Bayesian Models for Combining Information from Multiple Sources

by

Jiurui Tang

Department of Statistical Science
Duke University

Date: ______________________
Approved:

____
Jerome Reiter, Supervisor

____
Jason Xu

____
Rebecca Steorts

____
Sunshine Hillygus

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University

2022

ABSTRACT

Bayesian Models for Combining Information from Multiple Sources

by

Jiurui Tang

Department of Statistical Science
Duke University

Date: ______________________
Approved:

____
Jerome Reiter, Supervisor

____
Jason Xu

____
Rebecca Steorts

____
Sunshine Hillygus

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University

2022

Abstract

This dissertation develops Bayesian methods for combining information from multiple sources. I focus on developing Bayesian bipartite modeling for simultaneous regression and record linkage, as well as leveraging auxiliary information on marginal distributions for handling item and unit nonresponse and accounting for survey weights.

The first contribution is a Bayesian hierarchical model that allows analysts to perform simultaneous linear regression and probabilistic record linkage. This model allows analysts to leverage relationships among the variables to improve linkage quality. It also potentially offers more accurate estimates of regression parameters compared to approaches that use a two-step process, i.e., link the records first, then estimate the linear regression on the linked data. I propose and evaluate three Markov chain Monte Carlo algorithms for implementing the Bayesian model.

The second contribution is examining the performance of an approach for generating multiple imputation data sets for item nonresponse. The method allows analysts to use auxliary information. I examine the approach via simulation studies with Poisson sampling. I also give suggestions on parameter tuning.

The third contribution is a model-based imputation approach that can handle both item and unit nonresponse while accounting for auxiliary margins and survey weights. This approach includes an innovative combination of a pattern mixture model for unit nonresponse and a selection model for item nonresponse. Both unit and item nonresponse can be nonignorable. I demonstrate the model performance with simulation studies under the situations when the design weights for unit respondents are known and when they are not. I show that the model can generate multiple imputation data sets that both retain the relationship among survey variables and yield design-based estimates that agree with auxiliary margins. I use the model to

analyze voter turnout overall and across subgroups in North Carolina, with data from the 2018 Current Population Survey.

Acknowledgements

I would like to thank my advisor, Jerry Reiter, for his support and guidance through my Ph.D study. He is always patient, caring, inspiring and hard-working. He helped me overcome many obstacles in my research, and offered tremendous great suggestions in improving my writing. I feel grateful and fortunate to have him as my mentor.

I would like to thank everyone in the department of statistical science for a great experience. It feels like home. Thanks to Olanrewaju Akande and Mauricio Sadinle for sharing and explaining their past work. Thanks to Gabriel Madson and D. Sunshine Hillygus for their efforts in our collaborative work on analyzing data from the Current Population Survey.

I am grateful to my parents for having my back all these years. Without their support, I couldn't have come so far. Thanks to my friends for helping me in one way or another.

This research was partially supported by the National Science Foundation through grants SES1131897, SES1733835, SES1652431 and SES1534412.

Dedication

To those who gave me courage and strength to persist through adversity.

Contents

List of Figures

List of Tables

1

Introduction

As data processing equipment and techniques advance, researchers find it easier to combine information from various sources which means analysis does not need to be confined to a single data set. This dissertation concerns two areas of combining information across databases: merging data sets and handling missing data with auxiliary margins. For the first area, we have two data sets and would like to link them so that richer models can be built without additional costs of new data collection. For the second area, we have a data set with missing values and some population level information regarding the variables in the data set. We would like to generate completed data sets to better reflect reality with the help of this population level external information.

To provide context for the first area of this dissertation, increasingly analysts seek to link records across data sets to facilitate statistical analyses. As a prototypical example, a health researcher seeks to link data from a previously completed study to patients' electronic medical records to collect long-term outcomes, with the ultimate goal of estimating relationships between the long-term outcomes and baseline covariates (Tang et al., 2020). Such linkages are performed readily when unique identifiers, such as social security numbers, are available for all records in all data sets. Often, however, one or more of the data sets do not have unique identifiers, perhaps because they were never collected or are not made available due to privacy concerns. In such cases, analysts have to link records based on indirect identifiers, such as name, date of birth, and demographic variables (Christen, 2012, 2019). Generally, such indirect identifiers contain distortions and errors. As a result, they can differ across the data sets, which can make it difficult to determine the correct record linkages. This uncertainty should be quantified and propagated to statistical inferences, although typically this is not done.

In the statistics literature, the most popular method for linking records via indirect identifiers is based on the probabilistic record linkage (RL) approach of Newcombe et al. (1959), which was later extended and formalized by Fellegi and Sunter (1969). Many extensions to the Fellegi-Sunter (FS) model have been proposed (e.g., Winkler, 2006, 2014; Sadinle and Fienberg, 2013). A common drawback of these and other probabilistic record linkage methods (e.g., Larsen and Rubin, 2001; Hof et al., 2017) is the difficulty in quantifying linkage uncertainty. These limitations have led to developments of record linkage approaches from Bayesian perspectives (e.g., Dalzell and Reiter, 2018; Fortini et al., 2001; Larsen, 2001, 2002, 2005; Tancredi and Liseo, 2011; Zhao et al., 2012; Gutman et al., 2013; Sadinle, 2014; Steorts, 2015; Steorts et al., 2016; Zanella et al., 2016; Sadinle, 2017; Marchant et al., 2021). Frequently, obtaining a linked data set is not the ultimate goal of analysis. Researchers would like to perform downstream tasks with the linked data set, for example, using linear regression to study the relationships between variables belonging to different data sets before the linkage. In this dissertation, I present methods for performing record linkage and regression analysis simultaneously and show that it can improve inferences over doing each separately.

To provide context for the second area, many researchers encounter missing data, especially when dealing with survey data. It is inefficient to discard incomplete entries and use only the individuals with complete data. Moreover, this could lead to biased or even misleading conclusions. Multiple techniques have been proposed to deal with missing data. For example, a simple approach could be mean or median imputation of observed entries, which in special cases may give reasonable point estimates, but typically underestimates variances and attenuates associations. There also are methods such as hot and cold deck imputations. In this dissertation, I consider multiple imputation as proposed in Rubin (1987). Several completed data sets are generated to account for the uncertainty due to the missingness. Point estimates together with uncertainty quantification can be calculated via multiple imputation combination rules.

Auxiliary information, in particular population totals or percentages, can help with missing data imputation. When generating completed data sets for survey data in the presence of auxiliary information, we not only want the completed data sets to reflect relationships among the survey variables, but also want them to result in plausible design-based estimates of known margins. Akande et al. (2021) propose a framework for using auxiliary margins in multiple imputation; however, this framework only applies for simple random samples and does not consider complex survey weights. Akande and Reiter (ming) discuss an algorithm to achieve this goal for survey data from stratified simple random sampling, but not other complex survey designs. In this dissertation, I examine the performance of this algorithm for more general survey designs with item nonresponse. I also propose a new model-based imputation method to generate completed data sets that leverage the auxiliary margins to handle both item and unit nonresponse in complex surveys.

This dissertation consists of three chapters in addition to this introductory chapter. In Chapter 2, I propose a Bayesian hierarchical model that allows data analysts to perform simultaneous linear regression and probabilistic record linkage. This allows analysts to leverage relationships among the variables to improve linkage quality. Further, it enables analysts to propagate uncertainty in a principled way, while also potentially offering more accurate estimates of regression parameters compared to approaches that use a two-step process, i.e., link the records first, then estimate the linear regression on the linked data. I propose and evaluate three Markov chain Monte Carlo algorithms for implementing the Bayesian model, which I compare against a two-step process.

In Chapter 3, I review the algorithm in Akande and Reiter (ming) and conduct various simulation studies to test its performances in Poisson sampling when the data suffer from item nonresponse alone. In these simulations, I generate data sets with weights sampled from the inverse of a Beta distribution as well as weights from Current Population Survey (CPS) data. The results suggest that the method can use the auxiliary information to improve imputations and resulting inferences.

In Chapter 4, I propose a model-based imputation approach to account for both item and unit nonresponses in the presence of auxiliary margins. In particular, I use a pattern mixture model for unit nonresponse and a selection model for item nonresponse. I show that, with auxiliary margins, we can allow nonignorable missing in unit nonresponse or item nonresponse. I present results of simulation studies under situations when we have the design weights for unit respondents and when we do not. Lastly, I present a case study using CPS 2018 data and analyze subgroup voter turnout across gender, age, and ethnicity groups.

In the reminder of this chapter, I provide background material related to record linkage and missing data methods. In section 1.1, I review the Fellegi-Sunter record linkage model. In section 1.2, I define various missing data mechanisms and review the framework developed in (Akande et al., 2021) for handling missing data with auxiliary margins. In section 1.3, I review the multiple imputation combining rules in Rubin (1976) to calculate point and uncertainty estimates.

1.1 Record Linkage

Fellegi and Sunter (1969) proposed a model for record linkage that has become the most popular probabilistic approach over the decades. This model does not require training data or restrict the type of fields that need to be compared. In Chapter 2, I utilize a Bayesian version of the Fellegi-Sunter model. Thus, to provide context for that chapter, in this section I review the Fellegi-Sunter model.

Let A and B denote two files, and let $a \in A$ and $b \in B$ represent records in these files. The set of ordered pairs can be viewed as the union of two disjoint sets $M = \{(a,b); a = b, a \in A, b \in B\}$ and $U = \{(a,b); a \neq b, a \in A, b \in B\}$, where M is the set of matches and U the set of nonmatches. We call record pairs links if they are estimated to belong in M and nonlinks if they are estimated to belong in U. We define the comparision vector of record pair (a,b) as $\gamma(a,b) = \left(\gamma^1(a,b), \ldots, \gamma^F(a,b)\right)$ or simply as γ when no confusion arises. F denotes the number of linking variables or fields. I describe examples of linking variables and $\gamma(a,b)$ in the review of Bayesian record linkage in Chapter 2. Let Γ denote the comparison space, the set of all possible realizations of γ.

For each record pair, there are three possible decisions we could make. First, we declare $(a,b) \in M$ and denote this as A_1. Second, we declare $(a,b) \in U$ and denote this as A_3. Third, we are ambiguous about whether (a,b) is a link or nonlink, and denote this as A_2. Both A_1 and A_3 are definitive decisions, whereas A_2 is not. Let $m(\gamma) = \mathbb{P}(\gamma(a,b)|(a,b) \in M)$ and $u(\gamma) = \mathbb{P}(\gamma(a,b)|(a,b) \in U)$. Two types of error could happen: when an unmatched record pair is declared as link, and when a matched pair is declared as nonlink. Similar to traditional hypothesis testing, we denote the error levels as $\mathbb{P}(A_1|U) = \mu$ and $\mathbb{P}(A_3|M) = \lambda$.

Fellegi and Sunter (1969) present an optimal linkage rule in the sense that for fixed

μ and λ, the probability of unable to make a positive decision, i.e., declaring record pairs in A_2, is minimized. This procedure orders $\gamma(a, b)$ so that the corresponding sequence of $m(\gamma)/u(\gamma)$ is nonincreasing. An order is arbitrarily assigned when $m(\gamma) > 0$ but $u(\gamma) = 0$, and when there are ties in the value of $m(\gamma)/u(\gamma)$. This ordered set is indexed by the superscript i. Pick n and n' such that $\sum_{i \leq n-1} u(\gamma^i) < \mu \leq \sum_{i \leq n} u(\gamma^i)$ and $\sum_{i \geq n'} m(\gamma^i) \geq \lambda > \sum_{i \geq n'+1} u(\gamma^i)$. Those record pairs with $i \leq n-1$ are declared as links, those with $i \geq n' + 1$ are declared as nonlinks. The rest are undecided and need review by experts case by case.

1.2 Missing Data Mechanisms

Rubin (1976) is a seminal paper in missing data research. It proposed to treat missing data indicators as random variables and gave the weakest conditions under which ignoring the missing data mechanism always leads to correct inferences. Mealli and Rubin (2015) made clarification about interpretation of missing at random as a statement of conditional independence with the original notations and definitions. We restate the important definitions of missing data mechanisms here as shown in (Mealli and Rubin, 2015).

For ease of notation here, we define missingness mechanisms for single variables. The notation can be extended to multiple variables. Let $Y = (Y_1, \ldots, Y_n)$ denote the data of n units with probability density f_θ, where each Y_i is a scalar and θ is the parameter of this density. Let $R = (R_1, \ldots, R_n)$ be the missingness indicator for Y. More specifically, $R_i = 1$ if Y_i is missing, and $R_i = 0$ if Y_i is observed. Let $r = (r_1, \ldots, r_n)$ and $y = (y_1, \ldots, y_n)$ be realizations of R and Y. Let the probability that R takes the value r given that Y takes the value y be $p(R = r|Y = y, \phi)$. ϕ is the parameter associated with the missing data mechanism.

The ordered set $\{1, \ldots, n\}$ can be partitioned into two sets based on the value of r: $\{i : r_i = 1\}$ which we abbreviate as (1) and $\{i : r_i = 0\}$ which we abbreviate as (0). Y can also be partitioned into $Y_{(1)} = (Y_i : r_i = 1)$ and $Y_{(0)} = (Y_i : r_i = 0)$, and similarly y into the missing values $y_{(1)} = (y_i : r_i = 1)$ and the observed values $y_{(0)} = (y_i : r_i = 0)$. Let $\tilde{r}$ denote a particular sample realization of R, and $\tilde{y}_{(0)}$ the corresponding sample realization of $Y_{(0)}$.

1.2.1 Missing Completely At Random and Missing Always Completely At Random

The missing data are missing completely at random (MCAR) if

$$p(R = \tilde{r}|Y = y, \phi) = p(R = \tilde{r}|\phi) \tag{1.1}$$

for all y, and all ϕ.

The missing data are missing always completely at random (MACAR) if

$$p(R = r|Y = y, \phi) = p(R = r|\phi) \tag{1.2}$$

for all y, all r, and all ϕ.

1.2.2 Missing At Random and Missing Always At Random

The missing data are missing at random (MAR) if

$$p(R = \tilde{r}|Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}, \phi) = p(R = \tilde{r}|Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y'_{(1)}, \phi) \tag{1.3}$$

for all ϕ and all $y_{(1)}, y'_{(1)}$.

The missing data are missing always at random (MAAR) if

$$p(R = r|Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)}, \phi) = p(R = r|Y_{(0)} = y_{(0)}, Y_{(1)} = y'_{(1)}, \phi) \tag{1.4}$$

for all ϕ and all $r, y_{(0)}, y_{(1)}, y'_{(1)}$.

1.2.3 Missing Not At Random and Missing Not Always At Random

The missing data are missing not at random (MNAR) if

$$p(R = \tilde{r}|Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}, \phi) \neq p(R = \tilde{r}|Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y'_{(1)}, \phi) \tag{1.5}$$

for some ϕ, and some $y_{(1)} \neq y'_{(1)}$.

The missing data are missing not always at random (MNAAR) if

$$p(R = r|Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)}, \phi) \neq p(R = r|Y_{(0)} = y_{(0)}, Y_{(1)} = y'_{(1)}, \phi) \tag{1.6}$$

for some $\phi, r, y_{(0)}$ and some $y_{(1)} \neq y'_{(1)}$.

Based on the above definitions, we see that MCAR, MAR, and MNAR are weaker versions of MACAR, MAAR, and MNAAR. If a missing data mechanism satisfies MACAR, MAAR, or MNAAR, then any realizations of it will correspondingly satisfy MCAR, MAR, or MNAR. We will assume MCAR, MAR, and MNAR for the rest of the dissertation as considering all realizations of missing data mechanism is not the primary concern. Among the three missing data mechanisms we just defined, MNAR is also called nonignorable missing data mechanism because the missing data indicator depends on missing values. Ignoring the missing data mechanism will lead to biased inferences and incorrect conclusions. On the other hand, MCAR and MAR are ignorable missing mechanisms. I introduce two nonignorable missing data mechanisms that are highly related to my own work next.

1.2.4 Itemwise Conditionally Independent Nonresponse

Itemwise conditionally independent nonresponse (ICIN) is a nonignorable missing data mechanism proposed in (Sadinle and Reiter, 2017). ICIN is relevant for data sets with multiple variables. It happens when each variable and its missing data indicator are conditionally independent given other variables and their missing data indicators. It is a nonignorable missing data mechanism because the missing data indicators can depend on unobserved values of other variables. Formally, let $(Y_1, \ldots, Y_p)$ denote the data of p random variables, and for $j = 1, \ldots, p$, let R_j be the nonresponse indicator for Y_j. Define $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$ and likewise R_{-j}. Then, nonresponse is itemwise conditionally independent when

$$Y_j \perp R_j | Y_{-j}, R_{-j} \qquad (j = 1, \ldots, p).$$

1.2.5 Additive Nonignorable Model

The Additive Nonignorable (AN) model was developed originally to deal with longitudinal data with refreshment samples when the primary variable of interest has missing values. I use the AN model to work with the missing data problem in this dissertation by viewing the information from a refreshment sample as auxiliary data. In this section, I review the AN model in the general context of item nonresponse in a two-variable missing data example. In the original development by Hirano et al. (2001), the auxiliary information is in the form of record-level data. Here, and in the rest of this dissertation, we consider situations where the auxiliary information is known marginal totals or percentages.

Suppose we have two binary variables X_1 and Y_1. We have auxiliary information for X_1, and no auxiliary information for Y_1. Also, suppose X_1 suffers from item nonresponse but Y_1 is fully observed, so that R_1^x is the fully observed vector of item nonresponse indicators for X_1. There is no need to include a model for R_1^y since Y_1 is fully observed. The observed data for our two-variable example together with auxiliary information available on X_1 are presented in Table 1.1a, and the joint distribution of (X_1, Y_1, R_1^x) is shown in the form of a contigency table of 8 cells in Table 1.1b with auxiliary information excluded. We use "✓" to represent observed data, and "?" to represent missing data. We can see graphically the extent of sparsity in the full joint distribution from Table 1.1a and Table 1.1b.

Table 1.1: Two binary variables Y_1 and X_1: Y_1 is fully observed, X_1 suffers from item nonresponse. "✓" represents observed, "?" represents missing.

(a) Data

<table>
<tr><th>X_1</th><th>Y_1</th><th>R_1^x</th><th></th></tr>
<tr><td>✓</td><td rowspan="2">✓</td><td>0</td><td rowspan="2">}Observed Data</td></tr>
<tr><td>?</td><td>1</td></tr>
<tr><td>✓</td><td>?</td><td>?</td><td>⟶ Auxiliary Margin</td></tr>
</table>

(b) Contigency table

	$R_1^x = 0$		$R_1^x = 1$	
	$X_1 = 0$	$X_1 = 1$	$X_1 = 0$	$X_1 = 1$
$Y_1 = 0$	✓	✓	?	?
$Y_1 = 1$	✓	✓	?	?

Using a pattern mixture model factorization (Glynn et al., 1986; Little, 1993), we can write the joint distribution of (X_1, Y_1, R_1^x) as follows:

$$\begin{aligned} \mathbb{P}(X_1 = x, Y_1 = y, R_1^x = r) &= \mathbb{P}(X_1 = x | Y_1 = y, R_1^x = r) \\ &\quad \times \mathbb{P}(Y_1 = y | R_1^x = r) \times \mathbb{P}(R_1^x = r). \end{aligned} \tag{1.7}$$

As shown in Table 1.1b, there are in total $2^3 = 8$ joint probabilities to be estimated. Let $\theta_{yr} = \mathbb{P}(X_1 = 1|Y_1 = y, R_1^x = r)$, $\pi_r = \mathbb{P}(Y_1 = 1|R_1^x = r)$, and $q = \mathbb{P}(R_1^x = 1)$. We can estimate five out of seven parameters, namely $q, \pi_0, \pi_1, \theta_{00}, \theta_{10}$, directly from the observed data. More specifically, we can estimate q because R_1^x is fully observed. We can estimate π_0 and π_1 because the joint distribution of Y_1 and R_1^x is fully observed. We can estimate θ_{00} and θ_{10} because given $R_1^x = 0$, the joint distribution of X_1 and Y_1 is fully observed. The remaining two parameters, θ_{01}, θ_{11}, are not estimable based solely on observed data.

Auxiliary information about the marginal distribution of X_1 puts an additional constraint via the equation $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = 1, R_1^x = 0) + \mathbb{P}(X_1 = 1, R_1^x = 1)$. This gives a linear constraint on what values are possible for θ_{01} and θ_{11} simultaneously. We still experience identifiability issue with the presence of auxiliary margin on X_1, but it increases the number of estimable parameters from five to six. An additional constraint is needed to fully parameterize the joint distribution of (X_1, Y_1, R_1^x).

We re-write the joint distribution of (X_1, Y_1, R_1^x) in the form of selection model parameterization (Little, 1995) as in (Hirano et al., 2001):

$$(X_1, Y_1) \sim f(X_1, Y_1|\Theta), \tag{1.8}$$

$$\mathbb{P}(R_1^x = 1|X_1, Y_1) = h(\gamma_0 + \gamma_1 Y_1 + \gamma_2 X_1), \tag{1.9}$$

for some known, strictly increasing $h(a)$ such that $\lim_{a\to-\infty} h(a) = 0$ and $\lim_{a\to\infty} h(a) = 1$. Θ represents parameters in f, the joint distribution of X_1 and Y_1, and $(\gamma_0, \gamma_1, \gamma_2)$ parameters in h. Common choices of h can be logit or probit link function, between which (Hirano et al., 2001) finds little sensitivity in the choice of $h(\cdot)$.

The AN model permits nonignorability in R_1^x by allowing γ_2, the coefficient in front of X_1, to differ from 0, and hence, the missingness pattern of X_1 depends on the value of X_1. On the other hand, this model does not allow an interaction term between X_1 and Y_1. With additive non-ignorability, we have one more constraint on the joint distribution of (X_1, Y_1, R_1^x) and thus cannot identify an interaction term.

Several common models for missingness are special cases of the AN model. If $\gamma_1 = \gamma_2 = 0$ in (1.9), we obtain the MCAR mechanism. More generally, if $\gamma_2 = 0$ and $\gamma_1 \neq 0$, we obtain an MAR mechanism, where missingness depends on variables excluding oneself. When $\gamma_2 \neq 0$, we have the MNAR mechanism. When $\gamma_1 = 0$ and $\gamma_2 \neq 0$, we have a nonignorable missing mechanism called the Hausman and Wise model (Hausman and Wise, 1979). For more general cases and an identifiability proof of the AN model, see Hirano et al. (2001) for details. For incorporating multiple refreshment samples under different longitudinal data collection designs, see Deng et al. (2013). For nonignorable unit nonresponse both in the initial wave and in the refreshment samples, see Schifeling et al. (2015).

1.2.6 MD-AM Framework

The missing data with auxiliary margins (MD-AM) framework proposed in (Akande et al., 2021) can handle nonignorable nonresponse mechanisms such as the AN model, the ICIN model and ignorable nonresponse mechanisms like MAR and MCAR, while also using the information in one or more available margins. Because I build on the MD-AM framework, in this section I review its general ideas.

Let $\mathcal{D}$ comprise data from the survey of $i = 1, \ldots, n$ individuals, and $\mathcal{A}$ comprise data from the auxiliary database. The MD-AM framework proposes specifying the joint distribution of all variables in $\mathcal{D}$ and nonresponse indicators using a sequential factorization of identifiable conditional models. The models for the nonresponse indicators are some combination of nonignorable and ignorable models. By adding auxiliary information, we are able to specify identifiable models that maximize all available information. This framework consists of two steps quoted from (Akande et al., 2021) page 645.

1. "We begin with a model for the survey variables and nonresponse indicators that can be identifiable using the observed data alone, without any auxiliary information. Preferably, the model should allow for the maximum number of parameters identifiable from the observed data alone. This can be done using either a selection model or a pattern mixture factorization (Glynn et al., 1986; Little, 1993), according to the analysts preference. Generally, this step results in models, such as MAR models, that are often default choices for handling nonresponse in the missing data literature, absent auxiliary data.

2. We next find sets of parameters that can be added to the model in Step 1, so that the model still can be identified because of the auxiliary information. Typically, there are multiple identifiable models - determined by the nature of the auxiliary information - each representing different assumptions about the missingness process. Agencies can choose from among these models according to interpretability and plausibility for the data at hand."

To illustrate this framework, we use again the two binary variables example when introducing the AN model above. According to step 1, we should write down a MAR model for the item nonresponse indicator R_1^x. Since we never fully observe the joint distribution of X_1 and R_1^x absent the auxiliary information, we do not include X_1 as a main effect for the model of R_1^x. We can estimate five parameters directly from the observed data. The obvious choice of model with five parameters in total is

$$(X_1, Y_1) \sim f(X_1, Y_1|\Theta), \tag{1.10}$$

$$\mathbb{P}(R_1^x = 1|X_1, Y_1) = h(\gamma_0 + \gamma_1 Y_1). \tag{1.11}$$

γ_0 is identifiable because R_1^x is fully observed. As displayed in Table 1.1a, γ_1 is identifiable since the joint probabilities of R_1^x and Y_1 are known from the observed data alone. Under the conditional independence assumption between R_1^x and X_1 in (1.11), Θ is identifiable for common choices of f, for example, multinomial distribution or a sequence of logistic regression models. We can also specify an MCAR for R_1^x instead of (1.11), resulting in a model with four parameters in total. However, this option is not recommended as it does not allow for relationships between item nonresponse and Y_1.

Next, we move on to step 2 to add a main effect term of X_1 based on auxiliary margin of $\mathbb{P}(X_1 = 1)$. We add this main effect of X_1 to (1.11), resulting in

$$\mathbb{P}(R_1^x = 1|X_1, Y_1) = h(\gamma_0 + \gamma_1 Y_1 + \gamma_2 X_1). \tag{1.12}$$

(1.10) and (1.12) together have six identifiable parameters. Note that for the two binary variable example here, the model we write down following MD-AM framework steps is exactly the AN model.

We have illustrated how to apply the MD-AM framework under a selection model factorization. In fact, using a pattern mixture model factorization is not much different. Following step 1, the obvious choice under a pattern mixture model factorization with five parameters is

$$(Y_1, R_1^x) \sim f(Y_1, R_1^x|\Phi) \tag{1.13}$$

$$\mathbb{P}(X_1 = 1|Y_1, R_1^x) = h(\gamma_0 + \gamma_1 Y_1). \tag{1.14}$$

We cannot have any term capturing relationship between R_1^x and X_1 in step 1 based on observed data alone. In step 2, we add $\gamma_2 R_1^x$ term to (1.14), resulting in

$$\mathbb{P}(X_1 = 1|Y_1, R_1^x) = h(\gamma_0 + \gamma_1 Y_1 + \gamma_2 R_1^x). \tag{1.15}$$

Researchers can choose between pattern mixture and selection model factorization based on their preference.

1.3 Multiple Imputation

Multiple imputation (MI) is a common way to handle missing data. As the name suggests, analysts generate multiple completed data sets where the missing values are draws from predictive distributions. These completed data sets are then used for analysis or released for public use. Estimates for estimands of interest are computed in each completed data set, and these estimates and their corresponding variance estimates are combined based on Rubin's multiple imputation combining rules (Rubin, 1987) to account for uncertainty introduced by the imputation process. In this section, we review Rubin's combining rules, which are used in Chapter 3 and 4.

Let Q be the estimand of interest, for example, a population mean or regression parameter. Let q be the estimator of Q and u be the estimator of the variance of q in a complete data set. For $\ell = 1, \ldots, L$, let $q^{(\ell)}$ be the estimate of Q, and $u^{(\ell)}$ the estimate of variance of $q^{(\ell)}$ in the completed data set $\boldsymbol{Z}^{(\ell)}$. The estimate of Q is

$$\bar{q}_L = \sum_{\ell}^{L} \frac{q^{(\ell)}}{L}, \tag{1.16}$$

and its estimate of variance is

$$T_L = (1 + \frac{1}{L})b_L + \bar{u}_L, \tag{1.17}$$

where

$$b_L = \sum_{\ell=1}^{L} \frac{(q^{(\ell)} - \bar{q}_L)^2}{L-1} \tag{1.18}$$

and

$$\bar{u}_L = \sum_{\ell=1}^{L} \frac{u^{(\ell)}}{L}. \tag{1.19}$$

Inferences can be made based on the t-distribution

$$(\bar{q}_L - Q) \sim t_\nu(0, T_L), \tag{1.20}$$

where

$$\nu = (L-1)\Big(\frac{1 + \bar{u}_L}{(1+\frac{1}{L})b_L}\Big)^2 \tag{1.21}$$

is the degrees of freedom.

Chapter 3 of Rubin (1987) shows a Bayesian derivation of these inferences. Rubin (1996) presents a more detailed review of MI; Barnard and Meng (1999) review MI as used in medical studies; Reiter and Raghunathan (2007) describe various adaptations of MI rules not only in missing data methods, but also in synthetic data and measurement error correction; and, Akande et al. (2017) review and compare MI methods for categorical data.

2

Bayesian Modeling for Simultaneous Regression and Record Linkage

The presentation in this chapter closely follows Tang et al. (2020), where the research first appeared.

2.1 Introduction

In this chapter, we propose a Bayesian model for performing probabilistic RL and linear regression simultaneously. The proposed model quantifies uncertainty about the linkages and propagates this uncertainty to inferences about the regression parameters. We focus on bipartite RL—that is, the analyst seeks to merge two data sets—assuming that individuals appear at most once in each data set. As we illustrate, the model can leverage relationships among the dependent and independent variables in the regression to potentially improve the quality of the linkages. This also can increase the accuracy of resulting inferences about the regression parameters.

We use a Bayesian hierarchical model that builds on prior work by Sadinle (2017), who proposed a Bayesian version of the FS model for merging two data sets. In fact, one of our primary contributions is to turn the model in (Sadinle, 2017) into a procedure for jointly performing probabilistic RL and fully Bayesian inference for regression parameters. We also propose and evaluate the effectiveness of three algorithms for fitting the Bayesian hierarchical model, focusing on both the quality of the linkages and on the accuracy of the parameter estimates.

2.2 Review of Bayesian Probabilistic Record Linkage

In this section, we review the Bayesian bipartite RL model of (Sadinle, 2017). Consider two data sets $\mathbf{A}_1$ and $\mathbf{A}_2$, containing n_1 and n_2 records, respectively. Without loss of generality, assume $n_1 \geq n_2$. Our goal is to link records in $\mathbf{A}_1$ to records in $\mathbf{A}_2$. We further assume that $\mathbf{A}_1$ and $\mathbf{A}_2$ do not contain duplicate records; that is, each record in $\mathbf{A}_1$ corresponds to a single individual, as is the case for $\mathbf{A}_2$. We assume that some of the same individuals are in $\mathbf{A}_1$ and $\mathbf{A}_2$.

To characterize this, we define the random variable $\mathbf{Z} = (Z_1, \ldots, Z_{n_2})$ as the vector of matching labels for the records in $\mathbf{A}_2$. For $j = 1, \ldots, n_2$, let

$$Z_j = \begin{cases} i, \text{ if record } i \in \mathbf{A}_1 \text{ and } j \in \mathbf{A}_2 \text{ refer to the same entity;} \\ n_1 + j, \text{ if record } j \in \mathbf{A}_2 \text{ does not have a match in } \mathbf{A}_1. \end{cases}$$

Analysts determine whether a pair of records (i, j) is a link, i.e., whether or not $Z_j = i$, by comparing values of variables that are common to $\mathbf{A}_1$ and $\mathbf{A}_2$. Suppose we have F common variables, also known as *linking variables* or *fields*. For $f = 1, \ldots, F$, let γ_{ij}^f represent a score that reflects the similarity of field f for records i and j. For example, when field f is a binary variable, we can set $\gamma_{ij}^f = 1$ when record i agrees with record j on field f, and $\gamma_{ij}^f = 0$ otherwise. When field f is a string variable like name, we can calculate a similarity metric like the Jaro-Winkler distance (Jaro, 1989; Winkler, 1990) or the Levenshtein edit distance (Levenshtein, 1965). We can convert these string metrics to γ_{ij}^f by categorizing the scores into a multinomial variable, where the categories represent the strength of agreement. We illustrate this approach in section

2.4.

For each record (i, j) in $\mathbf{A}_1 \times \mathbf{A}_2$, let $\boldsymbol{\gamma}_{ij} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^F)$. We assume $\boldsymbol{\gamma}_{ij}$ is a realization of a random vector Γ_{ij} distributed as

$$\Gamma_{ij}|Z_j = i \overset{iid}{\sim} \mathcal{M}(\mathbf{m}), \qquad \Gamma_{ij}|Z_j \neq i \overset{iid}{\sim} \mathcal{U}(\mathbf{u}), \qquad \text{where}$$

$\mathcal{M}(\mathbf{m})$ represents the model for comparison vectors among matches, and $\mathcal{U}(\mathbf{u})$ represents the model for comparison vectors among non-matches. For each field f, we let $m_{f\ell} = \mathbb{P}(\Gamma_{ij}^f = \ell | Z_j = i)$ be the probability of a match having level ℓ of agreement in field f, and let $u_{f\ell} = \mathbb{P}(\Gamma_{ij}^f = \ell | Z_j \neq i)$ be the probability of a non-match having level ℓ of agreement in field f. Let $\mathbf{m}_f = (m_{f1}, \ldots, m_{fL_f})$ and $\mathbf{u}_f = (u_{f1}, \ldots, u_{fL_f})$; let $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_F)$ and $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_F)$.

For computational convenience, it is typical to assume the comparison fields are conditionally independent given the matching status of the record pairs. Let $\boldsymbol{\Theta} = (\mathbf{m}, \mathbf{u})$. The likelihood of the comparison data can be written as

$$\mathcal{L}(\mathbf{Z}|\boldsymbol{\Theta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n_1} \prod_{j=1}^{n_2} \prod_{f=1}^{F} \prod_{\ell=0}^{L_f} \left[m_{f\ell}^{I(Z_j=i)} u_{f\ell}^{I(Z_j \neq i)} \right]^{I(\gamma_{ij}^f = \ell)}, \tag{2.1}$$

where $I(\cdot) = 1$ when its argument is true and $I(\cdot) = 0$ otherwise.

Sometimes, there are missing values in the linking variables. Although we do not consider this possibility in our simulations, we summarize how to handle missing values in the model, assuming ignorable missing data. With conditional independence and ignorability, we can marginalize over the missing comparison variables. The likelihood of the observed comparison data can be written as

$$\mathcal{L}(\mathbf{Z}|\boldsymbol{\Theta}, \boldsymbol{\gamma}^{\text{obs}}) = \prod_{f=1}^{F} \prod_{\ell=0}^{L_f} m_{f\ell}^{a_{f\ell}(\mathbf{Z})} u_{f\ell}^{b_{f\ell}(\mathbf{Z})}, \qquad \text{where} \tag{2.2}$$

$a_{f\ell}(\mathbf{Z}) = \sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j = i)$ and $b_{f\ell}(\mathbf{Z}) = \sum_{i,j} I_{\text{obs}}(\gamma_{ij}^f) I(\gamma_{ij}^f = \ell) I(Z_j \neq i)$. For a given $\mathbf{Z}$, these represent the number of matches and non-matches with observed disagreement level ℓ in field f. Here, $I_{\text{obs}}(\cdot) = 1$ when its argument is observed, and $I_{\text{obs}}(\cdot) = 0$ when its argument is missing.

To define the prior distributions for $\mathbf{m}$ and $\mathbf{u}$, for all fields f, let $\boldsymbol{\alpha}_f = (\alpha_{f0}, \ldots, \alpha_{fL_f})$ and $\boldsymbol{\beta}_f = (\beta_{f0}, \ldots, \beta_{fL_f})$. We assume that $\mathbf{m}_f \sim \text{Dirichlet}(\boldsymbol{\alpha}_f)$ and $\mathbf{u}_f \sim \text{Dirichlet}(\boldsymbol{\beta}_f)$, where $\boldsymbol{\alpha}_f$ and $\boldsymbol{\beta}_f$ are known parameters. In our simulation studies, we set all entries of $\boldsymbol{\alpha}_f$ and $\boldsymbol{\beta}_f$ equal to 1 for every field.

We use the prior distribution for $\mathbf{Z}$ from (Sadinle, 2017). For $j = 1, \ldots, n_2$, let the indicator variable $I(Z_j \leq n_1) \mid \pi \overset{iid}{\sim} \text{Bernoulli}(\pi)$, where π is the proportion of matches expected a priori. Let $\pi \sim \text{Beta}(\alpha_\pi, \beta_\pi)$, where the prior mean $\alpha_\pi/(\alpha_\pi + \beta_\pi)$ represents the expected percentage of overlap. Let $n_{12}(\mathbf{Z}) = \sum_j I(Z_j \leq n_1)$ be the number of matches according to $\mathbf{Z}$. The prior specification implies that $n_{12}(\mathbf{Z}) \sim \text{Beta-Binomial}(n_2, \alpha_\pi, \beta_\pi)$ after marginalizing over π. That is,

$$\mathbb{P}(n_{12}(\mathbf{Z})) = \binom{n_2}{n_{12}(\mathbf{Z})} \frac{\mathrm{B}(n_{12}(\mathbf{Z}) + \alpha_\pi, n_2 - n_{12}(\mathbf{Z}) + \beta_\pi)}{\mathrm{B}(\alpha_\pi, \beta_\pi)}. \tag{2.3}$$

We assume that, conditional on the value of $n_{12}(\mathbf{Z})$, all the possible bipartite matchings are equally likely a priori. There are $n_1!/(n_1 - n_{12}(\mathbf{Z}))!$ such bipartite matchings. Thus, the prior distribution for $\mathbf{Z}$ is

$$\begin{align} \mathbb{P}(\mathbf{Z}|\alpha_\pi, \beta_\pi) &= \mathbb{P}(n_{12}(\mathbf{Z})|\alpha_\pi, \beta_\pi)\mathbb{P}(\mathbf{Z}|n_{12}(\mathbf{Z}), \alpha_\pi, \beta_\pi) \tag{2.4} \\ &= \frac{(n_1 - n_{12}(\mathbf{Z}))!}{n_1!} \frac{\mathrm{B}(n_{12}(\mathbf{Z}) + \alpha_\pi, n_2 - n_{12}(\mathbf{Z}) + \beta_\pi)}{\mathrm{B}(\alpha_\pi, \beta_\pi)}. \tag{2.5} \end{align}$$

2.3 The Bayesian Hierarchical Model for Simultaneous Regression and Record Linkage

In this section, we present the Bayesian hierarchical model for regression and RL, and propose three Markov chain Monte Carlo (MCMC) algorithms for fitting the model in practice. Throughout, we assume the explanatory variables $\mathbf{X}$ are in $\mathbf{A}_1$, and the response variable $\mathbf{Y}$ is in $\mathbf{A}_2$.

2.3.1 Model specification

We assume the standard linear regression, $\mathbf{Y}|\mathbf{X}, \mathbf{V}, \mathbf{Z} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$. Here, $\mathbf{V}$ are linking variables used in the RL model but not in the regression model. Analysts can specify prior distributions on (β, σ^2) that represent their beliefs. A full specification of the joint distribution of $(\mathbf{Y}, \mathbf{X}|\mathbf{V})$ requires analysts to specify some marginal model for $\mathbf{X}$, written generically as $f(\mathbf{X}|\mathbf{V})$. In some contexts, however, it is not necessary to specify $f(\mathbf{X}|\mathbf{V})$, as we explain in section 2.3.2. Critically, this model assumes that the distribution of $(\mathbf{Y}, \mathbf{X}|\mathbf{V})$ is the same for matches and non-matches. Finally, for the RL component of the model, we model $\mathbf{Z}$ using the Bayesian FS approach in section 2.2.

For the simulation studies, we illustrate computations with the Bayesian hierarchical model using univariate Y and univariate X. Assume $X \sim N(\mu, \tau^2)$. As a result, in the simulations, the random variable (X, Y) follows a bivariate normal distribution with

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\Big(\begin{bmatrix} \mu \\ \beta_0 + \beta_1\mu \end{bmatrix}, \begin{bmatrix} \tau^2 & \beta_1\tau^2 \\ \beta_1\tau^2 & \sigma^2 + \beta_1^2\tau^2 \end{bmatrix} \Big).$$

We assume a normal-Gamma prior on the regression parameters. Letting $\phi = 1/\sigma^2$, we have $\phi \sim G(.5, .5)$ and $\beta|\phi \sim N(b_0, \Phi_0\phi^{-1})$ where $b_0 = [3, 1]^T$ and Φ_0 is a 2×2 identity matrix. When needed, we assume Jeffrey's prior $p(\mu, \tau^2) \propto 1/\tau^2$.

2.3.2 Estimation Strategies

Even in the relatively uncomplicated set-up of the simulation study, it is not possible to compute the posterior distribution of the model parameters in closed form. Therefore, we consider three general strategies for MCMC sampling in order to approximate the posterior distribution.

First, we propose an MCMC sampler that uses only the linked records when estimating the full conditional distribution of the regression parameters in each iteration of the sampler. This method generates potentially different samples of linked records at each iteration. A key advantage of this method is that it does not require imputation of $\mathbf{X}$; hence, analysts need not specify $f(\mathbf{X}|\mathbf{V})$. We call this the joint model without imputation, abbreviated as JM. Second, we propose an MCMC sampler that imputes the missing values of $\mathbf{X}$ for those records in $\mathbf{A}_2$ without a link, and updates the regression parameters in each iteration using the linked pairs as well the imputed data. We call this the joint model with imputation, abbreviated as JMI. Third, we propose an MCMC sampler that is similar to JMI but uses an extra step when imputing the missing values of $\mathbf{X}$. Specifically, at each iteration, we (i) sample values of the regression parameters from a conditional distribution based on only the linked records, (ii) use the sampled parameters to impute missing values in $\mathbf{X}$, and (iii) update regression coefficients based on linked as well as imputed pairs. By adding step (i), we aim to reduce potential effects of a feedback loop in which less accurate estimates of regression parameters result in less accurate estimates of the conditional distribution of $\mathbf{X}$, and so on through the MCMC iterations. We call this the joint model with imputation and reduced feedback (JMIF).

For JMI and JMIF, inferences are based on every entity in $\mathbf{A}_2$, whereas for JM inferences are based on the subsets of linked pairs, which can differ across MCMC iterations. Analysts should keep these differences in mind when selecting an algorithm that suits their goals.

2.3.3 Details of the MCMC Samplers

In this section, we present the mathematical details for implementing the three proposed MCMC samplers. Before doing so, we present an algorithm for a two-step approach, where we perform RL and then use the linked data for regression. The three proposed MCMC samplers for the Bayesian hierarchical model utilize parts of the algorithm for the two-step approach.

Two Step Approach (TS).

Given the parameter values at iteration t of the sampler, we need to sample new values $\mathbf{m}_f^{[t+1]} = (m_{f0}^{[t+1]}, \ldots, m_{fL_f}^{[t+1]})$ and $\mathbf{u}_f^{[t+1]} = (u_{f0}^{[t+1]}, \ldots, u_{fL_f}^{[t+1]})$, where $f = 1, \ldots, F$. We then sample a new value $\mathbf{Z}^{[t+1]} = (Z_1^{[t+1]}, \ldots, Z_{n_2}^{[t+1]})$. The steps are as follows.

T.1 For $f = 1, \ldots, F$, sample

$$\mathbf{m}_f^{[t+1]} | \boldsymbol{\gamma}^{obs}, \mathbf{Z}^{[t]} \sim \text{Dirichlet}(a_{f0}(\mathbf{Z}^{[t]}) + \alpha_{f0}, \ldots, a_{fL_f}(\mathbf{Z}^{[t]}) + \alpha_{fL_f}), \tag{2.6}$$

$$\mathbf{u}_f^{[t+1]} | \boldsymbol{\gamma}^{obs}, \mathbf{Z}^{[t]} \sim \text{Dirichlet}(b_{f0}(\mathbf{Z}^{[t]}) + \beta_{f0}, \ldots, b_{fL_f}(\mathbf{Z}^{[t]}) + \beta_{fL_f}). \tag{2.7}$$

Collect these new draws into $\mathbf{\Theta}^{[t+1]}$. Here, each

$$a_{fl}(\mathbf{Z}) = \sum_{i,j} I_{obs}(\boldsymbol{\gamma}^f_{ij}) I(\gamma^f_{ij} = l) I(Z_j = i), \tag{2.8}$$

$$b_{fl}(\mathbf{Z}) = \sum_{i,j} I_{obs}(\boldsymbol{\gamma}^f_{ij}) I(\gamma^f_{ij} = l) I(Z_j \neq i). \tag{2.9}$$

T.2 Sample the entries of $\mathbf{Z}^{[t+1]}$ sequentially. Having sampled the first $j-1$ entries of $\mathbf{Z}^{[t+1]}$, we define $\mathbf{Z}^{[t+(j-1)/n_2]}_{-j} = (Z^{[t+1]}_1, \ldots, Z^{[t+1]}_{j-1}, Z^{[t]}_{j+1}, \ldots, Z^{[t]}_{n_2})$. We sample a new label $Z^{[t+1]}_j$, with the probability of selecting label $q \in \{1, \ldots, n_1, n_1 + j\}$ given by $p_{qj}(\mathbf{Z}^{[t+(j-1)/n_2]}_{-j} \mid \mathbf{\Theta}^{[t+1]})$. This can be expressed for generic $\mathbf{Z}_{-j}$ and $\mathbf{\Theta}$ as

$$p_{qj}(\mathbf{Z}_{-j}|\mathbf{\Theta}) \propto \begin{cases} \exp[w_{qj}] I(Z_{j'} \neq q, \forall j' \neq j), \text{ if } q \leq n_1 \\ [n_1 - n_{12}(\mathbf{Z}_{-j})] \dfrac{n_2 - n_{12}(\mathbf{Z}_{-j}) - 1 + \beta_\pi}{n_{12}(\mathbf{Z}_{-j}) + \alpha_\pi}, \text{ if } q = n_1 + j; \end{cases} \tag{2.10}$$

where $w_{qj} = \log[\mathbb{P}(\boldsymbol{\gamma}^{obs}_{qj}|Z_j = q, \mathbf{m})/\mathbb{P}(\boldsymbol{\gamma}^{obs}_{qj}|Z_j \neq q, \mathbf{u})]$ is equivalently

$$w_{qj} = \sum_{f=1}^{F} I_{obs}(\boldsymbol{\gamma}^f_{qj}) \sum_{l=0}^{L_f} \log\left(\frac{m_{fl}}{u_{fl}}\right) I(\gamma^f_{qj} = l). \tag{2.11}$$

The normalizing constant for $p_{qj}(\mathbf{Z}_{-j}|\mathbf{\Theta})$ is

$$\prod_{i=1}^{n_1}\prod_{k=1}^{n_2}\prod_{f=1}^{F}\prod_{\ell=0}^{L_f} \left[m_{f\ell}^{I(Z_k=i)} u_{f\ell}^{I(Z_k\neq i)}\right]^{I(\gamma^f_{ij}=\ell)} u_{fl}^{I(Z_j\neq i)}$$
$$\times \frac{\big(n_1 - (n_{12}(\mathbf{Z}_{-j}) + 1)\big)!}{n_1!} \frac{(n_{12}(\mathbf{Z}_{-j}) + \alpha_\pi)!(n_2 - (n_{12}(\mathbf{Z}_{-j}) + 1) + \beta_\pi - 1)!}{(n_2 + \alpha_\pi + \beta_\pi - 1)} \tag{2.12}$$

where $k \neq j$.

T.3 We now add the regression parameters to the sampler. For any draw of $\mathbf{Z}^{[t+1]}$, we sample $\beta^{[t+1]}$ and $(\sigma^2)^{[t+1]} = \phi^{-1}$ from

$$\beta^{[t+1]}|\phi, \mathbf{Y}, \mathbf{Z}^{[t+1]} \sim N(b_n, (\phi\Phi_n)^{-1}) \tag{2.13}$$

$$\phi|\mathbf{Y}, \mathbf{Z}^{[t]} \sim G\Big(\tfrac{n+1}{2}, \tfrac{1}{2}(SSE + 1 + \hat{\beta}^T\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\hat{\beta} + b_0^T\phi_0 b_0 - b_n^T\Phi_n b_n)\Big) \tag{2.14}$$

where $SSE = \widetilde{\mathbf{Y}}^T[\mathbf{I} - \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T]\widetilde{\mathbf{Y}}$, $\Phi_n = \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} + \Phi_0$, $b_n = \Phi_n^{-1}(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\hat{\beta} + \phi_0 b_0)$, and $\hat{\beta} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{Y}}$. Here, $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ are the subsets of $\mathbf{X}$ and $\mathbf{Y}$ belonging to only the linked cases at iteration t.

Steps **T.1** and **T.2** are the same as those used in (Sadinle, 2017); we add **T.3** to sample the regression parameters. Alternatively, and equivalently, analysts can run **T.1** and **T.2** until MCMC convergence, then apply **T.3** to each of the resulting draws of $\mathbf{Z}^{[t]}$ to obtain the draws of the regression parameters.

Joint Method without Imputation (JM).

The sampler for the JM method uses **T.1**, but it departs from **T.2** and **T.3**. As we shall see, in JM we need the marginal density $f(\mathbf{Y})$. This can be approximated with a standard univariate density estimator. Alternatively, one can derive it from $f(\mathbf{Y}|\mathbf{X})$ and extra assumptions about $f(\mathbf{X})$, although these extra assumptions obviate one of the advantages of JM compared to JMI and JMIF. In the simulations, for convenience we use the fact that (Y, X) are bivariate normal when computing the marginal density of $\mathbf{Y}$, as evident in step **J.2**. Step **J.2** can be omitted when using means to compute $f(\mathbf{Y})$ that do not leverage a joint model for $(\mathbf{Y}, \mathbf{X})$.

J.1 Sample $\mathbf{m}_f^{[t+1]}$ and $\mathbf{u}_f^{[t+1]}$ using **T.1**.

J.2 Sample $\mu^{[t+1]}$ and $(\tau^2)^{[t+1]}$ using $1/\tau^2 \sim G((n-1)/2, \sum(X_i - \bar{X})^2)$ and $\mu|\tau^2 \sim N(\bar{X}, \tau^2)$. We use all of $\mathbf{X}$ in this step.

J.3 Given $\mathbf{Z}^{[t]}$, sample $\beta^{[t+1]}$ and $(\sigma^2)^{[t+1]}$ from (2.13) and (2.14).

J.4 Sample $\mathbf{Z}^{[t+1]}$ sequentially. Having sampled the first $j-1$ entries of $\mathbf{Z}^{[t+1]}$, we define $\mathbf{Z}_{-j}^{[t+(j-1)/n_2]} = (Z_1^{[t+1]}, \ldots, Z_{j-1}^{[t+1]}, Z_{j+1}^{[t]}, \ldots, Z_{n_2}^{[t]})$. Then we sample a new label $Z_j^{[t+1]}$ with probability $p_{qj}(\mathbf{Z}_{-j}^{[t+(j-1)/n_2]} \mid \mathbf{\Theta}^{[t+1]}, \mathbf{X}, \mathbf{Y})$ of selecting label $q \in \{1, \ldots, n_1, n_1 + j\}$. For generic $(\mathbf{Z}_{-j}, \mathbf{\Theta}, \mathbf{X}, \mathbf{Y})$, we have $f(\mathbf{Z}_{-j}|\mathbf{\Theta}, \mathbf{X}, \mathbf{Y}) \propto f(\mathbf{Y}, \mathbf{X}|\mathbf{\Theta}, \mathbf{Z}_{-j}) f(\mathbf{Z}_{-j}|\mathbf{\Theta})$. For $q \le n_1$, and using the definition for w_{qj} in (2.11), we thus have

$$
\begin{aligned}
p_{qj}(\mathbf{Z}_{-j}|\mathbf{\Theta}, \mathbf{X}, \mathbf{Y}) &\propto \exp[w_{qj}] I(Z_{j'} \neq q, \forall j' \neq j) \prod_{i \neq q, i \in \mathbf{A}_{12}} f(X_i, Y_i | \mathbf{Z}_{-j}) \\
&\quad \times \prod_{i \neq q, i \in \mathbf{A}_{1-}} f(X_i) \prod_{i \neq q, i \in \mathbf{A}_{2-}} f(Y_i) f(Y_j, X_q) && (2.15) \\
&\propto \exp[w_{qj}] I(Z_{j'} \neq q, \forall j' \neq j) \frac{f(Y_j, X_q)}{f(Y_j) f(X_q)} && (2.16) \\
&= \exp[w_{qj}] I(Z_{j'} \neq q, \forall j' \neq j) \frac{f(Y_j | X_q)}{f(Y_j)}. && (2.17)
\end{aligned}
$$

Here, $\mathbf{A}_{12}$ is the set of matched records, $\mathbf{A}_{1-}$ is the set of records in $\mathbf{A}_1$ without a match in $\mathbf{A}_2$, and $\mathbf{A}_{2-}$ is the set of records in $\mathbf{A}_2$ without a match in $\mathbf{A}_1$.

For $q = n_1 + j$, after some algebra to collect constants, we have

$$p_{qj}(\mathbf{Z}_{-j}|\mathbf{\Theta}, \mathbf{X}, \mathbf{Y}) \propto [n_1 - n_{12}(\mathbf{Z}_{-j})] \frac{n_2 - n_{12}(\mathbf{Z}_{-j}) - 1 + \beta_\pi}{n_{12}(\mathbf{Z}_{-j}) + \alpha_\pi}.$$

Joint Method with Imputation (JMI).

The sampler for JMI is similar to the sampler for JM, except we impute $\mathbf{X}$ for non-matches in $\mathbf{A}_2$. Thus, we require a model for $\mathbf{X}$, which we also use to compute $f(\mathbf{Y})$. In accordance with the simulation set-up, we present the sampler with $X \sim N(\mu, \tau^2)$.

I.1 Sample $\mathbf{m}_f^{[t+1]}$ and $\mathbf{u}_f^{[t+1]}$ using **J.1**.

I.2 Sample $\mu^{[t+1]}$ and $(\tau^2)^{[t+1]}$ using **J.2**.

I.3 Impute $\mathbf{X}_{mis}$ for those records in $\mathbf{A}_2$ without a matched X. For the sampler in the simulation study, the predictive distribution is

$$X_{mis} \sim N\Big(\mu + \frac{\beta_1 \tau^2}{\sigma^2 + \beta_1^2 \tau^2}(Y - \beta_0 - \beta_1 \mu), \tau^2 - \frac{\beta_1^2 \tau^4}{\sigma^2 + \beta_1^2 \tau^2}\Big). \tag{2.18}$$

In JMI, we use the values of $(\beta^{[t]}, (\sigma^2)^{[t]}, (\tau^2)^{[t+1]})$ in (2.18). Once we have $\mathbf{X}_{mis}^{[t+1]}$, in the full conditional distributions for $(\beta^{[t+1]}, (\sigma^2)^{[t+1]})$ we use both the matched and imputed data for all records in $\mathbf{A}_2$, with the priors in section 2.3.3. As a result, we draw $\beta^{[t+1]}$ and $(\sigma^2)^{[t+1]}$ based on (2.13) and (2.14), but let $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ include both the linked pairs and imputed pairs in $\mathbf{A}_2$.

I.4 Sample $\mathbf{Z}^{[t+1]}$ sequentially using **J.4**.

Joint Method with Imputation and Reduced Feedback (JMIF).

The sampler for JMIF is like the sampler for JMI, but we use a different predictive model for $\mathbf{X}_{mis}$. We again present the sampler with $X \sim N(\mu, \tau^2)$.

F.1 Sample $\mathbf{m}_f^{[t+1]}$ and $\mathbf{u}_f^{[t+1]}$ using **J.1**.

F.2 Sample $\mu^{[t+1]}$ and $(\tau^2)^{[t+1]}$ using **J.2**.

F.3 Given $\mathbf{Z}^{[t]}$, take a draw (β^*, σ^{2*}) from the full conditional distributions in (2.13) and (2.14), using only the linked cases at iteration t. We impute $\mathbf{X}_{mis}$ for those records in $\mathbf{A}_2$ without a matched X using (β^*, σ^{2*}). For the sampler in the simulation study, we use (2.18) with (β^*, σ^{2*}) and $(\tau^2)^{[t+1]}$. Once we have $\mathbf{X}_{mis}^{[t+1]}$, in the full conditional distributions for $(\beta^{[t+1]}, \sigma^{[t+1]})$ we use both the matched and imputed data for all records in $\mathbf{A}_2$, with the priors in section 2.3.3. As a result, we draw $\beta^{[t+1]}$ and $(\sigma^2)^{[t+1]}$ based on (2.13) and (2.14), but let $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ include both the linked pairs and imputed pairs in $\mathbf{A}_2$.

F.4 Sample $\mathbf{Z}^{[t+1]}$ sequentially using (**J.4**).

2.3.4 MCMC Starting Values

Sadinle (2017) starts the MCMC sampler by assuming none of the records in file $\mathbf{A}_2$ have a link in file $\mathbf{A}_1$. We do not recommend this starting point for the hierarchical model, as it is beneficial to specify an initial set of links to determine sensible starting values for the linear regression parameters. Instead, we employ a standard FS algorithm—implemented using the `RecordLinkage` package in `R`—to determine a set of links to use as a starting point.

2.4 Simulation Studies

We generate simulated data sets using the `RLdata10000` data set from the `RecordLinkage` package in `R`. The `RLdata10000` contains 10,000 records; 10% of these records are duplicates belonging to 1,000 individuals. The `RLdata10000` includes linking variables, which the developers of the data set have distorted to create uncertainty in the RL task. To create $\mathbf{A}_2$, we first randomly sample $n_{12} = 750$ individuals from the 1,000 individuals with duplicates. We then sample 250 individuals from the remaining 8,000 individuals in RLdata10000. This ensures that each record in $\mathbf{A}_2$ belongs to one and only one individual. To create $\mathbf{A}_1$, we first take the duplicates for the 750 records in $\mathbf{A}_2$; these are true matches. Next, we sample another 250 records from the individuals in RLdata10000 but not in $\mathbf{A}_2$. Thus, we have 750 records that are true links, and 250 records in each file that do not have a match in the other file. We repeat this process independently in each simulation run.

In both files, in each simulation run, we generate the response and explanatory variables, as none are available in the RLdata10000. For each sampled record i, we generate $x_i \sim N(0, 1)$ and $y_i \sim N(\beta_0+\beta_1 x_i, \sigma^2)$ in each simulation run. We set $\beta_0 = 3$ and $\sigma^2 = 1$. We consider $\beta_1 \in \{.4, .65, .9\}$, to study how the correlation between X and Y affects performance of the methods.

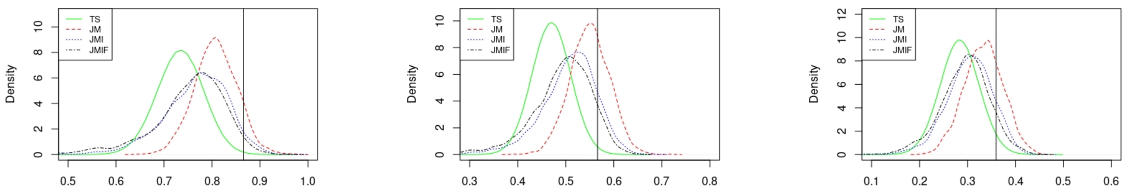
We use four linking variables: the first name and last name, and two constructed binary variables based on birth year and birth day. For the constructed variables, we create an indicator of whether the individual in the record was born before or after 1974, and another indicator of whether the individual in the record was born before or after the 16th day of the month.

To compare first and last name, we use the Levenshtein edit distance (LD), defined as the minimum number of insertions, deletions, or substitutions required to change one string into the other. We divide this distance by the length of the longest string to standardize it. The final measure is in the range of $[0, 1]$, where 0 represents total agreement and 1 total disagreement. Following (Sadinle, 2017), we categorize the LD into four levels of agreement. We set $f = 1$ and $\gamma_{ij}^f = 3$ when the first names for record i and j match perfectly ($LD = 0$); $\gamma_{ij}^f = 2$ when these names show mild disagreement ($0 < LD \leq .25$); $\gamma_{ij}^f = 1$ when these names show moderate disagreement ($.25 < LD \leq .5$); and, $\gamma_{ij}^f = 0$ when these names show extreme disagreement ($LD \geq .5$). The same is true for last names with $f = 2$. For the constructed binary variables based on birth day and year, we set $\gamma_{ij}^f = 1$ when the values for record i and j agree with each other, and $\gamma_{ij}^f = 0$ otherwise.

We create two scenarios to represent different strengths of the information available for linking. The strong linkage scenario uses all four linking variables, and the weak linkage scenario uses first and last name only.

2.4.1 Record Linkage Evaluation Metrics

Here, we review the definitions of the average numbers of correct links (CL), correct non-links (CNL), false negatives (FN), and false positives (FP). These allow one to calculate the false negative rate (FNR) and false discovery rate (FDR) (Steorts, 2015). For any MCMC iteration t, we define $\text{CL}^{[t]}$ as the number of record pairs with $Z_j \leq n_1$ and that are true links. We define $\text{CNL}^{[t]}$ as the number of record pairs with $Z_j > n_1$ that also are not true links. We define $\text{FN}^{[t]}$ as the number of record pairs that are true links but have $Z_j > n_1$. We define $\text{FP}^{[t]}$ as the number of record pairs that are not true links but have $Z_j \leq n_1$. In the simulations, the true number of true links is $\text{CL}^{[t]}+\text{FN}^{[t]} = 750$, and the estimated number of links is $\text{CL}^{[t]}+\text{FP}^{[t]}$. Thus, $\text{FNR}^{[t]} =$ is $\text{FN}^{[t]}/(\text{CL}^{[t]}+\text{FN}^{[t]})$. The $\text{FDR}^{[t]} = \text{FP}^{[t]}/(\text{CL}^{[t]}+\text{FP}^{[t]})$, where by convention we take $\text{FDR}^{[t]} = 0$ when both the numerator and denominator are 0. We report the FDR instead of the false positive rate (FPR), as an algorithm that does not link any records has a small FPR, but this does not mean that it is a good algorithm. Finally, for each metric, we compute the posterior means across all MCMC iterations, which we average across all simulations.



(a) $\beta_1 = 0.9$.

(b) $\beta_1 = 0.65$.

(c) $\beta_1 = 0.4$.

Figure 2.1: Posterior density of β_1 in one arbitrarily chosen data set for each value of β_1 under the strong linking information scenario. Posterior distribution for the two-step approach is in solid green, for JM is in dashed red, for JMI is in dotted blue, and for JMIF is in dotdash black. Vertical lines are estimates of β_1 when using all the correct links.

2.4.2 Results

Table 2.1 displays averages of the estimated regression coefficients over 100 simulation runs. Across all scenarios, on average the point estimates of β_1 from the Bayesian hierarchical model, regardless of the MCMC algorithm, are at least as close to the true β_1 as the point estimates from the two-step approach. The hierarchical model offers the largest improvements in accuracy over the two-step approach when the correlation between X and Y is strongest and the information in the linking variables is weakest. In this scenario, the model takes advantage of information in the relationship between the study variables that the two-step approach cannot. In contrast, when the correlation between X and Y is weakest and the information in the linking variables is strongest, there is little difference in the performances of the hierarchical model and two-step approach. These patterns are illustrated in Figure 2.1.

Generally, all the algorithms tend to underestimate β_1 in these simulations. It is practically impossible to identify all the true links. Therefore, the regression is estimated with some invalid pairs of (x_i, y_i). This attenuates the estimates of β_1.

The hierarchical model tends to overestimate β_0 slightly. The difference is most noticeable when the correlation between X and Y is strong. Generally, on average the two-step approach offers more accurate estimates of β_0, although the differences are practically irrelevant.

Among the hierarchical models, JM outperforms JMI and JMIF, with JMIF slightly better than JMI. This is because inaccuracies in the estimated distribution of $\mathbf{X}_{mis}$ in JMI and JMIF are propagated to the estimated distributions of (β_0, β_1). To illustrate this, suppose in a particular MCMC iteration the value of β_1 is somewhat attenuated, which in turn leads to inaccuracy in the parameters of the imputation model for $X|Y$. As a result, the imputed values of $\mathbf{X}_{mis}$ are not samples from an accurate representation of $f(X|Y)$. Thus, the full conditional distribution of (β_0, β_1) is estimated from completed-data that do not follow the relationship between X and Y. The inaccurate samples of (β_0, β_1) then create inaccurate imputations, and the cycle continues. In contrast, in any iteration, JM samples coefficients using only the records deemed to be links (in that iteration), thereby reducing the effects of feedback from imprecise imputations. This also explains why JMIF yields slightly more accurate estimates of β_1 than JMI when the correlation between X and Y is strong.

Table 2.2 displays averages across the 100 simulation runs of six standard metrics for the quality of the record linkages. These include the average numbers of CL, CNL, FN, and FP, as well as the FNR and FDR. These are formalized in section 2.4.1. The results in Table 2.2 indicate that the hierarchical model offers improved linkage quality over the two-step approach, regardless of the estimation algorithm. In particular, the hierarchical model tends to have smaller FP and larger CNL than the two-step approach. The difference in CNL is most apparent when the information in the linking variables is weak and when the correlation between X and Y is strong. The hierarchical model tends to have higher CL than the two-step approach, but the

Table 2.1: Summary of simulation results for regression coefficients. Results based on 100 runs per scenario. The true $\beta_0 = 3$ in all scenarios. For each reported average, the Monte Carlo standard errors are smaller than .01. "Strong" refers to scenarios where we use all four comparison fields, and "Weak" refers to scenarios where we use only two comparison fields.

	Results for β_1				Results for β_0			
	TS	JM	JMIF	JMI	TS	JM	JMIF	JMI
Strong								
$\beta_1 = .90$	.73	.85	.80	.79	3.01	3.05	3.04	3.04
$\beta_1 = .65$	.52	.60	.56	.55	3.00	3.02	3.01	3.01
$\beta_1 = .40$	.32	.36	.33	.32	2.99	3.00	3.00	3.00
Weak								
$\beta_1 = .90$	.60	.82	.78	.76	3.00	3.05	3.04	3.03
$\beta_1 = .65$	.42	.57	.53	.53	3.00	3.03	3.02	3.02
$\beta_1 = .40$	.27	.34	.32	.32	3.00	3.01	3.01	3.01

difference is practically important only when the linkage information is weak and the correlation is relatively strong ($\beta_1 = .9, \beta_1 = .65$). Overall, the hierarchical model has lower FDR compared to the two step approach.

As an additional simulation, we examine the performance of the hierarchical model in terms of linkage quality when we use a mis-specified regression. The true data generating model is $\log(\mathbf{Y})|\mathbf{X}, \mathbf{V}, \mathbf{Z} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$, but we incorrectly assume $\mathbf{Y}|\mathbf{X}, \mathbf{V}, \mathbf{Z} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ in the hierarchical model. Table 2.3 summarizes the measures of linkage quality when the linkage variables have weak information. Even though the regression component of the hierarchical model is mis-specified, the hierarchical model still identifies more correct non-matches than the two-step approach identifies, although the difference is less obvious than when we use the correctly specified regression. We see a similar trend when the information in the linking variables is strong, albeit with smaller differences between the two-step approach and the hierarchical model.

Table 2.2: Summary of linkage quality across 100 simulation runs. Averages in first four columns have standard errors less than 3. Averages in the last two columns have Monte Carlo standard errors less than .002.

		CL	CNL	FN	FP	FNR	FDeR
Strong							
$\beta_1 = .90$	JM	702	152	47	128	.06	.15
	JMIF	702	155	48	125	.06	.15
	JMI	702	155	48	125	.06	.15
	TS	697	123	53	167	.07	.19
$\beta_1 = .65$	JM	702	139	48	144	.06	.17
	JMIF	701	143	49	140	.06	.16
	JMI	701	143	49	141	.06	.17
	TS	698	119	52	172	.07	.20
$\beta_1 = .40$	JM	698	131	52	158	.07	.18
	JMIF	698	133	52	155	.07	.18
	JMI	698	133	52	155	.07	.18
	TS	697	121	53	170	.07	.20
Weak							
$\beta_1 = .90$	JM	636	114	114	223	.15	.26
	JMIF	636	116	114	222	.15	.26
	JMI	636	115	114	223	.15	.26
	TS	620	53	130	315	.17	.34
$\beta_1 = .65$	JM	632	93	118	256	.16	.29
	JMIF	631	94	119	255	.16	.29
	JMI	631	92	119	256	.16	.30
	TS	621	51	129	317	.17	.34
$\beta_1 = .40$	JM	625	69	125	291	.17	.32
	JMIF	624	69	126	291	.17	.32
	JMI	625	69	125	291	.17	.32
	TS	620	50	130	319	.17	.34

2.5 Discussion

The simulation results suggest that the Bayesian hierarchical model for simultaneous regression and RL can offer more accurate coefficient estimates than the two-step approach in which one first performs RL then runs regression on linked data. The hierarchical model is most effective when the correlation between the response and

Table 2.3: Results for simulation with mis-specified regression component in the hierarchical model. Entries summarize the linkage quality across 100 simulation runs. Averages in first four columns have standard errors less than 3. Averages in the last two columns have Monte Carlo standard errors less than .002.

		CL	CNL	FN	FP	FNR	FDeR
$\beta_1 = .90$	JM	625	69	125	292	.17	.32
	JMIF	624	70	126	291	.17	.32
	JMI	624	69	126	292	.17	.32
	TS	619	51	131	318	.17	.34
$\beta_1 = .65$	JM	626	62	124	299	.17	.32
	JMIF	626	62	124	299	.17	.32
	JMI	626	62	124	299	.17	.32
	TS	622	49	128	319	.17	.34
$\beta_1 = .40$	JM	623	56	127	309	.17	.33
	JMIF	623	56	127	309	.17	.33
	JMI	623	57	127	309	.17	.33
	TS	622	50	128	318	.17	.34

explanatory variable is strong. The hierarchical model also can improve linkage quality, in particular by identifying more non-links. This is especially the case when the information in the linking variables is not strong. In all scenarios, the relationship between the response and explanatory variable complements the information from the comparison vectors, which helps us declare record pairs more accurately.

As with any simulation study, we investigate only a limited set of scenarios. Our simulations have 75% of the individuals in the target file as true matches. Future studies could test whether the hierarchical model continues to offer gains with lower overlap rates, as well as different values of other simulation parameters. We used a correctly specified linear regression with only one predictor. We note that the hierarchical model without imputation for missing $\mathbf{X}$ extends readily to multivariate $\mathbf{X}$. When outcomes are binary, analysts can use probit regression in the hierarchical model. The model also can be modified for scenarios where one links to a smaller file containing explanatory variables. In this case, we use the marginal distribution of $\mathbf{X}$

and conditional distribution of $\mathbf{X}|\mathbf{Y}$ rather than those for $\mathbf{Y}$ and $\mathbf{Y}|\mathbf{X}$ in (2.17).

3

Leveraging auxiliary information on marginal distributions for handling item nonresponse and accounting for survey weights

3.1 Introduction

As discussed in Chapter 1, the MD-AM framework offers a method for using auxiliary margins to handle nonignorable missing data. To date, the MD-AM framework has been developed mainly in the context of simple random samples. An exception is Akande and Reiter (ming), who describe an approach to using the MD-AM framework in stratified simple randome samples. In this chapter, I investigate the usefulness of the approach described in (Akande and Reiter, ming) for more general unequal probability designs. Specifically, I examine the case of MD-AM imputation for item nonresponse. I discuss the case of simultaneous item and unit nonresponse in Chapter 4.

This chapter is organized as follows. In section 3.2, we set up notation and review the approach described in (Akande and Reiter, ming) for stratified simple random samples. In section 3.3, we carry out simulation studies and discuss the performance of the MD-AM framework under Poisson sampling survey designs. We also discuss implementation of the MD-AM framework with parameter tuning.

3.2 Background

We introduce relevant notation in section 3.2.1. We describe the model of (Akande and Reiter, ming) in section 3.2.2.

3.2.1 Notation

Let $\mathcal{D}$ comprise data from the survey of $i = 1, \ldots, n$ individuals, and let $\mathcal{A}$ comprise data from the auxiliary database. Let $X = (X_1, \ldots, X_k)$ represent the p variables in both $\mathcal{A}$ and $\mathcal{D}$, where each $X_k = (X_{1k}, \ldots, X_{nk})^T$ for $k = 1, \ldots, p$. Let $Y = (Y_1, \ldots, Y_q)$ represent the q variables in $\mathcal{D}$ but not in $\mathcal{A}$, where each $Y_k = (Y_{1k}, \ldots, Y_{nk})^T$ for $k = 1, \ldots, q$. We disregard variables in $\mathcal{A}$ but not in $\mathcal{D}$ as they are not of primary interest. We assume that $\mathcal{A}$ only contains sets of marginal distributions for variables in X based on some external database.

We introduce indicators for item nonresponse. For each $k = 1, \ldots, p$, let $R^x_{ik} = 1$ if individual i would not respond to the question on X_k in survey $\mathcal{D}$ and $R^x_{ik} = 0$ otherwise. Similarly, for each $k = 1, \ldots, q$, let $R^y_{ik} = 1$ if individual i would not respond to the question on Y_k in survey $\mathcal{D}$ and $R^y_{ik} = 0$ otherwise. Let $R^x = (R^x_1, \ldots, R^x_p)$ and $R^y = (R^y_1, \ldots, R^y_q)$ be the vectors of item nonresponse indicators for variables in X and Y respectively, where each $R^x_k = (R^x_{ik}, \ldots, R^x_{nk})^T$ and $R^y_k = (R^y_{ik}, \ldots, R^y_{nk})^T$. In this chapter, we assume all units respond to the survey, or that analysis can be based on the unit respondents possibly with adjusted design weights for unit nonrespondents. The discussion of handling item and unit nonresponse together is saved for Chapter 4.

3.2.2 Applying The MD-AM Framework In Stratified Simple Random Samples

In this section, we review the algorithm of (Akande and Reiter, ming) using a simple scenario as an example: there are two binary variables, one fully observed and the other suffering from item nonresponse. Following step 1 in MD-AM framework, we have:

$$(X_1, Y_1) \sim f(X_1, Y_1|\Theta), \tag{3.1}$$

$$\mathbb{P}(R_1^x = 1|X_1, Y_1) = h(\gamma_0 + \gamma_1 Y_1). \tag{3.2}$$

Based on the auxiliary margin on X, we can estimate one more parameter and turn (3.2) into

$$\mathbb{P}(R_1^x = 1|X_1, Y_1) = h(\gamma_0 + \gamma_1 Y_1 + \gamma_2 X_1). \tag{3.3}$$

This is basically the model described in Chapter 1; however, now we wish to incorporate the stratified sampling design.

Let N denote the population size from which the n survey units in $\mathcal{D}$ are sampled. Let $W = (w_1, \ldots, w_n)$, where each w_i is the weight for the ith unit in $\mathcal{D}$. Here, we let $w_i = 1/\pi_i$, where π_i is the probability of inclusion of the ith unit. We do not observe values of X_i or Y_i for nonsampled units in the population.

In practice, the most common auxiliary information available is the population total (or mean) of some of the variables. For example, for the total of some binary variable X, we have

$$T_X = \sum_{i=1}^{N} X_{i1} = N \times \mathbb{P}(X_1 = 1), \tag{3.4}$$

where $\mathbb{P}(X_1 = 1)$ is the true auxiliary marginal probability. In stratified samples, analysts also might know the population total in each stratum. For the remainder of this dissertation, I assume only knowledge of T_X. Akande and Reiter (ming) show that knowing the overall population margin or the population margins in each stratum give qualitatively similar results via simulation studies.

The Horvitz-Thompson estimator (Horvitz and Thompson, 1952) is a commonly-used design-unbiased estimator of T_X, defined as follows:

$$\hat{T}_X = \sum_{i \in \mathcal{D}} \frac{X_{i1}}{\pi_i} = \sum_{i \in \mathcal{D}} w_i X_{i1}. \tag{3.5}$$

Finite population central limit theorems ensure that when the sample size is large enough, $\hat{T}_X$ is approximately normally distributed around T_X, with a variance V_X. Thus, for fully observed data, we have

$$\sum_{i \in \mathcal{D}} w_i X_{i1} \sim N(T_X, V_X). \tag{3.6}$$

Analysts can estimate V_X using design-based principles (Fuller, 2009).

As noted in Akande and Reiter (ming), $\hat{T}_X$ cannot be computed directly when there is nonresponse in the data. However, this unobserved value of $\hat{T}_X$ should be distributed around T_X as governed by (3.6). Thus, when we impute the missing values for X_1, it is sensible to ensure further that any completed data set produces a reasonable $\hat{T}_X$ based on (3.6). This is the intuition behind the algorithm in Akande and Reiter (ming). We implement this logic as follows. For all $i \in \mathcal{D}$, let $X^\star_{i1} = X_{i1}$ when $R^x_{i1} = 0$, and let $X^\star_{i1}$ be an imputed value when $R^x_{i1} = 1$. Our imputation should fulfill this probabilistic constraint,

$$\sum_{i \in \mathcal{D}} w_i X^\star_{i1} \sim N(T_X, V_X). \tag{3.7}$$

We will refer to the model defined by (3.1), (3.2), and (3.7) as MD-AM rejection sampler, for reasons to be clear later.

We can estimate the parameters of the model, and hence generate completed datasets, using a MCMC sampler. At each MCMC iteration t, let the current draw of each X_{i1} be $X_{i1}^{(t)}$, and let $\hat{T}_X^{(t)} = \sum_{i \in \mathcal{D}} w_i X_{i1}^{(t)}$. The steps of MD-AM rejection sampler algorithm are as follows.

R1. For all $i \in \mathcal{D}$, i.e., $i = 1, \ldots, n$, set $X_{i1}^{\star} = X_{i1}$ when $R_{i1}^{x} = 0$. When $R_{i1}^{x} = 1$, generate a candidate $X_{i1}^{\star}$ for the missing X_{i1} by drawing from the posterior predictive distribution implied by (3.1) and (3.3).

R2. Let $\hat{T}_X^{\star} = \sum_{i \in \mathcal{D}} w_i X_{i1}^{\star}$. Calculate the acceptance ratio,

$$p = \frac{N(\hat{T}_X^{\star}; T_X, V_X)}{N(\hat{T}_X^{(t)}; T_X, V_X)}. \tag{3.8}$$

R3. Draw a value $u \sim Unif(0,1)$. If $u \leq p$, accept the proposed candidate $(X_{i1}^{\star}, \ldots, X_{in}^{\star})$ and set $X_{i1}^{(t+1)} = X_{i1}^{\star}$. Otherwise, reject the proposed candidate and set $X_{i1}^{(t+1)} = X_{i1}^{(t)}$ for $i = 1, \ldots, n$.

Akande and Reiter (ming) leave the issue of specifying V_X largely unaddressed. I discuss this specification in section 3.3.3.

3.3 Extending the MD-AM Framework to Account for Survey Weights

Conceptually, the MD-AM rejection sampler could be applied to other survey designs, including those with arbitrary weights. In this section, we examine its performance in simulation studies with Poisson sampling. In section 3.3.1, we present simulations where survey weights are generated from the inverse of a Beta distribution. In section 3.3.2 we use survey weights from the CPS 2012 data, and also compare the performance of MD-AM and MAR models. In section 3.3.3, we discuss how to select an estimate of V_X when X suffers from item nonresponse.

3.3.1 Poisson Sampling With Inverse Beta Distributed Weights

We set the population size $N = 50000$ across all simulation scenarios. We first generate simulated data sets where inclusion probabilities (inverse of survey weights) are independent and identically distributed draws from a Beta distribution. Then survey weights follow an inverted Beta distribution, which was proposed by Potter and Zheng (2015) for its right-tailed skewness. Dropping the subscripts on X_1 and Y_1 for simplicity, the generative model is as follows:

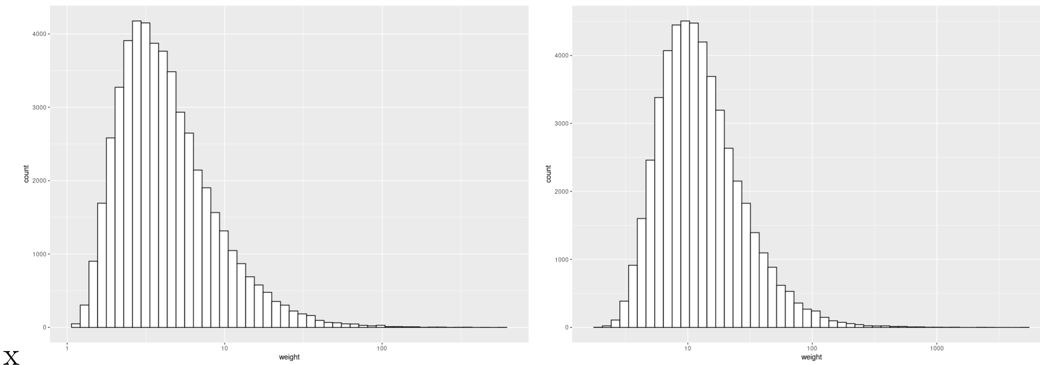
$$\frac{1}{W} \sim Beta(a, b) \tag{3.9}$$

$$Y \sim bern(\pi_y), \pi_y = \Phi(\beta_0 + \beta_1 \frac{1}{W}) \tag{3.10}$$

$$X|Y \sim bern(\pi_X), \pi_X = \Phi(\alpha_0 + \alpha_1 Y) \tag{3.11}$$

$$R^x|X, Y \sim bern(\pi_R), \pi_R = \Phi(\gamma_0 + \gamma_1 Y + \gamma_2 X). \tag{3.12}$$

We set $\alpha_0 = 0.3, \beta_0 = 0.5, \beta_1 = -0.15, \gamma_0 = -1.0, \gamma_1 = -0.5$. These values are picked so that the nonresponse rates across all the simulation settings are between 10% and 35%. For the remaining two parameters (α_1, γ_2), we have two values for each of them as we expect that these two variables may strongly affect the performance of the MD-AM rejection sampler. α_1 represents the relationship between X and Y. We pick α_1 to be either -0.5 or -2.0, the former for weak relationship and the latter for strong. γ_2 represents how strongly the nonresponse bias in X depends on X itself. We pick 0.4 and 1.5 to represent weak and strong relationships, respectively. The distribution of weights may also relate to the performance of MD-AM rejection sampler as the presence of more extreme weights is likely to result in lower acceptance rate. Hence, we generate inclusion probabilities from both $Beta(2, 5)$ and $Beta(2, 18)$. The distributions of weights generated are displayed in Figure 3.1 on $\log 10$ scale.



X

Figure 3.1: Distribution of weights generated from inverse of $Beta(2, 5)$ (left) and $Beta(2, 18)$ (right) on log 10 scale.

Weights generated from the inverse of $Beta(2, 18)$ on the right of Figure 3.1 have more extreme large values.

In this section and the next, we assume the true population V_X is known, which is obtained from the Horvitz-Thompson estimate as defined in (3.14) of sampled data before introducing item nonresponse at each simulation run. There are $2^3 = 8$ simulation scenarios in total generated by two levels for (α_1, γ_2) and the weight distributions. In each simulation run, we draw samples from the target population using Poisson sampling with inclusion probability $\pi_i = 1/w_i$ for each unit i. Then, we use the MD-AM rejection sampler to run 10000 MCMC iterations, discard the first 5000 runs as burn-in, and keep 5000 posterior samples. We create $L = 50$ multiple imputation data sets, $\boldsymbol{Z} = (\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(50)})$, from every 100 posterior samples. For each completed data set $\boldsymbol{Z}^{(\ell)}$, we compute the design-based estimates of $(T_X, \alpha_0, \alpha_1, \beta_0, \beta_1)$ along with the corresponding standard errors, using the survey-weighted generalized linear models option in the R package, "survey" (Lumley, 2021). We also compute estimates of $(\gamma_0, \gamma_1, \gamma_2)$, which do not depend on the design weights, along with the corresponding standard errors, using the generalized linear models option in the R package, "stats". We repeat this process 1000 times for each simulated population.

The population total of X is known from $T_X = \sum_{i=1}^{N} X_i$. An unbiased estimate of the theoretical variance of the Horvitz-Thompson estimator $\hat{T}_X$ is

$$\hat{Var}(\hat{T}_X) = \sum_{i \in \mathcal{D}} \frac{(\frac{x_i}{\pi_i})^2 \pi_i (1 - \pi_i)}{\pi_i} + 2 \sum_{i<j \in \mathcal{D}} \frac{\frac{x_i x_j}{\pi_i \pi_j}(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \tag{3.13}$$

where $\mathcal{D}$ includes all elements in sample. $\pi_i = \mathbb{P}(I_i = 1)$ are first order inclusion probabilities, and $\pi_{ij} = \mathbb{P}(I_i = 1, I_j = 1), i \neq j$ are second order inclusion probabilities. In the case of Poisson sampling, $\pi_{ij} = \pi_i \pi_j$ for all elements in the sample, so the second term in (3.13) becomes 0, and hence the Horvitz-Thompson estimator of the theoretical variance of $\hat{T}_X$ is

$$\hat{Var}(\hat{T}_x) = \sum_{i \in \mathcal{S}} \frac{(\frac{x_i}{\pi_i})^2 \pi_i (1 - \pi_i)}{\pi_i} = \sum_{i \in \mathcal{S}} \left(\frac{x_i}{\pi_i}\right)^2 (1 - \pi_i). \tag{3.14}$$

Simulation results are shown in Table 3.1 and Table 3.2. In those tables, "truth" includes the average of model parameters across 1000 sampled datasets without any item nonresponse, as well as the true population value of T_X. CI coverage is the 95% confidence interval coverage rate across 1000 samples. "Pre-SD" stands for pre-missing standard deviation, calculated based on the 1000 sampled full data sets, that is, before introduction of missing data. "SD" is the standard deviation of the 1000 multiple imputation point estimates for each estimand, whereas "avg. est. SD" (short for average estimated standard deviation) is calculated from taking the square root of the average of the 1000 multiple imputation variances calculated based on the multiple imputation combining rules (Rubin, 1987).

Point estimates of β_0 and β_1 are very accurate since there is no missingness in Y and W. Point estimates of α_0 and α_1 are very close to truth except for Table 3.2(d), where the estimate of α_1 is slightly biased. Under this particular setting, the nonresponse bias in X is large and the correlation between X and Y is strong.

Table 3.1: Simulation results across 1000 sampled data sets from simulated population using Poisson sampling. For each sampled data sets with missingness, 50 multiple imputed data sets were generated. Weights are generated from the inverse of $Beta(2,5)$ distribution.

	Truth	Estimate	CI Coverage %	Pre-SD	SD	Avg. Est. SD
(a) $\boldsymbol{\alpha_1 = -0.5, \gamma_2 = 0.4}$. Avg. nonresponse rate = 14.1%						
α_0	.289	.285	99.6	.021	.026	.036
α_1	-.490	.485	97.4	.026	.029	.033
β_0	.501	.501	97.4	.035	.035	.039
β_1	-.167	-.167	97.9	.089	.089	.100
γ_0	-1.004	-1.003	94.7	.022	.090	.093
γ_1	-.496	-.499	96.9	.023	.029	.033
γ_2	.403	.396	95.4	.021	.125	.126
T_X	24159	24135	100	346.5	58.9	844.0
(b) $\boldsymbol{\alpha_1 = -2.0, \gamma_2 = 0.4}$. Avg. nonresponse rate = 12.0%						
α_0	.289	.282	99.9	.022	.025	.039
α_1	-1.982	-1.972	99.2	.035	.037	.044
β_0	.509	.509	95.1	.040	.040	.040
β_1	-.180	-.180	95.7	.099	.099	.102
γ_0	-1.004	-.996	95.5	.025	.101	.109
γ_1	-.503	-.512	95.4	.029	.092	.099
γ_2	.377	.358	96.1	.028	.141	.152
T_X	11458	11433	100	230.7	37.0	686.9
(c) $\boldsymbol{\alpha_1 = -0.5, \gamma_2 = 1.5}$. Avg. nonresponse rate = 32.8%						
α_0	.299	.301	99.3	.023	.025	.035
α_1	-.502	-.510	97.8	.027	.028	.034
β_0	.491	.491	96.9	.039	.039	.040
β_1	-.135	-.135	96.3	.096	.096	.103
γ_0	-.987	-.991	97.0	.020	.080	.087
γ_1	-.512	-.500	97.1	.020	.023	.028
γ_2	1.504	1.498	96.3	.020	.091	.099
T_X	24178	24129	100	341.9	56.3	855.8
(d) $\boldsymbol{\alpha_1 = -2.0, \gamma_2 = 1.5}$. Avg. nonresponse rate = 21.8%						
α_0	.290	.286	99.8	.022	.022	.035
α_1	-2.010	-2.008	98.4	.035	.037	.047
β_0	.520	.520	96.5	.037	.037	.040
β_1	-.230	-.230	97.4	.094	.094	.102
γ_0	-.980	-.972	95.8	.025	.083	.087
γ_1	-.506	-.515	96.3	.028	.066	.072
γ_2	1.495	1.484	95.3	.028	.096	.100
T_X	11423	11392	100	235.3	34.8	685.5

Table 3.2: Simulation results across 1000 sampled data sets from simulated population using Poisson sampling. For each sampled data sets with missingness, 50 multiple imputed data sets were generated. Weights are generated from the inverse of $Beta(2, 18)$ distribution.

	Truth	Estimate	CI Coverage %	Pre-SD	SD	Avg. Est. SD
(a) $\boldsymbol{\alpha_1 = -0.5, \gamma_2 = 0.4}$. Avg. nonresponse rate = 14.0%						
α_0	.311	.298	99.4	.042	.051	.067
α_1	-.507	-.495	96.8	.051	.056	.061
β_0	.506	.506	95.2	.061	.061	.060
β_1	-.292	-.292	95.7	.401	.401	.298
γ_0	-.998	-1.000	96.7	.044	.180	.438
γ_1	-.512	-.522	96.8	.043	.054	.059
γ_2	.412	.393	96.8	.044	.245	.469
T_X	24258	24146	100	653.1	140.1	1605.0
(b) $\boldsymbol{\alpha_1 = -2.0, \gamma_2 = 0.4}$. Avg. nonresponse rate = 12.1%						
α_0	.300	.283	99.5	.043	.046	.070
α_1	-2.011	-1.990	97.4	.065	.068	.078
β_0	.521	.521	96.4	.059	.059	.060
β_1	-.353	-.353	96.0	.392	.392	.401
γ_0	-1.012	-1.002	96.1	.052	.840	.200
γ_1	-.477	-.490	96.5	.059	.167	.180
γ_2	.407	.365	97.3	.060	.263	.645
T_X	11194	11093	100	436.6	77.5	1289.6
(c) $\boldsymbol{\alpha_1 = -0.5, \gamma_2 = 1.5}$. Avg. nonresponse rate = 32.7%						
α_0	.311	.300	99.3	.041	.047	.064
α_1	-.511	-.508	97.3	.051	.054	.061
β_0	.508	.508	95.7	.060	.060	.060
β_1	-.170	-.170	96.3	.391	.391	.400
γ_0	-.968	-.958	95.3	.043	.148	.164
γ_1	-.521	-.524	97.0	.040	.044	.048
γ_2	1.498	1.481	95.8	.040	.170	.185
T_X	24159	24000	100	638.4	129.7	1614.8
(d) $\boldsymbol{\alpha_1 = -2.0, \gamma_2 = 1.5}$. Avg. nonresponse rate = 21.2%						
α_0	.298	.272	99.4	.044	.042	.065
α_1	-1.986	-1.950	94.9	.067	.070	.081
β_0	.489	.489	95.0	.060	.060	.060
β_1	-.029	-.029	95.0	.396	.396	.401
γ_0	-1.000	-.972	95.6	.051	.152	.161
γ_1	-.482	-.517	95.6	.055	.121	.128
γ_2	1.490	1.457	96.2	.058	.177	.187
T_X	11254	11130	100	459.6	75.1	1300.6

Since estimates of α_1 tend to be accurate in Table 3.1(d), more extreme weights sampled from the inverse of $Beta(2, 18)$ may cause the bias in estimating (α_0, α_1) in Table 3.2(d). Very large weights can amplify even small biases resulting from imputation modeling. Point estimates of $\boldsymbol{\gamma}$ are also quite accurate when weights are sampled from the inverse of $Beta(2, 5)$ in Table 3.1. When weights are sampled from the inverse of $Beta(2, 18)$, we observe that γ_2 is slightly underestimated in Table 3.2(b) and (d). These two scenarios correspond to strong correlation between X and Y. Since we set $\beta_1 = -0.15$, units with extremely large weights are more likely to have Y equal to 1 than those with smaller weights based on our generative model. Therefore, the proposals of $X^\star$ generated in the MD-AM rejection sampler are more likely to be influenced by Y and imputed as 0 more often than warranted, resulting in the underestimation of γ_2. The estimates of T_X fall in close range of the truth for all scenarios, signifying that the MD-AM rejection sampler successfully matches the estimated total of X to the auxiliary margin in the multiple imputed data sets.

The CI coverage rates for the estimates across the eight simulation settings are generally greater than 95%, meaning that we have at least nominal coverage. For T_X, the CI coverage rate is always 100%. This happens due to the accurate point estimates and overestimate of the true standard deviation. The true standard deviation for T_X is much smaller than the average estimated standard deviation calculated from Rubin's multiple imputation combining rule as we have extra information from the auxiliary margin to help with generating multiple imputed data sets. In fact, the true standard deviation for T_X obtained via MD-AM rejection sampler is even smaller than the standard deviation of original datasets before the nonresponse is introduced. The true standard deviations are roughly $1/\sqrt{m}$ times the pre-missing standard deviations because we average over m completed data sets to obtain the point estimate of T_X. The true standard deviation and the average estimated standard deviation of γ_0 and

γ_2 are larger than those of other parameters, especially in Table 3.2. γ_2 is the hardest parameter to estimate since the only information we have related to it is through the auxiliary margin. Uncertainty in estimating γ_2 affects the estimation of γ_0.

The acceptance rates of the proposed candidate in the R.3 step of the MD-AM rejection sampler across the eight simulation scenarios ranged between 82.8% and 89.3%. We do not observe significantly difference in acceptance rates across eight simulation scenarios. In general, with acceptance rate above 80%, researchers tend to worry that the MCMC sampler is not efficient in the sense that it makes too many small accepted steps, resulting in strong correlations between samples and low effective sample size. As commented in (Akande and Reiter, ming), we do not concern about high acceptance rate as (3.7) used in R.3 step is considered as a constraint rather than a target distribution. If desired, correlation between posterior samples can be reduced via thinning the chain. For example, we take every 100 posterior samples in generating multiple imputed data sets.

3.3.2 Poisson Sampling From the CPS Weights

In this section, we carry out additional simulation studies using a set of genuine weights from the 2012 Current Population Survey. More specifically, we extract all non-zero household weights from the CPS 2012 data. There are $N = 133427$ non-zero weights in total. We generate a population as in (3.10) - (3.12) and follow the procedures in 3.3.1 to generate and analyze simulated data sets. We set $(\alpha_0, \alpha_1) = (0.3, -0.5)$, $(\beta_0, \beta_1) = (0.5, -0.05)$, and $(\gamma_0, \gamma_1, \gamma_2) = (-0.25, 0.1, -1.5)$.

Without using auxiliary information, a natural model for R^x follows MAR mechanism. In order to see the benefits from incorporating auxiliary information, we generate multiple imputations from this model as well. To do so, we sample from posterior predictive distribution implied by (3.1) and (3.2) in step 1 of MD-AM rejection sampler, and skip the accept/rejection step. Without using auxiliary information, the posterior predictive distribution we sampled from is derived by MAR missing mechanism for X.

The results are shown in Table 3.3. Since the estimates of $\boldsymbol{\beta}$ are always very close to truth as we have observed in section 3.3.1, we do not show them from now on. All the 95% confidence interval coverage rates equal 0 under MAR except the intervals for α_1, where 79.2% coverage is much lower than the 96.2% coverage under MD-AM. Under MAR, α_0 is highly overestimated, whereas under MD-AM the estimate is close to truth. The propose-accept/reject steps R.2 and R.3 differentiates reasonable proposals of imputation from unreasonable ones. It leads to more accurate estimates in α_0 and not surprisingly in T_X as well. γ_2, which is not identifiable without the margin, is the hardest parameter to estimate because the only information available is through the margin. Under the MD-AM framework, the confidence interval coverage of γ_2 is 93.9%, the lowest among all the intervals but still reasonably close to the nominal rate of 95%.

Next we examine the true standard deviations and average estimated standard deviations. The true SD of $\hat{T}_X$ is much smaller than the average estimated SD under MD-AM. Reiter (2008) found that if there is more information used in estimating the imputation models than used for analysis, the variance estimator obtained from multiple imputation combination rules has positive bias. In our case, the multiple imputation combination rule is not aware of the population margin. By leveraging auxiliary information on T_X, we can reduce the variation on T_X in addition to getting a more accurate point estimate. Unfortunately, we were not able to derive a single unbiased variance estimator that accounts for the MD-AM imputation process. We discuss this further in Chapter 4.

Table 3.3: Simulation results across 1000 sampled data sets from simulated population with the CPS weights. For each sampled data sets with missingness, 50 multiple imputed data sets were generated. Algorithm with (MDAM) and without (MAR) margins are presented side-by-side.

	Estimate			CI Coverage %		SD			Avg. Est. SD	
	Truth	MDAM	MAR	MDAM	MAR	Pre-SD	MDAM	MAR	MDAM	MAR
α_0	.292	.296	.582	99.3	0	.030	.030	.034	.038	.031
α_1	-.493	-.495	-.453	96.2	79.2	.036	.036	.042	.038	.037
γ_0	-.267	-.271	-.928	94.3	0	.027	.047	.028	.048	.032
γ_1	.126	.124	.343	96.4	0	.029	.034	.026	.039	.030
γ_2	-1.504	-1.536	-	93.9	-	.033	.025	-	.027	-
T_X	64110	64210	80475	100	0	1180.4	362.5	1389.0	1610.1	1325.8

3.3.3 Role And Selection Of V_X

In practice, since there are missing values in the data set, the Horvitz-Thompson estimate $\hat{V}_X = \hat{Var}(\hat{T}_X)$ is not readily available. In fact, V_X can be used as a tuning parameter, which reflects how closely we want $\hat{T}_X$ to match T_X. If we adjust V_X to be relatively small, the accept/reject step will accept only imputations that result in $\hat{T}_X$ very close to the truth T_X. On the other hand, if we let $V_X \to \infty$, it's equivalent to imposing no probability constraint on the weighted total of imputed X at all.

In this section, we use simulations to examine the implications of specifying V_X. Since we cannot calculate the Horvitz-Thompson estimate $\hat{V}_X$ directly due to the presence of item nonresponse, we need to find a substitute. To do so, we first generate a completed data set by imputing X_i for units with $R_i^x = 1$ under the assumption of MAR mechanism in R^x. Then, we use (3.14) to calculate $\hat{V}_X$ on this completed data set. We use $\widetilde{V}_X$ to denote this estimate of V_X. One could repeat this step several times and average the completed data estimates of variance, which might provide some stability in the method.

To illustrate the impacts of using different variances, we repeat the simulation with parameters $(\alpha_0, \alpha_1) = (0.3, 0.5)$, $(\gamma_0, \gamma_1, \gamma_2) = (-0.25, 0.1, -1.5)$, and $(\beta_0, \beta_1) = (0.5, -0.15)$. We run 1000 simulations using three variances. These include $\hat{V}_X^L = \tilde{V}_X$, $\hat{V}_X^M = \tilde{V}_X/100$, and $\hat{V}^S = \tilde{V}_X/10000$. Table 3.4 summarizes features of the MCMC samplers for each V_X. The acceptance ratio from largest V_X to smallest V_X are 0.904, 0.393, and 0.054 respectively, demonstrating that when V_X is small, more proposals are rejected. With $\hat{V}_X^M$, the acceptance rate is the closest to the optimal acceptance rate. Hence, we observe largest effective sample size for all parameters in this case.

Table 3.5 displays repeated sampling properties of the MD-AM model under the three values of V_X. Comparing results from $\hat{V}_X^L$ to $\hat{V}_X^M$, the estimate of T_X becomes more accurate with smaller true standard deviation and average estimated standard deviation. Comparing results from $\hat{V}_X^M$ to $\hat{V}_X^S$, we do not see similar improvement in terms of point estimate and true standard deviation. Of course, the improvements could be more substantial in other contexts, particularly when $\tilde{V}$ is large. Judging by T_X alone, $\hat{V}_X^M$ seems to be the best choice. However, the estimates of model parameters lead to different conclusion. The three values of V_X give similar performance for estimating $\boldsymbol{\alpha}$. The difference lies in estimating $\boldsymbol{\gamma}$, especially γ_2. The 95% confidence interval coverages of γ_2 are low and the average estimated standard deviations are large with $\hat{V}_X^M$ and $\hat{V}_X^S$. To explain this phenomenon, we also report the median of the estimated standard deviation of γ_2 in Table 3.5. The medians are much smaller than means with $\hat{V}_X^M$ and $\hat{V}_X^S$, signifying that the distribution of posterior samples of γ_2 has outliers.

Since $\hat{T}_X$ for the pre-missing data in each sample does not match the auxiliary margin exactly, if we set V_X to be relatively small and force $\hat{T}_X$ in each imputed data set to be as close as possible to T_X, we can distort the relationships among variables, including missing data indicators. To illustrate this point, we examine the seven extreme cases across the 1000 simulations when the estimated standard deviation of γ_2 is greater than 5 when using $\hat{V}_X^M$. Their estimates of γ_2 are all more negative than the truth. Their weighted sum of X without introducing item nonresponse is at least 66806, more than three standard deviations away from from 64110. Therefore, in these cases, the MD-AM rejection sampler favors imputations of missing X with many more zeros than in other simulation runs, and subsequently increasing the estimated standard deviation of γ_2.

We suggest using $\hat{V}_X^L = \widetilde{V}_X$ in the MD-AM rejection sampler. Even though $\hat{V}_X^M$ yields largest effective sample size during MCMC, we focus more on generating multiple imputed data sets that agree with auxiliary information while preserving the

Table 3.4: Effective sample size for each paramter of 5000 MCMC iterations after burn-in with three levels of V_X used in MD-AM rejection sampler.

	$\hat{V}_X^L$	$\hat{V}_X^M$	$\hat{V}_X^S$
α_0	222	1946	1693
α_1	1940	1876	574
γ_0	107	1050	598
γ_1	362	1008	512
γ_2	36	224	118
Accept rate	.904	.393	.054

joint distribution of variables. The autocorrelation between posterior samples can be reduced by thinning, as we do when generating multiple imputed data sets.

Table 3.5: Summary of results across 1000 simulation runs with item nonresponse alone. Within each run, 50 multiple imputed data sets were generated. $\hat{V}_X^L$ means we use $\widetilde{V}_X$ as V_X. $\hat{V}_X^M$ means we use $\widetilde{V}_X \times 10^{-2}$. $\hat{V}_X^S$ means we use $\widetilde{V}_X \times 10^{-4}$. Numbers in parenthesis are the median of estimated standard deviations of γ_2.

	Estimate				CI Coverage %			SD			Avg. Est. SD		
	Truth	$\hat{V}_X^L$	$\hat{V}_X^M$	$\hat{V}_X^S$	$\hat{V}_X^L$	$\hat{V}_X^M$	$\hat{V}_X^S$	$\hat{V}_X^L$	$\hat{V}_X^M$	$\hat{V}_X^S$	$\hat{V}_X^L$	$\hat{V}_X^M$	$\hat{V}_X^S$
α_0	.292	.296	.293	.292	99.3	95.5	96.6	.030	.031	.030	.038	.031	.031
α_1	-.493	-.495	-.493	-.495	96.2	95.6	96.6	.036	.037	.035	.038	.038	.038
γ_0	-.267	-.271	-.266	-.266	94.3	82.9	82.1	.047	.047	.048	.048	.033	.033
γ_1	.125	.124	.122	.122	96.4	92.2	92.5	.034	.037	.037	.039	.036	.036
γ_2	-1.503	-1.536	-1.529	-1.562	93.9	39.6	35.1	.252	.304	.470	.656	1.179	2.208
											(.205)	(.051)	(.47)
T_X	64110	64210	64113	64115	100	100	100	362.5	33.1	63.0	1615	1189.7	1184.4

3.4 Conclusion

In this chapter, we assess the performance of MD-AM rejection sampler proposed in (Akande and Reiter, ming) in complex survey designs through simulation studies of Poisson sampling. We generate simulated populations with weights sampled from the inverse of a Beta distribution and from the CPS 2012 data. As with any simulation study, we can only investigate a limited number of scenarios. However, within the scenarios we studied, the MD-AM rejection sampler is able to generate multiple imputed data sets which yield accurate estimates of joint distribution of study variables and missing data indicators, and also match the auxiliary margin, when data suffer from item nonresponse.

4

Leveraging auxiliary information on marginal distributions for handling item and unit nonresponse and accounting for survey weights

4.1 Introduction

Many surveys suffer not only from item nonresponse as discussed in chapter 3, but also from unit nonresponse, i.e., when the unit is missing all values. Typical practice in such cases is to assume the unit nonresponse is ignorable, which can be problematic. In this chapter, I describe how to use the MD-AM framework to handle both unit and item nonresponse simultaneously in the presence of survey weights. I do so assuming that survey weights are available for the unit respondents but are not provided for unit nonrespondents, which is typical of public use survey data files. I also discuss approaches for when the respondents' survey weights are adjusted for unit nonresponse, which often is the case in practice.

The chapter is organized as follows. In section 4.2, I show how to use the MD-AM framework to specify models when the data suffer from both item and unit nonresponse. To illustrate ideas, I follow the approach in Akande et al. (2021) and use a two binary variable example. I also describe the MCMC algorithm for fitting the model. The MCMC sampler generates multiple imputations of the missing values as byproducts, which are then used for analysis. In section 4.3, I present simulation studies and discuss the results. In section 4.4, I apply the methods on data on voting from the 2018 Current Population Survey (CPS).

4.2 MD-AM Framework Incorporating Unit Nonresponse

Let U_i denote the unit nonresponse indicator for individual $i \in \{1, \ldots, n\}$. When $U_i = 1$, we say individual i suffers from unit nonresponse. More specifically, all the entries for individual i are missing. On the other hand, $U_i = 0$ means that we at least observe some information of individual i. Let $U = (U_1, U_2, \ldots, U_n)$. Following the MD-AM framework, I discuss an innovative model specification for a simple example with two binary variables in section 4.2.1 and how to deal with unknown weights for unit nonrespondents in section 4.2.2. I introduce an MCMC algorithm that I call the intercept matching algorithm in section 4.2.3.

4.2.1 Model Specification

I have introduced notation in chapter 3 that I continue to use in this chapter. Recall that $\mathcal{D}$ comprises data from the survey of $i = 1, \ldots, n$ individuals, and $\mathcal{A}$ comprises data from the auxiliary database. Let $X = (X_1, \ldots, X_k)$ represent the p variables in both $\mathcal{A}$ and $\mathcal{D}$, where each $X_k = (X_{1k}, \ldots, X_{nk})^T$ for $k = 1, \ldots, p$. Let $Y = (Y_1, \ldots, Y_q)$ represent the q variables in $\mathcal{D}$ but not in $\mathcal{A}$, where each $Y_k = (Y_{1k}, \ldots, Y_{nk})^T$ for $k = 1, \ldots, q$. For each $k = 1, \ldots, p$, let $R^x_{ik} = 1$ if individual i would not respond to the question on X_k in survey $\mathcal{D}$ and $R^x_{ik} = 0$ otherwise. Similarly, for each $k = 1, \ldots, q$, let $R^y_{ik} = 1$ if individual i would not respond to the question on Y_k in survey $\mathcal{D}$ and $R^y_{ik} = 0$ otherwise. Let $R^x = (R^x_1, \ldots, R^x_p)$ and $R^y = (R^y_1, \ldots, R^y_q)$ be the vectors of item nonresponse indicators for variables in

Table 4.1: Two binary variables X_1 and X_2: X_2 is fully observed, X_1 suffers from item nonresponse, and the data contain unit nonrespondents. "✓" represents observed, "?" represents missing.

X_1	X_2	R_1^x	U	
✓	✓	0	0	
?	✓	1		Observed Data
?	?	?	1	
✓	?	?	?	Auxiliary Margin
?	✓	?	?	

X and Y respectively, where each $R_k^x = (R_{ik}^x, \ldots, R_{nk}^x)^T$ and $R_k^y = (R_{ik}^y, \ldots, R_{nk}^y)^T$. Finally, let $\mathcal{R} = (R^x, R^y)$ comprise all the indicators of item nonresponse.

To illustrate the MD-AM framework for unit and item nonresponse, suppose we have two binary variables X_1 and X_2, both contained in $\mathcal{D}$. We have auxiliary information for X_1 and X_2 from $\mathcal{A}$. Also, suppose X_1 suffers from item nonresponse but X_2 is fully observed, so that R_1^x is the fully observed vector of item nonresponse indicators for X_1. There is no need to include a model for R_2^x since X_2 is fully observed. For units with $U_i = 1$, we observe neither X_{1i}, nor X_{2i}, nor R_{1i}^x. Note that U is fully observed. The observed data for this scenario together with auxiliary information available on X_1 and X_2 are presented in Table 4.1. We use "✓" to represent observed data, and "?" to represent missing data.

We can write the joint distribution of (X_1, X_2, R_1^x, U) as follows:

$$\begin{aligned}\mathbb{P}(X_1 = x, X_2 = y, R_1^x = r, U = u) &= \mathbb{P}(R_1^x = r | X_1 = x, X_2 = y, U = u) \\ &\quad \times \mathbb{P}(X_1 = x | X_2 = y, U = u) \\ &\quad \times \mathbb{P}(X_2 = y | U = u)\mathbb{P}(U = u) \qquad (4.1)\end{aligned}$$

for $x, y, r, u \in \{0, 1\}$. Let $\theta_{xyu} = \mathbb{P}(R_1^x = 0 | X_1 = x, X_2 = y, U = u)$, $\pi_{yu} = \mathbb{P}(X_1 = 0 | X_2 = y, U = u)$, $q_u = \mathbb{P}(X_2 = 0 | U = u)$ and $p = \mathbb{P}(U = 0)$. The joint distribution can be fully parameterized with fifteen parameters plus the constraint that probabilities sum to one. From the observed data alone, we can estimate six parameters, namely $p, q_0, \theta_{000}, \theta_{010}, \theta_{110}, \theta_{100}$.

Auxiliary information about the marginal distribution of X_1 puts an additional constraint via

$$\begin{aligned}
&\mathbb{P}(X_1 = 1) - \sum_{y=0}^{1} \mathbb{P}(X_1 = 1, X_2 = y, R_1^x = 0, U = 0) \\
&= \sum_{y=0}^{1} \mathbb{P}(X_1 = 1, X_2 = y, R_1^x = 1, U = 1) + \sum_{y=0}^{1} \mathbb{P}(X_1 = 1, X_2 = y, R_1^x = 0, U = 1) \\
&+ \sum_{y=0}^{1} \mathbb{P}(X_1 = 1, X_2 = y, R_1^x = 1, U = 0) \\
&= (1-p)\Big(q_1(1-\pi_{01})(1-\theta_{101}) \\
&+ (1-q_1)(1-\pi_{11})(1-\theta_{111})\Big) + (1-p)\Big(q_1\pi_{01}\theta_{101} + (1-q_1)(1-\pi_{11})\theta_{111}\Big) \\
&+ p\Big(q_0(1-\pi_{00})(1-\theta_{100}) + (1-q_0)(1-\pi_{10})(1-\theta_{110})\Big).
\end{aligned} \tag{4.2}$$

The left hand side of (4.2) can be estimated from the auxiliary margin and the observed data. Thus, (4.2) gives an additional linear constraint on what values are possible for the remaining parameters, allowing us to estimate one more parameter. Therefore, we can estimate seven out of fifteen parameters in total up to this point. We have one more constraint from the auxiliary information about the distribution of X_2,

$$\begin{aligned}
\mathbb{P}(X_2 = 1) &= 1 - \sum_{u=0}^{1} \mathbb{P}(X_2 = 0 | U = u)\mathbb{P}(U = u) \\
&= 1 - \Big(pq_0 + (1-p)q_1\Big).
\end{aligned} \tag{4.3}$$

This constraint allows to us to estimate one more parameter, increasing the total number of identifiable parameters up to eight. These derivations show that the margins provide extra information about the joint distribution that can be used to identify additional parameters in model specifications. This begs the question, what are some useful model specifications? Sadinle and Reiter (2019) show that certain sequential factorizations of conditional models are identifiable with auxiliary margins.

Akande et al. (2021) illustrate how to use the MD-AM framework to specify models with a selection model factorization, in which they have a model for $\mathcal{D}$, then models for $(U|\mathcal{D})$ and $(\mathcal{R}|U, \mathcal{D})$. Akande and Reiter (ming) propose extending this approach to stratified samples using a rejection sampling approach. In initial explorations with this model specification when seeking to account for survey weights and unit and item nonresponse simultaneously, I find that for many types of marginal information and missingness mechanisms for $(U|\mathcal{D})$ and $(\mathcal{R}|U, \mathcal{D})$, the acceptance rate of rejection sampling method is low. This potential difficulty in sampling motivates us to come up with another parameterization of the joint distribution of (X_1, X_2, R_2^x, U). Namely, I use a pattern mixture model for unit nonresponse and a selection model for item nonresponse. More specifically, I propose a model for U, then models for $(\mathcal{D}|U)$ and $(\mathcal{R}|U, \mathcal{D})$. I now illustrate this model specification in the context of the MD-AM framework.

Following step 1 in the MD-AM framework, we need to specify an identifiable model based on the observed data alone, not exceeding the maximum number of parameters identifiable from the observed data. In this case, it should be a model with six parameters. One natural choice is an ICIN (Sadinle and Reiter, 2017) model for item nonresponse in X_1 and an MCAR model for unit nonresponse:

$$\mathbb{P}(U = 1) = g(\nu_0) \tag{4.4}$$

$$f(X_1, X_2|U) \sim f(X_1, X_2|\Theta) \tag{4.5}$$

$$\mathbb{P}(R_1^x = 1|X_1, X_2, U) = h(\gamma_0 + \gamma_1 X_2). \tag{4.6}$$

Here, ν_0 is identifiable since we know the marginal distribution of U from the observed data. Assuming $\mathcal{R}$ is conditionally independent of U given $\mathcal{D}$, γ_0 and γ_1 are identifiable since the marginal distribution of R_1^x and the joint distribution between X_2 and R_1^x can be estimated from observed data alone, as shown in Table 4.1. Θ is also identifiable given the ICIN specification in (4.6).

As an example of specific models, we can use a sequence of logistic models to specify the joint distribution of (X_1, X_2), More specifically, we write (4.5) as follows:

$$X_2|U \sim bern(\pi_{x_2}),\ logit(\pi_{x_2}) = \alpha_0 \tag{4.7}$$

$$X_1|X_2, U \sim bern(\pi_{x_1}),\ logit(\pi_{x_1}) = \beta_0 + \beta_1 X_2, \tag{4.8}$$

in order to better illustrate how wo add terms related to auxiliary margins.

In step 2, we add one term related to X_1 and one related to X_2 respectively to our model specification. Since X_2 already exists in (4.6), we add U in (4.7):

$$X_2|U \sim bern(\pi_{x_2}),\ logit(\pi_{x_2}) = \alpha_0 + \alpha_1 U. \tag{4.9}$$

(4.9) implies that the distribution of X_2 differs for unit nonrespondents and unit respondents, and α_1 controls the strength of nonresponse bias. More specifically, for unit respondents, $U = 0$ and $logit(\pi_{x_2}) = \alpha_0$. For unit nonrespondents, $U = 1$ and

$logit(\pi_{x_2}) = \alpha_0 + \alpha_1$. As a result, with larger $|\alpha_1|$, the distribution of X_2 will be less homogeneous across unit nonrespondents and unit respondents.

Depending on how we choose to use the auxiliary margin on X_1, our final model can be either of two distributions. The first lets X_2 depend on U and specifies a nonignorable missing mechanism for R_1^x:

$$\mathbb{P}(U = 1) = g(\nu_0) \tag{4.10}$$

$$X_2|U \sim bern(\pi_{x_2}),\ logit(\pi_{x_2}) = \alpha_0 + \alpha_1 U \tag{4.11}$$

$$X_1|X_2, U \sim bern(\pi_{x_1}),\ logit(\pi_{x_1}) = \beta_0 + \beta_1 X_2 \tag{4.12}$$

$$\mathbb{P}(R_1^x = 1|X_1, X_2, U) = h(\gamma_0 + \gamma_1 X_2 + \gamma_2 X_1). \tag{4.13}$$

Thus, item nonresponse follows an AN model, which inherits the benefits of AN models described in Chapter 1. The second allows both X_1 and X_2 to depend on U, but allows item nonresponse for X_2 to be an ICIN model. We have

$$\mathbb{P}(U = 1) = g(\nu_0) \tag{4.14}$$

$$X_2|U \sim bern(\pi_{x_2}),\ logit(\pi_{x_2}) = \alpha_0 + \alpha_1 U \tag{4.15}$$

$$X_1|X_2, U \sim bern(\pi_{x_1}),\ logit(\pi_{x_1}) = \beta_0 + \beta_1 X_2 + \beta_2 U \tag{4.16}$$

$$\mathbb{P}(R_1^x = 1|X_1, X_2, U) = h(\gamma_0 + \gamma_1 X_2). \tag{4.17}$$

Since U is a binary variable being either 0 or 1, we can view (4.15) and (4.16) as models with different intercepts for unit respondents and nonrespondents, but the slope is preserved.

Choosing between the model specification of (4.10) - (4.13) and (4.14) - (4.17) is subjective to the rates of missing values and prior beliefs of missingness. If the unit nonresponse rate is higher than the item nonresponse rate or the primary concern is about unit nonresponse, we may prefer to use the auxiliary margins to help model unit nonresponse as in (4.14) - (4.17). Similarly, if we believe an additive nonignorable model describes item nonresponse, we would pick (4.10) - (4.13). Since in reality, we commonly have some observed variables that may potentially characterize the difference between item respondents and nonrespondents, but no information on unit nonrespondents except auxiliary margins, analysts may prefer (4.14) - (4.17) over (4.10) - (4.13) as a default, and hence I focus on this model specification for the rest of this chapter. Researchers can employ both of the model specifications and compare their results as a sensitivity analysis if no strong preference is present for either of them.

In the specifications in (4.10) - (4.13) and (4.14) - (4.17), we did not include other response indicators in the models for U and R_1^x. Researchers tend to assume nonresponse indicators depend on variables in the data set, rather than other nonresponse indicators (Sadinle and Reiter, 2019). Indeed, we will assume nonresponse indicators are conditionally independent given other variables. In our current scenario, it means that we assume U and R_1^x are conditionally independent. This is a reasonable assumption to make since we do not observe R_1^x for unit non-respondents. Also, there is rarely auxiliary information available on nonresponse indicators, R_1^x in this case, to help us specify a relationship between them. If X_2 also suffers from item nonresponse, analysts can assume that R_1^x and R_2^x are conditionally independent. As we can estimate eight parameters in total with auxiliary margins on both X_1 and X_2, we require additivity in X_1, X_2 and U, meaning that interactions are not allowed in (4.11) and (4.12) as well as in (4.15) and (4.16).

To summarize, I use a pattern mixture model parameterization for unit nonresponse and a selection model parameterization for item nonresponse. This model specification scheme can be readily extended to more variables. If we have more than two variables in $\mathcal{D}$, we can write down a model with more identifiable parameters in step 1. In step 2, we add terms related to the variables in both $\mathcal{A}$ and $\mathcal{D}$. Again we will have multiple plausible model specifications. We can choose to use all the auxiliary margins to help with unit nonresponse by adding an U term in modeling variables with auxiliary margins when we believe potential bias due to unit nonresponse is a bigger concern than potential bias due to item nonresponse. We can also choose to make some of the item nonresponse indicators follow additive nonignorable models by letting them depend on their corresponding survey variables.

4.2.2 Weights For Unit Nonrespondents

Recall that we denote the inclusion probability of unit i as π_i. The design weight w_i^d for unit i is the inverse of its inclusion probability, i.e., $w_i^d = 1/\pi_i$. Typically, agencies deem it unnecessary and inconvenient to publish unit nonrespondents' design weights. As for unit respondents, some agencies provide the design weights, while others only publish design weights adjusted for nonresponse. We refer to these as adjusted weights and denote as w_i^a henceforth. For convenience in notation, when using the weights in estimates, we set $w_i^d = w_i^a = 0$ whenever $U_i = 1$.

The MCMC sampler we will introduce in section 4.2.3 requires the availability of weights for all units. If the design weights for unit respondents are provided, we only need to create weights for unit nonrespondents. There are multiple ways to achieve this goal. For example, one could take a random sample from the existing weights, or set unknown weights all equal to the average of observed weights. We generate weights for unit nonrespondents so that summation of all the weights equals the population size N. Since nothing has been observed for unit nonrespondents, we treat them equally and let them have the same weights. Therefore, when design weights are available, the weights that we will use for analysis are as follows:

$$w_i = \begin{cases} w_i^d, & \text{if } U_i = 0 \\ \frac{N - \sum w_i^d}{\sum U_i}, & \text{if } U_i = 1. \end{cases} \tag{4.18}$$

It is a common practice for agencies to adjust design weights for unit nonresponse so that the respondents can weight up to the full population (Valliant et al., 2013). Weight adjustment methods include but are not limited to weighting class adjustment, propensity score adjustment, poststratification and raking adjustments (Valliant et al., 2013; Lohr, 2010). If we keep using 4.18 to generate weights for unit nonrespondents, we may end up with very small or even negative weights because the sum of adjusted weights $\sum w_i^a$ already might be close to or even exceed N. Therefore, we need another weight adjustment scheme to avoid this undesirable outcome. When only the adjusted weights w_i^a are published for unit respondents, we need to down-weight them to attempt to get back to the design weights. In some instances, agencies may release enough information to enable analysts to do so. However, many times agencies do not release sufficient information about the weight adjustments to allow analysts to reverse-engineer this process to get back the exact design weights.

In this dissertation, I employ an ad hoc and simple downweighting technique for practical use. The weights I generate for analysis are as follows:

$$w_i^* = \begin{cases} w_i^a \times (1 - \frac{\sum U_i}{n}), & \text{if } U_i = 0 \\ \frac{\sum w_i^a}{n}, & \text{if } U_i = 1, \end{cases} \tag{4.19}$$

where n is the sample size. We down-weight the design weights for unit respondents by the unit response rate, and assign the remaining weight evenly to unit nonrespondents. After this re-assignment step, $\sum w_i^* = \sum w_i^a$ remains the same as before, a reasonable estimate of the population size. Moreover, unit nonrespondents are no longer being represented by up-weighting the respondents.

We could also create weights for unit nonrespondents and let $\sum w_i^* = N$ via setting $w_i^* = \frac{N-\sum w_i^a \times (1-\frac{\sum U_i}{n})}{\sum U_i} = \frac{\sum w_i^a}{n} + \frac{N-\sum w_i^a}{\sum U_i}$. Setting w_i^* so that $\sum w_i^* = N$ or $\sum w_i^* = \sum w_i^a$ does not make much difference in our simulation study in section 4.3. Since the true population size N may not be available in all the studies, but an estimate of N can be obtained by $\sum w_i^a$, in this thesis, when w_i^d are unavailable I create weights according to (4.19).

4.2.3 MCMC Sampler

We present the MCMC sampler under the model specification from (4.14) - (4.17), where unit nonresponse is nonignorable and item nonresponse ICIN. We also use logistic regression models for $g(\cdot)$ and $h(\cdot)$. Hence, (4.14) and (4.17) become

$$\mathbb{P}(U=1) = \pi_U,\ logit(\pi_U) = \nu_0 \tag{4.20}$$

$$\mathbb{P}(R_1^x = 1|X_1, X_2, U) = \pi_R,\ logit(\pi_R) = \gamma_0 + \gamma_1 X_2. \tag{4.21}$$

We use a normal approximation for the distribution of the estimated coefficients of the logistic regression whenever the outcome variable is binary to simplify the computation process (Raghunathan et al., 2001). Drawing from the exact posterior distribution can be achieved via data augmentation using Pólya-Gamma latent variables (Polson et al., 2013).

Let n_A denote the number of unit respondents and n_B the number of unit nonrespondents. Assume auxiliary information on X_1 and X_2 are population totals, where $T_{X1} = \sum_{i=1}^N X_{i1}$ and $T_{X2} = \sum_{i=1}^N X_{i2}$. Based on the central limit theorem, an unbiased estimate of T_{X1} follows approximately a normal distribution with mean T_{X1}

and variance V_{X1} when the sample size is large enough, i.e., $\hat{T}_{X1} \sim N(T_{X1}, V_{X1})$. Similarly, $\hat{T}_{X2} \sim N(T_{X2}, V_{X2})$. Since V_{X1}, V_{X2} are generally not known, we employ a similar technique as in Chapter 3 to specify values for them after conducting the weight adjustment described in section 4.2.2. Ignoring unit nonrespondents, an Horvitz-Thompson estimate of V_{X2} can be readily calculated. To calculate Horvitz-Thompson estimate of V_{X1}, we generate a completed data set by imputing X_{i1} for units with $R^x_{i1} = 1$ under the assumption of MAR mechanism in R^x_1. Then, we can calculate Horvitz-Thompson estimate of V_{X1} on this completed data set. We also get starting values for the MCMC algorithm via imputing X_{i1} for units with $R^x_{i1} = 1$ under the assumption of MAR mechanism in R^x_1 for $U_i = 0$. We denote estimates for V_{X1} and V_{X2} as $\hat{V}_{X1}$ and $\hat{V}_{X2}$. Given the parameter values and X_1, X_2 at iteration t, we use the following sampler when design weights for unit respondents are published. When those weights are not available, we substitute w^*_i for w_i.

IM1. Draw a value $\hat{T}^{(t+1)}_{X2} \sim N(T_{X2}, \hat{V}_{X2})$.

IM2. Calculate the number of times $X^{(t+1)}_{i2}$ should be imputed as 1 when $U_i = 1$ so that the weighted sum of $X^{(t+1)}_2$ is as close to $\hat{T}^{(t+1)}_{X2}$ as possible and let this number be n_{2B}. More specifically,

$$n_{2B} = \lfloor \frac{\hat{T}^{(t+1)}_{X2} - \sum_i^n w_i X^{(t+1)}_{i2} I(U_i = 0)}{\sum_i^n w_i I(U_i = 1)/n_B} \rfloor. \tag{4.22}$$

IM3. Let $\hat{\alpha}_0$ and $V_{\hat{\alpha}_0}$ denote the maximum likelihood estimate and the variance of α_0 based on $U_i = 0$ observations. Sample $\alpha^{(t+1)}_0$ from its approximate posterior distribution $N(\hat{\alpha}_0, V_{\hat{\alpha}_0})$.

IM4. Calculate the proportion of unit nonrespondents that should be imputed as 1 for X_2. Denote it as $\hat{p_2} = \frac{n_{2B}}{n_B}$. Then, calculate $\alpha^{(t+1)}_1$ so that $\mathbb{E}[\alpha^{(t+1)}_0 + \alpha^{(t+1)}_1] =$

$logit(\hat{p}_2)$.

IM5. Draw imputations of X_2 for unit nonrespondents: $X_2|U = 1 \sim bern\Big(1/\big(1 + \exp(-\alpha_0^{(t+1)} - \alpha_1^{(t+1)})\big)\Big)$.

IM6. Draw a value $\hat{T}_{X1}^{(t+1)} \sim N(T_{X1}, \hat{V}_{X1})$.

IM7. Calculate the number of times $X_{i1}^{(t+1)}$ should be imputed as 1 when $U_i = 1$ so that the weighted sum of $X_1^{(t+1)}$ is as close to $\hat{T}_{X1}^{(t+1)}$ as possible and let this number be n_{1B}. More specifically,

$$n_{1B} = \lfloor \frac{\hat{T}_{X1}^{(t+1)} - \sum_i^n w_i X_{i1}^{(t)} I(U_i = 0)}{\sum_i^n w_i I(U_i = 1)/n_B} \rfloor. \tag{4.23}$$

IM8. Let $\hat{\beta}$ and $V_{\hat{\beta}}$ denote the maximum likelihood estimates and the covariance matrix of $\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ based on $U_i = 0$ observations. Sample $\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}^{(t+1)}$ from its approximate posterior distribution $N(\hat{\beta}, V_{\hat{\beta}})$.

IM9. Calculate the proportion of unit nonrespondents that should be imputed as 1 for X_1. Denote it as $\hat{p}_1 = \frac{n_{1B}}{n_B}$. Then, calculate $\beta_2^{(t+1)}$ so that $\mathbb{E}[\beta_0^{(t+1)} + \beta_1^{(t+1)} X_2 + \beta_2^{(t+1)}] = logit(\hat{p}_1)$.

IM10. Draw imputations of X_1 for unit nonrespondents: $X_1|U = 1 \sim bern\Big(1/\big(1 + \exp(-\beta_0^{(t+1)} - \beta_1^{(t+1)} X_2 - \beta_2^{(t+1)})\big)\Big)$.

IM11. Let $\hat{\gamma}$ and $V_{\hat{\gamma}}$ denote the maximum likelihood estimates and the covariance matrix of $\begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix}$ based on $U_i = 0$ observations. Sample $\begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix}^{(t+1)}$ from its approximate posterior distribution $N(\hat{\gamma}, V_{\hat{\gamma}})$.

IM12. Draw imputations of $X_1^{(t+1)}$ when $(R_{i1}^x = 0, U_i = 0)$: $X_1^{(t+1)}|R_1^x = 1, U = 0 \sim bern\Big(1/\big(1 + \exp(-\beta_0^{(t+1)} - \beta_1^{(t+1)} X_2)\big)\Big)$.

We use auxiliary margins to estimate α_1 and β_2. To do so, we adjust the intercepts in the models for X_1 and X_2, assuming that the estimates of (α_0, β_0, β_1) for unit nonrespondents and unit respondents are the same. We update X_2 first and use X_2 as a predictor to impute X_1 because in this setting, X_2 only suffers from unit nonresponse whereas X_1 also suffers from item nonresponse. If both X_2 and X_1 suffer from item nonresponse, we need to generate starting values for X_2 as well. We can use multiple imputation with chained equation (MICE) to obtain starting values (Azur et al., 2011).

The sequence of updating depends on researchers' model specification. We recommend updating variables with smaller item nonresponse rates early in the sequence.

This algorithm can be viewed as a generalized version of the one proposed in Pham et al. (2019) which incorporates an offset called the "calibrated-δ adjustment" calculated from auxiliary information. Pham et al. (2019) discusses this algorithm in handling item nonrespondents, whereas the intercept matching algorithm proposed in our work can handle item and unit nonresponse simultaneously.

For the alternative model specification in (4.10) - (4.13), we exchange the intercepting matching steps on X_1 for a rejection sampling step. More specifically, we skip IM7-9, and in IM10 we draw imputations of X_1 for unit nonrespondents based on $X_1|U = 1 \sim bern\Big(1/\big(1+\exp(-\beta_0^{(t+1)} - \beta_1^{(t+1)} X_2)\big)\Big)$. In order to draw imputations for item nonresponses in X_1, we set $X_{i1}^\star = X_{i1}$ when $(R_{i1}^x = 0, U_i = 0)$ and when $U_i = 1$, then generate a candidate $X_{i1}^\star$ when $(R_{i1}^x = 1, U_i = 0)$ from its posterior predictive distribution. We then decide either to accept or reject this candidate based on the acceptance ratio:

$$p = \frac{N(\sum w_i X_{i1}^{\star}; T_{X1}, V_{X1})}{N(\sum w_i X_{i1}^{(t)}; T_{X1}, V_{X1})}. \tag{4.24}$$

The intercept matching algorithm can be readily generalized to more than two variables with auxiliary margins if we use the margins to add unit nonresponse indicators to the models for the survey variables. This can be done via repeating IM1-5 or IM6-10: generate parameter draws from their full conditional based on $U_i = 0$ cases alone, use the auxiliary margin to adjust the intercept for unit nonrespondents, and then draw imputations for unit nonrespondents for this variable with the adjusted intercept. On the other hand, if we choose to use auxiliary margins to specify nonignorable models for item nonresponse, we repeat the rejection sampling for those variables and do not adjust intercepts when imputing unit nonresponses for them.

4.3 Simulation Studies

This section presents simulation studies that examine the repeated sampling performance of the modeling strategy and associated MCMC sampler. Section 4.3.1 describes eight simulation scenarios that we use for two situations, namely when design weights w_i^d are available for unit respondents and when only adjusted weights w_i^a are available for unit respondents. Section 4.3.2 presents results when design weights for unit respondents are available. Section 4.3.3 provides details on how we downweight the w_i^a to mimic when design weights for unit respondents are not provided, and shows corresponding results.

4.3.1 Simulation Set-up

We generate simulated data based on (4.25) - (4.28), presented below, using probit models to generate the data. We use weights w_i^d from CPS 2012 data as described in Chapter 3. In each simulation scenario, the population has 133427 individuals. We set $\nu_0 = -1.2$, leading to approximately 11% nonrespondents. This matches the percentage of households that do not respond in the CPS data. We set $(\gamma_0, \gamma_1) = (-1.05, 0.2)$, as this setting gives approximately 18% item nonresponse to match the item nonresponse rate of the vote variable in the CPS data. We set $(\alpha_0 = 0.3, \beta_0 = 0.2)$. For the remaining three parameters $(\alpha_1, \beta_1, \beta_2)$, we have two values for each of them, resulting in $2^3 = 8$ simulation scenarios in total. We expect that these three variables may affect the performance of the models. α_1 represents the relationship between unit nonresponse U and X_2. We pick α_1 to be either -0.5 or -2.0, one for weak relationship and the other for strong. β_1 represents the relationship between X_1 and X_2, for which we pick 0.4 and 2.0 to show weak and strong correlations. β_2 represents the relationship between unit nonresponse U and X_1. We pick β_2 to be either -0.5 or -2.0 as we do for α_1.

$$\mathbb{P}(U = 1) = \pi_U,\ \pi_U = \Phi(\nu_0) \tag{4.25}$$

$$X_2 | U \sim bern(\pi_{x_2}),\ \pi_{x_2} = \Phi(\alpha_0 + \alpha_1 U) \tag{4.26}$$

$$X_1 | X_2, U \sim bern(\pi_{x_1}),\ \pi_{x_1} = \Phi(\beta_0 + \beta_1 X_2 + \beta_2 U) \tag{4.27}$$

$$\mathbb{P}(R_1^x = 1 | X_1, X_2, U) = \pi_R,\ \pi_R = \Phi(\gamma_0 + \gamma_1 X_2). \tag{4.28}$$

Table 4.2 and Table 4.3 display the mean of X_1 and X_2 for unit respondents and nonrespondents across the eight simulation settings. When $\alpha_1 = -2.0$, the difference in the mean of X_2 for unit respondents and nonrespondents is much larger than when $\alpha_1 = -0.5$. The difference in the mean of X_1 is more complicated. When $(\alpha_1, \beta_1) = (-0.5, 0.4)$, we see the least difference in the mean of X_2 across unit respondents and non-respondents, around 0.2. When $(\alpha_1, \beta_1, \beta_2) = (-2.0, 2.0, -2.0)$, we see the largest difference in the mean of X_2 across unit respondents and non-respondents, which are 0.833 and 0.062 respectively. This represents a very substantial nonresponse bias. When $(\alpha_1, \beta_2) = (-0.5, -2.0)$, the difference in mean of X_2 are also quite large, which are around 0.6.

Table 4.2: Mean of X_2 in the eight simulation settings for unit respondents and nonrespondents.

	$\alpha_1 = -0.5$	$\alpha_1 = -2.0$
$U = 0$	.618	.618
$U = 1$	.421	.044

For each population generated from the parameter settings, we sample 1000 simulated data sets using Poisson sampling with inclusion probability $\pi_i = 1/w_i^d$ for unit i. Then, we fit the MD-AM model in (4.20), (4.15), (4.16), and (4.21), which are based on logistic regression models, with the intercept matching algorithm to run 10000 MCMC iterations, discard the first 5000 runs as burn-in, and keep 5000 posterior samples. We create $L = 50$ multiple imputed data sets, $\boldsymbol{Z} = (\boldsymbol{Z}^{(1)}, \ldots, \boldsymbol{Z}^{(50)})$, from every 100 posterior sample. For each completed data set $\boldsymbol{Z}^{(\ell)}$, we compute the estimates of

Table 4.3: Mean of X_1 in the eight simulation settings for unit respondents and nonrespondents.

		$\alpha_1 = -0.5$		$\alpha_1 = -2.0$	
		$\beta_1 = 0.4$	$\beta_1 = 2.0$	$\beta_1 = 0.4$	$\beta_1 = 2.0$
U = 0	$\beta_2 = -0.5$	.668	.830	.671	.833
	$\beta_2 = -2.0$	.670	.830	.672	.833
U = 1	$\beta_2 = -0.5$	.452	.625	.386	.409
	$\beta_2 = -2.0$	.053	.263	.038	.062

T_{X1}, T_{X2}, conditional margins and their standard errors with the "survey" package (Lumley, 2021) in R. We compute the design-based estimates of $(\alpha_0, \alpha_1, \beta_0, \beta_1, \beta_2)$ along with the corresponding standard errors using the survey-weighted generalized linear models option in the R package "survey". We also compute estimates of $(\gamma_0, \gamma_1, \gamma_2)$, which do not depend on the design weights, along with the corresponding standard errors, using the standard generalized linear models routine in R.

Together with the MD-AM model, we also analyze sampled data sets in each simulation setting without using auxiliary information, as specified in (4.5) - (4.8). Without auxiliary margins, we fit the exact same models for X_1 and X_2 for unit respondents and unit nonrespondents. There is no information available to adjust the intercept for unit nonrespondents as in the MD-AM model with margins. In the MCMC sampler for the model without margins, we skip IM1, IM2, IM4 and draw imputations from $X_2|U = 1 \sim bern\Big(1/\big(1 + \exp(-\alpha_0^{(t+1)})\big)\Big)$ when sampling X_2 for unit nonrespondents. Similarly, we skip IM6-8 and draw imputations of X_1 for nonresponse units from $X_1|U = 1 \sim bern\Big(1/\big(1+\exp(-\beta_0^{(t+1)} - \beta_1^{(t+1)} X_2)\big)\Big)$. We refer to analysis results without auxiliary margins as ICIN, meaning that the nonresponse mechanism we specify under this scenario is ICIN.

4.3.2 Simulations And Results When Design Weights Available

Starting with known design weights for unit respondents, we first look at results for survey variables, which are of primary interest in practice. Results for the eight simulation scenarios are shown in Table 4.4 and Table 4.5. Table 4.4 displays four simulation settings with $\alpha_1 = -0.5$, whereas Table 4.5 presents the rest with $\alpha_1 = -2.0$. For each simulation scenario, we summarize six quantities: estimates of T_{X1} and T_{X2} as well as the conditional distributions of X_1 given X_2 and X_2 given X_1. The first columns of those tables are called "truth", representing the true population quantities in each simulation setting. CI coverage is the 95% confidence interval coverage rate for each estimand across 1000 samples. "SD" is the standard deviation of 1000 sample means for each estimand. "Avg. est. SD" is the average of the standard deviations. We compute the average of the variances based on (Rubin, 1987)'s multiple imputation combining rules, taking the square root to get "avg. est. SD". "Pre-SD" stands for pre-missing standard deviation, the standard deviation of the 1000 point estimates calculated from the 1000 sampled full data sets with original weights. Recall that b_m is the between imputation standard deviation, part of Rubin's combining rule. The last column is obtained from taking the mean of b_m/m for each estimand and then taking its square root, where $m = 50$ is the number of multiple imputed data set generated for every data set. As we discuss later, having $\sqrt{b_m/m}$ is useful for assessing the multiple imputation variance estimates.

We begin by examining the results for T_{X1} and T_{X2}. Point estimates from MD-AM are less biased than those under ICIN when design weights for unit respondents are known. This is not surprising, as the auxiliary margins provide extra guidance for the imputation process. We do see MD-AM overestimate T_{X2} in Table 4.5. As we can see in Table 4.2, the mean of X_2 for unit respondents is much larger than that for unit non-respondents. Because of the large unit nonresponse bias, in the scenarios in Table 4.5 the weighted sum of X_2 of unit respondents alone could already be larger than T_{X2}, making it not possible for the imputation of X_2 for unit nonrespondents to adjust the estimate down sufficiently. This causes an upward bias in the multiple imputation estimates of T_{X2}. We note that this bias does not exist in Table 4.4, where the nonresponse bias is not so dramatic.

Across all eight simulation settings, the CI coverage rates for T_{X1} and T_{X2} are much higher under MD-AM framework than under ICIN. This happens because, first, MD-AM gives more accurate point estimates and second, the average estimated standard deviations based on Rubin's multiple imputation combining rule tend to be positively biased. Assuming homogeneous distribution of X_2 across unit respondents and nonrespondents, as done under ICIN model specification, is least problematic when $\alpha_1 = -0.5$ though results for the MD-AM model still outperform those for ICIN model. On the other hand, when we set $\alpha_1 = -2.0$, estimates under ICIN become so biased that all CI coverage rates of T_{X2} are 0. We see similar patterns in the coverage rates for T_{X1}.

As observed in Chapter 3, when we use auxiliary information in generating multiple imputed data sets, the true standard deviation of the point estimates tend to be smaller than the average estimated standard deviation, especially in estimating quantities that employ auxiliary margins directly. In our current simulation set-up, those are T_{X1} and T_{X2}. When we know the design weights for unit respondents, estimates of standard deviations from multiple imputation combining rules are at least three times as large as the truth under the MD-AM model specification. Larger true standard deviations in some simulation settings are due to weighted sums of unit respondents' values alone being close to auxiliary margin.

Since auxiliary margins enable us to constrain imputed data sets to have multiple imputation point estimates of T_{X1} and T_{X2} very close to the truth, the true standard deviations of those two quantities are even smaller than pre-missing standard deviation. We generate imputed data sets to match totals, and the average of those draws will converge to T_{X1} and T_{X2} due to the law of large numbers. The true standard deviations of T_{X1}, T_{X2} are in general between $\sqrt{b_m/m}$ and the average estimated standard deviation from the multiple combining rules. When design weights are known for unit respondents, we suggest using standard deviations calculated from Rubin's multiple imputation combining rules, even though these may be conservative, because $\sqrt{b_m/m}$ underestimates the uncertainty significantly when nonresponse bias is large or correlation between X_1 and X_2 is strong. True standard deviations under ICIN model are larger than those under MD-AM framework due to the employment of auxiliary margins in the latter.

Next, we examine estimates of conditional distributions. When design weights are known for unit respondents, estimates of conditional distributions obtained from MD-AM are more accurate than those under ICIN without using margins for all eight simulation settings overall. There are some instances when they achieve similar level of accuracy for one or two of the four conditionals, for example in Table 4.5 (d), but the MD-AM framework always performs better if we consider estimates of all four conditional margins. CI coverage rates under MD-AM are always higher than those under ICIN, due to more accurate point estimate and larger average estimated standard deviation as we have seen when analyzing results for T_{X1} and T_{X2}.

The average estimated multiple imputation standard deviations of the conditional margins under MD-AM are larger than true standard deviations, which may or may not be smaller than pre-missing standard deviations. This may because auxiliary margins available are on marginal distributions of X_1 and X_2. We do not have auxiliary information on the joint or conditional distribution of X_1 and X_2. The true standard deviations of the conditional margins are again in between the average estimated standard deviation and $\sqrt{b_m/m}$. For conditional margins under MD-AM, we again suggest using Rubin's multiple imputation combining rules. Average estimated standard deviations of conditionals under ICIN are similar to true standard deviations, which are also close to pre-missing standard deviations.

Although arguably of less interest than the totals and conditional means, simulation results for model parameters with known design weights for unit respondents are shown in Table 4.6 and 4.7. Here, the "truth" columns in those tables are obtained from averaging the estimated model parameters in 1000 pre-missing samples.

Estimates of $(\alpha_0, \gamma_0, \gamma_1)$ are similar and close to the "truth" values under MD-AM and ICIN as these three parameters are estimable with observed data alone. α_1 and β_2 are the most difficult parameters to estimate as the only information we have on them derives from auxiliary margins. Estimates of α_1 and β_2 under ICIN are close to 0 in all simulation settings as those two parameters are not identifiable without using margins, and our prior distributions for them are centered at zero. Estimates of α_1 under MD-AM are more accurate when the unit nonresponse bias in X_2 is weaker at $\alpha_1 = -0.5$ than when the unit nonresponse bias in X_2 is at $\alpha_1 = -2.0$. Under the same correlation between X_1 and X_2, estimates of β_2 under MD-AM are more accurate when the unit nonresponse bias in X_1 is weak than when the unit nonresponse bias in it is strong, except for Table 4.6 (c) and (d). Average estimated standard deviations under MD-AM and ICIN are similar except for α_1 and β_2, because (except for α_1 and β_2), estimating these model parameters does not directly involve auxiliary margins.

Table 4.4: Simulation results for totals and conditional margins assuming design weights for unit respondents are known and $\alpha_0 = 0.3, \alpha_1 = -0.5, \beta_0 = 0.2, \nu_0 = -1.2, \gamma_0 = -1.05, \gamma_1 = 0.2$. Results averaged over 1000 sampled data sets from simulated population using Poisson sampling and 50 multiple imputed data sets.

	Estimate			CI Coverage %		SD			Avg. Est. SD		$\sqrt{b_m/m}$
	Truth	MDAM	ICIN	MDAM	ICIN	Pre-SD	MDAM	ICIN	MDAM	ICIN	MDAM
(a) $\boldsymbol{\alpha_1 = -0.5, \beta_1 = 0.4, \beta_2 = -0.5}$											
T_{X2}	79477	79460	82450	100	36.2	1348.2	189.9	848.0	1976.7	1326.6	188.1
T_{X1}	85981	85985	89235	100	32.1	1331.3	206.6	828.3	1864.3	1425.9	202.8
$X_2 = 0 \mid X_1 = 0$	.514	.511	.483	100	31.8	.011	.010	.011	.019	.012	.002
$X_2 = 0 \mid X_1 = 1$	.344	.345	.332	100	67.7	.008	.005	.008	.012	.008	.001
$X_1 = 0 \mid X_2 = 0$	.452	.450	.419	100	16.4	.010	.009	.010	.019	.011	.002
$X_1 = 0 \mid X_2 = 1$	.290	.292	.277	100	63.7	.007	.006	.008	.012	.008	.001
(b) $\boldsymbol{\alpha_1 = -0.5, \beta_1 = 0.4, \beta_2 = -2.0}$											
T_{X2}	79499	79493	82545	100	31.7	1332.8	184.4	814.6	1853.3	1326.9	185.9
T_{X1}	79834	80248	89128	99.3	0	1333.8	495.0	819.1	1732.3	1424.5	157.3
$X_2 = 0 \mid X_1 = 0$	.515	.517	.489	100	42.3	.010	.012	.011	.025	.012	.003
$X_2 = 0 \mid X_1 = 1$	.329	.330	.328	98.7	96.7	.008	.007	.007	.009	.008	$< .001$
$X_1 = 0 \mid X_2 = 0$	.512	.509	.426	99.5	0	.010	.011	.010	.017	.011	.002
$X_1 = 0 \mid X_2 = 1$	.327	.323	.274	99.1	0	.008	.009	.007	.015	.008	.002
(c) $\boldsymbol{\alpha_1 = -0.5, \beta_1 = 2.0, \beta_2 = -0.5}$											
T_{X2}	79026	79021	82078	100	31	1283.6	181.9	822.4	1863.2	1321.9	188.5
T_{X1}	107451	106255	110517	100	50.5	1527.5	364.5	676.0	1838.9	1540.0	149.9
$X_2 = 0 \mid X_1 = 0$	.948	.946	.949	99.7	96.4	.006	.008	.007	.013	.008	.002
$X_2 = 0 \mid X_1 = 1$	.277	.270	.268	100	73.3	.006	.003	.006	.013	.007	.002
$X_1 = 0 \mid X_2 = 0$	.453	.472	.424	98.9	28.9	.010	.008	.011	.020	.011	.002
$X_1 = 0 \mid X_2 = 1$	.017	.019	.014	99.9	75.5	.002	.003	.002	.005	.002	.001
(d) $\boldsymbol{\alpha_1 = -0.5, \beta_1 = 2.0, \beta_2 = -2.0}$											
T_{X2}	79592	79577	82583	100	33.7	1272.5	186.1	804.6	1873.8	1328.3	189.7
T_{X1}	102095	103364	111004	100	0	1441.6	254.0	635.6	1948.9	1543.2	176.7
$X_2 = 0 \mid X_1 = 0$	.874	.900	.946	96	0	.008	.017	.007	.039	.008	.005
$X_2 = 0 \mid X_1 = 1$	.259	.259	.267	99.7	82	.006	.005	.006	.008	.007	.005
$X_1 = 0 \mid X_2 = 0$	.509	.502	.417	98.3	0	.008	.010	.010	.015	.011	.001
$X_1 = 0 \mid X_2 = 1$	.050	.038	.015	93.7	0	.003	.006	.002	.016	.002	.002

Table 4.5: Simulation results for totals and conditional margins assuming design weights for unit respondents are known and $\alpha_0 = 0.3, \alpha_1 = -2.0, \beta_0 = 0.2, \nu_0 = -1.2, \gamma_0 = -1.05, \gamma_1 = 0.2$. Results averaged over 1000 sampled data sets from simulated population using Poisson sampling and 50 multiple imputed data sets.

	Estimate			CI Coverage %		SD			Avg. Est. SD		$\sqrt{b_m/m}$
	Truth	MDAM	ICIN	MDAM	ICIN	Pre-SD	MDAM	ICIN	MDAM	ICIN	MDAM
(a) $\boldsymbol{\alpha_1 = -2.0, \beta_1 = 0.4, \beta_2 = -0.5}$											
T_{X2}	73451	73879	82167	99.3	0	1249.1	520.9	819.3	1582.5	1324.3	132.8
T_{X1}	84985	84983	89221	100	4.4	1333.2	203.0	819.7	1965.3	1424.4	201.3
$X_2 = 0 \mid X_1 = 0$	.581	.577	.491	99.7	0	.010	.012	.011	.018	.012	.002
$X_2 = 0 \mid X_1 = 1$	.375	.371	.331	98.9	0	.007	.009	.007	.014	.008	.002
$X_1 = 0 \mid X_2 = 0$	.469	.470	.423	100	1	.009	.010	.010	.024	.011	.003
$X_1 = 0 \mid X_2 = 1$	.277	.277	.274	98.5	97.1	.008	.008	.008	.009	.008	.001
(b) $\boldsymbol{\alpha_1 = -2.0, \beta_1 = 0.4, \beta_2 = -2.0}$											
T_{X2}	73410	73847	82216	99.1	0	1284.7	546.2	829.5	1580.7	1325.4	132.5
T_{X1}	79506	79995	89183	99.6	0	1306.0	570.4	821.4	1686.4	1425.3	147.3
$X_2 = 0 \mid X_1 = 0$	.617	.608	.488	98.8	0	.010	.011	.011	.019	.012	.002
$X_2 = 0 \mid X_1 = 1$	.336	.339	.332	97.4	93.8	.008	.009	.008	.011	.008	.001
$X_1 = 0 \mid X_2 = 0$	.555	.544	.422	97.9	0	.010	.010	.011	.019	.011	.002
$X_1 = 0 \mid X_2 = 1$	.281	.284	.276	98.7	92.4	.008	.009	.008	.012	.008	.001
(c) $\boldsymbol{\alpha_1 = -2.0, \beta_1 = 2.0, \beta_2 = -0.5}$											
T_{X2}	73808	74258	82693	99.2	0	1285.5	550.7	832.5	1571.9	1327.0	130.0
T_{X1}	104446	104198	110972	100	0	1512.8	304.5	655.0	2054	1542.6	198.2
$X_2 = 0 \mid X_1 = 0$	.965	.964	.949	98.2	54.5	.005	.005	.007	.006	.008	.001
$X_2 = 0 \mid X_1 = 1$	.303	.297	.265	99.2	0	.007	.005	.007	.015	.007	.002
$X_1 = 0 \mid X_2 = 0$	.469	.476	.420	100	1	.010	.005	.011	.027	.011	.004
$X_1 = 0 \mid X_2 = 1$	.014	.014	.014	98	96.7	.002	.002	.002	.002	.002	$< .001$
(d) $\boldsymbol{\alpha_1 = -2.0, \beta_1 = 2.0, \beta_2 = -2.0}$											
T_{X2}	73582	73986	82344	99.7	0	1208.3	501.8	854.8	1578.3	1324.3	132.0
T_{X1}	98814	99482	110485	99.7	0	1373.6	457.9	649.9	1883.2	1539.0	165.2
$X_2 = 0 \mid X_1 = 0$	.961	.957	.948	99.6	68.3	.005	.008	.007	.018	.008	.002
$X_2 = 0 \mid X_1 = 1$	.269	.271	.266	99.1	91.4	.007	.006	.007	.010	.007	.001
$X_1 = 0 \mid X_2 = 0$	.556	.546	.426	97.9	0	.010	.008	.011	.018	.011	.002
$X_1 = 0 \mid X_2 = 1$	.018	.020	.015	98.9	60	.002	.004	.002	.009	.002	.001

Table 4.6: Simulation results for model parameters assuming design weights for unit respondents are known and $\alpha_0 = 0.3, \alpha_1 = -0.5, \beta_0 = 0.2, \nu_0 = -1.2, \gamma_0 = -1.05, \gamma_1 = 0.2$. Results averaged over 1000 sampled data sets from simulated population using Poisson sampling and 50 multiple imputed data sets.

	Estimate			CI Coverage %		SD			Avg. Est. SD	
	Truth	MDAM	ICIN	MDAM	ICIN	Pre-SD	MDAM	ICIN	MDAM	ICIN
(a) $\boldsymbol{\alpha_1 = -0.5, \beta_1 = 0.4, \beta_2 = -0.5}$										
α_0	.300	.300	.300	95.5	95.5	.018	.018	.018	.018	.018
α_1	-.506	-.513	-.002	99.7	0	.037	.178	.012	.256	.057
β_0	.202	.205	.205	97	96.7	.026	.027	.027	.030	.030
β_1	.396	.388	.388	97.1	97.7	.033	.035	.035	.038	.038
β_2	-.498	-.493	-.004	99.8	0	.039	.167	.012	.268	.058
γ_0	-1.06	-1.06	-1.06	96	96	.025	.025	.025	.026	.026
γ_1	.210	.210	.210	96.6	96.6	.030	.030	.030	.032	.032
(b) $\boldsymbol{\alpha_1 = -0.5, \beta_1 = 0.4, \beta_2 = -2.0}$										
α_0	.302	.302	.302	96.5	96.5	.017	.017	.017	.018	.018
α_1	-.513	-.522	-.001	99.6	0	.037	.176	.012	.259	.057
β_0	.195	.189	.187	97.3	96.4	.026	.027	.027	.030	.030
β_1	.405	.410	.413	97.3	96.9	.034	.035	.035	.040	.038
β_2	-1.998	-3.055	-.002	90.9	0	.058	1.255	.012	1.673	.058
γ_0	-1.070	-1.070	-1.070	97.3	97.3	.024	.024	.024	.026	.026
γ_1	.221	.221	.221	97.4	97.4	.028	.028	.028	.032	.032
(c) $\boldsymbol{\alpha_1 = -0.5, \beta_1 = 2.0, \beta_2 = -0.5}$										
α_0	.292	.292	.292	95.9	95.9	.017	.017	.017	.018	.018
α_1	-.498	-.516	.003	99.7	0	.037	.170	.013	.252	.056
β_0	.198	.185	.194	95.8	95.9	.028	.028	.029	.030	.030
β_1	1.993	2.049	2.002	91	96	.057	.067	.065	.072	.068
β_2	-.477	-.782	-.004	97.4	0	.047	.128	.017	.332	.081
γ_0	-1.026	-1.026	-1.026	96.6	96.6	.024	.024	.024	.025	.025
γ_1	.174	.174	.174	95.8	95.8	.030	.030	.030	.032	.032
(d) $\boldsymbol{\alpha_1 = -0.5, \beta_1 = 2.0, \beta_2 = -2.0}$										
α_0	.303	.303	.303	96.8	96.8	.017	.017	.017	.018	.018
α_1	-.500	-.510	-.002	99.8	0	.038	.163	.012	.249	.057
β_0	.204	.199	.208	96.5	97.1	.025	.028	.027	.030	.030
β_1	1.982	2.024	1.977	94.3	98	.056	.067	.062	.072	.068
β_2	-2.035	-1.847	.006	84.4	0	.061	.507	.016	.975	.082
γ_0	-1.046	-1.046	-1.046	97.6	97.6	.023	.023	.023	.026	.026
γ_1	.191	.191	.191	97.1	97.1	.029	.029	.029	.032	.032

Table 4.7: Simulation results for model parameters assuming design weights for unit respondents are known and $\alpha_0 = 0.3, \alpha_1 = -2.0, \beta_0 = 0.2, \nu_0 = -1.2, \gamma_0 = -1.05, \gamma_1 = 0.2$. Results averaged over 1000 sampled data sets from simulated population using Poisson sampling and 50 multiple imputed data sets.

	Estimate			CI Coverage %		SD			Avg. Est. SD	
	Truth	MDAM	ICIN	MDAM	ICIN	Pre-SD	MDAM	ICIN	MDAM	ICIN
(a) $\boldsymbol{\alpha_1 = -2.0, \beta_1 = 0.4, \beta_2 = -0.5}$										
α_0	.295	.295	.295	95.6	95.6	.017	.017	.017	.018	.018
α_1	-2.027	-3.181	-.002	88.8	0	.063	1.255	.012	1.649	.056
β_0	.195	.193	.193	98.2	97.8	.026	.027	.026	.030	.030
β_1	.402	.406	.406	97.2	97.2	.035	.036	.035	.040	.038
β_2	-.491	-.520	.008	99.8	0	.041	.187	.011	.307	.058
γ_0	-1.061	-1.061	-1.061	97.3	97.3	.023	.023	.023	.026	.026
γ_1	.216	.216	.216	98	97.7	.028	.028	.028	.032	.032
(b) $\boldsymbol{\alpha_1 = -2.0, \beta_1 = 0.4, \beta_2 = -2.0}$										
α_0	.296	.296	.296	96.9	96.9	.017	.018	.017	.018	.018
α_1	-2.008	-3.164	-.005	88.4	0	.061	1.255	.012	1.645	.057
β_0	.196	.198	.198	96.1	96.1	.028	.028	.028	.030	.030
β_1	.401	.398	.399	96.7	96.8	.037	.037	.036	.040	.038
β_2	-1.965	-3.168	.001	89	0	.064	1.253	.012	1.692	.058
γ_0	-1.073	-1.073	-1.073	97	97	.024	.024	.024	.026	.026
γ_1	.224	.224	.224	96.8	96.9	.029	.029	.029	.032	.032
(c) $\boldsymbol{\alpha_1 = -2.0, \beta_1 = 2.0, \beta_2 = -0.5}$										
α_0	.305	.305	.305	94.7	94.7	.017	.017	.017	.018	.018
α_1	-2.050	-3.259	.003	87.3	0	.063	1.285	.012	1.646	.057
β_0	.198	.200	.201	98	96.9	.029	.029	.029	.031	.030
β_1	2.012	2.018	2.008	97	97.1	.063	.067	.064	.072	.069
β_2	-.486	-.633	.001	100	0	.045	.195	.018	.391	.082
γ_0	-1.059	-1.059	-1.059	97.7	97.7	.024	.024	.024	.026	.026
γ_1	.212	.212	.212	97.2	97.2	.030	.030	.030	.032	.032
(d) $\boldsymbol{\alpha_1 = -2.0, \beta_1 = 2.0, \beta_2 = -2.0}$										
α_0	.298	.298	.298	94.8	94.8	.018	.018	.018	.018	.018
α_1	-2.043	-3.173	.004	88.7	0	.062	1.247	.013	1.648	.057
β_0	.193	.186	.188	95.7	96.3	.028	.029	.027	.031	.030
β_1	1.983	2.009	1.998	96.4	97.2	.059	.065	.061	.072	.068
β_2	-2.048	-3.164	-.001	92.1	0	.067	1.274	.017	2.034	.081
γ_0	-1.033	-1.033	-1.033	97.9	97.9	.024	.024	.024	.026	.026
γ_1	.192	.192	.192	96.7	96.7	.029	.029	.029	.032	.032

4.3.3 Simulations And Results With Adjusted Weights

In this section, we examine how the MD-AM model works when we do not know the design weights for unit respondents. Thus, in our simulations, we need to mimic a nonresponse weighting adjustment to the unit respondents' weights. Although the CPS uses post–stratification, for the simulations we use a weight adjustment method that is similar to poststratification but does not require population counts in each post-stratum. The weight adjustment method we employ is known as weighting class adjustment. Sampled data are divided into weighting adjustment classes where unit respondents and nonrespondents are believed to be similar. In each weighting class, weights of unit respondents are increased to represent themselves as well as unit nonrespondents' share of the population (Lohr, 2010).

In the simulation study, we divide the sampled data into two weighting adjustment classes based on the value of X_1. We estimate the response probability for each class by

$$\hat{\phi}_c = \frac{\sum_{i=1}^{n} w_i^d I(X_1 = c, U = 0)}{\sum_{i=1}^{n} w_i^d I(X_1 = c)}, \tag{4.29}$$

where $c = \{0, 1\}$ and w_i^d are design weights for unit i. This is the ratio of the sum of weights for unit respondents over the sum of weights for all sampled units in class c. Then we multiply the design weights for unit respondents by $1/\hat{\phi}_c$ according to their X_1 values to obtain adjusted weights w_i^a, namely,

$$w_i^a = \begin{cases} w_i^d/\hat{\phi}_0, & \text{if } X_i = 0 \\ w_i^d/\hat{\phi}_1, & \text{if } X_i = 1. \end{cases} \tag{4.30}$$

To create analysis weights w_i^*, we follow (4.19) for the unit nonrespondents and unit respondents. We note that this attempt to "unadjust" the w_i^a does not get back to the corresponding w_i^d. One might expect this to create potential biases. Nonetheless, absent any information about the unit nonrespondents' weights or population totals for the weighting classes, the simple "unadjusted" weights formula can be viewed as a default procedure. Thus, it is worth evaluating the performance of the MD-AM model under this procedure.

Before presenting the simulation results, we return to a conjecture that we made in section 4.2.2. When design weights are unknown for unit respondents, whether we create weights for unit nonrespondents so that all the weights equal to true population size N or $\hat{N}$ does not make much difference in the simulation study results. Table 4.8 displays results of 100 repeated runs with parameter settings at $(\alpha_1, \beta_1, \beta_2) = (-0.5, 0.4, -0.5)$ when we use N to create weights and 1000 repeated runs when we use $\hat{N}$ to create weights. We barely observe any difference in the results between using N or $\hat{N}$, especially for results of survey variables. When we do not know the true population size, using $\hat{N}$ is an appropriate substitute to achieve comparable outcomes. Hence, we use this way to create weights for unit nonrespondents in the simulation studies in this section.

Table 4.9 and Table 4.10 display the simulation results when design weights for unit respondents are not provided. Table 4.9 displays four simulation settings with $\alpha_1 = -0.5$, whereas Table 4.10 presents the rest with $\alpha_1 = -2.0$. Point estimates of totals under MD-AM are less biased than under ICIN except T_{X2} in Table 4.9 (c). In this case with $(\alpha_1, \beta_1, \gamma_2) = (-0.5, 2.0, -0.5)$, nonresponse biases in X_1 and X_2 are small, and the correlation between X_1 and X_2 is strong. When imputing X_1 for unit nonrespondents, the correlation between X_1 and X_2 has a stronger effect on imputations than the nonresponse bias in X_1, resulting in less ones imputed for X_1 and hence an underestimate of T_{X1} in the MD-AM framework. Also because the nonresponse biases in X_1 and X_2 are small, ignoring them under ICIN is not terribly problematic, so that the estimate of T_{X1} happens to be closer to the truth in this setting. Comparing Table 4.5 and Table 4.10, we also notice that estimates of T_{X2} from MD-AM are closer to the truth when the nonresponse bias in X_2 is large with unknown design weights for unit respondents. When we do not know the design weights for unit respondents, we down-weight their weights to create weights for unit nonrespondents. In our simulation design, this weight adjustment lowers the weighted sum of X_2 for unit respondents, resulting in more accurate estimates of T_{X2}.

Across all eight simulation settings, the CI coverage rates for T_{X1} and T_{X2} are much higher under MD-AM than under ICIN, mirroring what we have seen when the design weights are known for unit respondents. Estimates of the standard deviations from multiple imputation combining rules are almost ten times as large as the truth under MD-AM model specification. When design weights are unknown for unit respondents, we suggest using $\sqrt{b_m/m}$ for the estimated standard deviations of the estimated totals for variables with margins. Indeed, $\sqrt{b_m/m}$ for T_{X1} and T_{X2} are much closer to the truth than average estimated standard deviation.

Next, we examine estimates of conditional margins. When design weights for unit respondents are not available, the performance of MD-AM in estimating conditional distributions is worse than when design weights for unit respondents are known, although still better than under ICIN except for the simulation setting in Table 4.10 (b). In this setting, the relationships between U and X_1 as well as U and X_2 are both strong, whereas the relationship between X_1 and X_2 is weak. It is the most difficult simulation scenario to get a good estimate for conditional margins. When the correlation between U and X_1 is weaker at $\beta_2 = -0.5$, CI coverage of conditional margins under MD-AM framework is higher than when the correlation between U and X_1 is at $\beta_2 = -2.0$, except for table 4.9 (a) and (b), where the performances under these two setting are similar, the latter slightly better. Stronger correlation between X_1 and X_2 leads to higher CI coverage in Table 4.10, where $\alpha_1 = -2.0$. We do not observe similar advantages of relationship between X_1 and X_2 when $\alpha_1 = -0.5$.

Comparing Table 4.4 and Table 4.5 to Table 4.9 and Table 4.10, the MD-AM model performs better when the design weights for unit respondents are available, especially in estimating conditional margins. The most obvious example is from Table 4.5 (b) to Table 4.10 (b). This suggests that agencies can help researchers by publishing the design weights of the unit nonrespondents, in addition to any adjusted weights.

We conclude this section on simulation results with a discussion of the model parameters. Table 4.11 displays the results for the model parameters in one simulation setting. The overall patterns in the simulation results for model parameters are quite similar across eight simulation scenarios. Recall that we do not know the weights for unit nonrespondents and have to do weight adjustment. This adjustment makes studying the quality of survey weighted estimates, namely $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, challenging. "Truth" column records the parameter values we used in data generating process, but they are not a good standard to compare with as we adjust the weights in every simulation run. Thus, our primary focus to present Table 4.11 is not to study how well we retrieve the parameters, but to discuss a few interesting points.

Compared to when we know design weights for unit respondents, α_1 and β_2 under MD-AM are more biased, especially β_2, signifying that after the weight adjustment, intercept matching algorithm needs to impute data in particular ways that agree with auxiliary information based on new weights. Both true standard deviation and average estimated standard deviation of α_1 and β_2 are quite large. As when design weights for unit respondents are known, the average estimated standard deviations of $\boldsymbol{\alpha}, \boldsymbol{\beta}$ based on multiple imputation combining rules are quite close to the corre-

Table 4.8: Simulation results of parameters in one setting at $(\alpha_1, \beta_1, \beta_2) = (-0.5, 0.4, -0.5)$. For each sampled data sets with missingness, 50 multiple imputed data sets were generated. Assuming design weights for unit respondents are unknown and we use N and $\sum w_i^a$ to create weights for unit nonrespondents, and MDAM model for analysis. "Truth" column refers to the parameters we used to generate this population for model parameters, and population truth for survey variables.

	Estimate			SD		Avg. Est. SD	
	Truth	N	$\sum w_i^a$	N	$\sum w_i^a$	N	$\sum w_i^a$
α_0	.300	.291	.291	.016	.017	.018	.018
α_1	-.500	-.423	-.416	.218	.243	.246	.241
β_0	.200	.144	.146	.026	.028	.030	.030
β_1	.400	.388	.392	.035	.036	.038	.039
β_2	-.500	.021	.010	.225	.267	.291	.316
γ_0	-1.05	-1.06	-1.06	.024	.024	.026	.026
γ_1	.200	.210	.211	.030	.029	.032	.032
T_{X2}	79477	79508	79477	202.8	179.6	1858.5	1844.8
T_{X1}	85981	85906	85961	198.2	189.8	1909.4	1906.6
$X_2 = 0 \vert X_1 = 0$	.514	.502	.504	.012	.014	.016	.016
$X_2 = 0 \vert X_1 = 1$	.344	.350	.349	.007	.007	.013	.013
$X_1 = 0 \vert X_2 = 0$	.452	.443	.443	.012	.014	.019	.018
$X_1 = 0 \vert X_2 = 1$	.290	.298	.296	.007	.008	.011	.011

sponding true standard deviation except for α_1 and β_2 because those estimates do not involve auxiliary margins directly.

Table 4.9: Simulation results for totals and conditional margins assuming design weights for unit respondents are not known and $\alpha_0 = 0.3, \alpha_1 = -0.5, \beta_0 = 0.2, \nu_0 = -1.2, \gamma_0 = -1.05, \gamma_1 = 0.2$. Results averaged over 1000 sampled data sets from simulated population using Poisson sampling and 50 multiple imputed data sets.

	Estimate			CI Coverage %		SD			Avg. Est. SD		$\sqrt{b_m/m}$
	Truth	MDAM	ICIN	MDAM	ICIN	Pre-SD	MDAM	ICIN	MDAM	ICIN	MDAM
(a) $\boldsymbol{\alpha_1 = -0.5, \beta_1 = 0.4, \beta_2 = -0.5}$											
T_{X2}	79477	79477	81986	100	55.2	1339.2	179.6	1357.2	1844.8	1320.0	184.8
T_{X1}	85981	85961	86717	100	92.5	1384.0	189.8	1380.1	1906.6	1388.0	193.3
$X_2 = 0 \mid X_1 = 0$	.514	.504	.484	93	34.5	.011	.014	.011	.016	.012	.002
$X_2 = 0 \mid X_1 = 1$	.344	.349	.332	99.8	70.4	.008	.007	.008	.013	.008	.002
$X_1 = 0 \mid X_2 = 0$	.452	.443	.440	97.9	83.9	.010	.014	.011	.018	.012	.002
$X_1 = 0 \mid X_2 = 1$	.290	.296	.293	98.6	96.4	.008	.008	.008	.011	.009	.001
(b) $\boldsymbol{\alpha_1 = -0.5, \beta_1 = 0.4, \beta_2 = -2.0}$											
T_{X2}	79499	79489	81227	100	75	1306.3	184.9	1349.5	1837.6	1316.9	183.4
T_{X1}	79834	79817	82176	100	59	1233.9	177.3	1330.3	1764.6	1320.4	176.3
$X_2 = 0 \mid X_1 = 0$	.515	.505	.492	97.1	52.9	.010	.012	.011	.017	.012	.002
$X_2 = 0 \mid X_1 = 1$	.329	.336	.328	99.8	96.1	.008	.007	.007	.012	.008	.001
$X_1 = 0 \mid X_2 = 0$	.512	.502	.483	96.7	28.7	.010	.012	.010	.016	.012	.002
$X_1 = 0 \mid X_2 = 1$	.327	.333	.320	99.6	91.9	.008	.007	.008	.012	.009	.001
(c) $\boldsymbol{\alpha_1 = -0.5, \beta_1 = 2.0, \beta_2 = -0.5}$											
T_{X2}	79026	79016	80115	100	89.2	1277.3	178.8	1281.4	1800.0	1288.1	179.7
T_{X1}	107451	105915	108007	100	95.1	1470.4	272.5	1453.4	2056.0	1503.9	201.9
$X_2 = 0 \mid X_1 = 0$	.948	.944	.950	100	94.9	.007	.007	.007	.013	.007	.002
$X_2 = 0 \mid X_1 = 1$	.277	.268	.270	100	83.9	.006	.003	.006	.014	.007	.002
$X_1 = 0 \mid X_2 = 0$	.453	.477	.453	85.9	97.7	.010	.016	.010	.026	.012	.003
$X_1 = 0 \mid X_2 = 1$	.017	.019	.016	100	92.3	.002	.003	.002	.005	.002	$< .001$
(d) $\boldsymbol{\alpha_1 = -0.5, \beta_1 = 2.0, \beta_2 = -2.0}$											
T_{X2}	79592	79572	77283	100	56.5	1344.3	176.4	1303.6	1727.2	1240.2	169.8
T_{X1}	102095	101788	104067	100	73.9	1557.1	336.7	1538.9	1746.4	1450.0	145.0
$X_2 = 0 \mid X_1 = 0$	.874	.900	.949	95.8	0	.009	.013	.007	.031	.008	.004
$X_2 = 0 \mid X_1 = 1$	.259	.250	.272	88.6	53.8	.006	.006	.006	.008	.007	$< .001$
$X_1 = 0 \mid X_2 = 0$	.509	.528	.496	74.1	86.6	.010	.013	.010	.014	.012	.001
$X_1 = 0 \mid X_2 = 1$	.050	.040	.020	96.5	0	.004	.005	.003	.013	.003	.002

Table 4.10: Simulation results for totals and conditional margins assuming design weights for unit respondents are not known and $\alpha_0 = 0.3, \alpha_1 = -2.0, \beta_0 = 0.2, \nu_0 = -1.2, \gamma_0 = -1.05, \gamma_1 = 0.2$. Results averaged over 1000 sampled data sets from simulated population using Poisson sampling and 50 multiple imputed data sets.

	Estimate			CI Coverage %		SD			Avg. Est. SD		$\sqrt{b_m/m}$
	Truth	MDAM	ICIN	MDAM	ICIN	Pre-SD	MDAM	ICIN	MDAM	ICIN	MDAM
(a) $\boldsymbol{\alpha_1 = -2.0, \beta_1 = 0.4, \beta_2 = -0.5}$											
T_{X2}	73451	73685	81569	100	0	1274.3	383.2	1347.4	1641.1	1315.9	147.2
T_{X1}	84985	84966	86193	100	86.6	1355.2	193.0	1370.1	1895.3	1376.5	192.8
$X_2 = 0 \mid X_1 = 0$	.581	.548	.492	55.3	0	.010	.017	.011	.017	.012	.002
$X_2 = 0 \mid X_1 = 1$	.375	.389	.332	97.8	0	.008	.007	.009	.015	.008	.002
$X_1 = 0 \mid X_2 = 0$	.469	.445	.448	89.9	58	.009	.017	.010	.023	.011	.003
$X_1 = 0 \mid X_2 = 1$	.277	.296	.294	46.6	50.8	.008	.008	.008	.009	.009	.001
(b) $\boldsymbol{\alpha_1 = -2.0, \beta_1 = 0.4, \beta_2 = -2.0}$											
T_{X2}	73410	73534	80869	100	0	1286.1	275.4	1349.4	1695.4	1315.7	158.5
T_{X1}	79506	79495	82028	100	54.6	1335.5	171.5	1363.5	1776.0	1319.1	178.9
$X_2 = 0 \mid X_1 = 0$	.617	.549	.491	1.9	0	.010	.016	.011	.016	.012	.002
$X_2 = 0 \mid X_1 = 1$	.336	.380	.333	1.6	94.2	.008	.007	.008	.014	.008	.002
$X_1 = 0 \mid X_2 = 0$	.555	.495	.480	14.2	0	.010	.016	.011	.021	.012	.003
$X_1 = 0 \mid X_2 = 1$	.281	.330	.323	0	0.2	.008	.008	.008	.010	.009	.001
(c) $\boldsymbol{\alpha_1 = -2.0, \beta_1 = 2.0, \beta_2 = -0.5}$											
T_{X2}	73808	73804	78826	100	2.6	1292.5	173.2	1302.2	1724.2	1262.2	170.9
T_{X1}	104446	103010	105932	100	83.6	1547.6	360.0	1548.0	2083.2	1473.7	212.1
$X_2 = 0 \mid X_1 = 0$	.965	.955	.950	84	52.7	.005	.007	.007	.007	.007	.001
$X_2 = 0 \mid X_1 = 1$	.303	.297	.269	100	0	.007	.004	.006	.017	.007	.002
$X_1 = 0 \mid X_2 = 0$	.469	.488	.479	99	89.3	.010	.014	.010	.029	.012	.004
$X_1 = 0 \mid X_2 = 1$	.014	.018	.017	76.5	79.9	.002	.003	.003	.003	.003	$< .001$
(d) $\boldsymbol{\alpha_1 = -2.0, \beta_1 = 2.0, \beta_2 = -2.0}$											
T_{X2}	73582	73572	75242	100	73.2	1283.7	163.3	1277.7	1661.6	1208.5	164.0
T_{X1}	98814	97638	101173	100	61.9	1492.8	381.4	1507.6	1685.2	1415.7	138.7
$X_2 = 0 \mid X_1 = 0$	.961	.942	.949	77.5	69.8	.005	.007	.007	.012	.007	.001
$X_2 = 0 \mid X_1 = 1$	.269	.268	.272	100	94.5	.007	.004	.006	.012	.007	.001
$X_1 = 0 \mid X_2 = 0$	.556	.563	.526	99.2	26.9	.010	.010	.010	.017	.012	.002
$X_1 = 0 \mid X_2 = 1$	.018	.028	.022	70.9	87.6	.002	.003	.003	.006	.003	.001

Table 4.11: Simulation results of parameters in one setting as indicated by first column "truth". For each sampled data sets with missingness, 50 multiple imputed data sets were generated. Algorithm with (MDAM) and without (ICIN) margins are presented side-by-side.

	Estimate			SD			Avg. Est. SD	
	Truth	MDAM	ICIN	Pre-SD	MDAM	ICIN	MDAM	ICIN
α_0	.3	.291	.291	.017	.017	.017	.018	.018
α_1	-.5	-.416	.008	.037	.243	.013	.241	.057
β_0	.2	.146	.146	.027	.028	.028	.030	.030
β_1	.4	.392	.392	.034	.036	.036	.039	.038
β_2	-.5	.010	.052	.034	.267	.012	.316	.058
γ_0	-1.05	-1.06	-1.06	.024	.024	.024	.026	.026
γ_1	.2	.211	.211	.029	.029	.029	.032	.032

4.4 Application to CPS Data

The Current Population Survey (CPS), sponsored jointly by the U.S Census Bureau and the U.S. Bureau of Labor Statistics, gathers demographic information. In its November supplement every year, sampled respondents are asked questions regarding voter registration and turnout. CPS is regarded as the most reliable source to study turnout in the U.S., which concerns many politicians and policy makers, due to its large sample size and detailed information collection. Researchers are interested in voter turnout not only from the national level, but also by states and within demographic subgroups, for example turnout among racial and ethnic minority groups (Ansolabehere et al., ming).

In spite of its advantages, CPS suffers from missing values. As mentioned in section 4.3.3, CPS adjusts weights of respondents to account for unit nonrespondents. For item nonresponse, CPS uses three imputation methods: relational imputation, longitudinal edits and hot deck allocation. CPS treats "Don't Know", "Refused", and "No Response" as not voting. Hur and Achen (2013) attributes the discrepancy between CPS 2008 turnout estimate and state vote counts to the Census imputation procedures. Therefore, appropriate handling of missing values in CPS data is essential for estimating voter turnout. We now demonstrate how to use the MD-AM framework to incorporate auxiliary information and generate multiple imputed data sets for analysis of voter turnout, especially in subgroups.

4.4.1 Data

We obtain CPS 2018 data for North Carolina from the IPUMS website (https://cps.ipums.org/cps/), and select three demographic variables - gender, race, and age - besides vote. Gender and vote are binary variables. Race is a categorical variable with four levels: white alone, black alone, Hispanic or Latino, and the rest. Age is a categorical variable with seven levels: $(0, 29]$, $(29, 39]$, $(39, 49]$, $(49, 59]$, $(59, 69]$, $(69, 79]$, and > 79.

Since we only have unit nonresponse rate at the houselevel from CPS data, we follow (Akande et al., 2021) and carry out the following steps in order to obtain unit nonresponse rate of persons. After referring to the household sampling data file for the reasons why sampled households failed to respond, we exclude “Type C” non-interview households marked as ineligible for the CPS survey. After getting the number of nonrespondents among eligible households, we estimate the average number of adult citizens per household. We use the Census Bureau's special tabulation of the Citizen Voting Age Population (CVAP) using 5-year American Community Survey (ACS) data by state and divided by the total number of households in North Carolina. Then, we multiply the number of eligible nonresponse household by the average number of adult citizens per household and round the result to the nearest person to get the unit nonresponse rate at the person-level.

After the pre-processing, there are 2926 observations in total, 913 unit nonrespondents and 2013 respondents. The percentages of missingness are shown in Table 4.12.

Table 4.12: Unit and item nonresponse rates. There are two observations with only weights not missing, we treat them as unit nonresponse with known weights.

	Unit	Item			
		Vote	Gender	Race	Age
NC	0.31	0.18	0	0.02	0.04

There are two observations with everything missing except weights. We treat them as unit nonrespondents, without modifying their weights. We use the total ballots counted voter-eligible population (VEP) turnout rate of North Carolina in the 2018 election as an auxiliary margin on voter turnout, which is 49% for North Carolina. For gender and race, we utilize auxiliary margins from the 2018 American Community Survey (ACS) by state. The margin for female in North Carolina is 52%. The margin for white alone is 69.9%, for black alone is 21.8%, for Hispanic is 3.9%, and for the rest is 4.4% in North Carolina. The standard errors of these statewide counts from the ACS are small enough that we can treat the survey estimates as population margins.

4.4.2 Model

Let A_i denote age, G_i denote gender, E_i denote race/ethnicity, V_i denote vote, and U_i denote unit nonresponse for $i = 1, \ldots, n$ individuals in the data. We use this notation instead of X_{ik} and Y_{ik} so that it is easier to identify the variables they represent. Note that only $A \notin \mathcal{A}$ and $G, E, V \in \mathcal{A}$ because we have auxiliary margins on them. Let R_i^A denote item nonresponse for age, R_i^E denote item nonresponse for race/ethnicity, and R_i^V denote item nonresponse for vote for $i = 1, \ldots, n$ individuals.

In terms of model specification, we first write down the joint distribution without using auxiliary margins. We write down models for item nonresponse using the selection model parameterization as illustrated in section 4.2. In the data, it turns out that whenever age suffers from item nonresponse, vote also suffers from item nonresponse. As a result, we cannot allow R_A to depend on vote, as there is not sufficient information to identify a coefficient of vote in a model for R_A. We let R_V be 1 when R^A is 1 to reflect the missingness pattern we observe in the data regarding vote and age. We add interactions between gender and race/ethnicity as well as gender and age in the model for vote as we are interested in voter turnout in these subgroups. For the first stage of the MD-AM framework, the model we use is as follows.

$$U_i \sim Bern(\pi_i^U),\, logit(\pi_i^U) = \nu_0 \tag{4.31}$$

$$G_i|- \sim Bern(\pi_i^G),\, logit(\pi_i^G) = \alpha_0^G \tag{4.32}$$

$$E_i|- \sim Multinom(\mathbb{P}[E_i = e]),\, logit(\mathbb{P}(E_i = e)) = \alpha_{0,r}^E + \alpha_{1,r}^E G_i \tag{4.33}$$

$$\begin{aligned} A_i|- \sim Multinom(\mathbb{P}[A_i = a]),\, logit(\mathbb{P}[A_i = a]) &= \alpha_{0,a}^A + \alpha_{1,a}^A G_i \\ &+ \alpha_{2,e}^A I[E_i = e] \end{aligned} \tag{4.34}$$

$$\begin{aligned} V_i|- \sim Bern(\pi_i^V),\, logit(\pi_i^V) &= \alpha_0^V + \alpha_1^V G_i + \alpha_{2,e}^V I[E_i = e] + \alpha_{3,a}^V I[A_i = a] \\ &+ \alpha_{4,e}^V I[G_i = 1, E_i = e] + \alpha_{5,a}^V I[G_i = 1, A_i = a] \end{aligned} \tag{4.35}$$

$$R_i^E|- \sim Bern(\pi_i^E),\, logit(\pi_i^E) = \gamma_0^E + \gamma_1^E G_i + \gamma_{2,a}^E I[A_i = a] + \gamma_{3,e}^E V_i \tag{4.36}$$

$$R_i^A|- \sim Bern(\pi_i^A),\, logit(\pi_i^A) = \gamma_0^A + \gamma_1^A G_i + \gamma_{2,e}^A I[E_i = e] \tag{4.37}$$

$$R_i^V = 1 | R_i^A = 1 \tag{4.38}$$

$$R_i^V | R_i^A = 0, - \sim Bern(\pi_i^V),\, logit(\pi_i^V) = \gamma_0^V + \gamma_1^V G_i + \gamma_{2,a}^V I[A_i = a] + \gamma_{3,e}^V I[E_i = e]. \tag{4.39}$$

Here, $(\cdot|-)$ denotes conditioning on other variables. We suppress the conditioning to save space. $I[\cdot]$ is the indicator function. We specify less parameters than the maximum allowed under MD-AM framework in theory in this step to simplify the modeling somewhat. Then we follow step 2 of MD-AM framework and add terms related to gender, race/ethnicity, and vote. Since we do not have an auxiliary margin on age, the model for A_i does not depend on U_i. We use a pattern mixture model to write down the joint distribution of unit nonresponses and other variables as follows:

$$G_i|- \sim Bern(\pi_i^G),\ logit(\pi_i^G) = \alpha_0^G + \alpha_1^G U_i \tag{4.40}$$

$$E_i|- \sim Multinom(\mathbb{P}[E_i = e]),\ logit(\mathbb{P}(E_i = e)) = \alpha_{0,r}^E + \alpha_{1,r}^E G_i + \alpha_{2,r}^E U_i \tag{4.41}$$

$$\begin{aligned} V_i|- \sim Bern(\pi_i^V),\ logit(\pi_i^V) &= \alpha_0^V + \alpha_1^V G_i + \alpha_{2,e}^V I[E_i = e] + \alpha_{3,a}^V I[A_i = a] \\ &+ \alpha_{4,e}^V I[G_i = 1, E_i = e] + \alpha_{5,a}^V I[G_i = 1, A_i = a] + \alpha_6^V U_i. \end{aligned} \tag{4.42}$$

We use the auxiliary margin of vote to add a U term in the model for vote rather than adding vote in the model for R^V because unit nonresponse rate is roughly two times of item nonresponse rate in vote. If one has strong reason to believe that item nonresponse in vote is nonignorable, using the alternative model to fit the data can be done as a sensitivity analysis.

The CPS file from IPUMS does not include the design weights. Rather, we only have weights adjusted for unit nonresponse. Thus, we apply the algorithm in (4.19) to create the analysis weights. Then, we employ intercept matching algorithm and fit the models consisting of (4.31), (4.34), (4.40) - (4.42), (4.36) - (4.39) using Bayesian MCMC with non-informative priors for all parameters. We run the MCMC sampler for 10,000 iterations and discard the first 5000 iterations as burn-in. The sampler converges judging from the traceplots of the model parameters. Among the 5000 posterior samples obtained, we retrieve every 100th posterior sample to create $L = 50$ multiple imputed data sets and do inference based on those completed data sets.

4.4.3 Results

Table 4.13 displays the estimated marginal distributions of gender, race, and age from 50 multiple imputed data sets generated under MD-AM. For comparisions, we use MICE to generate 50 multiple imputed data sets using only the unit respondents and their CPS adjusted weights on the IPUMS file. We also include weighted and unweighted estimates based on complete cases where we only look at item respondents. For the MD-AM model, we present standard deviations calculated from the multiple imputation combining rules as well as $\sqrt{b_m/m}$. Estimates of marginal probabilities under the MD-AM model match available auxiliary margins slightly more closely than those from MICE and complete case analysis. However, for gender and race where we have auxiliary margins, the true standard deviation should be smaller than those from MICE, closer to $\sqrt{b_m/m}$.

The benefits of handling item and unit nonresponse with the MD-AM framework become apparent when we examine the estimated proportion of voter turnout in subgroups. Those results are shown in Table 4.14. We include Census Bureau estimates of subgroup voter turnout for comparison as well. These voter turnouts across gender are taken directly from the Census Bureau's website under "Voting and Registration in the Election of November 2018" section (https://www.census.gov/data/tables/time-series/demo/voting-and-registration/p20-583.html). As we have different groupings for race/ethnicity and age, we do the following steps to obtain Census Bureau numbers reported in Table 4.14. We refer back to the original CPS data with unadjusted weights and unit respondents alone. We use the imputations for gender, race/ethnicity and age by Census Bureau, and code missingness in vote as not voting (Hur and Achen, 2013). Then, we use the "survey" package in R to obtain survey weighted estimates and uncertainty quantification.

Overall, the MD-AM framework estimates around 49.7% people are voters, around 10% lower than what MICE and complete cases analysis estimate. The true vote rate of North Carolina is 49%, close to the result of MD-AM imputed data sets. Also, estimates of voter turnout in all the subgroups under the MD-AM framework are smaller than those under MICE as well as weighted and unweighted complete case analysis, signifying that assuming nonignorable missingness leads to different estimates. The MD-AM framework shrinks voter turnout from unrealistic high rates under MICE and complete case anaysis to reasonable estimates. The MD-AM framework and Census Bureau's approach are similar in the sense that both of them use auxiliary margins. Hence, it is not surprising that estimates under MD-AM framework are close to albeit slightly lower than the Census Bureau's estimates by state. However, the results from the MD-AM framework have a model-based justification whereas the Census Bureau results make the relatively arbitrary assumption that all the individuals missing vote are non-voters. We also present voter turnout in more detailed subgroups, including gender by race, and gender by age, in Table 4.15. We observe similar pattern as in Table 4.14, where estimates under MD-AM framework are much smaller and more reasonable than under MICE and weighted and unweighted complete case analysis.

We cannot assess whether the nonresponse modeling assumptions are correct. However, we can assess how well the MD-AM model reproduces the observed data. To do so, we compare completed data with replicates generated from posterior predictive samples (He and Zaslavsky, 2012). Namely, we examine unweighted estimates of marginal and conditional probabilities for $\{(A_i, G_i, E_i, V_i) : U_i = 0, R_i^A = 0, R_i^E = 0, R_i^V = 0\}$. Using the 5000 draws of the parameters, we generate new values of all variables for all n units in the database. We construct 95% posterior predictive intervals for marginal probabilites of vote, gender, race/ethnicity, and age as well as subgroup voter turnout probabilities based on 5000 posterior samples. In Figure 4.1, the first panel shows marginal probabilities of vote, gender, race/ethnicity and age. The second panel shows voter turnout probabilites in various gender, race/ethnicity and age subgroups. The third panel displays voter turnout probabilities in gender crossed with race/ethnicity and gender crossed with age subgroups. All posterior predictive intervals contain the observed data estimates, giving us more confidence to believe the reasonableness of the MD-AM model, at least for the observable portion of the data.

As we have seen in the simulation studies in section 4.3.3, when design weights for unit respondents are not available, the performance of the MD-AM model in estimating voter turnout should be affected by the strength of unit nonresponse bias in vote as well as the other variables. Similarity of marginal probabilites of gender, race/ethnicity, and age estimated from MD-AM, MICE and complete case analysis in Table 4.13 suggest that, after weighting adjustments, the unit nonresponse biases in these three variables are not terrible. On the contrary, voter turnout estimated under MD-AM is much lower than MICE and complete case analysis. The unit nonresponse bias in vote is significant. When analyzing the CPS data, we are likely in the scenario in Table 4.9 (b) or (d) where results under MD-AM are more accurate than under ICIN. This also increases our preference of the MD-AM model over the MICE model.

4.5 Future Work

In this dissertation, we assume that there is no measurement error in the CPS data, and the discrepancy between auxiliary information and the observed data is due to nonresponse. However, research (Enamorado and Imai, 2019; Ansolabehere et al., ming) has found that overreporting of vote propensity also contributes to the bias since CPS is a self-reporting survey. Future direction of this research could explore

Table 4.13: Results from the CPS data analysis. Margins of gender, age, and race analyzed with 50 imputations generated via MD-AM framework and MICE. "Wgt" stands for weighted. "Unwgt" stands for unweighted. The auxiliary margin for female in North Carolina is .52, for White is .699, for Black is .218, for Hispanic is .039, for rest is .044. We do not have the population marginal distribution of age.

	MD-AM			MICE		Complete Cases	
	Prop	SD	$\sqrt{b_m/m}$	Prop	SD	Wgt	Unwgt
Male	.477	.015	.002	.470	.011	.466	.462
Female	.523	.015	.002	.530	.011	.534	.538
White	.699	.015	.002	.700	.011	.709	.743
Black	.218	.013	.001	.222	.010	.213	.175
Hispanic	.039	.006	.001	.035	.004	.033	.034
Rest	.044	.007	.001	.049	.005	.050	.048
$(0, 29]$	.206	.010	-	.215	.010	.205	.181
$(29, 39]$	.139	.008	-	.141	.008	.144	.139
$(39, 49]$	.161	.008	-	.162	.008	.150	.163
$(49, 59]$	.176	.009	-	.175	.009	.177	.181
$(59, 69]$	.165	.009	-	.161	.008	.168	.173
$(69, 79]$	.107	.007	-	.103	.007	.111	.114
> 79	.046	.005	-	.043	.004	.045	.048

incorporating measurement error caused by and underreporting and/or overreporting of vote in subgroups on top of what have been studied in this dissertation. We could employ similar strategy as used in (Schifeling et al., 2019), which leverages auxiliary information in data fusion for correcting measurement errors. We could view the observed data as an error-prone version of the error-free value of variable of interest. Conditional on the value of the truth and some covariates, we either observe the truth or the erroreous version of it. As simple example, Enamorado and Imai (2019) report that at least 30% of the nonvoters answer that they have voted in the election. Given the participant has not voted, there is a 30% chance that the participant did not tell the truth. We can again use auxiliary margins to help with missing data imputation, but instead of matching with the observed data, we match with the underlying truth. If the measurement error model could be added, we are able to study to what extent

Table 4.14: Proportion of "voted" in each subgroup of the CPS data (from top to bottom: gender, race, age), analyzed with 50 multiple-imputations via MD-AM framework and MICE. "Wgt" stands for weighted. "Unwgt" stands for unweighted. Results for complete case analyses and estimates released by the Census Bureau are also included. The auxiliary margin for "voted" in North Carolina is .49. As shown in simulation study, $\sqrt{b_m/m}$ tends to be a closer estimate for total/proportion when we have the particular margin available. $\sqrt{b_m/m}$ for proportion of "voted" is .002.

	MD-AM		MICE		Complete Cases		Census Bureau	
	Prop	SD	Prop	SD	Wgt	Unwgt	Prop	SD
Full	.497	.017	.633	.013	.637	.648	.524	.012
Male	.495	.023	.635	.018	.639	.651	.521	.017
Female	.498	.020	.631	.017	.636	.645	.526	.016
White	.505	.020	.646	.014	.648	.656	.544	.013
Black	.515	.033	.653	.031	.658	.679	.516	.027
Hispanic	.375	.062	.495	.071	.489	.510	.386	.061
Rest	.394	.061	.458	.058	.467	.493	.371	.049
$(0, 29]$	.324	.025	.423	.031	.421	.418	.331	.026
$(29, 39]$	.437	.033	.570	.034	.571	.573	.466	.031
$(39, 49]$	.527	.034	.676	.031	.681	.673	.517	.028
$(49, 59]$	.543	.032	.698	.026	.700	.701	.596	.026
$(59, 69]$	.640	.039	.795	.023	.794	.801	.664	.025
$(69, 79]$	.606	.042	.762	.030	.761	.762	.691	.031
> 79	.408	.055	.550	.056	.556	.560	.485	.051

the discrepancy is caused by nonresponse or false reporting.

Table 4.15: Proportion of "voted" in each subgroups of gender crossed with race/age of the CPS data, analyzed with 50 multiple-imputations generated via MD-AM framework and MICE. "Wgt" stands for weighted. "Unwgt" stands for unweighted. Results for complete case analyses and estimates released by the Census Bureau are also included.

	MD-AM		MICE		Complete Cases		Census Bureau	
	Prop	SD	Prop	SD	Wgt	Unwgt	Prop	SD
Male, White	.508	.028	.651	.021	.650	.661	.550	.040
Male, Black	.469	.044	.623	.045	.618	.629	.463	.018
Male, Hispanic	.450	.101	.568	.101	.597	.625	.443	.086
Male, Rest	.460	.094	.522	.086	.552	.576	.416	.074
Female, White	.502	.023	.642	.019	.647	.651	.539	.018
Female, Black	.554	.045	.677	.041	.687	.718	.559	.037
Female, Hispanic	.302	.081	.428	.095	.401	.407	.329	.085
Female, Rest	.334	.070	.401	.076	.397	.425	.332	.066
Male, $(0, 29]$	.311	.035	.413	.043	.408	.409	.317	.036
Male, $(29, 39]$	.439	.048	.576	.049	.573	.570	.460	.045
Male, $(39, 49]$	.488	.049	.646	.046	.639	.636	.482	.042
Male, $(49, 59]$	.529	.046	.698	.037	.689	.689	.596	.037
Male, $(59, 69]$	.649	.052	.810	.035	.810	.818	.656	.037
Male, $(69, 79]$	.685	.058	.835	.039	.840	.846	.781	.042
Male, > 79	.411	.085	.559	.092	.565	.567	.469	.083
Female, $(0, 29]$	.337	.036	.431	.042	.432	.427	.344	.037
Female, $(29, 39]$	.435	.044	.565	.046	.569	.575	.472	.042
Female, $(39, 49]$	.561	.044	.701	.037	.714	.702	.545	.037
Female, $(49, 59]$	.557	.041	.708	.037	.711	.712	.596	.037
Female, $(59, 69]$	.632	.049	.781	.033	.781	.788	.671	.035
Female, $(69, 79]$	.539	.052	.702	.043	.694	.694	.619	.044
Female, > 79	.408	.068	.545	.069	.551	.556	.494	.065

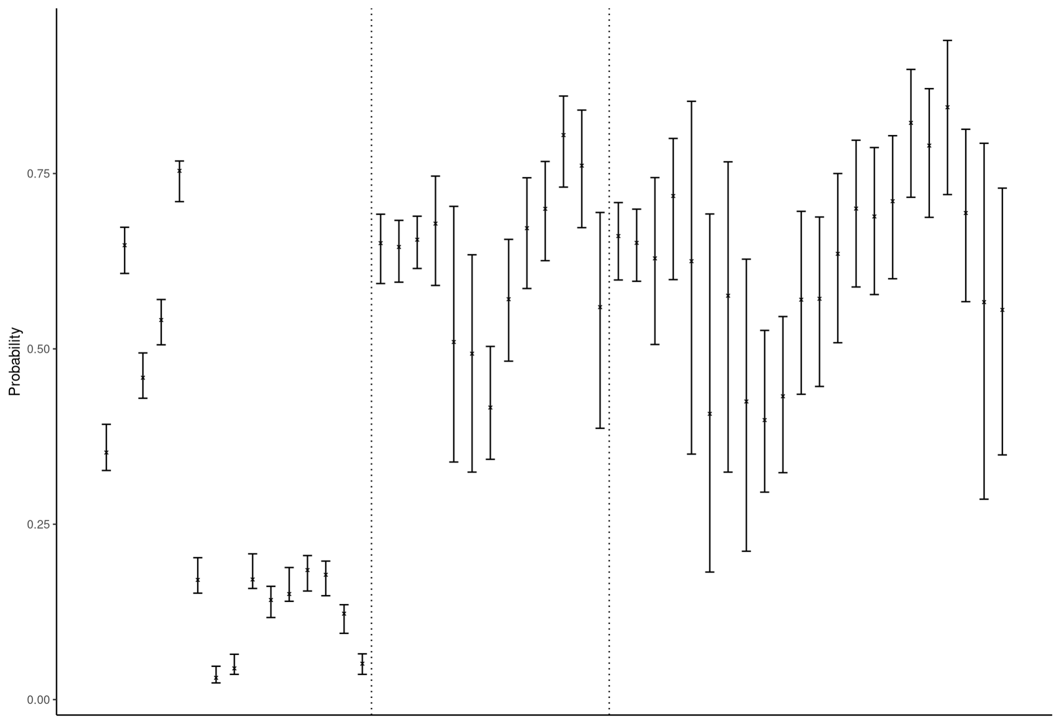


Figure 4.1: Posterior predictive intervals for all marginal as well as conditional probabilities involving vote for completed data only. Crosses are observed data estimates. Cross represents observed data estimates.

Bibliography

Akande, O. (2019). *Bayesian Models for Imputing Missing Data and Editing Erroneous Responses in Surveys*. PhD thesis, Duke University.

Akande, O., Li, F., and Reiter, J. P. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, 71:162–170.

Akande, O., Madson, G., Hillygus, D. S., and Reiter, J. P. (2021). Leveraging auxiliary information on marginal distributions in nonignorable models for item and unit nonresponse in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184:643–662.

Akande, O. and Reiter, J. P. (Forthcoming). Multiple imputation for nonignorable item nonresponse in complex surveys using auxiliary margins. In Carriquiry, A. L., Tanur, J. M., and Eddy, W. F., editors, *Statistics in the Public Interest: In Memory of Stephen E. Fienberg*. Springer, Berlin.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679.

Ansolabehere, S., Fraga, B. L., and Schaffner, B. F. (Forthcoming). The CPS voting and registration supplement overstates minority turnout. *Journal of Politics*.

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

Barnard, J. and Meng, X.-L. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods in Medical Research*, 8:17–36.

Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 24:1537–1555.

Christen, P. (2019). Data linkage: The big picture. *Harvard Data Science Review*, 1(2).

Dalzell, N. M. and Reiter, J. P. (2018). Regression modeling and file matching using possibly erroneous matching variables. *Journal of Computational and Graphical Statistics*, 27:728–738.

Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y., and Zheng, S. (2013). Handling attrition in longitudinal studies: The case for refreshment samples. *Statistical Science*, 28:238–256.

Enamorado, T. and Imai, K. (2019). Validating self-reported turnout by linking public opinion surveys with administrative records. *Public Opinion Quarterly*, 83:723–748.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64:1183–1210.

Fortini, M., Liseo, B., Nuccitelli, A., and Scanu, M. (2001). On Bayesian record linkage. *Research in Official Statistics*, 4:185–198.

Fuller, W. A. (2009). *Sampling Statistics*. John Wiley & Sons, Inc, New York.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples*, pages 115–142. Springer, New York, NY.

Gutman, R., Afendulis, C. C., and Zaslavsky, A. M. (2013). A Bayesian procedure for file linking to analyze end-of-life medical costs. *Journal of the American Statistical Association*, 108:34–47.

Hausman, J. A. and Wise, D. A. (1979). Attrition bias in experimental and panel data: The Gary income maintenance experiment. *Econometrica*, 47:455–473.

He, Y. and Zaslavsky, A. M. (2012). Diagnosing imputation models by applying target analyses to posterior replicates of completed data. *Statistics in Medicine*, 31:1–18.

Hirano, K., Imbens, G. W., Ridder, G., and Rubin, D. B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69:1645–1659.

Hof, M. H., Ravelli, A. C., and To, A. H. Z. (2017). A probabilistic record linkage model for survival data. *Journal of the American Statistical Association*, 112:1504–1515.

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

Hur, A. and Achen, C. H. (2013). Coding voter turnout responses in the current population survey. *Public Opinion Quarterly*, 77:985–993.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84:414–420.

Larsen, M. D. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96:32–41.

Larsen, M. D. (2002). Comments on hierarchical Bayesian record linkage. In *Proceedings of the Section on Survey Research Methods*, pages 1995–2000. American Statistical Association, Alexandria, VA.

Larsen, M. D. (2005). Advances in record linkage theory: Hierarchical Bayesian record linkage theory. In *Proceedings of the Section on Survey Research Methods*, pages 3277–3284. American Statistical Association, Alexandria, VA.

Larsen, M. D. and Rubin, D. B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96:32–41.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125–134.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90:1112–1121.

Lohr, S. L. (2010). *Sampling: Design and Analysis. 2nd Edition.* Brooks/Cole, Boston.

Lumley, T. (2021). *Analysis of Complex Survey Samples.* R (>= 3.5.0), grid, methods, Matrix, survival.

Marchant, N. G., Steorts, R. C., Kaplan, A., Rubinstein, B. I., and Elazar, D. N. (2021). d-blink: Distributed end-to-end Bayesian entity resolution. *Journal of Computational and Graphical Statistics*, 30:406–421.

Mealli, F. and Rubin, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, 102:995–1000.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959). Automatic linkage of vital records: Computers can be used to extract "follow-up" statistics of families from files of routine records. *Science*, 130:954–959.

Pham, T. M., Carpenter, J. R., Morris, T. P., Wood, A. M., and Petersen, I. (2019). Population-calibrated multiple imputation for a binary categorical covariate in categorical regression models. *Statistics in Medicine*, 38:792–808.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, 108:1339–1349.

Potter, F. J. and Zheng, Y. (2015). Methods and issues in trimming extreme weights in sample surveys. JSM2015-Survey Research Methods Section.

Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values. *Survey Methodology*, 27:85–95.

Reiter, J. P. (2008). Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika*, 95:933–946.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102:1462–1471.

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

Rubin, D. B. (1987). *Multiple Imputation For Nonresponse In Surveys*. John Wiley & Sons, New York.

Rubin, D. B. (1996). Multiple Imputation After 18+ years. *Journal of the American Statistical Association*, 91:473–489.

Sadinle, M. (2014). Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *The Annals of Applied Statistics*, 8:2404–2434.

Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112:600–612.

Sadinle, M. and Fienberg, S. E. (2013). A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record systems. *Journal of the American Statistical Association*, 108:385–397.

Sadinle, M. and Reiter, J. P. (2017). Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika*, 104:207–220.

Sadinle, M. and Reiter, J. P. (2019). Sequentially additive nonignorable missing data modelling using auxiliary marginal information. *Biometrika*, 106:889–911.

Scheuren, F. and Winkler, W. E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19:39–58.

Scheuren, F. and Winkler, W. E. (1997). Regression analysis of data files that are computer matched, part ii. *Survey Methodology*, 23:157–165.

Schifeling, T. A., Cheng, C., Reiter, J. P., and Hillygus, D. S. (2015). Accounting for nonignorable unit nonresponse and attrition in panel studies with refreshment samples. *Journal of Survey Statistics and Methodology*, 3:265–295.

Schifeling, T. A., Reiter, J. P., and DeYoreo, M. (2019). Data fusion for correcting measurement errors. *Journal of Survey Statistics and Methodology*, 7:175–200.

Steorts, R. C. (2015). Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10:849–875.

Steorts, R. C., Hall, R., and Fienberg, S. E. (2016). A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111:1660–1672.

Tancredi, A. and Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5:1553–1585.

Tang, J., Reiter, J. P., and Steorts, R. C. (2020). Bayesian modeling for simultaneous regression and record linkage. In *International Conference on Privacy in Statistical Databases*, pages 209–223. Springer.

Valliant, R., Dever, J. A., and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Sample Surveys*. Springer-Verlag, New York.

Winkler, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, pages 354–359. American Statistical Association, Alexandria, VA.

Winkler, W. E. (2006). Overview of record linkage and current research directions.

Technical Report Statistics #2006-2, U.S. Bureau of the Census.

Winkler, W. E. (2014). Matching and record linkage. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6:313–325.

Zanella, G., Betancourt, B., Wallach, H., Miller, J., Zaidi, A., and Steorts, R. C. (2016). Flexible models for microclustering with application to entity resolution. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1425–1433, NY, USA. Curran Associates Inc.

Zhao, B., Rubinstein, B. I. P., Gemmell, J., and Han, J. (2012). A Bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment*, 5:550–561.

Biography

Jiurui Tang received a scholarship from the Ministry of Education of Singapore and completed her high school at the NUS High School of Math and Science. After that, she attended college at the University of California, Los Angeles, and graduated with summa cum laude in 2017, double-majoring in statistics and math-economics. She plans to graduate with her Ph.D. in Statistical Science in May 2022, under the supervision of Professor Jerome P. Reiter. She also plans to graduate with a Certificate in College Teaching program in May 2022.