Bayesian Predictive Inference for a Non-probability Sample with Binary Responses from Small Areas

by

Zhiqing Xu

A PhD Dissertation

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

in

Statistics

by

September, 2020

APPROVED:

Professor Balgobin Nandram, Advisor Department of Mathematical Sciences Worcester Polytechnic Institute	Professor Joseph Sedransk University of Maryland
Associate Professor Fangfang Wang Department of Mathematical Sciences Worcester Polytechnic Institute	Professor Joseph Petruccelli Department of Mathematical Sciences Worcester Polytechnic Institute
Assistant Professor Andrea Arnold Department of Mathematical Sciences Worcester Polytechnic Institute	Dr. Jai Won Choi Statistical Consultant, Meho Inc., MD

Abstract

In a world of big data, non-probability samples are fast and easy to collect, and the responses can be binary. Traditional design surveys, where probability theory plays an important role, require enormous planning and are very expensive. Typically, to reduce cost and save time, large data sets are collected using haphazard methods instead of designed surveys. An issue with big data is that the selection probabilities are unknown and descriptive summaries are generally biased. It is often the case that selection probabilities are related to the covariates and the binary response variable, and selection is not at random (SNAR); the samples and non-samples are not random samples from the population. The main contribution of our research is a methodology for correcting selection bias in nonprobability samples with binary response and appropriate covariates.

To study binary response data and to deal with the sampling bias that comes from the SNAR mechanism in a single area, we propose a non-ignorable selection model that uses a double logistic regression to link the response model with the selection model. When selection is at random (SAR), a single logistic regression model could be used to serve as an ignorable selection model (a link to the selection mechanism is not needed). Both models are fit using full Bayesian methods. We use simulation studies to evaluate the ability of the non-ignorable selection model to adjust for the selection bias from the SNAR mechanism. The results show that when samples are SNAR, the non-ignorable selection model gives unbiased population proportion prediction, and when samples are SAR, the non-ignorable selection model performs similarly to the ignorable selection model. We also demonstrate the use of the model with real data from the Third National Health and Nutrition Examination Survey (NHANES III), where a binary version of body mass index is derived as the response with demographic covariates (age, race, sex). Additional work includes a study on priors and a methodology for situations where individual covariates are typically unknown for the non-sampled population, but other sources of data are integrated into the ensemble.

We extend the non-ignorable selection model to incorporate area level information, which is accommodated using random effects in the response sub-model and selection sub-model respectively. Small area estimation has become enormously important where inference from one area cannot be reliably made. Both the non-ignorable and ignorable selection models are applied to simulated data sets and real data from NHANES III with thirty-five counties. Furthermore, we develop two variations of this model using (a) more robust assumptions by assigning Dirichlet process priors to the random effects, and (b) a bivariate model to incorporate the correlation of the two sets of random effects.

Key Words: Bayesian hierarchical model, Dirichlet Process, Double logistic regression, Metropolis– Hastings sampler, Selectivity, Small area estimation.

Acknowledgments

First of all, I would like to express my deepest appreciation to my advisor, Professor Balgobin Nandram, without whose brilliant mind, enthusiasm and patient guidance this dissertation would not have been possible. I am very lucky to have Professor Nandram as my mentor since I was working on my master's degree. His passion for research and hard working keeps me stay motivated throughout my course of studies.

I would also like to extend my sincere thanks to my committee members, Professor Joe Sedransk, Dr. Jai Won Choi, Professor Joseph Petruccelli, Professor Fangfang Wang and Professor Andrea Arnold for their time and effort regarding my dissertation. Each of the members has provided me invaluable comments and insight in improving my work.

I very much appreciate the Department of Mathematrical Sciences at WPI for the education and supports. I am forever grateful for the kindness that all of the faculty members, staff, and students have shown me. I am especially indebted to Professor Petruccelli for his wise advice and for always managing to make me smile.

Special thanks to my family, especially my parents for giving me unconditional love.

Contents

1 Introduction **1**

1.1 Non-probability Sampling Methods 2

1.2 The Three Types of Missingness 4

1.3 State of the Art 5

1.3.1 Approaches for MAR 7

1.3.2 Approaches for MNAR 10

1.3.3 Multi-level Modelling under MNAR 13

1.4 Statistical Background 14

1.4.1 Binary Data 14

1.4.2 Logistic Regression 14

1.5 Notation 15

1.6 Contributions of this Research and Dissertation Organization 16

2 Bayesian Selection Model for Homogeneous Data **19**

2.1 Selection Model Specification 19

2.1.1 Nonignorable Selection Model 20

2.1.2 Ignorable Selection Model 23

2.2 Simulation Study 25

2.2.1 Dataset Description 25

2.2.2 Comparison and Results . . . 27
2.2.3 Repeated Simulation Runs on the Nonignorable Selection Model . . . 37
2.2.4 Model Diagnostic by LPML . . . 41
2.3 Application on Body Mass Index . . . 43
2.3.1 A More Compact Nonignorable Model . . . 43
2.3.2 Diagnostics and Results . . . 48
2.4 Study on Priors . . . 52
2.5 Study on Unknown Individual Covariates . . . 54
2.5.1 Simulation Study . . . 57
2.6 Conclusion . . . 59

3 Bayesian Selection Model for Heterogeneous Data 60

3.1 Hierarchical Model Specification . . . 61
3.1.1 Hierarchical Nonignorable Selection Model . . . 61
3.1.2 Hierarchical Ignorable Selection Model . . . 66
3.1.3 Simulation Study . . . 67
3.2 Application on Body Mass Index with Multiple Areas . . . 73
3.2.1 Model Reformulation of the Nonignorable Selection Model . . . 73
3.3 Model with Dirichlet Process Prior . . . 80
3.3.1 Slice Sampling . . . 81
3.3.2 Simulation Study . . . 85
3.4 Model with Bivariate Prior for Area Effects . . . 95
3.4.1 Simulation Study . . . 96
3.5 Conclusion . . . 100

4 Concluding Remarks and Future Work 101

4.1 Concluding Remarks . . . 102

4.2 Future work 104

A Conditional Posterior Distribution for y_i in Section 2.1.1 (iii) 106

B Propriety of the Posterior Density of a general case in (2.1.4) 108

C Bounds of free variables derived from (2.5.2) 111

D A preliminary approach to draw α 114

D.1 Posterior for α 114

List of Figures

2.1 Histogram of population proportion prediction by scenario. (a): nonignorable selection model (left), ignorable selection model (right). (b) and (c): nonignorable selection model (right), ignorable selection model (left). Dashed line: true population proportion 36

2.2 95% HPD interval and posterior mean of population proportion prediction in repeated simulation runs by scenario. (Dashed line: true population proportion. Red circle: sample proportion.) 40

2.3 Histogram of the finite population proportion for children with healthy BMI by ignorable (left) and nonignorable model (right) 51

2.4 Histogram of Population Proportion Prediction 58

3.1 Population proportion posterior mean and 95% HPD interval for each area using the nonignorable model. (Triangle: true population proportion. Circle: sample proportion.) 71

3.2 Histogram of the overall population proportion predictions by nonignorable selection model (right) and ignorable selection model (left). Dashed line: true population proportion. 72

3.3 Population proportion posterior mean and 95% HPD interval for each county under the nonignorable model. (Circle: sample proportion.) 78

3.4 Histogram of Population Proportion Prediction by nonignorable selection model (right) and ignorable selection model (left). 79
3.5 Population proportion posterior mean and 95% HPD interval for each area under the nonignorable model. (Triangle: true population proportion. Circle: sample proportion.) 88
3.6 Density of Population Proportion Predictions by Area under the Nonignorable Selection Model 89
3.7 Histogram of the overall population proportion predictions by nonignorable selection model (right) and ignorable selection model (left). Dashed line: true population proportion. 90
3.8 Density of $\boldsymbol{\nu}_1$ by group 91
3.9 Density of $\boldsymbol{\nu}_2$ by group 92
3.10 Histogram of k_1 93
3.11 Histogram of k_2 94
3.12 Population proportion posterior mean and 95% HPD interval for each area under the nonignorable model. (Triangle: true population proportion. Circle: sample proportion.) 98
3.13 Histogram of the overall population proportion predictions by nonignorable selection model (right) and ignorable selection model (left). Dashed line: true population proportion. 99
3.14 Diagnostics of ρ 100

List of Tables

2.1 Distributions for covariates 26

2.2 Posterior summaries and diagnostics of γ and β for the nonignorable selection model by scenario 29

2.3 Posterior summaries of $\bar{Y}$ for the nonignorable selection model by scenario 29

2.4 Posterior summaries and diagnostics of γ for the ignorable selection model by scenario 31

2.5 Posterior summaries of $\bar{Y}$ for the ignorable selection model by scenario 31

2.6 Results comparison by scenario 33

2.7 Results over 20 replicates by scenario 37

2.8 LPML of nonignorable and ignorable models under 3 scenarios 42

2.9 Posterior summaries and diagnostics for γ and β 49

2.10 Posterior summaries and diagnostics of γ for the ignorable selection model 50

2.11 Posterior summaries of $\bar{Y}$ 50

2.12 Results comparison by different prior for β 53

2.13 Data structure 54

2.14 Counts by levels for each covariate 55

2.15 Posterior summaries of $\bar{Y}$ 57

3.1 Distributions for covariates 67

3.2 Posterior summaries and diagnostics for β, γ, ν_1, ν_2, σ_1^2,σ_2^2 68

3.3 Posterior summaries of finite population proportions 70
3.4 Posterior summaries of finite population proportions 77
3.5 Posterior summaries of finite population proportions 87
3.6 Posterior summaries of finite population proportions 97

Chapter 1

Introduction

Traditionally, the objective of data collection is to obtain a sample that is an accurate representation of the target population in the sense that it reflects all aspects of the population adequately. The selection of such samples is guided by an underlying probabilistic mechanism which ensures that each and every population unit has a positive probability of being selected.

However, obtaining a representative probability sample is often prohibitively difficult in the real world due to various constraints. Even if a probability sampling plan is implemented, it is a difficult task to obtain the requisite responses from the selected sample units. According to the Pew Research Center, the response rate in telephone surveys dropped from 36% in 1997 to only 9% in 2012. Such low response rates cast doubts on the validity of probability samples as a proper representation of the population.

Those considerations along with an explosion of data being generated through various channels have led to an upsurge in the usage of non-probability sampling schemes. However, these obtained samples often suffer from selection and hidden biases. Beaumont (2020) and Rao (2020) review available methods that use data from a non-probability source, as well as the literature on integrating data of a probability and non-probability sample, and conclude that the recent methods are not reliable and general enough to eliminate the use of probability surveys. Therefore, this indicates

that more research is called for in this field.

The primary objective of this research is to propose a methodology that is aimed at reducing selection bias when there is a significant difference between sampled and non-sampled units. In this chapter, we discuss the main concepts in non-probability sampling and the currently used methodologies to follow. Some basic concepts relating to non-probability sampling and types of missingness are presented in Sections 1.1 and 1.2 respectively, a review of statistical approaches and some statistical backgrounds are presented in Sections 1.3 and 1.4, and notation is presented in Section 1.5. An overview of contributions of this research is discussed in Section 1.6.

1.1 Non-probability Sampling Methods

Since early work on observational studies of smoking and disease (Doll and Hill, 1964), non-random sampling (known as non-probability sampling) methods have increased in both complexity and use. In recent decades non-probability sampling holds the potential for providing a faster, cheaper, and easier way to collect data compared to the traditional probability sampling method. In order to explore the applicability of non-probability sampling schemes, the American Association of Public Opinion Research (AAPOR) conducted a task force that outlined various forms of non-probability sampling schemes (Baker, 2013).

Convenience sampling is a widespread sampling method in psychology (Presser, 1984), litigation research (Diamond, 2000), and medical research (Couper, 2007). It is a technique in which the ease that potential participants can be located or recruited is the primary consideration, i.e., the samples are selected based on their convenience for researchers. Sampling methods such as *mall-intercept samples*, *volunteer samples*, *river samples*, and *snowball samples*, are in the category of convenience sampling. Since those non-probability samples are often self-selected and/or homogeneous (e.g., river samples, where respondents are visiting the same website, are normally homogeneous), one could not use a randomization approach to make inference from such samples.

Quota sampling is another kind of non-probability sampling technique that uses sample matching and stratification. In quota sampling, the objective is to obtain the samples that 'mirror' the target population in terms of characteristics such as age, race and gender. In the UK, National Opinion Polls have consistently employed random samples while the Gallup Poll has used quota samples. Smith (1983) compared the survey results for the elections from 1959 to 1979 with the actual outcomes. From the comparison, they show that quota samples provide the same prediction accuracy as the random samples.

Network sampling is used to find members of hard-to-reach or hidden populations. The early foundations of network sampling in the statistical literature (e.g. Goodman 1961) are probability sampling methods. However, in more recent practice, the strategies, like *link-tracing network sampling*, where eligible sample members connect researchers to other eligible samples, has proved useful in cases where the probability sampling is not applicable, e.g., rare ethnic minorities (Welsh 1975), or people at risk for disease such as HIV (Klovdahl et al. 1994), or COVID-19 contact tracing (Firth etc., 2020).

Non-probability samples usually suffer from selection bias, because the sample is most likely differ from the nonsample in such a way that the sample cannot be accurately projected to the full population. Valliant and Dever (2011) defined three populations to describe three components of coverage survey bias. Those three populations are (1) the target population of interest, such as adults who currently use cigarettes; (2) the potentially covered population given the way that data are collected, such as people who have Internet access and visit the sites where study recruitment occurs; and (3) the actual covered population, in the same example is the set of people who participate in the study. The samples, those who are invited to participate and actually do, are biased due to the selection process.

In our research, we focus on the methodologies of dealing with selection-biased samples, which could be gathered by a non-probability sampling method or a probability sampling method with high nonresponse rate. We refer to both non-sampled units and sampled but non-respondents as

missing data. In the next section, we discuss the three broad types of missing data.

1.2 The Three Types of Missingness

Rubin (1976) introduced a formal framework for the field of incomplete data by classifying the missing data mechanisms into three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). When we talk about sample surveys, it is more natural to use selection completely at random (SCAR), selection at random (SAR), and selection not at random (SNAR), see Nandram et al. (2013). However, to be consistent with the literature across fields, we use MCAR, MAR, and MNAR in this document.

We say data are MCAR when the probability of a value being missing is unrelated to the observed and unobserved data on that unit. Denote the missingness indicator variable by $\boldsymbol{I}$, covariates by $\boldsymbol{X}$ and the response variable by $\boldsymbol{Y}$. Mathematically, MCAR can be represented by $P(I_i \mid X_i, Y_i) = P(I_i)$. When data are MCAR the observed sample is most likely to be representative of the population.

The data are MAR if given, or conditional on, the observed data the probability distribution of $\boldsymbol{I}$ is independent of the unobserved data. Mathematically, $P(I_i \mid X_i, Y_i) = P(I_i \mid X_i)$.

If the mechanism causing missing data is neither MCAR nor MAR, we say it is MNAR. Under an MNAR mechanism, the probability of an observation being missing depends on the underlying response, and this dependence remains even given the observed data. Mathematically, $P(I_i \mid X_i, Y_i) \neq P(I_i \mid X_i)$. For example, in a health-related survey study, it is reasonable to believe that nonresponse may be higher in those with worsening health status, which is the outcome of interest. Analysis under MNAR is considerably harder, because the probability of an observation being missing cannot be estimated directly from a model. We will introduce two approaches to bypass the issue in Section 1.3.2.

Next, we will review the methodologies that have been proposed in the missing data field.

1.3 State of the Art

The estimation procedures that have been used to deal with non-probability samples, missing data and selection biased samples can be classified into two broad categories: *Design-based* (sometimes referred to as *pseudo design-based*) and *Model-based* estimation.

Design-based estimation is associated with the estimated probability of being in the non-probability sample, which relies on the heuristic that each observation represents how much of other non-sampled or non-responding units, i.e., the pseudo weight of each observation. When the pseudo weight is formed, the estimators can be computed by substituting the sample weight in the traditional design-based sample with the estimated non-probability sample weight, and the pseudo weight can be applied to estimation for any Y. The method of estimating the pseudo weight differs from application to application and requires strong assumptions for the non-probability sampling recruitment methods. Also, since the pseudo weight does not account for the population structure of the Y, such an estimator can have a higher variance compared to model-based approaches and will be biased when the response depends on covariates.

Inverse probability of treatment weighting (IPTW) is a design-based approach that uses a propensity score. This allows one to obtain unbiased estimation if there are no residual systematic differences between groups after being weighted by the estimated inverse of the propensity score. A potential drawback is that the variance can be very large if the weights are extreme, i.e., if the estimated propensity scores are close to 0 or 1. Setting weights to be within a range has been proposed as one solution to this problem (Scharfstein, Rotnizky and Robins, 1999). One can Winsorize the weights above 1.5 IQR of the third quantile.

In our simulation study (Section 2.2.2), we use the inverse of the propensity score as the weight in the Horvitz-Thompson estimator and the Hajek Estimator to serve as comparisons with the model we proposed. When our simulation scenario is MAR, the Hajek estimator gives the most accurate prediction among all methods; but when scenarios are MNAR, neither estimator success-

fully adjusts the bias.

Model-based approaches rely on statistical models that describe the variable being estimated in the survey. Then, the variable of interest is assumed to be a random variable with that distribution. When the samples are observed, they are used to fit the model and the analysis is conducted. A typical use of this approach requires that the samples are independent draws from the distribution of interest. However, this requirement can be relaxed to exchangeability in the Bayesian paradigm. The Heckman selection model (1976) is an example of a model-based method from the econometrics literature. Our models are built on the selection model and we will see more about it in Section 1.3.2 and the following chapters.

Model-based approaches can be classified as ignorable nonresponse and nonignorable nonresponse models. When the selection indicators do not depend on the missing values, i.e., for MCAR or MAR, an ignorable nonresponse model could be used. When selection indicators depend on the missing values, i.e., for MNAR, a nonignorable selection model is appropriate to use. Ignorable selection is a strong assumption and evidence to support this assumption is typically nonexistent. Therefore, it is advisable to study the sensitivity of conclusions to plausible violations of ignorable selection models. In Sections 2.1 and 2.2, we will review ignorable selection models for MAR and nonignorable selection models for MNAR respectively.

We restrict our attention to the methodologies where only the response variable contains missing values and covariates are observed for the whole population. There are ample studies that have been done for other cases. Elliott and Valliant (2017) reviewed the approaches where covariates for individual nonsample units are unknown, or only population totals of the covariates are available. For the design-based approach, they used a probability sample as a reference survey with the goal of estimating the probability of a unit being in the nonprobability sample. This can be done by matching the covariates in the non-probability sample with the probability sample and applying the probability of the unit being in the probability sample to the derived formulation. When the covariates do not correspond precisely to the probability samples, regression methods can be used. The

superpopulation model approach requires that the population total of the covariates and the mean of the response variables are linearly related to the covariates. Multilevel regression and stratification (MRP) is another approach, which can be viewed as a weighted average of poststrata means, where the weight P_h can be estimated via a probability sample. MRP can be also done by Bayesian inference to incorporate the uncertainty of P_h, for example, Gelman and Little (1997) construct a hierarchical logistic regression model to fit many more cells compared to what is possible using classical methods.

Although our methodology is designed in the context of survey inference which involves mean or total estimation, literature in the causal context is also included. Despite the difference in the parameter of interest, the conditions that produce selection bias are similar (Mercer et al., 2017).

1.3.1 Approaches for MAR

The general idea in model-based estimation when adjusting for MAR is to assume the sample and nonsample units follow the same model, where model parameters can be estimated from the sample and used to make predictions for the nonsample cases. Valliant, Dorfman and Royall (2000) covered the *model-assisted calibration* extensively when the mean of a variable Y is linearly related to $\boldsymbol{X}$ (i.e., $E(Y_i \mid \boldsymbol{X}_i) = \boldsymbol{X}_i^T \boldsymbol{\beta}$). Regression models can be used with matching methods, which aim to equate or balance the distribution of covariates between groups. Stuart (2010) summarized the key advantages of matching methods: (1) matching methods are not in conflict with regression adjustment; (2) when there is not sufficient overlap between groups, matching methods perform better than selection models and regression models; and (3) matching methods have straightforward diagnostics by which their performance can be assessed.

Dealing with multiple covariates is a challenge. Chapin (1947) found that with initial pools of 671 treated and 523 controls there were only 23 pairs that matched exactly on six categorical covariates. Rosenbaum and Rubin (1983) made an important advance with the introduction of the propensity score, which is the conditional probability of assignment to a particular treatment

given a vector of observed covariates. The idea is that subjects in treatment or control groups with the same propensity scores will have similar distributions of observed covariates. Both large and small sample theory show that adjustment for the propensity score is sufficient to remove bias due to all observed covariates (Rosenbaum and Rubin, 1983). One problem with propensity score matching is that different covariates could have similar propensity scores. Also, one needs to have all possible covariates to construct propensity scores for nonsamples.

There are three types of matching methods, which are one to one matching, weighting, and subclassification. We briefly review those as follows.

Nearest neighbor matching

Nearest neighbor matching is the most commonly used and easiest to understand one to one matching method (Rubin, 1973). The idea is to match control (or sampled) individuals to the treated (or nonsampled) group and discard those who are not selected as matches. Since being discovered, several drawbacks have been reported. One concern is that when there are no control individuals with propensity scores similar to a given treated individual, the matching can lead to poor matches. To avoid poor matches, Rosenbaum and Rubin (1985) discussed trade-offs when imposing a caliper distance and only select a match if it is within the distance. Stuart (2010) reviewed the issues that come from optimal matching, ratio matching, and with or without replacement.

Post-stratification

Post-stratification forms groups of individuals who are similar, which is usually defined by quantiles of the propensity scores. The purpose of post-stratification is to correct for known differences between the sample and population. The implicit model of post-stratification is that data can be treated as a simple random sample within each poststratum or, more generally, the relative probabilities of selection are equal.

Gelman (2007) discussed several post-stratification models with the assumption that the co-

variates $\boldsymbol{X}$ are discrete and the categories include all the cross-classifications of $\boldsymbol{X}$, as well as the population size N_j of each category j is known. The *full post-stratification* model gives the population mean estimate as $\hat{\bar{Y}} = \dfrac{\sum_{j=1}^{J} N_j \bar{y}_j}{\sum_{j=1}^{J} N_j}$. This can be interpreted as weighted averages of cell means (Little, 1993). When the weights of the cells depend on the values of covariates $\boldsymbol{X}$, one can use the *classical regression* model. The post-stratified estimate can be expressed as $\hat{\bar{Y}}^{PS} = \dfrac{1}{n} w_j \bar{y}_j$, where $w_j = \dfrac{n}{N} \boldsymbol{N}' \boldsymbol{X}^{pop} (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}'$, here $\boldsymbol{N} = (N_1, .., N_J)$ and $\boldsymbol{X}$ is the $J \times k$ matrix of for the J post-stratification cells. The regression model can be generalized to a *hierarchical regression* model with $y \sim \mathrm{N}(X\beta, \Sigma_y)$ and a prior distribution on β with $\beta \sim \mathrm{N}(0, \Sigma_\beta)$.

In recent research, we use propensity scores to balance the sampled and non-sampled units and propose three modeling methods to address the issue of MAR in non-probability samples (Nandram, Cao, Xu, and Bhadra, 2020). The propensity score is given by $\pi(\boldsymbol{x}_i) = P(R_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{\beta}) = \dfrac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}, i = 1, ..., N$. We first draw 1,000 simulated values of $\boldsymbol{\beta}$ from its posterior distribution, where 1,000 is sufficient in representing the spread and location of $\boldsymbol{\beta}$ and has been used in many publications, e.g., see (Nandram, Chen, Fu, and Manandhar, 2018). Using fully Bayesian analysis and for each value of $\boldsymbol{\beta}$, we calculate the propensity scores for all units in the population. Then, 10 strata are formed by 10 intervals from deciles of all propensity scores. For each stratum and each $\boldsymbol{\beta}$, inferences for the non-sampled units can be made by the model $Y_{ij} \mid p_i \sim \text{Bernoulli}(p_i), i = 1, ..., J; j = n_i + 1, ..., N_i$, and the posterior distribution for p_i is $p_i \mid \boldsymbol{y}_i \sim \text{Beta}\left(\sum_{j=1}^{n_i} y_{ij}, n_i - \sum_{j=1}^{n_i} y_{ij}\right)$. At last, our target of inference – the finite population proportion is computed by $\dfrac{1}{N} \sum_{i=1}^{N} y_i$. This first model assumes that the strata are independent of each other. However, subjects close to the edges of two adjacent strata may have a non-negligible correlation. For this reason, we introduce a modified model that incorporates a spatial dimension that reflects the correlation between adjacent strata. The third model we propose is a more sophisticated spatial model with a monotonically weakening spatial correlation that leads to a more precise prediction of responses for the non-sampled units. In a simulation study where the true value of the population proportion is 0.4976, the 95% HPD interval for the first model is (0.455, 0.559), for the second

model that includes spatial effect it is (0.433, 0.517), for the third model with the modified spatial effect it is (0.456, 0.537). Since the modified spatial model gives the narrowest HPD interval and the true value lies near the center of it, we conclude it is the optimal model for prediction under such a situation.

The approaches of adjusting for MAR in non-probability samples have been extensively studied. Methods vary by situations where the covariates may or may not be known for individual nonsample cases; or the situations whether or not a probability reference sample is available (Little and Rubin, 2019). Our current research interest lies in making inferences from MNAR, where inferences to the full population are normally considered to be difficult or impossible.

1.3.2 Approaches for MNAR

In this section, we review the literature for the methods that are designed for missing not at random (MNAR). We focus on research using Bayesian methods. Little and Rubin (2019) distinguished between the selection model and pattern mixture model (PMM). PMM specifies the joint distribution through the distribution of the missing data given covariates, i.e., pattern for missingness, and the conditional distribution of the measurements given the pattern. One problem of the PMM is that the parameters associated with the nonresponse pattern are not identifiable because there are no data to estimate these parameters. While the selection model and the pattern mixture model lead to the same joint density of the selection indicators and the responses, we prefer to use the selection model because it is more convenient. The sample model is a simple adjustment of the population model.

Pattern Mixture Model

In the pattern mixture approach, the population is stratified into patterns of missingness, for example, respondents and nonrespondents. A *normal pattern mixture model* can be defined by

$$y_i \mid I_i = I, x_i, \beta_0^{(I)}, \beta^{(I)}, \sigma^{(I)2} \sim \text{N}\left(\beta_0^{(I)} + \beta^{(I)} x_i, \sigma^{(I)2}\right), \quad I = 0, 1$$
$$I_i \mid x_i, \beta_0^{(I)}, \beta^{(I)} \sim \text{Bernoulli}(\Phi(w_0 + w_1 x_i)), \quad i = 1, ...N;$$

where $I = 0$ representing respond, $I = 1$ representing missing and Φ is the CDF of a standard normal distribution (Little and Rubin, 2019). This model implies that the distribution of y_i is a mixture of two normal distributions,

$$(1 - \Phi(w_0 + w_1 x_i))\left(\beta_0^{(0)} + \beta^{(0)} x_i\right) + \Phi(w_0 + w_1 x_i)(\beta_0^{(1)} + \beta^{(1)} x_i)$$

The parameters $(\beta_0^{(0)}, \beta^{(0)}, \sigma^{(0)2})$ can be directly estimated from the respondent data, however, there is no data to estimate the distribution of the nonrespondent part, that is, parameters $(\beta_0^{(1)}, \beta^{(1))}, \sigma^{(1)2})$ are not identifiable. Thus, other assumptions are needed to allow estimates of $(\beta_0^{(1)}, \beta^{(1))}, \sigma^{(1)2})$.

Foster and Smith (1998) used a pattern mixture specification for multinomial data and expressed uncertainty about ignorability by centering a nonignorable selection model on an ignorable one, a method introduced by Rubin (1977). As expected, the centering parameter is not identified by the data. For this reason, sensitivity analysis is usually conducted with the pattern mixture model. For example, the difference between respondents and nonrespondents is characterized by δ, which has a simple interpretation as the difference in means. In the sensitivity analysis, one can vary the value of δ to see how the result changes accordingly.

Based on the normal pattern-mixture model, Little et al. (2019) proposed an index of the degree of departure from ignorable sample selection, which is called the standardized measure of unadjusted bias (SMUB). However, the index depends on an inestimable parameter $\phi \in [0, 1]$,

where $\phi = 0$ indicates the missing mechanism is MAR and $\phi = 1$ means it is MNAR. Thus, computing SMUB at 0 and 1 could provide a range of the index in a sensitivity analysis. Andridge et al. (2011, 2019) proposed a similar SMUB index of the potential selection bias in estimates of population proportions. They described both maximum likelihood and Bayesian estimation approaches.

Selection Model

In the selection approach, the joint distribution of the response variable and the selection indicator variable is decomposed by the distribution for the response variable and the model for the missingness mechanism (Heckman, 1976). Using similar notation as 2.2.1, a *probit selection model* is defined by

$$
\begin{aligned}
&y_i \mid x_i, \beta_0, \beta_1, \sigma^2 \sim \mathrm{N}\left(\beta_0 + \beta_1 x_i, \sigma^2\right) \\
&I_i \mid x_i, y_i, \psi_0, \psi_1, \psi_2 \sim \text{Bernoulli}(\Phi(\psi_0 + \psi_1 x_i + \psi_2 y_i)), \quad i = 1, ...N.
\end{aligned}
$$

As one might imagine, it is impossible to estimate the parameter ψ_2 by maximum likelihood methods when some of the y_i are not observed. However, as we mentioned at the beginning of this subsection, it is estimable under Bayesian paradigm since this parameter and data are not separated. Greenlees, Reece, and Zieschang (1982) developed a normal-logistic regression model, a nonignorable nonresponse model within the selection approach, to impute missing values in the Current Population Survey when the probability of response depends on the variable being imputed.

Nandram and Choi (2002a, 2002b) proposed a hierarchical Bayesian selection model for binary nonresponse data. To bypass the weak identifiability of the parameters, they 'borrow strength' across areas (states) as in small area estimation. The uncertainty about ignorability is accommodated by a centering parameter with different priors. They conclude that although there are differences for the centering parameters under different priors, inference about the proportion in each

area is very similar.

Nandram and Choi (2005) use both the pattern mixture approach and the selection approach to analyze the NHANES III data and study the finite population mean BMI. For the pattern mixture approach, they use a similar idea as Nandram and Choi (2002a) to "center" the nonignorable nonresponse model on the ignorable nonresponse model with some variation. They discover that the parameters become unidentifiable if there is a substantial difference between the respondents and the nonrespondents. For the selection approach, they use a normal-logistic model, where the response propensity is related to BMI only, and use a hierarchical Bayesian model to accommodate clustering within counties. They further refine the model by using a spline regression to model BMI.

Nandram and Choi (2010) extend the basic nonignorable nonresponse model of Nandram and Choi (2005) to assess the finite BMI population percentiles of domains formed by age, race, and sex within counties. They fit many models to investigate sensitivity and robustness to the assumption of nonignorable nonresponse and selection.

1.3.3 Multi-level Modelling under MNAR

Pfeffermann, Moura, and Silva (2006) proposed a model-dependent approach that accounts for selection not at random in multi-level sample surveys. The methodology they proposed first extracts the hierarchical model holding for the sample data given the selected sample, and then fit the resulting sample model using Bayesian methods. This approach allows the sample selection probabilities to serve as additional data that possibly strengthen the estimators. Even though in this paper the authors made inference only for superpopulation parameters, they also mentioned that the sample model can be used to predict the small area means in sampled and nonsampled areas, which is our target inference in Chapter 3.

Sverchkov and Pfeffermann (2018) extend the model to consider the MNAR at the unit level. The proposed extension consists of first identifying the model holding for the observed responses

and using the model for estimating the response probabilities, and then applying the approach of Pfeffermann and Sverchkov to the observed data with the unit sampling probabilities replaced by the products of the sampling probabilities and the estimated response probabilities. They further proposed a methodology to compute the estimates and calculate the empirical MSE over bootstrap samples.

1.4 Statistical Background

In this section, we discuss two statistical concepts that will be used in our model.

1.4.1 Binary Data

In statistics, binary dataset is a data type consisting of categorical data that can only take two possible values, such as "yes" and "no", or "presence" and "absence". A binary variable is a random variable of binary type that follows a Bernoulli distribution. A binary regression estimates a relationship between covariates and a single output binary variable. In binary regression, one of the two alternatives is considered as "success", and the probability of success is modeled. One of the most common binary regression models is logistic regression, and it is enormously popular in all sciences. In this dissertation, our main focus is in binary data from a nonprobability samples, i.e., "normal" or "abnormal" from BMI data.

1.4.2 Logistic Regression

Logistic regression is used in various fields, including machine learning, social sciences, and most medical fields. For example, it can be used in predicting the risk of developing a given disease, or used in calculating a customer's propensity to purchase a product.

The model assumes a linear relationship between the predictor variables (covariates) and the

log-odds of the event that $Y = 1$,

$$\ell = \log \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where ℓ is the log-odds, β_i are parameters of the model. This can also be written as $\text{logit}(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

By simple algebraic manipulation, the probability that $Y = 1$ is

$$P = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)},$$

which is the expression we use in our model.

1.5 Notation

Following Smith (1983), we consider the joint density of the response variables, $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_N)$ and the selection indicator variables, $\boldsymbol{I} = (I_1, I_2, ..., I_N)$. In our study, we assume the covariates $\boldsymbol{X}$ on the population level are fully observed or collected from external sources, for example, from the U.S. Census Bureau (https://www.census.gov). Suppose we have $p - 1$ covariates, then, the design matrix $\boldsymbol{X}$ has the dimension $N \times p$. Let n be the sample size and N be the population size. Without loss of generality, let $y_i, i = 1, ..., n$, denote the sample responses, and $y_i, i = n+1, ..., N$, denote the non-sampled units; accordingly, $I_i = 1$ for $i = 1, ...n$ and $I_i = 0, i = n+1, ..., N$. Note that we do not differentiate non-respondents from non-sampled units, that is, we treat the subjects that have been selected and do not respond as part of the non-sampled units. The finite population proportion, denoted by $\bar{Y} = (1/N) \sum_{i=1}^{N} y_i$, is the quantity of interest for the homogeneous model, and for the heterogeneous model the target of inferences are the finite population proportion for each area.

1.6 Contributions of this Research and Dissertation Organization

In our research, we present a methodology of using the selection model and fully Bayesian analysis for non-ignorable missing data mechanism. Our model setup is double logistic regresson, where the binary outcome y_i is modeled by

$$\text{logit}(P(y_i = 1)) = \boldsymbol{x}_i'\boldsymbol{\gamma};$$

and the model of missingness, which is denoted by I, is given by another logistic regression model with

$$\text{logit}(P(I_i = 1)) = \boldsymbol{x}_i'\boldsymbol{\beta} + \beta_{p+1}y_i.$$

The parameter β_{p+1} has been known for being 'non-identifiable' by many researchers. For example, Wang, Bartlett, and Ryan (2017) investigate the likelihood of the model parameters and show that the model is non-identifiable. Although the model setup is similar, our model incorporates covariates $\boldsymbol{x}$ which help to identify the parameter β_{p+1} by fully Bayesian analysis. We explore the performance of our methodology using simulated binary response data under different bias scenarios (sample proportion underestimate/overestimate the population proportion under MNAR). The simulation results show that the nonignorable selection model we proposed correctly specifies the parameter values and eliminates the sampling bias. Our preceding work has been presented in JSM 2019 and published (Xu and Nandram, 2019; Xu and Nandram, 2020).

We further apply our model to the BMI data from NHANES III (Nandram and Choi, 2010), in which it is reasonable to think that individuals will be less likely to respond if their BMI data are abnormal. In this application, we are using a sample of 761 subjects to make an inference for a population of 630,308 with three categorical covariates. To bypass the Markov chain Monte Carlo (MCMC) mixing problem that could come from the large population size and a small num-

ber of distinct covariates, we propose a more efficient MCMC algorithm that gives satisfactory performance.

When there are cluster area effects in the data and these are exchangeable, one could use a small area model (SAE) for the heterogeneity (Rao and Molina, 2015). Using double regression as the nonignorable selection model in SAE is an unexplored domain. We add area effects to the response model, denoted by ν_{1i}; and to the missingness model, denoted by ν_{2i}, and allow them to have their parametric distributions respectively. We use the same structure as in the model proposed previously, but update the model for responses by

$$\begin{aligned}
&\text{logit}(P(y_{ij}=1)) = \boldsymbol{x}_{ij}'\boldsymbol{\gamma} + \nu_{1i}, i = 1, ..., \ell, j = 1, ..., N_i, \\
&\nu_{1i} \mid \sigma_1^2 \overset{iid}{\sim} \text{N}(0, \sigma_1^2), \quad i = 1, ..., \ell,
\end{aligned}$$

and the model for missingness by

$$\begin{aligned}
&\text{logit}(P(I_{ij}=1)) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \beta_{p+1}y_{ij} + \nu_{2i}; i = 1, ..., \ell, j = 1, ..., N_i, \\
&\nu_{2i} \mid \sigma_2^2 \overset{iid}{\sim} \text{N}(0, \sigma_2^2), \quad i = 1, ..., \ell,
\end{aligned}$$

where i is the index for areas, and j is the index for units within areas.

Using a common parametric distribution facilitates a 'borrowing of strength' across the ensemble. However, when there is a preference for a more robust approach, one could prefer a nonparametric Bayesian approach (Nandram and Choi, 2004; Nandram and Yin, 2016a, 2016b; Yin and Nandram, 2020a, 2020b).

In **Chapter 2**, we focus on the Bayesian selection model for homogeneous data. We develop a nonignorable selection model and a MCMC algorithm to sample the unknown parameters and non-sampled units iteratively. We compare several other models with the proposed model, including an ignorable selection model, H-T estimator, and Hajek estimator through a simulation study. Through an application to BMI data, it is shown that the nonignorable selection model and

ignorable selection model give different predictions, indicating the sample bias is related to outcomes (MNAR). An extended research on scenarios where individual covariates are unknown is conducted, as well as the impact of priors in the model.

We extend the idea of the nonignorable model for homogeneous data to a model for heterogeneous data in **Chapter 3**, employing the formulation of small area estimation (SAE) in the two submodels of the nonignorable selection model. BMI data with 35 counties are adjusted with both the ignorable and nonignorable selection model, with a clearer separation in the finite population proportions. We explore two variations of the model, specifically incorporating the Dirichlet process prior and bivariate prior into the model.

We give concluding remarks and propose future work in **Chapter 4**. Specifically, we describe the main area in the dissertation, homogeneous and heterogeneous data, available in the form of a single nonprobability sample with appropriate covariates.

Chapter 2

Bayesian Selection Model for Homogeneous Data

Using Bayesian methodology for missing data has gained popularity over recent years. The Bayesian paradigm is appealing because missing data can be treated as hidden parameters and are estimated within each MCMC iteration. In this chapter, we focus on homogeneous data and propose both the ignorable and nonignorable Bayesian selection model (Section 2.1). In Section 2.2, a simulation study is conducted to test both models. In Section 2.3, we apply the BMI data from NHANES III to the two selection models proposed. In Section 2.4, we study how different priors affect the model's performance. Finally in Section 2.5, we discuss a more practical situation where covariates are unobserved for the non-sampled units.

2.1 Selection Model Specification

In this section, we first introduce a nonignorable selection model that is based on double regression and the selection model. Then, as a comparison, we design an ignorable selection model that uses a single logistic regression. We test the performance of both these models by simulated data in the

next section.

2.1.1 Nonignorable Selection Model

The selection model factors the joint likelihood of $\boldsymbol{Y}$ and $\boldsymbol{I}$ into the complete-data model for $\boldsymbol{Y}$ and the model for the missing data mechanism. The complete-data model specifies the relationship between responses and covariates from an underlying population. For binary response variable $\boldsymbol{Y}$, we use a logistic regression that can be expressed as follows,

$$y_i \mid \boldsymbol{\gamma} \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}\right\}, \quad i = 1, ..., N, \tag{2.1.1}$$

where $\boldsymbol{\gamma}$ is a coefficient vector with the length p. This is the population model and it gives a description of the responses where there is no selection bias.

The model of missingness is given by a logistic regression

$$I_i \mid \boldsymbol{y}, \boldsymbol{\beta} \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_i)}\right\}, \quad i = 1, ..., N. \tag{2.1.2}$$

We see the missing data mechanism is MNAR because the probability of missingness depends on all values of y, i.e., both the observed and unobserved values. Denote the vector of coefficients associated with $\boldsymbol{X}$ by $\boldsymbol{\beta}_{(p)}$ and the coefficient for $\boldsymbol{Y}$ in (2.1.2) by β_{p+1}. When it is not necessary to split into $\boldsymbol{\beta}_{(p)}$ and β_{p+1}, we use $\boldsymbol{\beta}$ to represent the combined parameter vector.

The joint likelihood of the i^{th} subject can be obtained by combing these two models, as follows,

$$
\begin{aligned}
& f(y_i, I_i|\boldsymbol{\beta}, \boldsymbol{\gamma}) \\
& \propto \left(\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}\right)^{y_i} \left(1-\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}\right)^{1-y_i} \left(\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)}+\beta_{p+1}y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)}+\beta_{p+1}y_i)}\right)^{I_i} \\
& \times \left(1-\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)}+\beta_{p+1}y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)}+\beta_{p+1}y_i)}\right)^{1-I_i}, \quad i=1,...,N.
\end{aligned} \tag{2.1.3}
$$

To improve the mixing performance in MCMC, we use a vague prior – multivariate normal $N(\hat{\boldsymbol{\beta}}, 10^2\hat{\Sigma}_{\boldsymbol{\beta}})$ to be the prior for $\boldsymbol{\beta}$, where $\hat{\boldsymbol{\beta}}$ is the mode of the likelihood for $\boldsymbol{\beta}$, and $\hat{\Sigma}$ is the inverse of the Hessian of its negative log-likelihood. A similar prior applies to $\boldsymbol{\gamma}$. Note we multiply the estimated covariance matrix by 10^2 to avoid the issues of the double use of data and making the prior vague. We tried different values aside from 10^2, i.e., 10, 10^3, and they all give similar results. Other forms of priors are investigated in Section 2.4.

Denote the non-sampled units by $\boldsymbol{y}_{ns}$, and the sampled ones by $\boldsymbol{y}_s$. Taking $I_i = 1$ for $i = 1, ..., n$, and $I_i = 0$ for $i = n+1, ..., N$, the joint posterior distribution is

$$
\begin{aligned}
& f(\boldsymbol{y}_{ns}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{I}, \boldsymbol{y}_s) \\
& \propto \prod_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)}+\beta_{p+1}y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)}+\beta_{p+1}y_i)} \prod_{i=n+1}^{N} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)}+\beta_{p+1}y_i)} \\
& \times \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(10^2\hat{\Sigma}_{\boldsymbol{\beta}})^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right) \exp\left(-\frac{1}{2}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})^T(10^2\hat{\Sigma}_{\boldsymbol{\gamma}})^{-1}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})\right).
\end{aligned} \tag{2.1.4}
$$

In Appendix B, we provide a proof for the propriety of a more general case of (2.1.4), that is, using a flat prior for $\boldsymbol{\gamma}$.

Next, our goal is to apply MCMC methods to draw samples of $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and the non-sampled units y_i, $i = n+1, ..., N$ from the posterior distribution. Our target of inference is the finite population

proportion that is based on the combination of the sampled Y values and the draws of the non-sampled Y values.

Conditional Posterior Distributions

To use the Metropolis-Hastings sampler to draw samples of the unknown parameters, we first need to write down their conditional posterior distributions. It is easy to derive the conditional posterior densities of vector $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The derivation of the non-sampled unit y_i is less straightforward, the proof of which is presented in Appendix A. The conditional posterior densities are

$$\text{(i)}\ \ f(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{I}) \propto \prod_{i=1}^{N} \frac{\exp\left\{(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i) I_i\right\}}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\beta}})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right);$$

$$\text{(ii)}\ \ f(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{I}) \propto \prod_{i=1}^{N} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma} y_i)}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \exp\left(-\frac{1}{2}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\gamma}})^{-1} (\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})\right);$$

$$\text{(iii)}\ \ P(y_i = 1 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}) \left\{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)})\right\}}{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}) \left\{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)})\right\} + 1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1})},$$

$$P(y_i = 0 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) = 1 - P(y_i = 1 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}), \quad i = n+1, ..., N.$$

Note that the density of $\boldsymbol{\beta}$ or $\boldsymbol{\gamma}$ is not in a simple form, corresponding to nonstandard probability densities, while the conditional posterior distribution of $y_i, i = n+1, ..., N$, has a standard probability density – Bernoulli distribution. Thus, the Metropolis-Hastings sampler is run in three blocks, one for $\boldsymbol{\gamma}$, one for $\boldsymbol{\beta}$, and one for y_i, the standard probability density. The detailed algorithm can be found in below.

Algorithm: For $(t = 1, ..., T)$, repeat:

1. Draw $\boldsymbol{\beta}^*$ from multivariate t-distribution, where the proposal mean vector is the mode of (i), and the covariance matrix is given by its inverse of the Hessian of the negative log-likelihood function. Keep the degrees of freedom ν fixed throughout the iterations. It is standard to tune the Metropolis step by varying ν. Denote this proposal density by $g(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{I})$.

2. Compute the acceptance ratio

$$r = \frac{f(\boldsymbol{\beta}^* \mid \boldsymbol{y}, \boldsymbol{I}) g(\boldsymbol{\beta}^{(t-1)} \mid \boldsymbol{y}, \boldsymbol{I})}{g(\boldsymbol{\beta}^* \mid \boldsymbol{y}, \boldsymbol{I}) f(\boldsymbol{\beta}^{(t-1)} \mid \boldsymbol{y}, \boldsymbol{I})}.$$

3. Draw u from uniform $\boldsymbol{U}(0,1)$, if $u \leq \min(1, r)$, set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^*$, else, set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$. Count the number of jumps.

4. Similarly, adapt steps 1-3 to draw $\boldsymbol{\gamma}^{(t)}$ based on formula (ii).

5. For each $i, i = n+1, ..., N$, calculate the probability of success by formula (iii), and draw the corresponding y_i from its Bernoulli distribution. Then, the finite population proportion, denoted by $\bar{Y}$, can be obtained by finding the proportion of combined samples and the predicted nonsamples.

We need to check if the jumping rates for $\boldsymbol{\beta}$ or $\boldsymbol{\gamma}$ are within the range of (25%, 75%). If the jumping rate is not acceptable, i.e. too low or too high, we tune the parameter by varying the corresponding ν and repeat the above steps. Note that the value of ν can be different or the same for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, here, we use the same ν. If the jumping rate is acceptable, then we further burn-in and thin the MCMC chain and check the diagnostics.

2.1.2 Ignorable Selection Model

If the missingness scheme is MAR or MCAR, we would use a single logistic regression as the ignorable selection model for binary response data. Here, we use it as our baseline model to compare results. The model with the prior for $\boldsymbol{\gamma}$ is

$$\begin{aligned} & y_i \mid \boldsymbol{\gamma} \overset{ind}{\sim} \text{Bernoulli} \left\{ \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\gamma})}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\gamma})} \right\}, \quad i = 1, ..., n, \\ & \pi(\boldsymbol{\gamma}) \propto 1. \end{aligned} \tag{2.1.5}$$

We note that others have used Jeffrey's prior for $\pi(\boldsymbol{\gamma})$, (for example, Chen, Ibrahim and Kim, 2008), but for ease of calculation we do not use it here.

The posterior distribution of $\boldsymbol{\gamma}$ is

$$f(\boldsymbol{\gamma} \mid \boldsymbol{y}_s) \propto \prod_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\gamma} y_i)}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\gamma})}. \tag{2.1.6}$$

The proof of its propriety is the same as a part of the proof for (2.1.4) using the idea of log-concavity, as shown in Appendix B.

Algorithm:

1. For $(t = 1, ..., T)$, repeat:

   (i) Draw $\boldsymbol{\gamma}^*$ from multivariate t-distribution, where the proposal mean vector is the mode of (2.1.6), and the covariance matrix is given by its inverse of the Hessian of the negative log-likelihood function. Keep the degrees of freedom ν fixed throughout the iterations. Denote this proposal density by $g(\boldsymbol{\gamma} \mid \boldsymbol{y}_s)$.

   (ii) Compute the acceptance ratio

   $$r = \frac{f(\boldsymbol{\gamma}^* \mid \boldsymbol{y}_s) g(\boldsymbol{\gamma}^{(t-1)} \mid \boldsymbol{y}_s)}{g(\boldsymbol{\gamma}^* \mid \boldsymbol{y}_s) f(\boldsymbol{\gamma}^{(t-1)} \mid \boldsymbol{y}_s)}.$$

   (iii) Draw u from uniform $\boldsymbol{U}(0, 1)$, if $u \leq \min(1, r)$, set $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^*$, else, set $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^{(t-1)}$. Count the number of jumps.

2. After finishing the T iterations, check if the jumping rate for $\boldsymbol{\gamma}$ is within the range of (25%, 75%). If the jumping rate is not acceptable, i.e. too low or too high, we tune the parameter by varying ν and repeat the above steps. If the jumping rate is acceptable, we further burn-in and thin the MCMC chain to obtain M sets of samples and check the diagnostics.

3. For each set of $\boldsymbol{\gamma}$, and each item in the non-sampled set, calculate the probability of success by formula (5), denoted by p_i, then draw the corresponding y_i from Bernoulli (p_i). When finished for all non-sampled units, one sample of $\bar{Y}$ can be obtained by finding the proportion of combined samples and y_i's, $i = n+1, ..., N$. Repeat this step for all sets of $\boldsymbol{\gamma}$ to obtain the posterior predictive distribution of population proportion.

2.2 Simulation Study

In this section, we conduct a simulation study with three different scenarios and compare results among the nonignorable selection model, the ignorable selection model, the Horvitz-Thompson estimator, and the Hajek estimator. In Section 2.2.1, we describe how datasets are generated under three scenarios. In Section 2.2.2, we give the results and diagnostics under the three scenarios by using different models. In Section 2.2.3, repeated simulations are provided by using the nonignorable selection model. In Section 2.2.4, we apply log pseudo marginal likelihood (LPML) to compare the nonignorable and the ignorable selection model.

2.2.1 Dataset Description

We run a simulation study to test our model. First, we generate values for the covariates, as an example, *age*, *race*, *sex*, *education* from distributions listed in Table 2.1. *Age* values are drawn from a normal distribution with mean 50 and standard deviation 5; *race*, *sex*, and *education* are drawn from Bernoulli distributions with probabilities of 0.45, 0.3, and 0.4 respectively. We draw $N = 10,000$ from each covariate to create the population.

Then, the values for the response variable $\boldsymbol{Y}$ are generated by the distribution below for all 10,000 units,

$$y_i|\boldsymbol{\gamma} \sim \text{Bernoulli}\{e^{\boldsymbol{x}_i'\boldsymbol{\gamma}}/(1+e^{\boldsymbol{x}_i'\boldsymbol{\gamma}})\}, i = 1, ..., N,$$

Table 2.1: Distributions for covariates

Covariate	Distribution
age	$N(50, 5^2)$
race	Bernoulli (0.45)
sex	Bernoulli (0.3)
education	Bernoulli (0.4)

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$, with the values of $\gamma_0 = 0.4, \gamma_1 = 3, \gamma_2 = 6, \gamma_3 = -2, \gamma_4 = -6$.

The samples are defined by $\{(y_i, \boldsymbol{x}_i) : I_i = 1\}$, and to test on different missingness mechanisms, we generate I_i by

$$I_i|\boldsymbol{\beta} \sim \text{Bernoulli}\{e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{(4)}+\beta_5 y_i}/(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_{(4)}+\beta_5 y_i})\}, i = 1, ..., N,$$

where the coefficients $\boldsymbol{\beta}$ are specified differently under 3 scenarios as follows.

Scenario 1: Nonignorable nonresponse (overestimation)

We take $\beta_0 = -2.2, \beta_1 = -0.6, \beta_2 = -1, \beta_3 = -0.5, \beta_4 = -1, \beta_5 = 1.5$. Since we have 4 covariates in this example, $\boldsymbol{\beta}_{(4)}$ is a 5 dimensional vector (including intercept), and β_5 is the β_{p+1} in the formulation for the nonignorable selection model. It has been noticed that this coefficient could be poorly identified, so we closely monitor it. In this case, $\beta_5 = 1.5$, a positive value, showing that the selection probability for $y = 1$ is greater than the probability for $y = 0$, which indicates the sample proportion will be greater than the population proportion.

As expected, the sample proportion is 0.69 with the sample size of $n = 1148$, while the true population proportion is 0.51. The sample proportion overestimates the population proportion by 35%.

Scenario 2: Ignorable nonresponse (MAR)

We take $\beta_0 = -2, \beta_1 = 2, \beta_2 = -3, \beta_3 = -1, \beta_4 = -0.5, \beta_5 = 0$. Here, $\beta_5 = 0$ means that the missingness indicator is not related to y, hence the missingness mechanism is MAR.

Using the selection submodel with above coefficients, we generated $n = 1078$ samples. The sample proportion is 0.67, and the true population proportion is 0.50. The samples are still biased because other coefficients in the selection submodel are not zero.

Scenario 3 Nonignorable nonresponse (underestimation)

We take $\beta_0 = -2, \beta_1 = 0.5, \beta_2 = 0, \beta_3 = -0.8, \beta_4 = 1, \beta_5 = -1$. As opposed to Scenario 1, here β_5 is negative, hence we expect the sample proportion will underestimate the population proportion.

In this scenario, with 1153 samples, the sample proportion becomes 0.33 while the population proportion is 0.50. The sample proportion underestimates the population proportion by 34%.

2.2.2 Comparison and Results

In this part, we run one simulation for each scenario on the nonignorable selection model, the ignorable selection model, the Horvitz-Thompson estimator, and the Hajek estimator. We focus on the detailed diagnostics and population proportion predictions among models. In the following subsection, we perform repeated simulations on the nonignorable selection model to see the consistency of the model's performance.

Nonignorable Model

Through trial and error, we use 30,000 iterations, burning-in the first 5,000, and thinning by 25 to obtain the MCMC samples of size 1,000, which gives a stationary distribution. The jumping rates are around 0.68 for $\boldsymbol{\gamma}$, and 0.61 for $\boldsymbol{\beta}$, which are acceptable as these are between 0.25 and 0.75. Our program is written using R language, and the execution time is about 110 minutes on my personal computer (processor is 3.7 GHz 6-Core Intel Core i5 and memory is 8 GB 2667 MHz DDR4).

Geweke's Diagnostics are also conducted to check the convergence of each parameter to its target distribution, with p-values and effective sample sizes (ESS) shown in Table 2.2. We see that all the p-values for the Z tests are greater than 0.05, and ESS for each parameter is adequate. It is

worth noting that the ESS is calculated by the formula for multivariate Markov chain as described in Vats et al. (2015), which is given by

$$\text{ESS} = n\frac{|\Lambda|^{1/p}}{|\Sigma|^{1/p}},$$

where Λ is the sample covariance matrix, and Σ is the asymptotic covariance matrix in the Markov chain central limit theorem, which is different from Λ since samples are correlated, and p is the dimension of the posterior.

Table 2.2 provides the posterior summaries, p-values for testing stationary in Geweke's Diagnostic as well as the ESS for all coefficients. Recall that the HPD intervals for β_5 have different relationships with 0 in these scenarios, which are greater than 0, containing 0 and less than 0 respectively. The results are in line with the missing data mechanisms in these three scenarios. Note that the p-value for γ_2 and β_3 in scenario 2 are a bit small, which indicates more iterations may required for scenario 2.

Table 2.3 depicts the posterior summaries of $\bar{Y}$, along with the true population proportions and sample proportions. The comparison of the population proportion prediction across all models are given in Table 2.6. We see that the HPD interval of posterior mean contains the true population proportion in each scenario, suggesting the models successfully adjust the sample bias and predict the population proportion.

Ignorable Model

For consistency purposes, we also use 30,000 as the number of iterations, then burn-in the first 5,000, thining by 25, and end up with 1,000 samples of γ. Then, we obtain 1,000 samples of $\bar{Y}$, with each one calculated by the draws of $\boldsymbol{y}_{ns}$ from the distribution with the corresponding γ value. The Geweke's test and effective sample sizes for $\boldsymbol{\gamma}$ are shown in Table 2.4. The whole procedure finished within a minute, and as we can see from Table 2.4, the MCMC chain mixes rapidly.

Table 2.2: Posterior summaries and diagnostics of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ for the nonignorable selection model by scenario

	$\boldsymbol{\gamma}$ (True)	PM (PSD)	95% HPD	p-value	$\boldsymbol{\beta}$ (True)	PM (PSD)	95% HPD	p-value
S1	γ_0 (0.4)	0.33 (0.20)	(-0.07, 0.72)	0.74	β_0(-2.2)	-2.19 (0.09)	(-2.37, -1.98)	0.81
	γ_1 (3)	3.13(0.22)	(2.67, 3.55)	0.20	β_1(-0.6)	-0.58 (0.04)	(-0.67, -0.49)	0.41
	γ_2 (6)	6.36 (0.47)	(5.51, 7.32)	0.11	β_2 (-1)	-1.09 (0.10)	(-0.28, -0.88)	0.80
	γ_3 (-2)	-2.07 (0.27)	(-2.61, -1.58)	0.49	β_3 (-0.5)	-0.52 (0.07)	(-0.66, -0.37)	0.23
	γ_4 (-6)	-6.89 (0.53)	(-7.87, -5.78)	0.53	β_4 (-1)	-0.92 (0.11)	(-1.11, -0.71)	0.63
	ESS	624			β_5 (1.5)	1.56 (0.17)	(1.21, 1.88)	0.59
					ESS	723		
S2	γ_0 (0.4)	0.79 (0.24)	(0.32, 1.22)	0.56	β_0 (-2)	-1.70 (0.17)	(-2.04, -1.38)	0.33
	γ_1 (3)	2.56 (0.22)	(2.15, 2.99)	0.54	β_1 (2)	2.01 (0.07)	(1.88, 2.15)	0.26
	γ_2 (6)	5.04 (0.85)	(3.52, 6.67)	0.06	β_2 (-3)	-2.97 (0.12)	(-3.20, -2.76)	0.99
	γ_3 (-2)	-1.81 (0.28)	(-2.39, -1.28)	0.28	β_3 (-1)	-1.03 (0.09)	(-1.21, -0.87)	0.01
	γ_4 (-6)	-5.65 (0.34)	(-6.30, -5.00)	0.35	β_4(-0.5)	-0.85 (0.17)	(-1.17, -0.52)	0.19
	ESS	515			β_5 (0)	-0.30 (0.24)	(-0.78, 0.12)	0.37
					ESS	1022		
S3	γ_0 (0.4)	0.28 (0.23)	(-0.18, 0.75)	0.74	β_0 (-2)	-1.86 (0.10)	(-2.04, -1.64)	0.20
	γ_1 (3)	3.12(0.21)	(2.74, 3.55)	0.31	β_1 (0.5)	0.50 (0.05)	(0.41, 0.59)	0.45
	γ_2 (6)	5.98 (0.39)	(5.22, 6.74)	0.84	β_2 (0)	0.13 (0.10)	(-0.05, 0.33)	0.89
	γ_3 (-2)	-2.01 (0.28)	(-2.57, -1.46)	0.82	β_3 (-0.8)	-0.85 (0.07)	(-1.00, -0.71)	0.83
	γ_4 (-6)	-5.69 (0.38)	(-6.47, -4.97)	0.56	β_4 (1)	0.80 (0.09)	(0.64, 0.98)	0.19
	ESS	758			β_5 (-1)	-1.14 (0.17)	(-1.49, -0.81)	0.29
					ESS	762		

Note: Scenario 1 (S1): nonignorable nonresponse (overestimation); Scenario 2 (S2): ignorable nonresponse; Scenario 3 (S3): nonignorable nonresponse (underestimation); PM: posterior mean; PSD: posterior standard deviation

Table 2.3: Posterior summaries of $\bar{Y}$ for the nonignorable selection model by scenario

	True $\bar{Y}$	$\bar{\boldsymbol{y}}_s$	Posterior Mean (SD)	95% HPD
Scenario 1	0.51	0.69	0.50 (0.02)	(0.47, 0.54)
Scenario 2	0.50	0.67	0.52 (0.03)	(0.46, 0.57)
Scenario 3	0.50	0.33	0.51 (0.01)	(0.48, 0.53)

The posterior mean of the predictive population proportions under three scenarios are 0.61, 0.50, and 0.44 respectively when their true values are 0.51, 0.50, and 0.50. In Table 2.5, we see that HPD intervals by the ignorable selection model do not contain the true population proportion in Scenario 1 and Scenario 3, the interval barely contains the true value in Scenario 2.

The histograms of population proportion predictions under both ignorable selection and nonignorable selection models in each scenario are given in Figure 2.1. We see that for Scenario 1 and 3, the predictions by the nonignorable selection model are centered around the true value (dashed line), while the predictions by the ignorable model are far away from the true population proportion. When the missingness mechanism is MAR, i.e., Scenario 2, both histograms cover the true value, but the one by the ignorable selection model is more equally balanced around the true value, indicating a better prediction.

Table 2.4: Posterior summaries and diagnostics of $\boldsymbol{\gamma}$ for the ignorable selection model by scenario

	$\boldsymbol{\gamma}$	PM (PSD)	95% HPD	p-value
S1	γ_0	1.46 (0.15)	(1.17, 1.74)	0.16
	γ_1	3.19 (0.22)	(2.75, 3.62)	0.29
	γ_2	7.03 (0.54)	(5.96, 8.15)	0.89
	γ_3	-1.60 (0.34)	(-2.27, -0.94)	0.10
	γ_4	-6.11 (0.53)	(-7.20, -5.14)	0.72
	ESS	1000		
S2	γ_0	0.53 (0.15)	(0.21, 0.82)	0.13
	γ_1	2.67 (0.21)	(2.27, 3.08)	0.21
	γ_2	5.04 (0.88)	(3.31, 6.77)	0.10
	γ_3	-1.89 (0.27)	(-2.45, -1.37)	0.84
	γ_4	-5.66 (0.34)	(-6.33, -5.01)	0.75
	ESS	903		
S3	γ_0	-0.74 (0.19)	(-1.09, 0.37)	0.06
	γ_1	3.23 (0.22)	(2.80, 3.66)	0.08
	γ_2	6.11 (0.42)	(5.38, 6.98)	0.10
	γ_3	-1.92 (0.30)	(-2.53, -1.37)	0.29
	γ_4	-5.77 (0.42)	(-6.58, -4.97)	0.31
	ESS	935		

Note: Scenario 1 (S1): nonignorable nonresponse (overestimation); Scenario 2 (S2): ignorable nonresponse; Scenario 3 (S3): nonignorable nonresponse (underestimation); PM: posterior mean; PSD: posterior standard deviation

Table 2.5: Posterior summaries of $\bar{Y}$ for the ignorable selection model by scenario

	True $\bar{Y}$	$\bar{\boldsymbol{y}}_s$	Posterior Mean (SD)	95% HPD
Scenario 1	0.51	0.69	0.61 (0.01)	(0.59, 0.63)
Scenario 2	0.50	0.67	0.50 (0.03)	(0.45, 0.55)
Scenario 3	0.50	0.33	0.44 (0.01)	(0.43, 0.46)

Horvitz-Thompson Estimator

The Horvitz -Thompson (H-T) estimator and the following Hajek estimator are both inverse probability weighted (IPW) estimators, where the probability could be defined by

$$Pr_i = \frac{n\pi_i}{\sum_{j=1}^{N} \pi_j}, i = 1, ..., n$$

Note that $\sum_{i=1}^{N} Pr_i = n$, and one can have $Pr_i > 1$. Here, π_i is the propensity score for each subject and it is calculated by

$$\pi_i = P(R_i = 1|x_i; \boldsymbol{\alpha}) = \frac{\exp(x_i'\boldsymbol{\alpha})}{1 + \exp(x_i'\boldsymbol{\alpha})}, i = 1, 2, ...N,$$

where $\boldsymbol{\alpha}$ is the maximum likelihood estimate from samples. Notice the selection probability formulation does not incorporate variable Y, hence the estimators using it are designed for MAR mechanisms.

Horvitz-Thompson (H-T) estimator (1952) has long been used as an unbiased estimator when the weights of samples are correctly specified. The expression to calculate the population mean is

$$\hat{\bar{Y}}_{HT} = \frac{1}{N} \sum_{i=1}^{n} \frac{y_i}{Pr_i},$$

where Pr_i could be the same probability defined above.

Results by the H-T estimator for all 3 scenarios are presented in Table 2.6. We see that the estimator corrects the sampling bias in the right direction for Scenario 1 and 3, but the performance is no better than the nonignorable nor the ignorable selection model. For Scenario 2, the estimator seems to overcorrect the bias.

Note the estimator gives only a single value, as opposed to an interval given by Bayesian methods. We obtain approximate standard errors and confidence intervals for the H-T and Hajek

estimators by bootstrapping $\{(Pr_i, y_i), i = 1, ..., n\}$.

Hajek Estimator

Hajek estimator of the population mean, defined by Sarndal, Swensson, and Wretman (2003) is

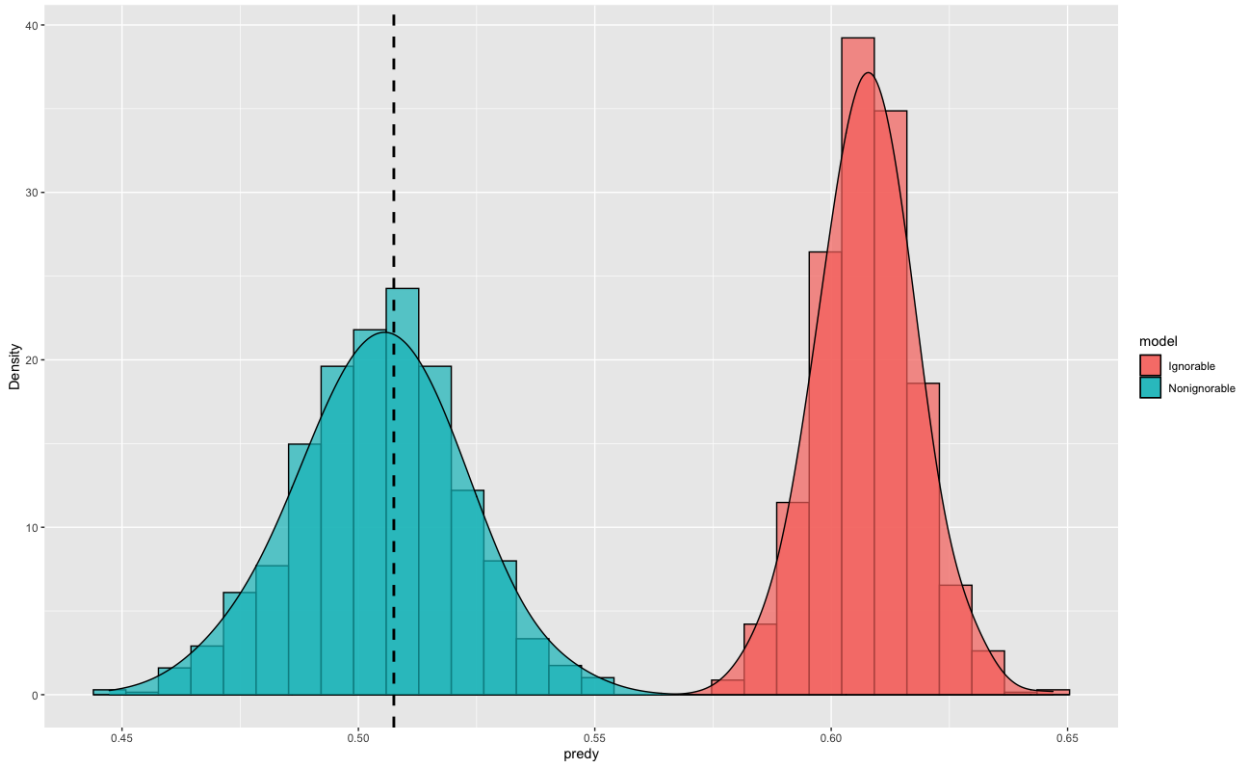
$$\hat{\bar{Y}}_H = \frac{\sum_{i=1}^{n} y_i/Pr_i}{\sum_{i=1}^{n} 1/Pr_i},$$

where Pr_i is the same selection probability used for H-T estimator. The results are provided in Table 2.6.

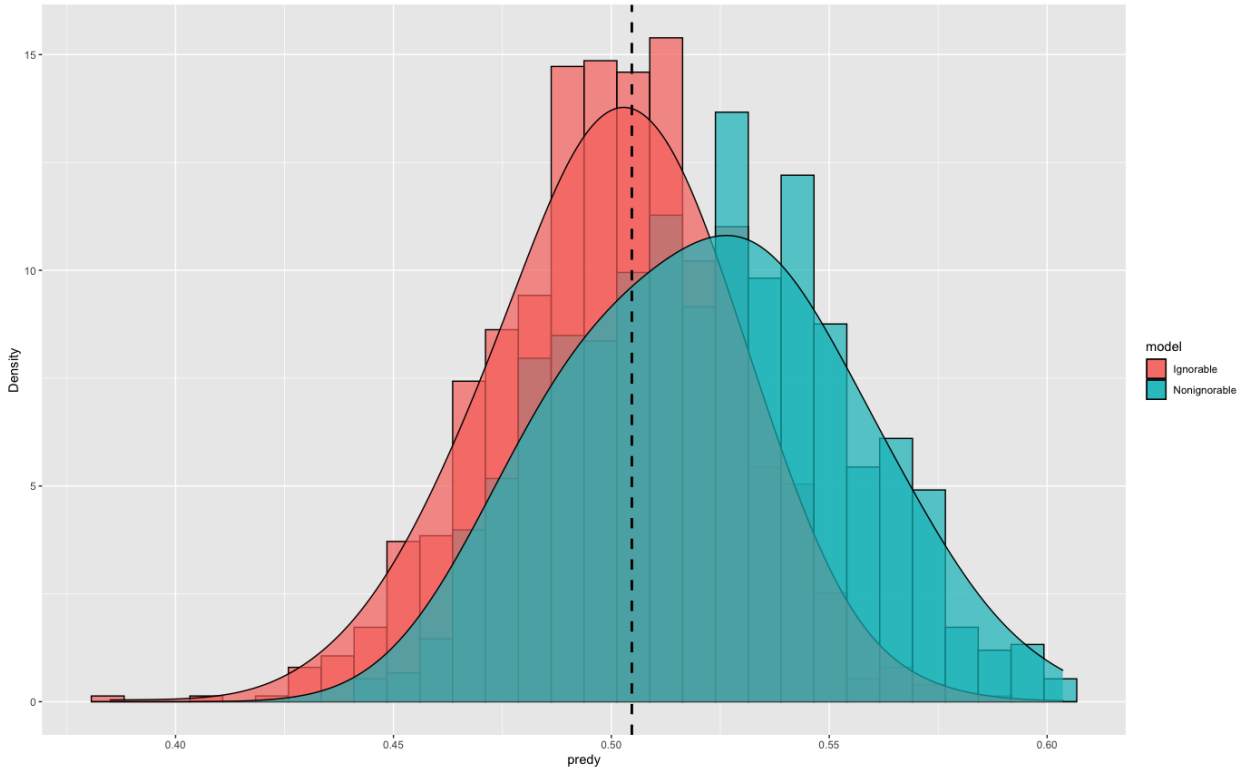
The estimates from the Hajek estimator are generally better or no worse than the H-T estimator. It is also worth noting that under Scenario 2, the Hajek estimator gives an estimate close to the ignorable model, but with a wider 95% HPD interval.

Table 2.6: Results comparison by scenario

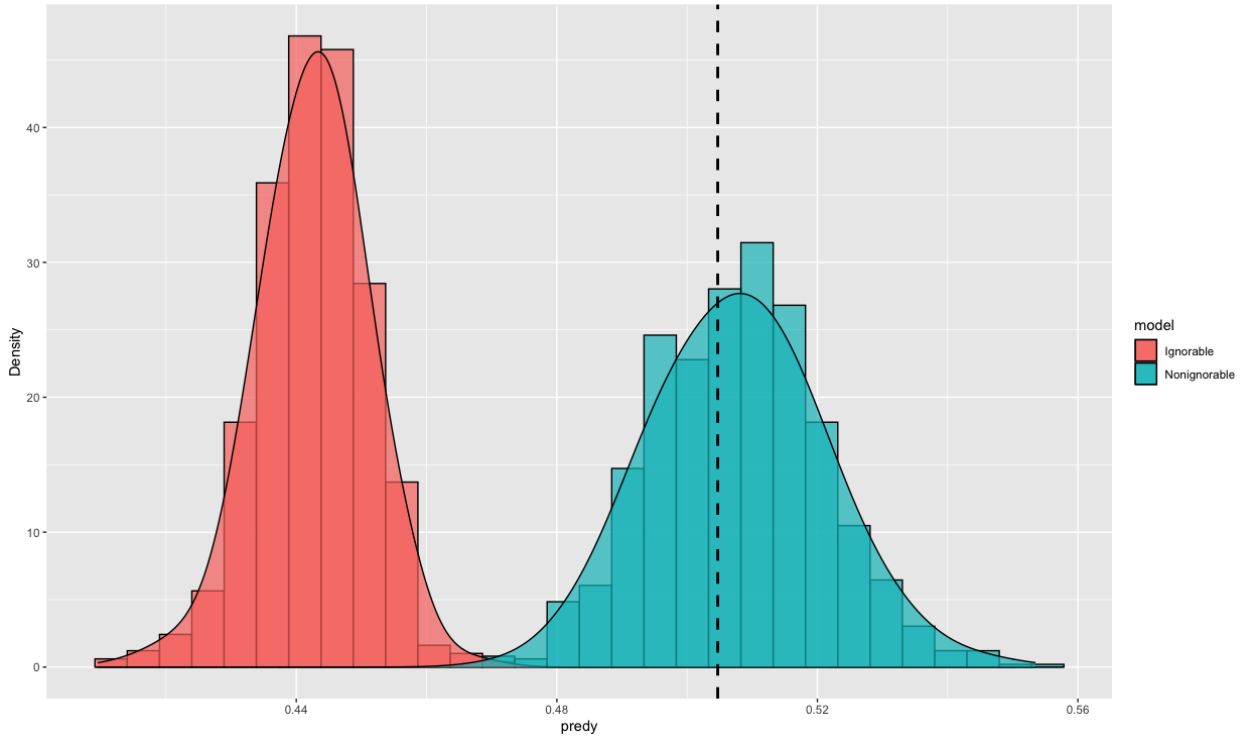
	Nonignorable model (95% HPD)	Ignorable model (95% HPD)	H-T (95% CI)	Hajek(95% CI)
S1 ($\bar{\boldsymbol{y}}_s : 0.69$)	0.51 (0.47, 0.54)	0.61 (0.59, 0.63)	0.64 (0.60, 0.69)	0.64 (0.60, 0.68)
S2 ($\bar{\boldsymbol{y}}_s : 0.67$)	0.52 (0.46, 0.57)	0.50 (0.45, 0.55)	0.44 (0.34, 0.55)	0.48 (0.36, 0.61)
S3 ($\bar{\boldsymbol{y}}_s : 0.33$)	0.51 (0.48, 0.53)	0.44 (0.43, 0.46)	0.40 (0.37, 0.44)	0.40 (0.36, 0.43)


40
30
20
10
0
Density
0.45
0.50
0.55
0.60
0.65
predy
model
Ignorable
Nonignorable

(a) Scenario 1


15
10
5
0
Density
0.40
0.45
0.50
0.55
0.60
predy
model
Ignorable
Nonignorable

(b) Scenario 2



(c) Scenario 3

Figure 2.1: Histogram of population proportion prediction by scenario. (a): nonignorable selection model (left), ignorable selection model (right). (b) and (c): nonignorable selection model (right), ignorable selection model (left). Dashed line: true population proportion

2.2.3 Repeated Simulation Runs on the Nonignorable Selection Model

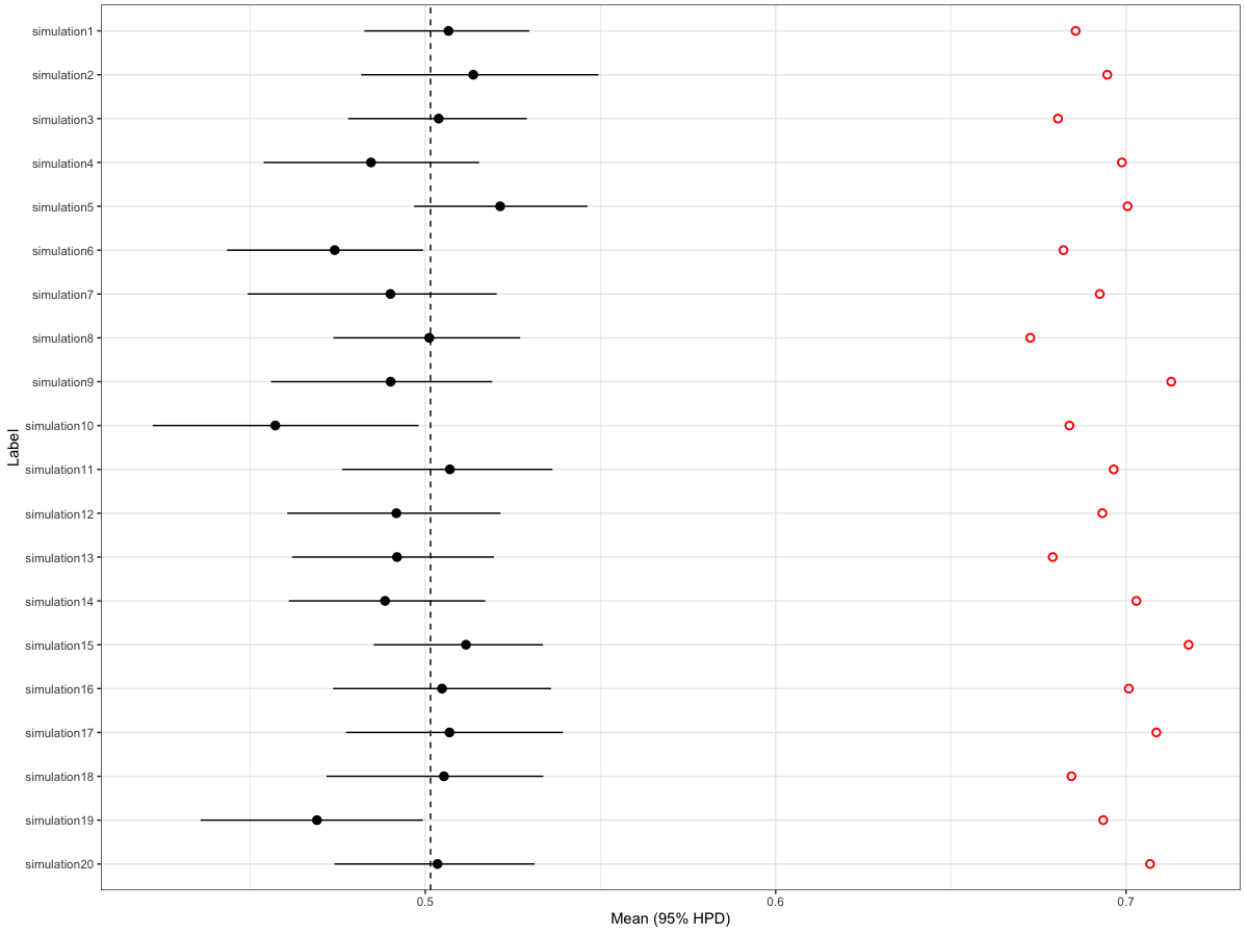
In this section, we performed 20 repeated simulation runs for each scenario and present the result of root mean square error (RMSE) and coverage rate in Table 2.7. Coverage rate is defined by the proportion of times the 95% HPD interval covers the true population proportion. RMSE is defined as the square root of the mean square error, which is calculated by

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}(\bar{Y}_t - \bar{Y}_{true})^2}{T}},$$

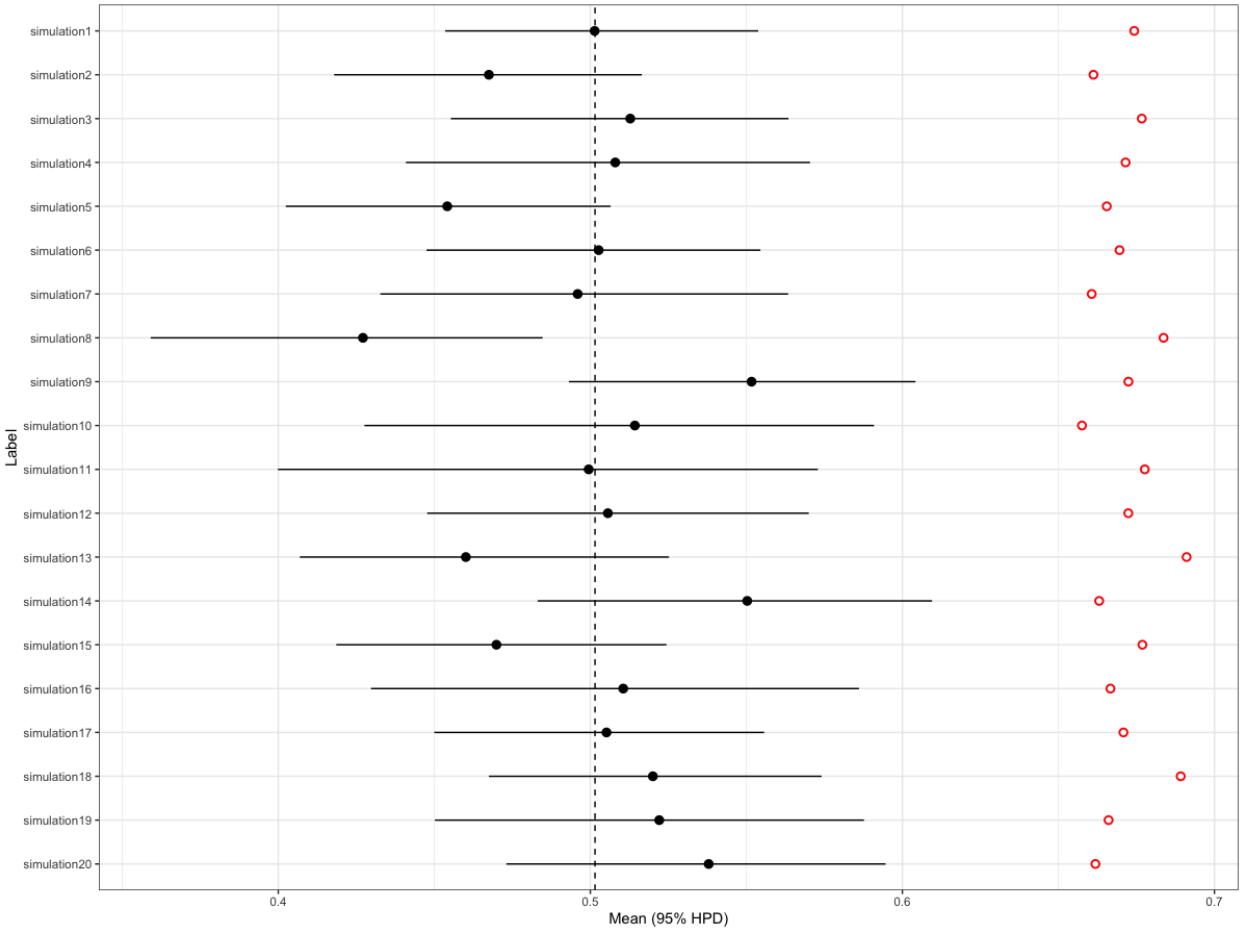
where T is the number of repetitions, $\bar{Y}_t$ is the posterior mean prediction for the repetition t, and $\bar{Y}_{true}$ is the true value of population proportion. Procedures are conducted in the same way across all simulation runs, that is, 30,000 iterations with a burn-in period of 5,000 and thinning by 25. From Table 2.7, we see that the RMSE for scenario 2 is a little bit greater than scenarios 1 and 3, this is consistent with the relatively small p-values noted in Table 2.2, indicating more interations might be needed for scenario 2. We also acknowledge that 20 repetitions might not be enough to observe the coverage probability. However, the procedure is computationally expensive, and we leave the large number of repetitions for future work. A visual of 95% HPD interval for each simulation run is shown in Figure 2.2, from which we see that the coverage counts are 17, 19, and 19 for scenarios 1 to 3 respectively. Overall, the results indicate that the model and procedure is consistent among all simulation runs.

Table 2.7: Results over 20 replicates by scenario

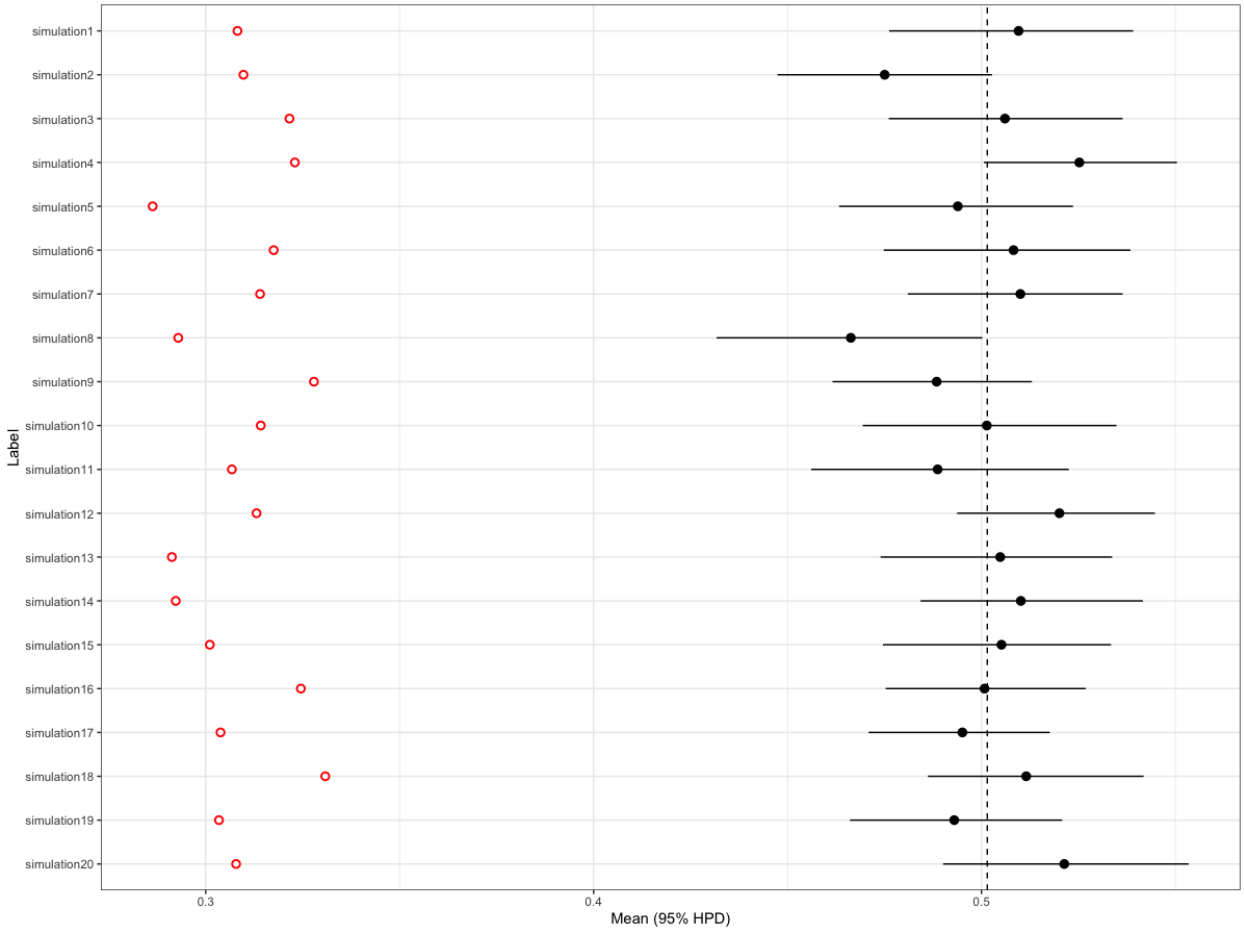
	Coverage Rate	RMSE
Scenario 1	0.85	0.016
Scenario 2	0.95	0.031
Scenario 3	0.95	0.014


simulation1
simulation2
simulation3
simulation4
simulation5
simulation6
simulation7
simulation8
simulation9
simulation10
simulation11
simulation12
simulation13
simulation14
simulation15
simulation16
simulation17
simulation18
simulation19
simulation20
Label
0.5
0.6
0.7
Mean (95% HPD)

(a) Scenario 1


simulation1
simulation2
simulation3
simulation4
simulation5
simulation6
simulation7
simulation8
simulation9
simulation10
simulation11
simulation12
simulation13
simulation14
simulation15
simulation16
simulation17
simulation18
simulation19
simulation20
Label
0.4
0.5
0.6
0.7
Mean (95% HPD)

(b) Scenario 2



(c) Scenario 3

Figure 2.2: 95% HPD interval and posterior mean of population proportion prediction in repeated simulation runs by scenario. (Dashed line: true population proportion. Red circle: sample proportion.)

2.2.4 Model Diagnostic by LPML

Log pseudo marginal likelihood (LPML) has been widely used as a criterion for comparing models, which is defined by conditional predictive ordinate (CPO) as below

$$LPML = \sum_{i=1}^{n} \log(CPO_i), \quad CPO_i = f(y_i^{obs} \mid y_{-i})$$

Large LPML value indicates a well fitted model.

The MCMC numerical approximation of CPO_i is calculated by

$$\widehat{CPO_i} = \left(M^{-1} \sum_{h=1}^{M} \frac{1}{f(y_i \mid \boldsymbol{\theta}_h)} \right)^{-1},$$

where y_i is the sample response, $\boldsymbol{\theta}_h$ are posterior draws of unknown parameters.

From the definition of CPO_i, we see that it measures how likely the sample data, which is subject to selection bias, is in line with the model. However, to use this well established methodology, we propose a revised CPO to evaluate the concordance of observed values with the proposed models. To adjust for the selection effect, we incorporate $I_i = 1$ to the revised CPO_i^*, which is

$$\widehat{CPO_i^*} = \left(M^{-1} \sum_{h=1}^{M} \frac{1}{f(y_i, I_i = 1 \mid \boldsymbol{\theta}_h)} \right)^{-1},$$

This allows the nonignorable and ignorable models a fair comparison. For the nonignorable selection model, $f(y_i, I_i = 1 \mid \boldsymbol{\theta}_h)$ is defined by

$$f(y_i, I_i = 1 \mid \boldsymbol{\theta}_h) = \left(\frac{\exp(\boldsymbol{x}_i' \boldsymbol{\gamma}_h)}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\gamma}_h)} \right)^{y_i} \left(\frac{1}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\gamma}_h)} \right)^{1-y_i} \frac{\exp(\boldsymbol{x}_i' \boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)}{1 + \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)}.$$

For the ignorable selection model, since we do not have $\boldsymbol{\beta}$ in the original model, we have to obtain that first to make LPML values comparable. So, samples of $\boldsymbol{\beta}$ are drawn from

$$f(\boldsymbol{\beta} \mid \boldsymbol{x}, \boldsymbol{I}) \propto \prod_{i=1}^{N} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} I_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}. \tag{2.2.1}$$

Then for the ignorable selection model, $f(y_i, I_i = 1 \mid \boldsymbol{\theta}_h)$ is defined by

$$f(y_i, I_i = 1 \mid \boldsymbol{\theta}_h) = \left(\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_h)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_h)}\right)^{y_i} \left(\frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}_h)}\right)^{1-y_i} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}.$$

Table 2.8 gives the estimated LPML for all three scenarios. Note that the greater value indicates better concordance of the observed values. We see that the nonignorable model has greater values for scenarios 1 and 3, but for scenario 2, both model have about the same values.

This indicates that the nonignorable selection model fits the data better for scenarios 1 and 3. However, we want to emphasize that this evaluation is useful to assess the quality of model predictions for the unobserved units to some degree, but not completely.

Table 2.8: LPML of nonignorable and ignorable models under 3 scenarios

	Nonignorable model	Ignorable model
Scenario 1	-1994.08	-2070.54
Scenario 2	-1573.87	-1572.23
Scenario 3	-2477.87	-2526.27

2.3 Application on Body Mass Index

We apply our model to the BMI data from NHANES III (Nandram and Choi, 2010). In the original datasets, there are sample weights, but we exclude using them since they are not available for non-samples and not compatible with the model we proposed. Our procedure provides an alternative to including the survey weights. The dataset we use contains *age*, *race* and *sex* as covariates, where *age* is collected as integers ranging from 1 to 19, *race* has the values of -1, 0, and 1 representing white non-Hispanic, black non-Hispanic and Hispanic, *sex* is presented by -1 and 1 for male and female. We use the data of the largest county from NHANES III. The sample includes 761 subjects, which is subject to MNAR at the selection and interview stages. The population size is 630,308, however, there are only $19 \times 3 \times 2 = 114$, denoted by C, unique sets of age-race-sex values in the population. This makes it inconvenient and time-consuming to directly apply the algorithm defined in Section 2.1.1. We propose a revised strategy to generate MCMC samples.

We are interested in estimating the proportion of children with healthy BMI, which is defined by greater than or equal to 20 and less than or equal to 25 in our application. Strictly speaking, normal weight status for children and teenagers should be based on BMI between 5^{th} and 85^{th} percentile growth data from the Centers for Disease Control and Prevention (CDC). However, our method is applicable to different criteria and studies.

2.3.1 A More Compact Nonignorable Model

We encounter the following two difficulties when applying the algorithm described in Section 2.1.1.

(1) When there are large numbers of replicates in the population (630,308 observations with only 114 distinct sets of values), the variance-covariance matrix of $\boldsymbol{\gamma}$ becomes very small, i.e., each item in the matrix is on the scale of 10^{-7}, which means the posterior is a very thin-tailed distribution. Such distributions cause the MCMC chain to mix slowly and fail to converge

within 100,000 iterations using our data.

(2) Notice that within each iteration, we need to draw the individual nonsampled units from their corresponding Bernoulli distribution, that is, $N - n = 629,547$ of them, which is computationally expensive.

To bypass the second difficulty, we can think of the $C(C = 114)$ different sets of values of covariates as different categories, denoted by $i = 1, ..., C$. For each category i, the probability of y_{ij} calculated by (iii) in Section 2.1 is the same for all j's, $j = 1, ..., n_i$, denoted by p_i. Thus, we can generate the sum of y_{ij} in category i by

$$\sum_{j=1}^{n_i} y_{ij} \sim \text{Binomial } (n_i, p_i).$$

Inspired by the idea of grouping observations into categories, we rewrite the posterior distribution of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in terms of quantities in category to deal with our first difficulty.

Consider the following 2×2 contingency table for any category i, $i = 1, ..., C$

	$y_i = 1$	$y_i = 0$	
$I_i = 1$	p_{i1}	p_{i2}	
$I_i = 0$	p_{i3}	p_{i4}	

Here we use y_i to denote the response variable in category i, and I_i the indicator variable in the same category. Let p_{i1}, p_{i2}, p_{i3}, and p_{i4} be the cell probabilities, with the constraint $p_{i1} + p_{i2} + p_{i3} + p_{i4} = 1$. Recall y_i and I_i follow the distributions,

$$\begin{aligned} y_i \mid \boldsymbol{\gamma} &\overset{ind}{\sim} \text{Bernoulli} \left\{ \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \right\}, \quad i = 1, ..., C; \\ I_i \mid \boldsymbol{y}, \boldsymbol{\beta} &\overset{ind}{\sim} \text{Bernoulli} \left\{ \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} \right\}, \quad i = 1, ..., C. \end{aligned} \tag{2.3.1}$$

Thus, it is straightforward to show that

$$
\begin{aligned}
p_{i1} &= P(y_i = 1, I_i = 1 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \times \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1})}, \\
p_{i2} &= P(y_i = 0, I_i = 1 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \times \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)})}, \\
p_{i3} &= P(y_i = 1, I_i = 0 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \times \frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1})}, \\
p_{i4} &= P(y_i = 0, I_i = 0 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \times \frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)})}.
\end{aligned}
\tag{2.3.2}
$$

One can verify that $p_{i1} + p_{i2} + p_{i3} + p_{i4} = 1$.

For the counts of each cell, we introduce the following notation,

	$y_i = 1$	$y_i = 0$	
$I_i = 1$	a_i	$n_i - a_i$	n_i
$I_i = 0$	z_i	$N_i - n_i - z_i$	$N_i - n_i$

with probability mass function as follows

$$
P(a_i, n_i - a_i, z_i, N_i - n_i - z_i) = \frac{N_i!}{a_i!(n_i - a_i)!z_i!(N_i - n_i - z_i)!} \times {p_{i1}}^{a_i} {p_{i2}}^{n_i - a_i} {p_{i3}}^{z_i} {p_{i4}}^{N_i - n_i - z_i}.
$$

It is worth noting that z_i is the summation of all the non-sampled y's in category i, and a_i, n_i, N_i are all observed. Using the weakly informative priors for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, the new form of the joint posterior density is

$$
\begin{aligned}
P(\boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{I}, \boldsymbol{y}_s) \propto & \prod_{i=1}^{C} \frac{N_i!}{a_i!(n_i - a_i)!z_i!(N_i - n_i - z_i)!} \times {p_{i1}}^{a_i} {p_{i2}}^{n_i - a_i} {p_{i3}}^{z_i} {p_{i4}}^{N_i - n_i - z_i} \\
& \times \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\beta}})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right) \exp\left(-\frac{1}{2}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\gamma}})^{-1} (\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})\right),
\end{aligned}
\tag{2.3.3}
$$

where p_{i1}, p_{i2}, p_{i3}, and p_{i4} are functions of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ as displayed in (2.3.2). We will explain how to

draw MCMC samples from it in the next part.

MCMC Algorithm

Looking at the joint posterior distribution (2.3.3), we realize it is very difficult to draw $\boldsymbol{z}, \boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ all together. Xu, Nandram and Manandhar (2020) construct a sampler called random sampler to draw samples for weakly identifiable parameters that improves the mixing performance. Here, we use a similar yet simpler idea since the variable z_i can be analytically summed out, thereby making drawing MCMC samples of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ much faster. We omit the priors in this derivation for the ease of reading. Thus,

$$
\begin{aligned}
\pi(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{I}, \boldsymbol{y}_s) &\propto \prod_{i=1}^{C} \sum_{z_i=0}^{N_i - n_i} \left\{ {p_{i1}}^{a_i} {p_{i2}}^{n_i - a_i} \times \frac{{p_{i3}}^{z_i} {p_{i4}}^{N_i - n_i - z_i}}{z_i!(N_i - n_i - z_i)!} \right\} \\
&= \prod_{i=1}^{C} \left\{ {p_{i1}}^{a_i} {p_{i2}}^{n_i - a_i} \times \frac{(p_{i3} + p_{i4})^{N_i - n_i}}{(N_i - n_i)!} \times (N_i - n_i)! \sum_{z_i=0}^{N_i - n_i} \frac{\left(\frac{p_{i3}}{p_{i3}+p_{i4}}\right)^{z_i} \left(\frac{p_{i4}}{p_{i3}+p_{i4}}\right)^{N_i - n_i - z_i}}{z_i!(N_i - n_i - z_i)!} \right\} \\
&= \prod_{i=1}^{C} \left\{ {p_{i1}}^{a_i} {p_{i2}}^{n_i - a_i} \times \frac{(p_{i3} + p_{i4})^{N_i - n_i}}{(N_i - n_i)!} \right\} \\
&\qquad \left(\text{because for a r.v. } x_i,\ x_i \sim \text{Binomial}\left(N_i - n_i, \frac{p_{i3}}{p_{i3} + p_{i4}}\right) \implies \sum_{x_i=0}^{N_i - n_i} x_i = 1, \right) \\
&\propto \prod_{i=1}^{C} \left\{ {p_{i1}}^{a_i} {p_{i2}}^{n_i - a_i} (p_{i3} + p_{i4})^{N_i - n_i} \right\}.
\end{aligned}
$$

Replacing p_{i1}, p_{i2}, p_{i3}, and p_{i4} with their corresponding formula in (2.3.2) and inserting the priors, we arrive at the following joint posterior density of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$,

$$
\begin{aligned}
&\pi(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{I}, \boldsymbol{y}_s) \propto \\
&\prod_{i=1}^{C} \left\{ \left(\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1})} \right)^{a_i} \left(\frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \times \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)})} \right)^{n_i - a_i} \right. \\
&\left. \times \left[\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}) \left(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)})\right) + 1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1})}{(1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})) \left(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1})\right) \left(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)})\right)} \right]^{N_i - n_i} \right\} \\
&\times \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\beta}})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right) \exp\left(-\frac{1}{2} (\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\gamma}})^{-1} (\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}) \right).
\end{aligned}
\tag{2.3.4}
$$

Therefore, the two conditional posterior distributions are,

$$
\begin{aligned}
&\pi(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \boldsymbol{I}, \boldsymbol{y}_s) \\
&\propto \prod_{i=1}^{C} \left\{ \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma} a_i) \left[\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}) \left(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)})\right) + 1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1})\right]^{N_i - n_i}}{\left(1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})\right)^{N_i}} \right\} \\
&\times \exp\left(-\frac{1}{2} (\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\gamma}})^{-1} (\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}) \right),
\end{aligned}
\tag{2.3.5}
$$

and

$$
\begin{aligned}
&\pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \boldsymbol{I}, \boldsymbol{y}_s) \\
&\propto \prod_{i=1}^{C} \left\{ \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} n_i + \beta_{p+1} a_i) \left[\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}) \left(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)})\right) + 1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1})\right]^{N_i - n_i}}{\left(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1})\right)^{N_i - n_i + a_i} \left(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)})\right)^{N_i - a_i}} \right\} \\
&\times \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\beta}})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right).
\end{aligned}
\tag{2.3.6}
$$

Algorithm:

1. For $(t = 1, ..., T)$, repeat:

   (i) Draw $\boldsymbol{\beta}^*$ from multivariate t-distribution, where the proposal mean vector is the mode

of (2.3.6), and the covariance matrix is given by its inverse of the Hessian of the negative log-likelihood function. Keep the degrees of freedom ν fixed throughout the iterations. It is standard to tune the Metropolis step by varying ν. Denote this proposal density by $g(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{I})$.

(ii) Compute the acceptance ratio

$$r = \frac{f(\boldsymbol{\beta}^* \mid \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{I}) g(\boldsymbol{\beta}^{(t-1)} \mid \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{I})}{g(\boldsymbol{\beta}^* \mid \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{I}) f(\boldsymbol{\beta}^{(t-1)} \mid \boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{I})}.$$

(iii) Draw u from uniform $\boldsymbol{U}(0,1)$, if $u \leq \min(1, r)$, set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^*$, else, set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$. Count the number of jumps.

(iv) Similarly, adapt steps (i)-(iii) to draw $\boldsymbol{\gamma}^{(t)}$ based on formula (2.3.5).

2. After finishing the T iterations, we check if the jumping rates for $\boldsymbol{\beta}$ or $\boldsymbol{\gamma}$ are within the range of (25%, 75%). If the jumping rate is not acceptable, i.e. too low or too high, we change the value for ν and repeat the above steps. If the jumping rate is acceptable, we further burn-in and thin the MCMC chain to obtain M sets of samples and check the diagnostics.

3. For each set of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and each $i, i = n+1, ..., C$, calculate the probability of success by formula (iii) in Section 2.1.1, denoted by p_i , then draw the corresponding z_i from its distribution, Binomial$(N_i - n_i, p_i)$. Then, the finite population proportion can be obtained by finding the proportion of combined samples and z_i's. Repeat this step for all sets of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to obtain the posterior predictive distribution of population proportion.

2.3.2 Diagnostics and Results

Nonignorable Selection Model

This new MCMC algorithm is much more efficient than applying the original one directly. The execution time is about 100 minutes for 30,000 iterations using the same computer as previously

specified. After burning-in the first 10,000 and thinning by 20, we obtain 1,000 sets of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and finite population proportions. Table 2.9 depicts the diagnostics and summaries for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. The diagnostics indicate adequate convergence. Since $\gamma_1, \gamma_2, \gamma_3$ correspond to variables Age, Sex and Race, then from the summaries we see that both age and sex are significant in predicting Y; and both race and Y are significant in the selection model.

The posterior mean of population proportion predictions by this nonignorable selection model is 0.63, with 95% HPD interval to be (0.46, 0.78) as shown in Table 2.11.

For completeness, we also want to know the proportion of children with unhealthy BMI. Since for each healthy BMI's proportion, the unhealthy BMI's proportion is just 1 minus that value, thus we can easily obtain the MCMC samples of unhealthy BMI proportion without doing the iteration again. The posterior mean for unhealthy BMI's proportion is 0.37, and the 95% HPD interval is (0.22, 0.54).

Table 2.9: Posterior summaries and diagnostics for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$

$\boldsymbol{\gamma}$	PM (PSD)	95% HPD	p-value	$\boldsymbol{\beta}$	PM (PSD)	95% HPD	p-value
γ_0	-1.86 (0.55)	(-2.89, -0.78)	0.62	β_0	-6.29 (0.12)	(-6.52, -6.04)	0.73
γ_1	0.28 (0.02)	(0.23, 0.32)	0.91	β_1	0.05(0.03)	(-0.01, 0.10)	0.52
γ_2	0.23 (0.10)	(0.04, 0.43)	0.07	β_2	0.03 (0.05)	(-0.06, 0.13)	0.72
γ_3	-0.09 (0.14)	(-0.38, 0.17)	0.13	β_3	1.20 (0.06)	(1.09, 1.34)	0.15
ESS	603			β_4	-2.00 (0.54)	(-3.09, -1.03)	0.56
				ESS	1034		

Note: PM: posterior mean; PSD: posterior standard deviation

We also apply the data to the ignorable model. The Geweke's Diagnostics in Table 2.10 indicate satisfactory mixing and convergence of the chain. The posterior mean prediction by the ignorable selection model is 0.31 as shown in Table 2.11, which is much less than the prediction by the nonignorable selection model, i.e., 0.63. This is because in the nonignorable selection model, the coefficient of y in the selection part is significant, which has the value of -2, while the ignorable selection model fails to incorporate this. The histograms of population proportions predicted by

these two models are shown in Figure 2.3, where we see that aside from the differences in the locations, the nonignorable selection model is more spread out than the ignorable selection model. This is sensible because, from Bayesian point of view, the nonignorable model has more variables and thus more variability; in frequentist view, this possibly be a case of bias-variance tradeoff.

Table 2.10: Posterior summaries and diagnostics of $\boldsymbol{\gamma}$ for the ignorable selection model

$\boldsymbol{\gamma}$	PM (PSD)	95% HPD	p-value
γ_0	-3.69 (0.25)	(-4.21, -4.24)	0.50
γ_1	0.26 (0.02)	(0.22, 0.30)	0.30
γ_2	0.32 (0.10)	(0.14, 0.51)	0.05
γ_3	-0.06 (0.13)	(-0.30, 0.22)	0.94
ESS	1000		

Note: PM: posterior mean; PSD: posterior standard deviation

Table 2.11: Posterior summaries of $\bar{Y}$

$\bar{\boldsymbol{y}}_s$	Nonignorable model PM (95% HPD)	Ignorable model PM (95% HPD)
0.23	0.63 (0.46, 0.78)	0.31 (0.26, 0.36)

Note: PM: posterior mean.

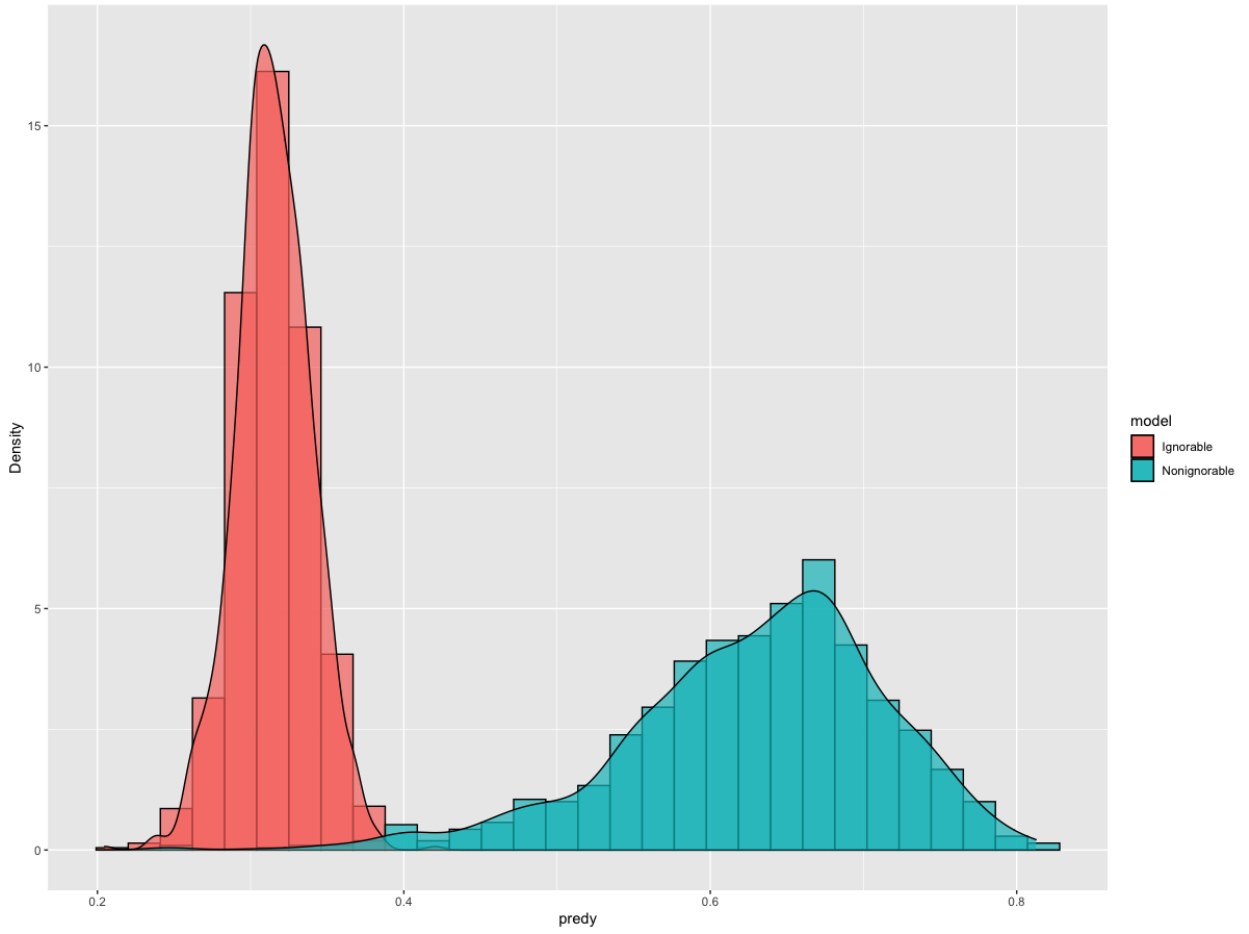


Figure 2.3: Histogram of the finite population proportion for children with healthy BMI by ignorable (left) and nonignorable model (right)

2.4 Study on Priors

In this section, we would like to find out the impact of the choice of the priors for $\boldsymbol{\beta}$, specifically, we want to know (1) the impact of prior on MCMC convergence; and (2) the estimate of the finite population proportion. We keep the prior for $\boldsymbol{\gamma}$ unchanged. To answer the first question, we use effective sample size of the posterior samples as the rule for deciding Markov chain convergence. ESS has a connection with many MCMC diagnostics, for example, Gong and Flegal (2016) and Vats et al. (2019) provide the connection between ESS and relative fixed-width stopping rule; Vats and Knudson (2018) draw correspondence between ESS and a version of the widely used Gelman-Rubin (GR) diagnostic. For the second question, we check the population proportion predictions for each tested prior.

We first propose a method of constructing the prior from the samples, this is based on the idea of creating the nonignorable selection model around the ignorble selection model.

Data Prior

1. Impute each $\boldsymbol{y}_{ns}$ by the nearest neighbor search, that is, impute the non-sampled y_j by y_i, where y_i is the point in sample that minimizes

$$(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + (x_{3i} - x_{3j})^2 + (x_{4i} - x_{4j})^2,$$

   assuming we use our previous simulation setup with 4 covariates.

2. Now we obtain a dataset with N observations. When N is large, the prior that is constructed on them could dominate the likelihood function, thus, we use 5% of the dataset to construct the prior. We randomly draw 5% of N, denoted by N_{sub}, then we find the mode and the

Hessian matrix of the following negative log-likelihood.

$$
\begin{aligned}
& I_i \mid \boldsymbol{y}, \boldsymbol{\beta} \overset{ind}{\sim} \text{Bernoulli} \left\{ \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} \right\}, \quad i = 1, ..., N_{sub}. \\
& \implies L(\boldsymbol{\beta}) = \prod_{i=1}^{N_{sub}} \frac{\exp\left\{(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i) I_i\right\}}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)}. \\
& \implies -logL(\boldsymbol{\beta}) = -\sum_{i=1}^{N_{sub}} \frac{\exp\left\{(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i) I_i\right\}}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)}.
\end{aligned} \tag{2.4.1}
$$

3. Denote the mode by $\boldsymbol{\beta}_0$ and the inverse of Hessian matrix by $\Sigma_{\boldsymbol{\beta}_0}$. The constructed prior is $\text{N}(\boldsymbol{\beta}_0, \Sigma_{\boldsymbol{\beta}_0})$

Simulation Results

Using the same method of conducting the simulation study as Section 2.2.1 scenario 3, we generate a sample of 1187 with the sample proportion of 0.34 and population proportion of 0.51. We varied the proportion of data used in constructing the data prior. The results are obtained from 30,000 iterations with burning in the first 10,000 and thinning by 19, of which presented in Table 2.12. The last row is the prior defined in Section 2.1.1. We see that all ESS for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are greater than 500, which indicates all of them are acceptable. However, we do not see one exceptional results and thus we cannot conclude which prior has the best performance in convergence. The population proportion predictions are also the same in terms of posterior mean and HPD interval.

Table 2.12: Results comparison by different prior for $\boldsymbol{\beta}$

Prior for $\boldsymbol{\beta}$	ESS for $\boldsymbol{\gamma}$	ESS for $\boldsymbol{\beta}$	Posterior Mean (95% HPD)
Data prior: 5%	568	957	0.51 (0.49, 0.53)
Data prior: 2%	692	1102	0.51 (0.48, 0.53)
Flat prior	535	832	0.51 (0.48, 0.53)
$\text{N}(\hat{\beta}, 10^2\hat{\Sigma}_\beta)$	758	762	0.51 (0.48, 0.53)

2.5 Study on Unknown Individual Covariates

In this section, we explore the case where individual covariates $\boldsymbol{X}$ are not available for non-sampled units. Here we perform a data fusion to obtain the nonsampled covariates, and propose a methodology to generate the individual discrete covariates based on information of total counts for each level. For illustration purpose, we simplify the application of BMI in Section 2.3, and use a level of 4 for the age variable, a level of 2 for both race and sex, which gives us a level of 16 in all possible combinations of covariates. The data can be presented by the table below (Table 2.13). Our goal is to use the total counts for each level of a variable (Table 2.14) to generate N_i, $i = 1, ..., 16$.

Table 2.13: Data structure

Age	Race	Sex	N
2 (0-4)	-1	-1	N_1
		1	N_2
	1	-1	N_3
		1	N_4
7 (5-9)	-1	-1	N_5
		1	N_6
	1	-1	N_7
		1	N_8
12 (10-14)	-1	-1	N_9
		1	N_{10}
	1	-1	N_{11}
		1	N_{12}
17 (15-19)	-1	-1	N_{13}
		1	N_{14}
	1	-1	N_{15}
		1	N_{16}

Table 2.14: Counts by levels for each covariate

Covariate	Level	Counts	Notation
Age	2	150108	A_1
	7	170516	A_2
	12	153041	A_3
	17	153588	A_4
Race	-1	459056	R_0
	1	168197	R_1
Sex	-1	305500	S_0
	1	321753	S_1
Total		627253	N

Goal: To draw samples of N_1, N_2, ..., N_{16}

The constraints are:

$$
\begin{aligned}
&N_1 + N_2 + N_3 + N_4 = A_1,\\
&N_5 + N_6 + N_7 + N_8 = A2,\\
&N_9 + N_{10} + N_{11} + N_{12} = A_3,\\
&N_{13} + N_{14} + N_{15} + N_{16} = A_4,\\
&N_1 + N_2 + N_5 + N_6 + N_9 + N_{10} + N_{13} + N_{14} = R_0,\\
&N_3 + N_4 + N_7 + N_8 + N_{11} + N_{12} + N_{15} + N_{16} = R_1,\\
&N_1 + N_3 + N_5 + N_7 + N_9 + N_{11} + N_{13} + N_{15} = S_0,\\
&N_2 + N_4 + N_6 + N_8 + N_{10} + N_{12} + N_{14} + N_{16} = S_1,\\
&N_1, N_2, ..., N_{16} \geq 0.
\end{aligned}
\tag{2.5.1}
$$

We have 16 variables with 10 constraints, while these constraints are not independent to each other, i.e., $\sum_{i=1}^{4} A_i = N$, $R_0 + R_1 = N$, $S_0 + S_1 = N$. From the constraints, one can obtain equations for 6 variables, which depend on 10 free variables. The Python module called Sympy is designed for symbolic calculation and can be used to obtain the following results. The equations

are

$$
\begin{aligned}
N_1 &= -A_2 - A_3 - A_4 + N + N_{10} + N_{12} + N_{14} + N_{16} - N_3 + N_6 + N_8 - S_1, \\
N_2 &= -N_{10} - N_{12} - N_{14} - N_{16} - N_4 - N_6 - N_8 + S_1, \\
N_5 &= A_2 + N_{11} + N_{12} + N_{15} + N_{16} + N_3 + N_4 - N_6 - R_1, \\
N_7 &= -N_{11} - N_{12} - N_{15} - N_{16} - N_3 - N_4 - N_8 + R_1, \\
N_9 &= A_3 - N_{10} - N_{11} - N_{12}, \\
N_{13} &= A_4 - N_{14} - N_{15} - N_{16}.
\end{aligned}
\tag{2.5.2}
$$

The free variables are $N_3, N_4, N_6, N_8, N_{10}, N_{11}, N_{12}, N_{14}, N_{15}, N_{16}$. Next, one can use the equations 2.5.2 to derive the bounds for free variables. To save space, we leave the mathematic formulations of the bounds in Appendix C, from where we arrive at the following algorithm for sampling $N_1, ..., N_{16}$.

Algorithm:

1. Give initial values to $N_1, ..., N_{16}$ by drawing from Multinomial $(N, \boldsymbol{p})$, where $\boldsymbol{p} = \boldsymbol{n}/n$, $\boldsymbol{n}$ is the vector of counts in each category in the sample and n is the sample size.

2. For $(t = 1, ..., T)$, repeat:

   (i) Draw $N_3^{(t)}$ from the truncated multinomial distribution Multinomial $(N, \boldsymbol{p})$ with bounds calculated in Appendix C and the current values for all other N_i 's, $i \neq 3$. Technically, this is implemented by the grid method, that is, calculating the probability of a sequence of values between bounds and drawing a value from the sequence with associated probabilities.

   (ii) Repeate (i) for $N_4^{(t)}, N_6^{(t)}, N_8^{(t)}, N_{10}^{(t)}, N_{11}^{(t)}, N_{12}^{(t)}, N_{14}^{(t)}, N_{15}^{(t)}, N_{16}^{(t)}$ with adapted ranges.

   (iii) Calculate values for $N_1^{(t)}, N_2^{(t)}, N_5^{(t)}, N_7^{(t)}, N_9^{(t)}, N_{13}^{(t)}$ based on formula (2.5.2).

   (iv) Back to (i) for the next loop.

2.5.1 Simulation Study

We use population covariates $\boldsymbol{X}$ from BMI data and coefficients $\boldsymbol{\gamma} = (-1.85, 0.27, -0.08, 0.23)$ to generate response $\boldsymbol{Y}$. Then, from the population and using coefficients $\boldsymbol{\beta} = (-5, 0.05, 0.2, 0.03, -1.5)$, we generate the sample and obtain the non-sample part. Values for $A_0, A_1, A_2, A_3, R_0, R_1, S_0, S_1$ could be summarized from the non-sample part and presented in the Table 2.13.

By iterating 5000 times and using the last one, we get the number for N_i presented in the table below, along with the true counts from simulation.

	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8
Generated	52731	50523	23715	23139	50985	56944	31585	31002
True	52340	55025	20962	21781	60742	63508	22571	23695
	N_9	N_{10}	N_{11}	N_{12}	N_{13}	N_{14}	N_{15}	N_{16}
Generated	64426	41637	22746	24232	68849	41430	22075	21234
True	55003	57703	19600	20735	55575	59160	18707	20146

We further use those generated N_i in the model and the predicted population proportions are summarized in the Table 2.15 with the visual in Figure 2.5. We see the 95% credible interval is a bit wider than the one in Section 2.3.2. This is probably because when the Age variable is grouped into 4 levels and Race grouped into 2 levels, some information is lost and thus makes the prediction less accurate.

Table 2.15: Posterior summaries of $\bar{Y}$

True $\bar{Y}$	$\bar{\boldsymbol{y}}_s$	Posterior Mean (PSD)	95% HPD
0.62	0.34	0.59 (0.11)	(0.37, 0.77)

Note: PSD: posterior standard deviation.

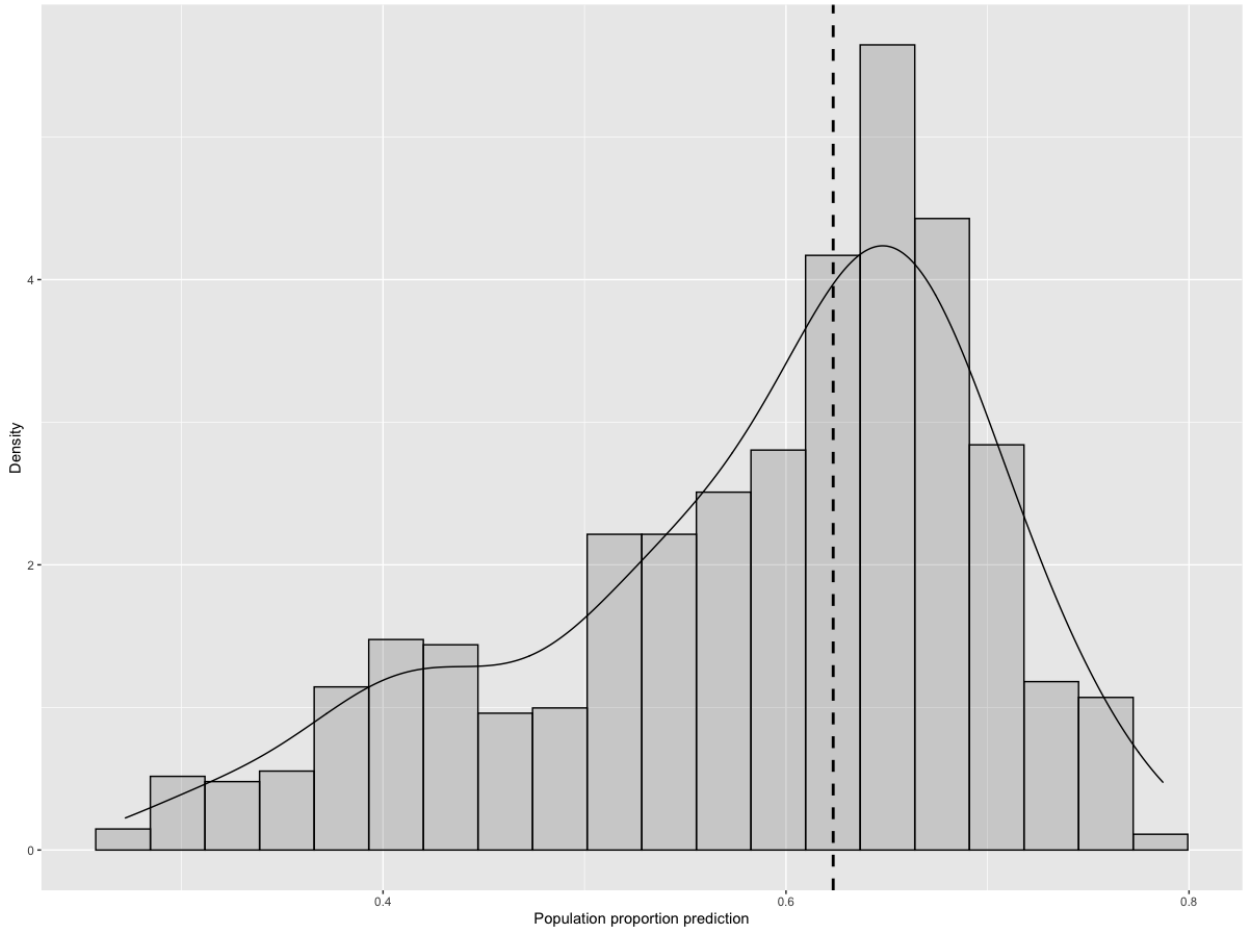


Figure 2.4: Histogram of Population Proportion Prediction

2.6 Conclusion

This chapter proposed a novel modeling framework that enables us to predict the non-sampled subjects' responses from biased samples when the missingness mechanism is MNAR. The underlying idea is to link the model of response with the model of selection by the unknown response variable, which can be obtained within MCMC iterations. We tested our model on the simulated datasets under three scenarios. A comparison of the predictions under various models with the true value indicates the superiority of the nonignorable selection model over the ignorable selection model, H-T and Hajek estimators under MNAR, and also indicates that the ignorable model performs best when missingness mechanism is MAR.

The proposed nonignorable selection model requires that the individual covariates are known for all units. When the individual covariates are unknown for non-sampled units, we provide a methodology to generate the individual discrete covariates.

It is usually difficult to learn about the missing data mechanism. However, incorporating the response variable into the selection model and testing its statistical significance gives us a sense of the selection mechanism. Moreover, by fitting the models to three scenarios, it is evident that our nonignorable model is robust and can automatically adjust to different missing data mechanisms (MAR or MNAR). We believe that the methodology we proposed, and the compact model applied on BMI data will contribute to ongoing research in this field.

Chapter 3

Bayesian Selection Model for Heterogeneous Data

Area information is widely used in practice, not only for the estimates of the subpopulation (area) but also for the total population of interest. A hierarchical model is often used to borrow strength from other areas. An attractive property of the hierarchical Bayesian model is that it takes care of extraneous variations among areas. Another benefit is that there is robustness in the model specifications at deeper levels beyond the sampling process. This has been found empirically (Nandram and Choi, 2005, 2010). At the first level of the hierarchy, these models assume that the individuals are exchangeable within an area but not across areas. At the second level of the hierarchy, the areas are assumed to be exchangeable, which acknowledges variation between areas. The nonignorable selection model we propose links two hierarchical submodels by variable Y.

In this chapter, we first introduce a model that includes clustering effects to account for heterogeneity among areas through the selection indicators and the response values, followed by a simulation study. Then, we propose some adjustments for the model, including the Dirichlet Process mixture model and the bivariate model.

3.1 Hierarchical Model Specification

In this section, we first introduce the hierarchical nonignorable selection model, which is an extension of the nonignorable selection model proposed in Chapter 2. Then, as a comparison, we propose the hierarchical ignorable selection model.

3.1.1 Hierarchical Nonignorable Selection Model

The hierarchical nonignorable selection model is built by two hierarchical submodels. The first submodel is for the response. Denote the area-specific random effect for the response by $\boldsymbol{\nu}_1$. The model is defined by a hierarchical Bayesian model $f(\boldsymbol{y} \mid \boldsymbol{\nu}_1, \boldsymbol{\gamma})$ and $f(\boldsymbol{\nu}_1 \mid \sigma_1^2)$.

The second submodel is to model the selection by $f(\boldsymbol{I} \mid \boldsymbol{y}, \boldsymbol{\nu}_2, \boldsymbol{\beta})$ and $f(\boldsymbol{\nu}_2 \mid \sigma_2^2)$, where $\boldsymbol{\nu}_2$ is the area-specific random effect for the selection mechanism. In this model, we assume $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ are independent. Each element in $\boldsymbol{I}$, denoted by I_{ij} for $i = 1, ..., \ell$ and $j = 1, ..., N_i$, is the selection indicator for the jth individual within the ith area in the population.

Now we present the formulations of these two submodels of the hierarchical nonignorable selection model.

The response submodel

Suppose ν_{1i}'s are the random effects for area i, $i = 1, ..., \ell$, which are independent and identically distributed (iid). Its variance σ_1^2 is a measure of homogeneity of the areas, which is unknown but given a non-informative prior. The hierarchical Bayesian model for the response is

$$
\begin{aligned}
& y_{ij} \mid \nu_{1i}, \boldsymbol{\gamma} \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{\exp(\boldsymbol{x}_{ij}'\boldsymbol{\gamma} + \nu_{1i})}{1 + \exp(\boldsymbol{x}_{ij}'\boldsymbol{\gamma} + \nu_{1i})}\right\}, \quad i = 1, ..., \ell, \quad j = 1, ..., N_i, \\
& \nu_{1i} \mid \sigma_1^2 \overset{iid}{\sim} \text{N}(0, \sigma_1^2), \quad i = 1, ..., \ell, \\
& \pi(\boldsymbol{\gamma}) \sim \text{N}(\hat{\boldsymbol{\gamma}}, 10^2\hat{\Sigma}_{\boldsymbol{\gamma}}), \\
& \pi(\sigma_1^2) = \frac{1}{(1 + \sigma_1^2)^2}.
\end{aligned}
\tag{3.1.1}
$$

Here the prior for σ_1^2 is the $F(2,2)$ distribution and holds good properties of a shrinkage prior. It is also mathematical convenient in such models because the form of this prior can be canceled with the Jacobian when transformed into a bounded variable, see (3.1.4).

The selection submodel

Similar to the nonignorable selection model in the previous chapter, we use $\boldsymbol{\beta}$ to represent the combined parameter vector for $\boldsymbol{X}$ and $\boldsymbol{Y}$, which is also given a vague prior - multivariate normal $N(\hat{\boldsymbol{\beta}}, 10^2\hat{\Sigma})$. The hierarchical Bayesian model for the selection mechanism is

$$
\begin{aligned}
&I_{ij} \mid y_{ij}, \nu_{2i}, \boldsymbol{\beta} \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_{ij} + \nu_{2i})}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_{ij} + \nu_{2i})}\right\}, \quad i = 1, ..., \ell, \quad j = 1, ..., N_i,\\
&\nu_{2i} \mid \sigma_2^2 \overset{ind}{\sim} \text{N}(0, \sigma_2^2), \quad i = 1, ..., \ell,\\
&\pi(\boldsymbol{\beta}) \sim \text{N}(\hat{\boldsymbol{\beta}}, 10^2\hat{\Sigma}_{\boldsymbol{\beta}})\\
&\pi(\sigma_2^2) = \frac{1}{(1+\sigma_2^2)^2}.
\end{aligned}
\tag{3.1.2}
$$

The joint posterior distribution

Putting 3.1.1 and 3.1.2 together, we derived the joint posterior distribution as

$$
\begin{aligned}
&f(\boldsymbol{y}_{ns}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \sigma_1^2, \sigma_2^2 \mid \boldsymbol{I}, \boldsymbol{y}_s)\\
&\propto \prod_{i=1}^{\ell}\prod_{j=1}^{N_i} \frac{\exp\left[(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})y_{ij}\right]}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})} \frac{\exp\left[(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_{ij} + \nu_{2i})I_{ij}\right]}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_{ij} + \nu_{2i})}\\
&\times \exp\left(-\frac{1}{2}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})^T(10^2\hat{\Sigma}_{\boldsymbol{\gamma}})^{-1}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})\right) \times \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(10^2\hat{\Sigma}_{\boldsymbol{\beta}})^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right)\\
&\times \prod_{i=1}^{\ell}(\sigma_1^2)^{-\frac{1}{2}} \exp\left(-\frac{\nu_{1i}^2}{2\sigma_1^2}\right) \times \prod_{i=1}^{\ell}(\sigma_2^2)^{-\frac{1}{2}} \exp\left(-\frac{\nu_{2i}^2}{2\sigma_2^2}\right) \times \frac{1}{(1+\sigma_1^2)^2} \times \frac{1}{(1+\sigma_2^2)^2}
\end{aligned}
\tag{3.1.3}
$$

One can develop a proof for (3.1.3) by extending the proof in Appendix B.

Bayesian computation

We first write down the conditional posterior distribution for each parameter given all other parameters. The conditional posterior distributions are

(i) $f(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{\nu}_1) \propto \prod_{i=1}^{\ell} \prod_{j=1}^{N_i} \frac{\exp[(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})y_{ij}]}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})} \times \exp\left(-\frac{1}{2}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})^T(10^2\hat{\Sigma}_{\boldsymbol{\gamma}})^{-1}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})\right);$

(ii) $f(\boldsymbol{\beta}, \mid \boldsymbol{y}, \boldsymbol{\nu}_2, \boldsymbol{I}) \propto \prod_{i=1}^{\ell} \prod_{j=1}^{N_i} \frac{\exp\left[(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_{ij} + \nu_{2i})I_{ij}\right]}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_{ij} + \nu_{2i})}$
$\times \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(10^2\hat{\Sigma})^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right);$

(iii) $f(\boldsymbol{\nu}_1 \mid \boldsymbol{y}, \boldsymbol{\gamma}, \sigma_1^2) \propto \prod_{i=1}^{\ell} \left\{ \prod_{j=1}^{N_i} \frac{\exp(\nu_{1i}y_{ij})}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})} \times \exp\left(-\frac{\nu_{1i}^2}{2\sigma_1^2}\right) \right\};$

(iv) $f(\boldsymbol{\nu}_2, \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{I}, \sigma_2^2) \propto \prod_{i=1}^{\ell} \left\{ \prod_{j=1}^{N_i} \frac{\exp(\nu_{2i}I_{ij})}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_{ij} + \nu_{2i})} \times \exp\left(-\frac{\nu_{2i}^2}{2\sigma_2^2}\right) \right\};$

(v) $f(\sigma_1^2 \mid \boldsymbol{\nu}_1) \propto \prod_{i=1}^{\ell} \left(\sigma_1^2\right)^{-\frac{1}{2}} \exp\left(-\frac{\nu_{1i}^2}{2\sigma_1^2}\right) \times \frac{1}{(1+\sigma_1^2)^2};$

(vi) $f(\sigma_2^2 \mid \boldsymbol{\nu}_2) \propto \prod_{i=1}^{\ell} \left(\sigma_2^2\right)^{-\frac{1}{2}} \exp\left(-\frac{\nu_{2i}^2}{2\sigma_2^2}\right) \times \frac{1}{(1+\sigma_2^2)^2};$

(vii) $P(y_{ij} = 1 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2)$

$= \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})\left(1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \nu_{2i})\right)}{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})\left(1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \nu_{2i})\right) + 1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1} + \nu_{2i})},$

$P(y_{ij} = 0 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2) = 1 - P(y_{ij} = 1 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2), \quad i = 1, ..., \ell; j \in \overline{S_i};$

where $\overline{S_i}$ denotes the non-sampled set of area i.

We use the Metropolis-Hastings sampling algorithm that runs in 7 blocks with one for each parameter.

For parameters $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and $\boldsymbol{y}_{ns}$, sampling methods are similar to the algorithm for the nonignorable selection model in Chapter 2, with formulations adjusted accordingly. For $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$, we could apply the same idea and use the multivariate t distribution as the proposal distribution for each of them. However, note that each component in $\boldsymbol{\nu}_1$ or $\boldsymbol{\nu}_2$, denoted by ν_{1i} and ν_{2i}, is independent of the others. Thus, we use the one-dimensional t distribution as the proposal distribution and draw the sample of each component individually. Using this method, we sacrifice some algorithm efficiency, but improve the mixing performance. Note that ν_{1i} and ν_{2i} can also be drawn by the grid method, which we use in the Dirichlet process prior model in Section 3.3.

As for σ_1^2 and σ_2^2, we use the grid method and first transform them onto the bounded range $(0, 1)$. We transform σ_1^2 by $\dfrac{1}{1+\sigma_1^2} = \tau_1$, $\tau_1 \in (0, 1)$, which results in the transformed density

$$f(\tau_1 \mid \boldsymbol{\nu}_1) \propto \left(\frac{\tau_1}{1-\tau_1}\right)^{\frac{\ell}{2}} \exp\left(-\frac{\sum_{i=1}^{\ell} \nu_{1i}^2}{2} \times \frac{\tau_1}{1-\tau_1}\right). \tag{3.1.4}$$

Then, a sample of σ_1^2 can be obtained by transforming back a sample of τ_1 based on the formula $\sigma_1^2 = \dfrac{1-\tau_1}{\tau_1}$.

Similarly, transform σ_2^2 by $\dfrac{1}{1+\sigma_2^2} = \tau_2$, $\tau_2 \in (0, 1)$, we have the transformed density

$$f(\tau_2 \mid \boldsymbol{\nu}_2) \propto \left(\frac{\tau_2}{1-\tau_2}\right)^{\frac{\ell}{2}} \exp\left(-\frac{\sum_{i=1}^{\ell} \nu_{2i}^2}{2} \times \frac{\tau_2}{1-\tau_2}\right); \tag{3.1.5}$$

and then using $\sigma_2^2 = \dfrac{1-\tau_2}{\tau_2}$ to get a sample of σ_2^2.

Algorithm:

For ($t = 1, ..., T$), repeat:

1. Draw $\boldsymbol{\gamma}^*$ from multivariate t-distribution, where the proposal mean vector is the mode of (i), and covariance matrix is given by its inverse of the Hessian of the negative log-likelihood function. Keep the degrees of freedom fixed through out the iterations. Denote this proposal density by $g(\boldsymbol{\gamma} \mid \boldsymbol{y}_s, \boldsymbol{\nu}_1)$.

2. Compute the acceptance ratio

$$r = \frac{f(\boldsymbol{\gamma}^* \mid \boldsymbol{y}_s, \boldsymbol{\nu}_1) g(\boldsymbol{\gamma}^{(t-1)} \mid \boldsymbol{y}_s, \boldsymbol{\nu}_1)}{g(\boldsymbol{\gamma}^* \mid \boldsymbol{y}_s, \boldsymbol{\nu}_1) f(\boldsymbol{\gamma}^{(t-1)} \mid \boldsymbol{y}_s, \boldsymbol{\nu}_1)}.$$

3. Draw u from uniform $\boldsymbol{U}(0, 1)$, if $u \leq \min(1, r)$, set $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^*,$ else, set $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^{(t-1)}$. Count the number of jumps.

4. Similarly, adapt steps 1-3 to draw $\boldsymbol{\beta}^{(t)}$ based on formula (ii).

5. For each $i = 1, ..., \ell$, adapt steps 1-3 to draw $\nu_{1i}^{(t)}$ and $\nu_{2i}^{(t)}$.

6. Use the grid method to sample τ_1 from (3.1.4) and transform it back to $\sigma_1^{2(t)}$ by $\sigma_1^2 = 1/\tau_1 - 1$.

7. Similarly, draw a sample from (3.1.5) and obtain a sample of $\sigma_2^{2(t)}$.

8. For each i, and each nonsampled index j, calculate the probability of success by formula (vii), and draw the corresponding y from its Bernoulli distribution. Then, one population proportion prediction, denoted by $\bar{Y}$, can be obtained by finding the proportion of the pooled samples and the predicted nonsamples. In addition, the finite population proportion for each area, denoted by $\bar{Y}_i, i = 1, ..., \ell$, could also be obtained.

After finishing T iterations, we check if the jumping rates for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, ν_{1i}, and ν_{2i}, $i = 1, ..., \ell$ are within the range of (25%, 75%). If the jumping rate is not acceptable, i.e. too low or too high, we

tune the parameter by varying the degree of freedom and repeat the above steps. If all jumpting rates are acceptable, we continue burning-in and thinning the MCMC chain to obtain M sets of samples and check the diagnostics.

3.1.2 Hierarchical Ignorable Selection Model

When the selection mechanism is ignorable, we can model the samples by a single response model, which is

$$
\begin{aligned}
&y_{ij} \mid \nu_{1i}, \boldsymbol{\gamma} \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{\exp(\boldsymbol{x}_{ij}'\boldsymbol{\gamma}+\nu_{1i})}{1+\exp(\boldsymbol{x}_{ij}'\boldsymbol{\gamma}+\nu_{1i})}\right\}, \quad i=1,...,\ell, \quad j=1,...,n_i,\\
&\nu_{1i} \mid \sigma_1^2 \overset{iid}{\sim} \text{N}(0,\sigma_1^2), \quad i=1,...,\ell,\\
&\pi(\boldsymbol{\gamma}) \sim \text{N}(\hat{\boldsymbol{\gamma}}, 10^2\hat{\Sigma}_{\boldsymbol{\gamma}}),\\
&\pi(\sigma_1^2) = \frac{1}{(1+\sigma_1^2)^2}.
\end{aligned}
$$

The joint posterior distribution is

$$
\begin{aligned}
f(\boldsymbol{\gamma}, \boldsymbol{\nu}_1, \sigma_1^2, \mid \boldsymbol{x}_s, \boldsymbol{y}_s) &\propto \prod_{i=1}^{\ell}\prod_{j=1}^{n_i}\frac{\exp\left[(\boldsymbol{x}_{ij}'\boldsymbol{\gamma}+\nu_{1i})y_{ij}\right]}{1+\exp(\boldsymbol{x}_{ij}'\boldsymbol{\gamma}+\nu_{1i})}\\
\times \exp\left(-\frac{1}{2}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})^T(10^2\hat{\Sigma}_{\boldsymbol{\gamma}})^{-1}(\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})\right) &\times \prod_{i=1}^{\ell}(\sigma_1^2)^{-\frac{1}{2}}\exp\left(-\frac{\nu_{1i}^2}{2\sigma_1^2}\right)\times\frac{1}{(1+\sigma_1^2)^2}.
\end{aligned}
$$

The algorithm can be adapted from the ignorable selection model in Section 2.1.2, which is, first run Metropolis-Hastings sampler to obtain, say 1000, sets of posterior samples of $\boldsymbol{\gamma}, \boldsymbol{\nu}_1$ and σ_1^2, then plug-in each set of values to the response model and sample $\boldsymbol{y}_{ns}$. Inferences of the finite population proportion can be derived further by $\boldsymbol{y}_s$ and $\boldsymbol{y}_{ns}$.

3.1.3 Simulation Study

We define the population size to be 12,000, that is N=12,000 with $\ell = 30$ areas, which gives the population size in each area to be $N_i = 400$. We first generate the covariates for the population, that is, *Age*, *Race*, *Gender*, *Education* from distributions listed in Table 3.1. *Age* is from the normal distribution with mean 50 and standard deviation 5; *Race*, *Gender*, and *Education* are from Bernoulli distribution with probabilities of 0.45, 0.3, and 0.4 respectively. Then, we generate area effects from $N(0, 1)$, denoted by ν_{1i}, $i = 1, ..., \ell$.

Table 3.1: Distributions for covariates

Covariate	Distribution
Age	$N(50, 5^2)$
Race	Bernoulli(0.45)
Gender	Bernoulli(0.3)
Education	Bernoulli(0.4)

Once the above covariate and area effect values are simulated, we generate the response $\boldsymbol{Y}$ by

$$y_{ij} \mid \nu_{1i}, \boldsymbol{\gamma} \overset{ind}{\sim} \text{Bernoulli} \left\{ \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})} \right\}, \quad i = 1, ..., l, \quad j = 1, ..., N_i.$$

We use $\boldsymbol{\gamma} = (3, 6, -2, -3)$.

Next, we simulate the selection indicator variable. The selection area effects, denoted by ν_{2i}, are generated from $N(0, 1)$. Then, the selection indicator is simulated by

$$I_{ij} \mid y_{ij}, \nu_{2i}, \boldsymbol{\beta} \overset{ind}{\sim} \text{Bernoulli} \left\{ \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_{ij} + \nu_{2i})}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_{ij} + \nu_{2i})} \right\}, \quad i = 1, ..., l, \quad j = 1, ..., N_i,$$

where $\boldsymbol{\beta} = (0.8, 0.8, -3, -2, -2.5)$.

Using the above procedure, we obtain the population with the population proportion of 0.56 The sample proportion is 0.46. The sample size by areas are 13, 11, 11, 63, 4, 41, 25, 104 , 52, 45,

15, 49, 16, 8, 16, 6, 47, 24, 68, 22, 31, 26, 10 , 16, 40, 10, 2, 36, 11, 36. We see for certain areas, the sample sizes are relatively small, i.e., 2, 4, 6, and 8.

Results

By running 30,000 iterations, burning-in the first 5,000, and thinning by 25, we obtain the MCMC samples of size 1,000. Posterior summaries and Geweke's Diagnostics are depicted in Table 3.2. To save space we do not list all 30 dimensions of $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$. The results indicate adequate convergence.

Table 3.2: Posterior summaries and diagnostics for $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, σ_1^2,σ_2^2

$\boldsymbol{\gamma}$	PM (PSD)	95% HPD	p-value	$\boldsymbol{\beta}$	PM (PSD)	95% HPD	p-value
γ_1	3.15(0.17)	(2.80, 3.48)	0.12	β_1	0.69 (0.04)	(0.60, 0.78)	0.78
γ_2	6.01 (0.30)	(5.48, 6.61)	0.49	β_2	0.67 (0.09)	(0.51, 0.88)	0.57
γ_3	-2.31 (0.35)	(-2.98, -1.67)	0.35	β_3	-3.00 (0.11)	(-3.24, -2.79)	0.30
γ_4	-3.06 (0.26)	(-3.56, -2.57)	0.24	β_4	-1.90 (0.07)	(-2.55, -2.03)	0.49
ESS	553			β_5	-2.28 (0.14)	(1.21, 1.88)	0.32
				ESS	735		
$\boldsymbol{\nu}_1$	PM (PSD)	95% HPD	p-value	$\boldsymbol{\nu}_2$	PM (PSD)	95% HPD	p-value
ν_{11}	-0.69 (0.63)	(-1.98, 0.45)	0.93	ν_{21}	-1.29 (0.30)	(-1.84, -0.67)	0.19
ν_{12}	0.43 (0.60)	(-0.73, 1.60)	0.63	ν_{22}	-0.98 (0.33)	(-1.56, -0.31)	0.64
ν_{13}	0.74 (0.65)	(-0.45, 2.09)	0.22	ν_{23}	-0.99 (0.32)	(-1.69, -0.42)	0.94
ν_{14}	-0.39 (0.32)	(-1.01, 0.22)	0.78	ν_{24}	1.23 (0.21)	(0.80, 1.59)	0.89
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		
σ_1^2	PM (PSD)	95% HPD	p-value	σ_2^2	PM (PSD)	95% HPD	p-value
	1.19 (0.41)	(0.46, 2.00)	0.74		1.27 (0.34)	(0.69, 1.95)	0.43
ESS	1000			ESS	1000		

Note: PM: posterior mean; PSD: posterior standard deviation.

Table 3.3 depicts the posterior proportion prediction for the overall population, as well as the prediction for each area under both the nonignorable selection model and the ignorable selection model. Under the nonignorable selection model, the overall posterior proportion is 0.55 and its 95% HPD interval is (0.52, 0.57). We see that the 95% HPD interval contains the true population

proportion, which is 0.56. For each area, a visual of posterior mean proportion with the 95% HPD interval, along with the true population proportion and sample proportion are presented in Figure 3.1. We see that the 95% HPD interval covers the true population proportion for each area with one exception (area 26). Under the ignorable model, the overall posterior proportion is 0.41 and the 95% HPD interval is (0.39, 0.42), therefore the interval does not contain the true population proportion. One can also check the predictive proportions and intervals on area level in Table 3.3.

Table 3.3: Posterior summaries of finite population proportions

			Nonignorable Model		Ignorable Model	
Area	True $\bar{Y}$	$\bar{\boldsymbol{y}}_s$	PM (PSD)	95% HPD	PM (PSD)	95% HPD
1	0.5	0.51	0.53 (0.029)	(0.47 , 0.58)	0.37 (0.028)	(0.32 , 0.43)
2	0.44	0.34	0.41 (0.029)	(0.36 , 0.47)	0.28 (0.025)	(0.24 , 0.33)
3	0.6	0.48	0.56 (0.025)	(0.52 , 0.61)	0.44 (0.023)	(0.39 , 0.48)
4	0.59	0.43	0.59 (0.05)	(0.5 , 0.69)	0.41 (0.057)	(0.29 , 0.52)
5	0.41	0.24	0.4 (0.038)	(0.32 , 0.46)	0.26 (0.032)	(0.19 , 0.32)
6	0.65	0.59	0.63 (0.022)	(0.59 , 0.67)	0.5 (0.021)	(0.46 , 0.54)
7	0.49	0.28	0.49 (0.043)	(0.42 , 0.58)	0.32 (0.034)	(0.26 , 0.39)
8	0.62	0.69	0.66 (0.037)	(0.58 , 0.73)	0.53 (0.03)	(0.47 , 0.59)
9	0.58	0.48	0.55 (0.019)	(0.52 , 0.59)	0.44 (0.019)	(0.4 , 0.47)
10	0.52	0.22	0.49 (0.041)	(0.42 , 0.58)	0.35 (0.036)	(0.29 , 0.42)
11	0.48	0.36	0.45 (0.018)	(0.42 , 0.48)	0.34 (0.019)	(0.3 , 0.38)
12	0.51	0.37	0.48 (0.018)	(0.44 , 0.51)	0.39 (0.017)	(0.36 , 0.42)
13	0.72	0.71	0.69 (0.03)	(0.63 , 0.74)	0.55 (0.034)	(0.48 , 0.61)
14	0.5	0.46	0.48 (0.023)	(0.43 , 0.52)	0.37 (0.02)	(0.33 , 0.41)
15	0.58	0.49	0.57 (0.023)	(0.53 , 0.62)	0.44 (0.023)	(0.4 , 0.48)
16	0.52	0.43	0.5 (0.017)	(0.47 , 0.54)	0.41 (0.017)	(0.37 , 0.44)
17	0.52	0.41	0.5 (0.014)	(0.48 , 0.53)	0.42 (0.015)	(0.39 , 0.44)
18	0.64	0.54	0.63 (0.022)	(0.58 , 0.67)	0.52 (0.024)	(0.46 , 0.56)
19	0.58	0.44	0.56 (0.023)	(0.52 , 0.6)	0.44 (0.022)	(0.4 , 0.48)
20	0.52	0.41	0.56 (0.036)	(0.49 , 0.63)	0.39 (0.037)	(0.32 , 0.46)
21	0.57	0.61	0.58 (0.016)	(0.55 , 0.62)	0.49 (0.015)	(0.46 , 0.51)
22	0.54	0.47	0.58 (0.052)	(0.49 , 0.69)	0.39 (0.052)	(0.29 , 0.49)
23	0.64	0.71	0.71 (0.052)	(0.61 , 0.8)	0.56 (0.062)	(0.44 , 0.68)
24	0.57	0.59	0.55 (0.039)	(0.48 , 0.62)	0.41 (0.037)	(0.34 , 0.48)
25	0.62	0.39	0.59 (0.048)	(0.49 , 0.67)	0.37 (0.043)	(0.28 , 0.45)
26	0.57	0.36	0.5 (0.036)	(0.42 , 0.56)	0.35 (0.029)	(0.3 , 0.4)
27	0.5	0.4	0.47 (0.02)	(0.43 , 0.51)	0.36 (0.02)	(0.33 , 0.41)
28	0.54	0.34	0.49 (0.036)	(0.41 , 0.55)	0.32 (0.035)	(0.26 , 0.39)
29	0.57	0.5	0.59 (0.034)	(0.52 , 0.66)	0.41 (0.035)	(0.34 , 0.47)
30	0.58	0.33	0.57 (0.038)	(0.49 , 0.64)	0.37 (0.046)	(0.28 , 0.46)
Overall	0.56	0.46	0.55 (0.01)	(0.52, 0.57)	0.41 (0.01)	(0.39, 0.42)

Note: PM: posterior mean; PSD: posterior standard deviation.

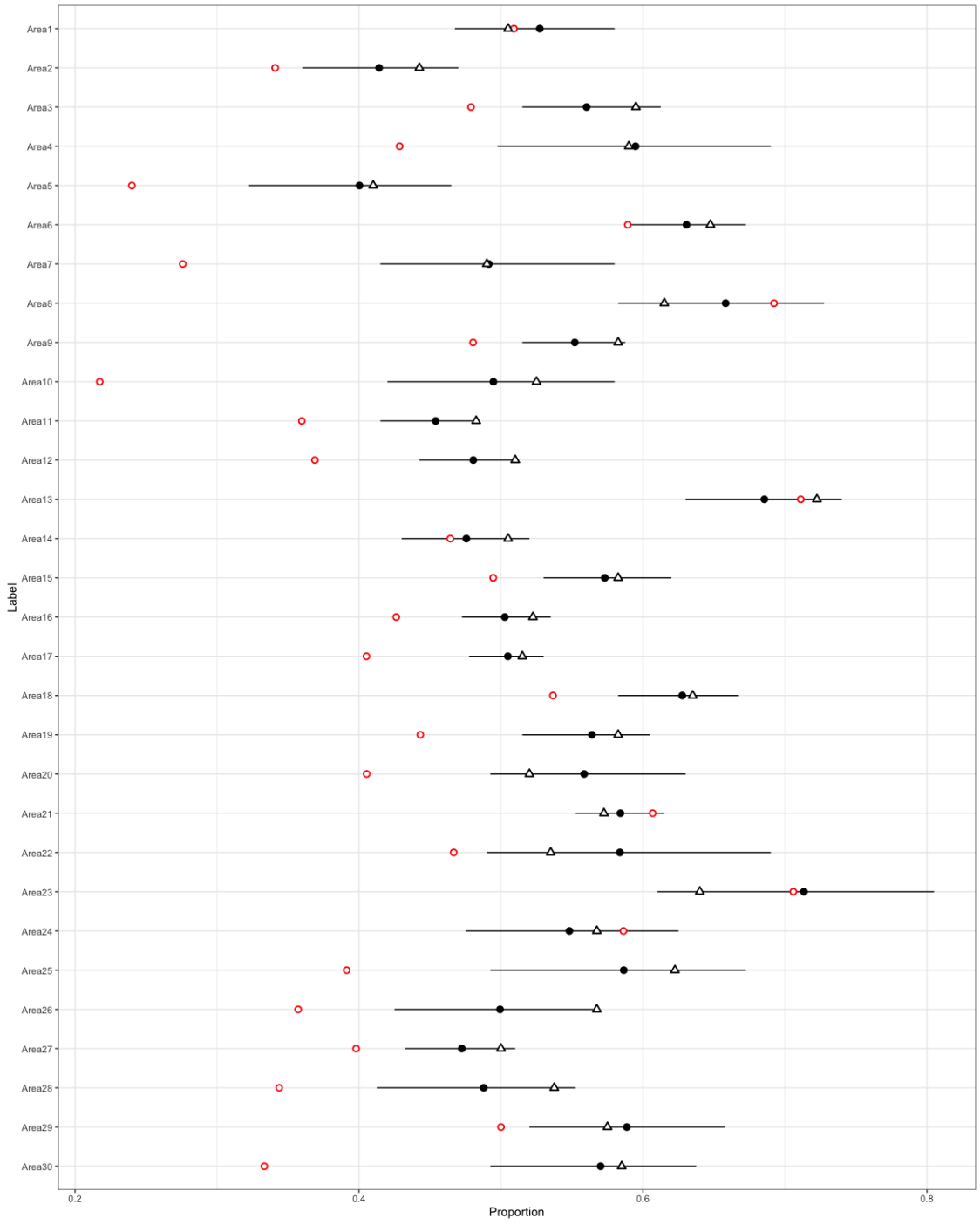


Figure 3.1: Population proportion posterior mean and 95% HPD interval for each area using the nonignorable model. (Triangle: true population proportion. Circle: sample proportion.)

A visual of histograms for both the nonignorable selection model and the ignorable selection model are presented in Figure 3.2. The histogram of the proportion predictions by the nonignorable selection model covers the true population proportion, while the one using the ignorable selection model does not. In addition to the location, the spread of predictions by the nonignorable selection mode is a bit wider than the ignorable selection model, which is likely caused by uncertainty from more parameters, but in a reasonable scale.

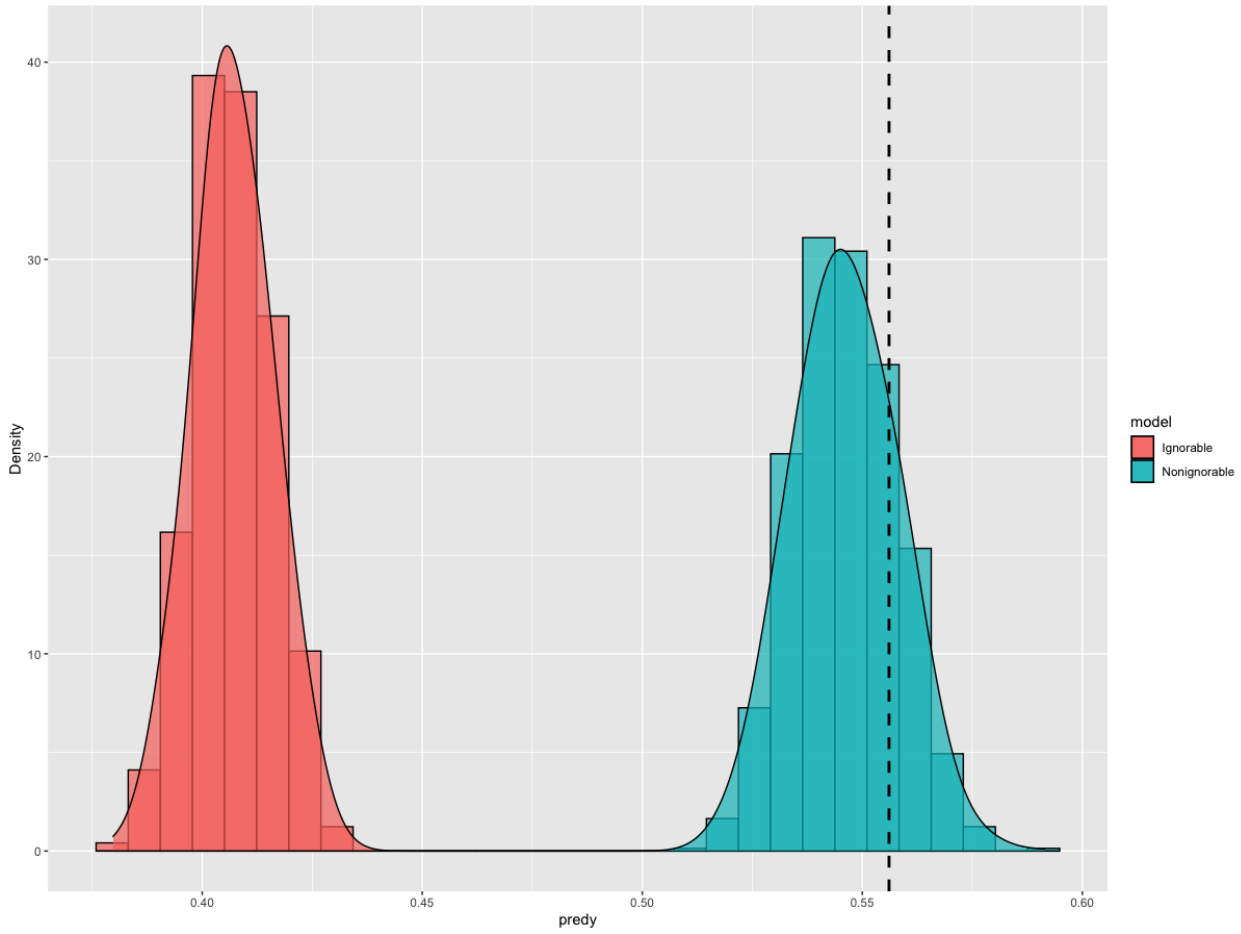


Figure 3.2: Histogram of the overall population proportion predictions by nonignorable selection model (right) and ignorable selection model (left). Dashed line: true population proportion.

3.2 Application on Body Mass Index with Multiple Areas

In this section, we apply both the nonignorable selection model and ignorable selection model on the BMI data with 35 areas (counties) and age, race, sex as covariates. Age values are grouped into 0-4, 5-9, 10-14, and 15-19; and we use 2,7,12,and 17 to represent the four age groups. Race has the value of 0 and 1, where 0 represents black and 1 white. Sex is presented by 0 and 1 for male and female respectively. The total sample size across all counties is 5185, while the population size is 23,106,185. Similar to Section 2.3, our target inference is the proportion of children with healthy BMI, which is defined by greater than or equal to 20 and less than or equal to 25.

3.2.1 Model Reformulation of the Nonignorable Selection Model

We adapt the idea in Section 2.3 to reformulate the model to a more compact form, to increase the algorithm's performance.

We use ℓ to denote the total number of areas and C_i the number of unique sets of age-race-sex values in each area i. Denote the index for areas by $i, i = 1, ..., \ell$; the index for categories, which are the different combinations of covariates, by $j, j = 1, ..., C_i$; and a item within a category by k. Accordingly, within category ij, sample size is denoted by n_{ij} and population size by N_{ij}.

Then, the notation for the count (probability) for each cell within category ij is

	$y_{ijk} = 1$	$y_{ijk} = 0$	
$I_{ijk} = 1$	a_{ij} (p_{ij1})	$n_{ij} - a_{ij}$ (p_{ij2})	n_{ij}
$I_{ijk} = 0$	z_{ij} (p_{ij3})	$N_{ij} - n_{ij} - z_{ij}$ (p_{ij4})	$N_{ij} - n_{ij}$

p_{ij1}, p_{ij2}, p_{ij3}, and p_{ij4} can be derived from

$$y_{ij} \mid \nu_{1i}, \boldsymbol{\gamma} \overset{ind}{\sim} \text{Bernoulli} \left\{ \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})} \right\}, \quad i = 1, ..., l, \quad j = 1, ..., N_i,$$

$$I_{ij} \mid y_{ij}, \nu_{2i}, \boldsymbol{\beta} \overset{ind}{\sim} \text{Bernoulli} \left\{ \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_{ij} + \nu_{2i})}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1}y_{ij} + \nu_{2i})} \right\}, \quad i = 1, ..., l, \quad j = 1, ..., N_i, \tag{3.2.1}$$

It can be shown that

$$\begin{aligned}
p_{ij1} &= P(y_i = 1, I_i = 1 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})} \times \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1} + \nu_{2i})}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1} + \nu_{2i})}, \\
p_{ij2} &= P(y_i = 0, I_i = 1 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{1}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})} \times \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \nu_{2i})}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \nu_{2i})}, \\
p_{ij3} &= P(y_i = 1, I_i = 0 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})} \times \frac{1}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1} + \nu_{2i})}, \\
p_{ij4} &= P(y_i = 0, I_i = 0 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{1}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})} \times \frac{1}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \nu_{2i})}.
\end{aligned} \tag{3.2.2}$$

One can verify that $p_{ij1} + p_{ij2} + p_{ij3} + p_{ij4} = 1$.

The probability mass function for the 2 by 2 contingency table is

$$\begin{aligned}
&P(a_{ij}, n_{ij} - a_{ij}, z_{ij}, N_{ij} - n_{ij} - z_{ij}) \\
&= \frac{N_{ij}!}{a_{ij}!(n_{ij} - a_{ij})!z_{ij}!(N_{ij} - n_{ij} - z_{ij})!} \times {p_{ij1}}^{a_{ij}} {p_{ij2}}^{n_{ij} - a_{ij}} {p_{ij3}}^{z_{ij}} {p_{ij4}}^{N_{ij} - n_{ij} - z_{ij}}.
\end{aligned}$$

Similar to the homogeneous derivation, we sum out z_{ij} and include all priors to derive the joint posterior distribution as

$$
\begin{aligned}
&\pi(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \sigma_1^2, \sigma_2^2 \mid \boldsymbol{I}, \boldsymbol{y}_s) \propto \\
&\prod_{i=1}^{\ell} \prod_{j=1}^{C_i} \left\{ {p_{ij1}}^{a_{ij}} {p_{ij2}}^{n_{ij}-a_{ij}} (p_{ij3}+p_{ij4})^{N_{ij}-n_{ij}} \right\} \times \exp\left(-\frac{1}{2} (\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\gamma}})^{-1} (\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}) \right) \\
&\times \exp\left(-\frac{1}{2} (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\beta}})^{-1} (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \right) \times \prod_{i=1}^{\ell} (\sigma_1^2)^{-\frac{1}{2}} \exp\left(-\frac{\nu_{1i}^2}{2\sigma_1^2} \right) \\
&\times \prod_{i=1}^{\ell} (\sigma_2^2)^{-\frac{1}{2}} \exp\left(-\frac{\nu_{2i}^2}{2\sigma_2^2} \right) \times \frac{1}{(1+\sigma_1^2)^2} \times \frac{1}{(1+\sigma_2^2)^2}.
\end{aligned}
\tag{3.2.3}
$$

Replacing p_{ij1}, p_{ij2}, p_{ij3}, and p_{ij4} by formula (3.2.2), we arrive at the joint posterior distribution

$$
\begin{aligned}
&\pi(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \sigma_1^2, \sigma_2^2 \mid \boldsymbol{I}, \boldsymbol{y}_s) \propto \\
&\prod_{i=1}^{\ell} \prod_{j=1}^{C_i} \left\{ \left(\frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma}+\nu_{1i})}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma}+\nu_{1i})} \times \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)}+\beta_{p+1}+\nu_{2i})}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)}+\beta_{p+1}+\nu_{2i})} \right)^{a_{ij}} \right. \\
&\times \left(\frac{1}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma}+\nu_{1i})} \times \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)}+\nu_{2i})}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)}+\nu_{2i})} \right)^{n_{ij}-a_{ij}} \\
&\times \left(\frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma}+\nu_{1i})}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma}+\nu_{1i})} \times \frac{1}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)}+\beta_{p+1}+\nu_{2i})} \right. \\
&\left. \left. + \frac{1}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma}+\nu_{1i})} \times \frac{1}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)}+\nu_{2i})} \right)^{N_{ij}-n_{ij}} \right\} \\
&\times \exp\left(-\frac{1}{2} (\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\gamma}})^{-1} (\boldsymbol{\gamma}-\hat{\boldsymbol{\gamma}}) \right) \\
&\times \exp\left(-\frac{1}{2} (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\beta}})^{-1} (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \right) \times \prod_{i=1}^{\ell} (\sigma_1^2)^{-\frac{1}{2}} \exp\left(-\frac{\nu_{1i}^2}{2\sigma_1^2} \right) \\
&\times \prod_{i=1}^{\ell} (\sigma_2^2)^{-\frac{1}{2}} \exp\left(-\frac{\nu_{2i}^2}{2\sigma_2^2} \right) \times \frac{1}{(1+\sigma_1^2)^2} \times \frac{1}{(1+\sigma_2^2)^2}.
\end{aligned}
\tag{3.2.4}
$$

Note that form (3.2.4) can be simplified analytically and then the conditional posterior dis-

tribution for each unknown parameter can be further derived. However, to save time we avoid simplifying the form and define a function in R to numerically calculate the conditional posterior density.

The summaries of the finite population proportion overall and in each area found by both the nonignorable and ignorable selection model are presented in Table 3.4. A visual of posterior mean with 95% HPD intervals under the nonignorable selection model versus sample proportion for each area are presented in Figure 3.3. We see that the overall inference from 35 counties are consistent with the result from the one county in Section 2.3. The individual inferences are consistent across counties, and seem not related with the sample proportion, but more affected by covariates structure.

From histograms of both models in Figure 3.4, we see that the histogram of the ignorable model is centered close to the sample proportion, which is 0.23, and with a small spread. By incorporating the selection submodel, the histogram of the nonignorable model is centered around 0.6 but with a relatively larger spread.

Table 3.4: Posterior summaries of finite population proportions

		Nonignorable Model		Ignorable Model	
County	$\bar{\boldsymbol{y}}_s$	PM (PSD)	95% HPD	PM (PSD)	95% HPD
4013	0	0.53 (0.058)	(0.41 , 0.63)	0.24 (0.023)	(0.19 , 0.28)
6001	0	0.64 (0.058)	(0.52 , 0.75)	0.31 (0.029)	(0.26 , 0.36)
6019	0	0.57 (0.061)	(0.46 , 0.69)	0.25 (0.025)	(0.21 , 0.3)
6037	0.5	0.61 (0.051)	(0.5 , 0.71)	0.29 (0.017)	(0.26 , 0.32)
6059	0.07	0.57 (0.055)	(0.46 , 0.67)	0.26 (0.027)	(0.21 , 0.31)
6071	0.09	0.55 (0.061)	(0.44 , 0.67)	0.24 (0.023)	(0.19 , 0.28)
6073	0.4	0.59 (0.06)	(0.48 , 0.71)	0.25 (0.025)	(0.2 , 0.3)
6085	0	0.64 (0.065)	(0.51 , 0.76)	0.3 (0.034)	(0.23 , 0.36)
6111	0.17	0.63 (0.06)	(0.51 , 0.74)	0.3 (0.028)	(0.24 , 0.35)
12025	1	0.66 (0.061)	(0.53 , 0.77)	0.33 (0.029)	(0.27 , 0.38)
12031	0.38	0.58 (0.059)	(0.47 , 0.69)	0.24 (0.026)	(0.19 , 0.29)
12099	0.88	0.58 (0.059)	(0.46 , 0.69)	0.27 (0.028)	(0.21 , 0.32)
17031	0	0.59 (0.054)	(0.49 , 0.7)	0.27 (0.02)	(0.23 , 0.31)
25017	0	0.55 (0.062)	(0.42 , 0.67)	0.24 (0.032)	(0.18 , 0.3)
26125	0.07	0.59 (0.062)	(0.47 , 0.7)	0.26 (0.027)	(0.21 , 0.31)
26163	0	0.58 (0.06)	(0.47 , 0.71)	0.26 (0.026)	(0.21 , 0.31)
29189	0	0.57 (0.062)	(0.45 , 0.69)	0.24 (0.029)	(0.18 , 0.29)
36029	0.08	0.57 (0.061)	(0.46 , 0.7)	0.24 (0.029)	(0.19 , 0.31)
36047	0.33	0.57 (0.064)	(0.46 , 0.7)	0.24 (0.027)	(0.2 , 0.3)
36059	0.3	0.59 (0.06)	(0.48 , 0.71)	0.26 (0.033)	(0.19 , 0.32)
36061	0	0.66 (0.058)	(0.56 , 0.78)	0.34 (0.03)	(0.28 , 0.39)
36081	0.5	0.59 (0.062)	(0.47 , 0.7)	0.26 (0.032)	(0.2 , 0.32)
36119	0.5	0.56 (0.062)	(0.45 , 0.69)	0.25 (0.029)	(0.2 , 0.31)
39035	0.4	0.59 (0.056)	(0.47 , 0.68)	0.27 (0.03)	(0.22 , 0.33)
39061	1	0.54 (0.064)	(0.42 , 0.67)	0.24 (0.026)	(0.19 , 0.29)
42003	1	0.63 (0.058)	(0.52 , 0.74)	0.3 (0.032)	(0.24 , 0.36)
42045	0.67	0.62 (0.059)	(0.51 , 0.73)	0.3 (0.031)	(0.25 , 0.37)
42101	0.86	0.58 (0.066)	(0.44 , 0.7)	0.24 (0.031)	(0.18 , 0.3)
44007	0	0.59 (0.058)	(0.49 , 0.71)	0.26 (0.031)	(0.21 , 0.32)
48029	0	0.62 (0.054)	(0.52 , 0.73)	0.29 (0.027)	(0.24 , 0.35)
48113	0	0.6 (0.057)	(0.49 , 0.71)	0.28 (0.027)	(0.24 , 0.34)
48141	0.1	0.61 (0.061)	(0.5 , 0.73)	0.27 (0.028)	(0.22 , 0.33)
48201	0	0.63 (0.056)	(0.53 , 0.75)	0.3 (0.025)	(0.25 , 0.35)
48439	0.2	0.63 (0.057)	(0.51 , 0.73)	0.3 (0.029)	(0.24 , 0.35)
53033	0	0.61 (0.058)	(0.5 , 0.72)	0.28 (0.028)	(0.23 , 0.34)
Overall	0.23	0.60 (0.05)	(0.50, 0.68)	0.28 (0.01)	(0.26, 0.29)

Note: PM: posterior mean; PSD: posterior standard deviation.

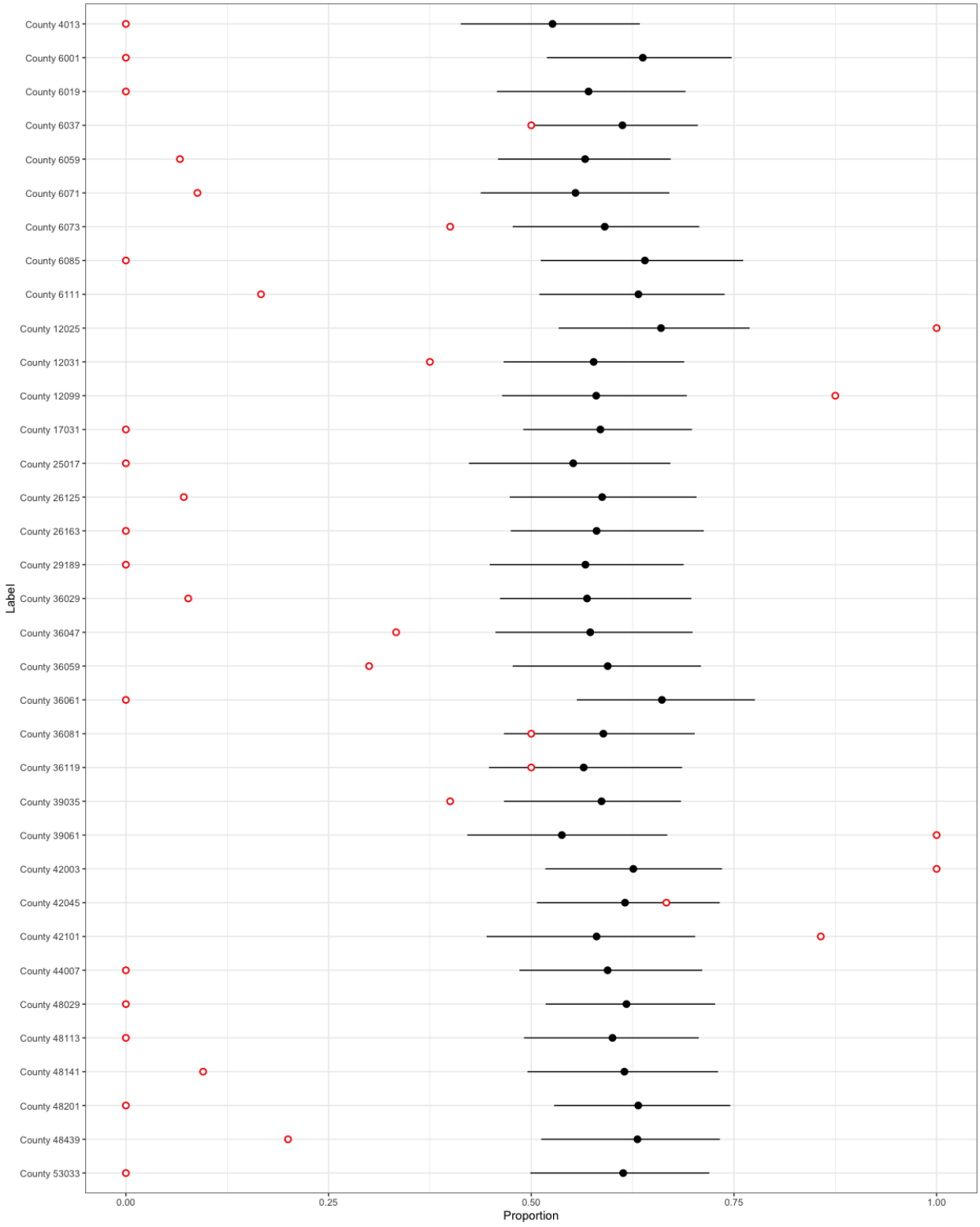


Figure 3.3: Population proportion posterior mean and 95% HPD interval for each county under the nonignorable model. (Circle: sample proportion.)

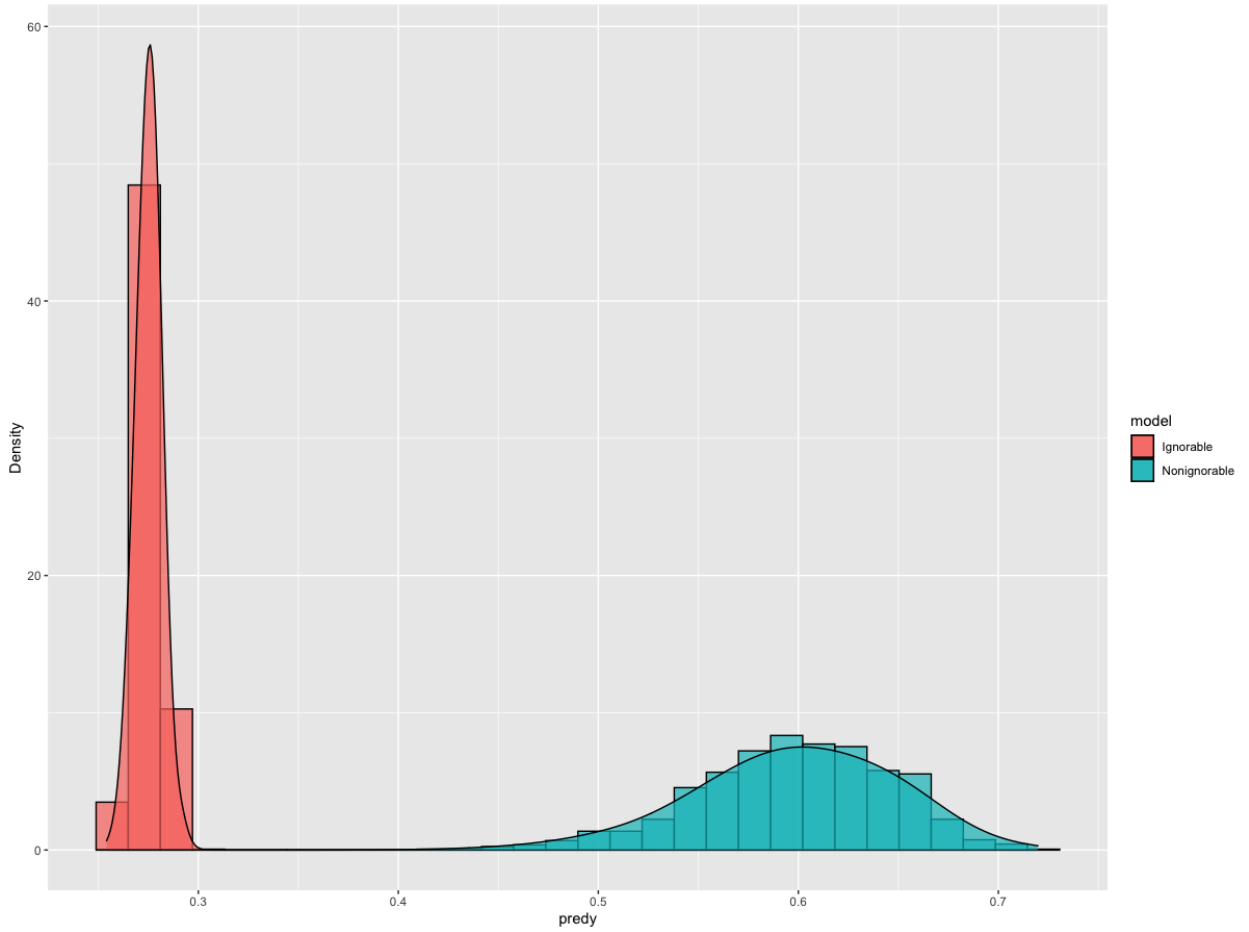


Figure 3.4: Histogram of Population Proportion Prediction by nonignorable selection model (right) and ignorable selection model (left).

3.3 Model with Dirichlet Process Prior

In this section, we extend the model to allow for sharing the same area effects across multiple areas. For example, areas close by could have similar area effects and thus can be grouped into clusters. Our approach is to place Dirichlet process (Ferguson 1973) priors on area variables ν_{1i} and ν_{2i}.

The Dirichlet process (DP) DP(α, G_0) has two parameters, a concentration parameter $\alpha > 0$, and a base probability measure, G_0. An explicit representation of a draw from a DP was given by Sethuraman (1994), who showed that if $G \sim DP(\alpha, G_0)$, then, with probability 1,

$$G(\nu) = \sum_{k=1}^{\infty} w_k \, \Delta_{\nu_k}(\nu),$$

where ν_k are independent random variables distributed according to G_0; Δ_{ν_k} is a point mass CDF concentrated at ν_k and weight w_k is random and defined by $w_1 = V_1$, $w_k = (1-V_1)(1-V_2)\ldots(1-V_{k-1})V_k$, $k \geq 2$, where V_k are independently drawn from Beta$(1, \alpha)$. This representation is called "stick-breaking" because it can be thought of as a stick-breaking procedure, where at each step, a stick of length w_k is broken off from a stick of unit length.

This representation shows that draws from a DP are discrete, which makes it is well suited for the problem of placing priors on mixture components in mixture modeling in survey sample and small area estimation. A significant amount of research has been conducted in DP mixture models (DPM) in survey sampling and small area estimation (Antoniak 1974; Escobar and West 1995; Maceachern and Muller 1998; Nandram and Yin 2016a, 2016b; Yin and Nandram 2020a, 2020b)

We maintain the structure in (3.1.1) and (3.1.2) but replace the priors for ν_{1i} and ν_{2i} with a DP prior, which corresponds to two DPM models.

Response submodel

$$
\begin{aligned}
&y_{ij} \mid \nu_{1i}, \boldsymbol{\gamma} \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma}+\nu_{1i})}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma}+\nu_{1i})}\right\}, \quad i=1,...,\ell, \quad j=1,...,N_i,\\
&\nu_{1i} \mid G_1 \overset{iid}{\sim} DP(\alpha_1, G_{01}), \quad i=1,...,\ell,\\
&\pi(\boldsymbol{\gamma}) \sim \text{N}(\hat{\boldsymbol{\gamma}}, 10^2\hat{\Sigma}_{\boldsymbol{\gamma}}),\\
&\pi(\sigma_1^2) \propto \frac{1}{(1+\sigma_1^2)^2},
\end{aligned}
\tag{3.3.1}
$$

where G_{01} is the CDF of a normal random variable with mean 0 and variance σ_1^2.

Selection submodel

$$
\begin{aligned}
&I_{ij} \mid y_{ij}, \nu_{2i}, \boldsymbol{\beta} \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)}+\beta_{p+1}y_{ij}+\nu_{2i})}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)}+\beta_{p+1}y_{ij}+\nu_{2i})}\right\}, \quad i=1,...,\ell, \quad j=1,...,N_i,\\
&\nu_{2i} \mid G_2 \overset{iid}{\sim} DP(\alpha_2, G_{02}), \quad i=1,...,\ell,\\
&\pi(\boldsymbol{\beta}) \sim \text{N}(\hat{\boldsymbol{\beta}}, 10^2\hat{\Sigma}_{\boldsymbol{\beta}}),\\
&\pi(\sigma_2^2) \propto \frac{1}{(1+\sigma_2^2)^2},
\end{aligned}
\tag{3.3.2}
$$

where G_{02} is the CDF of a normal random variable with mean 0 and variance σ_2^2.

3.3.1 Slice Sampling

Sampling from the DPM model can be difficult due to the infinite dimensional problem. An original algorithm by Escobar (1988), and some variations (MacEachern 1994, MacEachern and Muller 1998, Neal 2000) rely on integrating out the Dirichlet process from the model. These are usually referred to as "marginal" methods. Walker (2007) used slice sampling ideas and found a way to

sample a finite number of variables at each iteration, which is defined as the "conditional" method. Our sampling approach is developed from the slice sampling algorithm proposed by Kalli, Griffin and Walker (2011) and extended to include a sampler for α.

We focus on the details of sampling $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ as the sampling methods for other variables are unchanged. We first look at the hierarchical submodel that involves $\boldsymbol{\nu}_1$

$$\begin{aligned} &y_{ij} \mid \nu_{1i}, ... \overset{ind}{\sim} \text{Bernoulli}\left(p_{ij}(\nu_{1i})\right), \quad i = 1, ..., \ell, \quad j = 1, ..., N_i, \\ &\nu_{1i} \mid G_1 \overset{iid}{\sim} DP(\alpha_1, G_{01}), \quad i = 1, ..., \ell, \end{aligned} \tag{3.3.3}$$

where

$$p_{ij}(\nu_{1i}) = \frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})\left(1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \nu_{2i})\right)}{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})\left(1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \nu_{2i})\right) + 1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1} + \nu_{2i})}$$

from Section 3.1 (vii) and G_{01} is the CDF of a normal random variable with mean 0 and variance σ_1^2.

By using a stick-breaking representation,

$$G_1(\nu_1) = \sum_{k=1}^{\infty} w_k \, \delta_{\nu_{1k}}(\nu_1),$$

where ν_{11},ν_{12} ,ν_{13},... are independent and identically distributed from G_{01} and

$$w_1 = V_1, \quad w_k = V_k \prod_{\kappa<k}(1 - V_\kappa)$$

with V_k being independent and identically distributed from Beta$(1, \alpha_1)$, we can write

$$f(y \mid \boldsymbol{w}, \boldsymbol{\nu}_1, ...) = \sum_{k=1}^{\infty} w_k \text{Bernoulli}\left(p(\nu_{1k})\right). \tag{3.3.4}$$

As one can see, (3.3.4) is an infinite mixture model and we will describe how to change this

into a finite number of components. Our starting point is to introduce a random variable u such that the joint density of y and u satisfies

$$f(y, u \mid \boldsymbol{w}, \boldsymbol{\nu}_1, ...) = \sum_{k=1}^{\infty} \mathbf{1}(u < w_k) \text{Bernoulli}\left(p(\nu_{1k})\right). \tag{3.3.5}$$

Then given u, the number of components is finite, with the indices for the components being $A_u = \{k : w_k > u\}$.

To indicate which observations (area) belongs to which components, we introduce the vector $\boldsymbol{d}$ with the length of ℓ. For example, $d = (5, 2, 5, ...)$ means that the first observation (area) and the third belong to the same component (cluster) 5. Then, the joint posterior distribution is proportional to

$$f(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{d} \mid \boldsymbol{w}, \boldsymbol{\nu}_1, ...) \propto \prod_{i=1}^{\ell} \prod_{j=1}^{N_i} \mathbf{1}(u_i < w_{d_i}) \text{Bernoulli}\left(p_{ij}(\nu_{1d_i})\right), \tag{3.3.6}$$

Mixing rate can be further introduce by introducing a positive decreasing sequence $\xi_1, \xi_2, \xi_3, ...$ to the joint distribution (Kalli, Griffin and Walker, 2011). Our choice of the sequence is $\xi_j = (1-k)k^{j-1}$, $k \in (0, 1)$ and k is a tuning parameter, we find that k=0.75 strikes a good mixing. Finally, the joint posterior distribution is

$$f(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{d} \mid \boldsymbol{w}, \boldsymbol{\nu}_1, ...) \propto \xi_{d_i}^{-1} \prod_{i=1}^{\ell} \prod_{j=1}^{N_i} \mathbf{1}(u_i < \xi_{d_i}) w_{d_i} \text{Bernoulli}\left(p_{ij}(\nu_{1d_i})\right), \tag{3.3.7}$$

Next, we draw $\boldsymbol{\nu}_{1d_i}$ from this joint posterior distribution. The algorithm can be easily adapted for $\boldsymbol{\nu}_{2d_i}$.

Goal: To sample $\{\nu_{1k}, k = 1, 2, ...; (d_i, u_i), i = 1, ..., \ell\}$ at each sweep of the Gibbs sampler. Theoretically, the index k goes to infinity, but in practice, one can use the integer M where $M = \max_i\{M_i\}$ and M_i is the largest integer m for which $\xi_m > u_i$. Give a initial value for $\boldsymbol{d}$ and α_1.

Algorithm For $(t = 1, ..., T)$, repeat:

Step 1. Execute the algorithm in Section 3.1 steps 1 - 4 to draw samples of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$.

Step 2. Now start the stick-breaking process to draw $\boldsymbol{\nu}_1$:

1. Draw $\boldsymbol{u}$, where $u_i \sim U(0, \xi_{d_i}), i = 1, 2, ..., \ell$. Then, determine M such that M is the largest index for which $\xi_M > \min(\boldsymbol{u})$.

2. Based on the index $\boldsymbol{d}$, draw values of ν_{1k}, $k = 1, 2, ..., M$. For each k:

   – If there is d_i, $i = 1, ..., l$ equals to k, denote the set of all such i by A_k, draw ν_{1k} from its posterior distribution
   $f(\nu_{1k} \mid \boldsymbol{y}, \boldsymbol{\gamma}, \sigma_1^2) \propto \prod_{i \in A_k} \left\{ \prod_{j=1}^{N_i} \frac{\exp(\nu_{1i} y_{ij})}{1 + \exp(\boldsymbol{x}_{ij}' \boldsymbol{\gamma} + \nu_{1i})} \times \exp\left(-\frac{\nu_{1i}^2}{2\sigma_1^2} \right) \right\}$.

   – If there is no d_i equals to k, then draw ν_{1k} from the prior $\mathrm{N}(0, \sigma_1^2)$.

3. Draw $V_k, k = 1, 2, ..., M$ from $\mathrm{Beta}(a_k, b_k)$, where

$$a_k = 1 + \sum_{i=1}^{n} \mathbf{1}(d_i = k)$$
$$b_k = \alpha_1 + \sum_{i=1}^{n} \mathbf{1}(d_i > k)$$

   Then, $w_k, j = 1, 2, ..., M$ can be calculated by $w_1 = V_1$, $w_k = V_k \prod_{\kappa<k}(1 - V_\kappa)$ for $k > 1$.

4. Then, regroup the observations (areas) by first calculating the probability of the observation (area) i assigning to each component k:

$$P_{ik}(d_i = k|...) \propto \mathbf{1}(k : \xi_k > u_i) w_k / \xi_k \prod_{j=1}^{N_i} \text{Bernoulli}\left(y_{ij}; p(\nu_{1k})\right).$$

   For each i, scale the probability such that $\sum_{k=1}^{M} P_{ik} = 1$ and draw a integer from 1 to M with the rescaled probabilities; assign it to d_i and proceed for all the is.

5. Based on the updated $\boldsymbol{d}$, find the cluster number k. Use k to draw α_1 with details in

Appendix D.

Step 3. Adapt the Step 2 to $\boldsymbol{\nu}_2$, and proceed with the rest of the algorithms in Section 3.1 steps 6-8. Repeat from the Step 1.

3.3.2 Simulation Study

We conduct a similar simulation study as the one in Section 3.1.1. The population size is 12,000 with $\ell = 30$ areas. The coefficients for the population model are $\boldsymbol{\gamma}_{True} = (3, 6, -2, -3)$, and the terms ν_{1i} are sampled from a DP that gives 6 unique values. For the selection submodel, the coefficients are $\boldsymbol{\beta}_{True} = (0.8, 0.8, -3, -2, -2.5)$, and area effects ν_{2i} are drawn from a DP that gives 5 unique values.

The population proportion is 0.53, the sample proportion is 0.40, with the total sample size among all areas is 1282. We run 30,000 iterations; burn-in the first 5,000; thinning by 25 and end up with 1,000 samples.

Results

For simplicity, we omit presenting the diagnostics but only show the summaries of finite proportion predictions by area and overall under both the nonignorable and ignorable selection model (Table 3.5). We see that the 95% HPD interval from the nonignorable model covers the true population proportion for each area, which is also shown in Figure 3.5. The density of the finite population proportions for each area under the nonignorable selection model is presented in Figure 3.6. The histogram comparison of the overall population proportions under the two models is in Figure 3.7, which shows that the histogram from the nonignorable selection model covers the true value while the ignorable model does not.

We also show some graphics for the area effects $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ in Figures 3.8 and 3.9. Figure 3.8 gives us an idea of how values of ν_{1i} are distributed for each group, and how they are spread out

among groups, from which we can (vaguely) tell how they form the clusters. Figure 3.9 gives the curves for ν_{2i} by group. It is clearer that there are possibly 4 clusters. Figures 3.10 and 3.11 are the histograms for k_1 and k_2, which show the number of clusters for area effects of the response and selection submodel respectively. We see that the modes are at 6 and 5, which are the same as the true values (the numbers of the unique values in simulation), indicating that our model correctly discovers the clusters.

Table 3.5: Posterior summaries of finite population proportions

			Nonignorable Model		Ignorable Model	
Area	True $\bar{Y}$	$\bar{\boldsymbol{y}}_s$	PM (PSD)	95% HPD	PM (PSD)	95% HPD
1	0.54	0.47	0.54 (0.028)	(0.48 , 0.59)	0.4 (0.031)	(0.34 , 0.46)
2	0.5	0.5	0.54 (0.032)	(0.49 , 0.6)	0.37 (0.036)	(0.3 , 0.43)
3	0.51	0.38	0.5 (0.025)	(0.45 , 0.55)	0.33 (0.032)	(0.27 , 0.39)
4	0.52	0.3	0.51 (0.027)	(0.46 , 0.56)	0.35 (0.032)	(0.29 , 0.41)
5	0.55	0.4	0.57 (0.029)	(0.52 , 0.62)	0.39 (0.038)	(0.32 , 0.47)
6	0.53	0.3	0.55 (0.029)	(0.49 , 0.6)	0.29 (0.045)	(0.2 , 0.38)
7	0.57	0.44	0.55 (0.026)	(0.5 , 0.6)	0.44 (0.038)	(0.38 , 0.52)
8	0.53	0.46	0.55 (0.025)	(0.5 , 0.6)	0.42 (0.039)	(0.34 , 0.5)
9	0.52	0.31	0.54 (0.027)	(0.49 , 0.59)	0.37 (0.038)	(0.3 , 0.44)
10	0.5	0.32	0.52 (0.028)	(0.46 , 0.56)	0.31 (0.035)	(0.24 , 0.38)
11	0.47	0.3	0.48 (0.023)	(0.43 , 0.52)	0.31 (0.029)	(0.26 , 0.37)
12	0.49	0.42	0.55 (0.028)	(0.48 , 0.59)	0.4 (0.039)	(0.31 , 0.46)
13	0.52	0.27	0.5 (0.026)	(0.46 , 0.56)	0.24 (0.034)	(0.18 , 0.31)
14	0.57	0.34	0.53 (0.025)	(0.48 , 0.57)	0.39 (0.026)	(0.34 , 0.44)
15	0.52	0.33	0.53 (0.025)	(0.48 , 0.58)	0.28 (0.043)	(0.2 , 0.37)
16	0.48	0.39	0.53 (0.029)	(0.47 , 0.58)	0.38 (0.029)	(0.32 , 0.43)
17	0.52	0.43	0.53 (0.029)	(0.47 , 0.58)	0.3 (0.045)	(0.21 , 0.38)
18	0.57	0.56	0.57 (0.03)	(0.51 , 0.63)	0.41 (0.036)	(0.34 , 0.48)
19	0.56	0.47	0.58 (0.025)	(0.54 , 0.63)	0.4 (0.033)	(0.33 , 0.46)
20	0.56	0.37	0.53 (0.028)	(0.48 , 0.58)	0.37 (0.043)	(0.29 , 0.46)
21	0.58	0.44	0.56 (0.027)	(0.51 , 0.61)	0.42 (0.036)	(0.34 , 0.48)
22	0.52	0.34	0.55 (0.029)	(0.49 , 0.6)	0.34 (0.038)	(0.26 , 0.41)
23	0.56	0.48	0.57 (0.03)	(0.51 , 0.62)	0.34 (0.042)	(0.26 , 0.42)
24	0.49	0.42	0.51 (0.025)	(0.46 , 0.56)	0.35 (0.033)	(0.28 , 0.41)
25	0.56	0.53	0.57 (0.03)	(0.51 , 0.62)	0.42 (0.038)	(0.34 , 0.5)
26	0.52	0.38	0.53 (0.032)	(0.47 , 0.58)	0.34 (0.042)	(0.25 , 0.42)
27	0.6	0.48	0.56 (0.033)	(0.5 , 0.63)	0.37 (0.033)	(0.32 , 0.44)
28	0.56	0.42	0.57 (0.028)	(0.52 , 0.63)	0.4 (0.038)	(0.32 , 0.47)
29	0.5	0.38	0.52 (0.022)	(0.47 , 0.55)	0.36 (0.026)	(0.31 , 0.41)
30	0.52	0.51	0.58 (0.028)	(0.52 , 0.63)	0.46 (0.031)	(0.41 , 0.52)
Overall	0.53	0.40	0.54 (0.02)	(0.51, 0.57)	0.36 (0.01)	(0.34, 0.39)

Note: PM: posterior mean; PSD: posterior standard deviation.

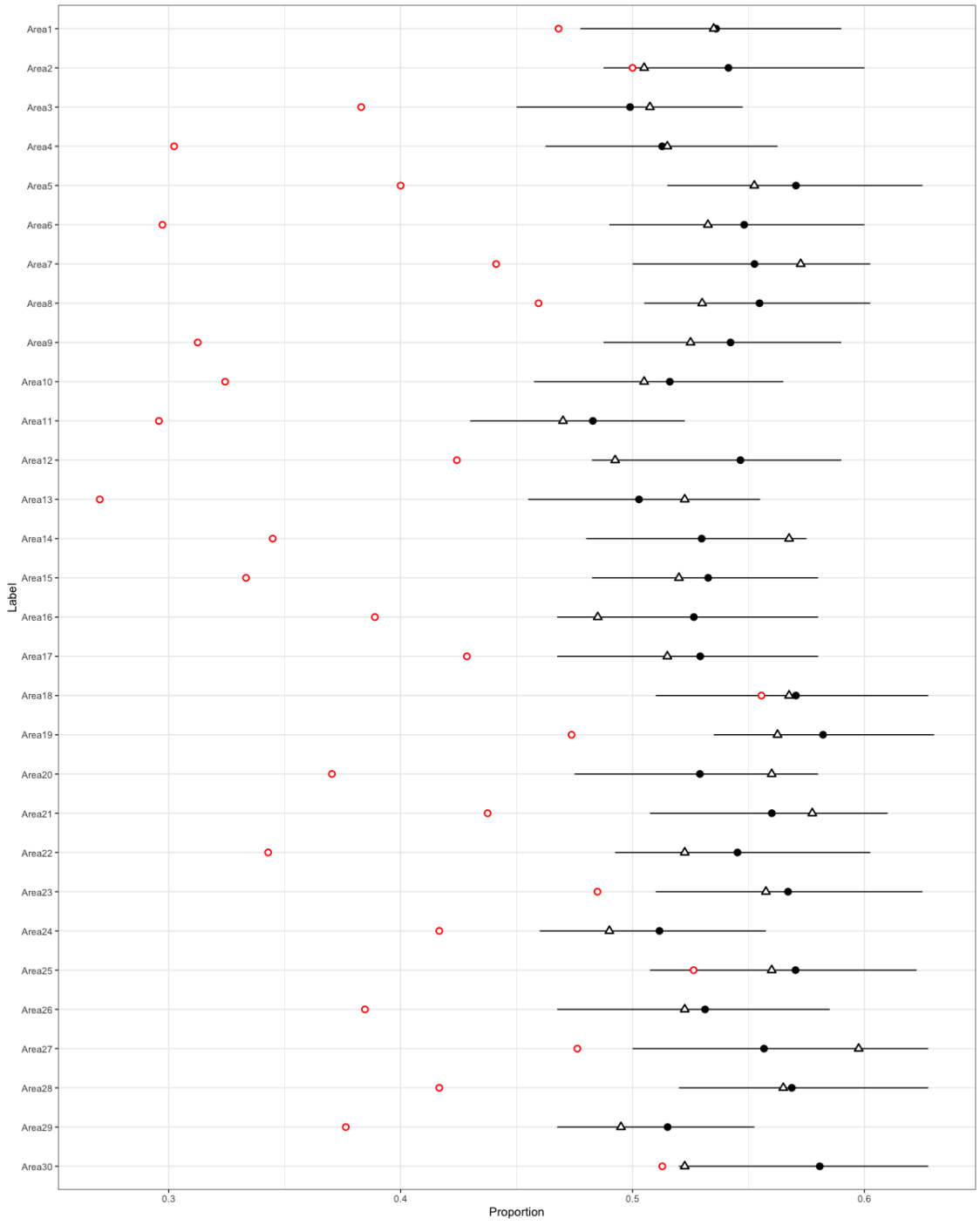


Figure 3.5: Population proportion posterior mean and 95% HPD interval for each area under the nonignorable model. (Triangle: true population proportion. Circle: sample proportion.)

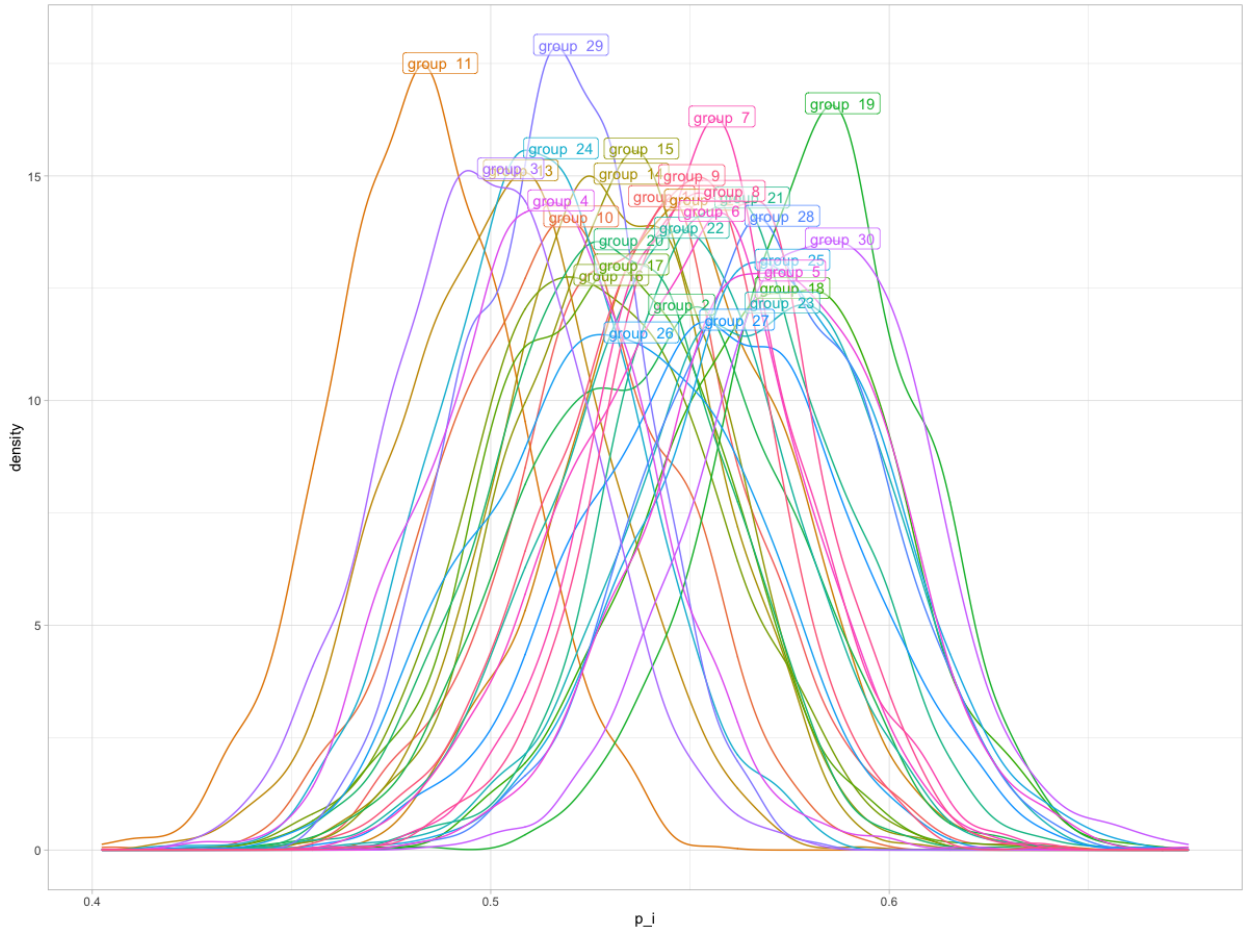


Figure 3.6: Density of Population Proportion Predictions by Area under the Nonignorable Selection Model

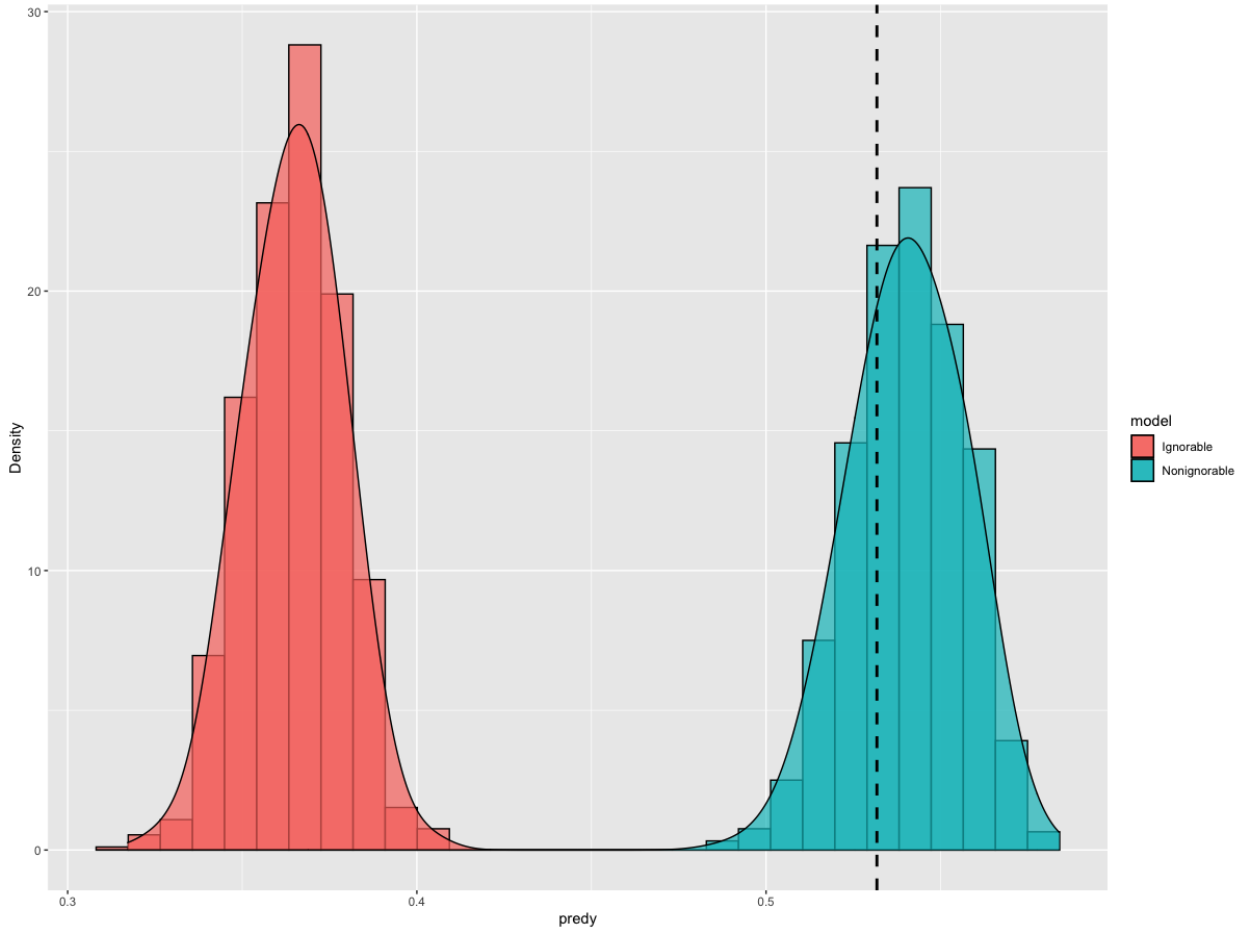


Figure 3.7: Histogram of the overall population proportion predictions by nonignorable selection model (right) and ignorable selection model (left). Dashed line: true population proportion.

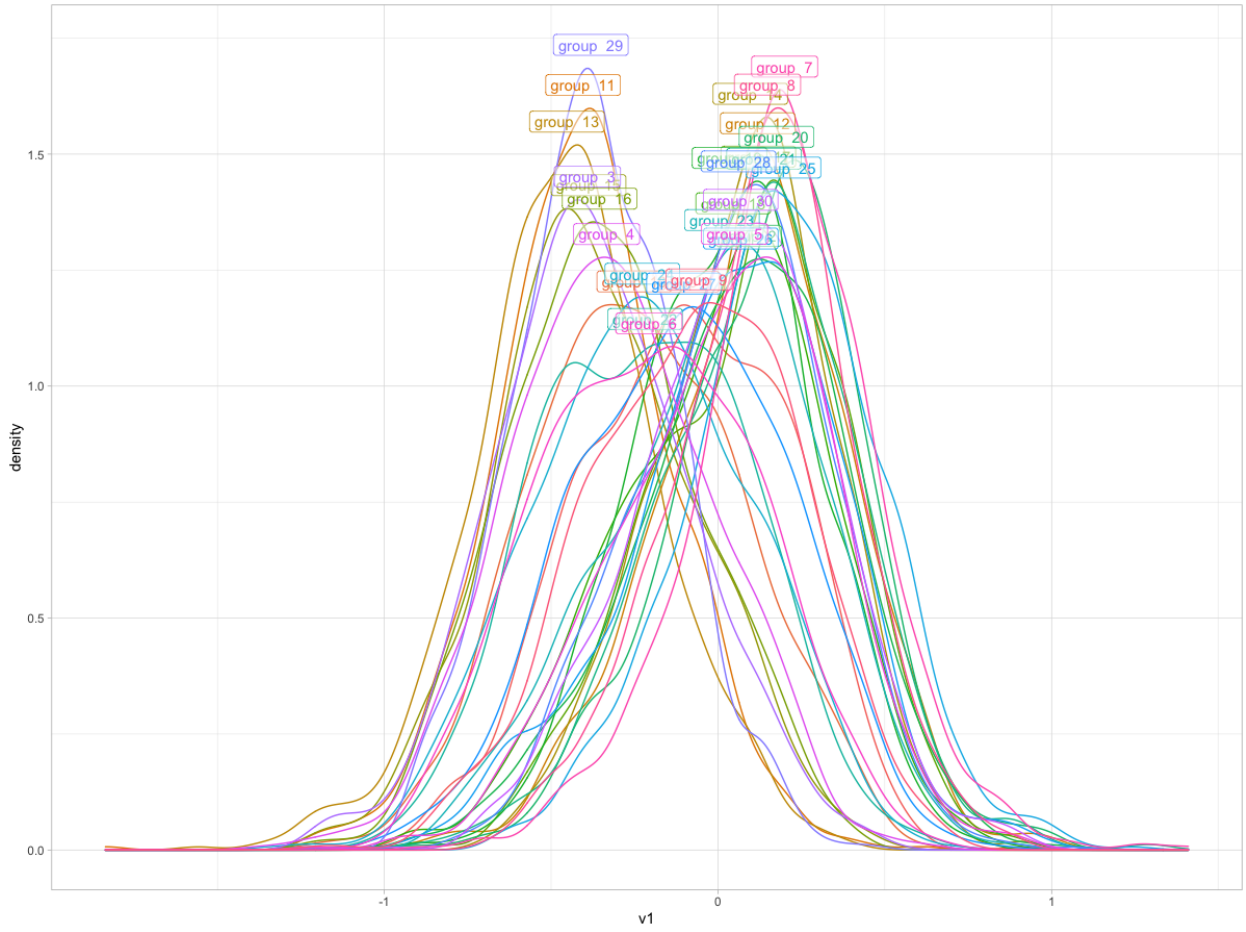


Figure 3.8: Density of $\boldsymbol{\nu}_1$ by group

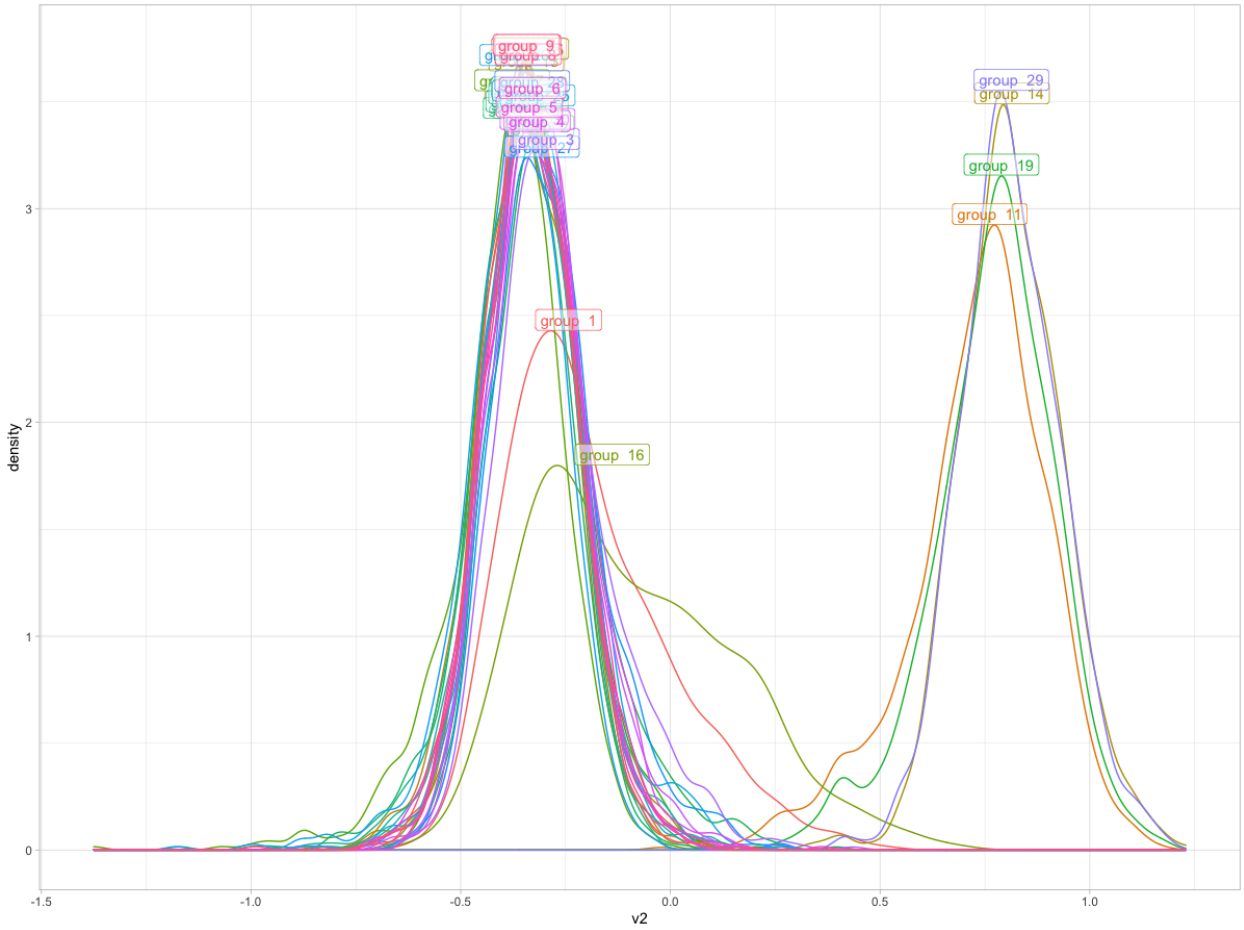


Figure 3.9: Density of $\boldsymbol{\nu}_2$ by group

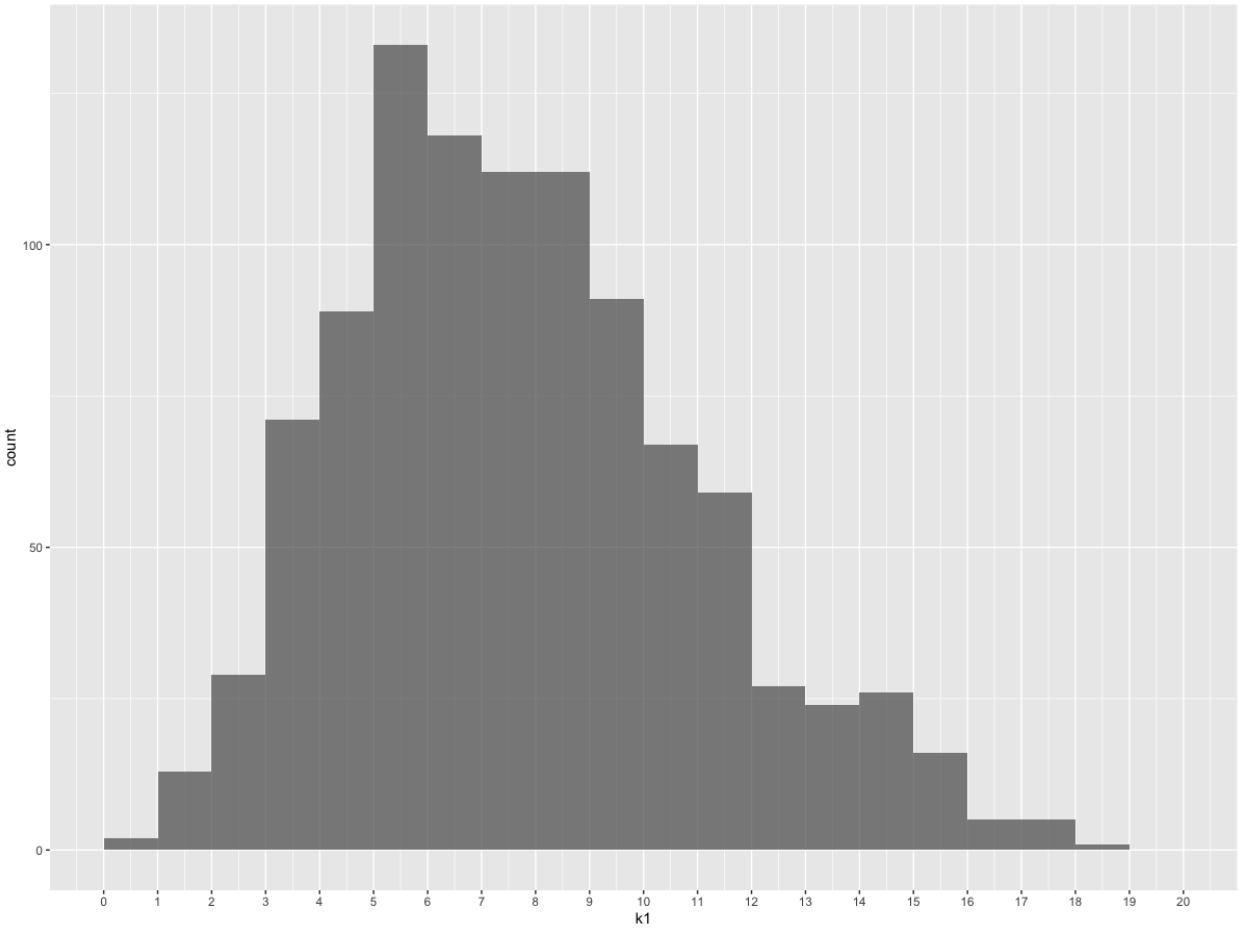


Figure 3.10: Histogram of k_1

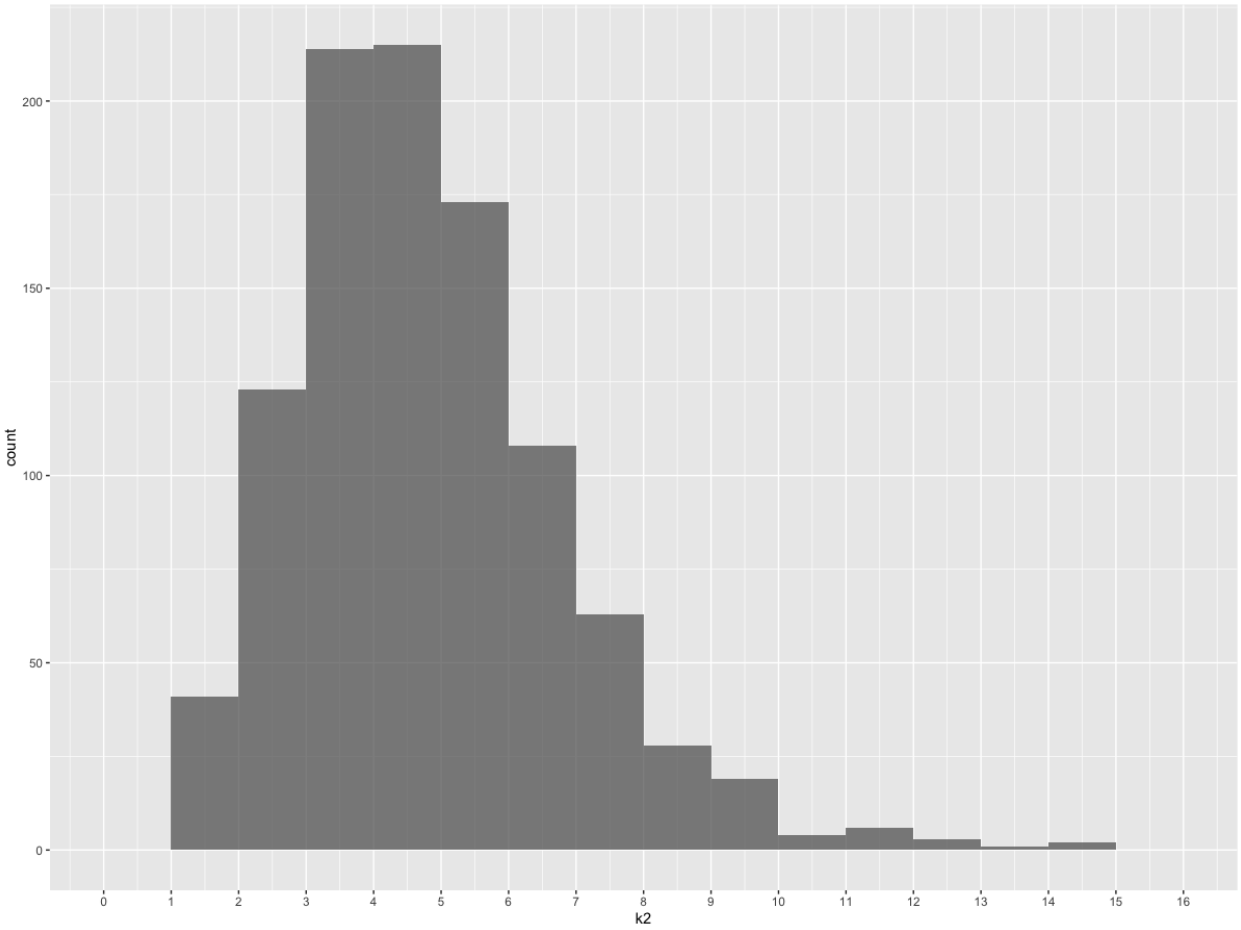


Figure 3.11: Histogram of k_2

3.4 Model with Bivariate Prior for Area Effects

In this section, we model (ν_{1i}, ν_{2i}) together with a bivariate normal distirbution as a prior. In this way, the area effects for the response model and the missingness model can be pooled together by the correlation between the two variables. The prior is

$$\begin{pmatrix} \nu_{1i} \\ \nu_{2i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right), i = 1, ..., \ell$$

It is straightforward to show that this prior can also be formulated as

$$\begin{aligned} &\nu_{1i} \sim \mathrm{N}(0, \sigma_1^2) \\ &\nu_{2i} \mid \nu_{1i} \sim \mathrm{N}\left(\rho\frac{\sigma_2}{\sigma_1}\nu_{1i}, (1-\rho^2)\sigma_2^2 \right), \quad i = 1, ..., \ell. \end{aligned} \tag{3.4.1}$$

We give ρ a flat prior on $(-1, 1)$.

Using the formulation of (3.4.1), the conditional posterior distributions of ν_{1i}, ν_{2i}, and ρ can be easily derived as

(i) $f(\nu_{1i} \mid \boldsymbol{y}, \boldsymbol{\gamma}, \sigma_1^2) \propto \prod_{j=1}^{N_i} \frac{\exp(\nu_{1i} y_{ij})}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\gamma} + \nu_{1i})} \times \exp\left(-\frac{\nu_{1i}^2}{2\sigma_1^2} \right), \quad i = 1, ..., \ell;$
(Note: this is unchanged from Section 3.1)

(ii) $f(\nu_{2i} \mid \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{I}, \sigma_2, \nu_{1i}, \rho)$
$$\propto \prod_{j=1}^{N_i} \frac{\exp(\nu_{2i} I_{ij})}{1 + \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_{ij} + \nu_{2i})} \times \exp\left(-\frac{1}{2}\left(\frac{\nu_{2i} - \rho\frac{\sigma_2}{\sigma_1}\nu_{1i}}{\sqrt{1-\rho^2}\sigma_2} \right)^2 \right), \quad i = 1, ..., l;$$

(iii) $f(\rho \mid \sigma_1, \sigma_2, \nu_{1i}, \nu_{2i}) \propto \left(\frac{1}{\sqrt{1-\rho^2}} \right)^l \prod_{i=1}^{l} \exp\left(-\frac{1}{2}\left(\frac{\nu_{2i} - \rho\frac{\sigma_2}{\sigma_1}\nu_{1i}}{\sqrt{1-\rho^2}\sigma_2} \right)^2 \right).$

The algorithm is adapted from the one in Section 3.1, with the formula for ν_{2i} replaced and for ρ added to the iteration, where ρ can be sampled by the grid method on $(-1, 1)$.

3.4.1 Simulation Study

A simulation study is conducted in a similar setup as in Section 3.2.2, with the only difference in the generation of ν_{1i} and ν_{2i}. ν_{1i} and ν_2 are drawn from the bivariate normal distribution with mean $(0, 0)'$, and marginal variances being 1 and the correlation ρ being 0.4.

The true population proportion is 0.55; the sample proportion is 0.47, with the total sample size of 1671.

Results

Summaries of population proportion predictions under both models are presented in Table 3.6, with the coverage for each area under the nonignorable selection model visually displayed in Figure 3.12. We see that in Area 5 the 95% HPD interval does not contain the true value. The histogram in Figure 3.13 shows the population proportion predictions under the nonignorable selection model are centered around the true value. We further investigate the diagnostic and summary of ρ (visually displayed in Figure 3.14). The p-value for Geweke's test is 0.36, the effective sample size is 793. The posterior mean is 0.37, while the true value is 0.4, indicating this correlation parameter can be correctly identified.

Table 3.6: Posterior summaries of finite population proportions

			Nonignorable Model		Ignorable Model	
Area	True $\bar{Y}$	$\bar{\boldsymbol{y}}_s$	PM (PSD)	95% HPD	PM (PSD)	95% HPD
1	0.56	0.42	0.5 (0.037)	(0.44 , 0.58)	0.32 (0.032)	(0.25 , 0.38)
2	0.55	0.47	0.56 (0.026)	(0.5 , 0.6)	0.4 (0.028)	(0.35 , 0.45)
3	0.42	0.22	0.42 (0.037)	(0.35 , 0.5)	0.23 (0.03)	(0.17 , 0.28)
4	0.7	0.54	0.66 (0.022)	(0.62 , 0.7)	0.53 (0.025)	(0.48 , 0.58)
5	0.41	0.38	0.55 (0.056)	(0.43 , 0.65)	0.33 (0.059)	(0.22 , 0.44)
6	0.46	0.4	0.47 (0.024)	(0.43 , 0.52)	0.35 (0.02)	(0.3 , 0.38)
7	0.65	0.42	0.6 (0.038)	(0.53 , 0.67)	0.43 (0.04)	(0.35 , 0.5)
8	0.49	0.27	0.49 (0.033)	(0.42 , 0.55)	0.3 (0.028)	(0.25 , 0.36)
9	0.49	0.41	0.53 (0.032)	(0.47 , 0.59)	0.34 (0.04)	(0.26 , 0.42)
10	0.7	0.64	0.64 (0.043)	(0.55 , 0.72)	0.43 (0.046)	(0.34 , 0.52)
11	0.64	0.63	0.68 (0.026)	(0.62 , 0.72)	0.54 (0.025)	(0.49 , 0.58)
12	0.46	0.31	0.45 (0.022)	(0.42 , 0.5)	0.3 (0.021)	(0.26 , 0.34)
13	0.66	0.65	0.64 (0.023)	(0.6 , 0.69)	0.51 (0.021)	(0.48 , 0.56)
14	0.6	0.56	0.61 (0.029)	(0.55 , 0.66)	0.44 (0.026)	(0.39 , 0.49)
15	0.44	0.34	0.44 (0.026)	(0.39 , 0.49)	0.32 (0.019)	(0.28 , 0.36)
16	0.48	0.38	0.51 (0.041)	(0.43 , 0.58)	0.31 (0.042)	(0.22 , 0.38)
17	0.56	0.49	0.55 (0.02)	(0.51 , 0.58)	0.41 (0.018)	(0.37 , 0.44)
18	0.57	0.37	0.54 (0.038)	(0.46 , 0.6)	0.34 (0.035)	(0.26 , 0.4)
19	0.65	0.64	0.66 (0.03)	(0.6 , 0.72)	0.54 (0.03)	(0.48 , 0.6)
20	0.47	0.44	0.5 (0.039)	(0.42 , 0.57)	0.31 (0.027)	(0.26 , 0.36)
21	0.43	0.18	0.46 (0.041)	(0.38 , 0.54)	0.25 (0.045)	(0.17 , 0.34)
22	0.7	0.67	0.67 (0.023)	(0.62 , 0.71)	0.54 (0.025)	(0.48 , 0.58)
23	0.52	0.3	0.48 (0.045)	(0.4 , 0.56)	0.23 (0.033)	(0.18 , 0.3)
24	0.57	0.63	0.59 (0.047)	(0.49 , 0.68)	0.42 (0.043)	(0.34 , 0.5)
25	0.58	0.53	0.59 (0.018)	(0.55 , 0.62)	0.49 (0.018)	(0.46 , 0.52)
26	0.55	0.47	0.57 (0.026)	(0.52 , 0.62)	0.41 (0.029)	(0.35 , 0.46)
27	0.48	0.35	0.49 (0.031)	(0.43 , 0.55)	0.32 (0.027)	(0.27 , 0.37)
28	0.51	0.5	0.54 (0.034)	(0.48 , 0.6)	0.39 (0.026)	(0.33 , 0.44)
29	0.56	0.37	0.54 (0.03)	(0.46 , 0.58)	0.31 (0.03)	(0.25 , 0.36)
30	0.52	0.45	0.51 (0.038)	(0.44 , 0.58)	0.34 (0.033)	(0.27 , 0.4)
Overall	0.55	0.47	0.55 (0.01)	(0.52, 0.57)	0.38 (0.01)	(0.36, 0.40)

Note: PM: posterior mean; PSD: posterior standard deviation.

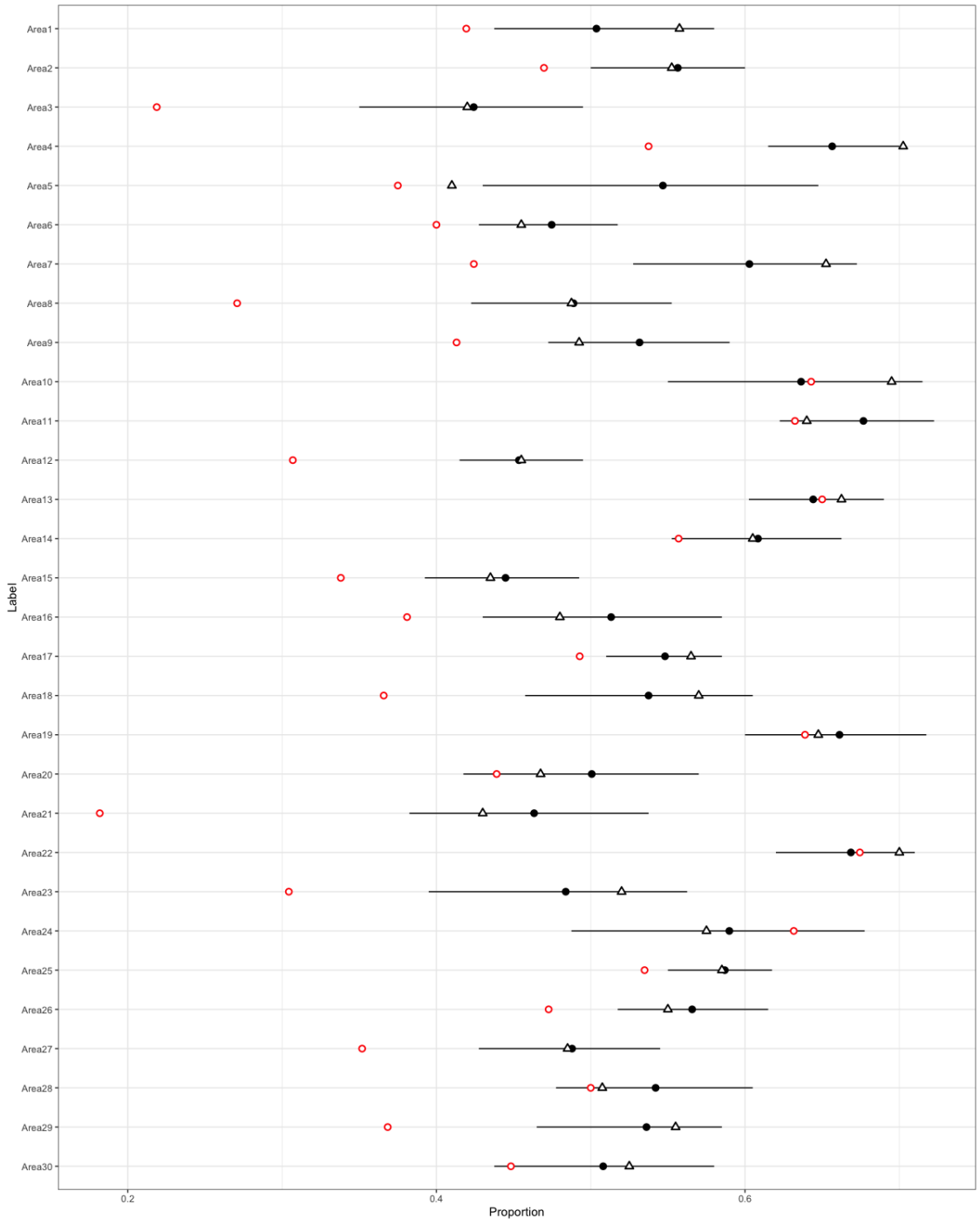


Figure 3.12: Population proportion posterior mean and 95% HPD interval for each area under the nonignorable model. (Triangle: true population proportion. Circle: sample proportion.)

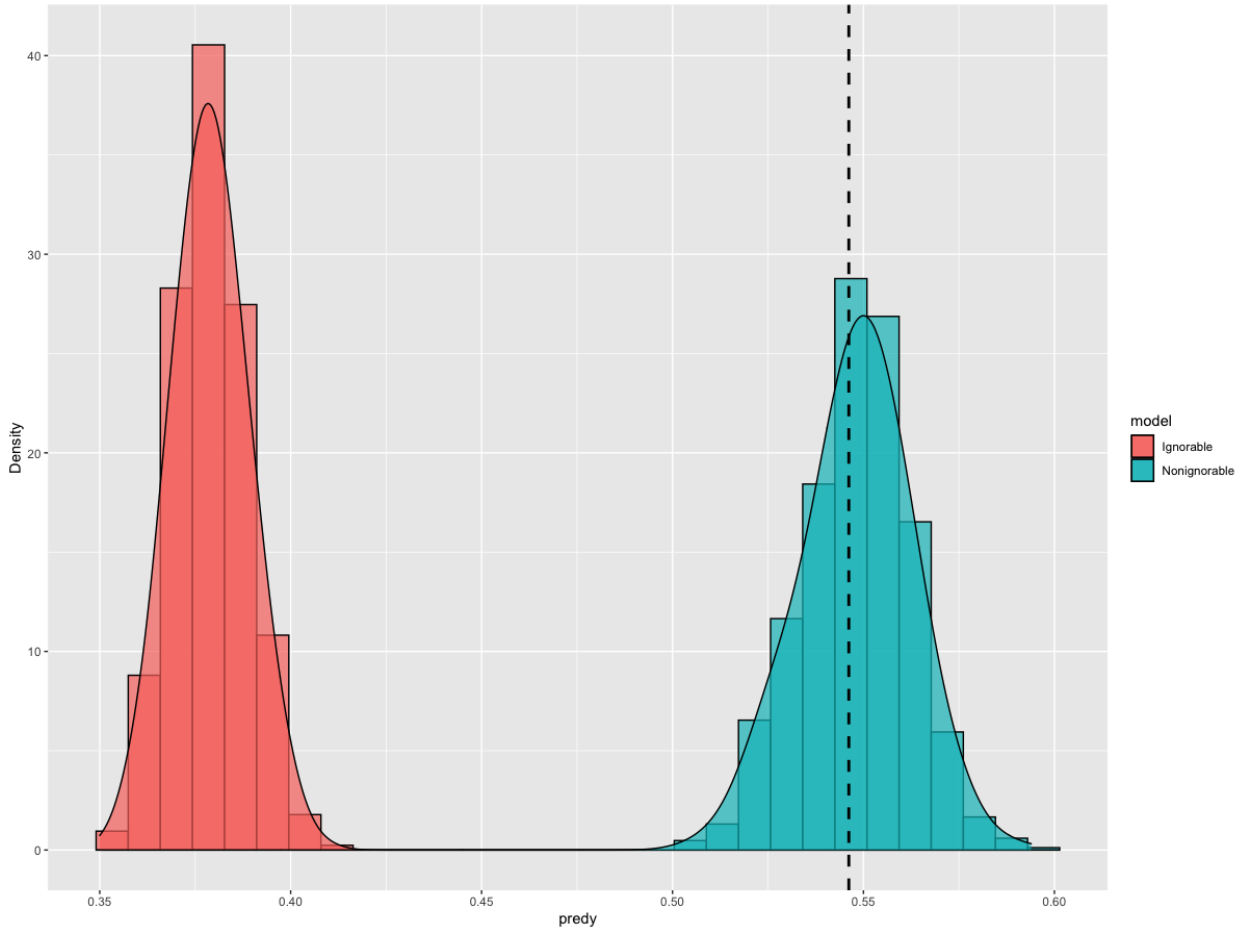


Figure 3.13: Histogram of the overall population proportion predictions by nonignorable selection model (right) and ignorable selection model (left). Dashed line: true population proportion.

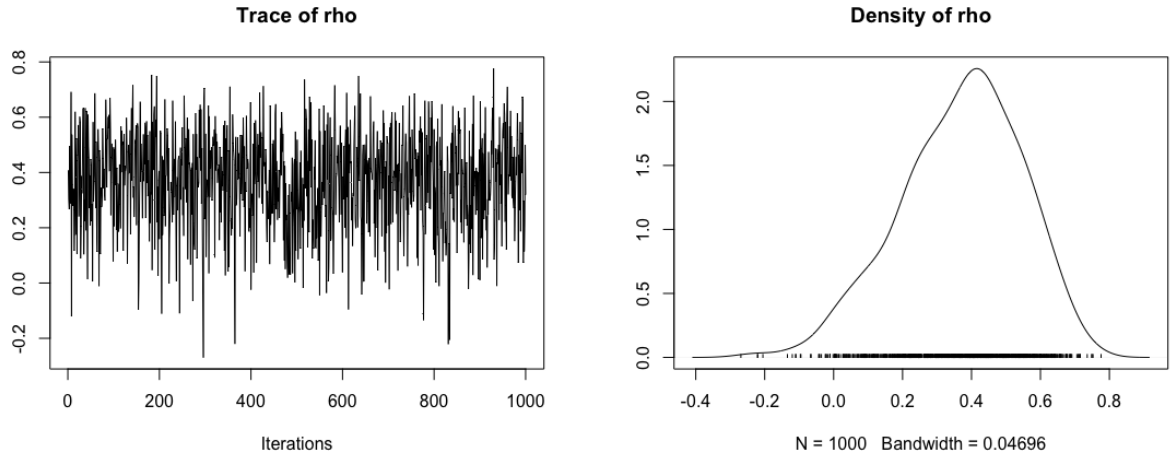


Figure 3.14: Diagnostics of ρ

3.5 Conclusion

In this chapter, we focus on the nonignorable selection model for heterogeneous data. Similar to the one in last chapter, this proposed model has two parts, the response submodel and selection submodel, which are connected by the response variable y. Differently, each of the two submodels has an area variable that accounts for heterogeneity among areas. The simulation study shows a decent population proportion prediction. Applying the model to BMI data provides us an real-world example of using heterogeneous data in prediction.

We also develop two variations of the model proposed, which are the Dirichlet process mixture model and the bivariate area effect model. The Dirichlet process prior works best when the area effects are discrete, for example, some areas share the same effect. The bivariate area effect model are designed when area effects for the response model and selection model are correlated, for example, the area effects for the same area in the response submodel and selection submodel are similar. These two models are tested by simulation studies respectively.

Chapter 4

Concluding Remarks and Future Work

This dissertation is motivated by the desire to make inference about the finite population proportion from biased samples, where the bias comes from selection not at random (SNAR), or missing not at random (MNAR). The parameter associated with the response variable in the selection model is usually deemed non-identifiable by some researchers, for example, in the paper by Wang, Bartlett, and Ryan (2017). Our contribution to this area is that by incorporating covariates, we conclude that this parameter can be identified, providing the model is correctly specified. The nonignorable selection models we proposed, along with an ignorable selection model as a comparison, are fit using both simulation data and real BMI data.

We used a single nonprobability sample to make inference about a finite population, and we incorporated appropriate covariates both for samples and nonsamples. When there are no covariates available for nonsamples, but there are population totals available from a census or administrative data, we can compute covariates using data fusion.

4.1 Concluding Remarks

In Chapter 2, we focused on the one-level Bayesian models that are designed for homogeneous data. We described the model specification for the nonignorable selection model, and proposed a MCMC algorithm for sampling the non-sampled units along with parameters from the joint posterior distribution. As a comparison, the ignorable selection model, Horvitz-Thompson estimator and Hajek estimator with weight calculated from propensity score are also provided. It was shown from simulation studies that when the missingness mechanism is MNAR, the nonignorable selection model provides decent population proportion predictions, with the 95% HPD interval covering true population proportion; when the missingness mechanism is MAR, all 95% HPD intervals from the 4 models cover the true value, but the ignorable selection model has the advantage of shortest interval range and quick execution time.

We applied both the nonignorable selection model and ignorable selection model in estimating the population proportion of children with healthy BMI in one county from NHANES III. The difficulty that arises from this real data application is the large population size (630,308) compared to a relatively small sample size (761). We addressed the problem by changing the way of presenting data while keep the model setup intact. The algorithm developed was shown to largely improve the MCMC chain's mixing performance in such a large population size within a reasonable execution time. Even though we do not have a true value to check the model's accuracy, the difference between population proportion predictions from the nonignorable selection model and the ignorable selection model indicates there is a huge selection bias in the data and the nonignorable selection model adjusts it.

We proposed a methodology to handle the case where individual covariates are not available for non-sampled units. This approach is particularly appealing because on many occasions only the total counts for each level of the discrete covariates are known using administrative data or a census. The method involves solving a large system of equations and inequalities, which could be

a dreadful experience if solved by hand. We employ a Python module to perform the symbolic calculation and pass the expressions to R for the following MCMC algorithms. The simulation study based on the BMI data indicates that finite population proportion is estimable under such conditions.

In Chapter 3, we extended the one-level Bayesian selection model to hierarchical Bayesian selection models for heterogeneous data. The hierarchical model relaxes the exchangeable requirement among individuals, and provides robustness in the model. As a trade-off, the complexity of the nonignorable selection model increases, along with the algorithm's running time. Simulation studies were conducted on both the nonignorable selection model and the ignorable selection model with results showing that the nonignorable model corrected the sampling bias overall and for each area, while the ignorable model did not. By applying the both models to the BMI data with 35 counties, the overall population proportion predictions by the two models displayed a clear distinction, which indicated a correction of the MNAR mechanism by the nonignorable model.

A variant of the model, which placed Dirichlet process (DP) priors on area effect variables was investigated. This model allowed discrete values of area effects to be sampled from the procedure, meaning different areas could share the same area effects. The difficulty came from finding the lower bounds of α, a concentration parameter that determines the number of distinct clusters, and small values of which would lead to poor mixing performance, see Nandram and Yin (2016a, 2016b). We proposed an original methodology that found a lower bound for α, which is derived from a prespecified highest possible value for the correlation of elements in the DP. The simulation study showed that, in addition to the correct prediction of finite population proportion, the number of distinct area effects can be correctly discovered as well.

Another variation is the model with a bivariate prior for area effects. This model is suitable for the scenario where for each area, area effects for the response submodel and the selection submodel are correlated. We used the bivariate normal distribution as the prior for the two area effects for each area. A good property of the bivariate normal distribution is that it could be reformulated

into two univariate normal distributions (3.4.1), which could make the algorithm easily adapted from the one in Section 3.1.1. From the simulation study, we saw that the correlation was correctly detected, as well as the finite population proportion overall and for each area.

Overall, this thesis provides novel contributions in making reference from MNAR scenarios when incorporating covariates information in the model. By treating the non-sampled units as unknown parameters and developing the MCMC algorithm, we are able to accurately predict the finite population proportions.

4.2 Future work

In future work, we aim to explore a new form of hierarchical Bayesian model, which is a model with random regression coefficients. The model structure proposed in this thesis allows only the intercept to vary among areas. A more practical model would be to allow all the regression coefficients to vary with the area (Rao and Molina, 2015). The main part of the model would be changed to

$$\text{logit}(P(y_{ij}=1)) = \boldsymbol{x}'_{ij}(\boldsymbol{\gamma}+\boldsymbol{\nu}_i), i=1,...,\ell, j=1,...,N_i,$$
$$\text{logit}(P(I_{ij}=1)) = \boldsymbol{x}'_{ij}(\boldsymbol{\beta}+\boldsymbol{\phi}_i)+\beta_{(p+1)}y_{ij}, i=1,...,\ell, j=1,...,N_i.$$

This type of modeling opens up a general framework to look at many practical problems.

First, $\boldsymbol{\nu}_i$ and $\boldsymbol{\phi}_i$ can have independent distributions as we have done for the single component (random intercept only). Second, $\boldsymbol{\nu}_i$ and $\boldsymbol{\phi}_i$ can be dependent, again as we have done here for an individual component (random intercept only). Third, we can have two Dirichlet processes independently on $\boldsymbol{\nu}_i$ and $\boldsymbol{\phi}_i$ (local pooling) or a single Dirichlet process on $(\boldsymbol{\nu}'_i, \boldsymbol{\phi}'_i)'$ (global pooling).

Another work we would like to conduct is increasing the number of repetitions in the simulation study to get a better sense of stability of the methodology. The computational burden to fit each of the models is severe, and we would like to know how to provide more efficient algorithms. In

Chapter 3 on Small Area Estimation, it would have been much better if the following three things could have been done.

- Draw data from the ignorable selection model and fit both the ignorable and nonignorable selection models. This can provide better support for the nonignorable selection model.
- Do a better job in estimating the concentration parameter of the Dirichlet process. This is still an open problem at large.
- Compare the three small area models to see how they might differ. This can be done by providing a neutral data set.

In this dissertation, we have studied a single nonprobability sample (homogeneous and heterogeneous) under nonignorability and ignorability. But it is now more useful to also add a small probability sample to help in estimation, e.g., Rao (2020); Beaumont (2020); Sakshaug et al. (2019); and Wisniowski et. al. (2020). Also, incorporating survey weights in non-probability samples has became a necessary part of data analysis, see also Chen, Li, and Wu (2019) for doubly robust inference.

Appendix A

Conditional Posterior Distribution for y_i in Section 2.1.1 (iii)

We know that Y is binary variable, which means it is from Bernoulli distribution and characterized by the success probability $P(y_i = 1 \mid ...), i = n+1, ..., N$. From (2.1.3) in Section 2.1.1, we see that

$$P(y_i = 1 \mid I_i = 0, \boldsymbol{x}_i, \boldsymbol{\gamma}, \boldsymbol{\beta}, \beta_{p+1}) = k \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{1}{1 + \exp(\boldsymbol{x}_i'\beta + \beta_{p+1})}, \tag{A.0.1}$$

where k is a constant of which the value is the product of the rest of fractions in the joint distribution. Similarly, by plugging in $y_i = 0$, we have

$$P(y_i = 0 \mid I_i = 0, \boldsymbol{x}_i, \boldsymbol{\gamma}, \boldsymbol{\beta}, \beta_{p+1}) = k \frac{1}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{1}{1 + \exp(\boldsymbol{x}_i'\beta)}, \tag{A.0.2}$$

where k is the product of the same fractions, and thus has the same value as in (A.0.1).

By using the property that $P(y_i = 1 \mid ...) + P(y_i = 0 \mid ...) = 1$ we can solve for k,

$$k = \frac{(1 + \exp(\boldsymbol{x}_i'\boldsymbol{\gamma}))\,(1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \beta_{p+1}))\,(1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}))}{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})(1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})) + 1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \beta_{p+1})}. \tag{A.0.3}$$

Putting (A.0.1) and (A.0.3) together, we arrived at

$$
\begin{aligned}
P(y_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{\gamma}, \boldsymbol{\beta}, \beta_{p+1}) &= \frac{(1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}))(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}+\beta_{p+1}))(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))}{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))+1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}+\beta_{p+1})} \\
&\quad \times \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{1}{1+\exp(\boldsymbol{x}_i'\beta+\beta_{p+1})} \\
&= \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))}{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))+1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}+\beta_{p+1})}.
\end{aligned}
$$

Appendix B

Propriety of the Posterior Density of a general case in (2.1.4)

We want to show the following density is proper

$$
\begin{aligned}
&f(\boldsymbol{y}_{ns}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{I}, \boldsymbol{y}_s) \\
&\propto \prod_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} \prod_{i=n+1}^{N} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} \\
&\times \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T (10^2 \hat{\Sigma}_{\boldsymbol{\beta}})^{-1} (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right).
\end{aligned}
\tag{B.0.1}
$$

That is, we need to show

$$
\begin{aligned}
\boldsymbol{B} = \int_{\boldsymbol{\gamma}} \int_{\boldsymbol{\beta}} \prod_{i=1}^{n} &\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} \\
&\times \prod_{i=n+1}^{N} \left\{ \sum_{y_i=0}^{1} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} \right\} \\
&\times \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T (10^2 \hat{\Sigma})^{-1} (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right) d\boldsymbol{\beta} d\boldsymbol{\gamma} < \infty.
\end{aligned}
$$

We proceed by the following three steps.

First, let us focus on the second row. It is worth noting that no matter what value y_i has, for $i = n+1, ..., N$, the term

$$\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} \leq 1.$$

Thus, we have the inequality

$$\prod_{i=n+1}^{N} \left\{ \sum_{y_i=0}^{1} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} \right\} < 2^{N-n} < \infty.$$

Replaced the term in B by the inequality, our goal is to prove $\boldsymbol{B}^* < \infty$, where

$$\begin{aligned}
\boldsymbol{B}^* &= \int_{\boldsymbol{\gamma}} \int_{\boldsymbol{\beta}} \prod_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (10^2 \hat{\Sigma})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right) d\boldsymbol{\beta} d\boldsymbol{\gamma} \\
&= \int_{\boldsymbol{\beta}} \prod_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (10^2 \hat{\Sigma})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right) d\boldsymbol{\beta} \\
&\times \int_{\boldsymbol{\gamma}} \prod_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} d\boldsymbol{\gamma}.
\end{aligned}$$

In the next two steps we are going to prove the $\int_{\boldsymbol{\beta}} d\boldsymbol{\beta}$ and $\int_{\boldsymbol{\gamma}} d\boldsymbol{\gamma}$ are both finite. We deal with $\boldsymbol{\beta}$ first.

$$\begin{aligned}
&\int_{\boldsymbol{\beta}} \prod_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (10^2 \hat{\Sigma})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right) d\boldsymbol{\beta} \\
&< \int_{\boldsymbol{\beta}} \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (10^2 \hat{\Sigma})^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right) d\boldsymbol{\beta} \quad \left(\text{since } \prod_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_{(p)} + \beta_{p+1} y_i)} < 1 \right) \\
&< \infty \quad (\text{since } N(\hat{\boldsymbol{\beta}}, 10^2 \hat{\Sigma}) \text{ is proper}).
\end{aligned}$$

Third, we need to show that

$$\int_{\boldsymbol{\gamma}} \prod_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} d\boldsymbol{\gamma} < \infty.$$

We are going to show that

$$\prod_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma} y_i)}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})} \tag{B.0.2}$$

is log-concave. A log-concave density has sub-exponential tails, see Borell (1983). This means any log-concave density is proper and the moment generating function exists.

Denote the log of (B.0.1) by $\Delta(\boldsymbol{\gamma})$, the gradient vector by $G(\boldsymbol{\gamma})$, and the Hessian matrix by $H(\boldsymbol{\gamma})$. To show log concavity, we need to show that $-H(\boldsymbol{\gamma})$ is positive definite (Dharmadhikari and Joag-Dev, 1988), where

$$\begin{aligned}
\Delta(\boldsymbol{\gamma}) &= \left(\sum_{i=1}^{n} \boldsymbol{x}_i' y_i\right)\boldsymbol{\gamma} - \sum_{i=1}^{n} \log(1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})), \\
G(\boldsymbol{\gamma}) &= \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}, \\
\text{and } H(\boldsymbol{\gamma}) &= -\sum_{i=1}^{n} \frac{\boldsymbol{x}_i\boldsymbol{x}_i' \exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{(1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}))^2}.
\end{aligned}$$

First note that if y_i are not all $\mathbf{0}$'s or $\mathbf{1}$'s, $G(\boldsymbol{\gamma}) = 0$ has at least one solution.

Then assuming $\boldsymbol{X}$ is full rank, and $-H(\boldsymbol{\gamma})$ is positive definite, because for any $\boldsymbol{a} \neq 0$, $-\boldsymbol{a}'H(\boldsymbol{\gamma})\boldsymbol{a} = \sum_{i=1}^{n} (\boldsymbol{a}'\boldsymbol{x}_i)^2 \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\gamma})}{(1+\exp(\boldsymbol{x}_i'\boldsymbol{\gamma}))^2} > 0$.

So we have established (B.1) is log-concave and thus the joint posterior distribution is proper.

Appendix C

Bounds of free variables derived from (2.5.2)

In Section 2.5, we propose an algorithm of sampling the 10 free variables from their respective truncated multinomial distribution. Here, we explain how the bounds for each of those free variables are derived. We start from the constraints in (2.5.2), which are

$$N_1 = -A_2 - A_3 - A_4 + N + N_{10} + N_{12} + N_{14} + N_{16} - N_3 + N_6 + N_8 - S_1$$

$$N_2 = -N_{10} - N_{12} - N_{14} - N_{16} - N_4 - N_6 - N_8 + S_1$$

$$N_5 = A_2 + N_{11} + N_{12} + N_{15} + N_{16} + N_3 + N_4 - N_6 - R_1$$

$$N_7 = -N_{11} - N_{12} - N_{15} - N_{16} - N_3 - N_4 - N_8 + R_1$$

$$N_9 = A_3 - N_{10} - N_{11} - N_{12}$$

$$N_{13} = A_4 - N_{14} - N_{15} - N_{16}$$

The free variables are $N_3, N_4, N_6, N_8, N_{10}, N_{11}, N_{12}, N_{14}, N_{15}, N_{16}$.

The bounds for each of the free variables is derived by making the right hand side of each and every constraint greater than 0. For example, by using $N_1 \geq 0$ we have the inequality$-A_2 - A_3 -$

$A_4 + N + N_{10} + N_{12} + N_{14} + N_{16} - N_3 + N_6 + N_8 - S_1 \geq 0$, from which we can get one upper bound for N_3, which is $N_3 \leq -A_2 - A_3 - A_4 + N + N_{10} + N_{12} + N_{14} + N_{16} + N_6 + N_8 - S_1$; so on and so forth for variables $N_{10}, N_{12}, N_{14}, N_{16}, N_6$ and N_8 in this constraint. This step can be implemented by Sympy as well, the output of which can then be passed to R environment. The constraints for each free variable are

Variable	Constraint
N_3	$N_3 < -N_{11} - N_{12} - N_{15} - N_{16} - N_4 - N_8 + R_1$
	$N_3 > -A_2 - N_{11} - N_{12} - N_{15} - N_{16} - N_4 + N_6 + R_1$
	$N_3 < -A_2 - A_3 - A_4 + N + N_{10} + N_{12} + N_{14} + N_{16} + N_6 + N_8 - S_1$
	$N_3 \geq 0$
N_4	$N_4 < -N_{10} - N_{12} - N_{14} - N_{16} - N_6 - N_8 + S_1$
	$N_4 < -N_{11} - N_{12} - N_{15} - N_{16} - N_3 - N_8 + R_1$
	$N_4 > -A_2 - N_{11} - N_{12} - N_{15} - N_{16} - N_3 + N_6 + R_1$
	$N_4 \geq 0$
N_6	$N_6 < -N_{10} - N_{12} - N_{14} - N_{16} - N_4 - N_8 + S_1$
	$N_6 < A_2 + N_{11} + N_{12} + N_{15} + N_{16} + N_3 + N_4 - R_1$
	$N_6 > A_2 + A_3 + A_4 - N - N_{10} - N_{12} - N_{14} - N_{16} + N_3 - N_8 + S_1$
	$N_6 \geq 0$
N_8	$N_8 < -N_{10} - N_{12} - N_{14} - N_{16} - N_4 - N_6 + S_1$
	$N_8 < -N_{11} - N_{12} - N_{15} - N_{16} - N_3 - N_4 + R_1$
	$N_8 > A_2 + A_3 + A_4 - N - N_{10} - N_{12} - N_{14} - N_{16} + N_3 - N_6 + S_1$
	$N_8 \geq 0$
N_{10}	$N_{10} < -N_{12} - N_{14} - N_{16} - N_4 - N_6 - N_8 + S_1$
	$N_{10} < A_3 - N_{11} - N_{12}$
	$N_{10} > A_2 + A_3 + A_4 - N - N_{12} - N_{14} - N_{16} + N_3 - N_6 - N_8 + S_1$

	$N_{10} \geq 0$
N_{11}	$N_{11} < -N_{12} - N_{15} - N_{16} - N_3 - N_4 - N_8 + R_1$
	$N_{11} < A_3 - N_{10} - N_{12}$
	$N_{11} > -A_2 - N_{12} - N_{15} - N_{16} - N_3 - N_4 + N_6 + R_1$
	$N_{11} \geq 0$
N_{12}	$N_{12} < -N_{10} - N_{14} - N_{16} - N_4 - N_6 - N_8 + S_1$
	$N_{12} < -N_{11} - N_{15} - N_{16} - N_3 - N_4 - N_8 + R_1$
	$N_{12} < A_3 - N_{10} - N_{11}$
	$N_{12} > -A_2 - N_{11} - N_{15} - N_{16} - N_3 - N_4 + N_6 + R_1$
	$N_{12} > A_2 + A_3 + A_4 - N - N_{10} - N_{14} - N_{16} + N_3 - N_6 - N_8 + S_1$
	$N_{12} \geq 0$
N_{14}	$N_{14} < -N_{10} - N_{12} - N_{16} - N_4 - N_6 - N_8 + S_1$
	$N_{14} < A_4 - N_{15} - N_{16}$
	$N_{14} > A_2 + A_3 + A_4 - N - N_{10} - N_{12} - N_{16} + N_3 - N_6 - N_8 + S_1$
	$N_{14} \geq 0$
N_{15}	$N_{15} < -N_{11} - N_{12} - N_{16} - N_3 - N_4 - N_8 + R_1$
	$N_{15} < A_4 - N_{14} - N_{16}$
	$N_{15} > -A_2 - N_{11} - N_{12} - N_{16} - N_3 - N_4 + N_6 + R_1$
	$N_{15} \geq 0$
N_{16}	$N_{16} < -N_{10} - N_{12} - N_{14} - N_4 - N_6 - N_8 + S_1$
	$N_{16} < -N_{11} - N_{12} - N_{15} - N_3 - N_4 - N_8 + R_1$
	$N_{16} < A_4 - N_{14} - N_{15}$
	$N_{16} > -A_2 - N_{11} - N_{12} - N_{15} - N_3 - N_4 + N_6 + R_1$
	$N_{16} > A_2 + A_3 + A_4 - N - N_{10} - N_{12} - N_{14} + N_3 - N_6 - N_8 + S_1$
	$N_{16} \geq 0$

Appendix D

A preliminary approach to draw α

D.1 Posterior for α

From Charles E. Antoniak's paper (1974), we know

$$P(K = k) = \frac{{}_n a_k \, \alpha^k \Gamma(\alpha)}{\Gamma(\alpha + n)},$$

where ${}_n a_k$ is the absolute value of Stirling number of the first kind.

Applying the prior for α, we have the posterior density as

$$\pi(\alpha|k) \propto \frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha + n)(\alpha + 1)^2}, \ \alpha \in (0, \infty). \tag{D.1.1}$$

A straightforward approach to sample α is the grid sampling method. We first map the parameter range: $\alpha \in (0, \infty)$ to $\phi \in (0, 1)$ by the transformation $\phi = \dfrac{1}{\alpha + 1}$. Then, we get the density of ϕ as follows,

$$\pi(\phi|k) \propto \frac{(\phi^{-1} - 1)^k \Gamma(\phi^{-1} - 1)}{\Gamma(\phi^{-1} - 1 + n)}, \ \phi \in (0, 1). \tag{D.1.2}$$

Thus, ϕ can be easily sampled by applying the grid sampling method. Correspondingly, α is computed by $\alpha = \dfrac{1}{\phi} - 1$.

However, in reality, this approach is problematic as it allows α to run from 0 to infinity. While in the limit of $\alpha \to 0$, the ν_j's are all concentrated at a single value, which makes it impossible for the algorithm to evaluate the lower level parameters, and thus produce the non-stationary chains. We provide a preliminary solution for this problem.

The idea is to introduce a lower bound of α that restrict it from reaching 0. The lower bound is derived from the correlation of any two elements in the Dirichlet process, which is $\frac{1}{1+\alpha}$. We do not want this correlation to be too large, otherwise we'll end up with drawing from the same urn over and over again (Polya Urn Scheme). Thus, we bound the correlation by the tuning parameter a, where a can be determined by try and error to ensure the single value occasion will not occurred at any iteration. We use $a = 0.5$ in our simulation. Note that

$$\frac{1}{1+\alpha} \leq a \Rightarrow \alpha \geq \frac{1-a}{a}.$$

Before we introduce the algorithm, we first derive some quantities of $\pi(\alpha|n,k)$, which will be used in the algorithm.

- Mode of $\pi(\alpha|n,k)$.

$$\begin{aligned}
\pi(\alpha|n,k) &\propto \frac{\alpha^k \Gamma(\alpha)}{\Gamma(\alpha+n)(\alpha+1)^2},\ \alpha \in (0,\infty) \\
&= \frac{\alpha^k}{\prod_{s=0}^{n-1}(\alpha+s)(\alpha+1)^2}, \\
log\pi(\alpha|n,k) &= k\, log\alpha - \sum_{s=0}^{n-1} log(\alpha+s) - 2log(\alpha+1) + c, \qquad \text{(D.1.3)} \\
\frac{\partial\, log\pi(\alpha|n,k)}{\partial\alpha} &= \frac{k}{\alpha} - \sum_{s=0}^{n-1}\frac{1}{\alpha+s} - \frac{2}{\alpha+1} = 0, \\
&\Rightarrow \alpha = \frac{k}{\frac{2}{\alpha+1} + \sum_{s=0}^{n-1}\frac{1}{\alpha+s}}.
\end{aligned}$$

The mode α^0 can be obtained by the fixed point iteration method.

- Variance and standard deviation (SD) of $\phi = \log \alpha$.

  We first transform α to ϕ by $\phi = \log \alpha$. Owing to the logconcavity of $\pi(\phi|n,k)$, we can obtain the approximated variance by the negative inverse of the Hessian matrix of $\pi(\phi|n,k)$ evaluated at mode as follows,

$$
\begin{aligned}
\pi(\phi|n,k) &\propto \frac{e^{\phi k}e^{\phi}}{\prod_{s=0}^{n-1}(e^{\phi}+s)(e^{\phi}+1)^2}, \quad \phi \in (-\infty,\infty), \\
log\pi(\phi|n,k) &= \phi(k+1) - \sum_{s=0}^{n-1} log(e^{\phi}+s) - 2log(e^{\phi}+1) + c, \\
\frac{\partial\, log\pi(\phi|n,k)}{\partial\phi} &= k+1 - \sum_{s=0}^{n-1}\frac{e^{\phi}}{e^{\phi}+s} - \frac{2e^{\phi}}{e^{\phi}+1}, \\
\frac{\partial^2\, log\pi(\phi|n,k)}{\partial\phi^2} &= -e^{\phi}\left(\frac{2}{(1+e^{\phi})^2} + \sum_{s=0}^{n-1}\frac{s}{(s+e^{\phi})^2}\right), \\
\widehat{var(\phi)} &= \frac{1}{\alpha^0\left(\frac{2}{(1+\alpha^0)^2} + \sum_{s=0}^{n-1}\frac{s}{(s+\alpha^0)^2}\right)}, \\
\widehat{SD(\phi)} &= \sqrt{\frac{1}{\alpha^0\left(\frac{2}{(1+\alpha^0)^2} + \sum_{s=0}^{n-1}\frac{s}{(s+\alpha^0)^2}\right)}}.
\end{aligned}
\tag{D.1.4}
$$

- Algorithm

  1. Obtain the mode α^0 and use it to compute $\widehat{SD(\phi)}$.

  2. Apply the grid sampling method to draw ϕ,
  where ϕ is defined on $\left(max\left(log\left(\frac{1-a}{a}\right), \phi - 10 * \widehat{SD(\phi)}\right), \phi + 10 * \widehat{SD(\phi)}\right)$.

  3. Transform ϕ back to α to obtain one sample of α.

Bibliography

Andridge R. R., and Little, R. J. A. (2011), "Proxy pattern-mixture analysis for survey nonresponse," *Journal of Official Statistics*, 27 (2), 153–180.

Andridge R. R., West, B. T., Little, R. J. A. Boonstra, P. S. and Leiton, F. A. (2019), "Indices of non-ignorable selection bias for proportions estimated from non-probability samples," *Applied Statistics*, 68 (5), 1465–1483.

Antoniak, C. E. (1974), "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, 2 (6), 1152–1174.

Austin, P. C. and Stuart E. A. (2015), "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies," *Statistics in Medicine*, 34, 3661–3679.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., Tourangeau, R. (2013), "Summary report of the AAPOR task force on Non-probability sampling," *Journal of Survey Statistics and Methodology*, 1, 90–143.

Bang, H. and Robin, J. M. (2005), "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61, 962–972.

Beaumont, J. (2020), "Are probability surveys bound to disappear for the production of official statistics," *Survey Methodology*, 46 (1), 1–28.

Borell, C. (1983), "Convexity of measures in certain convex cones in vector space σ-algebras," *Mathematica Scandinavica*, 53, 125–144.

Chen, Y., Li, P. and Wu, C.(2019), "Doubly robust inference with nonprobability survey samples," *Journal of the American Statistical Association*, 0, 1–11.

Chapin, F. S. (1947), *Experimental designs in sociological research*

Choi, S. M., and Kim, D. H. (2014), "Sensitivity analysis in Bayesian nonignorable selection model for binary responses," *Journal of the Korean Data & Information Science Society 2014*, 25, 187–194.

Couper, M. P. (2007), "Issues of representation in eHealth research with a focus on web surveys," *American Journal of Preventive Medicine*, 32, S83–S89.

Dharmadhikari, S. W., and Joag-Dev, K. (1988), *Unimodality, convexity, and appplication*, CA: Academic Press.

Diamond, S. S. (2000), "Reference guide on survey research." *Reference Manual on Scientific Evidence 2nd Edition*.

DiSogra, C., Cobb, C., Chan, E., and Dennis J. (2012) "Using rrobability-based online samples to calibrate non-probability opt-in samples." *Presentation at: 67th Annual Conference of the American Association for Public Opinion Research (AAPOR)*.

Doll, R. and Hill, A.B. (1964) "Mortality in relation to smoking: ten years' observations of British doctors," *Br Med*, 1, 1399–1410.

Elliott, M. R. and Valliant, R. (2017) "Inference for nonprobability samples" *Statistical Science*, 32(2), 249–264.

Escobar, M. D. (1988) "Estimating the means of several normal populations by nonparametric estimation of the distribution of the means" *Ph.D. dissertation, Yale University, Dept. of Statistics*.

Escobar, M. D., West, M. (1995) "Bayesian density estimation and inference using mixtures" *Journal of the American Statistical Association*, 90(430), 577–588.

Ferguson, T. S. (1973) "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, 1(2), 209–230.

Firth, F. A., Hellewell, J., Klepac, P., Kissler, S., CMMID COVID-19 Working Group, Kucharski, A. J., and Spurgin, L. G. (2020) "Using a real-world network to model localized COVID-19 control strategies," *Nature Medicine* .

Foster, J. J. and Smith, P. W. F. (1998) "Model-based inference for categorical survey data subject to non-ignorable non-response," *Journal of the Royal Statistical Society*, 60(1), 57–70.

Gelman A. (2007) "Struggles with survey weighting and regression modeling" *Statistical Science*, 22 (2).

Gelman A. and Little, T. (1997) "Poststratification into many categories using hierarchical logistic regression," *Survey Methodology*.

Gong, L., and Flegal, J. M. (2016) "A practical sequential stopping rule for high-dimensional Markov Chain Monte Carlo" *Journal of Computational and Graphical Statistics*, 25 (3): 684–700.

Goodman L. A. (1961) "Snowball sampling" *Annals of Mathematical Statistics*, 32: 148–70

Greenlees, J.S., Reece, W.S. and Zieschang, K.D. (1982) "Imputation of missing values when the probability of response depends on the variable being imputed, " *Journal of the American Statistical Association*, 77: 251–261.

Heckman, J. J. (1976), " The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models," *Annals of Economic and Social Measurement*, 5, 475–492.

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

Kalli, M., Griffin, J. E. and Walker, S. G. (2011), "Slice sampling mixture models" *Stat Comput*, 21, 93–105.

Klovdahl, A. S., John, J. Potterat, D. E., Woodhouse, J. B., Muth, S. Q., Muth, and William, W. D. (1994), "Social networks and infectious disease: the Colorado Springs study," *Social Science & Medicine*, 38 (1), 79–88.

Little, R. J. A. (1993), "Post-stratification: a modeler's perspective," *Journal of the American Statistical Association*, 88, 1001–1012.

Little, R. J. A. and Rubin (2019), *Statistical analysis with missing data*, John Wiley & Sons, New York.

Little, R. J. A., West B. T., Boonstra, P. S., and Hu J. (2019), "Measures of the degree of departure from ignorable sample selection," *Journal of Survey Statistics and Methodology*, 0, 1-33.

Mercer, A. W., Kreuter, F., Keeterr, S. and Stuart, E. A. (2017), "Theory and practice in nonprobability surveys – parallels between causal inference and survey inference," *Public Opinion Quarterly*, 81, 250–279.

Maceachern, S. N. (1994), "Estimating normal means with a conjugate style Dirichlet process prior," *Communication in Statistics: Simulation and Computation*, 23, 727–741.

Maceachern, S. N., and Muller, P. (1998), "Estimating mixture of Dirichlet process models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

Nandram, B., Cao, H., Xu, Z., and Bhadra, D. (2020), "Bayesian predictive inference for non-probability samples with spatial poststratification."

Nandram, B., Chen, L., Fu, S., and Manandhar, B. (2018), "Bayesian logistic regression for small areas with numerous households," *Statistics and Application*, 16, 171–205.

Nandram, B., and Choi, J. W. (2002a), "A Bayesian analysis of a proportion under non-ignorable non-response" *Statistics in Medicine*, 21, 1189–1212.

Nandram, B., and Choi, J. W. (2002b), "Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability," *Journal of the American Statistical Association*, 97(458), 381–388.

Nandram, B., and Choi, J. W. (2004), "Nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse," *Journal of Nonparametric Statistics*, 16(6), 821–839.

Nandram, B., and Choi, J. W. (2005), "Hierarchical Bayesian nonignorable nonresponse regression models for small areas: an application to the NHANES data," *Survey Methodology*, 31, 73–84.

Nandram, B., and Choi, J. W. (2010), "A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection" *Journal of the American Statistical Association*, 105, 120–135.

Nandram, B., Bhatta, D. Bhadra, D. Shen, G. (2013), "Bayesian predictive inference of a finite population proportion under selection bias" *Statistical Methodology*, 11, 1–21.

Nandram, B., and Yin, J. (2016a), "Bayesian predictive inference under a Dirichlet process with sensitivity to the normal baseline" *Statistical Methodology*, 28, 1–17.

Nandram, B., and Yin, J. (2016b), "A nonparametric Bayesian prediction interval for a finite population mean" *Journal of Statistical Computation and Simulation*, 86 (16), 3141–3157.

Pfeffermann, D., Moura, F. A. D. S., and Silva, P. L. D. N. (2006), "Multi-level modelling under informative sampling" *Biometrika*, 93, 943–959.

Presser, S. (1984), "Is inaccuracy on factual survey items item-specific or respondent-specific?" *Public Opinion Quarterly*, 48, 344–355.

Rao, J. N. K. (2020), "On making valid inferences by integrating data from surveys and other sources," *The Indian Journal of Statistics*.

Rao, J.N.K. and Molina, I. (2015), *Small Area Estimation*, NJ: Wiley.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985), "The bias due to incomplete matching," *Biometrika*, 41, 103–116.

Rubin, D. B. (1973), "Matching to remove bias in observational studies," *Biometrika*, 29, 159–184.

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

Rubin, D. B. (1977), "Assignment of treatment group on the basis of a covariate," *Journal of Educational Statistics*, 2, 1–26.

Sakshaug, J. W., Wisniowski, A., Ruiz, D. A. P., and Blom, A. G. (2019), "Supplementing small probability samples with nonprobability samples: a Bayesian approach," *Journal of Official Statistics*, 35(3), 653–681.

Sarndal, C. E., Swensson, B. and Wretman, J. (2003), *Model Assisted Survey Sampling*, NY: Springer.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999), "Adjusting for non-ignorable drop-out using semiparametric non-response models," *Journal of the American Statistical Association*, 94, 1096–1120.

Smith, T. M. F. (1983), "On the validity of Inferences from Non-random Samples," *Journal of the Royal Statistical Society*, 146(4), 394–403.

Stuart, E. A. (2010), "Matching methods for causal inference: a review and a look forward (2010)," *Statistical Science*, 25(1), 1–21.

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

Sverchkov, M., and Pfeffermann, D. (2018), "Small area estimation under informative sampling and not missing at random non-response," *Journal of the Royal Statistical Society*, 181(4), 981–1008.

Valliant, R. and Dever, J. A. (2011), "Estimating propensity adjustments for volunteer web surveys," *Sociological Methods & Research* , 40(1), 105–137.

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000), *Finite Population Sampling and Inference: A Prediction Approach.*

Vats, D., and Knudson, C. (2018), "Revisiting the Gelman-Rubin diagnostic," *arXiv:1812.09384*.

Vats, D., Flegal, J. M., and, Jones, G. L. (2015), "Multivariate output analysis for Markov chain Monte Carlo," *arXiv preprint arXiv:1512.07713*.

Walker, S. G. (2011), "Slice sampling mixture models," *Stat Comput*, 21, 93–105.

Walker, S. G. (2007), "Sampling the Dirichlet mixture model with slices," *Commun. Stat., Simul. Comput.*, 36, 45–54.

Wang, J.J.J., Bartlett, M. and Ryan, L. (2017), "Non-ignorable missingness in logistic regression," *Statistics in Medicine*, 36, 3005–3021.

Welch, S. (1975), "Sampling by referral in a dispersed population," *Public Opinion Quarterly*, 39, 237–45.

Wisniowski, A., Sakshaug, J., Ruiz, D. A. P., and Blom, A. G. (2020), "Intergrating probability and nonprobability samples for survey inference," *Journal of Survey Statistics and Methodology*, 8, 120–147.

Woo, N., Nandram, B., and Kim, D. (2018), "Bayesian small area models for three-way contingency tables with nonignorability," *Statistica Sinica*, 28, 1839–1866.

Xu, Z., Nandram, B., and Manandhar, B. (2020), "Bayesian inference of a finite population mean under length-biased sampling," in *Statistical Methods and Applications in Forestry and Environmental Sciences, Chapter 6*, Singapore: Springer Nature.

Xu, Z. and Nandram, B. (2019), "Bayesian inference of non-probability samples," *JSM Proceedings, 2585–2593*, Foundations in Bayesian Statistics Section. Alexandria, VA: American Statistical Association.

Xu, Z. and Nandram, B. (2020), "Bayesian inference for non-probability samples with binary responses" .

Yin, J. and Nandram, B. (2020a), "A Bayesian small area model with Dirichlet processes on the responses," *Statistics in Transition*.

Yin, J. and Nandram, B. (2020b), "A nonparametric Bayesian analysis of response data with gaps, ourliers and ties," *Statistics and Applications*, 18 (1), 1-21.