Missing Data Imputation for Voter Turnout Using Auxiliary Margins

by

Yangfan Ren

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Advisor

_____
Fan Li

_____
D. Sunshine Hillygus

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2020

Abstract

Missing Data Imputation for Voter Turnout Using Auxiliary Margins

by

Yangfan Ren

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Jerome P. Reiter, Advisor

_____
Fan Li

_____
D. Sunshine Hillygus

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science in the Graduate School of Duke University
2020

Abstract

Missing data is one of the essential problems in most data analysis. Typically, researchers are forced to make strong assumptions or constraints to handle missing data such as assuming the data are missing at random. However, these assumptions are generally improvable and could introduce bias when data are missing systematically. Under such circumstances, it is desirable to consider nonignorable missing data mechanisms. The missing data with auxiliary margins (MD-AM) framework proposed by Akande et al. (2019) provides a flexible method to characterize nonignorable missing models by combining auxiliary margins. Previous research applied the MD-AM framework on CPS voter turnout data with few variables. In this thesis, I apply the MD-AM framework on turnout data with nine primary variables. By changing the assumptions about how vote affects missingness, I specify models for each primary variables, their associated item missing indicators and unit nonresponse indicator. I illustrate sensitivity check and compare results.

To my parents who supported me in all things great and small.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to first thank my advisor, Jerry Reiter, who inspired me with this research and always guided me with a lot of patience. He is the best mentor I have ever met. Thanks to him for all the help in my graduate study, master's thesis and for the chance of being a teacher assistant for the interesting class. Thanks to D. Sunshine Hillygus and Gabriel Madson for their support and insight for this thesis. To Olanrewaju Michael Akande for his time and guidance. I also acknowledge the support from the National Science Foundation NSF (SES-1733835).

Thanks to Prof. David Siegel for providing me the precious opportunity to collaborate with such a wonderful team. Thank everyone in the Department of Statistical Science and Duke University who helped me with every little thing during these two years. This is a treasured and unforgettable experience in my life.

Special thank to my family for providing me with full support and continuous encouragement throughout these years.

1

Introduction

Missing data are unavoidable in research related to data analysis especially in large scale surveys. Response rates have dropped among many surveys due to declining study participation rates (Galea and Tracy, 2007). A lot of missing data problems arise because of nonresponse, i.e., respondents may not follow up longitudinal surveys after several periods, refusals to participate in a survey or unable to contact sampling individuals (unit nonresponse), or people do not provide acceptable answers for certain questions (Brand, 1999). There are also other facts leading to missing data such as records lost or discarded during the storing procedure. Mishandling missing values not only harms the quality of data but also casts doubt on the total analysis process based on the data.

Weight adjustment approaches are commonly used in survey estimates to reduce bias due to missingness (Brick and Kalton, 1996). A variety of ad hoc methods also have been proposed over the years including complete cases analysis, substituting missing values with observed data (for example, mean substitution) and regression based single imputation. Most of the time, these methods are not statistically valid (Stern, White, Carlin et al., 2009) which can cause serious bias if missingness is

systematic and can jeopardize the representativeness of the data. Furthermore, loss of information could result in loss of statistical power, e.g., large standard errors.

Imputation methods have been developed recent years. For example, hot deck imputation which replaces the missing data with values from respondents similar to the non-respondents based on the observed records (Andridge and Little, 2010) and multiple imputation which imputes multiple likely values for each missing entry (Rubin, 1987). However, researchers often apply constraints on the missingness mechanisms to generate inferences with respect to the observed data – for example, the values are missing at random (MAR) (Rubin, 1976). Without conclusive knowledge that these assumptions are proper, people can make implausible inferences based on that.

Under circumstances that are reasonable to consider nonignorable missing mechanisms, models could be unidentifiable with the observed data alone. Thus, it is desirable and reasonable to incorporate auxiliary information. Leveraging auxiliary information in inference models can adjust the bias from original collected data and provide additional information to construct nonignorable framework. For instance, suppose the true voter turnout rate in a state is 65%, but we have 75% total turnout rate among the respondents in the dataset. Then, the nonrespondents are more likely to be nonvoters. Without the true turnout rate, it would be difficult to account for the bias. Auxiliary information can be accessible from diverse data sources. For example, population margins for variables such as gender proportion are available on national censuses reports. Administrative databases like tax or medical records, as well as large national surveys, can provide potentially relevant margins for certain subgroups (Sadinle and Reiter, 2019). In general, one type of auxiliary information is sample-specific information which can link to individual responses; the other type is population-specific information such as marginal distributions of variables for certain groups in a survey (Akande et al., 2019). In this thesis, I work with population-specific information from the United States decennial census in 2010.

The imputation models specified in this thesis are based on the missing data with auxiliary margins (MD-AM) framework (Akande et al., 2019). The MD-AM framework is built on the additive nonignorable (AN) model of Hirano et al. (2001). The AN model weakens the unprovable assumptions about missing data mechanisms and specifies identifiable models by using additional data. This model has been applied to other research in longitudinal studies, e.g., Nevo (2003), Deng et al. (2013) and Schifeling et al. (2015). The MD-AM framework extends the AN model into more flexible settings with both unit nonresponse and item nonresponse as well as more than one marginal distributions available. Analysts can decide how to use auxiliary margins in the imputation models by specifying different missing mechanisms. For example, researchers can leverage the auxiliary margin to build either a nonignorable unit nonresponse model or a nonignorable item nonresponse model.

This thesis applies the MD-AM framework on Current Population Survey (CPS) data to estimate voter turnout. The CPS is one of the most accredited national surveys. It is one of the primary source for statistics on individuals and society. For example, the CPS collects data on demographic information (e.g., age, gender, race), labor force statistics, education, family income and election statistics. Nonetheless, the inevitable existence of missing data may induce biased estimates. Akande et al. (2019) has implemented the MD-AM imputation method on CPS data with four variables (age, sex, state and vote) and auxiliary data for age, sex and vote. However, more features than those considered by Akande et al. (2019) are related to voter turnout. Thus, this thesis applies the MD-AM framework using more variables of CPS data and incorporates more margins at the same time. To implement this application, I specify two chained models differing in the missing mechanism of vote. In each model, all variables and missing indicators – including a unit nonresponse indicator and item nonresponse indicators - are assigned a specific regression model. The estimation based on the chained models are achieved by Gibbs sampling within

a Bayesian framework. I implement Bayesian model checking along with multiple imputation (Rubin, 1987) to account for uncertainty for estimation among various subgroups and the overall population, which I use to compare models.

The remainder of this thesis is organized as follows. In chapter 2, I review related background knowledge including missing data mechanisms, multiple imputation, the AN model and the MD-AM framework. In Chapter 3, I describe the modeling for 2012 CPS data application with different assumptions. In Chapter 4, I present modeling results and uncertainty, as well as empirical analyses of voter turnout estimation. In Chapter 5, I present conclusions and a discussion.

2

Background

2.1 Multiple Imputation

In this thesis, I apply the framework of multiple imputation first introduced by Rubin (1987). Multiple imputation contains the following procedures. First, produce m completed data sets instead of one. Each completed data set contains the same observed data, and missing values imputed by sampling from their predictive distributions conditioning on the observed data. Schafer and Olsen (1998) suggested to accept $m = 3$ to $m = 5$, but later experiments show that more imputations ($m = 20$, even $m = 100$) are required to reach useful inferences under some circumstances (Graham, Olchowski and Gilreath, 2007). Second, apply standard statistical analysis to each completed data set. Last but not least, combine the multiple estimates and capture the variability across sets using the rules of Rubin (1987). These rules propagate the uncertainty from the missing data.

Multiple imputation holds many advantages. First, it is straightforward to implement and also flexible under different situations. Each completed data set can be treated separately by using common statistical methods. Second, it is statistically

efficient because it uses all available information from a data set and generates draws from a posited distribution representing the data (Rubin, 1987). Third, additional variability from missing values is taken into account. Thus analysts can produce better inferences.

The summary of methods for combining estimations follow the rules from Rubin (1987). Here I use notations consistent with Schafer (1997). First, calculate the point estimates by averaging over m imputed data set. Let Q be the quantity of interest, such as a population mean or regression coefficient. In completed dataset i, let Q_i be the point estimate of Q and let U_i be the variance associated with this estimate. We have

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} Q_i. \tag{2.1}$$

The variance of $\bar{Q}$ consists of within-imputation variance and between-imputation variance. The within-imputation variance is the average of variance over multiple imputations, namely

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} U_i, \tag{2.2}$$

where $\hat{U}_i$ is the variance of Q in completed data set i.

The between-imputation variance is the variability of parameter estimations across multiple procedures. We have

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(Q_i - \bar{Q})^2. \tag{2.3}$$

The total variance is a combination of within and between imputation variances. We have

$$T = \bar{U} + \frac{m}{m+1}B. \tag{2.4}$$

The term $\frac{m}{m+1}$ is used for correction under small m.

The ratio of $\frac{m}{m+1}\bar{B}$ and T reflects the amount of missing information. For small m, a t-distribution is adopted with degrees of freedom,

$$\nu = (m-1)(1+\frac{m}{m+1}\frac{\bar{U}}{B})^2. \tag{2.5}$$

A $100(1-\alpha)\%$ confidence interval is obtained by

$$\bar{Q} \pm t_{\nu,\frac{\alpha}{2}}\sqrt{T}. \tag{2.6}$$

The degrees of freedom are related to the fraction of missing information. Thus, a smaller ν suggests a more accurate estimate. If ν is large, it suggests that more imputations can be beneficial for inferences (Schafer and Olsen, 1998).

2.2 Handling Nonignorable Nonresponses

The default applications of multiple imputation are based on the assumption that the reason for missingness is not related to the missing value itself. However, when this assumption fails, default applications can produce bias. Therefore, specifying a missing data mechanism and applying appropriate model design is helpful to avoid misleading results and assess sensitivity of results to different assumptions.

2.2.1 Missing Data Mechanisms

Let $Y = (y_{ij})$ represent the complete data of n individuals and k variables, with $y_i = (y_{i1}, ..., y_{ik})$ where y_{ij} is the value of jth variable for individual i. Let $Y_{obs} = (y_{1obs}^T, ..., y_{nobs}^T)^T$ and $Y_{mis} = (y_{1mis}^T, ..., y_{nmis}^T)^T$ represent observed parts and missing parts of data, respectively. Then the ith row $y_i = (y_{iobs}, y_{imis})$. We define the missing indicator matrix $R = (r_{ij})$, with $r_{ij} = 1$ when y_{ij} is missing and $r_{ij} = 0$ otherwise. We define $Y_{-j} = (Y_1, ..., Y_{j-1}, Y_{j+1}, .., Y_k)$ and $R_{-j} = (R_1, ..., R_{j-1}, R_{j+1}, .., R_k)$, where Y_j and R_j represent the value and missing indicator for the jth variable, respectively.

Then we can use the distribution of R conditioning on Y to describe the missing data mechanism, as $f(R \mid Y, \phi)$ where ϕ represents the parameters.

We consider three common missing data mechanisms of Rubin (1987) and the itemwise conditionally independent nonresponse mechanism of Sadinle and Reiter (2017).

1. Missing completely at random (MCAR): The missing values are independent of both the observed and missing data, such as flipping a coin before answering a question. That is,

$$f(R \mid Y, \phi) = f(R \mid \phi) \text{ for all } Y, \phi. \tag{2.7}$$

2. Missing at random (MAR): Given observed data, data are missing independently of unobserved data, e.g., male participants refuse to respond to depression questions but it does not depend on their level of depression. That is,

$$f(R \mid Y, \phi) = f(R \mid Y_{obs}, \phi) \text{ for all } Y_{mis}, \phi. \tag{2.8}$$

3. Missing not at random (MNAR): The distribution of missingness depends on the missing values. For example, people who do not vote are less likely to respond to the question about voting. That is,

$$f(R \mid Y, \phi) = f(R \mid Y_{obs}, Y_{mis}, \phi) \text{ for all } Y, \phi. \tag{2.9}$$

4. Itemwise conditionally independent nonresponse (ICIN): Given other items and missingness indicators, the missingness of each item is independent of the value itself. That is,

$$Y_j \perp\!\!\!\perp R_j \mid Y_{-j}, R_{-j} \quad (j = 1, ..., k). \tag{2.10}$$

Here, we can assume that the missing data mechanism is ignorable under MCAR and MAR, while it is nonignorable unber MNAR and ICIN.

Table 2.1: Data structure for an example of the AN model.

Observations		X	Y_1	Y_2	R
N_p	N_{p1}	✓	✓	✓	0
	N_{p2}	✓	✓	?	1
N_r		✓	?	✓	?

2.2.2 The AN Model

The AN model was developed originally for panel (longitudinal) data with refreshment samples. This section reviews the AN model using a simple two wave-panel with N_p original samples and N_r refreshment samples. Let $\mathbf{X} = (\mathbf{X_1}, ..., \mathbf{X_q})$ represent q time invariant variables observed in both the first and the second wave. Let Y_1 be a binary variable in wave 1, and assume no missing exists in Y_1. Let Y_2 be a binary variable measured in wave 2. Among the N_p individuals, N_{p1} individuals provide Y_2 in wave 2; for these individuals, we set $R_i = 0$. The remaining $N_{p2} = N_p - N_{p1}$ individuals drop out of the panel and thus generate missing data in wave 2. These individuals have $R_i = 1$. The N_r refreshment samples only have observations of $\mathbf{X}$ and Y_2; Y_1 and R are missing since they do not participate in wave 1. The data structure is shown in detail in Table 2.1, where the "✓" represents observed data and the "?" represents missing data.

Since $\mathbf{X}$ is observable for all individuals, we can specify the joint distribution of (Y_1, Y_2, R) as a pattern mixture factorization (Glynn, Laird and Rubin, 1986) with $y_1, y_2, r \in \{0, 1\}^3$. We have

$$\begin{aligned} Pr(Y_1 = y_1, Y_2 = y_2, R = r) =& Pr(Y_2 = y_2 \mid Y_1 = y_1, R = r) \\ & Pr(Y_1 = y_1, R = r). \end{aligned} \tag{2.11}$$

Denote the conditional probability $Pr(Y_2 = 1 \mid Y_1 = y, R = r)$ as q_{yr} and the joint probability $Pr(Y_1 = y, R = r)$ as z_{yr}, for $y, r \in \{0, 1\}$. Among these eight

parameters, $q_{00}, q_{10}, z_{00}, z_{10}, z_{01}, z_{11}$ can be estimated using the observed data only. By adding the margin $Pr(Y_2 = y_2)$, we can estimate either q_{01} or q_{11}. In order to estimate all eight parameters, one more constraint needs to be applied to this model.

Hirano et al. (2001) suggest using a chain of selection models (Little, 1995) to characterize the joint distribution of (Y_1, Y_2, R). We have

$$\begin{aligned} Y_{1i} \mid \mathbf{X}_i &\sim Bern(\pi_{1i}) \\ logit(\pi_{1i}) &= \alpha_0 + \boldsymbol{\alpha}_x \mathbf{X}_i \end{aligned} \tag{2.12}$$

$$\begin{aligned} Y_{2i} \mid Y_{1i}, \mathbf{X}_i &\sim Bern(\pi_{2i}) \\ logit(\pi_{2i}) &= \beta_0 + \beta_1 Y_{1i} + \boldsymbol{\beta}_x \mathbf{X}_i \end{aligned} \tag{2.13}$$

$$\begin{aligned} R_i \mid Y_{1i}, Y_{2i}, \mathbf{X}_i &\sim Bern(\pi_{Ri}) \\ logit(\pi_{Ri}) &= \gamma_0 + \gamma_1 Y_{1i} + \gamma_2 Y_{2i} + \boldsymbol{\gamma}_x \mathbf{X}_i. \end{aligned} \tag{2.14}$$

The interaction term between Y_1 and Y_2 is not allowed due to lack of information.

MAR and MCAR are two special cases for the AN model with $(\gamma_1 = 0, \gamma_2 = 0)$ and $(\gamma_1 \neq 0, \gamma_2 = 0)$, respectively. If $(\gamma_1 \neq 0, \gamma_2 \neq 0)$, it follows a MNAR mechanism. Specifically, $(\gamma_1 = 0, \gamma_2 \neq 0)$ forms the nonignorable model by Hausman and Wise (1979) which can be denoted as HW model. Deng et al. (2013) present comparisons among MAR, HW model and AN model as well as further exploration about interaction terms.

2.2.3 The MD-AM Framework

The missing data with auxiliary margins (MD-AM) framework (Akande et al., 2019) extends the AN model to be more flexible for managing both item nonresponse and unit nonresponse. Specifically, we can specify a sequence of identifiable conditional models of all variables and their missing indicators to form a joint distribution. By combining auxiliary marginal information, we are able to maximize the information included in the models. The construction of the framework follows two steps:

1. Specify an identifiable model based only on the observed data. To achieve the identifiability, we assume that the missing is under an ignorable mechanism such as MAR or MCAR. Akande et al. (2019) suggests to use the maximum number of parameters under some identifiable assumption.

2. Introduce auxiliary margins and enhance the model by adding corresponding parameters. The auxiliary information can be used differently due to different assumptions. Interactions involving the variables of interest in the indicator models can be estimated only if joint margins are available.

One of the advantages of the MD-AM framework is that it can handle item nonresponse and unit nonresponse simultaneously, as well as more than one available margins. I now illustrate the implementation of the framework by an example from Akande et al. (2019) of two binary variables suffering from both item and unit nonresponse. The notation follows the former section. Additionally, U represents the unit nonresponse indicator, $U_i = 1$ implies the individual i does not provide any respons, and $U_i = 0$ otherwise. In this case, both Y_1 and Y_2 have item nonresponses. So there are two item nonresponse indicators, R_1 and R_2 linked to Y_1 and Y_2 respectively. For simplicity, the always observed covariates $\mathbf{X}$ are ignored in this example. The data structure is illustrated in Table 2.2.

To construct the joint distribution for all variables, we can denote

$$\begin{aligned}
\pi^1_{yr_2r_1u} &= Pr(Y_2 = 1 \mid Y_1 = y, R_2 = r_2, R_1 = r_1, U = u) \\
\pi^2_{r_2r_1u} &= Pr(Y_1 = 1 \mid R_2 = r_2, R_1 = r_1, U = u) \\
\pi^3_{r_1u} &= Pr(R_2 = 1 \mid R_1 = r_1, U = u) \\
\pi^4_u &= Pr(R_1 = 1 \mid U = u) \\
\pi^5 &= Pr(U = u).
\end{aligned} \tag{2.15}$$

Among these parameters, we can estimate $(\pi^1_{0000}, \pi^1_{1000}, \pi^2_{000}, \pi^2_{100}, \pi^3_{00}, , \pi^3_{10}, \pi^4_0, \pi^5)$

Table 2.2: Data structure for an example of the MD-AM framework.

Data	Y_1	Y_2	R_1	R_2	U
Original Data	✓	✓	0	0	0
	?	✓	1		
	✓	?	0	1	
	?	?	1		
	?	?	?	?	1
Auxiliary Margin	✓	?	?	?	?
Auxiliary Margin	?	✓	?	?	?

based on the observed data alone. We can start with the models characterized by the observed data itself. First we specify a joint distribution for (Y_1, Y_2),

$$(Y_1, Y_2) \sim f(Y_1, Y_2 \mid \Theta), \tag{2.16}$$

where Θ denotes the parameters in the joint distribution. We can specify a MCAR + ICIN mechanism for the missing indicators. For example, the unit nonresponse indicator is under a MCAR mechanism and the item nonresponse indicators are under an ICIN mechanisms.

$$U_i \mid Y_1, Y_2 \sim Bern(\pi_{Ui}), \quad logit(\pi_{Ui}) = \alpha_0 \tag{2.17}$$

$$R_{1i} \mid Y_1, Y_2 \sim Bern(\pi_{R1i}), \quad logit(\pi_{R1i}) = \beta_0 + \beta_1 Y_2 \tag{2.18}$$

$$R_{2i} \mid Y_1, Y_2 \sim Bern(\pi_{R2i}), \quad logit(\pi_{R2i}) = \gamma_0 + \gamma_1 Y_1. \tag{2.19}$$

By adding $Pr(Y_1)$ and $Pr(Y_2)$ from the margins, two more parameters can be introduced without losing identifiability. We can specify four different models from different missing data mechanism assumptions, leaving at least one of R_1, R_2, U under ICIN mechanisms.

- Assume (2.17) for U, and replace (2.18) and (2.19) by

$$R_{1i} \mid Y_1, Y_2 \sim Bern(\pi_{R1i}), \quad logit(\pi_{R1i}) = \beta_0 + \beta_1 Y_2 + \beta_2 Y_1 \tag{2.20}$$

$$R_{2i} \mid Y_1, Y_2 \sim Bern(\pi_{R2i}), \quad logit(\pi_{R2i}) = \gamma_0 + \gamma_1 Y_1 + \gamma_2 Y_2. \tag{2.21}$$

- Assume (2.18) for R_1, and replace (2.19) by (2.21), (2.17) by

$$U_i \mid Y_1, Y_2 \sim Bern(\pi_{Ui}), \quad logit(\pi_{Ui}) = \alpha_0 + \alpha_1 Y_1. \tag{2.22}$$

- Assume (2.19) for R_2, and replace (2.18) by (2.20), (2.17) by

$$U_i \mid Y_1, Y_2 \sim Bern(\pi_{Ui}), \quad logit(\pi_{Ui}) = \alpha_0 + \alpha_1 Y_2. \tag{2.23}$$

- Assume (2.18) and (2.19) for R_1 and R_2, and replace (2.17) by

$$U_i \mid Y_1, Y_2 \sim Bern(\pi_{Ui}), \quad logit(\pi_{Ui}) = \alpha_0 + \alpha_1 Y_1 + \alpha_2 Y_2. \tag{2.24}$$

Analysts can use logistic regression to estimate the model. In genuine applications, analysts can decide models to fit the distribution of data as well as missing data mechanisms.

3

Application to CPS Voter Turnout Data

Since voter turnout is one of the reflections of the wellness of democratic health (Fieldhouse, Tranmer and Russell, 2007), researchers tend to measure voter turnout for policy adjusting and population analysis. However, the measurements for voter turnout are not straightforward, especially by subgroups (Akande et al., 2019).

Voter turnout can be estimated by calculating the overall voter number over the total voting eligible population. The United States Elections Project (USEP) collects the aggregate level voting data by state, and voting age population (VAP) estimates are available from U.S. Census Bureau. Nevertheless, VAP includes people who are older than 18 but ineligible to vote (e.g., non-citizens). Thus, researchers turn to using other measures such as citizen voting age population (CVAP) and voting eligible population (VEP). The latter is preferred as it excludes non-citizens, felons and other populations that are not eligible to vote.

While overall turnout can be estimated from the sources above, researchers are also interested in turnout discrepancies among different subgroups. For example, turnout among 18 to 29-year-old young adults are evidently lower than for older populations (Holbein and Hillygus, 2016), and turnout among different racial and

ethnic groups and education levels could be unequal as well. In terms of the concerns about different patterns across subpopulation, public survey data can be another choice of estimating voter turnout.

The CPS survey is the most widely used for voter turnout estimation among national public surveys. One of the advantages is that CPS has been gathering demographic information connected to each voter for relatively large sample sizes. However, it still suffers from missing values that could lead to biased estimation. Several nonresponse bias correction methods are implemented by CPS. For unit nonresponse, CPS applies weight adjustments to upweight respondents; for item nonresponse, CPS utilizes three imputation methods - rational imputation, longitudinal edits and hot deck allocation; for voters with answers as "No response", "Don't know" or "Refused", CPS counts all these people as nonvoters (United States Census Bureau, 2016, at "Imputation of Unreported Data Items"; United States Elections Projects, at "CPS Vote Over-Report and Non-Response Bias Correction").

An analysis of the respondents to the turnout question, without any corrections for nonresponse, results in estimates that are higher than the actual turnout rates. For Washington DC, the vote over-report bias is over 20% in 2012 by subtracting VEP turnout rates from CPS turnout estimates (United States Elections Projects, at "2012 November General Election Turnout Rates"). This is partially due to the neglect of missing patterns among nonresponses and different patterns between unit nonresponse and item nonresponse. Therefore, in this thesis I apply the MD-AM framework on voter turnout concerning these problems.

3.1 Data Preparation

In order to consider more possible patterns for voter turnout, except for Age, Sex and Vote considered in previous research (Akande et al., 2019), I introduce nine variables from CPS data. The data are restricted to North Carolina in 2012. I filter

Table 3.1: Description of variables and missing rates in the CPS data analysis

Variables	Categories
VOTERESP (0.000)	1 = Self Respondent, 2 = Proxy Respondent, 3 = NIU
SEX (0.000)	0 = Male, 1 = Female
HISPAN (0.002)	0 = Not Hispanic, 1 = Hispanic
RACE (0.007)	1 = White, 2 = Black, 3 = Asian, 4 = Other, 5 = Multi
EDUC (0.018)	1 = High School-, 2 = College, 3 = Advanced
AGE (0.027)	1 = 19 - 29, 2 = 30 - 49, 3 = 50 - 69, 4 = 70+
VOTERES (0.094)	0 = Residence at current address for less than a year
	1 = Residence at current address for a year and more
FAMINC (0.183)	1 = Family income Less than 25000, 2 = 25000 - 49999,
	3 = 50000 - 74999, 4 = 75000+
VOTED (0.106)	0 = Did not vote, 1 = Voted

the ineligible voting samples and individuals with VOTE responded as "NIU (Not in Universe)" from the original data set. Table 3.1 presents the descriptions of variables in this application along with their missing rates.

Among these variables, VOTERESP and SEX have zero missing value in North Carolina. FAMINC bears the largest missing rate up to 18.3%. Our target variable VOTE has noticeable missing at 10.6%. Missing rate provides important evidence for model specification and choice of missing mechanism in the following section. From the ICPSR (Inter-university Consortium for Political and Social Research) version of CPS data, the unit nonresponse rates in NC for 2012 is 13.1%. We assume all unit nonrespondents are individuals that are eligible for voting in this application. To combine this information, I add completely missing observations to the data in order to ensure the unit nonresponse rate consistent with the real value. Then I create item nonresponse indicators for each variable with missing (ALL except VOTERESP and SEX) and a unit nonresponse indicator. The data have $n = 2191$ observations and $p = 17$ features.

Table 3.2: Variable margins in the CPS data analysis

Variables	Margins
SEX, AGE	(0,1) = 0.109, (0,2) = 0.180, (0,3) = 0.145, (0,4) = 0.046
	(1,1) = 0.107, (1,2) = 0.186, (1,3) = 0.160, (1,4) = 0.068
HISPAN	1 = 0.084
RACE	1 = 0.685, 2 = 0.215, 3 = 0.022, 4 = 0.057, 5 = 0.002
VOTED	1 = 0.648

I obtain marginal information for five of the nine variables, namely SEX, AGE, HISPAN, RACE and VOTE. For SEX, AGE, HISPAN and RACE, the marginal data is from the 2010 United States Census. I choose the 2010 Census data instead of other surveys in 2012 because of the accuracy of decennial census and the assumption of consistent information in two years duration (Akande et al., 2019). Based on the completeness of SEX, I use joint marginals for SEX and AGE to help with the nonresponses for AGE. For VOTE, I use the VEP (voter-eligible population) and aggregate counts for highest office. The voter turnout in NC for 2012 is 64.8% (The United States Elections Project, 2014). Table 3.2 shows the marginals for each available variable.

To incorporate marginal information, one approach is to sample a large number of synthetic observations for a desired variable from its known margin and leave the other variables missing (Schifeling and Reiter, 2016). This process is repeated for each variable with known margin and they are all appended to the original data. The size of the synthetic data is chosen to ensure the consistency between empirical distributions and margins as well as the appropriate impact for nonresponse. In this application, I generate $N = 3n$ synthetic samples resulting in $n^* = 28483$ as the final data size.

In this thesis, we use a number of unit nonrespondents equal to the number of

Table 3.3: Notations of variables used in model for CPS data analysis

Variables	Variable Notations	Missing Indicators
VOTERESP	PR_i	-
SEX	S_i	-
HISPAN	H_i	R_i^H
RACE	R_i	R_i^R
EDUC	E_i	R_i^E
AGE	A_i	R_i^A
VOTERES	RS_i	R_i^{RS}
FAMINC	F_i	R_i^F
VOTED	V_i	R_i^V
Unit Nonresponse	U_i	-

households that do not respond. Ideally, we instead would use the number of unit nonrespondents who are eligible to vote. Unfortunately that count is not available. Thus, all results from our analyses should not be interpreted substantively. As our purpose is to illustrate methodology, we proceed with the analysis using the incorrect number of unit nonrespondents.

3.2 Modeling

Following the notation of Akande et al. (2019), I notate variables and missing indicators in Table 3.3. I rank variables based on the missing rates and build a sequence of models for the distributions of all features following (3.1) to (3.9). These chain models form a joint distribution for all variables. Since VOTERESP and SEX are all observed, I adopt MCAR models to relieve computational burden and introduce them as covariates in all other models. I apply logistic regression for binary variables and multinomial logistic regression for categorical ones. I do not include any interaction term in the chained models.

$$
\begin{aligned}
PR_i &\sim Bern(\pi_i^{PR}): \\
&\quad logit(\pi_i^{PR}) = \beta_1^{PR}
\end{aligned} \tag{3.1}
$$

$$
\begin{aligned}
S_i &\sim Bern(\pi_i^{S}): \\
&\quad logit(\pi_i^{S}) = \beta_1^{S}
\end{aligned} \tag{3.2}
$$

$$
\begin{aligned}
H_i \mid PR_i, S_i &\sim Bern(\pi_i^{H}): \\
&\quad logit(\pi_i^{H}) = \beta_1^{H} + \beta_2^{H} S_i + \beta_3^{H} PR_i
\end{aligned} \tag{3.3}
$$

$$
\begin{aligned}
R_i \mid PR_i, S_i, H_i &\sim Multinom(Pr[R_i = r]): \\
&\quad logit(Pr[R_i = r]) = \beta_1^{R} + \beta_2^{R} S_i + \beta_3^{R} H_i + \beta_4^{R} PR_i
\end{aligned} \tag{3.4}
$$

$$
\begin{aligned}
E_i \mid PR_i, ..., H_i, R_i &\sim Multinom(Pr[E_i = e]): \\
&\quad logit(Pr[E_i = e]) = \beta_1^{E} + \beta_2^{E} S_i + \beta_3^{E} H_i + \beta_4^{E} PR_i \\
&\quad + \beta_5^{E} \mathbb{1}[R_i = r]
\end{aligned} \tag{3.5}
$$

$$
\begin{aligned}
A_i \mid PR_i, ..., R_i, E_i &\sim Multinom(Pr[A_i = a]): \\
&\quad logit(Pr[A_i = a]) = \beta_1^{A} + \beta_2^{A} S_i + \beta_3^{A} H_i + \beta_4^{A} PR_i \\
&\quad + \beta_5^{A} \mathbb{1}[R_i = r] + \beta_6^{A} \mathbb{1}[E_i = e]
\end{aligned} \tag{3.6}
$$

$$
\begin{aligned}
RS_i \mid PR_i, ..., E_i, A_i &\sim Bern(\pi_i^{RS}): \\
&\quad logit(\pi_i^{RS}) = \beta_1^{RS} + \beta_2^{RS} S_i + \beta_3^{RS} H_i + \beta_4^{RS} PR_i \\
&\quad + \beta_5^{RS} \mathbb{1}[R_i = r] + \beta_6^{RS} \mathbb{1}[E_i = e] + \beta_7^{RS} \mathbb{1}[A_i = a]
\end{aligned} \tag{3.7}
$$

$$
\begin{aligned}
F_i \mid PR_i, ..., A_i, RS_i &\sim Multinom(Pr[F_i = f]): \\
&\quad logit(Pr[F_i = f]) = \beta_1^{F} + \beta_2^{F} S_i + \beta_3^{F} H_i + \beta_4^{F} PR_i \\
&\quad + \beta_5^{F} \mathbb{1}[R_i = r] + \beta_6^{F} \mathbb{1}[E_i = e] + \beta_7^{F} \mathbb{1}[A_i = a] \\
&\quad + \beta_8^{F} RS_i
\end{aligned} \tag{3.8}
$$

$$
\begin{aligned}
V_i \mid PR_i, ..., RS_i, F_i \sim Bern(\pi_i^V) : \\
& logit(\pi_i^V) = \beta_1^V + \beta_2^V S_i + \beta_3^V H_i + \beta_4^V PR_i \\
& + \beta_5^V \mathbb{1}[R_i = r] + \beta_6^V \mathbb{1}[E_i = e] + \beta_7^V \mathbb{1}[A_i = a] \\
& + \beta_8^V RS_i + \beta_9^V \mathbb{1}[F_i = f].
\end{aligned} \tag{3.9}
$$

The next step is to specify models for item and unit nonresponse indicators. In order to retain the identifiability, each margin can be used only in either the unit nonresponse model or the respective item nonresponse model. Based on the significant unit nonresponse rate (13.1%) and relatively low missing rate for age (2.7%), Hispanic (0.2%) and race (0.7%), I use these margins to improve prediction of U by putting terms for these variables in the unit nonresponse model but not their respective item nonresponse models. Since the nonresponse rate for Hispanic is close to zero, I apply a MCAR model for R^H to lessen the model complexity. For the rest of missing indicator models except for R^V, no other margins are available. Therefore, I adopt ICIN models (Sadinle and Reiter, 2017) for them with maximum information included.

Since the most substantively important variable VOTE has a comparable missing rate (11.7%) with U, I consider two different assumptions to test possible mechanisms. First, I assign the margin for vote to the model for U to account for the relationship between unit nonresponse and vote behavior. In this case, I have to ignore the effect of vote decision on R^V. Second, as an alternative, I use the margin for vote to form a nonignorable model for R^V and leave U as irrelevant to vote. The detailed chain models under the first assumption are as follows (3.10) to (3.17).

$$
\begin{aligned}
U_i \mid S_i, H_i, R_i, A_i, V_i \sim Bern(\pi_i^U) : \\
& logit(\pi_i^U) = \gamma_1^U + \gamma_2^U S_i + \gamma_3^U H_i \\
& + \gamma_4^U \mathbb{1}[R_i = r] + \gamma_5^U \mathbb{1}[A_i = a] + \gamma_6^U V_i
\end{aligned} \tag{3.10}
$$

$$
\begin{aligned}
R_i^H &\sim Bern(\pi_i^{R_i^H}) : \\
& \\
& logit(\pi_i^{R_i^H}) = \gamma_1^{R_i^H}
\end{aligned}
\tag{3.11}
$$

$$
\begin{aligned}
R_i^R \mid S_i, ..., V_i &\sim Bern(\pi_i^{R_i^R}) : \\
& \\
& logit(\pi_i^{R_i^R}) = \gamma_1^{R_i^R} + \gamma_2^{R_i^R} S_i + \gamma_3^{R_i^R} H_i \\
& + \gamma_4^{R_i^R} PR_i + \gamma_5^{R_i^R} \mathbb{1}[E_i = e] + \gamma_6^{R_i^R} \mathbb{1}[A_i = a] \\
& + \gamma_7^{R_i^R} RS_i + \gamma_8^{R_i^R} \mathbb{1}[F_i = f] + \gamma_9^{R_i^R} V_i
\end{aligned}
\tag{3.12}
$$

$$
\begin{aligned}
R_i^E \mid S_i, ..., V_i &\sim Bern(\pi_i^{R_i^E}) : \\
& \\
& logit(\pi_i^{R_i^E}) = \gamma_1^{R_i^E} + \gamma_2^{R_i^E} S_i + \gamma_3^{R_i^E} H_i \\
& + \gamma_4^{R_i^E} PR_i + \gamma_5^{R_i^E} \mathbb{1}[R_i = r] + \gamma_6^{R_i^E} \mathbb{1}[A_i = a] \\
& + \gamma_7^{R_i^E} RS_i + \gamma_8^{R_i^E} \mathbb{1}[F_i = f] + \gamma_9^{R_i^E} V_i
\end{aligned}
\tag{3.13}
$$

$$
\begin{aligned}
R_i^A \mid S_i, ..., V_i &\sim Bern(\pi_i^{R_i^A}) : \\
& \\
& logit(\pi_i^{R_i^A}) = \gamma_1^{R_i^A} + \gamma_2^{R_i^A} S_i + \gamma_3^{R_i^A} H_i \\
& + \gamma_4^{R_i^A} PR_i + \gamma_5^{R_i^A} \mathbb{1}[R_i = r] + \gamma_6^{R_i^A} \mathbb{1}[E_i = e] \\
& + \gamma_7^{R_i^A} RS_i + \gamma_8^{R_i^A} \mathbb{1}[F_i = f] + \gamma_9^{R_i^A} V_i
\end{aligned}
\tag{3.14}
$$

$$
\begin{aligned}
R_i^{RS} \mid S_i, ..., V_i &\sim Bern(\pi_i^{R_i^{RS}}) : \\
& \\
& logit(\pi_i^{R_i^{RS}}) = \gamma_1^{R_i^{RS}} + \gamma_2^{R_i^{RS}} S_i + \gamma_3^{R_i^{RS}} H_i \\
& + \gamma_4^{R_i^{RS}} PR_i + \gamma_5^{R_i^{RS}} \mathbb{1}[R_i = r] + \gamma_6^{R_i^{RS}} \mathbb{1}[E_i = e] \\
& + \gamma_7^{R_i^{RS}} \mathbb{1}[A_i = a] + \gamma_8^{R_i^{RS}} \mathbb{1}[F_i = f] + \gamma_9^{R_i^{RS}} V_i
\end{aligned}
\tag{3.15}
$$

$$
\begin{aligned}
R_i^F \mid S_i, ..., V_i \sim Bern(\pi_i^{R_i^F}) : \\
& logit(\pi_i^{R_i^F}) = \gamma_1^{R_i^F} + \gamma_2^{R_i^F} S_i + \gamma_3^{R_i^F} H_i \\
& + \gamma_4^{R_i^F} PR_i + \gamma_5^{R_i^F} \mathbb{1}[R_i = r] + \gamma_6^{R_i^F} \mathbb{1}[E_i = e] \\
& + \gamma_7^{R_i^F} \mathbb{1}[A_i = a] + \gamma_8^{R_i^F} RS_i + \gamma_9^{R_i^F} V_i \\
R_i^V \mid S_i, ..., F_i \sim Bern(\pi_i^{R_i^V}) : \\
& logit(\pi_i^{R_i^V}) = \gamma_1^{R_i^V} + \gamma_2^{R_i^V} S_i + \gamma_3^{R_i^V} H_i \\
& + \gamma_4^{R_i^V} PR_i + \gamma_5^{R_i^V} \mathbb{1}[R_i = r] + \gamma_6^{R_i^V} \mathbb{1}[E_i = e] \\
& + \gamma_7^{R_i^V} \mathbb{1}[A_i = a] + \gamma_8^{R_i^V} RS_i + \gamma_9^{R_i^V} \mathbb{1}[F_i = f].
\end{aligned}
\tag{3.16}
\tag{3.17}
$$

To switch into the second assumption, I replace (3.10) and (3.17) with (3.18) and (3.19) as follows.

$$
\begin{aligned}
U_i \mid S_i, H_i, R_i, A_i \sim Bern(\pi_i^U) : \\
& logit(\pi_i^U) = \gamma_1^U + \gamma_2^U S_i + \gamma_3^U H_i \\
& + \gamma_4^U \mathbb{1}[R_i = r] + \gamma_5^U \mathbb{1}[A_i = a] \\
R_i^V \mid S_i, ..., F_i, V_i \sim Bern(\pi_i^{R_i^V}) : \\
& logit(\pi_i^{R_i^V}) = \gamma_1^{R_i^V} + \gamma_2^{R_i^V} S_i + \gamma_3^{R_i^V} H_i \\
& + \gamma_4^{R_i^V} PR_i + \gamma_5^{R_i^V} \mathbb{1}[R_i = r] + \gamma_6^{R_i^V} \mathbb{1}[E_i = e] \\
& + \gamma_7^{R_i^V} \mathbb{1}[A_i = a] + \gamma_8^{R_i^V} RS_i + \gamma_9^{R_i^V} \mathbb{1}[F_i = f] + \gamma_9^{R_i^V} V_i.
\end{aligned}
\tag{3.18}
\tag{3.19}
$$

The "..." notation represents omitted terms among all primary variables except for one related to the indicator. For example, R_i^A conditions on S_i, H_i, PR_i, R_i, E_i, RS_i, F_i, V_i but not A_i.

3.3 Implementation

I implement all models by Bayesian MCMC sampling. Since the posterior distribution for logistic regression does not have a closed form, methods like bootstrap and augmenting data can be applied to this problem. Based on the high complexity of the chain model, I instead use a normal approximation for the posterior probability (Rubin, 1987). Although there are several small subgroups that are not following normal distribution, they would have limited influence on the results in terms of the modest fraction they make.

When using a normal approximation for logistic regression, perfect separation problem could arise due to completely separated group samples (Albert and Anderson, 1984; Lesaffre and Albert, 1989; White, Daniel and Royston, 2010). In this case, the model will not be identifiable, thus infinite maximum likelihood estimates for coefficients will occur, which produces senseless results. Gelman et al. (2008) proposed a weak prior distribution for logistic regression to address this problem. They use a default choice of independent Cauchy distribution that centers 0 with scale 2.5. However, the heavy tails of Cauchy distribution can lead to an extremely large estimation. Instead, we can use a lighter tailed prior such as Student-t distribution with larger degrees of freedom (Ghosh, Li and Mitra, 2018). I choose a t_7 prior for predictors, and a Cauchy prior with 0 mean and scale 10 for the intercept.

For categorical variables, it is both theoretically and computationally challenging to implement these priors (Gelman et al., 2008). Nonetheless, rather than running a multinomial logistic regression, I adopt separate logistic regression for each category against all other levels following the approach of Su et al. (2011). For example, let Y be a categorical variable with $\{1, 2, 3\}$ three levels and $\mathbf{X}$ be covariates. We define three binary variables W^1, W^2, W^3 as indicators for each level. We set $W_i^1 = 1$ when $Y_i = 1$ and $W_i^1 = 0$ when $Y_i = 2$ or $Y_i = 3$, likewise W_i^2 and W_i^3. Then a multinomial

logistic regression on Y can be seperated into three models in (3.20). After running the models, the probability of each category are normalized to sum to one.

$$\begin{aligned} W_i^1 \mid \mathbf{X} &\sim Bern(\pi_{W_i^1}), \quad logit(\pi_{W_i^1}) = \beta_0^1 + \beta_1^1 \mathbf{X} \\ W_i^2 \mid \mathbf{X} &\sim Bern(\pi_{W_i^2}), \quad logit(\pi_{W_i^2}) = \beta_0^2 + \beta_1^2 \mathbf{X} \\ W_i^3 \mid \mathbf{X} &\sim Bern(\pi_{W_i^3}), \quad logit(\pi_{W_i^3}) = \beta_0^3 + \beta_1^3 \mathbf{X} \end{aligned} \tag{3.20}$$

I run the Gibbs sampler on the models for 5000 iterations, and drop the first 2000 as burn-in. Most estimates for parameters have decent convergence after 5000 iterations. Parameters related to variables with small sample size have thicker tail posterior distributions. Estimates for paramters in unit nonresponse models display a trend of slow convergence since we do not have any information in the original data. The results in next Chapter are based on the inferences from the 3000 iteration posterior samples.

4

Results

4.1 Voter Turnout Estimations

Table 4.1 presents the estimated coefficients of model (3.9) with 95% confidence intervals. Results include the estimates under (3.10) and (3.17), as well as estimates under (3.18) and (3.19). We call (3.10) and (3.17) the VOTE-UNIT model, and (3.18) and (3.19) the VOTE-ITEM model. The two sets of results provide similar patterns with slightly different performances for specific groups. The odds of voting for intercept is 26.4% (18.4%, 38.7%) by VOTE-UNIT model and 22.0% (14.6%, 30.0%) by VOTE-ITEM model. Whether it is a self or proxy respondent shows an impact on the imputation models, it does not make a difference to the chance of voting. There is no strong evidence that a specific gender is more inclined to vote. Both models predict that Hispanic are less likely to vote than non-Hispanic. The VOTE-UNIT model even estimates that the odds of voting for a person who is Hispanic is only 18.1% (12.2%,26.4%) times the odds for same one who is not Hispanic.

Black people are most likely to vote among all races based on the results from both models. This is due to the the high proportion of voters in black people from

Table 4.1: Coefficient estimates for VOTE (The references for categories and meanings of variables are in Table 3.1).

Variable	VOTE-UNIT	VOTE-ITEM
Intercept	-1.022 (-1.491, -0.504)	-1.265 (-1.765, -0.847)
VOTERESP2	-0.004 (-0.092, 0.082)	-0.008 (-0.088, 0.074)
VOTERESP3	0.002 (-0.132, 0.135)	-0.003 (-0.127, 0.121)
SEX1	-0.059 (-0.311, 0.203)	-0.173 (-0.399, 0.064)
HISPAN1	-1.710 (-2.107, -1.330)	-1.106 (-1.512, -0.687)
RACE2	0.756 (0.374 , 1.146)	1.230 (0.894, 1.565)
RACE3	-0.989 (-2.078, 0.262)	-0.091 (-1.431, 1.386)
RACE4	-1.266 (-1.989, -0.577)	-0.271 (-1.199, 0.625)
RACE5	-0.210 (-1.070, 0.833)	0.932 (-0.127, 1.991)
EDUC2	0.956 (0.693 , 1.244)	0.859 (0.601, 1.114)
EDUC3	1.878 (1.265 , 2.658)	1.762 (1.111, 2.468)
AGE2	0.378 (0.029 , 0.728)	0.435 (0.096, 0.735)
AGE3	0.968 (0.601 , 1.400)	0.921 (0.582, 1.250)
AGE4	1.475 (1.037 , 1.908)	1.501 (1.063, 1.929)
VOTERES1	0.607 (0.258 , 0.944)	0.520 (0.202, 0.827)
FAMINC2	0.417 (0.043 , 0.801)	0.537 (0.198, 0.920)
FAMINC3	0.624 (0.203 , 1.019)	0.770 (0.348, 1.190)
FAMINC4	0.936 (0.542 , 1.371)	1.013 (0.589, 1.431)

the original data set. For Asian, other races and multi-races, since these observations occupy only a small fraction of the data, the estimates for these groups have high variances with little consistency between results from the two models. We see compatible estimates for subgroups as age, education, residence and family income in two models: both models suggest that older people, people with higher level of education, people with more stable residence or people with higher amount of family income are more likely to vote.

Table 4.2 displays the overall estimated voter turnouts for each subgroup by

Table 4.2: Predicted voter turnout by groups using multiple imputation compared to the complete cases. The sample size is appended for each group in the complete data set. The point estimates with standard errors in the parentheses.

			VOTE-UNIT	VOTE-ITEM
	Size	CC	Est(SD)	Est(SD)
Overall	1701	0.774	0.629 (.012)	0.650 (.011)
Self	102	0.782	0.778 (.028)	0.757 (.022)
Proxy	678	0.763	0.756 (.030)	0.713 (.035)
Male	774	0.756	0.625 (.015)	0.649 (.014)
Female	927	0.790	0.632 (.015)	0.652 (.015)
Non-Hispan	1661	0.778	0.630 (.012)	0.652 (.011)
Hispan	39	0.641	0.619 (.029)	0.641 (.029)
White	1258	0.755	0.667 (.024)	0.638 (.020)
Black	376	0.867	0.704 (.068)	0.803 (.045)
Asian	12	0.750	0.328 (.197)	0.554 (.271)
Other	25	0.560	0.152 (.082)	0.354 (.164)
Multi-Races	19	0.737	0.345 (.197)	0.593 (.191)
College-	659	0.663	0.483 (.027)	0.526 (.031)
College	878	0.828	0.719 (.033)	0.719 (.028)
College+	153	0.941	0.844 (.072)	0.864 (.075)
<30	286	0.636	0.480 (.056)	0.492 (.047)
30-49	572	0.762	0.582 (.036)	0.632 (.038)
50-69	562	0.822	0.727 (.041)	0.726 (.031)
70+	281	0.843	0.784 (.064)	0.798 (.050)
Res <1 year	197	0.629	0.389 (.049)	0.483 (.081)
Res 1 year+	1486	0.793	0.704 (.033)	0.685 (.028)
<25k($)	363	0.645	0.501 (.053)	0.518 (.052)
25k-49k($)	366	0.751	0.607 (.058)	0.632 (.055)
50k-74k($)	305	0.797	0.652 (.063)	0.687 (.065)
75k+($)	423	0.861	0.755 (.042)	0.761 (.045)

multiple imputation compared to the complete cases. I stored 50 imputed data sets among 5000 iterations to do the analysis. Similar to the coefficient tables above, the VOTE-UNIT and the VOTE-ITEM model have analogous behaviors overall and within groups. As we can see, all the turnout estimates are lower than the complete case analyses. This reflects the effects of unit and item nonresponse in the data. Because of the discrepancy between the turnout rate from VEP and from the original data set, the models will specify most nonrepondents as nonvoters. Both models have a 95% confidence interval for overall turnouts that covers the marginal turnout rate from VEP.

For groups with relatively low levels of certain features such as education, age and family income, the estimated turnout rates under both models are lower than those among the respondents. This is reasonable because those groups of younger age, low education or low family income have a higher fraction of nonreponse. For groups with small sample size, e.g., small population races, the VOTE-UNIT model tends to predict a low turnout rate while the VOTE-ITEM model estimates those closer to the complete set. One potential problem is the correlation between residence status and age. As Younger people have a higher mobility, the estimated turnout for individuals with residence at current address for less than a year is affected more by younger age people. Thus the predicted turnout for high mobility group is evidently lower than the complete case. Out of this consideration, introducing an interaction term between residence status and age can be a possible solution.

Figure 4.1 displays the posterior predictive distributions of turnout for overall and for respondents in the complete data set. Compared to the turnout margin from VEP (64.8%) and the turnout rate from respondents in the original data set (77.4%), the VOTE-ITEM model characterizes those information better than the VOTE-UNIT model. The VOTE-UNIT model raises the turnout rate on respondents to compensate for the extremely low turnout estimate on unit nonrespondents. I will

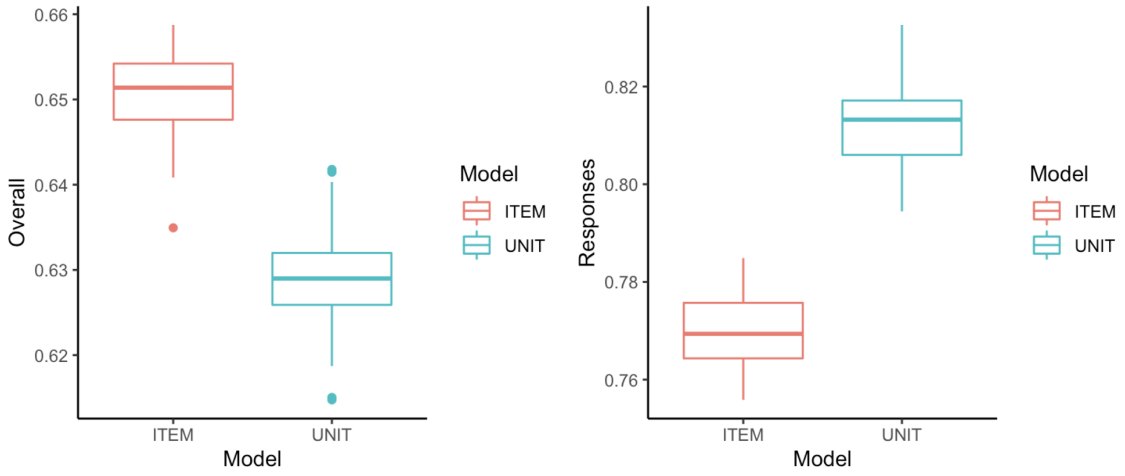

FIGURE 4.1: The left boxplot displays the overall posterior predictive distributions for turnout. The right boxplot displays the turnout posterior predictive distributions for respondents in the original set

present detailed differences for nonresponses between two models in the section 4.2.

4.2 Unit and Item Nonresponse Estimates

Figure 4.2 displays the posterior predictive distributions of turnout for unit and item nonrespondents. In both cases, the VOTE-UNIT model results in lower turnout rates. Especially for unit nonresponse, the estimated turnout is close to zero which implies almost all unit nonrespondents are nonvoters. This is because the considerable discrepancy between observed turnout rate and marginal turnout rate. The following sections will discuss the results for unit nonresponse and item nonresponse respectively.

4.2.1 Unit Nonresponse

Table 4.3 shows the coefficient estimates for unit noresponse for the two models. It seems that sex and Hispanic do not have much impact on the unit nonresponse. The two models agree on the estimates of Asian, other races or multi-races, apparently

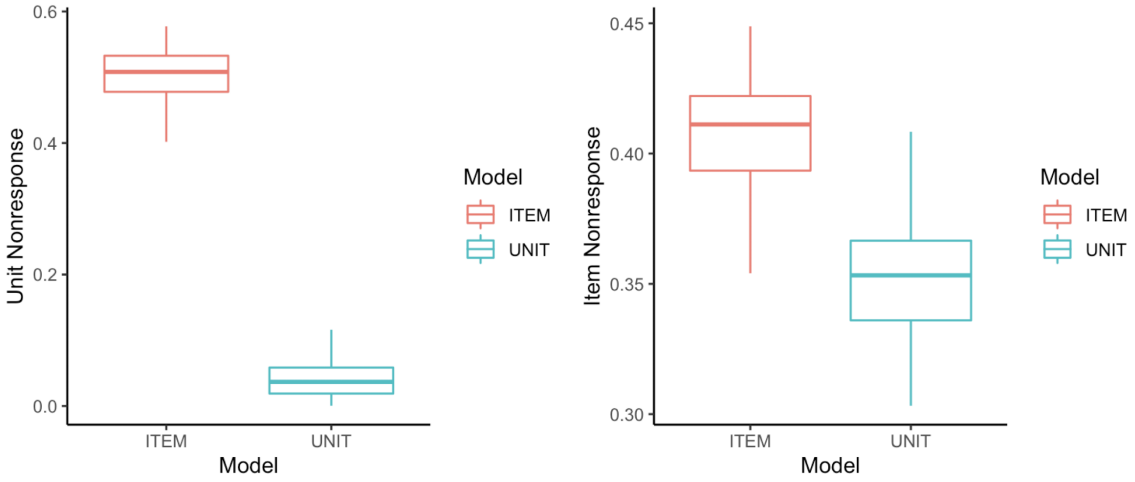

FIGURE 4.2: The left boxplot displays the turnout posterior predictive distributions for unit nonresponse. The right boxplot displays the turnout posterior predictive distributions for item nonresponse.

Table 4.3: Coefficient estimates for unit nonresponse.

Variable	VOTE-UNIT	VOTE-ITEM
Intercept	-1.980 (-2.966, -1.085)	-2.126 (-3.483, -1.330)
SEX1	0.004 (-0.160, 0.170)	-0.013 (-0.162, 0.131)
HISPAN1	0.117 (-0.130, 0.363)	0.123 (-0.083, 0.317)
RACE2	2.057 (1.214, 2.885)	-0.638 (-4.209, 1.888)
RACE3	3.458 (2.104, 5.024)	3.788 (2.792, 5.065)
RACE4	3.968 (3.057, 4.850)	4.088 (3.218, 5.226)
RACE5	3.502 (2.227, 4.805)	3.116 (2.064, 4.547)
AGE2	0.409 (-0.714, 1.662)	-0.725 (-1.716, 0.341)
AGE3	-0.574 (-1.944, 0.777)	-1.012 (-2.044, 0.049)
AGE4	-0.932 (-3.099, 0.599)	-2.679 (-5.697, -0.491)
VOTED	-4.515 (-6.979, -3.020)	-

Table 4.4: Turnout estimates for unit nonrespondents by groups.

	VOTE-UNIT	VOTE-ITEM
Overall	0.042 (.061)	0.495 (.165)
Male	0.042 (.061)	0.497 (.167)
Female	0.042 (.062)	0.494 (.166)
Non-Hispan	0.041 (.061)	0.495 (.166)
Hispan	0.046 (.067)	0.495 (.173)
White	0.030 (.048)	0.569 (.176)
Black	0.070 (.104)	0.758 (.266)
Asian	0.050 (.119)	0.537 (.318)
Other	0.021 (.033)	0.330 (.186)
Multi-Races	0.052 (.105)	0.565 (.258)
<30	0.025 (.041)	0.404 (.144)
30-49	0.047 (.068)	0.503 (.183)
50-69	0.049 (.074)	0.616 (.210)
70+	0.058 (.110)	0.656 (.328)

individuals in races with small sample size are more inclined to be unit nonrespondents.

The VOTE-UNIT model suggests that black people are more likely to not respond, whereas the VOTE-ITEM model does not present such trend. In addition to the shared variables between two models, the VOTE-UNIT model predicts strongly that individuals who do not vote are likely to be unit nonrespondents.

This phenomenon also can be observed from Table 4.4, which displays the estimated turnout rates for unit nonrespondents by groups. Most unit nonrespondents are predicted to be nonvoters by the VOTE-UNIT model. The VOTE-UNIT model considers vote as a factor affecting unit nonresponse, thus it classifies most unit nonresponse as nonvoters to diminish the gap between the true margin and the observed data.

Table 4.5: Coefficient estimates for vote item nonresponse.

Variable	VOTE-UNIT	VOTE-ITEM
Intercept	-2.861 (-3.906, -1.895)	-2.948 (-4.101, -2.024)
VOTERESP2	0.013 (-0.155, 0.187)	0.012 (-0.172, 0.198)
VOTERESP3	0.003 (-0.269, 0.269)	0.006 (-0.266, 0.288)
SEX1	0.850 (0.334, 1.341)	0.824 (0.289, 1.440)
HISPAN1	7.553 (6.434, 8.869)	9.402 (7.533, 11.359)
RACE2	-0.739 (-1.414, -0.040)	0.009 (-0.819, 0.776)
RACE3	-1.311 (-3.853, 1.284)	-1.435 (-4.292, 1.130)
RACE4	-0.397 (-2.676, 1.330)	-0.592 (-3.127, 1.461)
RACE5	-2.056 (-4.578, 0.074)	-1.955 (-4.732, 1.078)
EDUC2	0.130 (-0.559, 0.683)	0.674 (0.070, 1.311)
EDUC3	-0.354 (-1.385, 0.803)	0.610 (-0.949, 2.091)
AGE2	-0.327 (-1.054, 0.381)	-0.058 (-0.734, 0.731)
AGE3	-0.411 (-1.064, 0.303)	0.088 (-0.672, 0.884)
AGE4	-1.383 (-2.455, -0.372)	-0.831 (-1.998, 0.275)
VOTERES1	0.134 (-0.673, 0.951)	0.529 (-0.428, 1.583)
FAMINC2	-0.771 (-1.698, 0.061)	-0.629 (-1.707, 0.450)
FAMINC3	-1.118 (-2.196, -0.076)	-0.597 (-1.659, 0.584)
FAMINC4	-0.714 (-1.534, 0.168)	0.178 (-0.731, 1.249)
VOTED	-	-3.839 -5.590 -2.535

4.2.2 Item Nonresponse

Table 4.5 shows the coefficient estimates for item nonresponse of vote for the two models. Different from those of unit nonresponse, there is no evidence that race has much effect on item nonresponse. On the contrary, sex and Hispanic have large influence on the propensity to respond to the vote item. A male is more likely to be an item nonresponse to turnout than a female. Furthermore, a Hispanic individual has a noticeably larger chance of not answering voting question. In the VOTE-ITEM model, vote also highly contributes to item nonresponse, i.e., It is highly possible that

an individual who did not vote also did not answer vote question.

From the table in Appendix A, almost for all groups the turnout estimates for vote item nonresponse by the VOTE-UNIT model are smaller than those by the VOTE-ITEM model. This is consistent with the right boxplot in Figure 4.2.

5

Discussion

In this application, I do not introduce any interaction terms and I use a selection model on missing indicators for model clarity and computational simplicity. As mentioned in Chapter 4, the margins computed by multiple imputation indicate a collinearity between age and residence mobility. It is also commented by Akande et al. (2019) that the existence of the monotone missing pattern for age and sex indicates the need to build a relationship between missing indicators for age and sex. According to this, it is suggested to check the data pattern first in order to specify a model that can explain the data better.

In the CPS data, analyzing sixteen variables in total, including primary variables and missing indicators results in a complicated model specification and estimation task. Especially for categorical features, additional computational expenses are introduced by transforming those variables into binary ones. Even using normal approximation to handle logistic regression, it still takes long time for the models to run to convergence. This suggests a thought about incorporating nonparametric approaches such as Bayesian additive regression trees (BART) on this framework for future research.

6

Conclusion

In this thesis, I apply the MD-AM framework on CPS voter turnout data with multiple variables. When expanding to more variables and more margins by MD-AM framework, it is still straightforward to codify a joint distribution for the primary variables as well as models for missing indicators. Since vote is the variable of interest with available margin, we can choose to incorporate the margin into the unit nonresponse model or the item nonresponse model, based on the missing mechanism assumption.

From the estimates of the two models, we observe that the VOTE-UNIT model tends to predict most of the unit nonresponses as nonvoters. It also predicts a less than 40% turnout rate for item nonrespondents. This is due to the large differences between the margin from VEP and the turnout rate from the observed data. For the VOTE-ITEM model, as it does not consider the effect of vote for unit nonresponse, it generates estimates for both unit and item nonresponse around 40-50%. Since we only have one margin for vote, we need to consider which of the two sets of assumptions are more plausible.

Appendix A

Estimates for Vote Nonresponse by Groups

Table A.1: Turnout estimates for vote nonrespondents by groups.

	VOTE-UNIT	VOTE-ITEM
Overall	0.355 (.045)	0.409 (.044)
Male	0.352 (.049)	0.410 (.046)
Female	0.357 (.047)	0.409 (.048)
Non-Hispan	0.355 (.046)	0.410 (.044)
Hispan	0.351 (.062)	0.406 (.059)
White	0.409 (.064)	0.401 (.054)
Black	0.451 (.098)	0.586 (.098)
Asian	0.221 (.200)	0.428 (.324)
Other	0.104 (.067)	0.275 (.164)
Multi-Races	0.199 (.148)	0.401 (.248)
College-	0.226 (.046)	0.288 (.058)
College	0.454 (.075)	0.477 (.061)
College+	0.658 (.139)	0.716 (.146)

30	0.234 (.070)	0.272 (.064)
30-49	0.310 (.056)	0.402 (.066)
50-69	0.479 (.082)	0.504 (.073)
70+	0.532 (.125)	0.584 (.094)
Res 1 year	0.261 (.053)	0.308 (.099)
Res 1 year+	0.434 (.066)	0.442 (.052)
25k($)	0.234 (.061)	0.254 (.064)
25k-49k($)	0.339 (.068)	0.400 (.081)
509k-74k($)	0.386 (.095)	0.463 (.090)
75k+($)	0.495 (.085)	0.533 (.076)

Appendix B

Estimates for Vote by Specific Groups

Table B.1: Predicted voter turnout by certain subgroups using multiple imputation compared to the complete cases. The sample size is appended for each group in the complete data set. The point estimates with standard errors in the parentheses.

			VOTE-UNIT	VOTE-ITEM
	Size	CC	Est(SD)	Est(SD)
Nonhis White	1230	0.755	0.668 (.024)	0.640 (.021)
Young Black	74	0.784	0.576 (.092)	0.682 (.076)
Young Nonhis White	186	0.613	0.513 (.061)	0.463 (.061)
Res $<$ 1 year White	140	0.600	0.413 (.061)	0.452 (.081)
Res $<$ 1 year Black	51	0.725	0.478 (.092)	0.662 (.089)
Noncoll White	449	0.621	0.523 (.037)	0.497 (.040)
Young Noncoll Low Inc	48	0.417	0.256 (.079)	0.286 (.072)

Bibliography

Albert, A. and Anderson, J. A. (1984), "On the existence of maximum likelihood estimates in logistic regression models," *Biometrika*, 71, 1–10.

Andridge, R. R. and Little, R. J. (2010), "A review of hot deck imputation for survey non-response," *International Statistical Review*, 78, 40–64.

Brand, J. (1999), *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*, Erasmus Universiteit Rotterdam.

Brick, J. M. and Kalton, G. (1996), "Handling missing data in survey research," *Statistical Methods in Medical Research*, 5, 215–238.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010), "mice: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, 45.

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

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

Fichman, M. and Cummings, J. N. (2003), "Multiple imputation for missing data: Making the most of what you know," *Organizational Research Methods*, 6, 282–308.

Fieldhouse, E., Tranmer, M., and Russell, A. (2007), "Something about young people or something about elections? Electoral participation of young people in Europe: Evidence from a multilevel analysis of the European Social Survey," *European Journal of Political Research*, 46, 797–822.

Galea, S. and Tracy, M. (2007), "Participation rates in epidemiologic studies," *Annals of Epidemiology*, 17, 643–653.

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*, ed. H. Wainer, Springer, New York, NY.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007), "How many imputations are really needed? Some practical clarifications of multiple imputation theory," *Prevention Science*, 8, 206–213.

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

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.

Holbein, J. B. and Hillygus, D. S. (2016), "Making young voters: the impact of preregistration on youth turnout," *American Journal of Political Science*, 60, 364–382.

Lesaffre, E. and Albert, A. (1989), "Partial separation in logistic discrimination," *Journal of the Royal Statistical Society: Series B (Methodological)*, 51, 109–116.

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

Nevo, A. (2003), "Using weights to adjust for sample selection when auxiliary information is available," *Journal of Business & Economic Statistics*, 21, 43–52.

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

Rubin, D. B. (2004), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons.

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.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall/CRC.

Schafer, J. L. and Olsen, M. K. (1998), "Multiple imputation for multivariate missing-data problems: A data analyst's perspective," *Multivariate Behavioral Research*, 33, 545–571.

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., et al. (2016), "Incorporating marginal prior information in latent class models," *Bayesian Analysis*, 11, 499–518.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009), "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *BMJ*, 338, b2393.

Su, Y.-S., Gelman, A. E., Hill, J., and Yajima, M. (2011), "Multiple imputation with diagnostics (mi) in R: Opening windows into the black box," *Journal of Statistical Software*, 40.

White, I. R., Daniel, R., and Royston, P. (2010), "Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables," *Computational Statistics & Data Analysis*, 54, 2267–2275.