International Journal of Statistics and Probability; Vol. 13, No. 2; June 2024
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Supplementing a Non-probability Sample With a Probability Sample to Predict the Finite Population Mean

Zihang Xu[1], Balgobin Nandram[2]

[1] Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA, USA

[2] Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA, USA

Correspondence: Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA, USA. E-mail: balnan@wpi.edu

Received: January 23, 2024 Accepted: April 7, 2024 Online Published: May 29, 2024

doi:10.5539/ijsp.v13n2p16 URL: https://doi.org/10.5539/ijsp.v13n2p16

Abstract

We show how to analyze a non-probability sample (nps) with limited information from a small probability sample (ps). The most practical case is when the nps has auxiliary variables and study variable but no survey weights and the ps has known weights, auxiliary variables, but no study variable. Two samples are taken from the same population and the variables are common to both the nps and the ps. A large non-probability sample can reduce the cost but will give biased estimator with small variance, the small probability sample can provide supplemental information. Following this, we apply these weights to fit a mixture model, enhancing the robustness of the results and enabling the estimation of the finite population mean. Additionally, we present a method to enhance the efficiency of the Gibbs sampler.

Keywords: adjusted survey weight, Gibbs sampling, logistic regression, missing data, propensity score, robust model

1. Introduction

Consider a finite population of size N with covariates $x_i, i = 1, ..., N$. The objective is to calculate the population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. A common application is to estimate the population mean value of BMI (Body Mass Index), $\bar{Y}$, to gain a direct view of the health conditions of adults older than 20 years. The aim is to achieve a BMI value within the range 20-25, indicating better health. However, practical challenges arise due to the potentially large value of N or the unknown y_i for all units. To address this, we utilize the samples to estimate the population mean.

Probability samples with known weights are an excellent choice for accurate estimation. Probability sampling methods have been widely accepted since they were first introduced in the 1950s. However, as the required sample size increases over time, the cost of conducting a large probability sample has risen rapidly. A significant challenge arises in balancing the cost and potential bias in results. It is well-known that insufficient samples can introduce bias to inference results. To address this issue, statisticians explore alternative methods that strike a balance.

One such alternative is the consideration of non-probability samples, a concept that has been well-explored in numerous recent studies (Marella, 2023, Salvatore, Biffignandi, Sakshaug, Wiśniowski, & Struminskaya, 2023, Nandram & Rao, 2024). The primary advantage of non-probability sampling lies in its cost-effectiveness. However, the drawback is evident: It may not adequately represent the entire population, leading to biased inference. The central problem now becomes how to strike a balance between the cost-effectiveness of probability samples and the potential bias introduced by non-probability samples.

One method to address this issue is to utilize a large-size non-probability sample (nps) in conjunction with a small-size probability sample (ps). Both samples are drawn from the same population and share common auxiliary variables. The non-probability sample, denoted as nps, includes the observed study variable y without weights. Meanwhile, the probability sample, denoted as ps, has the weights but lacks the study variable y. To implement this method effectively, unknown propensity scores or selection probabilities for the non-probability sample need to be estimated. Chen, Li & Wu (2020) employed non-probability sample, S_1, and a probability sample, S_2, to fit logistic regression with Horvitz-Thompson (HT) estimator, estimating the propensity scores under the ignorable assumption; the selection probabilities $p(R_i = 1|x_i, y_i) = p(R_i = 1|x_i)$, $i = 1, ..., N$, where R_i is an indicator variable for unit i being included in the sample .This assumption is related to missing at random (MAR) defined by Rubin (1976), Little and Rubin (2002) which is widely used in general missing data studies (Nandram & Choi, 2002a, 2002b, Nandram & Choi, 2006, Nandram & Choi, 2010, Nandram & Rao, 2021, Nandram, Choi, & Liu, 2021).

This article introduces a Bayesian approach to the method of Chen et al., (2020) under the ignorable assumption. Fur-

thermore, we extend this method to the non-ignorable assumption, which corresponds to missing not at random (MNAR) defined by Rubin (1976): selection probability $p(R_i = 1|\underset{\sim}{x_i}, y_i)$ is related to our study variable. This assumption is also widely used in many existing work (Nandram & Choi, 2010, Nandram, Bhatta & Shen, 2013).

Given the absence of the y value in the ps sample, a straightforward imputation method is to use classification and regression tree (CART). CART, introduced by Breiman, Friedman, Olshen, & Stone (1984) is a well-known and effective method for addressing challenges associated with missing data problems. Another notable Bayesian CART model is BART, proposed by Chipman, George, & McCulloch (1998) and Chipman, George, & McCulloch (2010). Given the substantial size of our non-probability sample and the smaller size of the probability sample, CART proves to be particularly advantageous, leveraging the information within the non-probability sample effectively. More details about CART are given in Appendix C. In addition to utilizing the link function to estimate weights, we introduce a method using Euclidean distance to impute survey weights after imputing y values. This approach is referred to as the double imputation method. Further details will be provided in the Section 2.

Another method in this article involves running a logistic regression on the combined sample of nps and ps, with the non-probability sample (nps) designated as the selected sample. Unlike the previous approach, this method does not involve running the model on the entire population. We can determine the regression coefficients θ for the model and subsequently calculate the propensity scores for all units within our combined sample. Given that the ps sample has known weights, we can then impute the weights for the nps sample by comparing the calculated propensity scores. This method can be easily implemented under the ignorable assumption, but it introduces a challenge related to the missing y problem in the ps sample under the non-ignorable assumption, we can use the classification and regression tree to impute the y values or we can use the other method. Since we will employ the Gibbs sampler to find the coefficient values, we will also leverage the Gibbs sampler to generate samples from the posterior distributions of the missing y given the informative prior to do the imputation. More details will be given in the Section 2.

After obtaining the weights $W_i, i = 1, ..., n$, a common approach to infer the study variable is using the adjusted weights model. The adjusted weights model, an extension of the standard weighted regression model, involves the replacement of weights with adjusted weights, Nandram, Choi, & Liu (2021) used the similar way for solving the non-probability problem. Potthoff, Woodbury, & Manton (1992) obtained the adjusted weights w_i corresponding to the original weights W_i,

$$w_i = \frac{\hat{n} W_i}{\sum_{i=1}^{n} W_i}, i = 1, ..., n. \tag{1}$$

Here, $\hat{n}$, called effective sample size (ESS), is calculated using the formula,

$$\hat{n} = \frac{(\sum_{i=1}^{n} W_i)^2}{\sum_{i=1}^{n} W_i^2}, i = 1, ..., n. \tag{2}$$

We consider a independence population where $y_i \sim Normal(\mu_i, \sigma^2), i = 1, ..., N$. We know that the population mean will follow $\bar{Y} \sim Normal(\bar{\mu}, \sigma^2/N)$. Under these assumptions, we can use the data y_i with adjusted weights w_i to estimate the values of μ_i and σ^2 using $y_i \sim Normal(\mu_i, \sigma^2/w_i)$. We can estimate the weights for our non-probability sample based on the previous methods. Subsequently, we can use these weights with the non-probability sample to conduct inferences about study variables.

Utilizing the adjusted weights to fit the model often results in a more reasonable variance when compared to the original weights model. Our primary focus will be on using the non-probability sample with estimated weights to fit the adjusted weights model. Additionally, we will leverage the probability sample to estimate the population size, σ^2 and $\bar{\mu}$, enhancing the reliability of our model.

The adjusted weights model discussed previously might encounter stability issues in real-world applications. There may be some outliers or sub-population mixed with the population, the single model can work well for most parts of the population but cannot solve the extreme parts. To make our model more robust we can consider using the mixture model or we call it robust regression. We assume that the residuals follow a mixture of normal distributions:

$$e_i \sim (1-p)\text{Normal}(0, \sigma_1^2) + p\text{Normal}(0, \sigma_2^2), i = 1, ..., n \tag{3}$$

That is residuals have probability $1 - p$ of coming from a normal distribution with variance σ_1^2 and probability p of coming from a normal distribution with variance σ_2^2, where p should be smaller than 0.5 and σ_1^2 is smaller than σ_2^2. Now explore a similar concept by considering a population denoted as $y_i \sim (1-p)\text{Normal}(\mu_i, \sigma_1^2) + p\text{Normal}(\mu_i, \sigma_2^2)$, where $i = 1, ..., N$, we utilize the data $y_i, i = 1, ..., n$, with known weights to fit the robust adjusted weights model. This model estimates p, σ_1^2, and σ_2^2. Then, we can make inference about the finite population mean. Chakraborty, Datta, & Mandal (2019) and Goyal, Datta, & Mandal (2020) use the similar idea to fit the model. In our article, we model μ_i using auxiliary

variables, specifically $\mu_i = \underset{\sim}{x}_i'\underset{\sim}{\phi}$, and leverage probability sampling to determine $\bar{x}$, which is related to $\bar{\mu} = \bar{\underset{\sim}{x}}'\underset{\sim}{\phi}$. Employing probability sampling enhances the robustness of our model.

To sample the parameters from the posterior distributions obtained, we employ the Gibbs sampler. For parameters with complex forms, the grid method serves as a practical approach for generating samples. This method is straightforward, involving the generation of samples in the range (0,1), with subsequent simple transformations to ensure our target parameter falls within the desired range. However, challenges associated with the grid method includes computational time and the potential for getting stuck at boundary values. In this article, we introduce a new grid method known as the bisection grid method. Unlike traditional approaches, this method generates samples from the interval (0,1) with the unique advantage of minimizing the need for repeated density calculations and addressing boundary-related challenges. This modification enhances computational efficiency, making it a valuable alternative in parameter estimation. The main idea for the bisection grid method is to find the area under the curve and cut the area by the half point of the range, then we randomly select one side area in probability then repeat the process. This way can decrease the target range and computation time efficiently because we do not need to repeat the calculation many times and the area under the curve can be approximated in many methods. More details are given in the Appendix B.

The primary contribution of this article lies in supplementing a non-probability sample with a probability sample to make inferences about a finite population mean using various methods through robust adjusted weights model. The specific contributions include:

(1) Double Imputation Model: This model employs both tree and Euclidean distance to impute missing $\underset{\sim}{y}$ values for the probability sample and missing survey weights for the non-probability sample.

(2) We explore two Bayesian approaches to enhance the method of Chen et al. (2020) for estimating survey weights. The first approach focuses on the ignorable assumption, utilizing Bayesian techniques to derive survey weights for the probability sample. The second approach extends the method to accommodate a non-ignorable assumption, incorporating tree imputation strategies.

(3) We explore various Bayesian approaches for estimating survey weights under both ignorable and non-ignorable assumptions without using the Horvitz-Thompson (HT) estimator. In the ignorable assumption setting, we employ a Bayesian method with logistic regression, utilizing a matching approach to determine survey weights. For the non-ignorable assumption, we investigate two different strategies: one involves tree imputation within the logistic regression, and the other employs Bayesian imputation in the same logistic regression framework on the combined sample.

(4) We investigate the efficacy of the robust adjusted weights model, enhancing the overall robustness of the results across all the models employed.

(5) Efficiency improvement with Bisection Grid method: Utilizing the bisection grid method to enhance the efficiency of the Gibbs sampler in estimating parameters.

The article is organized as follows. In Section 2, we present a review of the relevant methods to be employed in this article, providing an understanding of the background. In Section 3, we introduce various Bayesian approaches to find survey weights. In Section 4, we outline the application of the robust adjusted weights model. Finally, in Section 5, we employ BMI data, incorporating three covariates to illustrate and compare differences among the models.

2. Review of the Relevant Methodology

In Section 2, we will provide a review of the method of Chen et al. (2020) and introduce the concept of the adjusted weights model. Subsequently, we will outline the process of double imputation.

2.1 Review of the Method of Chen et al. (2020)

Now suppose a finite population with a size of N, where the unit in the population is associated with covariates $\underset{\sim}{x}_i, i = 1, ..., N$. We have nps sample S_1 with size n_1 and ps sample S_2 with size n_2. The indicator variable $R_i = 1$ signifies that the unit is selected into the non-probability sample, S_1. In Appendix A we provide the likelihood function and log-likelihood function with $\pi(\underset{\sim}{x}_i, \underset{\sim}{\theta}) = exp(\underset{\sim}{x}_i'\underset{\sim}{\theta})/\{1 + exp(\underset{\sim}{x}_i'\underset{\sim}{\theta})\}$ as follows,

$$p(\underset{\sim}{R}|\underset{\sim}{x}, \underset{\sim}{\theta}) = \prod_{i=1}^{N} \pi(\underset{\sim}{x}_i, \underset{\sim}{\theta})^{R_i}\{1 - \pi(\underset{\sim}{x}_i, \underset{\sim}{\theta})\}^{1-R_i}, \tag{4}$$

$$\ell(\underset{\sim}{\theta}) = \sum_{i \in S_1} \underset{\sim}{x}_i'\underset{\sim}{\theta} - \sum_{i \in S_2} d_i \log\{1 + exp(\underset{\sim}{x}_i'\underset{\sim}{\theta})\}. \tag{5}$$

The conventional approach involves using numerical methods, such as the Newton-Raphson method, to find the estimated value of θ. Once the values of θ have been obtained, we can utilize them to estimate the propensity score, $\pi(\underset{\sim}{x}_i, \underset{\sim}{\theta})$, for the unit in the non-probability sample.

2.2 Adjusted Weights Model

After having the weights, a very common way to learn the study variable is using the adjusted weights model. Now suppose we have population $y_i \sim \text{Normal}(\underset{\sim}{x}_i'\underset{\sim}{\phi}, \sigma^2), i = 1, ..., N$, then we know the population mean will have $\bar{Y} \sim \text{Normal}(\bar{x}'\underset{\sim}{\phi}, \sigma^2/N)$. Now under this assumption, we can use the data $y_i, i = 1, ..., n_1$, from our nps sample with known weights $\tilde{W}_i$ to find the value of ϕ and σ^2 in Bayesian way given the prior distribution of ϕ and σ^2 using adjusted weights model $y_i|\underset{\sim}{\phi}, \sigma^2 \sim \text{Normal}(\underset{\sim}{x}_i'\underset{\sim}{\phi}, \sigma^2/w_i), i = 1, ..., n_1$. Here the adjusted weights are constructed through the weights W_i by equation (1) and (2). Then we can generate samples from the $\bar{Y} \sim \text{Normal}(\bar{x}'\hat{\phi}, \hat{\sigma^2}/\hat{N})$ to learn the mean of the population. $\hat{N} = \sum d_i, i = 1, ..., n_2$, $\bar{x} = \sum_{i=1}^{n_2} d_i x_i/\hat{N}$ and $d_i, i = 1, ..., n_2$ are the known weights from ps sample. This sampling process for $\bar{Y}$ is termed the surrogate sampling approach (Nandram 2007). Our sampling process relies on two models: one for the sample, which aids in estimating $\underset{\sim}{\phi}$ and σ^2, the other for the population, which is built on the estimated parameters we have obtained.

In the following section we will talk about three cases about how to find the weights which is based on the whole population and the combined sample. The first case involves the double imputation method, after having the survey weights $\hat{d}_i$ by matching from the probability sample, we can use the estimated $\hat{N} = \sum d_i, i = 1, ..., n_2$, from ps sample to re-scale the weights to make sure the sum of the weights is equal to the population size using equation (6) and then fit the adjusted weights model. Here,

$$W_i = \frac{\hat{N}\hat{d}_i}{\sum_{i=1}^{n_1} \hat{d}_i}, i = 1, ..., n_1 \tag{6}$$

In the second case, the process is similar. After we have the propensity scores $\pi_i, i = 1, ..., n_1$, we can easily have the weights $\hat{d}_i = 1/\pi_i$. Then use the above process to re-scale the weights and fit the adjusted weights model to find the population mean.

In our third case, we compute the propensity scores, $\pi_{i1}, i = 1, ..., n_1$, for the non-probability sample (nps) and $\pi_{i2}, i = 1, ..., n_2$, for the probability sample (ps). Leveraging the known weights in the ps sample, we proceed with imputing the unknown weights of the nps sample by matching them with the closest values of propensity scores. Once we obtain the weights for the non-probability sample (nps), we proceed with re-scaling them by $\hat{N}$ and calculate the adjusted weights for the adjusted weights model to estimate the population mean.

It is crucial to note the difference in estimating weights among the three cases. In the second case, after obtaining the propensity score, we directly estimate the weights using the inverse of the propensity score, which is an estimation and not the actual survey weights. However, in the first and third cases, we impute the weights for the non-probability sample from the probability sample, resulting in the real survey weights.

2.3 Double Imputation

Double imputation stands out as a more direct method when compared to following approaches. It enables the direct estimation of survey weights for the non-probability sample. This method is applicable only in non-ignorable cases, as there are no missing y values in the ignorable scenario. After we use the CART to impute missing y values of probability sample, we have a probability sample with known auxiliary variables, estimated study variable and survey weights and a non-probability sample with known auxiliary variables and study variable. Add study variable as an additional covariate $\underset{\sim}{x}_i^* = (\underset{\sim}{x}_i, y_i)$, Euclidean distance on the new covariates serves as a measure to identify the minimum distance between each set of observations between the probability sample (ps) and non-probability sample (nps). Subsequently, the survey weights for the nps sample can be imputed using the corresponding weights from the ps sample with the minimum Euclidean distance. Our approach does not involve direct weight estimation through calculation; instead, we estimate the weights by matching from the true survey weights.

3. Bayesian Approaches for Estimating Weights

In Section 3, we will discuss the utilization of the Bayesian Approach for the method of Chen et al. (2020) and the Bayesian Approach on a combined sample without using the Horvitz-Thompson (HT) estimator. Additionally, we will introduce the bisection grid method.

3.1 Bayesian Approach for the Method of Chen et al. (2020)

Alternatively, we can adopt a Bayesian approach to achieve the same goal under ignorable assumption. To simplify our

sampling process, it is essential to re-scale θ to the range $(0, 1)$ by utilizing $\beta_k = \frac{e^{\theta_k}}{1+e^{\theta_k}}, k = 1, ..., K$. This transformation on the likelihood function can be performed directly,

$$p(\underset{\sim}{R}|\underset{\sim}{x}, \underset{\sim}{\beta}) = \prod_{i=1}^{N} \pi(\underset{\sim}{x_i}, \underset{\sim}{\beta})^{R_i}\{1 - \pi(\underset{\sim}{x_i}, \underset{\sim}{\beta})\}^{1-R_i}. \tag{7}$$

Then using the non-informative prior, Uniform(0, 1), we can have the log posterior distribution of $\underset{\sim}{\beta}$,

$$\ell(\underset{\sim}{\beta}|\underset{\sim}{x}, \underset{\sim}{R}) \propto \sum_{i\in S_1}\sum_{k=1}^{K} x_{ik}\log(\frac{\beta_k}{1-\beta_k}) - \sum_{i\in S_2}\sum_{k=1}^{K} d_i \log\{1 + exp\{x_{ik}\log(\frac{\beta_k}{1-\beta_k})\}\}. \tag{8}$$

Given the posterior distribution we can generate random sample from it, also find the selection probability based on the sample.

This method can also be done easily under the non-ignorable assumption, just need to add y_i as an additional covariate $x_i{}^* = (x_i, y_i)$ then repeat the process above. Given that the probability sample (ps) is lacking the study variable $\underset{\sim}{y}$, we can use CART to do the imputation.

3.2 Bayesian Approach on a Combined Sample

Now we still consider the finite population with a size N, we have a combined sample $S_3 = S_1 \cup S_2$ from the population. We use indicator $R_i = 1$ for the units in S_1 and $R_i = 0$ for the units in S_2. Then under the ignorable assumption we can have the logistic regression model,

$$p(\underset{\sim}{R}|\underset{\sim}{x}, \underset{\sim}{\theta}) = \prod_{i\in S_3} \pi(\underset{\sim}{x_i}, \underset{\sim}{\theta})_i^R (1 - \pi(\underset{\sim}{x_i}, \underset{\sim}{\theta}))^{1-R_i}. \tag{9}$$

The log-likelihood function is,

$$\begin{aligned}\log((\underset{\sim}{R}|\underset{\sim}{x}, \underset{\sim}{\theta})) &= \sum_{i\in S_3} R_i \log\{\pi(\underset{\sim}{x_i}, \underset{\sim}{\theta})\} + (1 - R_i)\log\{(1 - \pi(\underset{\sim}{x_i}, \underset{\sim}{\theta}))\} \\ &= \sum_{i\in S_1}\log\{\pi(\underset{\sim}{x_i}, \underset{\sim}{\theta})\} + \sum_{i\in S_2}\log\{(1 - \pi(\underset{\sim}{x_i}, \underset{\sim}{\theta}))\}.\end{aligned} \tag{10}$$

We can repeat the previous process for sampling θ from the posterior distribution, getting the propensity score for each unit. Next, impute the weights for the non-probability sample (nps) by matching the propensity score with the minimum difference between the nps sample and ps sample.

For the non-ignorable assumption, two scenarios arise. In the first scenario, similar to the earlier method, we impute the y_i value using the CART method. Subsequently, we estimate the parameters and determine the weights by matching the propensity score with the minimum difference between the non-probability sample (nps) and probability sample (ps).

In the second scenario, we impute the missing values for y_i in Bayesian way. In this case our log-likelihood function will be,

$$\log(P(\underset{\sim}{R}|\underset{\sim}{x}, \underset{\sim}{\theta}, \underset{\sim}{y_1}, \underset{\sim}{y_2})) = \sum_{i\in S_1}\log\{\pi(\underset{\sim}{x_i^*}, \underset{\sim}{\theta})\} + \sum_{i\in S_2}\log\{(1 - \pi(\underset{\sim}{x_i}, \underset{\sim}{\theta}, y_{2i}))\}. \tag{11}$$

We use y_{1i} and y_{2i} to represent the related values from nps and ps samples. In the equation, we make the assumption that $\underset{\sim}{y_2}$ is missing and treat it as an unknown parameter, similar to θ. We can use the previous process to transform θ to $\underset{\sim}{\beta}$, then use the Uniform(0,1) as the prior for it. Given the prior distribution $p(\underset{\sim}{\beta})$ for $\underset{\sim}{\beta}$ and $p(y_{2i})$ for y_{2i}, we can derive the joint posterior distribution,

$$\begin{aligned}\log(P(\underset{\sim}{\beta}, \underset{\sim}{y_2}|R, \underset{\sim}{x}, \underset{\sim}{y_1})) &\propto \sum_{i\in S_1}\log\{\pi(\underset{\sim}{x_i^*}, \underset{\sim}{\beta})\} + \log(p(\underset{\sim}{\beta})) \\ &+ \sum_{i\in S_2}\{\log\{(1 - \pi(\underset{\sim}{x_i}, \underset{\sim}{\beta}, y_{2i}))\} + \log(p(y_{2i}))\}.\end{aligned} \tag{12}$$

To construct the prior distribution of y_{2i}, we leverage the non-probability sample (nps). We assume there are a total of n_1 observations in the nps sample, with a total of n_{1k} unique observations of y. Imputation of y_{2i} is then carried out using the discrete unique values $y^* = (y_{11}, \ldots, y_{1k})$. We can run a linear regression model based on all y_1 and x from nps. Then give the prior distribution $y_i^* \sim Normal(x_i{}'\hat{\alpha}, \hat{\sigma}^2)$, $\hat{\alpha}$ and $\hat{\sigma}^2$ are from the linear regression model, generate samples of $\underset{\sim}{y_2}$ and $\underset{\sim}{\beta}$ from the posterior distributions can be performed. Given the distinct values of $x_i, i = 1, ..., n_2$, it is important to note that the prior distribution for each $y_{2i}, i = 1, ..., n_2$, will vary. Once we obtain $\underset{\sim}{\beta}$ and $\underset{\sim}{y_2}$, we can then iterate through the previous process to obtain the propensity scores and impute weights using the propensity scores.

3.3 Sampling from the Posterior Distributions

Next, we will demonstrate the application of the Gibbs sampler for sampling our posterior distributions. As the posterior distribution of y_2 is a discrete distribution based on y^* and straightforward to sample, our focus shifts to the posterior distribution of $\tilde{\theta}$. Once we have the posterior distribution $p(\theta|x, R)$ (or $p(\theta|x, R, y)$), let θ be a vector with p elements. For each θ_i, we need to sample from the distribution $p(\theta_i|x, R, \theta_{(i)})$. Here, $\theta_{(i)}$ denotes the vector without the ith element, resulting in a $p-1$ length vector. This sampling process is repeated N times to obtain $(\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)})$. Subsequently, we discard the first n samples during a burn-in period, retaining $N-n$ samples $(\theta^{(n+1)}, \theta^{(n+2)}, ..., \theta^{(N)})$. To ensure independence and mitigate strong correlations, we collect samples at cycles t, denoted as $(\theta^{(n+1)}, \theta^{(n+1+t)}, ..., \theta^{(N)})$, resulting in a total of N_t samples. Evaluation techniques such as effective sample size, Geweke test, trace plots, or auto-correlation are employed to confirm the independence of the collected samples. We will employ the bisection grid method, offering increased computational efficiency. Further details can be found in Appendix B.

4. Robust Adjusted Weights Model

Now, we delve into a robust variant known as the robust adjusted weights model, an extension of the adjusted weights model discussed in Section 2.

In this model, we introduce a latent variable z_i alongside the population $y_i, i = 1, ..., N$. The variable z_i remains unobserved in the real data. When $z_i = 1$, we have the model,

$$y_i|z_i = 1, \underset{\sim}{x_i}, \underset{\sim}{\phi}, \sigma_2^2 \sim \text{Normal}(\underset{\sim}{x_i}'\underset{\sim}{\phi}, \sigma_2^2),$$

representing outliers. On the other hand, when $z_i = 0$, we have the model,

$$y_i|z_i = 0, \underset{\sim}{x_i}, \underset{\sim}{\phi}, \sigma_1^2 \sim \text{Normal}(\underset{\sim}{x_i}'\underset{\sim}{\phi}, \sigma_1^2),$$

representing typical data. Notably, σ_1^2 is chosen to be smaller than σ_2^2. Assuming,

$$z_i|p \sim \text{Bernoulli}(p),$$

then we can express y_i as,

$$y_i|p, \underset{\sim}{x_i}, \underset{\sim}{\phi}, \sigma_1^2, \sigma_2^2 \sim p\text{Normal}(\underset{\sim}{x_i}'\underset{\sim}{\phi}, \sigma_2^2) + (1-p)\text{Normal}(\underset{\sim}{x_i}'\underset{\sim}{\phi}, \sigma_1^2).$$

For simplicity we assume the population mean, $\bar{Y}$, follows the model,

$$\bar{Y}|p, \underset{\sim}{x_i}, \underset{\sim}{\phi}, \sigma_1^2, \sigma_2^2 \sim p\text{Normal}(\underset{\sim}{\bar{x}}'\underset{\sim}{\phi}, \frac{\sigma_2^2}{N}) + (1-p)\text{Normal}(\underset{\sim}{\bar{x}}'\underset{\sim}{\phi}, \frac{\sigma_1^2}{N}).$$

See Appendix D for an improvement of the procedure. We have done the procedure and find that the results are very accurate.

Suppose we have samples $y_i, i = 1, ..., n_1$, from a non-probability sample with estimated weights W_i. Similar to the adjusted weights model in Section 2, we formulate the model $y_i|z_i = 1 \sim \text{Normal}(\underset{\sim}{x_i}'\underset{\sim}{\phi}, \frac{\sigma_2^2}{w_i})$ and $y_i|z_i = 0 \sim \text{Normal}(\underset{\sim}{x_i}'\underset{\sim}{\phi}, \frac{\sigma_1^2}{w_i})$. Here, we denote the probability density function (pdf) of $\text{Normal}(\underset{\sim}{x_i}'\underset{\sim}{\phi}, \frac{\sigma_2^2}{w_i})$ as N_{i2} and the pdf of $\text{Normal}(\underset{\sim}{x_i}'\underset{\sim}{\phi}, \frac{\sigma_1^2}{w_i})$ as N_{i1}, where w_i represents the adjusted weights. Consequently, the model for y_i becomes $y_i \sim pN_{i2} + (1-p)N_{i1}$. The population mean $\bar{Y}$ retains the previously mentioned model, but N can be calculated using the probability sample as $\hat{N} = \sum_{i=1}^{n_2} d_i$. The vector $\bar{x}$ can be determined using the same method illustrated in Section 2.

We incorporate the prior $\frac{1}{\sigma_1^2\sigma_2^2}$ and set $p \sim U(0, 0.5)$, $\sigma_1^2 < \sigma_2^2$. This leads to the following posterior distribution,

$$p(\underset{\sim}{\phi}, \sigma_1^2, \sigma_2^2, p, \underset{\sim}{z}|\underset{\sim}{y}) \propto \frac{1}{\sigma_1^2\sigma_2^2} I_{(0,1/2)}(p) \prod_{i=1}^{n_1} [N_{i1}^{1-z_i} N_{i2}^{z_i} p^{z_i} (1-p)^{1-z_i}] I_{(\sigma_1^2<\sigma_2^2)}. \tag{13}$$

The reasons we give the prior of p as $U(0, 1/2)$ are: (a). We expect a small numer of outliers in the data, as we said $z_i = 1$ will be related to the model $y_i \sim \text{Normal}(x_i'\phi, \sigma_2^2)$ which is seen as outlier so we expect p value should be smaller than 0.5. (b). We want to avoid label switching problem by giving the order of these two clusters.

To help us have an easier sampling algorithm using block Gibbs sampler, we can consider the model integrating out z_i,

$$y_i|p, \underset{\sim}{\phi}, \sigma_1^2, \sigma_2^2 \sim (1-p)N_{i1} + pN_{i2}. \tag{14}$$

Then we can have the posterior distribution for p,

$$p|\underset{\sim}{y}, \underset{\sim}{\phi}, \sigma_1^2, \sigma_2^2 \propto I_{(0,1/2)}(p) \prod_{i=1}^{n_1} \{(1-p)N_{i1} + pN_{i2}\}. \tag{15}$$

Then we get posterior distribution for z_i from equation (13),

$$z_i|\underset{\sim}{y}, \underset{\sim}{\phi}, \sigma_1^2, \sigma_2^2, p \sim \text{Bernoulli}\{\frac{pN_{i2}}{pN_{i2} + (1-p)N_{i1}}\}. \tag{16}$$

Now we can sample the $p, z_i, i = 1, ..., n_1$, from equation (15) and (16).
For $\underset{\sim}{\phi}$ and σ_1^2, σ_2^2 we re-write equation (13),

$$\begin{aligned} &\frac{1}{\sigma_1^2}^{\frac{n_1 - \sum_{i=1}^{n_1} z_i}{2} + 1} \frac{1}{\sigma_2^2}^{\frac{\sum_{i=1}^{n_1} z_i}{2} + 1} \\ &\times \exp\{-\frac{1}{2} \sum_{i=1}^{n_1} (y_i - \underset{\sim}{x}_i' \underset{\sim}{\phi})^2 (\frac{w_i}{\sigma_1^2} - \frac{w_i z_i}{\sigma_1^2} + \frac{w_i z_i}{\sigma_2^2})\} I_{(\sigma_1^2 < \sigma_2^2)}. \end{aligned} \tag{17}$$

We can rewrite the exponential part without $\frac{1}{2}$ as $\sum_{i=1}^{n_1} t_i (y_i - \underset{\sim}{x}_i' \underset{\sim}{\phi})^2$. Here $t_i = \frac{w_i}{\sigma_1^2} - \frac{w_i z_i}{\sigma_1^2} + \frac{w_i z_i}{\sigma_2^2}$. Then we can define $\hat{\underset{\sim}{\phi}} = (XTX')^{-1}(XT\underset{\sim}{y})$. $\underset{\sim}{y}$ is $n_1 \times 1$ vector, X is $k \times n_1$ matrix, T is $diag(t_1, t_2, ..., t_{n_1})$. Now our exponential part can be written as,

$$\begin{aligned} &(\underset{\sim}{y} - X'\underset{\sim}{\phi})' T (\underset{\sim}{y} - X'\underset{\sim}{\phi}) = \\ &(\underset{\sim}{y} - X'\hat{\underset{\sim}{\phi}} + X'\hat{\underset{\sim}{\phi}} - X'\underset{\sim}{\phi})' T (\underset{\sim}{y} - X'\hat{\underset{\sim}{\phi}} + X'\hat{\underset{\sim}{\phi}} - X'\underset{\sim}{\phi}). \end{aligned} \tag{18}$$

After some simplification we can have $(\underset{\sim}{y} - X'\underset{\sim}{\phi})' T (\underset{\sim}{y} - X'\underset{\sim}{\phi}) = (\underset{\sim}{y} - X'\hat{\underset{\sim}{\phi}})' T (\underset{\sim}{y} - X'\hat{\underset{\sim}{\phi}}) + (\hat{\underset{\sim}{\phi}} - \underset{\sim}{\phi})' XTX' (\hat{\underset{\sim}{\phi}} - \underset{\sim}{\phi})$. Then we will have posterior distribution for $\underset{\sim}{\phi}$,

$$\underset{\sim}{\phi} \sim \text{Normal}(\hat{\phi}, (XTX')^{-1}). \tag{19}$$

Now if we can write the equation (17) into product of two parts which are related to σ_1^2 and σ_2^2,

$$\begin{aligned} &\frac{1}{\sigma_1^2}^{\frac{n_1 - \sum_{i=1}^{n_1} z_i}{2} + 1} \exp\{-\frac{1}{\sigma_1^2} \sum_{i=1}^{n_1} (y_i - \underset{\sim}{x}_i' \underset{\sim}{\phi})^2 (\frac{w_i(1-z_i)}{2})\} \\ &\times \frac{1}{\sigma_2^2}^{\frac{\sum_{i=1}^{n_1} z_i}{2} + 1} \exp\{-\frac{1}{\sigma_2^2} \sum_{i=1}^{n_1} (y_i - \underset{\sim}{x}_i' \underset{\sim}{\phi})^2 (\frac{w_i z_i}{2})\} I_{(\sigma_1^2 < \sigma_2^2)}. \end{aligned} \tag{20}$$

Now we can have posterior distribution for σ_1^2, σ_2^2 which are two truncated Inverse Gamma distributions from equation (20),

$$\sigma_1^2 \sim \text{InvGam}(\frac{n_1 - \sum_{i=1}^{n_1} z_i}{2}, \frac{\sum_{i=1}^{n1} w_i (1-z_i)(y_i - \underset{\sim}{x}_i' \underset{\sim}{\phi})^2)}{2}), \tag{21}$$

truncated at the right by σ_2^2,

$$\sigma_2^2 \sim \text{InvGam}(\frac{\sum_{i=1}^{n_1} z_i}{2}, \frac{\sum_{i=1}^{n1} w_i z_i (y_i - \underset{\sim}{x}_i' \underset{\sim}{\phi})^2)}{2}), \tag{22}$$

truncated at the left by σ_1^2.

Then we can do the Gibbs sampler using equation (19), (21) and (22) to draw samples for $\underset{\sim}{\phi}, \sigma_1^2$ and σ_2^2. We can make inference about the population mean $\bar{Y}$ based on these samples.

5. Data Analysis

In this section, various methods will be employed to estimate the population mean BMI value. We use the part of the BMI data from National Center for Health Statistics (NCHS). The data totally have 1866 probability observations from

California with known weights, BMI values and some auxiliary variables. We will divide these 1866 observations into two parts: non-probability part with size 1566 and probability part with size 300. We just select the small size probability sample randomly and keep their weights and auxiliary variables, then the others will belong to non-probability sample and we will eliminated their weights. For the auxiliary variables we will keep three of them which are age, race, sex and we will give one more term called intercept; which is all 1 in our model. We want to learn the information of the mean value of BMI. We conducted a total of 12 methods, with the initial 6 employing the adjusted weights model for estimating the population mean, while the remaining 6 utilized the robust adjusted weights model. In all Bayesian approaches, we ensure that samples of related parameters in models from the posterior distributions exhibit independence, indicated by an effective sample size (ESS) of 1000 and large p-values of Geweke tests. The methods are:

1. Double Imputation Model (DI);
2. Bayesian Approach for the method of Chen et al. (Ignorable Assumption)(BIC);
3. Bayesian Approach for the method of Chen et al. (Non-Ignorable Assumption with Tree Imputation)(BNCT);
4. Bayesian Approach on a Combined Sample (Ignorable Assumption)(BI);
5. Bayesian Approach on a Combined Sample (Non-Ignorable Assumption with Tree Imputation)(BNT);
6. Bayesian Approach on a Combined Sample (Non-Ignorable Assumption with Bayesian Imputation)(BNB);
7. Double Imputation Model-Robust (DI-R);
8. Bayesian Approach for the method of Chen et al. (Ignorable Assumption)-Robust(BIC-R);
9. Bayesian Approach for the method of Chen et al. (Non-Ignorable Assumption with Tree Imputation)-Robust(BNCT-R);
10. Bayesian Approach on a Combined Sample (Ignorable Assumption)-Robust(BI-R);
11. Bayesian Approach on a Combined Sample (Non-Ignorable Assumption with Tree Imputation)-Robust(BNT-R);
12. Bayesian Approach on a Combined Sample (Non-Ignorable Assumption with Bayesian Imputation)-Robust(BNB-R).

Table 1. Posterior summaries.: Summaries are posterior mean (PM), posterior standard deviation (PSD) and 95% highest posterior density interval (95% HPDI). Model 1=DI; Model 2=BIC; Model 3=BNCT; Model 4=BI; Model 5=BNT; Model 6=BNB; Model 7=DI-R; Model 8=BIC-R; Model 9=BNCT-R; Model 10=BI-R; Model 11=BNT-R; Model 12=BNB-R

Model	PM	PSD	95% HPDI
1	27.307	0.138	(27.080, 27.537)
2	27.310	0.133	(27.089, 27.534)
3	27.938	0.146	(27.694, 28.179)
4	27.566	0.146	(27.324, 27.802)
5	27.664	0.145	(27.423, 27.900)
6	27.562	0.143	(27.324, 27.799)
7	26.517	0.121	(26.316, 26.709)
8	26.853	0.123	(26.654, 27.055)
9	27.153	0.129	(26.951, 27.363)
10	26.589	0.131	(26.368, 26.807)
11	26.598	0.131	(26.393, 26.807)
12	26.589	0.126	(26.383, 26.796)

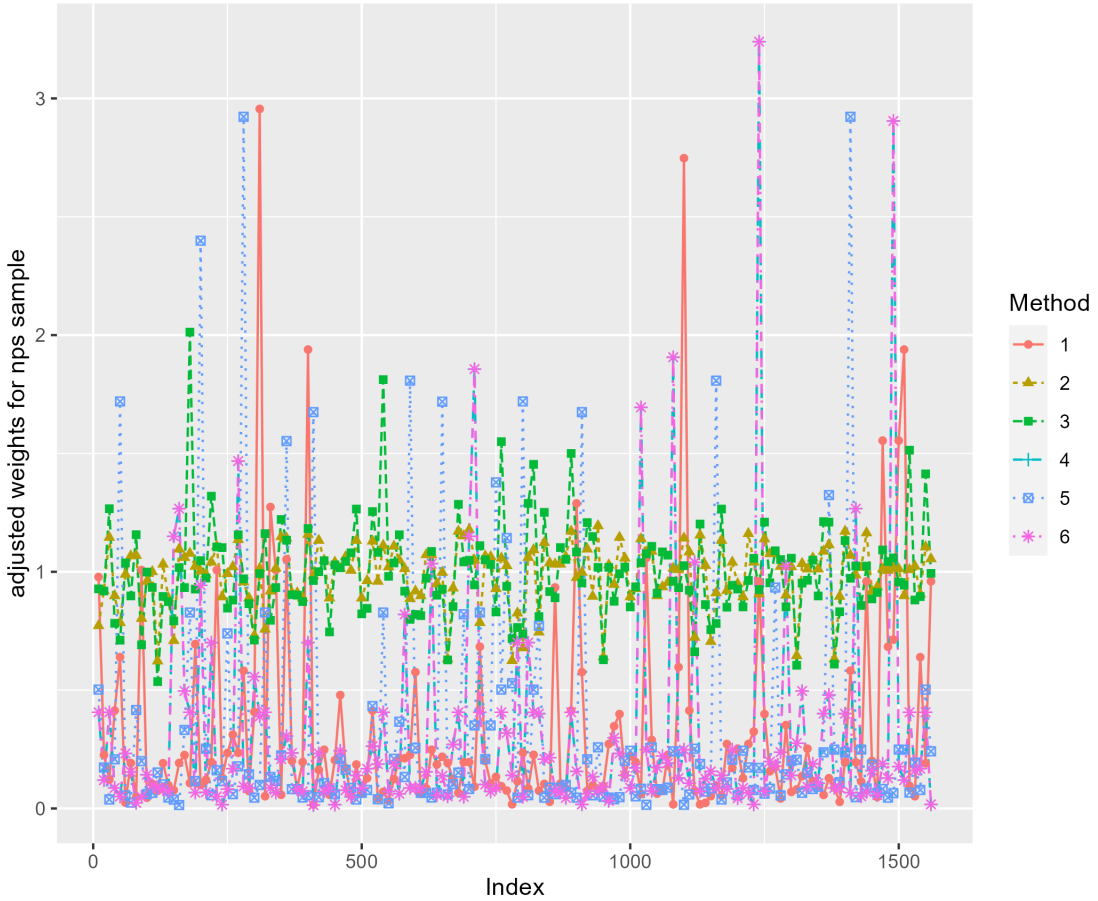

Figure 1. Adjusted weights across 6 models

A comparison of the adjusted weights across the six methods is depicted in Figure 1. Given that both the robust and non-robust approaches utilize the same adjusted weights from the initial six models, the comparison of adjusted weights is limited to these six models. Two distinct trends in the adjusted weights are evident. Methods 2 and 3 exhibit similar values, while methods 1, 4, 5 and 6 show a different set of similar values. Methods 2 and 3 have larger values and exhibit less variation compared to the other trend. This observation is logical, as Methods 2 and 3 both utilize the propensity score to estimate survey weights. In contrast, the other methods employ a matching approach to impute missing weights from the actual survey weights, leading to more variability in values. Upon scrutinizing individual comparisons, we observe that Methods 2 and 3 suggest that the ignorable case is more stable than the non-ignorable case. Additionally, Methods 4 and 6 show a similar trend when compared to Method 1.

Next, we examine the results of the population mean. Figure 2 illustrates the outcomes of 1000 samples from the distribution of $\bar{Y}$, while Figure 3 displays the distribution of the population mean across the first 6 models which use the adjusted weights model. Analyzing the results, we observe some similarities in the adjusted weights. When comparing Methods 2 and 3, Method 2 exhibits a smaller posterior mean (PM) and a smaller posterior standard deviation (PSD), this phenomenon may be attributed to a possible collinearity problem in the non-ignorable case. When estimating weights under a non-ignorable assumption $p(R_i|y_i, x_i)$, there may exist a relationship between y_i and x_i, leading to collinearity issues, particularly when imputing the missing y_i using CART based on x_i. Interestingly, Methods 4 and 6 reveal almost identical values for both posterior mean (PM) and posterior standard deviation (PSD). It appears that the Bayesian imputation method may not fully account for the effect of collinearity when compared to Method 3. Additionally, we uncover intriguing results. Method 1, employing double imputation, boasts the smallest posterior mean (PM) and the second smallest posterior standard deviation (PSD). In contrast, Method 2, the Bayesian Approach for the method of Chen et al. (2020) under the ignorable assumption, displays the smallest PSD and the second smallest PM. On the other hand, the other methods involving tree imputation do not appear to perform as effectively. Specifically, both Method 3 and Method 5 yield posterior means (PM) of 27.938 and 27.664, respectively.

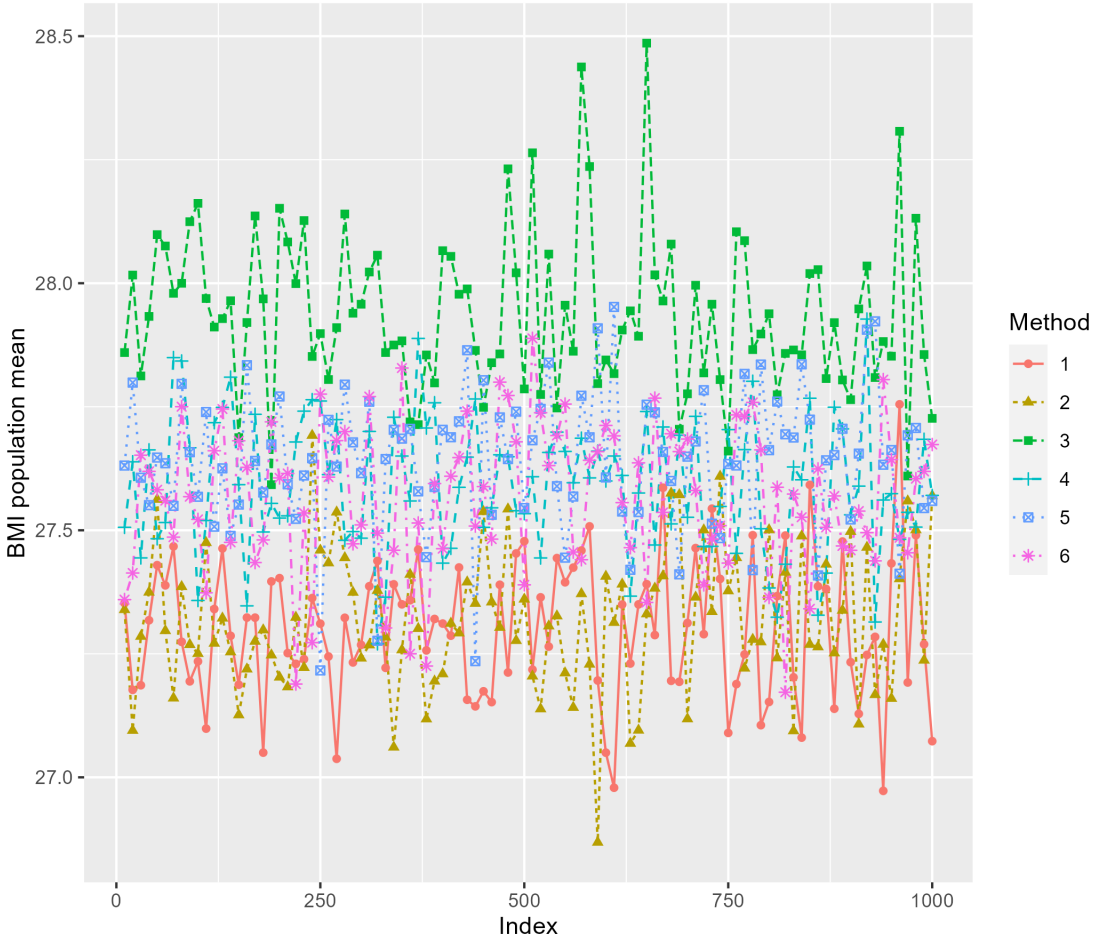

Figure 2. Comparison for the posterior distributions of the finite population mean by different models

The results obtained from the robust adjusted weights model, as presented in Table 1, Table 2, Figure 6 and Figure 7, highlight its notable effectiveness in mitigating bias. Table 1, Figure 6 and Figure 7 reveal that all posterior mean (PM) values decrease compared to the original results without using the robust model. Additionally, all highest posterior density intervals (HPDI) decrease, indicating that our robust model efficiently reduces the impact of outliers. The BI (Bayesian Approach on a Combined Sample (Ignorable Assumption)) method and BNB (Bayesian Approach on a Combined Sample (Non-Ignorable Assumption with Bayesian Imputation)) method exhibit similar outcomes, with both showing a reduction from 27.6 to 26.6.

Table 2 presents the results of the relative decrease (RD) by comparing the posterior means (PM) between the robust models and the original models.

$$\text{RD} = \frac{\text{PM of original model-PM of the related robust model}}{\text{PM of original model}}.$$

Interestingly, the BNT (Bayesian Approach on a Combined Sample (Non-Ignorable Assumption with Tree Imputation)) method exhibits the most significant decrease after adopting the robust adjusted weights model, showing similar results

Table 2. Comparison of the robust adjusted weights model and adjusted weights model,the Relative Decrease (RD) is calculated by comparing the PM of model using the robust adjusted weights model with the model using the adjusted weights model. Model 1=DI; Model 2=BIC; Model 3=BNCT; Model 4=BI; Model 5=BNT; Model 6=BNB.

Model	1	2	3	4	5	6
RD	2.9%	1.7%	2.8%	3.5%	3.9%	3.5%

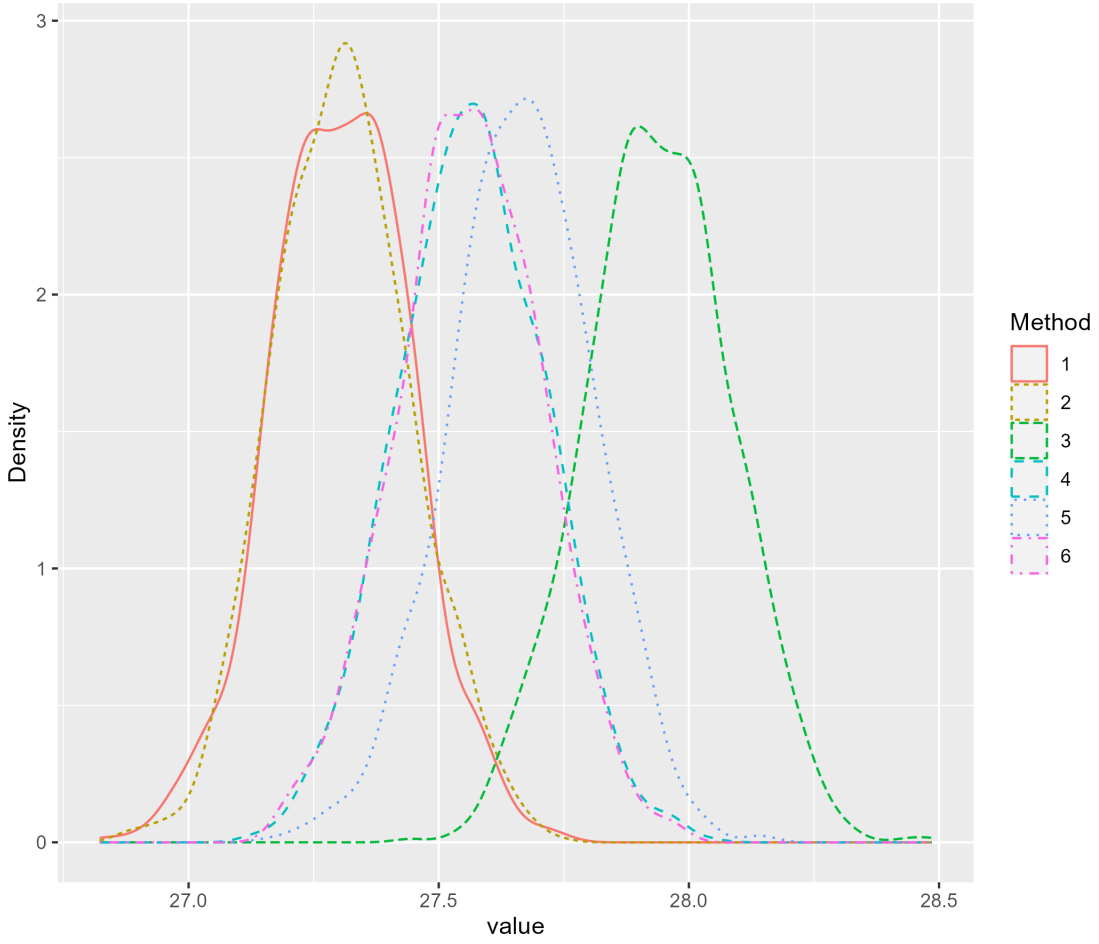

Figure 3. Comparison for the posterior distributions of the finite population mean by different models

when compared with the BI and BNB methods. Although the DI (Double imputation) method continues to have the smallest posterior mean, the advantage is not considerably large. An interesting observation is that the BIC (Bayesian Approach for the method of Chen et al. (2020) with Ignorable Assumption) method shows the least decrease after employing the robust adjusted weights model. Additionally, the posterior mean (PM) value becomes larger than most other values in this context. It appears that both the BIC and BNCT (Bayesian Approach for the method of Chen et al. (2020) with Non-Ignorable Assumption with Tree Imputation) methods yield the least favorable results. The HT estimator used by these two methods, while providing stable outcomes, may also introduce a notable bias when compared to other methods.

6. Concluding Remarks

This paper has demonstrated, under two scenariosłconsidering the entire population and utilizing a combined samplełhow to implement non-ignorable and ignorable assumptions in both non-probability and probability samples using Bayesian methods. Additionally, in cases where the study variable is not available in the probability sample, we illustrate the imputation process, highlighting the matching technique based on propensity scores. Moreover, the application of the robust adjusted weights model is demonstrated to enhance the robustness of estimations and mitigate the outlier effect across various methods, aiming to bring the population mean closer to the range of 20-25.

The diverse outcomes presented in the study demonstrate that the robust adjusted weights model significantly enhances the robustness of estimations efficiently. It is notable that the robust adjusted weights model consistently reduces the posterior mean by approximately 3% across all methods, further emphasizing its effectiveness in improving estimation robustness. Under the robust adjusted weights model, both Bayesian imputation and tree imputation exhibit noteworthy effectiveness, yielding results around a posterior mean of 26.5. This underscores the positive impact of these imputation techniques in enhancing the overall performance of the methods.

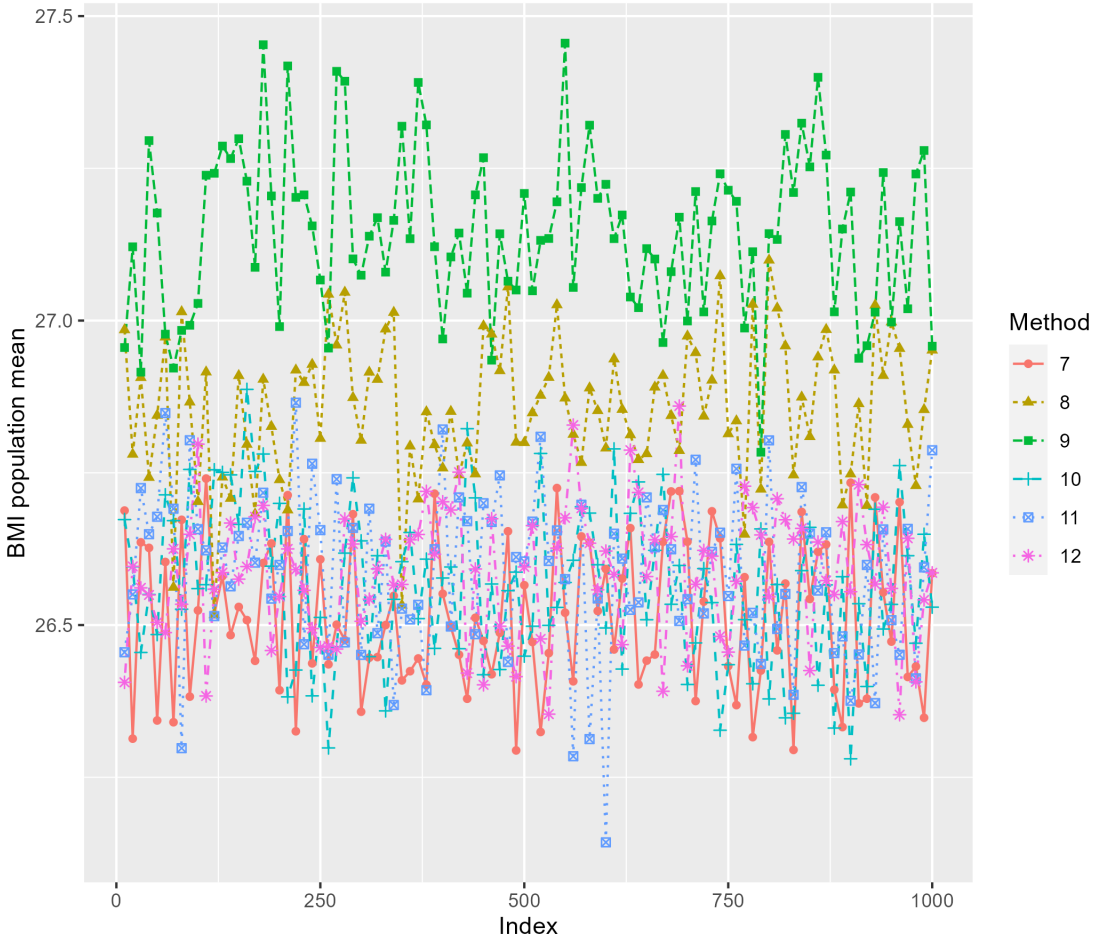

Figure 4. Comparison for the posterior distributions of the finite population mean by different models

Future work can be divided into two main areas. First, enhancements to the imputation method for the missing study variable are worth exploring. This may involve incorporating more robust priors or adopting non-parametric approaches to ensure the imputation process is more reliable. Additionally, to further improve the accuracy of the analysis, two potential avenues could be explored. Firstly, adopting robust regression methods that consider alternative distributions instead of the logit link may be beneficial. This is particularly noteworthy given the effective performance of tree imputation in the double imputation method and its comparatively poor performance in other approaches. Secondly, exploring a three-component mixture model for the robust adjusted weights model might prove advantageous in achieving enhanced accuracy.

Appendix A: Review of logit link for estimating propensity scores

Suppose the selection probability can be modeled parametrically as $\pi_i = \pi(\underset{\sim}{x}_i, \underset{\sim}{\theta})$, then we can have:

$$p(R|\underset{\sim}{x}, \underset{\sim}{\theta}) = \prod_{i=1}^{N} \pi(\underset{\sim}{x}_i, \underset{\sim}{\theta})^{R_i} \{1 - \pi(\underset{\sim}{x}_i, \underset{\sim}{\theta})\}^{1-R_i}.$$

The population log-likelihood is:

$$\ell(\underset{\sim}{\theta}) = \log(p(R|\underset{\sim}{x}, \underset{\sim}{\theta})) = \sum_{i=1}^{N} R_i \log\{\frac{\pi(\underset{\sim}{x}_i, \underset{\sim}{\theta})}{1 - \pi(\underset{\sim}{x}_i, \underset{\sim}{\theta})}\} + \sum_{i=1}^{N} \log\{1 - \pi(\underset{\sim}{x}_i, \underset{\sim}{\theta})\}.$$

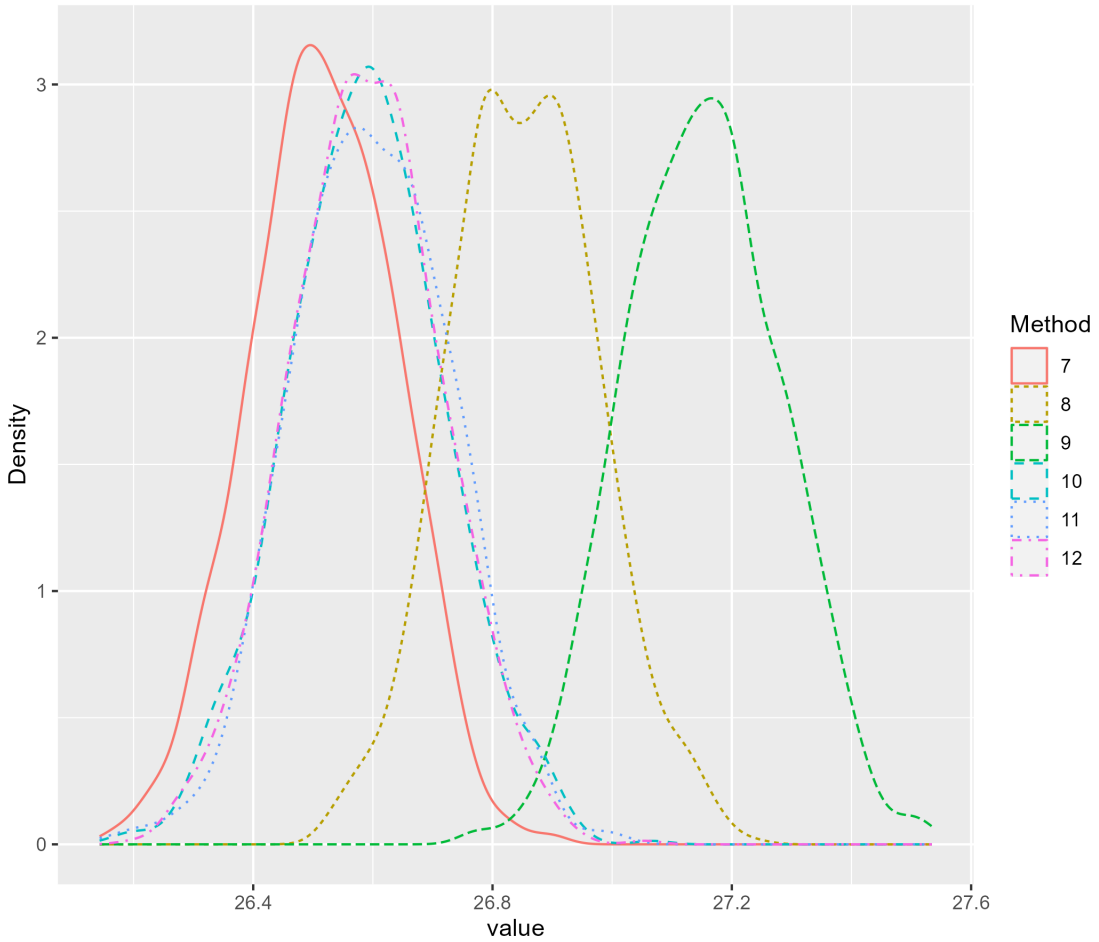

Figure 5. Comparison for the posterior distributions of the finite population mean by different models

Now we can see as R_i indicates if the unit i is from the non-probability sample, we can rewrite the equation as:

$$\ell(\underset{\sim}{\theta}) = \log(p(R|\underset{\sim}{x}, \underset{\sim}{\theta})) = \sum_{i \in S_1} \log\{\frac{\pi(\underset{\sim}{x_i}, \underset{\sim}{\theta})}{1 - \pi(\underset{\sim}{x_i}, \underset{\sim}{\theta})}\} + \sum_{i=1}^{N} \log\{1 - \pi(\underset{\sim}{x_i}, \underset{\sim}{\theta})\}.$$

Then Chen, Li and Wu. (2020) used the Horvitiz Thompson estimator from probability sample to replace the second part which is a pseudo log-likelihood model:

$$\ell(\underset{\sim}{\theta}) = \log(p(R|\underset{\sim}{x}, \underset{\sim}{\theta})) = \sum_{i \in S_1} \log\{\frac{\pi(\underset{\sim}{x_i}, \underset{\sim}{\theta})}{1 - \pi(\underset{\sim}{x_i}, \underset{\sim}{\theta})}\} + \sum_{i \in S_2} d_i \log\{1 - \pi(\underset{\sim}{x_i}, \underset{\sim}{\theta})\}.$$

Here the d_i is the known weights from probability sample. One basic way is to use the logit link function to model the $\pi(\underset{\sim}{x_i}, \underset{\sim}{\theta}) = exp(\underset{\sim}{x_i}'\underset{\sim}{\theta})/\{1 + exp(\underset{\sim}{x_i}'\underset{\sim}{\theta})\}$, so we can have model:

$$\ell(\underset{\sim}{\theta}) = \sum_{i \in S_1} \underset{\sim}{x_i}'\underset{\sim}{\theta} - \sum_{i \in S_2} d_i \log\{1 + exp(\underset{\sim}{x_i}'\underset{\sim}{\theta})\}.$$

Then we can use the above function to estimate the propensity score for nps. Subsequently, we can employ the inverse probability weighted estimator for the population mean: $\frac{1}{N}\sum_{i \in S_1} \frac{y_i}{\pi_i}$. This estimator involves weighting each observation in the non-probability sample by the inverse of its estimated propensity score and then computing the average.

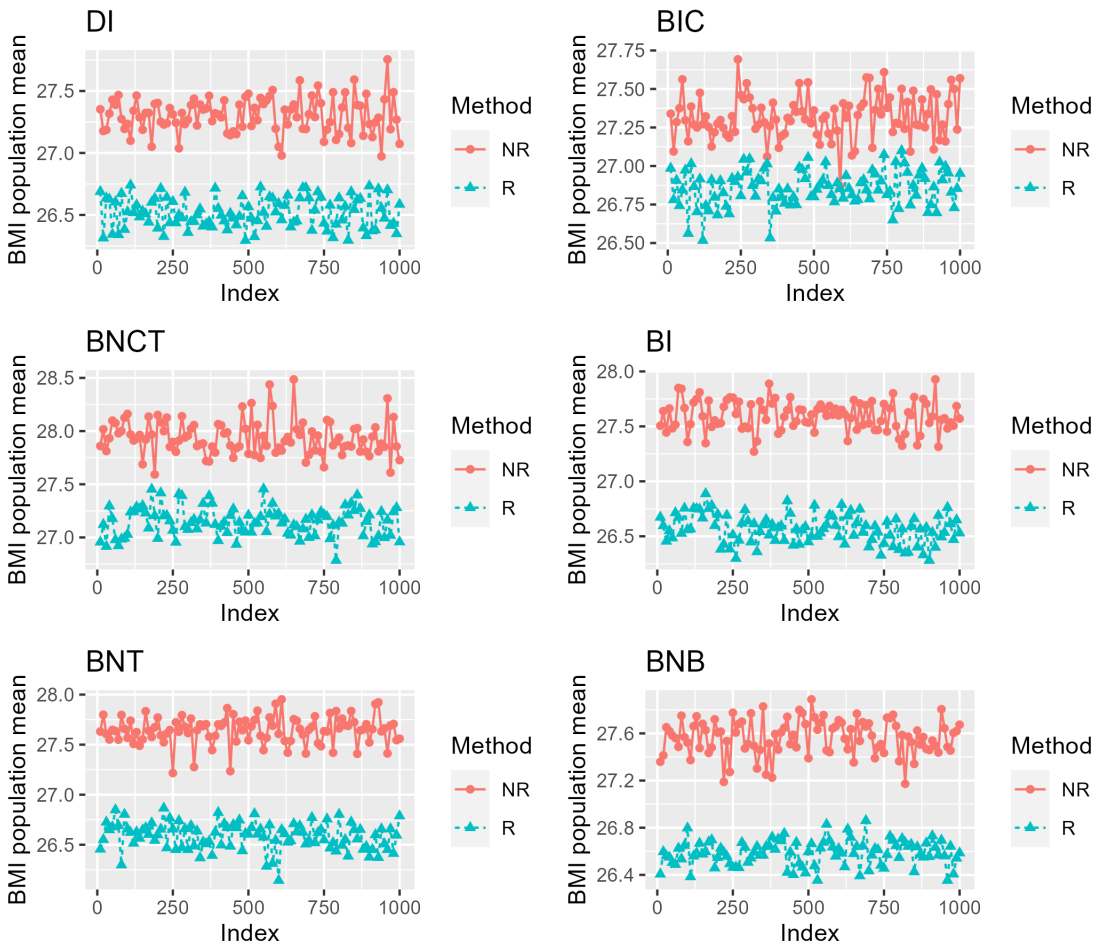

Figure 6. Comparison for the posterior distributions of the finite population mean by different models between robust and non-robust model

Appendix B: Bisection Grid Method

One problem in our Gibbs sampler is: the posterior distribution is not a well known form. So it will be hard to sample from it. We can do a transformation on each θ_i by $\alpha_i = \exp(\theta_i)/(1 + \exp(\theta_i))$. Then α_i will fall in the range $(0, 1)$, we can use the bisection grid method to draw samples. Bisection grid method is based on the idea of grid method but it is more efficient. We can run it in the following way:

1. We give two initial points $a = 0, b = 1$ and their mid point $\frac{(a+b)}{2}$,
2. We find the area under the curve for$(a, \frac{(a+b)}{2})$ and $(\frac{(a+b)}{2}, b)$,
   we call the left area as A_l and right one A_r,
3. Sample one area with related probability $(\frac{A_l}{A_l + A_r}, \frac{A_r}{A_l + A_r})$,
4. If sample the left area than let $a = a, b = \frac{(a+b)}{2}$
   otherwise $a = \frac{(a+b)}{2}, b = b$,
5. Repeat the above process 2-4 for 5-10 times,
6. Finally draw a sample from (a, b),
   which is just generating sample from simple uniform distribution.

From the outlined process, it is evident that for each sample, repeating the calculations 5-10 times is sufficient. Precision

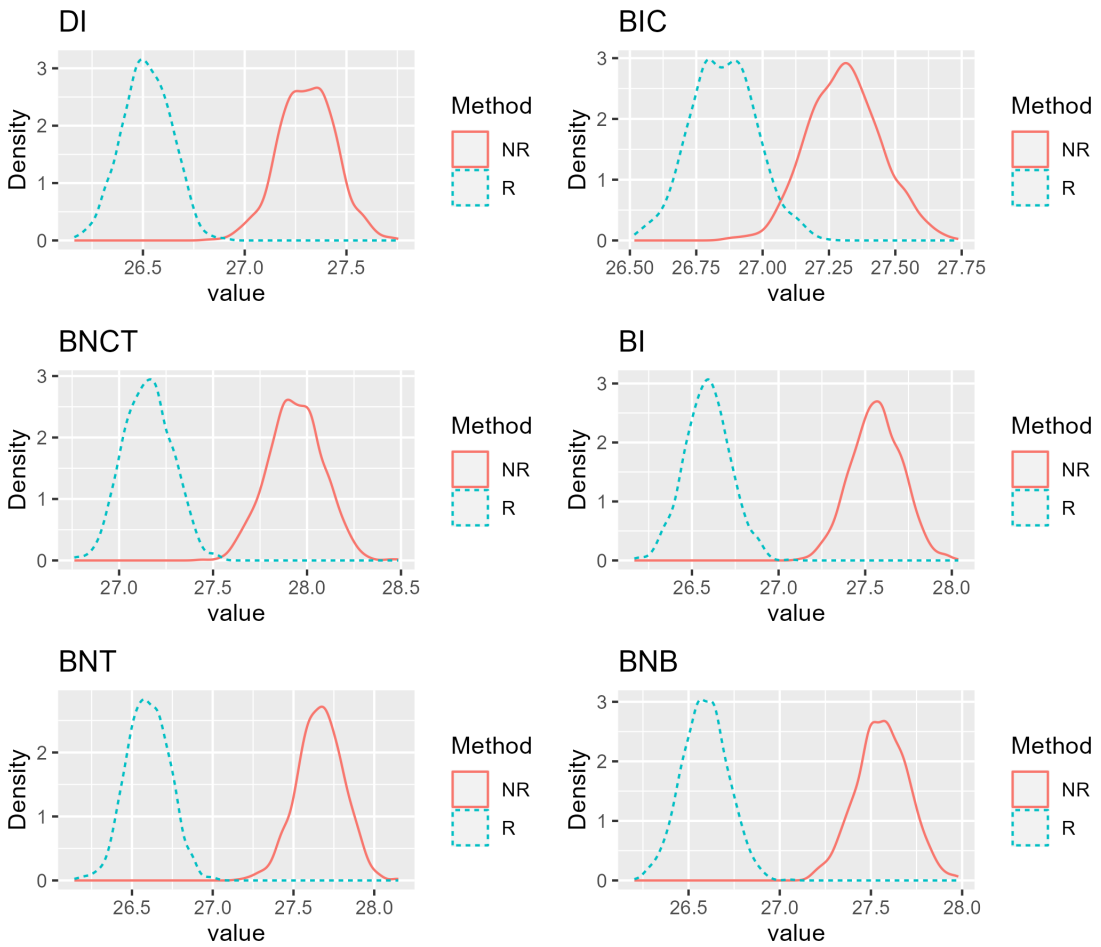

Figure 7. Comparison for the posterior distributions of the finite population mean by different models between robust and non-robust model

in the values of the areas is not crucial, estimation for both sides can be efficiently achieved through numeric integration. We specifically utilize GaussCLegendre quadrature for estimating the areas, a task easily accomplished with computer assistance. This approach not only simplifies the process but also significantly reduces computation time.

Appendix C: Classification and regression tree (CART) model

Our focus lies in tackling a regression problem using CART. The problem involves a single quantitative response variable, BMI value for each unit, along with multiple predictors. Our samples are divided into two parts: non-probability part (nps) and probability part (ps). The challenge arises from the missing response variable $\underset{\sim}{y}$ in ps, necessitating the use of CART for imputation. To address this, we employ the $\underset{\sim}{y}$ and $\underset{\sim}{x}_i$ from nps to fit the tree and subsequently perform imputation using the $\underset{\sim}{x}_i$ from ps. The CART algorithm aims to find the optimal data split based on predictors, minimizing the final sum of squares. This involves two key steps: determining the best split for each predictor and identifying the best overall split.
In the quest for the optimal predictor split, each potential binary split uniquely partitions the data. The sum of squares for the response variable is independently calculated in each partition and then aggregated. For instance, when using covariate age with a split point of 25, the data is divided into $\underset{\sim}{y}_1$ (age < 25) and $\underset{\sim}{y}_2$ (age ≥ 25), each with mean values $\bar{y}_1$ and $\bar{y}_2$. The best split for a predictor is the one that maximizes the reduction in sum of squares. Thus, the task is to identify the optimal split point for each predictor. Identifying the best overall split involves comparing predictors based on this reduction, with the winner being the one yielding the most substantial decrease‡resembling predictor selection. After acquiring the optimal data split, prediction can be directly made using subset data summary statistics. Alternatively, viewing the splitting process as a linear basis expansion with indicator variables from the best splits, a regression of the response on these basis functions yields coefficients and fit statistics in a standard manner.

In our approach, we leverage the covariates $\underset{\sim}{x_i}$ and y_i for $i = 1, ..., n_1$ from the non-probability sample to identify the best split point for each predictor and construct the overall optimal tree. Once the tree is established, we employ the covariates $\underset{\sim}{x_i}, i = 1, ..., n_2$, from the probability sample to impute the missing values of $\underset{\sim}{y}$. The entire process is executed in R programming using a package called mice (Stef van Buuren, Karin Groothuis-Oudshoorn 2011), which facilitates imputation using CART and ensures ease of implementation.

Appendix D: Prediction in Robust Model

In our problem, we aim to find the population mean $\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$ of a finite population with size N. The sum of the population $\sum_{i=1}^{N} y_i$ can be re-written as $\sum_{i=1}^{N}(z_i + 1 - z_i)y_i = \sum_{i=1}^{N} z_i y_i + \sum_{i=1}^{N}(1 - z_i)y_i = \sum_{z_i=1} y_i + \sum_{z_i=0} y_i$ given the model,

$$y_i|z_i = 1, \underset{\sim}{x_i}, \underset{\sim}{\phi}, \sigma_2^2 \sim \text{Normal}(\underset{\sim}{x_i'}\underset{\sim}{\phi}, \sigma_2^2),$$

$$y_i|z_i = 0, \underset{\sim}{x_i}, \underset{\sim}{\phi}, \sigma_1^2 \sim \text{Normal}(\underset{\sim}{x_i'}\underset{\sim}{\phi}, \sigma_1^2),$$

$$z_i|p \sim \text{Bernoulli}(p).$$

Since we can sample y_i from the related Normal distribution, the problem lies in obtaining all $\underset{\sim}{x_i}, z_i, i = 1, ..., N$. In our scenario, we have a non-probability sample of size n_1 with unknown weights, known y_i and a supplementary probability sample of size n_2 with known weights d_i and unknown y_i, both containing $\underset{\sim}{x_i}$. For the entire population, we know that the summation of each covariate should be a constant value. We can leverage the probability sample to estimate it,

$$\sum_{i}^{N} x_{ij} = \sum_{i=1}^{n_1} x_{ij} + \sum_{i=n_1+1}^{N} x_{ij} = \sum_{i=1}^{n_2} d_i x_{ij} = C_j, j = 1, ..., k$$

where k represents the number of covariates. Now we need to find the values $\underset{\sim}{x_i}, i = n_1 + 1, ..., N$. We can use the bootstrap to generate sufficient number of samples $(\underset{\sim}{x_{n_1+1}}, ..., \underset{\sim}{x_N})$ from $\underset{\sim}{x_i}, i = 1, ..., n_1$. Then we retain the set of samples that minimize the difference for each j,

$$|\sum_{i=n_1+1}^{N} x_{ij} - (C_j - \sum_{i=1}^{n_1} x_{ij})|, j = 1, ..., k.$$

Subsequently, we can generate samples of $z_i, i = 1, ..., N$ using a Bernoulli distribution with probability p. In Section 4, we illustrated how to generate samples from the posterior distributions of $\underset{\sim}{\phi}, \sigma_1^2, \sigma_2^2$, and p. Once we have $\underset{\sim}{x_i}, z_i, i = 1, ..., N$, we can sample $\underset{\sim}{y_i}, i = n_1 + 1, ..., N$ from the related Normal distribution. Then, we can make inference about the population mean based on these samples.

Acknowledgments

The authors thank the reviewers for their informative comments. Balgobin Nandram was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram).

Authors Contributions: Balgobin Nandram served as advisor and helped with the technical details. Zihang Xu, assisted by Balgobin Nandram, did most of the work, computing and wrote the paper.
Funding: None. Competing interests: None.

References

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees(1st ed.).* Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470

Chakraborty, A., Datta, G. S., & Mandal, A. (2019). A robust hierarchical bayes small area estimation for nested error linear regression model. *International Statistical Reviews, 87*, S1, S158-S156. https://doi.org/10.1111/insr.12283

Chen, Y., Li, P., & Wu, C. (2020). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association, 115*(532), 2011-2021. https://doi.org/10.1080/01621459.2019.1677241

Chipman, H. A., George, E. I., & McCulloch, R. E. (1998). Bayesian CART Model Search. *Journal of the American Statistical Association, 93*(443), 935-948. http://dx.doi.org/10.1080/01621459.1998.10473750

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics, 4*(1), 266-298. http://dx.doi.org/10.1214/09-AOAS285

Goyal, S., Datta, G. S., & Mandal, A. (2020). A Hierarchical Bayes Unit-Level Small Area Estimation Model for Normal Mixture Populations. *Sankhya, Series B*, S1-S27. http://dx.doi.org/10.1007/s13571-019-00216-8

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis With Missing Data.*(2nd ed.) Wiley Interscience, New York. https://doi.org/10.1002/9781119013563

Marella, D. (2023). Adjusting for Selection Bias in Nonprobability Samples by Empirical Likelihood Approach. *Journal of Official Statistics, 39*(2), 151-172. http://dx.doi.org/10.2478/JOS-2023-0008

Nandram, B., & Choi, J. W. (2002). Hierarchical Bayesian Nonresponse Models for Binary Data From Small Areas With Uncertainty About Ignorability. *Journal of the American Statistical Association, 97*(458), 381-388. http://dx.doi.org/10.1198/016214502760046934

Nandram, B., & Choi, J. W. (2002b). A Bayesian Analysis of a Proportion under Nonignorable Nonresponse. *Statistics in Medicine, 21,* 1189-1212. https://doi.org/10.1002/sim.1100

Nandram, B., & Choi, J. W. (2006). Hierarchical Bayesian Nonignorable Nonresponse Regression Models for Small Areas: An Application to the NHANES III Data. *Survey Methodology, II(1)*,73-84

Nandram, B. (2007). Bayesian Predictive Inference Under Informative Sampling via Surrogate Samples. In: S.K. Upadhyay, Umesh Singh, Dipak K. Dey (Eds.),*Bayesian Statistics and Its Applications, Anamaya, New Delhi*, 356-374 Chapter 25. https://doi.org/10.1002/asmb.650

Nandram, B., & 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*(489), 120-135. http://dx.doi.org/10.1198/jasa.2009.ap08443

Nandram, B., Bhatta, D.,& Shen, G. (2013). Bayesian Predictive Inference of a Finite Population Proportion Under Selection Bias. *Statistical Methodology, 11*, 1-21. http://dx.doi.org/10.1016/j.stamet.2012.08.003

Nandram, B., & Rao, J. (2021). A Bayesian approach for integrating a small probability sample with a non-probability sample. In JSM Proceedings, *Survey Research Methods Section,* Alexandria, VA: American Statistical Association, 1568-1603

Nandram, B., Choi, J. W., & Liu, Y. (2021). Integration of nonprobability and probability samples via survey weights. *International Journal of Statistics and Probability, Canadian Center of Science and Education, 10*(6), 1-5

Nandram, B., & J, N, K, Rao. (2024). Bayesian Integration for Small Areas by Supplementing a Probability Sample with a Non-probability Sample. *Statistical and Applications, 20*(10), 1-32.

Potthoff, R. F., Woodbury, M. A., & Manton, K. G. (1992). Equivalent Sample Size and Equivalent Degrees of Freedom Refinements for Inference Using Survey Weights Under Superpopulation Models. *Journal of American Statistical Association, 87*(418), 383-396. https://doi.org/10.2307/2290269

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, 63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Salvatore, C., Biffignandi, S., Sakshaug, J. W., Wiśniowski, A., & Struminskaya, B. (2023). Bayesian Integration of Probability and Non-probability Samples for Logistic Regression. *Journal of Survey Statistics and Methodology (2023)00*, 1-35 https://doi.org/10.1093/jssam/smad041

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1-67. https://doi.org/10.18637/jss.v045.i03