Statistics and Applications {ISSN 2454-7395 (online)}
Special Issue on Recent Advances in Bayesian Statistics and Machine Learning
Volume 24, No. 2, 2026 (New Series), pp 187–210
http://www.ssca.org.in/journal

Bayesian Predictive Inference in Logistic Regression with Covariates and Survey Weights

Lingli Yang[1] **and Balgobin Nandram**[2]

[1]*Department of Statistics and Quantitative Sciences*
Takeda Pharmaceutical Company Limited, MA, USA
[2]*Department of Mathematical Sciences*
Worcester Polytechnic Institute, MA, USA

Received: 09 November 2025; Revised: 31 July 2026; Accepted: 04 August 2026

Abstract

In this study, we present a comprehensive Bayesian predictive inference framework for binary responses, using of probability survey samples with auxiliary data. The proposed framework integrates survey weights into a logistic regression model. We explore six distinct models comprising both unnormalized and normalized weighted likelihoods, with three variations of adjusted survey weights: original, trimmed, and calibrated. The logistic regression model is implemented utilizing the Metropolis-Hastings sampler. Subsequently, we employ the surrogate sampling and stratification approach to make predictive inferences on finite population parameters. The comparative performance and efficacy of the six models are assessed via a simulation study and a real-world dataset on body mass index. Our results reveal that models incorporating normalized density functions with adjusted trimmed weights demonstrate superior estimation accuracy while maintaining adherence to Bayesian principles, particularly in the presence of survey weight outliers. These findings underscore the significance of the proposed framework and the excellence of adjusted trimmed weights for effective predictive inference in survey-based studies.

Key words: Sampling theory; Survey weights; Bayesian inference; Pseudo-likelihood; Generalized linear models.

AMS Subject Classifications: 62D05, 62F15, 62J12

1. Introduction

In statistics, survey sampling is a vital methodology for obtaining valid and reliable data from a representative subset of a population. The goal of survey sampling is to reduce cost and obtain a sample with similar characteristics to the population in terms of key variables of interest. By carefully selecting a sample that is representative of the population, researchers can make predictions and inferences based on the sample data. Also, survey

Corresponding Author: Balgobin Nandram
Email: balnan@wpi.edu

sampling is utilized to make accurate estimates and predictions about the population as a whole.

In survey sampling, survey weights are used to adjust for any sampling biases that may be present in the sample data. They are used to ensure that the sample is representative of the population from which it was drawn. The use of survey weights is essential to ensure that inferences and estimates made from the sample data are accurate and unbiased (Lohr, 2009). However, incorrect or improper use of survey weights can lead to biased estimates and inaccurate inferences. The opening sentence in Gelman (2007) is, "Survey weighting is a mess. It is not always clear how to use weights in estimating anything more complicated than a simple mean or ratio, and standard errors are tricky even with simple weighted means." Lohr (2007) responded in her comments, "I do not think that survey weighting is a mess, but I do think that many people ask too much of the weights." We agree with Lohr (2007) that there is a limit to how much can be gained from survey weights, but we should make every effort to do so.

The process of weighting involves the calculation of weight for each unit in the sample, which is done by taking the inverse of the probability of selection. This allows for a correction of bias that may be present in the sample data due to the sampling design. For example, if some groups in the population have a lower probability of being selected, then the weight for each unit in those groups will be adjusted upward to ensure that the sample is representative of the population. In data sets from population surveys, the weights attached to the units can include adjustments for unit non-response, and post-stratification to match the distributions of auxiliary variables with known distributions in the population (Chen *et al.*, 2017). Also, it is possible to generate survey weights by using a single nonprobability sample. Elliott and Valliant (2017) used propensity scores to obtain surrogates for survey weights for nonprobability sample. Chen *et al.* (2020) used weighted likelihoods to obtain propensity scores, and Robbins *et al.* (2021) concentrated on design-based methods such as calibration and propensity score weighting (without parametric models). Nandram and Rao (2021) and Nandram *et al.* (2021) showed a full Bayesian approach to incorporate nonprobability sample and a probability sample. However, our concern is not about how to estimate survey weights, but rather to show the comparisons of different types of adjusted weights into a parametric model. Note that Yang *et al.* (2023) discussed the performance of nine methods but they did not take auxiliary information into consideration.

When the study variable is binary, Bayesian logistic regression is a widely used probabilistic model that holds the relationship between covariates and the study variable. It utilizes Bayesian inference to calculate the posterior distribution of the model parameters given the observed data. In contrast to traditional (frequentist) logistic regression, Bayesian logistic regression explicitly models the uncertainty in the model parameters and produces a distribution over the possible parameter values. This distribution can be used to make probabilistic predictions and to qualify the uncertainty in the model's predictions. Although it is practical to weight units in order to help mitigate the effects of differential inclusion into the sample, the estimates that result can often be very ineffective.

Auxiliary information refers to additional data or variables that are not the primary focus of a statistical analysis but can provide valuable support or enhancement to the analysis (Srivastava, 1967). In our context, covariates (independent variables) are often considered

auxiliary information. While the response variable is the main focus, including covariates in the model can help account for additional variability and better capture the relationships within the data. The covariates not only provide auxiliary information, but also are an essential component in the modeling that focuses on regression and the estimation of the corresponding regression coefficients. Roberts *et al.* (1987) and Archer and Lemeshow (2006) discussed different test statistics for logistic regression analysis that are adapted to include the survey design. Rader *et al.* (2017) proposed weighted estimating equations for logistic regression models to correct bias of estimates. Chen and Nandram (2023) described a hierarchical Bayesian logistic regression model to capture the variation in the binary data. The same idea is shown in the work of Barasa and Muchwanju (2015), but they only analyzed the regression coefficients, rather than making predictions and inferences about a finite population quantity.

Predictive inference can be obtained using surrogate sampling. This model is adjusted using the survey weights to account for selection bias. After this adjustment, we obtain the posterior distribution of the model parameters, which are interpreted as population parameters. At this stage, the original sample data and their weights are no longer needed. Using the posterior parameters, we then generate the full population from the population model, which is referred to as surrogate sampling (Nandram, 2007). For example, suppose we have a population model with the population size N,

$$y_1, \ldots, y_N | \underset{\sim}{\theta} \overset{ind}{\sim} f(y \mid \underset{\sim}{\theta}).$$

With the observed sample data $(w_i, y_i), i = 1, \ldots, n$ and a prior on $\underset{\sim}{\theta}$, we are able to obtain the posterior density $\pi(\theta|\underset{\sim}{y_s})$. Suppose the focus is on the mean of the finite population, denoted as $\bar{Y}$. Subsequently, Bayesian predictive inference regarding the mean of the finite population is acquired through the following process,

$$\pi(\bar{Y}|\underset{\sim}{y_s}) = \int f(\bar{Y}|\theta)\pi(\underset{\sim}{\theta}|\underset{\sim}{y_s})d\underset{\sim}{\theta},$$

where $f(\bar{Y}|\underset{\sim}{\theta})$, the population model, does not depend on $\underset{\sim}{y_s}$; see Nandram and Rao (2021). We simply draw $\underset{\sim}{\theta}$ from the posterior density $\pi(\underset{\sim}{\theta}|\underset{\sim}{y_s})$ and draw $\bar{Y}$ from the population model. In general, the available information in this study is the study variable and the covariates of the survey sample. Based on the surrogate sampling method, we incorporate the survey weights into the population model to obtain the sample distribution.

The main goal of this study is to evaluate the performance of six models to do Bayesian predictive inference of finite population interests. In Section 2, we explain three types of adjusted survey weights: adjusted original survey weights, adjusted trimmed survey weights and adjusted calibrated survey weights. Then we incorporate weights into the unnormalized model and normalized model to access Bayesian predictive inference. In Section 3, a simulation study is carried out to further understand these six models by the absolute relative bias (ARB), the posterior standard deviation (PSD), the posterior root mean squared error (PRMSE), the coverage probability (CP), and the width of highest posterior density interval (Wid). In Section 4, we present an application using body mass index (BMI) data from the Third National Health and Nutrition Examination Survey (NHANES III) and compare the posterior means and posterior standard deviations of these six models. Finally, Section 5 gives some recommendations for further research after reviewing the study's advantages and disadvantages in greater depth.

2. Bayesian methodology

In this section, we establish a general framework for statistical inference of a binary variable from probability survey samples when pertinent auxiliary data are available. In particular, we demonstrate how to include survey weights in a logistic regression model. The survey weights are included using six different methods, which use either unnormalized or normalized weighted likelihoods. We employ three different kinds of adjusted survey weights, based on the original survey weights, trimmed survey weights, and calibrated survey weights.

2.1. Adjusted survey weights

Assuming that the selection probability only depends on the covariates that are observed (*i.e.*, ignorable selection), for unit i, let y_i be a vector of response; $\underset{\sim}{x_i}$ be a vector of observed covariates. When a sampling plan is used in a finite population of size N to draw a sample of size n, with given survey weights in the sample, $W_1, \ldots, W_n$, Horvitz-Thompson estimators of population total and population size are,

$$\hat{T} = \sum_{i=1}^{n} W_i y_i, \quad \hat{N} = \sum_{i=1}^{n} W_i.$$

These estimators are typically applied in a similar ways and are design-unbiased.

Weight trimming is a technique used to handle outliers in a dataset by replacing extreme values with less extreme ones (Haziza and Beaumont, 2017). This can reduce the impact of outliers on statistical analyses. Specifically, weight trimming replaces extreme values with either the next smallest or next largest non-outlier value in the dataset, up to a certain percentile or quantile threshold. It can help improve the robustness and reliability of statistical analyses; see Rao (1966) and Basu (2010).

In our work, $Q_3 + 1.5(Q_3 - Q_1)$ is considered as the threshold, where Q_1 is the first quartile, Q_3 is the third quartile. Let $\underset{\sim}{W^*}$ be weights after trimming and W_0 be the threshold, where $W_0 = Q_3 + 1.5(Q_3 - Q_1)$, then

$$W_i^* = \begin{cases} W_0, & W_i \geq W_0 \\ aW_i, & W_i < W_0 \end{cases}, \tag{1}$$

where a is a rescaling parameter so that $\sum_{i=1}^{n} W_i^* = \sum_{i=1}^{n} W_i = \hat{N}$, ensuring the summation of weights would not be changed after the trimming. This rescaling step is important to keep the HT estimator of population size the same, but we can not trim too much. The choice of W_0 has been discussed by Potter (1990); see also Chen *et al.* (2017) for an explanation of weight trimming processes. In our case, we use $Q_3 + 1.5(Q_3 - Q_1)$, which is the common rule in boxplot construction to detect outliers. It is true that weight trimming can lead to biased estimates, but it will increase the effective sample size.

Calibrated survey weights were used to adjust for the under-representation of certain population groups in the sample (Haziza and Beaumont, 2017). This was necessary to improve the accuracy and reliability of the study's estimates of obesity prevalence in the finite population. Calibrated survey weights are a common technique used in survey research to adjust for non-representative samples or under-representation of specific subgroups.

Assume we can find the totals of each covariate, $\underset{\sim}{t}$, through a census, government documents, or online scraping. By the calibration equations, $\sum_{j=1}^{n}\tilde{W}_j x_j = \underset{\sim}{t}$, survey weights are calibrated as $\tilde{W}$. As proposed by Haziza and Beaumont (2017), a distance function $G(u)$ is used to make sure $\tilde{W}$ are as close as possible to W, where $G(u) \geq 0$ and $G(1) = 0$; $G(u)$ is differentiable, say $g(u) = G'(u), g(1) = 0$, and strictly convex. The coefficient q_j is the scale factor indicating the importance of unit j in the distance calculation. In most practical situations, the factor q_j is set to 1. Let the q_j be set to 1, then the calibrated weights are

$$\begin{aligned} &\underset{\tilde{W}_j}{\operatorname{argmin}} \sum_{j=1}^{n} \frac{\tilde{W}_j}{q_j} G\left(\frac{\tilde{W}_j}{W_j}\right) \\ &\text{s.t.} \quad \sum_{j=1}^{n} \tilde{W}_j x_j = \underset{\sim}{t}. \end{aligned} \tag{2}$$

To find the minimal value of a function subject to equality constraint, Lagrangian multipliers, $\underset{\sim}{\lambda} = (\lambda_1, \ldots, \lambda_p)'$, are commonly used. We have,

$$\phi\left(\tilde{W}_1, \ldots, \tilde{W}_n, \underset{\sim}{\lambda}\right) = \sum_{j=1}^{n} \frac{\tilde{W}_j G\left(\frac{\tilde{W}_j}{W_j}\right)}{q_j} - \underset{\sim}{\lambda}'\left(\sum_{j=1}^{n} \tilde{W}_j x_j - \underset{\sim}{t}\right). \tag{3}$$

After we obtain $\hat{\underset{\sim}{\lambda}}$ (Haziza and Beaumont, 2017; Nandram *et al.*, 2021), the calibrated weights are straightforward,

$$\tilde{W}_j = W_j g^{-1}\left(q_j \hat{\underset{\sim}{\lambda}}' \underset{\sim}{x}_j\right), j = 1, \ldots, n. \tag{4}$$

In our study, we consider the simple Euclidean distance, $G(u) = \frac{1}{2}(u-1)^2$, to have the closed-form solutions. Let

$$A = \sum_{j=1}^{n} q_j W_j \underset{\sim}{x}_j \underset{\sim}{x}_j', \quad \underset{\sim}{b} = \underset{\sim}{t} - \sum_{j=1}^{n} W_j \underset{\sim}{x}_j. \tag{5}$$

Therefore, by solving $A\underset{\sim}{\lambda} = \underset{\sim}{b}$ for $\hat{\underset{\sim}{\lambda}}$ and assuming A is invertible, we have $\hat{\underset{\sim}{\lambda}} = A^{-1}\underset{\sim}{b}$ and the calibrated weights are,

$$\tilde{W}_j = W_j\left(1 + q_j \hat{\underset{\sim}{\lambda}}' \underset{\sim}{x}_j\right), j = 1, \ldots, n. \tag{6}$$

It is worth noting that if $\hat{\underset{\sim}{\lambda}} = \underset{\sim}{0}$, there are no differences between original survey weights and calibrated ones. In addition, negative calibrated weights can occur when the sample selected for the survey is not representative of the population in some way. This can happen for a variety of reasons, such as non-response bias, sampling error, or the use of an inappropriate weighting adjustment procedure. When this happens, the calibrated weights may be negative for some observations in the sample.

To avoid the negative calibrated weights, we rescale the negative weights to be positive. This is done by setting the weights that less than 1 to 1, because the smallest representative size should be greater or equal to 1, then normalized them to sum to the original total weights so that the sample size is preserved.

Effective sample size is a measure of the amount of independent information contained in a sample of data. In statistics, it is used to adjust for the fact that the observed sample size may not reflect the actual amount of independent information in the data due to issues such as dependence between observations, clustering of data, or non-random sampling (Potthoff *et al.*, 1992). A larger effective sample size indicates that the sample contains more independent information, and therefore provides more accurate estimates and narrower confidence intervals. Conversely, a smaller effective sample size indicates that the sample contains less independent information, which can result in less precise estimates and wider confidence intervals. Let n_e denote the effective sample size, and $\underset{\sim}{w}$ be adjusted original weights, $\underset{\sim}{w}^*$ be adjusted trimmed weights, $\underset{\sim}{\tilde{w}}$ be adjusted calibrated weights. For $i = 1, \ldots, n$, we construct three types of adjusted weights,

$$n_e = \frac{\left(\sum_{j=1}^n W_j\right)^2}{\sum_{j=1}^n W_j^2}, \quad w_i = \frac{n_e W_i}{\sum_{j=1}^n W_j}, \tag{7}$$

$$n_e^* = \frac{\left(\sum_{j=1}^n W_j^*\right)^2}{\sum_{j=1}^n W_j^{*2}}, \quad w_i^* = \frac{n_e^* W_i^*}{\sum_{j=1}^n W_j^*}, \tag{8}$$

$$\tilde{n}_e = \frac{\left(\sum_{j=1}^n \tilde{W}_j\right)^2}{\sum_{j=1}^n \tilde{W}_j^2}, \quad \tilde{w}_i = \frac{\tilde{n}_e \tilde{W}_i}{\sum_{j=1}^n \tilde{W}_j}. \tag{9}$$

So far, we described the details of these three types of adjusted weights; adjusted original weights $\underset{\sim}{w}$, adjusted trimmed weights $\underset{\sim}{w}^*$, and adjusted calibrated weights $\underset{\sim}{\tilde{w}}$. They are more accurate and reliable than unadjusted ones. In addition, adjusted trimmed weights are not able to be unduly influenced by a few extreme observations, and can help improve the robustness of the analysis to data that may not confirm to certain assumptions, such as the normality of the data or the homogeneity of variance (Lohr, 2009). Survey weights are designed to reflect the representativeness of sampled units about the broader population. Extreme weights may result from various factors, such as nonresponse, oversampling of certain groups, or errors in the weight calculation process. For example, if the extreme survey weights are caused by oversampling of certain groups, the adjusted trimmed weights downplay the contribution of these values, reducing their impact on the analysis. In this case, the impact could be the normality of the data or the homogeneity of variance. Designed to adjust for differences between the sample and the population on certain variables of interest, such as age, gender, or education level, calibrated weights can help correct for potential biases and improve the representativeness of survey data.

2.2. Weighted density

Suppose that a probability sample of size n is taken from a finite population,

$$y_1, \ldots, y_n \mid \underset{\sim}{\theta} \overset{iid}{\sim} f(y \mid \underset{\sim}{\theta}).$$

With given survey weights of each unit in the sample, $\underset{\sim}{W}$, the Horvitz-Thompson estimator of the log-likelihood for the entire population, $L(\underset{\sim}{\theta})$, is

$$\widehat{L(\underset{\sim}{\theta})} = \sum_{i=1}^{n} W_i \log\left\{f(y_i|\underset{\sim}{\theta})\right\},$$

where we use W_i to illustrate our models, but replace the weights with adjusted survey weights w_i in application.

The density function is $g(y_i|\theta) \propto f(y_i \mid \underset{\sim}{\theta})^{W_i}, i = 1, \ldots, n.$

However, $f(y_i \mid \underset{\sim}{\theta})^{W_i}$ is an unnormalized 'density' function. We need to insert the normalization constant, $\int f(y_i \mid \underset{\sim}{\theta})^{W_i} dy_i$, such that $g(y_i \mid \underset{\sim}{\theta})$ integrates to 1. Therefore, we have,

$$g(y_i|\underset{\sim}{\theta}) = \frac{(f(y_i|\underset{\sim}{\theta}))^{W_i}}{\int (f(y_i|\underset{\sim}{\theta}))^{W_i} dy_i}.$$

When the normalization constant is a function of $\underset{\sim}{\theta}$, it is important to keep this part to follow the rules of Bayesian theory (see more in Nandram and Rao, 2024). The denominator can be a function of $\underset{\sim}{\theta}$ and can not be ignored.

In this study, we focus on the binary response; see more examples in Yang *et al.* (2023) and details in Appendix B. We assume logistic regression with p covariates, including an intercept, for the population model,

$$y_i \mid \underset{\sim}{\beta} \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}{1 + e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}\right\}, i = 1, \ldots, N, \tag{10}$$

where N, the population size, and the non-sampled $\underset{\sim}{x}_i$ may be unknown, and these can come from an external source or from a calibrated bootstrap of the sampled covariates. Note that there are no survey weights in the population model.

For the sample, we have $(W_i, \underset{\sim}{x}_i, y_i), i = 1, \ldots, n$. Without the normalization,

$$f_1(y_i \mid \underset{\sim}{\beta}) = \left(\frac{e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}{1 + e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}\right)^{W_i y_i} \left(\frac{1}{1 + e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}\right)^{W_i(1-y_i)}, i = 1, \ldots, n, \tag{11}$$

independently, but note that this is not probability mass function in y_i, despite the fact that it is typically employed in model-based analysis of data derived from survey sampling. With normalization,

$$f_2(y_i|\underset{\sim}{\theta}) = \frac{(f_1(y_i|\underset{\sim}{\theta}))^{W_i}}{f_1(y_i=0|\underset{\sim}{\theta}))^{W_i} + f_1(y_i=1|\underset{\sim}{\theta}))^{W_i}} \tag{12}$$

$$= \frac{\left(\frac{e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}{1+e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}\right)^{y_iW_i}\left(\frac{1}{1+e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}\right)^{(1-y_i)W_i}}{\left(\frac{e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}{1+e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}\right)^{W_i} + \left(\frac{1}{1+e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}\right)^{W_i}} \tag{13}$$

$$= \left(\frac{e^{W_i\underset{\sim}{x}_i'\underset{\sim}{\beta}}}{1+e^{W_i\underset{\sim}{x}_i'\underset{\sim}{\beta}}}\right)^{y_i}\left(\frac{1}{1+e^{W_i\underset{\sim}{x}_i'\underset{\sim}{\beta}}}\right)^{(1-y_i)}. \tag{14}$$

Therefore,

$$y_i \mid \underset{\sim}{\beta} \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{e^{W_i\underset{\sim}{x}_i'\underset{\sim}{\beta}}}{1+e^{W_i\underset{\sim}{x}_i'\underset{\sim}{\beta}}}\right\}, i = 1,\ldots,n. \tag{15}$$

Once more, the Bayesian paradigm favors the normalized form (*i.e.* the covariates and weights) to provide coherent.

Then, using a flat prior on $\underset{\sim}{\beta}$, $\pi(\underset{\sim}{\beta}) = 1$, the joint pseudo-posterior density and the joint posterior density are,

$$\pi_1(\underset{\sim}{\beta} \mid \underset{\sim}{y}) \propto \left\{\frac{e^{\sum_{i=1}^n y_iW_i\underset{\sim}{x}_i'\underset{\sim}{\beta}}}{\prod_{i=1}^n(1+e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}})^{W_i}}\right\},\ \underset{\sim}{\beta} \in R^p, \tag{16}$$

$$\pi_2(\underset{\sim}{\beta} \mid \underset{\sim}{y}) \propto \left\{\frac{e^{\sum_{i=1}^n y_iW_i\underset{\sim}{x}_i'\underset{\sim}{\beta}}}{\prod_{i=1}^n(1+e^{W_i\underset{\sim}{x}_i'\underset{\sim}{\beta}})}\right\},\ \underset{\sim}{\beta} \in R^p. \tag{17}$$

For both posterior distributions, the prior is assumed to be a uniform distribution on $\underset{\sim}{\beta}$; see Chen *et al.* (2008) for more details about Jeffreys' prior.

Note that the Pólya–Gamma data augmentation for logistic regression (Polson *et al.*, 2013) cannot be applied generally in our setting. In the expression for $\pi_2(\underset{\sim}{\beta} \mid \underset{\sim}{y})$ in (17), we can rewrite $\underset{\sim}{x}_i'$ as $\underset{\sim}{x}_i$, in a form that allows the Pólya–Gamma augmentation to be used entirely within (17). However, this reparameterization is not possible for the expression in (16), and therefore the Pólya–Gamma data augmentation cannot be applied there.

In this application, for the sake of generality, we instead use a more flexible and general Metropolis–Hastings sampler to obtain samples of $\underset{\sim}{\beta}$; see Appendix A for more details. For both posteriors, large values of W_i will give unacceptably small variance, so it is necessary to replace the original weights with adjusted weights. After we replace the original survey weights with three adjusted weights in these two posteriors, we have six models, and they are illustrated in the following flowchart; see Figure 1.

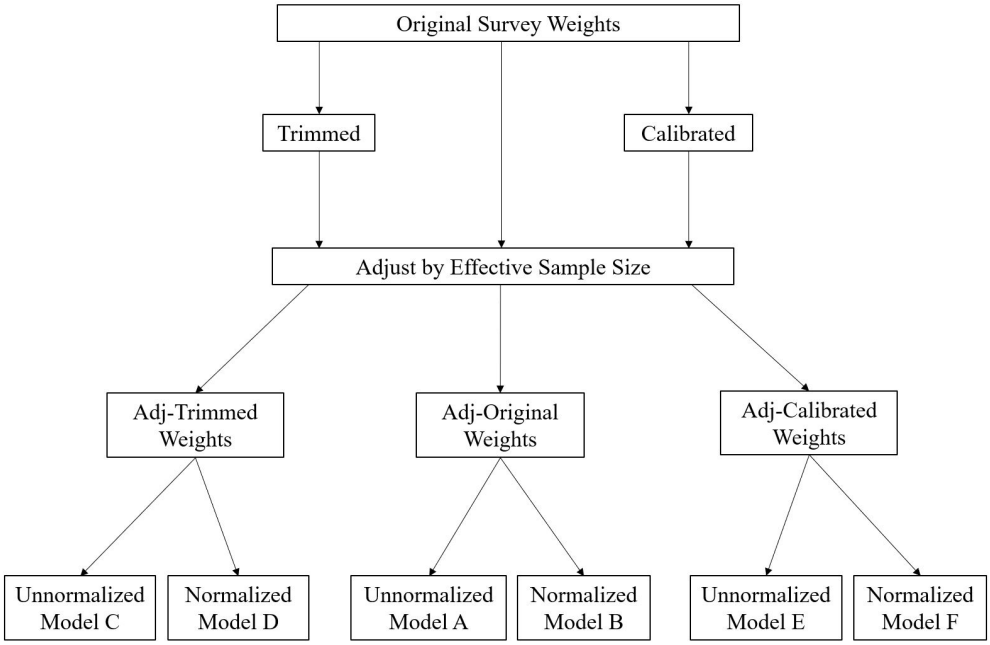

Figure 1: A flowchart of the six models

2.3. Bayesian predictive inference for logistic regression

To perform Bayesian predictive inference, we first specify a prior distribution for the model parameters $\beta_1, \beta_2, \ldots, \beta_{p-1}$. This prior can be chosen based on our prior beliefs or knowledge about the relationship between the predictor variables and the outcome. We then update this prior using Bayes' theorem to obtain the posterior distribution over the parameters given the observed data,

$$\pi(\underset{\sim}{\beta}|y, \underset{\sim}{x}) \propto \pi(y|\underset{\sim}{x}, \underset{\sim}{\beta})\pi(\underset{\sim}{\beta}),$$

where $\pi(y|\underset{\sim}{x}, \underset{\sim}{\beta})$ is the likelihood function, which specifies the probability of the observed data (y) given the predictor variables ($\underset{\sim}{x}$) and the model parameters ($\underset{\sim}{\beta}$), and $\pi(\underset{\sim}{\beta})$ is the prior distribution over the parameters. It can be shown that under mild conditions, $\pi(\underset{\sim}{\beta}|y, \underset{\sim}{x})$ is proper.

Bayesian predictive inference for logistic regression is a method used to make predictions about the probability of a binary outcome (*e.g.*, success/failure) based on a set of predictor variables. In this approach, we start with a prior distribution over the model parameters and update this distribution based on the observed data using Bayes' theorem. The resulting posterior distribution over the parameters can then be used by the stratification method to make predictive inferences about the population.

Specifically, we use a logistic regression model of the form,

$$\log \frac{p(y=1|x)}{1-p(y=1|x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1},$$

where $p(y = 1|x)$ is the probability of the binary outcome ($y = 1$) given the predictor variables ($\underset{\sim}{x}$), β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are the coefficients for the predictor variables

$x_1, x_2, \ldots, x_{p-1}$, and the logistic function maps the linear combination of predictor variables to a probability between 0 and 1. Once we have obtained the posterior distribution over the parameters, we can use it to make predictions about the probability of the binary outcome for new observations.

In this study, one of the primary challenges is the estimation of non-sampled variables. These are essential for predicting the proportion of the finite population but are typically unknown. To address this, we use a form of stratification to make predictive inferences about covariates of the population.

In this approach, the population is divided into K distinct strata or cells. Under the assumption that the sample covariates sufficiently represent the population, particularly when the sample size is large, unobserved variables can be structurally excluded, because the sampled covariates are likely to capture a broad spectrum of the population characteristics.

For the k-th stratum, the corresponding variables $\underset{\sim}{x}_k$ are known and every unit has the same covariates. Once the successful stratification of the population, the reliance on the estimation of non-sampled variables, which is a hard and resource-intensive task, is efficiently eliminated. Then, we can use the Horvitz-Thompson estimator for each stratum by aggregating the original survey weights of each unit in the stratum, denoted by $\hat{N}_k = \sum_{i \in \kappa} W_i$, where κ is the set of indexes in stratum k.

Surrogate sampling (Nandram, 2007) in Bayesian predictive inference involves obtaining predictions for a target parameter or quantity of interest by drawing samples from the posterior distribution of the model parameters. Instead of directly simulating from the predictive distribution, surrogate sampling simplifies the process by breaking it into two steps. Firstly, obtaining samples from the posterior distribution of the model parameters given the observed data by using Bayesian inference techniques, such as Markov chain Monte Carlo (MCMC) or other sampling methods. Secondly, for each set of sampled parameters, draw predictions for the quantity of interest using the population model. The population model represents the assumed distribution of the data given the model parameters. The term "surrogate" reflects the fact that the samples are used as substitutes for direct samples from the predictive distribution, and the whole population is sampled.

Therefore, according to the population distribution (Equation 10), we have,

$$t_k \mid \underset{\sim}{\beta} \overset{ind}{\sim} \text{Binomial}\left\{\hat{N}_k, \frac{e^{\underset{\sim}{x}_k'\underset{\sim}{\beta}}}{1+e^{\underset{\sim}{x}_k'\underset{\sim}{\beta}}}\right\}, k = 1, \ldots, K, \tag{18}$$

where t_k is the total of the binary outcome ($y = 1$) in this stratum, and K denotes the number of unique combinations of covariates. For example, in our study, we only focus on age (20-90), race (0 or 1), and sex (0 or 1), which means there are 284 combinations of these covariates in the population. The estimate of the finite population proportion is then given by,

$$\hat{T} = \frac{\sum_{k=1}^{K} t_k}{\sum_{k=1}^{K} \hat{N}_k}.$$

Moreover, this method, which is a form of stratified analysis, is particularly feasible when all covariates are discrete, each having only a few levels, or when a specific continuous

variable can be discretized to a limited number of levels. The discretization of continuous variables can simplify their handling, enabling the application of methods designed for discrete data. It also helps to reduce the complexity and computational load of the methods.

Furthermore, when auxiliary information is available in the form of known marginal counts, the raking approach can be used as an alternative and more robust procedure within strata. However, this auxiliary information is typically not readily available in most instances, and survey weights are normally used as we have done, so prediction using the already fitted logistic regression is preferred. For example, consider a single stratum with s successes and f failures, a sample size of $n = s + f$, with a stratum size of N. Then, raking allocates N units proportionally to get $(s/n)N$ and $(f/n)N$. But, of course, this proportionality assumption may not be true; see Deville *et al.* (1993) for general multi-way tables using iterative proportional fitting (IPF).

To summarize, in this approach, the population is divided into K distinct strata or cells. Under the assumption that the sample covariates sufficiently represent the population, particularly when the sample size is large, unobserved variables can be structurally excluded, because the sampled covariates are likely to capture a broad spectrum of the population characteristics. Then, we use the estimated sizes (sum of survey weights) for each stratum. By dividing sampled variables into distinct strata, we can efficiently make predictive inferences for a finite population. This is achieved using the covariates and the cumulative weights of each stratum, and the entire stratum has just one covariate vector. This stratification strategy enables us to effectively use sampled variables for estimation, thus eliminating the need for estimating non-sampled variables, simplifying the process, and bolstering the reliability of our results.

3. Simulation

In this section, we generate the finite population and the sample using the design-based method, where we also vary the strength of the association between the response and covariates, in accordance with Nandram and Rao (2021) and Chen *et al.* (2020). Age, race, and sex, the three covariates, determine the structure of the simulation data as follows:

$$z_i = 24.2449 + .0559x_{1i} + 1.2656x_{2i} + 1.2525x_{3i} + \epsilon_i, \tag{19}$$

$$y_i = \begin{cases} 1, & z_i \geq 30 \\ 0, & z_i < 30 \end{cases}, \quad i = 1 \ldots N, \tag{20}$$

where,

$$x_{1i} \overset{ind}{\sim} \text{DiscreteUniform}(20, 90),$$
$$x_{3i} \overset{ind}{\sim} \text{Bernoulli}(0.5),$$
$$x_{2i} \mid x_{1i}, x_{3i} \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{e^{a_i}}{1+e^{a_i}}\right\},$$
$$\epsilon_i \overset{iid}{\sim} \text{Normal}\left(0, \sigma^2\right),$$

with $a_i = \{24.2449 + .0559x_{1i} + 1.2525x_{3i}\}^{\frac{1}{10}}$. In order to control the correlation coefficient ρ between z_i and the linear predictor $\underset{\sim}{x}_i'\underset{\sim}{\beta}$ at 0.2, 0.3, 0.5, and 0.8 for the simulation studies,

the values of σ are selected as 6.8638, 4.4551, 2.4267, and 1.0508, respectively. A Monte Carlo check with $N = 10^6$ reproduces the target correlations to within 0.0003.

Then, the probability sample with the target sample size $n = 200$ is taken from the population with population size, $N = 10,000$, by the randomized systematic probability-proportional-to-size (PPS) sampling method with the selection probabilities,

$$\pi_i = \frac{nb_i}{\sum_{i=1}^{N} b_i}, \quad i = 1 \dots N,$$

with $b_i = \theta + x_{1i} + 0.2x_{2i} + 0.1x_{3i}$, and θ is chosen by trial and error to control the variation of the survey weights such that $\max\{b_i\} / \min\{b_i\} \approx 50$. After we use the PPS sampling method with $\underset{\sim}{\pi}$ to select samples, the survey weight of each unit in the sample is $W_i = 1/\pi_i, i = 1, \dots, n$.

Two scenarios are considered: scenario 1 using the generated samples and scenario 2 using the generated samples with manually added weights outliers. For scenario 2, we randomly select 10 samples where the response variable is 0 and multiply the corresponding weights by 5. These ten units are considered outliers and the outlier fraction is 5% when sample size is 200. Then, we can calculate these three adjusted weights based on $\underset{\sim}{W}$. The outliers are manually generated to create the bias that may be caused by nonresponse, oversampling of certain groups, or errors in the weight calculation process. We do not make any changes on the population. For the adjusted calibrated weights, assume the totals of the population are obtainable. Note that we include the survey weights in the logistic regression model because the key variables used in the model do not fully account for the complex survey design, such as the nonresponse issue. Nandram and Choi (2010) have shown that incorporating weights into the model can provide more accurate assessments of obesity and overweight in children and adolescents.

In this simulation, our parameter of interest is the finite population proportion. For a given estimator, its performance is evaluated through the absolute relative bias (ARB), the posterior standard deviation (PSD), the posterior root mean squared error (PRMSE), the coverage probability (CP), and the width of the highest posterior density interval (Wid) computed as

$$\text{ARB} = \frac{1}{H} \sum_{h=1}^{H} \mid \frac{PM^{(h)} - T}{T} \mid,$$

$$\text{PSD} = \frac{1}{H} \sum_{h=1}^{H} PSD^{(h)},$$

$$\text{PRMSE} = \frac{1}{H} \sum_{h=1}^{H} \sqrt{(PM^{(h)} - T)^2 + (PSD^{(h)})^2},$$

$$\text{CP} = \frac{1}{H} \sum_{h=1}^{H} I\left(C_{025}^{(h)} \le T \le C_{975}^{(h)}\right),$$

$$\text{Wid} = C_{975}^{(h)} - C_{025}^{(h)},$$

where T is the true finite population proportion, $PM^{(h)}$ is the posterior mean, $PSD^{(h)}$ is the posterior standard deviation, $I(\cdot)$ is the indicator function, $C_{025}^{(h)}$ and $C_{975}^{(h)}$ denote the

endpoints, not necessarily percentiles. All were obtained from the h^{th} simulated sample with $H = 1,000$ as the total number of simulation runs.

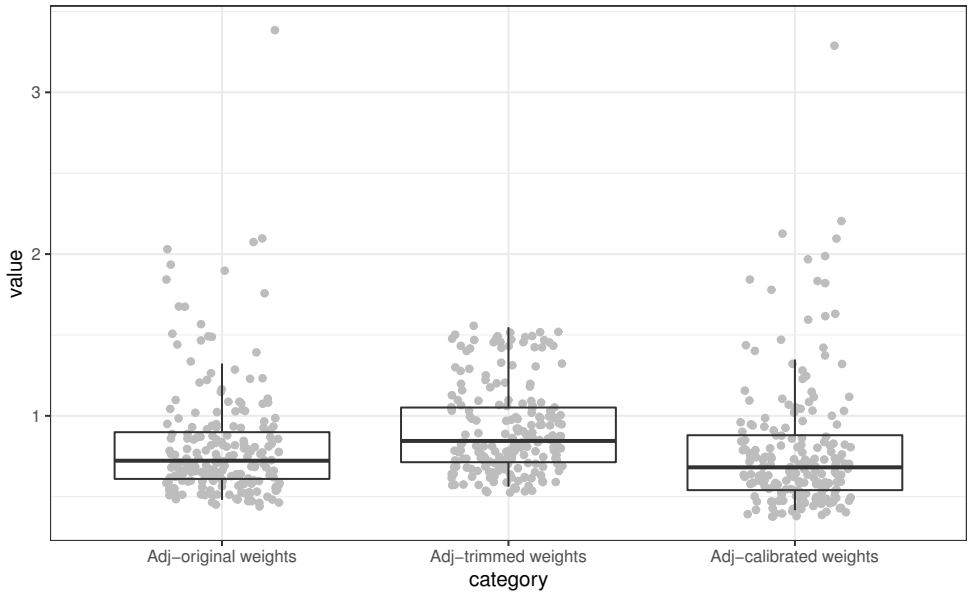

Figure 2: Comparison boxplots with three different weights with outliers using one simulation data when $n = 200$ and there is 5% weights outliers.

Figure 2 is the boxplot of one simulation run under scenario 2 with weights outliers, which indicates that the median, quartiles, and range of the adjusted calibrated weights are very similar to those of the adjusted original weights. Note that this calibration method could be useful in other cases when the weights need more adjustment. As we trimmed some weights outliers and reassigned weights to ensure $\sum_{i=1}^{n} W_i^* = \sum_{i=1}^{n} W_i = \hat{N}$, the boxplot of adjusted trimmed weights is a little higher and more consistent than others, without any outliers. Their corresponding effective sample sizes are $n_e = 162, n_e^* = 184, \tilde{n}_e = 147$ out of the real sample size $n = 200$ in this run.

Table 1 displays simulation comparisons in the absence of weights outliers. When there are no outliers of weights in the selected samples, models with unnormalized densities (model A, C, and E) outperform models with normalized densities (model B, D, and F). However, in surveys with complex designs, large-scale surveys covering diverse populations, or surveys experiencing high rates of nonresponse or measurement error, we must consider the occurrence of survey weights outliers (Nandram and Choi, 2010).

Table 2 presents results under the scenario with 5% weights outliers. In this scenario, it is reasonable that the assessment metrics of the model ignoring weights are better in general compared to when both weights with outliers are incorporated. As we generate samples using the PPS sampling method and specifically introduce weights outliers to the sample weights, we focus on examining the discrepancies arising from different types of weights and their associated densities. It shows that model D appears to be a better choice compared to other models (A, B, C, E, and F) because it consistently demonstrates lower bias, better precision, and higher overall performance in estimating the parameter of interest. Models using unnormalized densities with adjusted original weights (model A) and adjusted

Table 1: Comparison of six models using simulation data with different correlations and weights without outliers

	ρ	no wgt	A	B	C	D	E	F
ARB	0.2	0.076	0.087	0.126	0.077	0.080	0.086	0.132
	0.3	0.076	0.082	0.130	0.076	0.079	0.082	0.128
	0.5	0.106	0.110	0.171	0.106	0.131	0.109	0.172
	0.8	0.096	0.096	0.127	0.096	0.126	0.097	0.134
PSD	0.2	0.027	0.031	0.041	0.025	0.025	0.032	0.043
	0.3	0.026	0.033	0.044	0.025	0.025	0.032	0.042
	0.5	0.023	0.029	0.033	0.022	0.023	0.029	0.035
	0.8	0.016	0.022	0.020	0.017	0.017	0.023	0.020
PRMSE	0.2	0.046	0.054	0.073	0.046	0.046	0.054	0.076
	0.3	0.043	0.050	0.073	0.042	0.043	0.049	0.071
	0.5	0.047	0.051	0.073	0.046	0.055	0.051	0.074
	0.8	0.034	0.038	0.044	0.034	0.042	0.038	0.046
CP	0.2	0.946	0.951	0.933	0.945	0.941	0.952	0.939
	0.3	0.956	0.965	0.948	0.962	0.941	0.958	0.960
	0.5	0.905	0.960	0.842	0.918	0.796	0.921	0.863
	0.8	0.910	0.953	0.819	0.923	0.810	0.919	0.841
Wid	0.2	0.107	0.124	0.161	0.100	0.100	0.127	0.168
	0.3	0.105	0.129	0.174	0.098	0.098	0.125	0.167
	0.5	0.092	0.113	0.133	0.088	0.093	0.114	0.139
	0.8	0.064	0.089	0.078	0.066	0.068	0.089	0.080

calibrated weights (model E) have similar results across all ρ levels, which suggests that there are no big differences between adjusted original weights $\underset{\sim}{w}$ and adjusted calibrated weights $\tilde{\underset{\sim}{w}}$, and the same situation for normalized densities with adjusted original weights (model B) and adjusted calibrated weights (model F). The models with adjusted trimmed weights (model C, and D), however, exhibit more substantial difference in these measurements, especially at lower ρ levels (0.2 and 0.3), with lower ARB, PRMSE, and higher CP. With the setting of weights outliers, the adjusted trimmed weights show their advantage in this simulation.

It is also notable that at lower ρ levels (0.2 and 0.3), normalized model with adjusted trimmed weights (model D) performs better than unnormalized model with adjusted trimmed weights (model C), where model D tends to have relatively low or close ARB values, and higher CP values.

In our context, outliers refer specifically to survey units with extreme weights, meaning weights that are much larger (or smaller) than those of the other sampled units. Such extreme weights are common in survey data and, in binary-outcome settings, can strongly influence predictive inference for the population.

To examine the impact of these weight outliers, we conduct a simulation in which we generate synthetic data under varying levels of weight contamination, with the fraction of outlier units ranging from 0 to 0.5. For each scenario, we randomly select a subset of sampled units according to the specified outlier fraction and inflate their original weights

Table 2: Comparison of six models using simulation data with different correlations and weights with 5% outliers

	ρ	no wgt	A	B	C	D	E	F
ARB	0.2	0.076	0.145	0.240	0.074	0.068	0.145	0.222
	0.3	0.094	0.149	0.253	0.092	0.088	0.151	0.237
	0.5	0.090	0.158	0.247	0.091	0.088	0.159	0.251
	0.8	0.116	0.176	0.195	0.124	0.114	0.176	0.210
PSD	0.2	0.028	0.036	0.045	0.026	0.026	0.035	0.044
	0.3	0.027	0.035	0.045	0.025	0.025	0.034	0.043
	0.5	0.023	0.034	0.037	0.023	0.024	0.032	0.036
	0.8	0.016	0.025	0.020	0.016	0.017	0.024	0.020
PRMSE	0.2	0.046	0.076	0.118	0.044	0.042	0.076	0.110
	0.3	0.049	0.073	0.117	0.048	0.046	0.074	0.110
	0.5	0.042	0.069	0.099	0.042	0.042	0.068	0.100
	0.8	0.039	0.058	0.061	0.041	0.039	0.058	0.065
CP	0.2	0.965	0.962	0.835	0.973	0.982	0.945	0.841
	0.3	0.962	0.941	0.840	0.966	0.968	0.881	0.808
	0.5	0.899	0.954	0.794	0.942	0.975	0.940	0.776
	0.8	0.820	0.910	0.648	0.809	0.872	0.853	0.641
Wid	0.2	0.110	0.143	0.177	0.103	0.102	0.138	0.173
	0.3	0.105	0.140	0.176	0.100	0.099	0.135	0.169
	0.5	0.092	0.134	0.147	0.089	0.093	0.125	0.141
	0.8	0.063	0.098	0.081	0.065	0.068	0.096	0.081

by a factor of five. We then apply our full procedure to obtain posterior means under six competing models and compare their performance using average relative bias (ARB).

The performance of six models at various outlier fraction levels is shown in Figure 3. When the outlier fraction is 0, the result shows that there are no significant differences between any of the six models. The models using adjusted trimmed weights (models C and D) perform better than the other models when the outlier fraction rises. This result was anticipated since weight trimming can reduce the influence of weights with outliers.

It is worth noting that when outlier fractions at a lower lever (0.1, 0.2, and 0.3), normalized density (model D) appears to have a lower ARB values than unnormalized density (model C). Nevertheless, the normalized densities (models B and F) have higher ARB values than the unnormalized densities (models A and E) across all the outlier fractions. There are no significant differences between unnormalized models using adjusted original weights (model A) and adjusted calibrated weights (model E), and the same applies to normalized models (models B and F).

4. Application on body mass index

The proportion of the population that has been identified as being obese is an important indicator to research when looking at the health of a finite population. In this section, we illustrate our six models by using sample BMI data of individuals from eight counties

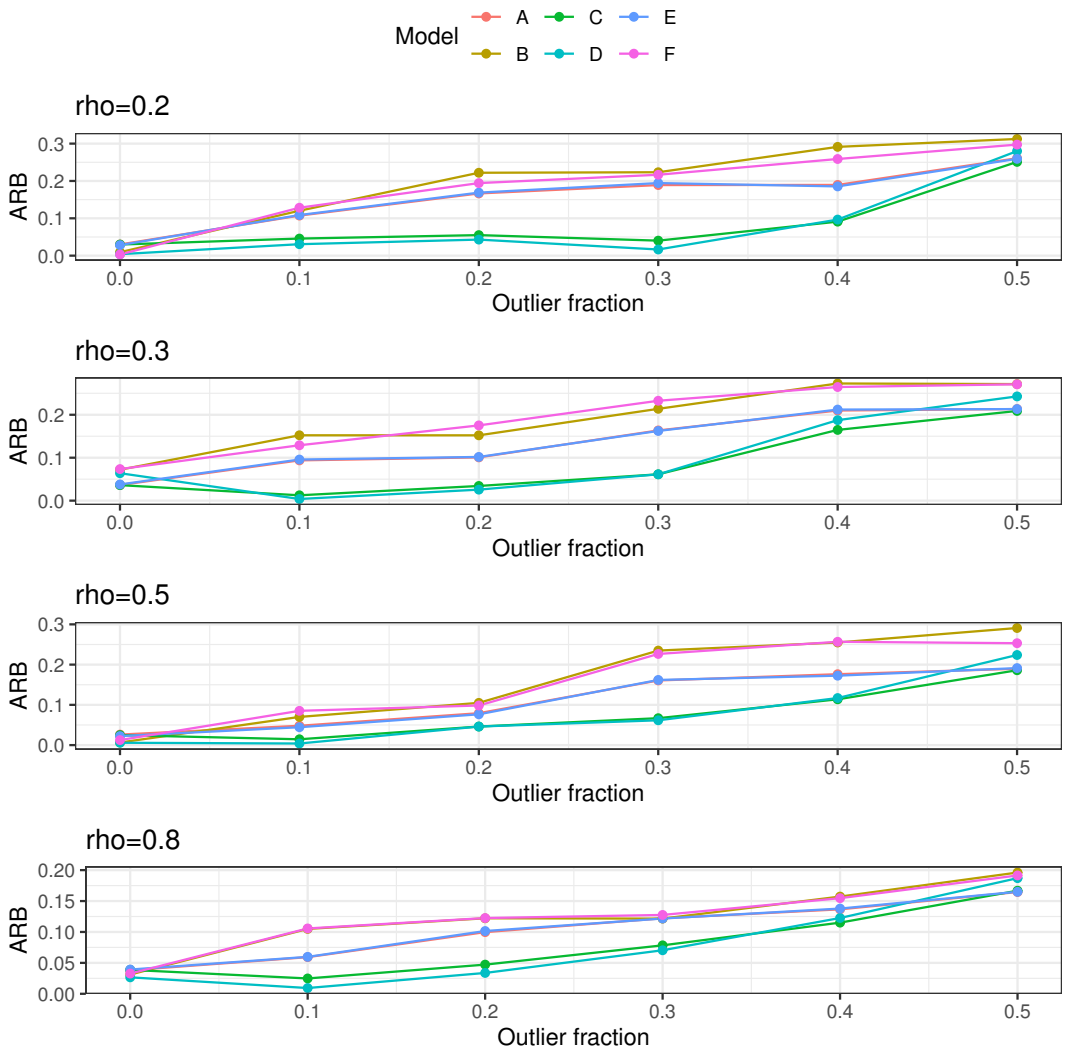

Figure 3: ARB Comparison of six models using simulation data with $\rho = 0.2, 0.3, 0.5, 0.8$ and different outlier fractions.

in California from the Third National Health and Nutrition Examination Survey (NHANES III). It contains 1,867 observations, which are large enough to create the contingency table to represent the finite population covariates if we only focus on age, race, and sex. The three covariates that are typically utilized with BMI data are discrete and include age, race, and sex. Age has about 71 levels (age runs from 20 to 90) and race and sex, each have two levels, so there are 284 distinct vectors $\underset{\sim}{x}$ of covariates.

For this example on eight counties in California, we use web scraping and the 1990 census report of California to obtain the totals of each covariate. From the Census, we can find the population size N, the average age, and race and sex proportions of the population, which are $N = 4,035,862$, average age $= 36.7$, race proportion $= .719$, sex proportion $= .497$. Then, we can make comparisons between the results of these various models.

First of all, the incorporation of jittered points in the boxplot may give the impression that some data points lie below the 0 line, however, it should be noted that all points are in

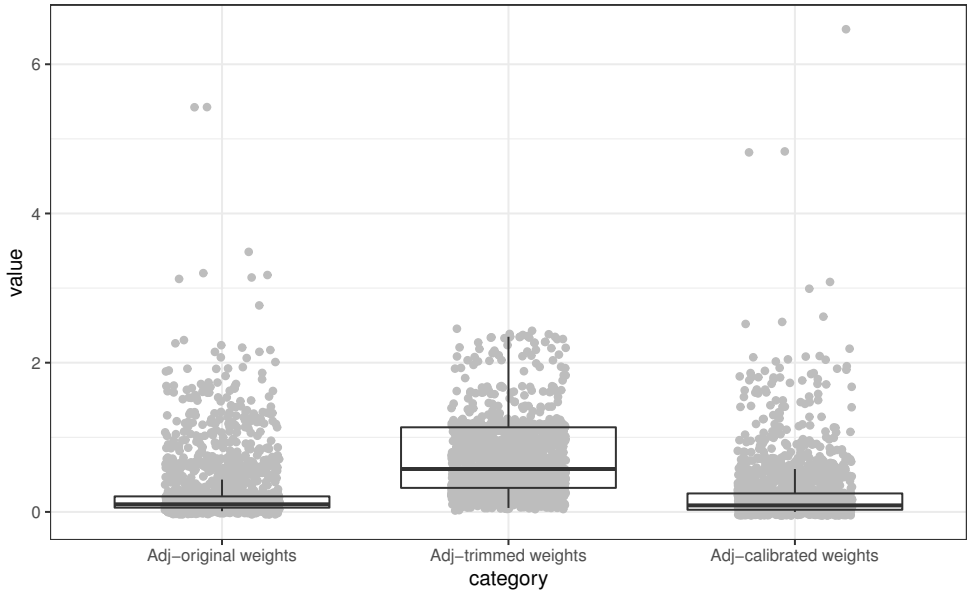

Figure 4: Comparison boxplot with jittered points of three different weights using BMI data

fact positive. In Figure 4, the boxplots of adjusted original weights and adjusted calibrated weights have a heavy tail, which means there may be some outliers of weights above the maximum boundary and can affect the estimates a lot. Their corresponding effective sample size are $n_e = 498, n_e^* = 1301, \tilde{n}_e = 498$.

As we know, outliers can have a significant impact on the effective sample size. For example, if a dataset with a large number of outliers is used to estimate the mean of a population, the outliers can pull the mean away from the true value, resulting in a larger variance and a smaller effective sample size. In this case, the effective sample size would be lower than the actual sample size, indicating that there is less useful information in the data for making inferences about the population. For the BMI dataset, the large amount of outliers decreases the two effective sample sizes derived by adjusted original weights and adjusted calibrated weights.

The distributions of PMs in Figure 5 show that these normalized models with adjusted original weights (model B) and adjusted calibrated weights (model F) are to be more to the right than other models. As a result, it is possible to challenge the accuracy of their estimates. To thoroughly investigate this issue, additional analysis of the table is required; see Appendix A for convergence tests.

In Table 3, we present posterior summaries of these six models of the finite population proportion of obesity. These six models are compared using four metrics, posterior mean (PM), posterior standard deviation (PSD), posterior coefficient of variation (PCV) and 95% credible interval. Based on the table, we can observe that for adjusted original weights and adjusted calibrated weights, the models with unnormalized densities (models A and E) offer more robustness against weights with outliers and provide more stable estimates of the population proportion in the presence of extreme weights than the models using normalized densities (model B and F). The proportion of obese individuals should be a little more around

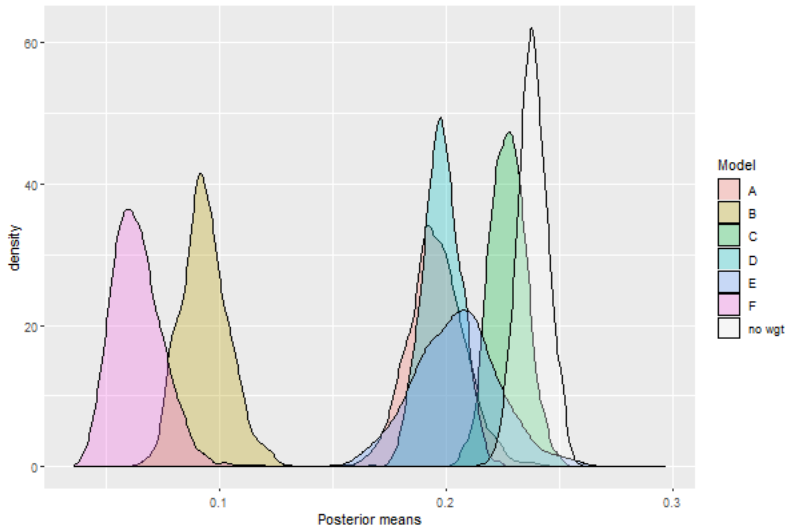

Figure 5: Comparison of densities of posterior means (PM) using BMI data

Table 3: Comparison of six models using BMI data

Model	PM	PSD	PCV	95% CI
no wgt	0.238	0.007	0.028	(0.226, 0.251)
A	0.196	0.017	0.089	(0.160, 0.232)
B	0.095	0.016	0.172	(0.065, 0.127)
C	0.226	0.012	0.055	(0.202, 0.251)
D	0.198	0.012	0.059	(0.175, 0.222)
E	0.195	0.018	0.092	(0.162, 0.233)
F	0.094	0.017	0.178	(0.063, 0.128)

0.2. If we exclude those weights, the posterior density of the population proportion shifts towards the right, resulting in a higher estimated population proportion.

An examination of responses associated with extreme weights reveals that most of them are 0's, which decreases the PMs of Models B and F to 0 in Table 3 and Figure 5. However, when we implement weight trimming, an approach designed to mitigate the influence of these weights with outliers, Models C and D have the higher PMs.

In addition, Models C and D exhibit the smallest PSD (0.012), implying the least amount of uncertainty in their estimates. This is a critical feature in statistical analysis, as reduced uncertainty typically translates to more reliable results. The low PCVs for Models C and D (0.055 and 0.059, respectively) further support their robustness, suggesting the lowest relative variability in their estimates. Contrastingly, when adjusted original weights and adjusted calibrated weights are used (as in Models A, B, E, and F), the resultant interval is wider than that of Models C and D. This interval being wider could potentially introduce a greater degree of uncertainty into the analysis; see Table 3 and Figure 5.

Figure 6 shows the posterior distributions of $\underset{\sim}{\beta}$. It is clear to state that when the sample size is small, the posterior densities of $\underset{\sim}{\beta}$ have a large variance. But when we used the normalized model in the small sample size cases, the distributions shrink a little bit. For

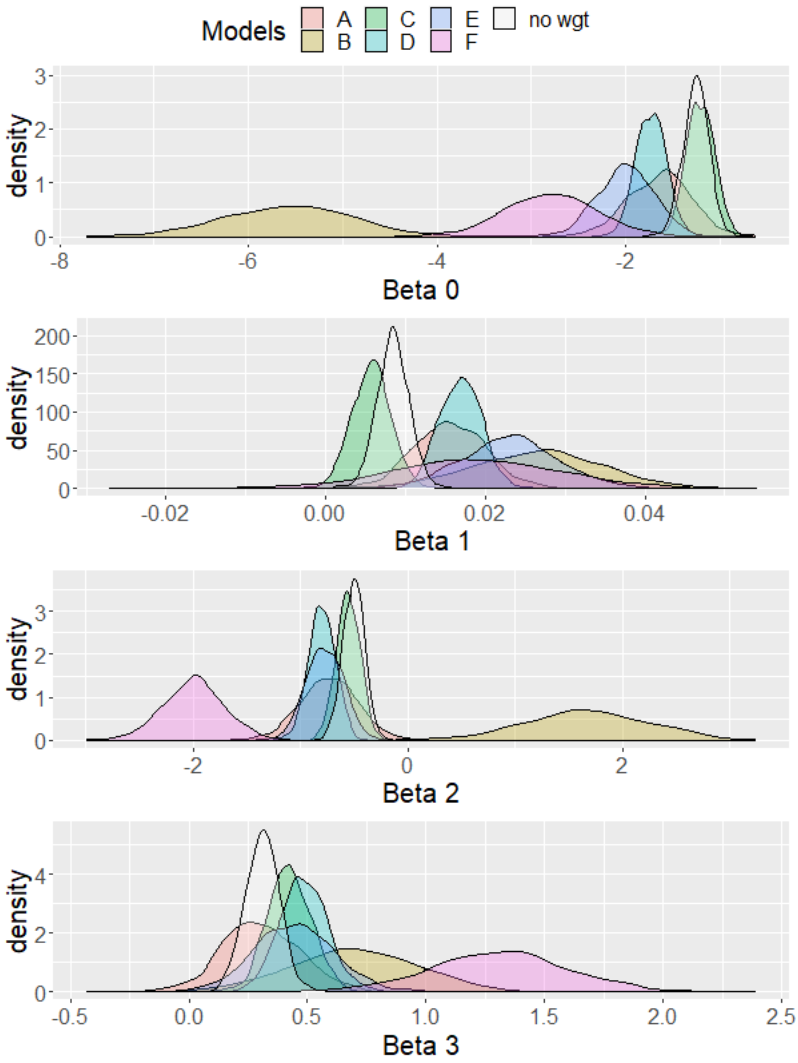

Figure 6: Comparison of posterior densities of β using BMI data

the large effective sample size, models with the adjusted trimmed survey weights, there is no significant difference between the unnormalized model and the normalized model.

5. Concluding remarks

Integrating the relevant auxiliary data, our study presents a framework to conduct Bayesian predictive inference on binary responses using probability survey samples. We introduce three distinct adjusted survey weights: the adjusted original survey weights, the adjusted trimmed survey weights, and the adjusted calibrated survey weights. These weights are integrated into both unnormalized and normalized densities within the Bayesian framework. Subsequently, the logistic regression model is implemented utilizing the Metropolis-Hastings sampler, and the predictive inference of finite population interests is made by surrogate sampling and the stratification approach.

The simulation and BMI dataset demonstrate the advantages of models incorporating adjusted trimmed weights into normalized density functions, particularly when the

correlation coefficient is small. These models not only excel in estimation accuracy but also adhere to the principles of the Bayesian paradigm, ensuring proper posterior distributions. In scenarios with exceedingly large population sizes, the normalized posterior that employs adjusted trimmed weights showcases increased robustness against weights with outliers, ultimately resulting in more precise estimations of the parameters of interest. Thus, such models should be carefully considered for their potential benefits when dealing with small correlation coefficients and large population sizes.

There will be many future works on the Bayesian predictive inference with covariates and survey weights. The proposed framework can be further expanded and applied to other types of response variables, such as categorical or continuous responses. It can also be adapted to handle more complex survey data, such as nonresponse and nonprobability sampling. One possible extension of this weighted logistic regression model is the Bayesian nonparametric priors based on the stick-breaking process, to increase model's robustness. The stick-breaking construction provides a rich framework for capturing complex patterns in data, making it a powerful tool in Bayesian nonparametric. Ren *et al.* (2011) introduced a logistic stick-breaking process as a non-parametric clustering method for general spatially- or temporally-dependent data, and the underlying assumption is that closely located data points are more likely to be grouped together. In survey sampling, the clusters often arises in complex survey sampling methods, especially those involving stratification and clustering.

Acknowledgments

I am indeed grateful to the Editors for their guidance and counsel. I am very grateful to the reviewer for valuable comments and suggestions of generously listing many useful references.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

APPENDIX A. Metroplis-Hastings algorithm for $\underset{\sim}{\beta}$

The Metropolis-Hastings algorithm is a Markov chain Monte Carlo (MCMC) method for generating samples from a target probability distribution that is difficult to sample from directly. The algorithm is widely used in Bayesian statistics and other applications where sampling from a complex distribution is necessary. The basic idea of the Metropolis-Hastings algorithm is to simulate a Markov chain whose stationary distribution is the target distribution of interest. The algorithm proceeds in a series of iterations, where at each iteration, a candidate state is proposed and accepted or rejected based on a acceptance probability. The acceptance probability is designed to ensure that the chain converges to the target distribution, even if the initial distribution is far from it.

In this paper, we use the Metropolis-Hastings algorithm to obtain a sample of $\underset{\sim}{\beta}$, with $\pi(\underset{\sim}{\beta} \mid \underset{\sim}{y})$ in (16) as the target density. Then, we need to set a candidate generating (proposal) density, denoted by $q(\underset{\sim}{\beta})$. The choice of proposal distribution can have a significant impact on the performance of the algorithm. Ideally, the proposal distribution should have high

probability density in regions where the target distribution has high probability density, to minimize the number of rejected proposals. However, it should also have some probability density in regions where the target distribution has low probability density, to allow the chain to explore the entire space.

It is reasonable to consider a normal approximation distribution when the target density is uni-modal. Specifically, by using the mode of our target density as the mean, $\hat{\underset{\sim}{\beta}}$, and the negative of the inverse-Hessian matrix at the mode as the covariance matrix, $\hat{\Sigma}$, we can obtain the normal distribution. Therefore,

$$\underset{\sim}{\beta} \mid \underset{\sim}{y} \overset{app}{\sim} \text{Normal}\left\{\hat{\underset{\sim}{\beta}}, \sigma^2\hat{\Sigma}\right\},$$

$$\frac{\gamma}{\sigma^2} \sim \chi^2_\gamma,$$

where γ is the degrees of freedom of the chi-square distribution. We also tune the degrees of freedom to ensure the jumping rate is between $25\% - 75\%$.

The algorithm works as follows,

1. Start with an initial value, $\underset{\sim}{\beta}_0$.
2. At each iteration t:
   (a) Generate a new state, $\underset{\sim}{\beta}_t$, from $q(\underset{\sim}{\beta})$.
   (b) Calculate the acceptance probability, $\alpha(\underset{\sim}{\beta}_t, \underset{\sim}{\beta}_{t-1}) = \min\{1, \frac{\pi(\underset{\sim}{\beta}_{t-1})q(\underset{\sim}{\beta}_t)}{\pi(\underset{\sim}{\beta}_t)q(\underset{\sim}{\beta}_{t-1})}\}$, where $\underset{\sim}{\beta}_{t-1}$ is the current state.
   (c) Generate a uniform random variable u from [0,1].
   (d) If $u \leq \alpha(\underset{\sim}{\beta}_t, \underset{\sim}{\beta}_{t-1})$, accept the proposed state and set $\underset{\sim}{\beta}_{t+1} = \underset{\sim}{\beta}_t$. Otherwise, reject the proposed state and set $\underset{\sim}{\beta}_{t+1} = \underset{\sim}{\beta}_{t-1}$.
3. Repeat step 2 until convergence.

For the simulation and BMI example, the degrees of freedom are set equal to 8, and we get jumping rates, $(0.468, 0.460, 0.461, 0.462, 0.469, 0.468)$, for all models in BMI. Additionally, we set the iteration times to 15,000 with burn-in at 5,000 and take every tenth value to have the posterior samples of parameters. Then, we use trace plots, auto-correlations, the Geweke test of stationary, and the effective sample sizes to check their convergence. In the model without survey weights, the acceptance rate is 0.466, and the Geweke test yields p-values of $(0.482, 0.253, 0.289, 0.446)$ for each parameter, all with corresponding effective sample sizes of 1,000.

Table 4 presents the results of a Geweke test and the Effective Sample Size (ESS) for different parameters ($\underset{\sim}{\beta}$) under six different models in the BMI application. In the Geweke test, a higher p-value suggests that the chain has converged well. In our table, all p-values are greater than 0.05, apart from β_4 of Model E (unnormalized density with adjusted calibrated weights). Effective Sample Size (ESS) is another diagnostic measure used in MCMC analysis.

Table 4: P-values and effective sample size for models A-F in BMI

	Model A		Model B	
	P-value	ESS	P-value	ESS
Beta 1	0.623	1000	0.620	1000
Beta 2	0.383	1000	0.492	1000
Beta 3	0.435	1000	0.908	1000
Beta 4	0.592	1000	0.926	1000
	Model C		Model D	
	P-value	ESS	P-value	ESS
Beta 1	0.296	1079	0.124	1000
Beta 2	0.078	902	0.421	1164
Beta 3	0.543	762	0.662	1000
Beta 4	0.302	1099	0.606	1000
	Model E		Model F	
	P-value	ESS	P-value	ESS
Beta 1	0.454	1096	0.447	1000
Beta 2	0.706	1000	0.537	1000
Beta 3	0.848	1000	0.069	1000
Beta 4	0.028	1000	0.907	1000

In our case, the ESS for all parameters of each model is very close to or exceeds 1000 (the length of $\beta_i, i = 1, \ldots, 4$), except β_3 of Model C (unnormalized density with adjusted trimmed weights), and a few more runs can do, but will virtually not change the predictive inference.

According to all these mentioned measurements, we can conclude that all models across all parameters have good convergence of the MCMC chains and the results obtained from these models can be considered reliable.

APPENDIX B. Illustration of impact of normalization constant

Consider the following models for y,

$$f\left(y \mid \theta\right) = \theta e^{-\theta y}, \quad \theta, y > 0.$$

Let $W > 0$ be a fixed real number. Then,

$$\left(f(y \mid \theta)\right)^W = \theta^W e^{-\theta W y}, \quad y > 0,$$

which is not a density function of y. In this way, the density function of y is,

$$\frac{(f(y|\theta))^W}{\int (f(y|\theta))^W dy} = \frac{\theta^W e^{-\theta W y}}{\int \theta^W e^{-\theta W y} dy}, \quad y > 0.$$

If we assume the prior of θ, $\pi(\theta) = 1, \theta > 0$, we have two forms of the posterior

distribution of θ,

$$\pi(\theta \mid y) \propto \theta^W e^{-\theta W y}, \quad \textit{i.e. } \theta \mid y \sim \text{Gamma}(W+1, Wy),$$
$$\pi(\theta \mid y) \propto \theta W e^{-\theta W y}, \quad \textit{i.e. } \theta \mid y \sim \text{Gamma}(2, Wy).$$

Obviously,

$$E_1(\theta \mid y) = \frac{W+1}{Wy} = (W+1)a, \quad E_2(\theta \mid y) = \frac{2}{Wy} = 2a,$$
$$Var_1(\theta \mid y) = \frac{W+1}{(Wy)^2} = (W+1)a^2, \quad E_2(\theta \mid y) = \frac{2}{(Wy)^2} = 2a^2,$$

where $a = 1/(W+1)$ is an irrelevant value.

If $W = 1$, there is no difference between the posterior distribution of unnormalized and normalized density. If $W < 1$, $E_1(\theta|y) < E_2(\theta|y)$ and $Var_1(\theta|y) < Var_2(\theta|y)$. If $W > 1$, the opposite situation holds. Now, it is clear when the normalization constant is related to the parameter, how it will affect our estimation.

References

Archer, K. J. and Lemeshow, S. (2006). Goodness-of-fit test for a logistic regression model fitted using survey sample data. *The Stata Journal*, **6**, 97–105.

Barasa, K. S. and Muchwanju, C. (2015). Incorporating survey weights into binary and multinomial logistic regression models. *Science Journal Applied Mathematics and Statistics*, **3**, 243–249.

Basu, D. (2010). An essay on the logical foundations of survey sampling, part one. In DasGupta, A., editor, *Selected Works of Debabrata Basu*, pages 167–206. Springer, New York, NY.

Chen, L. and Nandram, B. (2023). Bayesian logistic regression model for sub-areas. *Stats*, **6**, 209–231.

Chen, M.-H., Ibrahim, J. G., and Kim, S. (2008). Properties and implementation of Jeffreys's prior in binomial regression models. *Journal of the American Statistical Association*, **103**, 1659–1664.

Chen, Q., Elliott, M. R., Haziza, D., Yang, Y., Ghosh, M., Little, R. J., Sedransk, J., and Thompson, M. (2017). Approaches to improving survey-weighted estimates. *Statistical Science*, **32**, 227–248.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, **115**, 2011–2021.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, **88**, 1013–1020.

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

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, **22**, 153–164.

Haziza, D. and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, **32**, 206–226.

Lohr, S. (2007). Comment: Struggles with survey weighting and regression modeling. *Statistical Science*, **22**, 175–178.

Lohr, S. (2009). *Sampling: Design and Analysis*. Nelson Education.

Nandram, B. (2007). Bayesian predictive inference under informative sampling via surrogate samples. *Bayesian Statistics and its Applications*, **1**, 356–374.

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., Choi, J. W., and Liu, Y. (2021). Integration of nonprobability and probability samples via survey weights. *International Journal of Statistics and Probability*, **10**, 1–5.

Nandram, B. and 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*, **1**, 1568–1603.

Nandram, B. and Rao, J. (2024). Bayesian integration for small areas by supplementing a probability sample with a non-probability sample. *Statistics and Applications* {*ISSN 2454-7395 (online)*}, **22**, 343–374.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, **108**, 1339–1349.

Potter, F. J. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, volume 225230. American Statistical Association Washington, DC.

Potthoff, R. F., Woodbury, M. A., and Manton, K. G. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association*, **87**, 383–396.

Rader, K. A., Lipsitz, S. R., Fitzmaurice, G. M., Harrington, D. P., Parzen, M., and Sinha, D. (2017). Bias-corrected estimates for logistic regression models for complex surveys with application to the united states' nationwide inpatient sample. *Statistical Methods in Medical Research*, **26**, 2257–2269.

Rao, J. (1966). Alternative estimators in pps sampling for multiple characteristics. *Sankhyā: The Indian Journal of Statistics, Series A*, **28**, 47–60.

Ren, L., Du, L., Carin, L., and Dunson, D. B. (2011). Logistic stick-breaking process. *Journal of Machine Learning Research*, **12**.

Robbins, M. W., Ghosh-Dastidar, B., and Ramchand, R. (2021). Blending probability and nonprobability samples with applications to a survey of military caregivers. *Journal of Survey Statistics and Methodology*, **9**, 1114–1145.

Roberts, G., Rao, N., and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, **74**, 1–12.

Srivastava, S. K. (1967). An estimator using auxiliary information in sample serveys. *Calcutta Statistical Association Bulletin*, **16**, 121–132.

Yang, L., Nandram, B., and Choi, J. W. (2023). Bayesian predictive inference under nine methods for incorporating survey weights. *International Journal of Statistics and Probability*, **12**, 33–53.