00	0.03	0.12 (0.18)	0.00	0.05	0.16 (0.23)	-0.01	0.06	0.17 (0.23)
υ_6 = -0.14	-0.01	0.04	0.13 (0.19)	0.16	0.04	0.17 (0.34)	-0.02	1.14	0.46 (1.77)
υ_7 = 0.14	0.00	0.03	0.11 (0.17)	0.00	0.07	0.18 (0.28)	0.00	0.07	0.16 (0.26)
υ_8 = 0.14	0.01	0.02	0.08 (0.13)	0.01	0.02	0.10 (0.16)	0.01	0.02	0.10 (0.16)
υ_9 = 0.14	-0.01	0.03	0.12 (0.17)	-0.08	0.05	0.16 (0.21)	-0.03	0.06	0.16 (0.21)
υ_{10} = 0.14	-0.01	0.07	0.16 (0.25)	-0.01	0.09	0.19 (0.30)	0.00	0.10	0.19 (0.30)
ω_0 = -0.10	-0.01	0.20	0.30 (0.44)	0.07	0.30	0.37 (0.56)	0.12	0.30	0.35 (0.53)
ω_1 = -0.10	0.00	0.02	0.10 (0.15)	0.00	0.03	0.12 (0.20)	-0.01	0.03	0.12 (0.19)
ω_2 = -0.10	0.00	0.01	0.06 (0.08)	-0.01	0.01	0.08 (0.12)	-0.01	0.01	0.08 (0.12)
ω_3 = 0.10	0.00	0.01	0.07 (0.10)	0.01	0.02	0.09 (0.14)	0.01	0.02	0.09 (0.13)
ω_4 = 0.10	0.00	0.00	0.05 (0.07)	-0.01	0.01	0.08 (0.12)	-0.01	0.01	0.07 (0.11)
ω_5 = -0.10	0.01	0.02	0.10 (0.15)	0.01	0.04	0.13 (0.20)	0.00	0.04	0.12 (0.19)
ω_6 = 0.10	0.00	0.02	0.10 (0.16)	-0.02	0.05	0.14 (0.23)	-0.03	0.05	0.15 (0.22)
ω_7 = 0.10	0.00	0.02	0.09 (0.14)	-0.10	0.02	0.10 (0.26)	-0.36	0.20	0.36 (0.26)
ω_8 = -0.10	0.00	0.01	0.07 (0.11)	0.00	0.02	0.08 (0.13)	0.00	0.02	0.08 (0.13)

Continued on next page

TABLE IX (Continued)
SIMULATION RESULTS FOR THE PARAMETERS OF MISSING DATA MECHANISM,
CASE 3 (M = 1,000 REPLICATIONS; N = 1,000 OBSERVATIONS)

	Full Case			MICE			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
ω_9 = -0.10	-0.01	0.02	0.10 (0.15)	-0.04	0.03	0.12 (0.18)	0.03	0.03	0.12 (0.18)
ω_{10} = 0.10	0.00	0.04	0.14 (0.22)	0.01	0.07	0.18 (0.26)	0.03	0.07	0.18 (0.26)
ω_{11} = 0.10	0.01	0.03	0.11 (0.17)	0.01	0.04	0.13 (0.20)	0.03	0.04	0.13 (0.20)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error.

TABLE X
SIMULATION RESULTS FOR THE PARAMETERS OF SAMPLE SELECTION MECHANISM, CASE 3
(M = 1,000 REPLICATIONS; N = 1,000 OBSERVATIONS)

	Full Case			Complete Case			MICE			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
δ_0 = -0.50	-0.01	0.86	0.63 (0.92)	-0.03	1.59	0.84 (1.24)	-0.11	0.88	0.66 (0.96)	-0.05	0.89	0.65 (0.95)
δ_1 = 0.10	0.00	0.00	0.01 (0.02)	0.00	0.00	0.02 (0.03)	0.00	0.00	0.01 (0.02)	0.00	0.00	0.01 (0.02)
δ_2 = -1.00	-0.01	0.03	0.13 (0.19)	-0.02	0.06	0.17 (0.25)	0.01	0.04	0.13 (0.19)	-0.01	0.04	0.14 (0.19)
δ_3 = -1.00	-0.01	0.01	0.05 (0.07)	-0.02	0.01	0.07 (0.10)	0.02	0.01	0.06 (0.08)	-0.02	0.01	0.06 (0.08)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error.

It appears, nonetheless, that the additional complexity in the Case 3 simulation has no effect on the speed of convergence of the proposed method relative to the Case 2 simulation. The median iteration is 72 ($(Q_1, Q_3) = (35, 200)$). Such distribution completely overlaps that of the previous simulation. In evaluating the performance of the proposed method here, I decide not to compare it with the augmentation-CR method. Despite having no convergence issue in the preliminary trials, implementation of this modified algorithm almost always results in a Louis information matrix as in Equation 3.17 which is not positive definite for some parameters. With regard to the fraction of missing values, I observe that the actual proportion generally follows the simulation setup, except for X_7 and X_8. Of course, this is mainly caused by their extra constraint of being non-observable (Table II). The rounded average $\bar{N}_{Y_{mis}} = 305$ (31%), $\bar{N}_{X_{5,mis}} = 410$ (41%), and $\bar{N}_{X_{6,mis}} = 107$ (11%). Meanwhile, $\bar{N}_{X_{7,mis}} = 459$ (46%), and $\bar{N}_{X_{8,mis}} = 586$ (59%), which are in extreme contrast to their setup value, 20% and 35%, respectively. The average sample size of this simulation is $\bar{n} = 695$ after rounding to the closest integer. Overall, only an average of 180 from the originally 1000 observations available for a complete case analysis of Y and $X_2, \cdots, X_8$ relationship.

The properties of the estimated parameters for the outcome model (Table VII and Figure 3), the missing covariates distribution (Table VIII), the missing data mechanism (Table IX), and the sampling probability (Table X) within the Case 3 simulation are largely similar to those exhibited in the previous simulation. The proposed method overall tends to perform slightly better than MICE, and is substantially superior to a complete case analysis. This latter approach should be avoided at all cost when estimating the parameters of interest in such case

of survey data, particularly if it is feasible to implement a missing data technique like the augmentation method; the estimates of a complete case analysis are very unstable (Table VII and Figure 3), clearly due to both the non-ignorability of missingness and an extreme loss of data with incomplete observations. As in the Case 2 simulation, however, the estimates for the missing data mechanism based on the proposed method are not quite good for the parameters attached to the corresponding missing covariates, particularly if the covariates are binary.

4.5 Simulation of Case 1 Survey

Traditional surveys are commonly conducted on the population of size much larger than the total samples. This is particularly true for national surveys and a number of periodical cross-sectional studies. Typically, the data from these surveys are only observed among the samples. It is almost impossible to get any information about the non-samples given both the geographical coverage of the survey and an immense population size. The sampling information may sometimes be made available for public to access. However, the chance is much higher that such details are provided to a limited description. For what it is worth, simulation of data from these survey settings is critical for testing the proposed method, because they represent the majority of survey application in the real world.

4.5.1 Situation 1: Sampling Mechanism is Known

On average, the sample size for each simulation based on Case 1 Situation 1 scenario is about $\bar{n} = 1283$. The empirical missing observations among the samples, rounded to the nearest integer, are $\bar{n}_{Y_{mis}} = 0$ (0%), $\bar{n}_{X_{5,mis}} = 568$ (44%), $\bar{n}_{X_{6,mis}} = 125$ (10%), $\bar{n}_{X_{7,mis}} = 246$ (19%), and $\bar{n}_{X_{8,mis}} = 446$ (35%). Two important things to note here. First, I use the notation

n instead of N to refer to the use of sample size as the denominator for obtaining the actual fraction of missing values; in the preceding scenarios, the denominator was all observations in the simulated data. Second, all analyses concerning the incompletely observed data (complete case, MICE, augmentation-CR, and augmentation methods) use the sampled part of the simulated dataset, n; however, the full case analysis is implemented on all observations N. The reason for this choice is Equation 3.7 as provided in the Methodology chapter. To elaborate, the models shown in Table II, Equation 4.1, and Equation 4.2 under Case 1 gives rise to the joint likelihood analog to Equation 3.6 (with notational adjustment), which in terms of estimating the parameters of interest $\boldsymbol{\beta}$, only requires the sampled part of N. For the analysis of the relationship between Y and $X_2, \cdots, X_8$ using complete cases, the study has an average of 344 observations per simulation run, or approximately 27% of samples. There is more variability in the distribution of total iterations before convergence in the proposed method (median 51, (Q_1, Q_3) = (11, 232)) as compared to the previous scenarios. It is obviously not so with the augmentation-CR method, its modification. The latter still requires a median of 16 ((Q_1, Q_3) = (14, 17)) iterations to converge, which is basically identical to its distribution in the Case 2 simulation (recall that the augmentation-CR method was not implemented in the Case 3 simulation).

I observe in this simulation that the augmentation method maintains its superiority over an analysis based on complete cases for the estimates of $\boldsymbol{\beta}$ (Table XI and Figure 4) and those of the missing covariates distribution (Table XII). It also competes well with MICE in all estimates, including the estimated parameters of the missing data mechanism (Table XIII),

and occasionally the augmentation method produces a better estimate than MICE. The issues with the GLM parameters relating R_k and the corresponding missing covariate X_k, however, remains. Removal of the weight component with constant values, which is the approach applied on the augmentation-CR method, appears to circumvent this problem. This alternative version of the proposed method has the benefit of producing fairly stable estimates in all fitted models of the current simulation. Its results are in general closer to MICE than the augmentation method. The parameters of the sample selection are not estimated here, because other than the full case analysis, the computation only involves the sampled portion of the simulated dataset for the reason explained above.

4.5.2 Situation 2: Sampling Mechanism is Not Known

The sample size for each run on the Case 1 Situation 2 simulation averages almost equal to Case 1 Situation 1, where here $\bar{n} = 1282$. Such a similarity also extends to the empirical fraction of missing data: $\bar{n}_{Y_{mis}} = 0$ (0%), $\bar{n}_{X_{5,mis}} = 569$ (44%), $\bar{n}_{X_{6,mis}} = 125$ (10%), $\bar{n}_{X_{7,mis}} = 246$ (19%), and $\bar{n}_{X_{8,mis}} = 446$ (35%). Note that the calculation of these missing proportions follows the same fashion as the previous simulation. In contrast to the Case 1 Situation 1 simulation, however, all analyses (including the full case analysis) utilize the sampled part of the simulated dataset, n. The decision to restrict the full case analysis only to samples is to avoid redundancy. Without this restriction, the results will very likely duplicate those of Case 1 Situation 1 regardless of it being reruns using the simulated data of the present simulation, because of a large N. Besides, opting for n increases the comparability with the

rest of implemented methods. There are approximately 343 observations, on average, available for a complete case analysis of the relationship between Y and $X_2, \cdots, X_8$.

Computation wise, this simulation case seems relatively more straightforward for the proposed method to accomplish. Overall, it only needs a median of 13 $(Q_1, Q_3) = (10, 77)$) iterations for convergence. One may notice that across all simulation cases, this is also the scenario of which the iteration variability of the augmentation method is the least. The speed of convergence of its alternative algorithm, the augmentation-CR method, is interestingly not affected at all. It still completes its estimation by a median of 16 $((Q_1, Q_3) = (15, 18))$ iterations; these numbers are almost identical to the rest of the implemented scenarios. Note that, as stated in Section 4.2, MICE is excluded from the comparison because the current version of R package `mice` does not allow for estimation using survey weight.

Table XIV and Figure 5 show the simulation results for the parameters of interest. The estimates of all non-full case methods appear to be slightly biased either downward or upward (Figure 5). Those of the proposed method, however, remain fairly close to zero for all $\hat{\boldsymbol{\beta}}$. Table XV demonstrates the estimated parameters of the missing covariates distribution. Table XVI shows that the proposed method struggles to estimate particularly $\zeta_0, \cdots, \zeta_4$, ν_5, and ω_7. For the rest of the $\mathbf{R}$ parameters, nevertheless, the estimates from the augmentation method are generally stable.

TABLE XI
SIMULATION RESULTS FOR THE PARAMETERS OF INTEREST, CASE 1 SITUATION 1 ($M = 1{,}000$ REPLICATIONS; $N = 100{,}000$ OBSERVATIONS; $\bar{n} = 1{,}283$ SAMPLED OBSERVATIONS)

	Full Case			Complete Case			MICE			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE(ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
$\beta_0 =$ 0.10	0.00	0.00	0.02 (0.03)	0.24	0.29	0.40 (0.50)	0.01	0.08	0.19(0.28)	0.01	0.08	0.19 (0.28)	0.00	0.08	0.19 (0.28)
$\beta_1 =$ -0.10	0.00	0.00	0.01 (0.01)	-0.05	0.04	0.13 (0.19)	-0.05	0.02	0.08(0.11)	-0.05	0.02	0.08 (0.11)	-0.05	0.01	0.08 (0.11)
$\beta_2 =$ -0.10	0.00	0.00	0.00 (0.01)	0.05	0.01	0.08 (0.11)	0.05	0.01	0.06(0.07)	0.05	0.01	0.06 (0.07)	0.05	0.01	0.06 (0.07)
$\beta_3 =$ -0.10	0.00	0.00	0.00 (0.01)	-0.06	0.02	0.09 (0.12)	-0.05	0.01	0.06(0.08)	-0.05	0.01	0.06 (0.08)	-0.05	0.01	0.06 (0.08)
$\beta_4 =$ -0.10	0.00	0.00	0.00 (0.01)	0.06	0.01	0.07 (0.09)	0.05	0.01	0.06(0.06)	0.05	0.01	0.06 (0.06)	0.05	0.01	0.06 (0.06)
$\beta_5 =$ -0.10	0.00	0.00	0.01 (0.01)	0.06	0.03	0.13 (0.18)	-0.01	0.01	0.07(0.11)	-0.01	0.01	0.07 (0.11)	0.00	0.01	0.07 (0.11)
$\beta_6 =$ -0.10	0.00	0.00	0.01 (0.01)	-0.04	0.04	0.13 (0.19)	0.02	0.01	0.08(0.12)	0.02	0.01	0.08 (0.11)	0.02	0.01	0.08 (0.11)
$\beta_7 =$ 0.10	0.00	0.00	0.01 (0.01)	-0.06	0.03	0.11 (0.17)	0.01	0.02	0.08(0.13)	0.01	0.02	0.08 (0.12)	0.02	0.02	0.08 (0.12)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error.

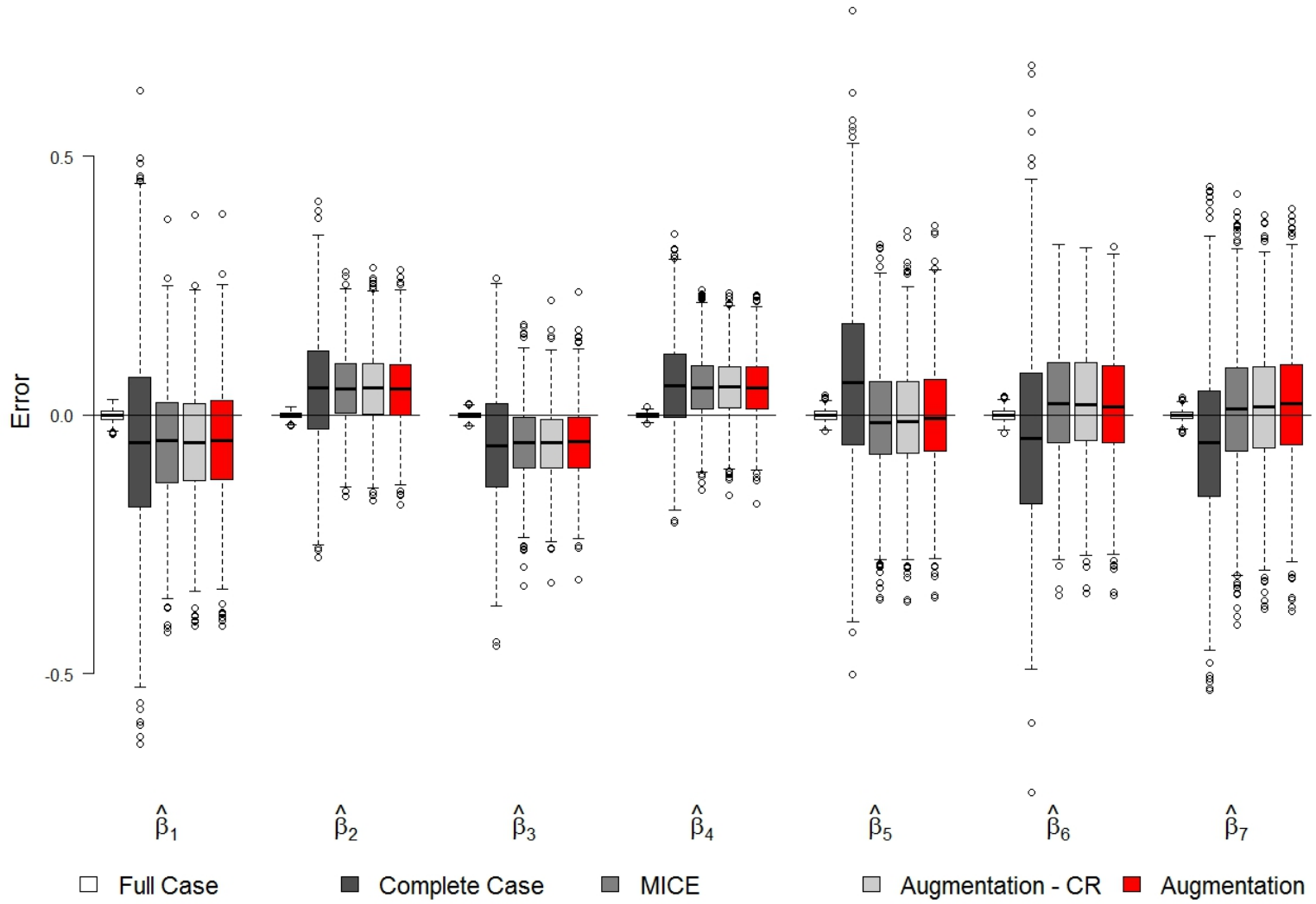

Figure 4. Distribution of Errors in the Parameters of Interest on Case 1 Situation 1 Simulation

TABLE XII
SIMULATION RESULTS FOR THE COVARIATES PARAMETERS, CASE 1 SITUATION 1 ($M = 1{,}000$ REPLICATIONS; $N = 100{,}000$ OBSERVATIONS; $\bar{n} = 1{,}283$ SAMPLED OBSERVATIONS)

	Full Case			Complete Case			MICE			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
α_0 = -1.00	0.00	0.00	0.01 (0.02)	-0.01	0.04	0.14 (0.20)	0.03	0.05	0.15 (0.21)	0.01	0.04	0.14 (0.19)	0.02	0.05	0.15 (0.19)
α_1 = 1.00	0.00	0.00	0.00 (0.01)	0.00	0.01	0.05 (0.07)	-0.01	0.01	0.05 (0.08)	0.00	0.01	0.05 (0.07)	0.00	0.01	0.05 (0.07)
α_2 = -1.00	0.00	0.00	0.00 (0.00)	0.01	0.00	0.02 (0.03)	0.01	0.00	0.02 (0.03)	0.01	0.00	0.02 (0.03)	0.01	0.00	0.02 (0.03)
α_3 = 1.00	0.00	0.00	0.00 (0.00)	-0.01	0.00	0.03 (0.04)	-0.01	0.00	0.03 (0.04)	-0.01	0.00	0.02 (0.03)	-0.01	0.00	0.03 (0.03)
$\sigma^2_{X_5}$ = 1.00	0.00	0.00	0.00 (.)	-0.07	0.01	0.07 (.)	.	.	. (.)	-0.07	0.01	0.07 (0.05)	-0.07	0.01	0.07 (0.05)
γ_0 = 0.14	0.00	0.00	0.03 (0.04)	-0.30	0.34	0.39 (0.49)	-0.08	0.16	0.27 (0.39)	-0.09	0.16	0.26 (0.39)	-0.15	0.20	0.29 (0.38)
γ_1 = -0.14	0.00	0.00	0.01 (0.02)	0.05	0.04	0.14 (0.20)	0.06	0.03	0.12 (0.17)	0.06	0.03	0.12 (0.16)	0.06	0.03	0.12 (0.16)
γ_2 = -0.14	0.00	0.00	0.01 (0.01)	-0.05	0.01	0.08 (0.11)	-0.06	0.01	0.08 (0.11)	-0.06	0.01	0.08 (0.10)	-0.05	0.01	0.07 (0.10)
γ_3 = 0.14	0.00	0.00	0.01 (0.01)	0.07	0.02	0.10 (0.12)	0.06	0.02	0.09 (0.11)	0.07	0.02	0.09 (0.11)	0.07	0.02	0.09 (0.11)
γ_4 = 0.14	0.00	0.00	0.00 (0.01)	-0.06	0.01	0.07 (0.09)	-0.06	0.01	0.08 (0.09)	-0.06	0.01	0.07 (0.09)	-0.05	0.01	0.07 (0.09)
ι_0 = -0.25	0.00	0.00	0.03 (0.04)	0.16	0.41	0.41 (0.59)	-0.05	0.19	0.30 (0.42)	-0.05	0.18	0.29 (0.41)	-0.02	0.24	0.31 (0.41)
ι_1 = 0.25	0.00	0.00	0.01 (0.02)	-0.05	0.06	0.17 (0.23)	-0.05	0.03	0.13 (0.18)	-0.05	0.03	0.12 (0.17)	-0.06	0.03	0.12 (0.17)
ι_2 = 0.20	0.00	0.00	0.01 (0.01)	0.07	0.02	0.11 (0.14)	0.06	0.02	0.09 (0.12)	0.06	0.02	0.09 (0.11)	0.06	0.02	0.09 (0.11)
ι_3 = 0.25	0.00	0.00	0.01 (0.01)	-0.05	0.03	0.11 (0.15)	-0.05	0.02	0.10 (0.13)	-0.05	0.02	0.09 (0.12)	-0.05	0.02	0.10 (0.12)
ι_4 = -0.25	0.00	0.00	0.00 (0.01)	0.06	0.01	0.08 (0.11)	0.06	0.01	0.08 (0.11)	0.06	0.01	0.08 (0.10)	0.06	0.01	0.08 (0.10)
ι_5 = 0.15	0.00	0.00	0.01 (0.02)	0.06	0.06	0.16 (0.22)	-0.02	0.03	0.12 (0.17)	-0.03	0.03	0.12 (0.17)	-0.02	0.03	0.12 (0.16)
κ_0 = -0.25	0.00	0.00	0.03 (0.04)	0.41	0.64	0.57 (0.67)	0.07	0.21	0.30 (0.45)	0.08	0.20	0.32 (0.43)	0.20	0.25	0.34 (0.43)
κ_1 = -0.10	0.00	0.00	0.01 (0.01)	-0.06	0.06	0.17 (0.25)	-0.06	0.04	0.13 (0.18)	-0.06	0.03	0.12 (0.17)	-0.05	0.03	0.13 (0.17)
κ_2 = -0.10	0.00	0.00	0.01 (0.01)	0.06	0.03	0.11 (0.15)	0.06	0.02	0.10 (0.12)	0.07	0.02	0.10 (0.11)	0.07	0.02	0.10 (0.11)
κ_3 = 0.10	0.00	0.00	0.01 (0.01)	-0.07	0.03	0.12 (0.16)	-0.06	0.02	0.10 (0.13)	-0.07	0.02	0.09 (0.12)	-0.07	0.02	0.10 (0.12)
κ_4 = -0.10	0.00	0.00	0.00 (0.01)	0.06	0.02	0.09 (0.12)	0.06	0.01	0.08 (0.10)	0.06	0.01	0.08 (0.10)	0.05	0.01	0.08 (0.10)
κ_5 = -0.10	0.00	0.00	0.01 (0.01)	0.06	0.07	0.18 (0.24)	-0.01	0.03	0.12 (0.18)	-0.01	0.03	0.11 (0.17)	0.00	0.03	0.12 (0.17)
κ_6 = -0.10	0.00	0.00	0.01 (0.01)	-0.06	0.07	0.18 (0.25)	0.03	0.03	0.12 (0.18)	0.03	0.03	0.11 (0.17)	0.02	0.03	0.11 (0.17)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error; { . }, not estimated.

TABLE XIII
SIMULATION RESULTS FOR THE PARAMETERS OF MISSING DATA MECHANISM, CASE 1
SITUATION 1 ($M = 1{,}000$ REPLICATIONS; $N = 100{,}000$ OBSERVATIONS; $\bar{n} = 1{,}283$ SAMPLED OBSERVATIONS)

	Full Case			MICE			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
ζ_0 = -0.52	0.00	0.00	0.03 (0.04)	0.33	0.24	0.37 (0.37)	0.34	0.24	0.37 (0.37)	0.30	0.23	0.34 (0.37)
ζ_1 = 0.52	0.00	0.00	0.01 (0.02)	-0.52	0.29	0.52 (0.15)	-0.53	0.30	0.53 (0.15)	-0.51	0.29	0.50 (0.15)
ζ_2 = -0.52	0.00	0.00	0.01 (0.01)	0.51	0.26	0.51 (0.10)	0.51	0.27	0.51 (0.10)	0.50	0.26	0.49 (0.10)
ζ_3 = 0.52	0.00	0.00	0.01 (0.01)	-0.52	0.27	0.52 (0.11)	-0.52	0.28	0.52 (0.11)	-0.50	0.27	0.50 (0.11)
ζ_4 = -0.52	0.00	0.00	0.01 (0.01)	0.52	0.28	0.53 (0.09)	0.53	0.28	0.53 (0.09)	0.51	0.28	0.52 (0.09)
ζ_5 = -0.52	0.00	0.00	0.01 (0.02)	-0.04	0.02	0.10 (0.15)	-0.04	0.02	0.10 (0.15)	-0.04	0.02	0.10 (0.15)
ζ_6 = 0.52	0.00	0.00	0.01 (0.02)	0.11	0.03	0.13 (0.15)	0.11	0.03	0.13 (0.14)	0.10	0.03	0.12 (0.14)
ζ_7 = 0.52	0.00	0.00	0.01 (0.01)	0.02	0.03	0.11 (0.16)	0.03	0.03	0.11 (0.15)	0.03	0.03	0.11 (0.15)
ζ_8 = 0.52	0.00	0.00	0.01 (0.01)	0.03	0.01	0.07 (0.11)	0.03	0.01	0.07 (0.11)	0.03	0.01	0.08 (0.11)
ν_0 = 0.18	0.00	0.00	0.04 (0.06)	0.00	0.34	0.39 (0.61)	-0.02	0.33	0.41 (0.70)	-0.10	0.50	0.44 (0.81)
ν_1 = 0.18	0.00	0.00	0.02 (0.02)	0.00	0.06	0.16 (0.24)	0.00	0.06	0.16 (0.24)	0.00	0.06	0.16 (0.24)
ν_2 = 0.18	0.00	0.00	0.01 (0.01)	0.01	0.03	0.11 (0.16)	0.01	0.03	0.11 (0.16)	0.03	0.03	0.11 (0.16)
ν_3 = 0.18	0.00	0.00	0.01 (0.01)	-0.01	0.03	0.11 (0.17)	-0.01	0.03	0.10 (0.17)	-0.02	0.03	0.11 (0.17)
ν_4 = 0.18	0.00	0.00	0.01 (0.01)	0.01	0.02	0.10 (0.15)	0.00	0.02	0.09 (0.14)	0.00	0.02	0.09 (0.14)
ν_5 = -0.18	0.00	0.00	0.02 (0.02)	0.17	0.04	0.17 (0.33)	0.18	0.04	0.19 (0.64)	0.41	0.77	0.32 (0.85)
ν_6 = 0.18	0.00	0.00	0.02 (0.02)	0.00	0.06	0.17 (0.25)	0.00	0.06	0.16 (0.24)	-0.02	0.06	0.16 (0.24)
ν_7 = 0.18	0.00	0.00	0.01 (0.02)	0.00	0.07	0.18 (0.27)	0.01	0.07	0.18 (0.26)	0.00	0.07	0.18 (0.26)
ν_8 = 0.18	0.00	0.00	0.01 (0.02)	0.01	0.03	0.12 (0.17)	0.01	0.03	0.12 (0.17)	0.01	0.03	0.12 (0.17)
ν_9 = 0.18	0.00	0.00	0.01 (0.02)	-0.07	0.05	0.15 (0.21)	-0.07	0.05	0.15 (0.21)	-0.03	0.05	0.15 (0.22)
υ_0 = -0.14	0.00	0.00	0.03 (0.05)	0.00	0.23	0.33 (0.49)	0.00	0.23	0.33 (0.50)	0.11	0.38	0.37 (0.75)
υ_1 = 0.14	0.00	0.00	0.01 (0.02)	0.00	0.03	0.13 (0.18)	0.00	0.03	0.12 (0.18)	0.00	0.03	0.12 (0.18)
υ_2 = 0.14	0.00	0.00	0.01 (0.01)	-0.01	0.01	0.08 (0.12)	-0.01	0.01	0.08 (0.12)	-0.01	0.02	0.08 (0.12)
υ_3 = 0.14	0.00	0.00	0.01 (0.01)	0.00	0.02	0.08 (0.13)	0.00	0.02	0.08 (0.13)	0.00	0.02	0.08 (0.13)
υ_4 = 0.14	0.00	0.00	0.01 (0.01)	0.00	0.01	0.07 (0.11)	0.00	0.01	0.07 (0.10)	0.00	0.01	0.07 (0.10)
υ_5 = -0.14	0.00	0.00	0.01 (0.02)	-0.01	0.03	0.12 (0.18)	-0.01	0.03	0.12 (0.18)	-0.01	0.03	0.12 (0.18)
υ_6 = -0.14	0.00	0.00	0.01 (0.02)	0.15	0.03	0.15 (0.24)	0.14	0.02	0.14 (0.36)	0.04	0.59	0.21 (0.68)
υ_7 = 0.14	0.00	0.00	0.01 (0.02)	0.01	0.04	0.14 (0.20)	0.02	0.04	0.14 (0.19)	0.01	0.04	0.14 (0.19)

Continued on next page

TABLE XIII (Continued)
SIMULATION RESULTS FOR THE PARAMETERS OF MISSING DATA MECHANISM, CASE 1 SITUATION 1 ($M = 1{,}000$ REPLICATIONS; $N = 100{,}000$ OBSERVATIONS; $\bar{n} = 1{,}283$ SAMPLED OBSERVATIONS)

	Full Case			MICE			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
$\upsilon_8 = 0.14$	0.00	0.00	0.01 (0.01)	0.02	0.02	0.08 (0.13)	0.02	0.02	0.08 (0.13)	0.02	0.02	0.08 (0.13)
$\upsilon_9 = 0.14$	0.00	0.00	0.01 (0.02)	-0.09	0.03	0.12 (0.15)	-0.09	0.03	0.12 (0.16)	-0.08	0.04	0.13 (0.16)
$\upsilon_{10} = 0.14$	0.00	0.00	0.02 (0.02)	-0.01	0.06	0.16 (0.24)	-0.01	0.06	0.16 (0.24)	0.00	0.06	0.17 (0.24)
$\omega_0 = -0.10$	0.00	0.00	0.03 (0.04)	0.08	0.17	0.30 (0.42)	0.08	0.17	0.29 (0.42)	0.16	0.20	0.29 (0.42)
$\omega_1 = -0.10$	0.00	0.00	0.01 (0.01)	-0.01	0.02	0.10 (0.15)	-0.01	0.02	0.10 (0.15)	-0.02	0.02	0.10 (0.15)
$\omega_2 = -0.10$	0.00	0.00	0.01 (0.01)	0.00	0.01	0.06 (0.10)	0.00	0.01	0.06 (0.09)	0.00	0.01	0.06 (0.09)
$\omega_3 = 0.10$	0.00	0.00	0.01 (0.01)	0.00	0.01	0.07 (0.10)	0.00	0.01	0.07 (0.10)	0.01	0.01	0.07 (0.10)
$\omega_4 = 0.10$	0.00	0.00	0.00 (0.01)	0.00	0.01	0.06 (0.09)	0.00	0.01	0.05 (0.08)	0.00	0.01	0.06 (0.08)
$\omega_5 = -0.10$	0.00	0.00	0.01 (0.01)	-0.01	0.02	0.11 (0.15)	-0.01	0.02	0.11 (0.15)	-0.01	0.02	0.11 (0.15)
$\omega_6 = 0.10$	0.00	0.00	0.01 (0.02)	0.01	0.02	0.10 (0.15)	0.01	0.02	0.10 (0.15)	0.00	0.02	0.10 (0.15)
$\omega_7 = 0.10$	0.00	0.00	0.01 (0.01)	-0.10	0.01	0.10 (0.19)	-0.10	0.01	0.10 (0.21)	-0.36	0.19	0.37 (0.21)
$\omega_8 = -0.10$	0.00	0.00	0.01 (0.01)	0.00	0.01	0.07 (0.10)	0.00	0.01	0.07 (0.10)	0.00	0.01	0.07 (0.10)
$\omega_9 = -0.10$	0.00	0.00	0.01 (0.01)	-0.03	0.02	0.08 (0.13)	-0.03	0.02	0.08 (0.13)	0.00	0.02	0.09 (0.13)
$\omega_{10} = 0.10$	0.00	0.00	0.01 (0.02)	0.00	0.04	0.14 (0.20)	0.00	0.04	0.14 (0.20)	0.02	0.04	0.14 (0.21)
$\omega_{11} = 0.10$	0.00	0.00	0.01 (0.02)	0.00	0.02	0.11 (0.15)	0.00	0.02	0.11 (0.15)	0.01	0.03	0.11 (0.16)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error.

TABLE XIV
SIMULATION RESULTS FOR THE PARAMETERS OF INTEREST, CASE 1 SITUATION 2 (M = 1,000 REPLICATIONS; N = 100,000 OBSERVATIONS; $\bar{n}$ = 1,282 SAMPLED OBSERVATIONS)

	Full Case			Complete Case			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE(ASE)	Bias	MSE	MAE (ASE)
β_0 = 0.10	0.00	0.11	0.22 (0.32)	0.27	0.43	0.45 (0.56)	0.04	0.11	0.22(0.06)	0.03	0.11	0.22 (0.06)
β_1 = -0.10	-0.01	0.02	0.08 (0.13)	-0.06	0.06	0.17 (0.21)	-0.06	0.02	0.09(0.04)	-0.05	0.02	0.09 (0.04)
β_2 = -0.10	0.00	0.01	0.05 (0.07)	0.04	0.02	0.10 (0.12)	0.05	0.01	0.07(0.13)	0.04	0.01	0.07 (0.13)
β_3 = -0.10	0.00	0.01	0.05 (0.08)	-0.06	0.02	0.10 (0.14)	-0.06	0.01	0.07(0.31)	-0.05	0.01	0.07 (0.31)
β_4 = -0.10	0.00	0.00	0.04 (0.06)	0.05	0.01	0.07 (0.10)	0.05	0.01	0.06(0.20)	0.05	0.01	0.06 (0.20)
β_5 = -0.10	0.01	0.01	0.08 (0.13)	0.07	0.05	0.15 (0.21)	0.00	0.02	0.08(0.05)	0.00	0.02	0.08 (0.05)
β_6 = -0.10	0.00	0.02	0.08 (0.12)	-0.04	0.05	0.15 (0.21)	0.02	0.02	0.09(0.05)	0.02	0.02	0.09 (0.05)
β_7 = 0.10	0.00	0.01	0.08 (0.12)	-0.06	0.04	0.14 (0.19)	0.01	0.02	0.09(0.05)	0.01	0.02	0.09 (0.05)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error.

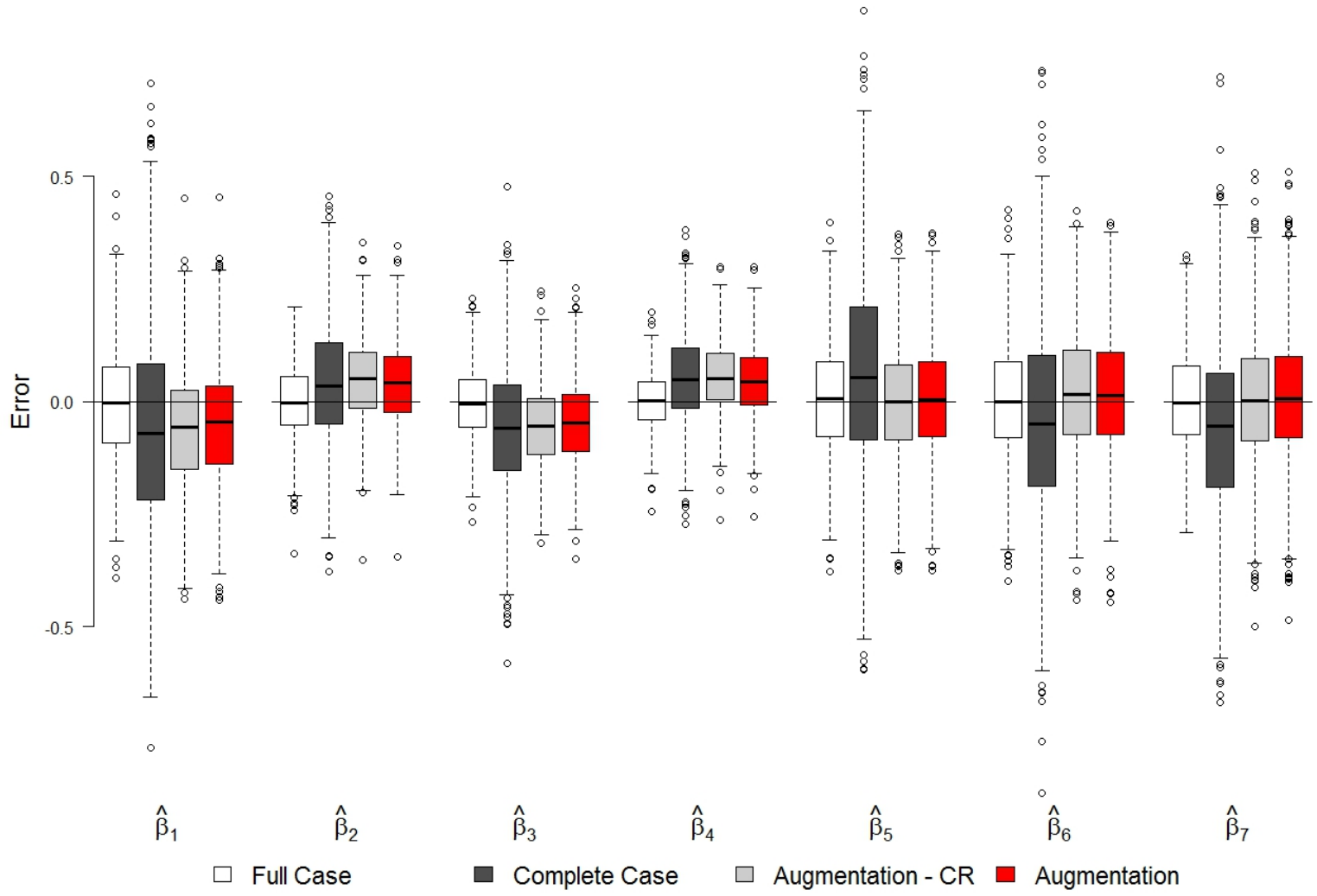

Figure 5. Distribution of Errors in the Parameters of Interest on Case 1 Situation 2 Simulation

TABLE XV
SIMULATION RESULTS FOR THE COVARIATES PARAMETERS, CASE 1 SITUATION 2 (M = 1,000 REPLICATIONS; N = 100,000 OBSERVATIONS; $\bar{n}$ = 1,282 SAMPLED OBSERVATIONS)

	Full Case			Complete Case			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
α_0 = -1.00	0.00	0.04	0.14 (0.19)	-0.20	0.11	0.23 (0.25)	-0.18	0.10	0.22 (0.13)	-0.15	0.09	0.21 (0.13)
α_1 = 1.00	0.00	0.00	0.05 (0.07)	-0.01	0.01	0.06 (0.09)	-0.01	0.01	0.06 (0.10)	0.00	0.01	0.06 (0.10)
α_2 = -1.00	0.00	0.00	0.02 (0.03)	0.01	0.00	0.03 (0.03)	0.01	0.00	0.03 (0.28)	0.01	0.00	0.03 (0.28)
α_3 = 1.00	0.00	0.00	0.02 (0.04)	-0.01	0.00	0.03 (0.05)	-0.01	0.00	0.03 (0.70)	-0.01	0.00	0.03 (0.70)
$\sigma^2_{X_5}$ = 1.00	0.00	0.00	0.03 (.)	-0.02	0.01	0.05 (.)	-0.02	0.01	0.05 (0.09)	-0.02	0.01	0.05 (0.09)
γ_0 = 0.14	-0.01	0.19	0.30 (0.44)	-0.28	0.45	0.45 (0.59)	-0.09	0.24	0.34 (0.05)	-0.14	0.27	0.36 (0.05)
γ_1 = -0.14	-0.01	0.03	0.12 (0.18)	0.05	0.06	0.16 (0.24)	0.06	0.05	0.14 (0.03)	0.05	0.05	0.14 (0.03)
γ_2 = -0.14	0.00	0.01	0.07 (0.10)	-0.06	0.02	0.09 (0.13)	-0.06	0.02	0.08 (0.10)	-0.05	0.02	0.09 (0.10)
γ_3 = 0.14	0.00	0.01	0.07 (0.11)	0.07	0.03	0.11 (0.15)	0.07	0.02	0.10 (0.23)	0.06	0.02	0.10 (0.24)
γ_4 = 0.14	0.00	0.01	0.06 (0.08)	-0.06	0.02	0.09 (0.11)	-0.06	0.02	0.08 (0.15)	-0.05	0.02	0.09 (0.15)
ι_0 = -0.25	0.02	0.19	0.30 (0.43)	0.18	0.58	0.51 (0.70)	-0.02	0.25	0.35 (0.05)	0.01	0.31	0.39 (0.05)
ι_1 = 0.25	0.00	0.03	0.10 (0.17)	-0.05	0.08	0.19 (0.27)	-0.05	0.04	0.14 (0.04)	-0.05	0.04	0.14 (0.04)
ι_2 = 0.20	0.00	0.01	0.07 (0.10)	0.06	0.03	0.12 (0.16)	0.05	0.02	0.10 (0.09)	0.04	0.02	0.10 (0.09)
ι_3 = 0.25	0.00	0.01	0.07 (0.11)	-0.05	0.04	0.13 (0.18)	-0.05	0.02	0.11 (0.26)	-0.04	0.02	0.11 (0.26)
ι_4 = -0.25	0.00	0.01	0.05 (0.08)	0.05	0.02	0.09 (0.13)	0.05	0.02	0.09 (0.19)	0.04	0.02	0.09 (0.19)
ι_5 = 0.15	0.00	0.03	0.12 (0.17)	0.06	0.08	0.18 (0.27)	-0.02	0.04	0.13 (0.04)	-0.01	0.04	0.13 (0.04)
κ_0 = -0.25	0.00	0.18	0.29 (0.41)	0.41	0.86	0.64 (0.80)	0.09	0.30	0.36 (0.06)	0.20	0.33	0.38 (0.06)
κ_1 = -0.10	0.01	0.03	0.11 (0.16)	-0.06	0.10	0.20 (0.30)	-0.06	0.05	0.14 (0.04)	-0.04	0.05	0.14 (0.05)
κ_2 = -0.10	0.00	0.01	0.06 (0.09)	0.06	0.04	0.13 (0.17)	0.06	0.03	0.11 (0.14)	0.06	0.03	0.11 (0.14)
κ_3 = 0.10	0.00	0.01	0.07 (0.10)	-0.07	0.05	0.15 (0.19)	-0.07	0.03	0.11 (0.32)	-0.06	0.03	0.11 (0.32)
κ_4 = -0.10	0.00	0.01	0.05 (0.07)	0.05	0.03	0.11 (0.14)	0.06	0.02	0.10 (0.21)	0.05	0.02	0.10 (0.21)
κ_5 = -0.10	0.00	0.03	0.10 (0.16)	0.06	0.09	0.20 (0.29)	-0.01	0.04	0.14 (0.05)	0.00	0.04	0.14 (0.05)
κ_6 = -0.10	0.00	0.02	0.10 (0.15)	-0.06	0.09	0.21 (0.30)	0.02	0.04	0.14 (0.05)	0.01	0.04	0.14 (0.05)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error; { . }, not estimated.

TABLE XVI
SIMULATION RESULTS FOR THE PARAMETERS OF MISSING DATA MECHANISM, CASE 1 SITUATION 2 (M = 1,000 REPLICATIONS; N = 100,000 OBSERVATIONS; $\bar{n}$ = 1,282 SAMPLED OBSERVATIONS)

	Full Case			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
ζ_0 = -0.52	-0.02	0.20	0.30 (0.45)	0.42	0.36	0.45 (0.05)	0.31	0.31	0.39 (0.05)
ζ_1 = 0.52	0.01	0.03	0.11 (0.17)	-0.53	0.30	0.52 (0.04)	-0.44	0.24	0.46 (0.04)
ζ_2 = -0.52	-0.01	0.01	0.07 (0.10)	0.51	0.27	0.51 (0.10)	0.43	0.22	0.47 (0.10)
ζ_3 = 0.52	0.01	0.01	0.07 (0.11)	-0.52	0.28	0.52 (0.25)	-0.43	0.22	0.47 (0.25)
ζ_4 = -0.52	-0.01	0.01	0.06 (0.08)	0.53	0.28	0.53 (0.17)	0.44	0.23	0.50 (0.17)
ζ_5 = -0.52	0.00	0.03	0.12 (0.17)	-0.04	0.03	0.12 (0.04)	-0.03	0.03	0.12 (0.04)
ζ_6 = 0.52	0.01	0.03	0.11 (0.16)	0.11	0.04	0.13 (0.04)	0.08	0.04	0.13 (0.04)
ζ_7 = 0.52	0.01	0.02	0.11 (0.15)	0.03	0.04	0.13 (0.03)	0.03	0.04	0.13 (0.03)
ζ_8 = 0.52	0.01	0.02	0.09 (0.13)	0.03	0.02	0.10 (0.03)	0.02	0.02	0.10 (0.03)
ν_0 = 0.18	0.03	0.48	0.49 (0.69)	-0.01	0.51	0.49 (0.03)	-0.12	0.63	0.54 (0.03)
ν_1 = 0.18	0.00	0.07	0.17 (0.26)	0.00	0.08	0.19 (0.02)	0.00	0.09	0.19 (0.02)
ν_2 = 0.18	0.00	0.02	0.10 (0.15)	0.02	0.04	0.13 (0.06)	0.03	0.04	0.13 (0.06)
ν_3 = 0.18	0.01	0.03	0.11 (0.17)	-0.01	0.04	0.13 (0.15)	-0.01	0.04	0.14 (0.15)
ν_4 = 0.18	0.00	0.02	0.08 (0.12)	0.00	0.03	0.12 (0.10)	0.00	0.03	0.12 (0.10)
ν_5 = -0.18	-0.02	0.08	0.19 (0.27)	0.17	0.04	0.19 (0.04)	0.39	0.61	0.30 (0.04)
ν_6 = 0.18	-0.01	0.06	0.18 (0.25)	0.00	0.09	0.20 (0.02)	-0.02	0.09	0.20 (0.02)
ν_7 = 0.18	-0.01	0.06	0.16 (0.24)	0.00	0.11	0.21 (0.02)	0.00	0.11	0.21 (0.02)
ν_8 = 0.18	0.02	0.05	0.14 (0.21)	0.02	0.05	0.14 (0.02)	0.02	0.05	0.14 (0.02)
ν_9 = 0.18	0.01	0.06	0.17 (0.25)	-0.06	0.06	0.17 (0.02)	-0.01	0.06	0.16 (0.02)
υ_0 = -0.14	0.01	0.32	0.38 (0.56)	0.01	0.34	0.40 (0.04)	0.12	0.57	0.42 (0.04)
υ_1 = 0.14	0.00	0.04	0.13 (0.20)	-0.01	0.05	0.14 (0.03)	-0.01	0.05	0.15 (0.03)
υ_2 = 0.14	0.00	0.01	0.08 (0.11)	0.00	0.02	0.10 (0.08)	0.00	0.02	0.10 (0.08)
υ_3 = 0.14	0.00	0.02	0.09 (0.13)	-0.01	0.02	0.11 (0.20)	-0.01	0.03	0.12 (0.20)
υ_4 = 0.14	0.00	0.01	0.06 (0.09)	0.01	0.02	0.09 (0.13)	0.01	0.02	0.09 (0.13)
υ_5 = -0.14	-0.01	0.04	0.13 (0.21)	-0.01	0.05	0.15 (0.03)	-0.01	0.05	0.15 (0.03)
υ_6 = -0.14	-0.01	0.04	0.13 (0.20)	0.15	0.03	0.14 (0.04)	0.02	0.72	0.19 (0.04)
υ_7 = 0.14	0.00	0.04	0.13 (0.18)	0.01	0.06	0.17 (0.03)	0.00	0.06	0.17 (0.03)
υ_8 = 0.14	0.02	0.03	0.11 (0.15)	0.02	0.03	0.11 (0.03)	0.02	0.03	0.11 (0.03)
υ_9 = 0.14	0.00	0.04	0.13 (0.19)	-0.08	0.04	0.14 (0.03)	-0.06	0.05	0.15 (0.03)
υ_{10} = 0.14	-0.01	0.09	0.17 (0.28)	-0.01	0.09	0.17 (0.04)	0.00	0.09	0.18 (0.04)
ω_0 = -0.10	0.03	0.25	0.31 (0.49)	0.12	0.27	0.35 (0.05)	0.19	0.30	0.37 (0.05)
ω_1 = -0.10	0.00	0.03	0.11 (0.16)	0.00	0.03	0.12 (0.03)	-0.01	0.03	0.12 (0.03)
ω_2 = -0.10	0.00	0.01	0.07 (0.09)	0.00	0.01	0.08 (0.10)	-0.01	0.01	0.08 (0.10)
ω_3 = 0.10	0.00	0.01	0.07 (0.11)	0.00	0.02	0.08 (0.24)	0.00	0.02	0.08 (0.24)
ω_4 = 0.10	0.00	0.01	0.05 (0.08)	0.00	0.01	0.07 (0.16)	0.00	0.01	0.07 (0.16)
ω_5 = -0.10	-0.01	0.03	0.12 (0.17)	-0.01	0.03	0.12 (0.04)	-0.01	0.03	0.12 (0.04)
ω_6 = 0.10	0.00	0.03	0.11 (0.16)	-0.01	0.04	0.13 (0.04)	-0.02	0.04	0.13 (0.04)
ω_7 = 0.10	0.00	0.02	0.10 (0.15)	-0.10	0.01	0.10 (0.04)	-0.34	0.18	0.35 (0.04)

Continued on next page

TABLE XVI (Continued)
SIMULATION RESULTS FOR THE PARAMETERS OF MISSING DATA MECHANISM, CASE 1 SITUATION 2 (M = 1,000 REPLICATIONS; N = 100,000 OBSERVATIONS; $\bar{n}$ = 1,282 SAMPLED OBSERVATIONS)

	Full Case			Augmentation-CR			Augmentation		
Parameter	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)	Bias	MSE	MAE (ASE)
ω_8 = -0.10	0.00	0.02	0.08 (0.12)	0.00	0.02	0.07 (0.03)	0.01	0.02	0.08 (0.03)
ω_9 = -0.10	0.00	0.03	0.11 (0.16)	-0.03	0.03	0.11 (0.03)	0.01	0.03	0.11 (0.03)
ω_{10} = 0.10	-0.01	0.06	0.16 (0.24)	0.00	0.06	0.16 (0.04)	0.02	0.06	0.16 (0.04)
ω_{11} = 0.10	0.00	0.03	0.12 (0.18)	0.00	0.03	0.12 (0.04)	0.01	0.04	0.13 (0.04)

[a] MSE, mean squared error; MAE, median absolute error; ASE, asymptotic standard error.

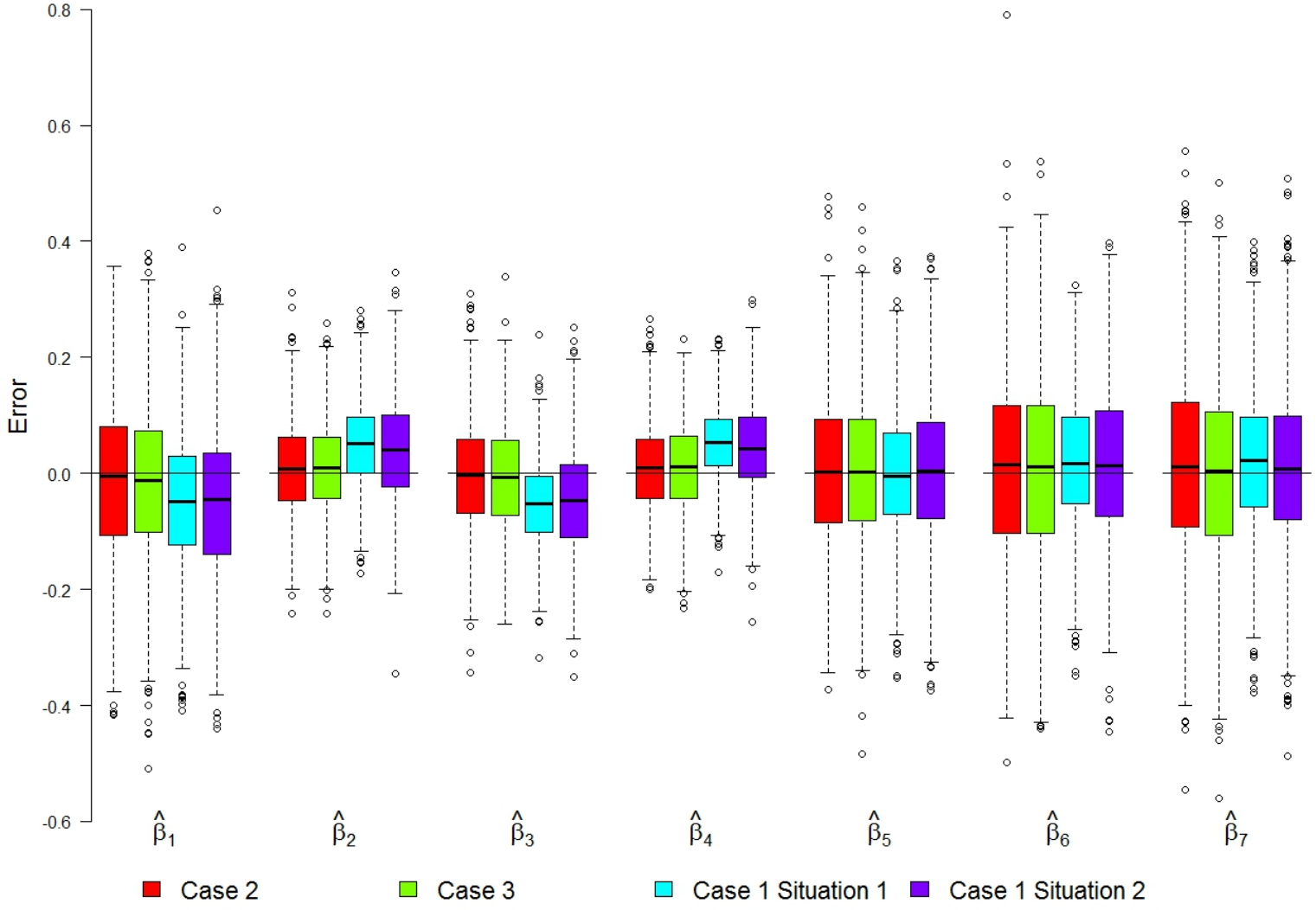

Figure 6. Distribution of Errors in the Parameters of Interest on All Survey Cases for the Augmentation Method

Figure 6 sums up the results using the proposed method in terms of the parameters $\boldsymbol{\beta}$ on all simulated cases. Those of Case 2 and 3 are the least biased estimates. In Case 1, the estimates seem to behave reasonably, although not as good as in the previous two survey classes.

4.6 Discussion

I present in this chapter the simulation studies of survey data with non-ignorable missing covariates. Three classes of survey data are defined: those of which none (Case 1), all (Case 2), or some (Case 3) of the covariates are observable outside the samples. For Case 1, two situations regarding data analyst's knowledge of sampling design are devised and simulated. These includes such situations where the functional form of sample selection is known (Situation 1) and when it is not(Situation 2). The purpose of these exercises is to evaluate the performance of the proposed method, which concerns a procedure for analyzing survey data with missing covariates based on the likelihood approach. This procedure, termed the augmentation assisted EM algorithm, or simply the augmentation method, is argued to have the desirable properties of the maximum likelihood estimation, while flexible enough to handle both continuous and categorical missing covariates, and can adapt the use of survey weight to improve inference.

Overall, the proposed method indicates a reliable performance throughout all classes of survey data that are evaluated. In terms of unbiasedness of $\widehat{\boldsymbol{\beta}}$, the estimated parameters of interest, it tends to produce a bias closer to zero than the multiple imputation technique MICE particularly in Case 2 and Case 3, and is generally far superior to a complete case analysis. The same is true for the nuisance parameters. Note, however, that it may be unusual for practitioners to apply MICE for estimating the whole "true" parameters of a joint missing

data likelihood; MICE is conceptually opposing the joint likelihood approach (13; 19; 17), and its algorithm proceeds by creating parameters for the fully conditional models that may not correspond to the true parameters in the joint distribution of the missing data. I use MICE for computing the nuisance parameters merely for a comparison purpose. The efficiency of the proposed method appears to be comparable with MICE, but of course, extremely better than a complete case analysis. These findings are a good showcase of the relative benefits from applying the proposed method on a survey dataset in the presence of potentially non-ignorable missing covariates. One of the possible reasons for the good behavior of the proposed method is related to its implementation of a maximum likelihood approach. Another is the use of an artificially complete dataset obtained through augmentation, which facilitates a convenient application of the EM algorithm. In addition, a correct specification of the joint missing data likelihood during the simulation may also eventually be instrumental to the proposed method achieving good results.

It is open for argument that after all, the proposed method in these simulation studies does not really show its edge against the established MICE. But this is a classic debate of choosing between maximum likelihood versus multiple imputation. Both are acknowledged to have attractive statistical properties (30). As I put forth in the literature review, however, MICE is in general vulnerable to models incompatibility and thus, it lacks the theory to support its use. Like any multiple imputation method, it also entails far more analysis decisions, such as the choice of iterative algorithm, imputation model for each incomplete variable, number of data replications, total iterations, and prior distribution. The proposed method, on the

other hand, is built on a well-known theoretical basis and fairly straightforward to implement. Moreover, given the same set of data, the proposed method will produce an identical result; this will not necessarily be the case for MICE.

An area where the proposed method seems to struggle with is the estimation of parameters with the variable consisting of constant values. Under the non-ignorable missingness framework, this relates to the model for the missing data mechanism. In particular, while the missing covariate is augmented with all or most likely values, its missing data indicator remains constant (in the notation of this chapter, $\mathsf{R}_{ik} = 0$). To circumvent this problem, I make a slight modification in the algorithm. There is no change in assumption of the joint distribution for the non-ignorable missing data; however, the weight component containing the missing data mechanism for the corresponding missing covariate is selectively removed from the conditional distribution of missing covariate given the observed data (that is, the missing data weight, see Equation 4.5 for example) during the iteration of the EM algorithm. This approach, which I named the augmentation-CR method, appears to address the issue well. The estimates based on this modification for Case 1 and Case 2 are quite promising, sitting between the augmentation method itself and MICE with regard to performance. Most importantly, it achieves convergence substantially faster than the augmentation method, where the gain is relatively stable across the cases it was tested. It thus is a good option when computation time is of great concern. Unfortunately, the augmentation-CR method lacks the ability of obtaining a valid variance estimate via Louis method (86) in the simulation of Case 3. Perhaps an alternative

procedure using, for instance, bootstrap or other techniques, may help solving such variance estimation problem in the future.

Several aspects of the presented studies may limit the generalizability of the results. To begin with, a simulation is what the word suggests. It is artificial and may not necessarily represent the real world scenario. I partly address this concern in the next chapter, by extending the application of the proposed method and its competing approaches into a real survey datasets and re-contrasting the performance. Another limitation is that, despite the number of replications, these simulation studies can not cover all possible scenarios for each case and situation of survey data with non-ignorable missing covariates. What I try to accomplish is establishing the relative benefits of the proposed method for some sensible and likely scenarios of survey data in the real world settings, particularly within the realm of health care, public health, and medicine. This explains the selection of variables and their distribution in the simulation studies. Finally, the findings are obtained under the constraint that the joint distribution of the missing data is correctly specified. This is probably an unrealistic assumption, but not impossible in the analysis of survey data.

CHAPTER 5

REAL DATA APPLICATION: HOUSEHOLD DETERMINANTS OF INFANT MORTALITY IN INDONESIA

I demonstrate in this chapter the application of the proposed augmentation method for handling missing data on an actual survey dataset. Data of the Indonesia Demographic and Health Survey (IDHS) of 2012, as mentioned in Chapter 1, motivate the proposed method and thus I use them for this illustration. The analysis also has a secondary objective of identifying the household-level determinants of infant mortality in Indonesia.

Infant mortality refers to the death of a child before one year of age (or, less than exact 12 months) (4). It can be further specified into perinatal (death of fetus in 22 weeks gestation to birth, or newborn during the first seven days postpartum), neonatal (death within 28 days after birth), or post-neonatal (death between 29 days and exact age 1) mortality, depending on the study objective and data quality. The occurrence of infant mortality is well documented in children with infections (such as acute respiratory infection and diarrhea) (128), premature birth and low birth weight (128; 129), complications during delivery (128), and tight birth spacing (130; 131). Household socioeconomic status has been also suggested to be a determinant (132; 133; 134; 135). Infant mortality is an important indicator for monitoring public health programs and policies (4).

At the Millennium Summit of 2000, 191 members of the United Nations (UN) and at least 22 international organizations committed to achieve the Millennium Development Goals (MDGs)

by 2015. Goal 4, which was the reduction of child mortality, had fostered the global effort to achieve the set objective of reducing child mortality rates, including infant mortality, by two-thirds between 1990 and 2015. As of 2016, the UN has replaced MDGs with the Sustainable Development Goals (SDGs). Goal 3 of SDGs is about ensuring healthy lives and promoting well-being for all at all ages, where one of the targets is to end preventable deaths of newborns and children under 5 years of age by 2030.

The Republic of Indonesia is an archipelago country with approximately 17,000 islands. It lies between Asia and Australia, bounded in the north by the South China Sea and the Pacific Ocean, in the east by the Pacific Ocean, and in the south and west by the Indian Ocean. Over 80 percent of the country's territory is water, leaving a total land area of about 1.9 million square kilometers. Indonesia has a tropical climate with two seasons throughout the year: the dry season in May to October, and the rainy season in November to April. Administratively, this country is divided into provinces. Each province has districts and municipalities. Below them, there are subdistricts, which in turn are also divided into villages. In 2012, Indonesia consisted of 33 provinces, 399 districts and 98 municipalities, 6,793 subdistricts, and 79,075 villages. The population size was 237.6 million in the 2010 Census, making Indonesia the fourth most populous country after the Peoples Republic of China, India, and the United States of America. An estimated 118.3 million (50 %) of the population lived in urban areas. The average annual growth rate of population between 2000 and 2010 was 1.44 %. Nationally in 2010, the population density was 124 persons per square kilometer.

Over the years, there have been seven national surveys conducted in Indonesia under the auspices of the Demographic and Health Surveys program. These include the National Indonesia Contraceptive Prevalence Survey of 1987, and the Indonesia Demographic and Health Survey (IDHS) of 1991, 1994, 1997, 2002-03, 2007, and 2012. Originally, the surveys only covered ever-married women aged 15-49. Since 2002-2003, however, IDHS started to sample currently married men aged 15-54, and never-married women and men aged 15-24. Then in the IDHS of 2012, the survey included all women aged 15-49 regardless of the marital status. The childhood mortality rates, which include the infant (and all its sub-classifications), the child (12-59 months), and the under-five (from birth to 59 months) mortality rates, have been computed and reported since the IDHS of 1991.

The infant mortality rate in the five-year period preceding the survey as recorded by the IDHSs show a gradual decline from 68 deaths per 1,000 live births in the IDHS of 1991 to 32 deaths per 1,000 live births in the IDHS of 2012. While this reduction seems descent, the rate barely moves in the last three IDHSs: 35, 34, and 32 deaths per 1,000 live births, respectively, in the IDHS of 2002-2003, 2007, and 2012. The target of MDGs Goal 4, in particular, was 23 deaths per 1,000 live births by 2015. A stagnant reduction in the recent years is acknowledged in the 2012 IDHS report (4). It is suggested that the knowledge about the factors associated with infant mortality has to be updated such that a further mortality reduction can be satisfactorily achieved.

The choice of household level in this analysis is almost heuristic. However, there are several reasons that may justify its use instead of the individual child level. First, the IDHS sampled

households; all births within the selected household are simply included. Accordingly, it is safer to assume independence at the household level than at the individual child level. Second, the information about wealth index, access to water sources and sanitation facilities, and residence, to name a few, which have been suggested as the determinants of childhood mortality in the previous studies (134; 135), are household measurements in IDHS, and thus, it is more natural to do the analysis at the household level. Third, a number of individual-level variables that may be used as proxies for the leading causes of infant mortality, including any presence of upper respiratory tract infection and diarrhea, and vaccination history (135; 128), are only collected among living children in the IDHS. Aggregation of these events into the households will certainly allow a convenient use of such information. And perhaps the most compelling reason, which is related to the second and third, is the structure of the IDHS of 2012 data. Information about each child was recorded in varying details in the original datasets. Survival status and age at death (if applicable), for instance, were collected for all live births, but the data related to pregnancy and delivery of a child was only available for children born in the last five years. Meanwhile, certain pieces of information such as the presence of any infection and vaccination history were by design limited to children born in the last five years who were still alive at the time of survey. Thus, it is quite complicated to take advantage of all relevant information about infant mortality in the IDHS of 2012 data if the analysis is set on the individual child level.

Presentation of this chapter follows the following organization. Section 2 describes the overview of the methodology used for analysis, which includes the variables, assumption of the outcome, the covariates, and the missing data models, and the competing techniques for

analyzing the missing data. Section 3 presents the results. And finally, Section 4 discusses the results with respect to the performance of the proposed methods, and the determinants of infant mortality in Indonesia.

5.1 Methods

There were 43,852 households sampled in the IDHS of 2012. Of them, 29,648 households had a record of at least one live birth. The analysis, however, was restricted on the 22,809 households of which the child was born in the last 10 years prior to the survey. I opted for this study population following the IDHS recommendation. In their report (4), the investigators of the IDHS of 2012 recommended using a 10-year period preceding the survey for analysis involving stratification by covariates, to ensure the stability of infant mortality estimates; without any stratification, they suggested it fine to use a shorter, 5-year period preceding the survey.

The outcome of interest was the number of children died before their first birthday within the household. Thus, a count variable. For covariates, a list of potential variables were considered based on their relevance and the findings of the previous studies (134; 135; 128). Appendix A shows these variables together with their description that include, if applicable, the approach to aggregate them into the household level. Selection of covariates for the outcome was then conducted in a stepwise forward fashion. While significance of the variables and both AIC and BIC values were important factors, the ultimate goal of the covariates selection was obtaining the most parsimonious model that still allowed illustration of the proposed method (that is, some of the covariates, preferably consisted of both continuous and categorical variables, were subject to missing observations), and was able to represent three domains of interest, which

were household characteristics, health-related factors, and birth history. Table XVII shows the selected covariates. Two of them, preceding birth interval and baby weight at birth, were missing in 27.8% and 43.8%, respectively, of the households. It is quite reasonable to assume that their missingness was non-ignorable. For instance, the households might fail to keep a record and thus to report the birth weight when they thought it was usual or normal weight. This is very likely in home deliveries or births by a non-skilled attendant, which are still prevalent in Indonesia. On the other hand, it would not be surprising if the households held the information about the unusual or abnormal birth weight due to fear of humiliation. In terms of preceding birth intervals, the value of 0 was missing by design for the household having single births and twins. Also, a few households might have a problem to recall it because the births were too far apart.

It is obvious that the missing data in this analysis fits the description of Case 1 in Chapter 3. That is, information about the outcome and covariates is available only for the samples, while the values of some covariates among these samples are also subject to presumably non-ignorable missingness. The likelihood-based inference of the outcome and covariates relationship, therefore, needs to follow Equation 3.6.

TABLE XVII
LIST OF COVARIATES IN THE OUTCOME MODEL

Variable	Short Description	Type	Distribution	n_{obs}	n_{mis} (%)
Residence	Residence: urban, rural	Categorical	Binomial	22,809	0 (0.0)
Wealth index	Category of wealth index: poorest, poorer, middle, richer, richest	Categorical	Multinomial	22,809	0 (0.0)
Median children	Median of total children ever born to each woman in the household	Categorical	Poisson	22,809	0 (0.0)
Children $\leq$ 5 yr	Total children aged 5 or less in the household	Categorical	Poisson	22,809	0 (0.0)
Diarrhea/URTI[a]	Any child with diarrhea or URTI in past 2 weeks: yes, no	Categorical	Binomial	22,809	0 (0.0)
Vaccination	Any child with any vaccination: yes, no	Categorical	Binomial	22,809	0 (0.0)
Preceding birth	Median preceding birth interval of children in the household (in log-scale)	Continuous	Normal	16,478	6,331 (27.8)
Baby weight	Mean weight at birth of children in the household, categorized as: normal, abnormal (small/large)	Categorical	Binomial	12,809	10,000 (43.8)

[a] URTI, upper respiratory tract infection.

I considered a Poisson GLM for modeling the outcome y_i = the number of infant deaths in household i given the covariates $\mathbf{X}_i$ and the unknown quantities $\boldsymbol{\beta} = (\beta_0, ..., \beta_8)'$. $\mathbf{X}$ included all the variables in Table XVII. Hence,

$$\begin{aligned}\log\big(\mathrm{E}[y_i|\boldsymbol{x}_i;\boldsymbol{\beta}]\big) =& \beta_0 + \beta_1 \text{residence}_i + \boldsymbol{\beta}_2 \text{wealth index}_i + \beta_3 \text{total children}_i + \\ & \beta_4 \text{total under5}_i + \beta_5 \text{diarrhea URTI}_i + \beta_6 \text{vaccination}_i + \\ & \beta_7 \text{preceding birth}_i + \beta_8 \text{baby weight}_i\end{aligned} \tag{5.1}$$

where $\text{residence}_i \in \{\text{urban, rural}\}$, $\text{baby weight}_i = \in \{\text{abnormal, normal}\}$, and $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23}, \beta_{24})'$ were the parameters of the dummy variables created for each category of wealth index minus the reference (poorest), which consisted of poorer, middle, richer, and richest categories.

A sequence of one-dimensional conditional distributions was then constructed for the joint distribution of the missing covariates. In particular, I conditioned them on the other covariates that were fully observed among the samples. Thus,

$$\begin{aligned}&\mathrm{f}(\boldsymbol{x}_{\text{mis},i} \mid \boldsymbol{x}_{\text{obs},i}; \boldsymbol{\alpha}) = \\ &\quad \mathrm{f}(\text{baby weight}_i \mid \text{preceding birth}_i, \text{residence}_i, \text{wealth index}_i, \text{total children}_i, \\ &\quad \text{total under5}_i, \text{diarrhea URTI}_i, \text{vaccination}_i; \boldsymbol{\alpha}_2) \ \times \\ &\quad \mathrm{f}(\text{preceding birth}_i \mid \text{residence}_i, \text{wealth index}_i, \text{total children}_i, \\ &\quad \text{total under5}_i, \text{diarrhea URTI}_i, \text{vaccination}_i; \boldsymbol{\alpha}_1).\end{aligned} \tag{5.2}$$

The distributional assumption for baby weight and preceding birth (in log-scale) as stated in Table XVII was used to specify the conditional distributions, each with its canonical link function in GLM (logit for binomial, identity for Gaussian).

Let $\mathbf{R} = (\mathbf{R}_1, \mathbf{R}_2)$ be an indicator matrix of the two missing covariates. $R_{ik} = 1$ represents the situation when X_{ik} are observed for household i while $R_{ik} = 0$ otherwise, and $k = 1, 2$ respectively indexes preceding birth and baby weight. Following Equation 3.5, I structured the joint distribution of the missing data mechanism also as a sequence of one-dimensional conditional distributions and depending on both the outcome and covariates. Hence,

$$
\begin{aligned}
&f(\mathbf{r}_i \mid y_i, \mathbf{x}_i; \boldsymbol{\gamma}) = \\
&\quad f(r_{i2} \mid r_{i1}, y_i, \text{baby weight}_i, \text{preceding birth}_i, \text{residence}_i, \text{wealth index}_i, \text{total children}_i, \\
&\quad \text{total under5}_i, \text{diarrhea URTI}_i, \text{vaccination}_i; \boldsymbol{\gamma}_2) \ \times \\
&\quad f(r_{i1} \mid y_i, \text{baby weight}_i, \text{preceding birth}_i, \text{residence}_i, \text{wealth index}_i, \text{total children}_i, \\
&\quad \text{total under5}_i, \text{diarrhea URTI}_i, \text{vaccination}_i; \boldsymbol{\gamma}_1).
\end{aligned} \tag{5.3}
$$

There were two approaches taken with regard to the sample selection model $p_i \equiv \Pr\{i \in \mathscr{S}\}$, where $\mathscr{S}$ denoting the sample set. First, assuming the information about sampling mechanism in the IDHS of 2012 report was complete, I set

$$p_i = f(I_i = 1 | \text{residence}_i, \text{province}_i; \boldsymbol{\delta}),$$

where $I_i = 1$ indicating the household $i \in \mathscr{S}$. Second, I supposed that I decided instead to rely on the weight variable w_i (sampling weighted approach), which was available for all $i \in \mathscr{S}$ in the IDHS, based on the argument that only the original investigators knew the true functional form of p_i. The first approach operated under Situation 1 in the model development of Chapter 3, while the second was analog to Situation 2 of the same chapter. In both approaches, as demonstrated by Equation 3.7 and Equation 3.8, estimation of the parameters of interest $\boldsymbol{\beta}$ may proceed without the need of modeling p_i. However, because the IDHS report clearly mentioned urban/rural classification of the household's residence as part of the sampling stratification (and thus, it should be part of the IDHS weight model), the variable residence in Table XVII was removed from the covariates in the second approach.

The likelihood-based estimation of the parameters was conducted using the data augmentation algorithm outlined in Section 3.6 of Chapter 3. Its computational modification, the augmentation-CR method, where the component with constant values is removed from the joint distribution during estimation, was also implemented for comparison. The convergence criterion for both augmentation algorithms was set as $\left\|\boldsymbol{\theta}_{XY}^{(t+1)} - \boldsymbol{\theta}_{XY}^{(t)}\right\|^2 < 10^{-3}$, where $\boldsymbol{\theta}'_{XY} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$, and t was the iteration index. A stricter convergence criterion tended to increase the computation time exponentially but subtly improved the estimates. Separate estimates were obtained for the two assumptions of sample selection mechanism as mentioned above. Additionally, I performed a complete-case analysis in both sampling assumptions, and MICE estimation for the models where the sample selection was assumed known. It should be noted that the underlying assumption of the missing data mechanism is MCAR in the complete-case analysis, and MAR

in MICE. I conducted all data analyses in the R statistical software. For MICE, I used the default settings of the `mice` library in R. This package, however, has yet provided a practical way of using survey weight.

I also run sensitivity analyses in the likelihood-based inference to evaluate my specification of the missing covariate distribution and the missing data mechanism against several alternatives. This is important since there was no guarantee that the parametric forms I used were correct, and they would not be testable using the data. To check the sensitivity of the joint modeling scheme to the specification of the missing covariates distribution, I fixed the missing data mechanism as it is shown in Equation 5.3 and varied the missing covariates model as follows:

$$
\begin{aligned}
XM_1: \quad & f(\text{baby weight}_i \mid \text{residence}_i, \text{wealth index}_i, \text{total children}_i, \\
& \text{total under5}_i, \text{diarrhea URTI}_i, \text{vaccination}_i; \boldsymbol{\alpha}_2^*) \quad \times \\
& f(\text{preceding birth}_i \mid \text{residence}_i, \text{wealth index}_i, \text{total children}_i, \\
& \text{total under5}_i, \text{diarrhea URTI}_i, \text{vaccination}_i; \boldsymbol{\alpha}_1^*) \\
XM_2: \quad & f(\text{baby weight}_i \mid \text{preceding birth}_i, \text{residence}_i, \text{wealth index}_i, \\
& \text{total under5}_i, \text{vaccination}_i; \boldsymbol{\alpha}_2^*) \quad \times \\
& f(\text{preceding birth}_i \mid \text{residence}_i, \text{wealth index}_i, \\
& \text{total under5}_i, \text{vaccination}_i; \boldsymbol{\alpha}_1^*) \\
XM_3: \quad & f(\text{baby weight}_i \mid \text{preceding birth}_i, \text{wealth index}_i; \boldsymbol{\alpha}_2^*) \quad \times \\
& f(\text{preceding birth}_i \mid \text{wealth index}_i; \boldsymbol{\alpha}_1^*).
\end{aligned}
$$

On the other hand, the sensitivity with regard to the specification of the missing data mechanism was evaluated by fixing the covariates distribution in the form of Equation 5.2, while the missing data mechanism was re-parametrized into one of the following

$$
\begin{aligned}
RM_1: \quad & f(r_{i2} \mid r_{i1}, y_i, \text{residence}_i, \text{wealth index}_i, \text{total children}_i, \\
& \text{total under5}_i, \text{diarrhea URTI}_i, \text{vaccination}_i; \boldsymbol{\gamma}_2^*) \quad \times \\
& f(r_{i1} \mid y_i, \text{residence}_i, \text{wealth index}_i, \text{total children}_i, \\
& \text{total under5}_i, \text{diarrhea URTI}_i, \text{vaccination}_i; \boldsymbol{\gamma}_1^*) \\
RM_2: \quad & f(r_{i2} \mid y_i, \text{baby weight}_i, \text{preceding birth}_i, \text{residence}_i, \text{wealth index}_i, \\
& \text{total children}_i, \text{total under5}_i, \text{diarrhea URTI}_i, \text{vaccination}_i; \boldsymbol{\gamma}_2^*) \quad \times \\
& f(r_{i1} \mid y_i, \text{baby weight}_i, \text{preceding birth}_i, \text{residence}_i, \text{wealth index}_i, \\
& \text{total children}_i, \text{total under5}_i, \text{diarrhea URTI}_i, \text{vaccination}_i; \boldsymbol{\gamma}_1^*) \\
RM_3: \quad & f(r_{i1} \mid r_{i2}, y_i, \text{baby weight}_i, \text{preceding birth}_i, \text{residence}_i, \text{wealth index}_i, \\
& \text{total children}_i, \text{total under5}_i, \text{diarrhea URTI}_i, \text{vaccination}_i; \boldsymbol{\gamma}_2^*) \quad \times \\
& f(r_{i2} \mid y_i, \text{baby weight}_i, \text{preceding birth}_i, \text{residence}_i, \text{wealth index}_i, \\
& \text{total children}_i, \text{total under5}_i, \text{diarrhea URTI}_i, \text{vaccination}_i; \boldsymbol{\gamma}_1^*) \\
RM_4: \quad & f(r_{i2} \mid r_{i1}, y_i, \text{baby weight}_i, \text{preceding birth}_i; \boldsymbol{\gamma}_2^*) \quad \times \\
& f(r_{i1} \mid y_i, \text{baby weight}_i, \text{preceding birth}_i; \boldsymbol{\gamma}_1^*) \\
RM_5: \quad & f(r_{i2} \mid r_{i1}, y_i, \text{baby weight}_i; \boldsymbol{\gamma}_2^*) \quad \times \quad f(r_{i1} \mid y_i, \text{preceding birth}_i; \boldsymbol{\gamma}_1^*).
\end{aligned}
$$

Throughout the sensitivity analyses, the outcome model remained as that in Equation 5.1. Note again that for the second approach of the sample selection mechanism (that is, the sampling weighted modeling), the variable residence was removed from the joint distribution.

5.2 Results

Table XVIII shows the Poisson regression estimates of the log relative prevalence of infant mortality using the complete case, MICE, the augmentation, and the augmentation-CR methods. In terms of computation, the augmentation-CR method needed much less iterations to converge than the augmentation method: 15 versus 81. There were quite notable differences between the complete case estimates and those of the other methods. And though they generally agreed to each other with regard to the variables significance, the estimates from the augmentation method were still distinct from the augmentation-CR and MICE estimates, suggesting that the non-ignorable assumption might have its justification.

The variables median children, total kids aged 5 or less, vaccination, preceding birth, and baby weight were highly significant in all analyses. All methods also confirmed the significance of residence. Contrasting with the category poorest as the reference, only the category richest of the variable wealth index was significant in all analyses. The complete case analysis found no significance on the other categories, but they were at least marginally significant in the other method; MICE even determined that the contrast richer versus poorest was significant. The augmentation method indicated a highly significant effect of the presence of diarrhea or upper respiratory tract infection (URTI) on infant mortality, which was confirmed by MICE and the augmentation-CR method, but was failed to detect by the complete case analysis. For every

additional child aged 5 or less in the household, the augmentation method estimated that there was on average a 0.262 (SE 0.039, $\mathrm{p} < 0.001$) point reduction in the log relative prevalence of infant mortality; this estimate was close to that of the augmentation-CR method (-0.258, SE 0.039, $\mathrm{p} < 0.001$) and MICE (-0.247, SE 0.040, $\mathrm{p} < 0.001$), but was appreciably different from the complete case estimate (-0.932, SE 0.063, $\mathrm{p} < 0.001$). For the two variables with missing observations, the augmentation method also produced estimates that were different from the complete case analysis: their absolute effect difference was about 0.185 and 0.495, respectively, for preceding birth and baby weight.

Table XIX shows the survey weighted Poisson regression estimates using the complete case and the two augmentation methods. As it was before, the augmentation-CR method required fewer iterations than the augmentation method to converge, that is, 23 versus 67. Overall, the conclusion about which variable was significant in the analyses assuming the sample selection mechanism known did not change almost at all. The only difference was now all methods agreed that none of the contrasts for the variable wealth index, except for richest versus poorest, was significant. Another thing was the estimates in the survey weighted analyses tended to be higher in magnitude. In a way, the survey weighted modeling can be considered as a sensitivity analysis for a more complicated approach involving a joint distribution of the sampling probabilitiy and the missing data model.

The apparent differences between the results of the complete case analysis and the other methods in both inferences assuming the sample selection mechanism known and not known raised a strong doubt about an MCAR assumption. On the other hand, the fairly distinct

estimates of the augmentation method provided a substantial support for the non-ignorable assumption. What still required caution was the parametric specifications of the augmentation method, particularly of the missing data and covariates models.

Tables XXII-XXV reveal the findings from the sensitivity analyses. For the missing data mechanism, the Poisson regression estimates seemed quite robust against the different specifications (Table XX and Table XXI). Significance of the variables did not change at all across the alternative parameterizations, whether the sample selection was assumed known or not. The estimated direction of effect on infant mortality was identical for all variables in both inferences, and their magnitude of effect was reasonably close in each scenario of the sample selection. In addition, the estimated standard errors barely changed. A similar situation was noted when the missing data mechanism was fixed and the covariates model was varied (Table XXII and Table XXIII). There was no change in conclusion regarding the significance of the variables over the different specifications of the missing covariates models. This brings some confident in interpreting the analysis results using the augmentation method.

5.3 Discussion

Survey samples are commonly selected according to the values of some potential covariates in regression analysis. For instance, the IDHS of 2012 implements a sampling design that involves a stratification by province and urban/rural area, and the analysis demonstrated in this chapter uses the residential urban/rural classification of household as a covariate for the regression model. If sample selection is only related to the covariates values, then the standard regression methods produce valid estimates. However, this assumes a correct specification of the

sampling probability, which in turn requires appropriate knowledge about the survey design. Alternatively, one uses the survey weight. Standard methods may no longer be valid when the survey data also have missing variables. To deal with such circumstances, I propose the augmentation method, which works by augmenting the missing elements with all or certain set of possible values, and weight the augmented data using the conditional probability of the missing given the observed variables. This method is flexible to the missing data mechanisms, including non-ignorable missingness. Illustration in this chapter shows that the augmentation method is able to improve the analysis of infant mortality in Indonesia using the dataset from the IDHS of 2012, where some of the covariates are missing most likely in a non-ignorable fashion.

A clear gain of the augmentation method as compared to the complete case analysis is an increased efficiency. Such an improvement proved to be substantial for the current illustration, as it changes the significance of a variable in the outcome model. The proposed method may have also avoided bias introduced by analyzing only the complete observations. The benchmark method for this demonstration and one of the most popular algorithms for missing data in the literature, MICE, produces estimated effects of the covariates that are closer to the augmentation method than the complete case analysis, which may confirm the accuracy of this proposed method. It is important to note that in simulation studies the augmentation method competes well with MICE. Hence, its application on survey data such as the IDHS of 2012 may overall lead to improvement in both bias and efficiency.

There are caveats in this real data application. First, it assumes that the functional forms of missing data mechanisms and covariates distribution are correctly specified. Sensitivity analyses are performed to partly address this issue by varying the parameters of the missing data mechanisms and the missing covariate distribution. Nevertheless, verification of the models may still require further measures, such as application of different models for the outcome, which is a topic for future research. Second, the probability of missingness is conditioned on the potentially missing covariates, causing the method prone to identifiability problem. This also deserves further studies. Third, the illustrated dataset has a fully observed outcome and a set of auxiliary variables without missing covariates, which may not hold for survey data in general. The proposed method, however, can be readily extend to models without these conditions.

Limitations of the study with respect to subject matter include those inherent to survey data. Retrospective information, such as birth history in the IDHS of 2012, is notorious for its susceptibility to recall errors. Naturally, human has a better memory for recent than distant events. To minimize misreporting of child deaths due to recall errors, investigators of the survey encourage a time period in the recent past where the problem of biased mortality estimates tend to be less serious than that in a more distant past. Another limitation is that the IDHS of 2012 only collected data on women aged 15-49 and were still living at the time of survey. Thus, no information is available for children of those who had died. The report (4), however, claims that the resulting bias in mortality estimates should be negligible since the difference of fertility rates between surviving and non-surviving women in Indonesia is statistically low.

Meanwhile, women aged 50 or older at the time of survey who gave birth during the period under consideration (for the current illustration, ten years preceding the survey) could not report the child's survival status. This provides another reason to limit the time period for mortality analysis, because as the coverage extends further into the past, censoring of survival information becomes more severe. As a rule of thumb, the investigators of the IDHS recommend a period of no longer than 15 years prior to the survey. They also caution not to utilize an interval shorter than five years. The low fertility levels of this country have led to relatively few cases of infant mortality. Accordingly, an estimate based on a very short interval tends to be unstable.

TABLE XVIII
POISSON REGRESSION ESTIMATES OF LOG RELATIVE PREVALENCE OF INFANT MORTALITY AMONG HOUSEHOLDS IN INDONESIA, IDHS OF 2012

	Complete Case			MICE			Augmentation-CR			Augmentation		
Variable	Est	SE	p-value	Est	SE	p-value	Est	SE	p-value	Est	SE	p-value
Intercept	4.488	0.264	<0.001	2.371	0.276	<0.001	2.697	0.217	<0.001	2.759	0.218	<0.001
Rural residence	0.227	0.089	0.011	0.151	0.069	0.029	0.157	0.067	0.019	0.162	0.067	0.016
Wealth Index: Poorest (reference)												
Poorer	-0.101	0.115	0.379	-0.154	0.083	0.064	-0.123	0.079	0.121	-0.134	0.079	0.089
Middle	0.020	0.119	0.870	-0.172	0.091	0.059	-0.155	0.089	0.083	-0.166	0.089	0.061
Richer	0.089	0.126	0.478	-0.207	0.101	0.041	-0.167	0.099	0.091	-0.181	0.098	0.065
Richest	-0.356	0.158	0.024	-0.576	0.122	<0.001	-0.561	0.122	<0.001	-0.576	0.122	<0.001
Median children	0.149	0.024	<0.001	0.174	0.016	<0.001	0.185	0.014	<0.001	0.185	0.014	<0.001
Children $\leq$ 5 yr	-0.932	0.063	<0.001	-0.247	0.040	<0.001	-0.258	0.039	<0.001	-0.262	0.039	<0.001
Diarrhea/URTI	0.004	0.081	0.965	0.384	0.071	<0.001	0.368	0.068	<0.001	0.356	0.068	<0.001
Vaccination	-1.220	0.105	<0.001	-1.006	0.094	<0.001	-0.968	0.094	<0.001	-0.979	0.094	<0.001
Preceding birth	-1.426	0.066	<0.001	-1.180	0.059	<0.001	-1.239	0.051	<0.001	-1.241	0.051	<0.001
Normal baby weight	-0.659	0.108	<0.001	-0.960	0.123	<0.001	-1.136	0.095	<0.001	-1.154	0.096	<0.001

NOTE: CR, constant removed; Est, estimate; SE, standard error; URTI, upper respiratory tract infection.

TABLE XIX
SURVEY-WEIGHTED POISSON REGRESSION ESTIMATES OF LOG RELATIVE PREVALENCE OF INFANT MORTALITY AMONG HOUSEHOLDS IN INDONESIA, IDHS OF 2012

	Complete Case			Augmentation-CR[a]			Augmentation		
Variable	Est	SE	p-value	Est	SE	p-value	Est	SE	p-value
Intercept	5.224	0.408	<0.001	3.042	0.311	<0.001	3.149	0.311	<0.001
Wealth Index: Poorest (reference)									
Poorer	-0.007	0.147	0.960	-0.025	0.109	0.816	-0.040	0.109	0.714
Middle	0.083	0.151	0.582	-0.036	0.130	0.784	-0.051	0.130	0.694
Richer	0.150	0.160	0.350	-0.014	0.135	0.916	-0.035	0.135	0.797
Richest	-0.477	0.218	0.029	-0.474	0.151	0.002	-0.497	0.151	0.001
Median children	0.138	0.036	<0.001	0.209	0.018	<0.001	0.208	0.018	<0.001
Children $\leq$ 5 yr	-1.090	0.109	<0.001	-0.314	0.056	<0.001	-0.323	0.056	<0.001
Diarrhea/URTI	0.142	0.112	0.207	0.544	0.106	<0.001	0.521	0.106	<0.001
Vaccination	-1.184	0.138	<0.001	-0.923	0.132	<0.001	-0.938	0.132	<0.001
Preceding birth	-1.505	0.100	<0.001	-1.312	0.079	<0.001	-1.316	0.079	<0.001
Normal baby weight	-0.860	0.144	<0.001	-1.322	0.077	<0.001	-1.356	0.084	<0.001

[a] CR, constant removed; Est, estimate; SE, standard error; URTI, upper respiratory tract infection.

TABLE XX
AUGMENTATION ESTIMATES (STANDARD ERRORS) FOR LOG RELATIVE PREVALENCE OF INFANT MORTALITY USING SEVERAL ALTERNATIVES OF MISSING DATA MODEL

Variable	RM_{used}	RM_1	RM_2	RM_3	RM_4	RM_5
Intercept	2.759(0.218)***	2.697(0.216)***	2.798(0.218)***	2.724(0.217)***	2.860(0.213)***	2.851(0.213)***
Rural residence	0.162(0.067)*	0.157(0.067)*	0.163(0.067)*	0.160(0.067)*	0.158(0.067)*	0.159(0.067)*
Wealth Index: Poorest (ref)						
Poorer	-0.134(0.079)	-0.120(0.079)	-0.137(0.079)	-0.129(0.079)	-0.117(0.079)	-0.118(0.079)
Middle	-0.166(0.089)	-0.152(0.089)	-0.170(0.089)	-0.161(0.089)	-0.148(0.089)	-0.151(0.089)
Richer	-0.181(0.098)	-0.162(0.099)	-0.185(0.098)	-0.174(0.098)	-0.159(0.098)	-0.162(0.098)
Richest	-0.576(0.122)***	-0.557(0.122)***	-0.581(0.122)***	-0.569(0.122)***	-0.565(0.122)***	-0.568(0.122)***
Median children	0.185(0.014)***	0.187(0.014)***	0.183(0.014)***	0.186(0.014)***	0.163(0.014)***	0.161(0.014)***
Children $\leq$ 5 yr	-0.262(0.039)***	-0.256(0.039)***	-0.264(0.039)***	-0.260(0.039)***	-0.260(0.039)***	-0.260(0.039)***
Diarrhea/URTI	0.356(0.068)***	0.369(0.068)***	0.353(0.068)***	0.362(0.068)***	0.405(0.069)***	0.403(0.069)***
Vaccination	-0.979(0.094)***	-0.962(0.094)***	-0.983(0.094)***	-0.975(0.094)***	-0.931(0.095)***	-0.936(0.095)***
Preceding birth	-1.241(0.051)***	-1.247(0.051)***	-1.245(0.050)***	-1.241(0.051)***	-1.285(0.049)***	-1.284(0.049)***
Normal baby weight	-1.154(0.096)***	-1.120(0.095)***	-1.156(0.097)***	-1.144(0.096)***	-1.038(0.093)***	-1.023(0.094)***

NOTE: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; ref, reference; URTI, upper respiratory tract infection.

TABLE XXI
AUGMENTATION ESTIMATES (STANDARD ERRORS) FOR SURVEY-WEIGHTED LOG RELATIVE PREVALENCE OF INFANT MORTALITY USING SEVERAL ALTERNATIVES OF MISSING DATA MODEL

Variable	RM_{used}	RM_1	RM_2	RM_3	RM_4	RM_5
Intercept	3.149(0.311)***	3.059(0.310)***	3.202(0.310)***	3.131(0.311)***	3.193(0.302)***	3.186(0.302)***
Wealth Index: Poorest (ref)						
Poorer	-0.040(0.109)	-0.020(0.109)	-0.044(0.110)	-0.040(0.110)	-0.010(0.109)	-0.011(0.109)
Middle	-0.051(0.130)	-0.031(0.130)	-0.055(0.130)	-0.050(0.130)	-0.021(0.129)	-0.022(0.129)
Richer	-0.035(0.135)	-0.009(0.135)	-0.039(0.135)	-0.035(0.135)	0.007(0.136)	0.004(0.136)
Richest	-0.497(0.151)**	-0.467(0.151)**	-0.504(0.151)***	-0.497(0.151)**	-0.464(0.151)**	-0.466(0.151)**
Median children	0.208(0.018)***	0.210(0.018)***	0.206(0.018)***	0.210(0.018)***	0.186(0.019)***	0.184(0.019)***
Children $\leq$ 5 yr	-0.323(0.056)***	-0.313(0.055)***	-0.325(0.056)***	-0.321(0.056)***	-0.305(0.055)***	-0.303(0.055)***
Diarrhea/URTI	0.521(0.106)***	0.546(0.106)***	0.514(0.106)***	0.520(0.106)***	0.592(0.107)***	0.590(0.107)***
Vaccination	-0.938(0.132)***	-0.915(0.132)***	-0.943(0.131)***	-0.940(0.132)***	-0.885(0.132)***	-0.890(0.132)***
Preceding birth	-1.316(0.079)***	-1.323(0.079)***	-1.321(0.079)***	-1.315(0.080)***	-1.358(0.076)***	-1.358(0.076)***
Normal baby weight	-1.356(0.084)***	-1.314(0.076)***	-1.362(0.086)***	-1.352(0.083)***	-1.248(0.073)***	-1.231(0.073)***

NOTE: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; ref, reference; URTI, upper respiratory tract infection.

TABLE XXII
AUGMENTATION ESTIMATES (STANDARD ERRORS) FOR LOG RELATIVE PREVALENCE OF INFANT MORTALITY USING SEVERAL ALTERNATIVES OF COVARIATE MODEL

Variable	XM_{used}	XM_1	XM_2	XM_3
Intercept	2.851(0.213)***	2.684(0.216)***	2.512(0.217)***	2.477(0.215)***
Rural residence	0.159(0.067)*	0.160(0.067)*	0.156(0.067)*	0.146(0.067)*
Wealth Index: Poorest (ref)				
Poorer	-0.118(0.079)	-0.134(0.079)	-0.133(0.079)	-0.136(0.079)
Middle	-0.151(0.089)	-0.166(0.089)	-0.164(0.089)	-0.163(0.089)
Richer	-0.162(0.098)	-0.180(0.098)	-0.177(0.098)	-0.179(0.098)
Richest	-0.568(0.122)***	-0.574(0.122)***	-0.567(0.122)***	-0.568(0.122)***
Median children	0.161(0.014)***	0.185(0.014)***	0.201(0.013)***	0.205(0.013)***
Children $\leq$ 5 yr	-0.260(0.039)***	-0.258(0.039)***	-0.253(0.039)***	-0.236(0.039)***
Diarrhea/URTI	0.403(0.069)***	0.361(0.068)***	0.340(0.068)***	0.339(0.068)***
Vaccination	-0.936(0.095)***	-0.977(0.094)***	-0.979(0.094)***	-1.039(0.094)***
Preceding birth	-1.284(0.049)***	-1.232(0.051)***	-1.197(0.051)***	-1.192(0.051)***
Normal baby weight	-1.023(0.094)***	-1.120(0.097)***	-1.155(0.096)***	-1.156(0.097)***

NOTE: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; ref, reference; URTI, upper respiratory tract infection.

TABLE XXIII
AUGMENTATION ESTIMATES (STANDARD ERRORS) FOR SURVEY-WEIGHTED LOG RELATIVE PREVALENCE OF INFANT MORTALITY USING SEVERAL ALTERNATIVES OF COVARIATES MODEL

Variable	XM_{used}	XM_1	XM_2	XM_3
Intercept	3.149(0.311)***	3.151(0.311)***	2.844(0.316)***	2.827(0.312)***
Wealth Index: Poorest (ref)				
Poorer	-0.040(0.109)	-0.040(0.109)	-0.039(0.110)	-0.038(0.110)
Middle	-0.051(0.130)	-0.051(0.130)	-0.045(0.131)	-0.041(0.132)
Richer	-0.035(0.135)	-0.034(0.135)	-0.028(0.134)	-0.030(0.134)
Richest	-0.497(0.151)**	-0.497(0.151)**	-0.481(0.151)**	-0.479(0.151)**
Median children	0.208(0.018)***	0.208(0.018)***	0.229(0.017)***	0.232(0.017)***
Children $\leq$ 5 yr	-0.323(0.056)***	-0.323(0.056)***	-0.312(0.056)***	-0.294(0.056)***
Diarrhea/URTI	0.521(0.106)***	0.521(0.106)***	0.504(0.106)***	0.503(0.106)***
Vaccination	-0.938(0.132)***	-0.937(0.132)***	-0.933(0.133)***	-0.991(0.132)***
Preceding birth	-1.316(0.079)***	-1.317(0.079)***	-1.264(0.081)***	-1.262(0.081)***
Normal baby weight	-1.356(0.084)***	-1.357(0.084)***	-1.362(0.082)***	-1.371(0.084)***

NOTE: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; ref, reference; URTI, upper respiratory tract infection.

CHAPTER 6

CONCLUSION AND REMARKS

6.1 Conclusion

I propose here a method to analyze survey data where the primary interest is modeling a count outcome with potentially non-ignorable covariates. The method is particularly developed to address the limitations in standard routines of the major statistical software when the following situations are encountered simultaneously: the model of interest has a mixture of categorical and continuous missing covariates, the analysis needs to incorporate the sampling design under different assumptions about its functional form, and there is a demand for manageable computation time in practical sense. To meet this objective, I modify the EM algorithm by inserting a step to augment the missing elements of data, and then weighting the E-step with the conditional probability of the missing variables given the observed data. The algorithm proceeds as a full likelihood procedure if the sampling probability function is known for all observations, but it becomes a quasi-likelihood approach when instead the quantity of survey weight is the only available information about sample selection. Thus, the proposed method from the perspective of survey analysis may be considered as both model and design based inference. There are three classes of survey data considered during the development, which include those of which none (Case 1), all (Case 2), or some (Case 3) of the covariates are observable outside the samples. Two situations are further defined on each of them, that is,

whether the functional form of sample selection is known (Situation 1) or unknown (Situation 2). Given its construction, the proposed method, termed the augmentation assisted EM algorithm or simply the augmentation method, retains the desirable properties of the maximum likelihood estimates, while flexible enough to handle both continuous and categorical missing covariates, and can adapt the use of survey weight to improve inference.

The simulation studies indicates that the proposed method performs reliably well across all classes of survey data. In terms of unbiasedness, it is competitive with and may occasionally outperform the multiple imputation technique MICE. Efficiency of its estimates are also comparable to MICE. In the real data application using the dataset from the Indonesia Demographic and Health Survey of 2012, the proposed method successfully estimates the demographic, health, and birth-related factors associated with the infant mortality. Most importantly, it is able to improve the results of complete case analyses by both correcting the magnitude of effect size and increasing the power of analysis to detect the variable significance. Sensitivity analyses confirm the stability of the proposed method findings.

Therefore, it can be concluded that the proposed method is a useful option for survey data analysis where the modeling is complicated by the presence of potentially non-ignorable missing covariates. As it has been stated earlier, the method has a particular value when along with the mixed nature of the missing covariates, the investigator needs to account for the sampling process, as it may be related to the covariates with missingness. And although this will also depend on the analyst's programming skill and the capacity of the statistical software, the proposed method is relatively less time consuming, because there is no need for stochastic

sampling in its computation. Such choice in turn differentiates this proposed method from the data augmentation technique by Tanner and Wong (82). I note, in addition, that they implicitly limit their original development to continuous variables. This method also bears resemblance to the missing data approach promoted by Ibrahim et al. (70; 72; 2) and Lipsitz et al. (71; 73). I have acknowledged this in the Methodology chapter. What makes the present algorithm distinct from Ibrahim et al.'s and Lipsitz et al.'s approach is that it considers the incorporation of sampling information from various classes of survey data.

6.2 Remarks

Applied statisticians and survey analysts may find this proposed method very useful when dealing with missingness in a variety of survey data. The method is developed with such investigator background in mind. It is also hoped that the findings can enrich the literature of missing data, computational statistics, and survey analysis with respect to the addressed issues. As the propensity for dual model and design based inference becomes more mainstream in the statistics community, the proposed method may also provide leverage to connect the two schools of thought.

National surveys have become an increasingly crucial source of public health data. In the past, their extensive use might have been hampered by the question about their data quality, the cross-sectional nature of information, a limited computing power, a lack of understanding on how to properly handle survey data, restricted routines, and computational intensity of survey analysis in the standard software. Yet by time all these problems seem to gradually disappear. In fact, with the accumulation of experience among the survey administrators, continuously

tested instruments, steadily improved and standardized procedures for data collection, along with the traditional benefits such as wide geographical coverage and large sample size, national surveys have grown to be the ultimate data source for many health measures. Moreover, these surveys are normally administered by the government agencies. Hence, they are often available for public use at no cost. Their accessibility is further accentuated by the worldwide spread of internet, even in the most remote places on earth. It becomes, therefore, an important task to facilitate the researchers taking the most advantages from large survey data.

Several issues, however, require further research. Identification of the parameters for missing data mechanisms remains a big concern. If the proportion of missing data is really large, it is perhaps not a bad idea to resort to MAR assumption. That will lead to a certain degree of bias, but it has less problem of computation convergence. Otherwise, one may need to use other techniques. Meanwhile, augmentation of a missing count variable using the "most likely" values may not always be convenient to implement. For instance, in a Poisson distributed variable with an expected rate sufficiently far from zero, such values will cover a range that could be too cumbersome for the algorithm to tackle. A more capable fashion of selecting the values is thus necessary to devise. It is also advisable to improve the flexibility of the algorithm for augmenting different types of variables without intensifying computation. A Gibbs sampler or any Monte Carlo based sampling (71; 73; 72; 2) are great options, but their use for missing variables consisted of categorical and continuous mixture is still an active research area. For continuous variables, their application in general demands a more expensive computation than Gaussian quadrature due to the need of assuring independent random draws from the joint posterior

predictive distribution. During the method development, I have actually considered the use of a quasi Gibbs sampler as implemented in MICE (or, the fully conditional specification approach) to facilitate the augmentation of the missing observations. Such quasi-Gibbs sampler procedure is relatively more flexible for various types of variables. But its use will separate the data augmentation and the maximization into two different steps. To produce valid estimates, I need multiple runs of EM algorithm. It is critical to recall that MICE still lacks the theory supporting its application. Finally, the proposed method relies on inverting Louis information matrix to obtain the variance estimates. Unfortunately, the matrix may not always be invertible for some parameters in the joint missing data model. Louis method also requires an additional step in the algorithm for computing the expected of squared first derivative of the Q function given the observed data. In the future, it might be worth considering to resort to some alternative methods such as bootstrapping or other approaches that do not need matrix inversion.

CITED LITERATURE

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

2. Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H.: Missing-Data Methods for Generalized Linear Models. Journal of the American Statistical Association, 100(469):332–346, March 2005.

3. Little, R. J. A.: Regression with Missing X's: A Review. Journal of the American Statistical Association, 87(420):1227–1237, December 1992.

4. Statistics Indonesia, Indonesia National Population and Family Planning Board, Indonesia Ministry of Health, and MEASURE DHS ICF International: Indonesia Demographic and Health Survey 2012. Jakarta, Indonesia, BPS, BKKBN, Kemenkes, and ICF International, August 2013.

5. Little, R. J. and Rubin, D. B.: Statistical analysis with missing data. John Wiley & Sons, 2014.

6. Glynn, R. J. and Laird, N. M.: Regression estimates and missing data: complete case analysis. Cambridge MA: Harvard School of Public Health, Department of Biostatistics, 1986.

7. Chen, Q., Ibrahim, J. G., Chen, M.-H., and Senchaudhuri, P.: Theory and Inference for Regression Models with Missing Responses and Covariates. Journal of multivariate analysis, 99(6):1302–1331, July 2008.

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

9. Rubin, D. B.: Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In Proceedings of the survey research methods section of the American Statistical Association, volume 1, pages 20–34. American Statistical Association, 1978.

10. Rubin, D. B.: Multiple imputation after 18+ years. Journal of the American statistical Association, 91(434):473–489, 1996.

11. Rubin, D. B.: Multiple imputation for nonresponse in surveys, volume 81. John Wiley & Sons, 2004.

12. Rubin, D. B. and Schenker, N.: Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association, 81(394):366–374, 1986.

13. Van Buuren, S. and Oudshoorn, K.: Flexible multivariate imputation by MICE. Leiden, The Netherlands: TNO Prevention Center, 1999.

14. Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P.: A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey methodology, 27(1):85–96, 2001.

15. Nielsen, S. F.: Proper and improper multiple imputation. International Statistical Review, 71(3):593–607, 2003.

16. Buuren, S. and Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in R. Journal of statistical software, 45(3), 2011.

17. Van Buuren, S.: Flexible imputation of missing data. CRC press, 2012.

18. Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B.: Fully conditional specification in multivariate imputation. Journal of statistical computation and simulation, 76(12):1049–1064, 2006.

19. Van Buuren, S.: Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in medical research, 16(3):219–242, 2007.

20. Van Buuren, S., Boshuizen, H. C., Knook, D. L., and others: Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in medicine, 18(6):681–694, 1999.

21. Kennickell, A. B.: Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. In Proceedings of the Survey Research Methods Section of the American Statistical Association, volume 1, 1991.

22. Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., and Kadie, C.: Dependency networks for inference, collaborative filtering, and data visualization. Journal of Machine Learning Research, 1(Oct):49–75, 2000.

23. Rubin, D. B.: Nested multiple imputation of NMES via partially incompatible MCMC. Statistica Neerlandica, 57(1):3–18, 2003.

24. Gelman, A.: Parameterization and Bayesian modeling. Journal of the American Statistical Association, 99(466):537–545, 2004.

25. Buuren, S. v. and Oudshoorn, C. G. M.: Multivariate imputation by chained equations: MICE V1. 0 user's manual. Technical report, TNO, 2000.

26. Schafer, J. L.: Analysis of incomplete multivariate data. CRC press, 1997.

27. Rubin, D. B. and Schafer, J. L.: Efficiently creating multiple imputations for incomplete multivariate normal data. In Proceedings of the Statistical Computing Section of the American Statistical Association, volume 83, page 88, 1990.

28. Chen, H. Y., Xie, H., and Qian, Y.: Multiple imputation for missing values through conditional Semiparametric odds ratio models. Biometrics, 67(3):799–809, September 2011.

29. Allison, P. D.: Missing data. Sage Thousand Oaks, CA, 2012.

30. Allison, P. D.: Handling missing data by maximum likelihood. In SAS global forum, volume 23 of 312, pages 1–21, Orlando, FL, 2012. SAS Institute Inc.

31. Swamy, P. and Mehta, J. S.: On Bayesian estimation of seemingly unrelated regressions when some observations are missing. Journal of Econometrics, 3(2):157–169, 1975.

32. Press, S. J. and Scott, A. J.: Missing variables in Bayesian regression, II. Journal of the American Statistical Association, 71(354):366–369, 1976.

33. Guttman, I. and Menzefricke, U.: Bayesian Inference in Multivariate Regression with Missing Observations on the response variables. Journal of Business & Economic Statistics, 1(3):239–248, 1983.

34. Chen, C. F.: A Bayesian approach to nested missing-data problems. In Bayesian Inference and Decision Techniques: Essays in Honor of B. de Finetti, eds. P. K. Goel and A. Zellner, pages 355–361. New York City, NY, Elsevier, 1986.

35. Little, R. J.: Approximately calibrated small sample inference about means from bivariate normal data with missing values. Computational Statistics & Data Analysis, 7(2):161–178, 1988.

36. Huang, L., Chen, M.-H., and Ibrahim, J. G.: Bayesian analysis for generalized linear models with nonignorably missing covariates. Biometrics, 61(3):767–780, 2005.

37. Mason, A., Best, N., Richardson, S., and Plewis, I.: Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods. Technical report, National Center for Research Methods, London, 2010.

38. Gelfand, A. E. and Smith, A. F.: Sampling-based approaches to calculating marginal densities. Journal of the American statistical association, 85(410):398–409, 1990.

39. Gilks, W. R. and Wild, P.: Adaptive rejection sampling for Gibbs sampling. Applied Statistics, pages 337–348, 1992.

40. Casella, G. and George, E. I.: Explaining the Gibbs sampler. The American Statistician, 46(3):167–174, 1992.

41. Gilks, W. R., Best, N. G., and Tan, K. K. C.: Adaptive rejection Metropolis sampling within Gibbs sampling. Applied Statistics, pages 455–472, 1995.

42. Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H.: Missing-Data Methods for Generalized Linear Models: A Comparative Review. Journal of the American Statistical Association, 100(469):332–346, 2005.

43. Robins, J. M., Rotnitzky, A., and Zhao, L. P.: Estimation of regression coefficients when some regressors are not always observed. Journal of the American statistical Association, 89(427):846–866, 1994.

44. Robins, J. M. and Rotnitzky, A.: Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association, 90(429):122–129, 1995.

45. Robins, J. M., Ritov, Y., and others: Toward A Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-Parametric Models. Statistics in medicine, 16(3):285–319, 1997.

46. Robins, J. M. and Rotnitzky, A.: Comment on Inference for semiparametric models: Some questions and an answer, by PJ Bickel and J. Kwon. Statist. Sinica, 11:920–936, 2001.

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

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

49. Kang, J. D. Y. and Schafer, J. L.: Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. Statistical Science, 22(4):523–539, November 2007.

50. Chen, B. and Zhou, X.-H.: Doubly robust estimates for binary longitudinal data analysis with missing response and missing covariates. Biometrics, 67(3):830–842, 2011.

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

52. Rotnitzky, A. and Robins, J. M.: Semiparametric Regression Estimation in the Presence of Dependent Censoring. Biometrika, 82(4):805–820, 1995.

53. Zhao, L. P., Lipsitz, S., and Lew, D.: Regression analysis with missing covariate data using estimating equations. Biometrics, pages 1165–1182, 1996.

54. Rotnitzky, A., Robins, J. M., and Scharfstein, D. O.: Semiparametric regression for repeated outcomes with nonignorable nonresponse. Journal of the American Statistical Association, 93(444):1321–1339, 1998.

55. Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P.: A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. Journal of the American Statistical Association, 94(448):1147–1160, 1999.

56. Scharfstein, D. O., Rotnitzky, A., and Robins, J. M.: Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association, 94(448):1096–1120, 1999.

57. Robins, J. M., Rotnitzky, A., and Scharfstein, D. O.: Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Statistical models in epidemiology, the environment, and clinical trials, pages 1–94. Springer, 2000.

58. Parzen, M., Lipsitz, S. R., Ibrahim, J. G., and Lipshultz, S.: A weighted estimating equation for linear regression with missing covariate data. Statistics in Medicine, 21(16):2421–2436, August 2002.

59. Scharfstein, D. O. and Irizarry, R. A.: Generalized additive selection models for the analysis of studies with potentially nonignorable missing outcome data. Biometrics, 59(3):601–613, 2003.

60. Herring, A. H., Ibrahim, J. G., and Lipsitz, S. R.: Non-ignorable missing covariate data in survival analysis: a case-study of an International Breast Cancer Study Group trial. Journal of the Royal Statistical Society: Series C (Applied Statistics), 53(2):293–310, 2004.

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

62. McCullagh, P. and Nelder, J. A.: Generalized linear models, volume 37. CRC press, 1989.

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

64. Wilks, S. S.: Moments and distributions of estimates of population parameters from fragmentary samples. The Annals of Mathematical Statistics, 3(3):163–195, 1932.

65. Anderson, T. W.: Maximum Likelihood Estimates for a Multivariate Normal Distribution when Some Observations are Missing. Journal of the American Statistical Association, 52(278):200–203, June 1957.

66. Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), pages 1–38, 1977.

67. Gourieroux, C. and Monfort, A.: On the problem of missing data in linear models. The Review of Economic Studies, 48(4):579–586, 1981.

68. Little, R. J. A. and Schluchter, M. D.: Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values. Biometrika, 72(3):497–512, 1985.

69. Schluchter, M. D. and Jackson, K. L.: Log-linear analysis of censored survival data with partially observed covariates. Journal of the American Statistical Association, 84(405):42–52, 1989.

70. Ibrahim, J. G.: Incomplete Data in Generalized Linear Models. Journal of the American Statistical Association, 85(411):765–769, 1990.

71. Lipsitz, S. R. and Ibrahim, J. G.: A conditional model for incomplete covariates in parametric regression models. Biometrika, 83(4):916–922, 1996.

72. Ibrahim, J. G., Lipsitz, S. R., and Chen, M.-H.: Missing Covariates in Generalized Linear Models When the Missing Data Mechanism Is Non-Ignorable. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 61(1):173–190, 1999.

73. Lipsitz, S. R., Ibrahim, J. G., Chen, M. H., and Peterson, H.: Non-ignorable missing covariates in generalized linear models. Statistics in Medicine, 18(17-18):2435–2448, September 1999.

74. Ibrahim, J. G., Lipsitz, S. R., and Horton, N.: Using Auxiliary Data for Parameter Estimation with Non-Ignorably Missing Outcomes. Journal of the Royal Statistical Society. Series C (Applied Statistics), 50(3):361–373, 2001.

75. Chen, T. and Fienberg, S. E.: The analysis of contingency tables with incompletely classified data. Biometrics, pages 133–144, 1976.

76. Fuchs, C.: Maximum likelihood estimation and model selection in contingency tables with missing data. Journal of the American Statistical Association, 77(378):270–278, 1982.

77. Vach, W.: Logistic regression with missing values in the covariates, volume 86. Springer Science & Business Media, 2012.

78. Pepe, M. S. and Fleming, T. R.: A Nonparametric Method for Dealing with Mismeasured Covariate Data. Journal of the American Statistical Association, 86(413):108–113, March 1991.

79. Reilly, M. and Pepe, M. S.: A mean score method for missing and auxiliary covariate data in regression models. Biometrika, 82(2):299–314, 1995.

80. Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J.: Semiparametric methods for response-selective and missing data problems in regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(2):413–438, 1999.

81. Tang, G., Little, R. J., and Raghunathan, T. E.: Analysis of multivariate missing data with nonignorable nonresponse. Biometrika, 90(4):747–764, 2003.

82. Tanner, M. A. and Wong, W. H.: The calculation of posterior distributions by data augmentation. Journal of the American statistical Association, 82(398):528–540, 1987.

83. Tanner, M. A. and Wong, W. H.: From EM to data augmentation: the emergence of MCMC Bayesian computation in the 1980s. Statistical science, 25(4):506–516, 2010.

84. Geman, S. and Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence, 20(6):721–741, 1984.

85. Wei, G. C. and Tanner, M. A.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American statistical Association, 85(411):699–704, 1990.

86. Louis, T. A.: Finding the Observed Information Matrix when Using the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 44(2):226–233, 1982.

87. Bartlett, J. W., Seaman, S. R., White, I. R., and Carpenter, J. R.: Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. Statistical Methods in Medical Research, 24(4):462–487, 2015.

88. Little, R. J. A.: Survey inference with weights for differential sample selection or nonresponse. Proceedings of the Survey Research Methods Section, American Statistical Association 1989, pages 62–71, 1989.

89. Pfeffermann, D.: The role of sampling weights when modeling survey data. International Statistical Review/Revue Internationale de Statistique, pages 317–337, 1993.

90. Kish, L. and Frankel, M. R.: Inference from complex samples. Journal of the Royal Statistical Society. Series B (Methodological), pages 1–37, 1974.

91. Jnrup, H. and Rennermalm, B.: Regression analysis in samples from finite populations. Scandinavian Journal of Statistics, pages 33–36, 1976.

92. Shah, B. V., Holt, M. M., and Folsom, R. E.: Inference about regression models from sample survey data. Bulletin of the International Statistical Institute, 47(3):43–57, 1977.

93. Skinner, C. J., Holt, D., and Smith, T. F.: Analysis of complex surveys. John Wiley & Sons, 1989.

94. Fuller, W. A.: Sampling statistics, volume 560. John Wiley & Sons, 2011.

95. Godambe, V. P.: A new approach to sampling from finite populations. I Sufficiency and linear estimation. Journal of the Royal Statistical Society. Series B (Methodological), pages 310–319, 1966.

96. Basu, D.: An essay on the logical foundations of survey sampling, Part-I: In Foundations of statistical Inference, Ed. By Godambe VP and Spott DA. Toronto: Holt, Rinehart & Winston, 1971.

97. Casella, G. and Berger, R. L.: Statistical inference, volume 2. Duxbury Pacific Grove, CA, 2002.

98. Lehmann, E. L.: Elements of large-sample theory. Springer Science & Business Media, 1999.

99. Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M. P.: Panel surveys. Wiley New York, 1989.

100. Binder, D. A.: On the variances of asymptotically normal estimators from complex surveys. International Statistical Review/Revue Internationale de Statistique, pages 279–292, 1983.

101. Godambe, V. P. and Thompson, M. E.: Parameters of superpopulation and survey population: their relationships and estimation. International Statistical Review/Revue Internationale de Statistique, pages 127–138, 1986.

102. Krieger, A. M. and Pfeffermann, D.: Maximum likelihood estimation from complex sample surveys. Survey Methodology, 18(2):225–239, 1992.

103. Breckling, J. U., Chambers, R. L., Dorfman, A. H., Tam, S. M., and Welsh, A. H.: Maximum Likelihood Inference from Sample Survey Data. International Statistical Review / Revue Internationale de Statistique, 62(3):349–363, 1994.

104. Little, R. J.: Models for nonresponse in sample surveys. Journal of the American statistical Association, 77(378):237–250, 1982.

105. Sugden, R. A. and Smith, T. M. F.: Ignorable and informative designs in survey sampling inference. Biometrika, 71(3):495–506, 1984.

106. Kish, L.: Weighting: Why, when, and how. In Proceedings of the survey research methods section, pages 121–130, 1990.

107. Chambers, R. L., Dorfman, A. H., and Wang, S.: Limited information likelihood analysis of survey data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(2):397–411, January 1998.

108. Alexander, C. H.: A model-based justification for survey weights. In Proceedings of the Section on Survey Research Methods, pages 183–188. American Statistical Association, 1987.

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

110. Nathan, G. and Holt, D.: The effect of survey design on regression analysis. Journal of the Royal Statistical Society. Series B (Methodological), pages 377–386, 1980.

111. DuMouchel, W. H. and Duncan, G. J.: Using sample survey weights in multiple regression analyses of stratified samples. Journal of the American Statistical Association, 78(383):535–543, 1983.

112. Pfeffermann, D. and Holmes, D. J.: Robustness considerations in the choice of a method of inference for regression analysis of survey data. Journal of the Royal Statistical Society. Series A (General), pages 268–278, 1985.

113. Rao, J. N. K., Kovar, J. G., and Mantel, H. J.: On estimating distribution functions and quantiles from survey data using auxiliary information. Biometrika, 77(2):365–375, 1990.

114. Rubin, D. B.: The use of propensity scores in applied Bayesian inference. In Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting : September 6/10, 1983, eds. J. M. Bernardo, M. H. Degroot, D. V. Lindley, and Smith, pages 463–472. Elsevier Science Ltd, September 1985.

115. Patil, G. P. and Rao, C. R.: Weighted Distributions and Size-Biased Sampling with Applications to Wildlife Populations and Human Families. Biometrics, 34(2):179–189, 1978.

116. Rao, C. R.: Weighted Distributions Arising Out of Methods of Ascertainment: What Population Does a Sample Represent? In A Celebration of Statistics, eds. A. C. Atkinson and S. E. Fienberg, pages 543–569. Springer New York, 1985. DOI: 10.1007/978-1-4613-8560-8_24.

117. Chambless, L. E. and Boyle, K. E.: Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. Communications in Statistics-Theory and Methods, 14(6):1377–1392, 1985.

118. Fuller, W. A.: Least squares and related analyses for complex survey designs. Survey Methodology, 10(97):112, 1984.

119. Skinner, C. and Mason, B.: Weighting in the regression analysis of survey data with a cross-national application. Canadian Journal of Statistics, 40(4):697–711, 2012.

120. Nelder, J. A. and Wedderburn, R. W. M.: Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135(3):370–384, 1972.

121. Wedderburn, R. W. M.: Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. Biometrika, 61(3):439–447, 1974.

122. Cameron, A. C. and Trivedi, P. K.: Regression analysis of count data, volume 53. Cambridge university press, 2013.

123. Dean, C., Lawless, J. F., and Willmot, G. E.: A mixed poissoninverse-gaussian regression model. Canadian Journal of Statistics, 17(2):171–181, 1989.

124. Lambert, D.: Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics, 34(1):1–14, 1992.

125. Ibrahim, J. G. and Weisberg, S.: Incomplete data in generalized linear models with continuous covariates. Australian Journal of Statistics, 34(3):461–470, 1992.

126. Stubbendick, A. L. and Ibrahim, J. G.: Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. Statistica Sinica, 16(4):1143–1167, October 2006.

127. Ibrahim, J. G. and Molenberghs, G.: Missing data methods in longitudinal studies: a review. TEST, 18(1):1–43, May 2009.

128. Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., Lawn, J. E., Cousens, S., Mathers, C., and Black, R. E.: Global, regional, and national causes of under-5 mortality in 2000-15: an updated systematic analysis with implications for the Sustainable Development Goals. Lancet (London, England), 388(10063):3027–3035, 2016.

129. Sebayang, S. K., Dibley, M. J., Kelly, P. J., Shankar, A. V., Shankar, A. H., and SUMMIT Study Group: Determinants of low birthweight, small-for-gestational-age and preterm birth in Lombok, Indonesia: analyses of the birthweight cohort of the SUMMIT trial. Tropical medicine & international health: TM & IH, 17(8):938–950, August 2012.

130. Norton, M.: New evidence on birth spacing: promising findings for improving newborn, infant, child, and maternal health. International Journal of Gynaecology and Obstetrics: The Official Organ of the International Federation of Gynaecology and Obstetrics, 89 Suppl 1:S1–6, April 2005.

131. Rutstein, S. O.: Effects of preceding birth intervals on neonatal, infant and under-five years mortality and nutritional status in developing countries: evidence from

the demographic and health surveys. International Journal of Gynaecology and Obstetrics: The Official Organ of the International Federation of Gynaecology and Obstetrics, 89 Suppl 1:S7–24, April 2005.

132. Hodge, A., Firth, S., Marthias, T., and Jimenez-Soto, E.: Location matters: trends in inequalities in child mortality in Indonesia. Evidence from repeated cross-sectional surveys. PloS One, 9(7):e103597, 2014.

133. Adewuyi, E. O., Zhao, Y., and Lamichhane, R.: Risk factors for infant mortality in rural and urban Nigeria: evidence from the national household survey. Scandinavian Journal of Public Health, 45(5):543–554, July 2017.

134. Hanf, M., Nacher, M., Guihenneuc, C., Tubert-Bitter, P., and Chavance, M.: Global determinants of mortality in under 5s: 10 year worldwide longitudinal study. BMJ : British Medical Journal, 347, 2013.

135. Wang, L.: Determinants of child mortality in LDCs: empirical findings from demographic and health surveys. Health Policy (Amsterdam, Netherlands), 65(3):277–299, September 2003.

APPENDIX

LIST OF VARIABLES CONSIDERED FOR COVARIATES IN THE REAL DATA APPLICATION

Variable	Description	Level	Base
Household Characteristics			
Region (province)	Region of residence in which the household resides. There were 33 provinces: Aceh, North Sumatera, West Sumatera, Riau, Jambi, South Sumatera, Bengkulu, Lampung, Bangka Belitung, Riau Islands, Jakarta, West Java, Central Java, Yogyakarta, East Java, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, Papua.	Household	All household
Residence	Type of place of residence where the household resides: urban or rural.	Household	All household
Total members	Total number of household members based on the number of entries in the household data.	Household	All household
Total de jure members	The number of household members that usually lived in the household.	Household	All household
Total de facto members	The number of household members that slept in the household the previous night, including visitors.	Household	All household
Total women 15-49	Total number of women aged between 15 and 49.	Household	All household
Total children aged 5 or less	Number of children resident in the household and aged 5 and under. Visiting children are not included.	Household	All household

Continued on next page

APPENDIX (Continued)

Variable	Description	Level	Base
Median children ever born	Median number of children ever born among women aged 15-49 with any birth in the household.	Woman	Women 15-49 with any birth
Mean maternal age at first birth	Mean age at first birth of women aged 15-49 with any birth in the household.	Woman	Women 15-49 with any birth
Median living children	Median number of living children of women aged 15-49 with any birth in the household.	Woman	Women 15-49 with any birth
Mode of mother's education	The mode of educational attainment of women aged 15-49 with any birth in the household. Categorized into the following: None, incomplete primary, complete primary, incomplete secondary, complete secondary, higher education.	Woman	Women 15-49 with any birth
Wealth index category	A composite measure of a household's cumulative living standard. The wealth index is calculated using easy-to-collect data on a households ownership of selected assets, such as televisions and bicycles; materials used for housing construction; and types of water access and sanitation facilities. Generated using principal components analysis, the wealth index first placed individual households on a continuous scale of relative wealth. IDHS then separated all interviewed households into five wealth quintiles : poorest, poorer, middle, richer, richest.	Household	All household
Wealth index score	Wealth index factor score (5 decimals)	Household	All household
Health-related Factors			
Drinking water	Major source of drinking water for members of the household. Categorized into: tap water, well, spring/river/lake/pond, rainwater, bottled water, refill water, and other.	Household	All household

Continued on next page

APPENDIX (Continued)

Variable	Description	Level	Base
Time to water source	Time taken to get to the water source for drinking water. Households with drinking water either piped to, or available from a well in, the residence, yard or plot or who use rainwater or bottled water, had time set to 0.	Household	All household
Place to wash hands	Place where household members wash their hands: observed, not observed.	Household	All household
Water source	Location of source for water: in own dwelling, in own yard/plot, or elsewhere.	Household	All household
Toilet facility	Type of toilet facility in the household: private/shared/public WC, pit latrine, open area, other.	Household	All household
Access to electricity	Whether the household has electricity.	Household	All household
Own transportation vehicle	Whether the household has any of these: a bicycle, a motorcycle, or a car.	Household	All household
Any child with diarrhea in past 2 weeks	Whether any child aged 5 or less in the household had diarrhea within the last two weeks.	Child	Living children born in the last 5 years
Any child with fever in past 2 weeks	Whether any child aged 5 or less in the household had fever within the last two weeks.	Child	Living children born in the last 5 years
Any child with a cough in past 2 weeks	Whether any child aged 5 or less in the household had suffered from a cough in the last two weeks.	Child	Living children born in the last 5 years
Any child with a cough with rapid breathing in past 2 weeks	Whether any child aged 5 or less in the household had suffered from rapid breathing when he/she had the cough in the past two weeks.	Child	Living children born in the last 5 years
Any child with diarrhea or URTI in past 2 weeks	Whether any child in the household had diarrhea, fever, or suffered from a cough with or without rapid breathing in the last two weeks.	Child	Living children born in the last 5 years
Mode of contraception use	The mode of contraception use among women 15-49 in the household: ever vs never. Contraception was defined broadly as the use of anything or attempt to delay or avoid getting pregnant.	Woman	All women 15-49

Continued on next page

APPENDIX (Continued)

Variable	Description	Level	Base
Mode of contraceptive method	The mode of the current contraceptive method used by women 15-49 in the household, categorized as either a modern method, a traditional method, or a folkloric method.	Woman	All women 15-49
Proportion of children with complete vaccination	Proportion of living children born in the last 5 years within the household with a complete schedule of the country's basic vaccination for children, which included BCG, polio 0-3, DPT 1-3, and measles vaccination.	Child	Living children born in the last 5 years
Mode of children with complete vaccination	A binary mode (yes/no) of whether the living children born in the last 5 years within the household had a complete schedule of the country's basic vaccination for children, which included BCG, polio 0-3, DPT 1-3, and measles vaccination.	Child	Living children born in the last 5 years
Proportion of children with any vaccination	Proportion of living children born in the last 5 years within the household who ever had any of the basic vaccination (BCG, polio 0-3, DPT 1-3, and measles).	Child	Living children born in the last 5 years
Mode of children with any vaccination	A binary mode of whether the living children born in the last 5 years within the household ever had any of the basic vaccination (BCG, polio 0-3, DPT 1-3, and measles).	Child	Living children born in the last 5 years
Members smoking in house	Frequency of household members smoke inside the house	Household	All household
Birth History			
Median preceding birth interval	Median difference in months between the current birth and the previous birth of children born in the last 5 years within the household, counting twins as one birth.	Child	Children born in the last 5 years
Proportion of women ever had pregnancy complication	Proportion of women 15-59 in the household ever had complications during pregnancy.	Woman	All women 15-49

Continued on next page

APPENDIX (Continued)

Variable	Description	Level	Base
Proportion of women ever had terminated pregnancy	Proportion of women 15-59 in the household ever had a pregnancy that was terminated in a miscarriage, abortion, or still birth, that is, did not result in a live birth.	Woman	All women 15-49
Proportion of births with ANC by SBA	Proportion of births in the last 5 years that had antenatal care (ANC) by a skilled birth attendant (SBA). SBA included doctor, obstetrician, nurse, midwife, or village midwife.	Child	Children born in the last 5 years
Median timing of first ANC	Median months between the start of the pregnancy and the first antenatal visit for the pregnancies related to any birth in the last 5 years within the household.	Child	Children born in the last 5 years
Median number of ANC	Median number of ANC during pregnancy of those led to any birth in the last 5 years within the household.	Child	Children born in the last 5 years
Proportion of births delivered by SBA	Proportion of births in the last 5 years within the household that were assisted by either doctor, obstetrician, nurse, midwife, or village midwife.	Child	Children born in the last 5 years
Mode of delivery place	Most common place of delivery of children born in the last 5 years within the household, categorized into: public hospital/clinic, private hospital/clinic, home, or other.	Child	Children born in the last 5 years
Proportion of births by C-section	Proportion of children born in the last 5 years by caesarian section	Child	Children born in the last 5 years
Proportion of births with PNC	Proportion of children born in the last 5 years with postnatal check within 2 months	Child	Children born in the last 5 years
Proportion of births with PNC by SBA	Proportion of children born in the last 5 years with postnatal checkup performed by medical personnel.	Child	Children born in the last 5 years
Mode of PNC place	Most common place of postnatal checkup among children born in the last 5 years. This variable was grouped into 4 categories: home, public sector, private sector, and other.	Child	Children born in the last 5 years
Mode of perceived child size	Most common category of size of child reported subjectively by the mother among children born in the last 5 years.	Child	Children born in the last 5 years
Mean of child weight (grams)	Mean weight at birth of children born in the last 5 years, given in gram metric.	Child	Children born in the last 5 years

Continued on next page

APPENDIX (Continued)

Variable	Description	Level	Base
Proportion of births with any problem during pregnancy	Proportion of births of children in the household where the mother had any problem during pregnancy. The problems probed including wrong position baby, faint, breathlessness, and tiredness.	Child	All children
Proportion of births with any problem during labor/delivery	Proportion of births of children in the household where the mother had any problem during labor/delivery. The problems probed including water broke too soon, fever, long labor, faint, convulsions, and placenta did not come out.	Child	All children
Proportion of births with any problem after birth	Proportion of births of children in the household where the mother had any problem after giving birth/during seclusion. The problems probed including excessive bleeding, convulsions, fever, foul-smelling discharge, sore breast, and sadness/depression.	Child	All children
Proportion of births with any issue on time of birth	Proportion of births in the last 5 years within the household where the mother had any problem in time of birth. The problems probed including prolonged labor, fever and foul smelling vaginal discharge, convulsions, and water broke $>$ 6 hours before delivery.	Child	Children born in the last 5 years

VITA

NAME	Fima Lanra Fredrik Gerarld Langi
EDUCATION	B.A., Medicine, Universitas Sam Ratulangi, Manado, Indonesia, 1997
	Dokter (M.D.), General Practice, Universitas Sam Ratulangi, Manado, Indonesia, 1999
	M.Med.Stats (M.S., Medical Statistics), University of Newcastle, Newcastle, Australia, 2006
	Ph.D., Biostatistics, University of Illinois at Chicago, Chicago, Illinois, 2017
WORK	Graduate Research Assistant, Institute on Disability and Human Development (Research Fund: Grant from Division of Developmental Disabilities, the State of Illinois Department of Human Services), College of Applied Health Sciences, University of Illinois at Chicago, Chicago, Illinois, February - June 2017
	Graduate Research Assistant, Department of Disability and Human Development (Research Fund: Grant from Division of Rehabilitation Services, the State of Illinois Department of Human Services), College of Applied Health Sciences, University of Illinois at Chicago, Chicago, Illinois, June 2012 - October 2016
	Graduate Research Assistant, Division of Epidemiology and Biostatistics (Research Fund: the Mtoto Msafi Mbili Study), School of Public Health, University of Illinois at Chicago, Chicago, Illinois, August 2015 - August 2016
	Teaching Assistant (BSTT 400 and BSTT 401), Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, Chicago, Illinois, August 2012 - May 2014, August - December 2016
	Instructor, Department of Epidemiology and Biostatistics, School of Public Health, Universitas Sam Ratulangi, Manado, Indonesia, December 2001 - July 2009
	Medical Officer, Community Health Center of Kendahe, the District of Sangihe Talaud Department of Health, Tahuna, Indonesia, August 1999 - July 2001

HONORS

Supersemar Award, the Republic of Indonesia Ministry of Education and Culture, 1995-1999

Australian Development Scholarship, AusAID and the Government of Australia, 2004-2006

Fulbright Presidential Scholarship, USAID and the United States Department of State, 2009-2014

PUBLICATION

Langi, F. L. F. G., and Balcazar, F. E. Risk Factors for Failure to Enter Vocational Rehabilitation Services among Individuals with Disabilities. Disability and Rehabilitation, 39(26):2640-2647, 2017.

Bailey, R. C., Adera, F., Mackesy-Amiti, M. E., Adipo, T., Nordstrom, S. K., Mehta, S. D., Jaoko, W. **Langi, F. L. F. G.**, Obiero, W., Obat, E., Otieno, F. O., and Young, M. R. Prospective comparison of two models of integrating early infant male circumcision with maternal child health services in Kenya: the Mtoto Msafi Mbili Study. PloS One, (in press).

Langi, F. L. F. G., Oberoi, A., Balcazar, F. E., and Awsumb, J. Vocational Rehabilitation of Transition-Age Youth with Disabilities: A Propensity-Score Matched Study. Journal of Occupational Rehabilitation, 27(1):15-23, 2017.

Langi, F. L F. G., Oberoi, A. K., and Balcazar, F. E. Toward a Successful Vocational Rehabilitation in Adults with Disabilities: Does Residential Arrangement Matter? Journal of Prevention and Intervention in the Community, 45 (2):124-137, 2017.

Keshtgarpour, M., Tan, W. S., Zwanziger, J., Awadalla, S., **Langi, F. G.**, and Dudek, A. Z. Prognostic Value of Serum Proteomic Test and Comorbidity Index in Diversified Population with Lung Cancer. Anticancer Research, 36(4):1759-65, 2016.

Oberoi, A., Balcazar, F. E., Suarez-Balcazar, Y., **Langi, F. L. F. G.**, and Lukyanova, V. Employment outcomes among African American/Black and White Women with Disabilities: Examining the Inequalities. Women, Gender, and Families of Color, 3(2):144-64, 2015.

Langi, F. L. F. G., and Langi, G. Barriers to delivery care by skilled attendants in Sulawesi Utara. Journal of Public Health and Development, 7(3):1-13, 2009.