Topics in Bayesian Design and Analysis for Sampling

Yutao Liu

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

©2021
Yutao Liu
All Rights Reserved

ABSTRACT

Topics in Bayesian Design and Analysis for Sampling

Yutao Liu

Survey sampling is an old field, but it is changing due to recent advancement in statistics and data science. More specifically, modern statistical techniques have provided us with new tools to solve old problems in potentially better ways, and new problems arise as data with complex and rich information become more available nowadays. This dissertation is consisted of three parts, with the first part being an example of solving an old problem with new tools, the second part solving a new problem in a data-rich setting, and the third part from a design perspective. All three parts deal with modeling survey data and auxiliary information using flexible Bayesian models.

In the first part, we consider Bayesian model-based inference for skewed survey data. Skewed data are common in sample surveys. Using probability proportional to size sampling as an example, where the values of a size variable are known for the population units, we propose two Bayesian model-based predictive methods for estimating finite population quantiles with skewed sample survey data. We assume the survey outcome to follow a skew-normal distribution given the probability of selection, and model the location and scale parameters of the skew-normal distribution as functions of the probability of selection. To allow a flexible association between the survey outcome and the probability of selection, the first method models the location parameter with a penalized spline and the scale parameter with a polynomial function, while the second method models both the location and scale parameters with penalized splines. Using a fully Bayesian approach, we obtain the posterior predictive distributions of the non-sampled units in the population, and thus the posterior distributions of the finite population quantiles. We show through simulations that our proposed methods are more efficient and yield shorter credible intervals with better coverage rates than the conventional weighted method in estimating finite population quantiles. We demonstrate

the application of our proposed methods using data from the 2013 National Drug Abuse Treatment System Survey.

In the second part, we consider inference from non-random samples in data-rich settings where high-dimensional auxiliary information is available both in the sample and the target population, with survey inference being a special case. We propose a regularized prediction approach that predicts the outcomes in the population using a large number of auxiliary variables such that the ignorability assumption is reasonable while the Bayesian framework is straightforward for quantification of uncertainty. Besides the auxiliary variables, inspired by Little and An (2004), we also extend the approach by estimating the propensity score for a unit to be included in the sample and also including it as a predictor in the machine learning models. We show through simulation studies that the regularized predictions using soft Bayesian additive regression trees (SBART) yield valid inference for the population means and coverage rates close to the nominal levels. We demonstrate the application of the proposed methods using two different real data applications, one in a survey and one in an epidemiology study.

In the third part, we consider survey design for multilevel regression and post-stratification (MRP), a survey adjustment technique that corrects the known discrepancy between sample and population using shared auxiliary variables. MRP has been widely applied in survey analysis, for both probability and non-probability samples. However, literature on survey design for MRP is scarce. We propose a closed form formula to calculate theoretical margin of errors (MOEs) for various estimands based on the variance parameters in the multilevel regression model and sample sizes in the post-strata. We validate the theoretical MOEs via comparisons with the empirical MOEs in simulations studies covering various sample allocation plans. The validation procedure indicates that the theoretical MOEs based on the formula aligns with the empirical results for various estimands. We demonstrate the application of the sample size calculation formula in two different survey design scenarios, online panels that utilize quota sampling and telephone surveys with fixed total sample sizes.

Table of Contents

List of Figures

List of Tables

Acknowledgments

My greatest gratitude goes to my dissertation advisors, Dr. Qixuan Chen and Dr. Andrew Gelman. Qixuan has always been so available whenever I need support or mentorship, and has witnessed every single step as I toddle along my way into statistical research. Andrew has inspired me via his blog and recorded presentations online since I was in college, even before graduate school. It has been a true privilege working with and learning from him.

Many thanks to Dr. Ying Wei, Dr. Thomas D'Aunno and Dr. Lauren Kennedy for being on my committee and offering helpful comments. Ying also offered me a research opportunity through which I gained training in high performance computing.

I am grateful to the Department of Biostatistics for the friendly environment and fellowship support, with acknowledgments to all faculty, fellow students and staff. I learn a little bit of life wisdom every single day.

I appreciate the research assistantships with Dr. Guohua Li, Dr. John Santelli and Dr. Jessica Justman. The collaborative research projects expose me to important scientific questions and interesting real data problems, and I also learn to work with public health researchers as an applied practicing statistician.

Special thanks to my intern mentors and intern fellows at Boehringer Ingelheim, Google and Facebook. I would like to thank Dr. Qiqi Deng for offering the internship opportunity when I was a first year Ph.D. student, before taking the qualify exams, so that I had the chance to see statistical research in industry early on. And the internship experience at Google and Facebook changes my mindset substantially.

Last but not least, I would like to thank friends and family for love and companion over the years.

Chapter 1

Introduction

1.1 Overview

Sample surveys are widely used to collect information about various characteristics of a finite population of interest. In complex surveys, units are selected with unequal probabilities and non-response presents in almost all surveys. Such survey design features and practical issues need to be considered to perform valid statistical inference. In this chapter, we review statistical methods for estimating population total with survey data drawn from a finite population, including both design-based and model-based inferential approaches. In the design-based approach, estimators can be constructed using either survey weights or model-assisted approaches that incorporate auxiliary information with the intention of improving efficiency. In model-assisted estimators, we also review methods using modern statistical learning techniques when more flexible methods other than linear model is desired or high-dimensional auxiliary information becomes a challenge. For the model-based approach, we review both super-population and Bayesian approaches but focus on Bayesian modeling in

various sampling schemes.

1.2 Statistical Analysis of Complex Surveys

1.2.1 Framework and Notation

In finite population scenarios, we consider a finite population of size $N < \infty$ and the set, consisted of units in the population, is denoted as $U = \{1, \ldots, N\}$. We are interested in a particular survey variable Y, representing some certain characteristic of the population, with value y_i corresponding to the ith unit. A survey sample s of size n is drawn from the finite population, in other words, $s \subset U$, according to a given sampling design $p(\cdot)$, where $p(s)$ is the probability of selecting sample s. For $i, j \in U$, the first-order probabilities of selection are given by $\pi_i = \Pr[i \in s] = \sum_{s \subset U: i \in s} p(s)$ and the second-order probabilities of selection are given by $\pi_{ij} = \Pr[i, j \in s] = \sum_{s \subset U: i,j \in s} p(s)$. Let I_i be the sample inclusion indicator of unit i, with value 1 if unit i is included in the sample and value 0 if otherwise. Without the presence of survey nonresponse, $I_i = 1$ if $i \in s$ and $I_i = 0$ if otherwise; and inclusion probabilities are equal to probabilities of selection $\Pr[I_i = 1] = \Pr[i \in s] = \pi_i$. Very often, we have information on other characteristics of the population, denoted by variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$, prior to conducting a survey. Lohr (2009) defines auxiliary variable as any variable available prior to sampling. Such auxiliary information can be used to improve survey inference. A subset of the auxiliary variables $\boldsymbol{Z} \subset \{x_1, x_2, \ldots, x_p\}$ could be used for survey design, e.g. size variable in probability proportional to size (PPS) sampling design. Such variables are called design variables. Usually the purpose of the survey is to make "descriptive" inference about the finite population quantity $Q(\boldsymbol{Y}, \boldsymbol{Z})$, a function of

$\boldsymbol{Y} = \{y_1, y_2, \ldots, y_N\}$ and potentially $\boldsymbol{Z}$, e.g. domain estimation. For now, we focus on inference on the population total $Q(\boldsymbol{Y}) = t_y = \sum_{i \in U} y_i = \sum_{i=1}^{N} y_i$, considering that more complex finite population quantities can be written as functions of t_y (Breidt and Opsomer, 2017).

1.2.2 Design-Based Inference

In the design-based approach, the reference distribution is the distribution of sample inclusion indicator I. The survey outcomes of the units $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_N\}$ are treated as fixed and inference is based on the statistical distribution of $\boldsymbol{I} = \{I_1, I_2, \ldots, I_N\}$. The statistical uncertainty that needs to be quantified comes from repeated sampling from the finite population using the sampling design. The statistical inference procedure consists of the following parts:

(a) choosing an estimator $\hat{Q} = \hat{Q}(\boldsymbol{Y}_s, \boldsymbol{I})$, a function of the observed survey outcomes $\boldsymbol{Y}_s = \{y_i\}_{i \in s}$ and sampling inclusion indicators, which enjoys certain statistical properties, e.g. unbiasedness or consistency, with respect to the distribution of $\boldsymbol{I}$.

(b) choosing a variance estimator $\hat{V} = \hat{V}(\boldsymbol{Y}_s, \boldsymbol{I})$ which is unbiased or approximately unbiased for the variance of $\hat{Q}$, with respect to the distribution of $\boldsymbol{I}$.

Then the point estimate is given by $\hat{Q}(\boldsymbol{Y}_s, \boldsymbol{I})$ and $(1-\alpha)$ level confidence interval could be constructed based on large sample approximation using the standard normal distribution, $(\hat{Q} - z_\alpha \sqrt{\hat{V}}, \hat{Q} + z_\alpha \sqrt{\hat{V}})$. For small sample scenarios, confidence intervals based on resampling methods are available for some sampling designs (Rao, Wu, and Yue, 1992; Rao and Wu, 1988).

1.2.2.1 Weighted Estimators without Auxiliary Information

Among estimators not utilizing auxiliary information, two weighted estimators play a central role in design based inference, the Horvitz and Thompson (1952) estimator and the Hájek (1971) estimator (Chen et al., 2017).

Note that, in the absence of survey nonresponse, inclusion probabilities are equal to probabilities of selection. Weighting the units in the sample using inverse of the inclusion probabilities, the Horvitz and Thompson (1952) estimator of finite population total t_y takes the form

$$\hat{t}_{y,\text{HT}} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in U} I_i \frac{y_i}{\pi_i}.$$

The estimator is design-unbiased in the sense that when taking expectation with respect to sampling design $p(\cdot)$, the following holds

$$\text{E}_p(\hat{t}_{y,\text{HT}}) = \text{E}_p\left(\sum_{i \in U} I_i \frac{y_i}{\pi_i}\right) = \sum_{i \in U} \text{E}_p(I_i) \frac{y_i}{\pi_i} = \sum_{i \in U} \pi_i \frac{y_i}{\pi_i} = \sum_{i \in U} y_i = \sum_{i=1}^{N} y_i = t_y.$$

The variance of the estimator is given by

$$\text{Var}_p(\hat{t}_{y,\text{HT}}) = \text{Var}_p\left(\sum_{i \in U} I_i \frac{y_i}{\pi_i}\right) = \sum_{i \in U} \sum_{j \in U} \text{Cov}_p(I_i, I_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j},$$

which involves second-order inclusion probabilities. For $\pi_{ij} > 0$, an unbiased estimator of the variance is

$$\hat{V}(\hat{t}_{y,\text{HT}}) = \hat{\text{Var}}_p(\hat{t}_{y,\text{HT}}) = \sum_{i \in U} \sum_{j \in U} I_i I_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} = \sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

(Sen, 1953).

It can be seen from the variance formula that the Horvitz and Thompson (1952) estimator could be very inefficient when some units have large values for survey outcome y_i's or very low probabilities of selection π_i's, leading to extreme weights, given a sampling design. Various approaches have been proposed to improve efficiency of weighted estimator, which will be discussed later. An alternative weighted estimator is the Hájek (1971) estimator which is given by

$$\hat{t}_{y,\text{HA}} = \frac{\hat{t}_{y,\text{HT}}}{\hat{N}_{\text{HT}}} N,$$

where $\hat{N}_{\text{HT}} = \sum_{i \in U} I_i/\pi_i$. The Hájek (1971) estimator is design-consistent.

1.2.2.2 Model-Assisted Estimators

In many settings, auxiliary variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ are available at the population level. Model-assisted estimators can be used to account for such auxiliary information when it is available. All model-assisted estimators in this section can be viewed as types of difference estimator under various working models (Breidt and Opsomer, 2017; Särdnal et al., 1992, chapter 6) . The difference estimator takes the following form

$$\hat{t}_{y,\text{diff}} = \sum_{i \in U} m(\boldsymbol{x}_i) + \sum_{i \in s} \frac{y_i - m(\boldsymbol{x}_i)}{\pi_i} = \sum_{i \in U} m(\boldsymbol{x}_i) + \sum_{i \in U} I_i \frac{y_i - m(\boldsymbol{x}_i)}{\pi_i},$$

where $m(\cdot)$ is some method that predicts y_i using $\boldsymbol{x}_i$ (Breidt and Opsomer, 2017). Note that the second term is the Horvitz and Thompson (1952) estimator of finite population total $t_y - \sum_{i \in U} m(\boldsymbol{x}_i)$ for the constructed population $\{y_1 - m(\boldsymbol{x}_1), y_2 - m(\boldsymbol{x}_2), \ldots, y_N - m(\boldsymbol{x}_N)\}$ and is, therefore, design-unbiased for $t_y - \sum_{i \in U} m(\boldsymbol{x}_i)$. Hence, the difference estimator is design-unbiased for finite population total t_y. In terms of variance of estimator, in design-based approach, only the second term is random under the sampling design, as the reference

distribution is the distribution of sample inclusion indicators. The variance of the difference estimator is given by

$$\begin{aligned}\mathrm{Var}_p(\hat{t}_{y,\mathrm{diff}}) &= \mathrm{Var}_p\left(\sum_{i\in U} I_i \frac{y_i - m(\boldsymbol{x}_i)}{\pi_i}\right) = \sum_{i\in U}\sum_{j\in U} \mathrm{Cov}_p(I_i, I_j)\frac{y_i - m(\boldsymbol{x}_i)}{\pi_i}\frac{y_j - m(\boldsymbol{x}_j)}{\pi_j} \\ &= \sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)\frac{y_i - m(\boldsymbol{x}_i)}{\pi_i}\frac{y_j - m(\boldsymbol{x}_j)}{\pi_j}.\end{aligned}$$

By weighting the difference instead of the original survey outcomes, the different estimator potentially improves efficiency when the method $m(\cdot)$ has good performance in predicting survey outcomes and the difference is small.

The method $m(\cdot)$ needs to be specified in the different estimator and model-assisted approach introduces a working model to predict survey outcomes y_i's with auxiliary information $\boldsymbol{x}_i$'s. The general form of the working model is written as

$$y_i = m(\boldsymbol{x}_i) + \epsilon_i,$$

where $\mathrm{E}(\epsilon_i) = 0$ and the survey outcomes y_i's are modeled as realizations from superpopulation. The working model does not have to be correctly specified to model the relationship as long as it has some predictive power for survey outcomes. With an observed survey sample, $m(\cdot)$ can be estimated with some estimator $\hat{m}(\cdot)$ and plugging $\hat{m}(\cdot)$ into the difference estimator leads to the model-assisted estimator

$$\hat{t}_{y,\mathrm{MA}} = \sum_{i\in U}\hat{m}(\boldsymbol{x}_i) + \sum_{i\in s}\frac{y_i - \hat{m}(\boldsymbol{x}_i)}{\pi_i} = \sum_{i\in U}\hat{m}(\boldsymbol{x}_i) + \sum_{i\in U} I_i\frac{y_i - \hat{m}(\boldsymbol{x}_i)}{\pi_i}.$$

Note that there is statistical uncertainty in $\hat{m}(\boldsymbol{x}_i)$ under the sampling design for $\hat{m}(\boldsymbol{x}_i)$ depends on the sample and, more specifically, the sample inclusion indicators. The estimator

could be further re-written in the following form

$$\begin{aligned}\hat{t}_{y,\mathrm{MA}} &= \sum_{i\in U}\hat{m}(\boldsymbol{x}_i) + \sum_{i\in U} I_i \frac{y_i - \hat{m}(\boldsymbol{x}_i)}{\pi_i} \\ &= \sum_{i\in U}\hat{m}_N(\boldsymbol{x}_i) + \sum_{i\in U} I_i \frac{y_i - \hat{m}_N(\boldsymbol{x}_i)}{\pi_i} + \sum_{i\in U}\left(\hat{m}(\boldsymbol{x}_i) - \hat{m}_N(\boldsymbol{x}_i)\right)\left(1 - \frac{I_i}{\pi_i}\right),\end{aligned}$$

where $\hat{m}_N(\cdot)$ is the population-level fit estimated using data from all units in the finite population. Note that $\hat{m}_N(\cdot)$ is fixed under the sampling design, as it does not depend on the sample inclusion indicators. The first part, consisted of the first two terms, is the different estimator based on the population-level fit $\hat{m}_N(\cdot)$ which is design-unbiased for the finite population total t_y, for any $\hat{m}_N(\cdot)$. Therefore, as long as the third term can be shown to be negligible relative to the difference estimator, asymptotic design-unbiasedness can be claimed for the model-assisted estimator.

Specifying a linear working model with heteroskedastic errors for the difference estimator yields the generalized regression estimator. More specifically, the model is written as

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim (0, \sigma_i^2),$$

where $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, for $i \neq j$. The population-level model fit can be obtained using weighted least squares

$$\hat{m}_N(\boldsymbol{x}_i) = \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}_N = \boldsymbol{x}_i^{'}\left(\sum_{j\in U}\frac{\boldsymbol{x}_j^{'}\boldsymbol{x}_j}{\sigma_j^2}\right)^{-1}\sum_{j\in U}\frac{\boldsymbol{x}_j y_j}{\sigma_j^2}.$$

Since only data from the sample is available, plugging in the Horvitz and Thompson (1952) estimator for the finite population totals yields

$$\hat{m}(\boldsymbol{x}_i) = \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}} = \boldsymbol{x}_i^{'}\left(\sum_{j\in s}\frac{\boldsymbol{x}_j^{'}\boldsymbol{x}_j}{\pi_j\sigma_j^2}\right)^{-1}\sum_{j\in s}\frac{\boldsymbol{x}_j y_j}{\pi_j\sigma_j^2}$$

and plugging in the estimated model fit leads to the generalized regression estimator

$$\hat{t}_{y,\text{GREG}} = \sum_{i\in U} \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}} + \sum_{i\in s} \frac{y_i - \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}}{\pi_i} = \sum_{i\in U} \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}} + \sum_{i\in U} I_i \frac{y_i - \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}}{\pi_i}.$$

The idea of choosing working models that better predict survey outcomes motivates the direction of research in moving beyond linear models. Flexible models that capture non-linearity have been considered. Statistical learning techniques have been used to construct model-assisted survey estimators. The motivations of using statistical learning techniques in survey inference include modeling potentially nonlinear relationship between survey variable and auxiliary variables as well as handling high-dimensional issues if a large number of auxiliary variables are available. The rest of the section consists a few such examples.

Kernel Methods. Kernel methods assume that the model is locally simple, constant or linear, but globally smooth and then estimate the local regression function using nearby points determined by a kernel weighting function. Breidt and Opsomer (2000) consider modeling $m(\cdot)$ as a smooth function of a single auxiliary variable X estimated by local polynomial regression. The smooth function is approximated locally at x_i by a q-th order polynomial, fitted at the finite population level via weighted least squares using weights given by a kernel function centered at x_i:

$$m_N(x_i) = (1, 0, \ldots, 0) \cdot (\boldsymbol{X}_{Ui}^{'}\boldsymbol{W}_{Ui}\boldsymbol{X}_{Ui})^{-1}\boldsymbol{X}_{Ui}^{'}\boldsymbol{W}_{Ui}\boldsymbol{y}_U$$

where $\boldsymbol{X}_{Ui} = [1 \quad x_j - x_i \quad \ldots (x_j - x_i)^q]_{j\in U}$, $\boldsymbol{W}_{Ui} = \text{diag}\left\{\frac{1}{h}K\left(\frac{x_j - x_i}{h}\right)\right\}_{j\in U}$ and $\boldsymbol{y}_U^{'} = [y_1, y_2, \ldots, y_N]$. Let $\boldsymbol{X}_{si} = [1 \quad x_j - x_i \quad \ldots (x_j - x_i)^q]_{j\in s}$, $\boldsymbol{W}_{si} = \text{diag}\left\{\frac{1}{h}K\left(\frac{x_j - x_i}{h}\right)\right\}_{j\in s}$ and $\boldsymbol{y}_s^{'} = [y_i]_{j\in s}$, the estimated model fit with a sample is given by

$$\hat{m}_{\text{LPR}}(x_i) = (1, 0, \ldots, 0) \cdot (\boldsymbol{X}_{si}^{'}\boldsymbol{W}_{si}\boldsymbol{X}_{si})^{-1}\boldsymbol{X}_{si}^{'}\boldsymbol{W}_{si}\boldsymbol{y}_s.$$

Breidt and Opsomer (2000) also show that the survey estimator of finite population total can be written as weighted average of survey outcomes $\hat{t}_{y,\text{LPR}} = \sum_{i\in s} w_{is} y_i$ with weights w_{is} independent of y and is calibrated to powers of x,

$$\sum_{i\in s} w_{is} x^l = \sum_{i\in U} x^l \quad (l = 0, 1, \ldots, p).$$

Alternatively, Breidt, Opsomer, Johnson, and Ranalli (2007) consider a working model the mean of which is a semiparametric additive model. Disadvantages of kernel-based methods include the difficulties of adapting the kernel to incorporate multiple covariates, especially the combination of categorical and continuous variables.

Penalized Spline. Breidt, Claeskens, and Opsomer (2005) consider penalized spline functions of covariates and control model complexity via penalization/regularization. The working model is a linear mixed model

$$y_i = \beta_0 + \sum_{l=1}^{p} \beta_l x_i^l + \sum_{k=1}^{K} b_k (x_i - m_k)_+^p + \epsilon_i$$
$$\boldsymbol{b} \sim N(\mathbf{0}, \lambda^{-1} \boldsymbol{I}_K)$$

where the constants $m_1, \ldots, m_K$ are K selected fixed knots and λ is chosen a priori to give specified degrees of freedom in the smooth. As $\lambda \to 0$, the model becomes a piecewise pth-order polynomial, while as $\lambda \to \infty$, the model approaches a global pth-order polynomial. Similar to local polynomial regression survey estimator, the penalized spline survey estimator can be also written in weighted form, with weights independent from the survey outcomes. The weights are calibrated to the powers of x, $\{x^l\}_{l=0}^p$, but not the truncated polynomial basis functions. The survey asymptotic is discussed by McConville and Breidt (2013).

Neural network is a very popular statistical learning method that handles nonlinearity by specifying the mean response using nonlinear functions of new covariates derived as linear

combinations of original covariates. Montanari and Ranalli (2005) develop a model-assisted estimator using a feedforward neural network with skip-layer connections,

$$m(\boldsymbol{x}_i) = \boldsymbol{x}_i^{'}\boldsymbol{\beta} + \sum_{j=1}^{M} \alpha_j a(\boldsymbol{\gamma}_j^{'}\boldsymbol{x}_i),$$

where $\boldsymbol{\gamma}_j^{'}\boldsymbol{x}_i$ are derived new covariates, $a(\cdot)$ is a known nonlinear activation function and $\boldsymbol{\beta}$, $\{\alpha_j\}_{j=1}^{M}$, $\{\boldsymbol{\gamma}_j\}_{j=1}^{M}$ are unknown parameters to be estimated. Design consistency and asymptotic normality of the model-assisted estimator is proven.

LASSO. The least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996) has been widely used since developed. The method simultaneous performs variable selection and regularization using $L1$ penalty. Mcconville, Jay Breidt, Lee, and Moisen (2017) consider a linear working model with homogeneous variance and high-dimensional auxiliary variables and propose survey-weighted lasso estimator for the regression coefficients

$$\hat{\boldsymbol{\beta}}_s^{(L)} = \text{argmin}_{\beta}(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta})^T\boldsymbol{\Pi}_s^{-1}(\boldsymbol{Y}_s - \boldsymbol{X}_s\boldsymbol{\beta}) + \lambda\sum_{i=1}^{p}|\beta_i|,$$

where $\boldsymbol{\Pi}_s = \text{diag}(\pi_j)_{j\in s}$. Design consistency was established for the LASSO survey regression estimator of finite population total.

1.2.3 Model-Based Inference

In the model based approach, the survey outcomes $\{y_1, y_2, \ldots, y_N\}$ are viewed as realizations of random variables $\{Y_1, Y_2, \ldots, Y_N\}$. A statistical model is specified to model the (random) survey variable and to predict survey outcomes for units not included in the sample. Inference of finite population quantities is viewed as a prediction problem. The finite population total

t_y can be partitioned into two terms

$$t_y = \sum_{i \in U} y_i = \sum_{i \in s} y_i + \sum_{i \in U/s} y_i,$$

and the problem of estimating finite population total can be solved by predicting the survey outcomes $\{y_i\}_{i \in U/s}$. A probability distribution $p(Y|\boldsymbol{Z}, \boldsymbol{\theta})$ indexed by parameter $\boldsymbol{\theta}$ is specified for predictive purpose and inference is based on the joint distribution of survey variable Y and sample inclusion indicator I. Rubin (1976) demonstrates that, under probability sampling, inference can be based on the distribution of survey variable Y alone, as long as the design variables $\boldsymbol{Z}$ are included in the model and the distribution of I given Y is independent of the distribution of Y conditional on the design variables, $p(I|Y, \boldsymbol{Z}) = p(I|\boldsymbol{Z})$. The model-based approach includes two variants, super-population modeling and Bayesian modeling.

1.2.3.1 Super-Population Modeling

In the super-population model-based approach, the population survey outcomes $\{Y_i\}_{i \in U}$ are assumed to be a random sample from a super-population model. A probability distribution $p(Y|\boldsymbol{Z}, \boldsymbol{\theta})$ index by parameter $\boldsymbol{\theta}$ is specified for the survey variable (Little, 2004). The underlying assumption is that the model holds for both the population and the sample.

Example 1: Hospital Discharges. Consider estimation of total number of patients discharged (Y) during a given month in all the hospitals in the sampling frame consisting a finite population. For each hospital i in a sample drawn from the finite population, we observe number of patients discharged y_i and number of beds x_i. It is reasonable to believe that the number of patient discharged is roughly proportional to the number of beds in each

hospital and the belief naturally leads to the following model specification

$$\mathrm{E}_M[Y_i] = \beta x_i, \quad \mathrm{Var}_M(Y_i) = \sigma^2 x_i, \quad i = 1, \ldots, N$$

where $\mathrm{Cov}_M(Y_i, Y_j) = 0$, for $i \neq j$. Under the model above, the best linear unbiased estimator (BLUE) of β can be obtained via Weighted Least Squares (WLS) estimator

$$\hat{\beta} = \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i},$$

and the model-based estimator is given by

$$\hat{t_y} = \sum_{i \in s} y_i + \sum_{i \in U/s} \hat{y}_i = \sum_{i \in s} y_i + \sum_{i \in U/s} \hat{\beta} x_i = \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i} \sum_{i \in U} x_i = \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i} t_x,$$

which is the ratio estimator. Given a sample, the estimator is unbiased under the model specification in the sense that

$$\mathrm{E}_M[\hat{t_y} - t_y] = \mathrm{E}_M[\sum_{i \in U/s} \hat{\beta} x_i - \sum_{i \in U/s} y_i] = 0.$$

(Royall, 1992)

1.2.3.2 Bayesian Modeling

Bayesian modeling requires prior specification on the parameters and inference of finite population quantities is based on the posterior predictive distribution of the survey variables for non-sampled units $p(\boldsymbol{Y}_{ns}|\boldsymbol{Y}_s, \boldsymbol{Z})$, where $\boldsymbol{Y}_s = \{Y_i\}_{i \in s}$ consists of survey outcomes in the selected sample and $\boldsymbol{Y}_{ns} = \{Y_i\}_{i \in U/s}$ consists of survey outcomes in the reminder of the population. Specification of the prior distribution can be achieved by specifying a model $p(Y|\boldsymbol{Z}, \boldsymbol{\theta})$ on the survey variable Y indexed by parameter $\boldsymbol{\theta}$ and a prior distribution $p(\boldsymbol{\theta}|\boldsymbol{Z})$

on the parameters. With observations in sample s, the model can be fitted and the posterior distribution of $\boldsymbol{\theta}$ computed via Bayes's theorem

$$p(\boldsymbol{\theta}|\boldsymbol{Y}_s, \boldsymbol{Z}) = \frac{p(\boldsymbol{Y}_s|\boldsymbol{Z}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{Z})}{\int p(\boldsymbol{Y}_s|\boldsymbol{Z}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{Z})\, d\boldsymbol{\theta}}.$$

Therefore, the posterior predictive distribution is given by

$$p(\boldsymbol{Y}_{ns}|\boldsymbol{Y}_s, \boldsymbol{Z}) \propto \int p(\boldsymbol{Y}_{ns}|\boldsymbol{Y}_s, \boldsymbol{Z}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{Y}_s, \boldsymbol{Z})\, d\boldsymbol{\theta},$$

which induces the posterior distribution of finite population quantity $p(Q(\boldsymbol{Y}, \boldsymbol{Z})|\boldsymbol{Y}_s, \boldsymbol{Z})$. (Little, 2004) The model formulation does not include the sample inclusion indicator I, which is justified when the sampling mechanism is ignorable given the design variables $\boldsymbol{Z}$, $p(\boldsymbol{I}|\boldsymbol{Y}, \boldsymbol{Z}) = p(\boldsymbol{I}|\boldsymbol{Z})$ which leads to $p(\boldsymbol{Y}_{ns}|\boldsymbol{Y}_s, \boldsymbol{Z}, \boldsymbol{I}) = p(\boldsymbol{Y}_{ns}|\boldsymbol{Y}_s, \boldsymbol{Z})$ (Gelman et al., 2014; Rubin, 1983). Actually the condition holds for all probability sampling design, which avoids the need to specify a model for sampling mechanism (Rubin, 1983). Below we review Bayesian model-based inference under various sampling designs.

Example 2: Bayesian Model-Based Inference for the Mean from a Stratified Random Sample. If the survey variable of interest takes very different values in different subpopulations, stratified random sampling is considered to improve precision of the estimates of population quantities. The population is partitioned into H disjoint strata so that the survey variable takes similar value within each stratum and then a simple random sample of size n_h is independently taken within each stratum. Therefore, parametric models with distinct parameters assigned to different strata are considered to reflect strata differences for such stratified samples. Denote Y_{hi} the survey outcome for unit i in stratum h. A common baseline model for continuous outcome assumes normal distributions for Y_{hi} with distinct

parameters mean μ_h and σ_h^2 for strata $h = 1, \ldots, H$. A simple Bayesian non-informative prior specification leads to the following model

$$p(y_{hi}|z_{hi} = h, \mu_h, \sigma_h^2) \overset{\text{iid}}{\sim} N(\mu_h, \sigma_h^2), \quad p(\mu_h, \log \sigma_h^2) = \text{const.}$$

With known variances $\{\sigma_h^2\}_{h=1}^H$, standard Bayesian calculations indicate that the posterior distribution of finite population mean $\bar{Y}$ given $\boldsymbol{Y}_s$, $\boldsymbol{I}$ and $\{\sigma_h^2\}_{h=1}^H$ is normal with mean and variance

$$\mathrm{E}(\bar{Y}|\boldsymbol{Y}_s, \boldsymbol{I}, \sigma_h^2) = \sum_{h=1}^{H} \frac{N_h}{N} \left(\frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \right),$$

$$\mathrm{Var}(\bar{Y}|\boldsymbol{Y}_s, \boldsymbol{I}, \sigma_h^2) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \sigma_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right),$$

where N_h is the sub-population size and n_h is the sample size in stratum h. Note that the posterior mean is the stratified mean from design-based inference and, if replacing $\{\sigma_h^2\}_{h=1}^H$ with sample variance $\{s_h^2\}_{h=1}^H$ in each stratum, the posterior variance equals the design-based variance. The fully Bayesian inference quantifies the uncertainty in estimating the variances $\{\sigma_h^2\}_{h=1}^H$ by integrating out of the posterior distribution of $\bar{Y}|\boldsymbol{Y}_s, \boldsymbol{I}, \sigma_h^2$ over the posterior of distribution $\sigma_h^2|\boldsymbol{Y}_s, \boldsymbol{I}$.

Example 3: A Non-robust Model for Disproportionate Stratified Sampling. If the model fails to differentiate the strata by assuming the same distribution for all strata, the validity of inference is compromised. In the setting of stratified sampling, assigning same parameters for all strata leads to the following misspecification

$$p(y_{hi}|z_{hi} = h, \mu, \sigma^2) \overset{\text{iid}}{\sim} N(\mu, \sigma^2), \quad p(\mu, \log \sigma^2) = \text{const.}$$

The posterior mean under the model is unweighted sample mean

$$\mathrm{E}(\bar{Y}|\boldsymbol{Y}_s, \boldsymbol{I}, \sigma^2) = \sum_{h=1}^{H} \frac{n_h}{n} \left(\frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \right) = \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} y_{hi},$$

which is biased if the probabilities of selection vary across the strata. Therefore, attention needs to be drawn to the limitation of model-based inference that it is subject to bias if the model is miss-specified.

Example 4: Bayesian Model-Based Inference for Probability Proportional to Size Sample. When units in a survey population are of very different sizes, units with large sizes often contribute more to population quantities than units with smaller sizes. In probability proportional to size (PPS) sampling design, units of large sizes are selected with higher probabilities by assigning the probability of selection π is proportional to the value of a size variable X available for all population units. Such design can be considered for efficient estimation of population mean of a survey variable if the variance of which increases with size of the unit. Consider a finite population of size N, a PPS sample of size n is drawn by assigning probabilities of selection $\pi_i = nX_i / \sum_{i=1}^{N} X_i$ to unit i. The size variable here is a design variable and, therefore, should be included in the model to construct model-based predictive estimators. Model-based estimators, as discussed above, are subject to bias when the underlying model is miss-specified and such limitation motivates the development of flexible models that are robust against model miss-specification. As a Bayesian extension of Zheng and Little (2003), Chen et al. (2010) propose a Bayesian penalized spline predictive (BPSP) estimator for finite population proportion in unequal probability sampling. Denote Y a binary survey variable of interest and $p = N^{-1} \sum_{i=1}^{N} Y_i$ be the population quantity, the proportion of population units for which $Y = 1$. A probit truncated polynomial penalized

spline regression model is considered

$$\Phi^{-1}(\mathrm{E}(Y_i|\boldsymbol{\beta}, \boldsymbol{b}, \pi_i)) = \beta_0 + \sum_{l=1}^{p} \beta_l \pi_i^l + \sum_{k=1}^{K} b_k (\pi_i - m_k)_+^p$$

$$\boldsymbol{b}|\tau_b^2 \sim N(\mathbf{0}, \tau_b^2 \boldsymbol{I}_K)$$

where $\Phi^{-1}(\cdot)$ denotes the inverse CDF of standard normal distribution and the constants $m_1, \ldots, m_K$ are K selected fixed knots. The function $(\pi_i - m_k)_+^p$ is called truncated polynomial spline basis function with power p, where $(u)_+^p = \{u \times I(u > 0)\}^p$ for $u \in \mathbb{R}$. Penalty is imposed by specifying a normal distribution for the coefficients $\boldsymbol{b} = (b_1, \ldots, b_K)'$ for truncated polynomials, which is equivalent to smoothing via penalized likelihood. In fully Bayesian inference, a weak informative prior $N(0, \varphi^2 = (10^3)^2)$ is specified for the polynomial coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ and an inverse-Gamma distribution or improper uniform prior for the variance component for the truncated polynomial coefficients τ_b^2 and the posterior predictive distribution of finite population proportion is obtained by generating a large number of draws using Markov chain Monte Carlo simulations.

Zangeneh and Little (2015) consider Bayesian finite population inference of finite population total when only summary information of the aggregate size of non-sample units is available.

While stratified sampling and probability proportional to size sampling is limited to design variables that are known prior to survey design and data collection, post-stratification provides a way to combine data collected in the survey with aggregate data at population-level from other sources. Post-stratification can improve efficiency of survey estimates and can correct for bias, in the presence of differential nonresponse, by correcting for known differences between sample and population. In basic formulation, joint distribution of post-

stratification discrete variables $\boldsymbol{X}$ is known and the population can be partitioned into J subpopulations according to possible categories of $\boldsymbol{X}$, each category labeled as post-stratification cell j with population size N_j and sample size n_j. The total population size $N = \sum_{j=1}^{J} n_j$ and the sample size $n = \sum_{j=1}^{J} n_j$. The implicit assumption is that the data are collected by simple random sampling or, more generally, the relative probabilities of selection are equal, within each of the J post-strata. Further assume that the population size N_j of each post-stratification cell j is known. The population mean of any survey response can be written as

$$\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j},$$

where θ_j denotes subpopulation mean of each cell. And the corresponding estimates

$$\hat{\theta}^{\text{PS}} = \frac{\sum_{j=1}^{J} N_j \hat{\theta}_j}{\sum_{j=1}^{J} N_j}.$$

(Gelman, 2007; Little, 1993)

Example 4: Bayesian Multilevel Regression Post-Stratification. Gelman (2007) considers a continuous survey variable Y for the CBS/New York Times polls and assume a normal hierarchical regression model $\boldsymbol{Y} \sim N(X\boldsymbol{\beta}, \Sigma_y)$ with a prior distribution on regression coefficients $\boldsymbol{\beta} \sim M(0, \Sigma_\beta)$. The following predictors are included

- A constant term
- An indicator for sex (1 if female, 0 if male)
- An indicator for ethnicity (1 if black, 0 otherwise)
- Sex $\times$ ethnicity

- 4 indicators for age categories
- 4 indicators for education categories
- 16 age $\times$ education categories.

For simplicity, conditional independence is assumed for the components of $\boldsymbol{\beta}$ in the prior distribution, conditioning on the hyperparameters for the variance components. Therefore, the prior precision matrix Σ_β^{-1} is diagonal, with zeros for non-hierarchical regression coefficients (coefficients for the first 4 terms including the constant term) and 3 different parameters for the last 3 groups of coefficients. The regression coefficients can be estimated using posterior mean $\hat{\boldsymbol{\beta}} = (X^t\Sigma_y^{-1}X + \Sigma_\beta^{-1})^{-1}X^t\Sigma_y^{-1}\boldsymbol{Y}$. Denote X the $n \times k$ matrix of predictors in the data, X^{pop} the $J \times k$ matrix of predictors for the J post-stratification cells and label the vector of post-stratum populations as $N^{\text{pop}} = (N_1, \ldots, N_J)$. The Bayesian post-stratification estimator is given by

$$\hat{\theta}^{\text{PS}} = \frac{1}{N}(N^{\text{pop}})^t X^{\text{pop}}\hat{\boldsymbol{\beta}} = \frac{1}{N}(N^{\text{pop}})^t X^{\text{pop}} \times (X^t\Sigma_y^{-1}X + \Sigma_\beta^{-1})^{-1}X^t\Sigma_y^{-1}\boldsymbol{Y}.$$

Si, Pillai, and Gelman (2015) consider a scenario where inverse-probability weights are available for sample units only. They uses a hierarchical Bayesian approach to model the distribution of the weights of the nonsample units and simultaneously include the weights as predictors in a nonparametric Gaussian process regression.

In some applications, a large number of auxiliary variables are available with some variables potentially not significantly related to survey variable interest. At the same time, some of the the auxiliary variables are highly correlated. In such scenario, it is natural to consider

variable selection and shrinkage methods to improve the model when constructing survey estimator.

Shrinkage Priors for Sparse Bayesian Estimation. The spike-and-slab prior proposed by Mitchell and Beauchamp (1988) and George and McCulloch (1993) is often considered as the "gold standard" for sparse Bayesian estimation. For a p-dimensional vector of regression coefficients $\boldsymbol{\beta}$, the prior can be written as a two-component discrete mixtures

$$\beta_j|\lambda_j, c, \epsilon \sim \lambda_j N(0, c^2) + (1 - \lambda_j)N(0, \epsilon^2),$$

$$\lambda_j \sim \text{Ber}(\pi), \quad j = 1, \dots p,$$

where $\epsilon \ll c$ and the indicator variable $\lambda_j \sim \{0, 1\}$ denotes whether the coefficient β_j is close to zero ($\lambda_j = 0$) or nonzero ($\lambda_j = 1$). Carvalho, Polson, and Scott (2010) proposed a continuous horseshoe prior that is easy to implement and has been shown comparable performance to the spike-and-slab prior

$$\beta_j|\lambda_j, \tau \sim N(0, \tau^2\lambda_j^2),$$

$$\lambda_j \sim C^+(0, 1), j = 1, \dots, p.$$

The horseshoe is one of the so called global-local shrinkage priors. The global hyperparameter τ shrinks all the parameters towards zero while the local hyperparameters λ_j allow some coefficients to escape the shrinkage.

Chapter 2

Bayesian Inference of Finite Population Quantiles for Skewed Survey Data Using Skew-Normal Penalized Spline Regression

2.1 Introduction

Skewed data commonly arise in sample surveys. In such scenario, it is of more interest to draw inference of population quantiles, especially the lower and upper tails of the distribution, than the population mean. In this paper, we study the inference of population quantiles with skewed survey data from unequal probability sampling. Inference of finite population quantities can be either design-based or model-based, in which the survey design is incorporated into the statistical analysis in different ways (Kish, 1995; Little, 2004; Smith, 1976, 1994). In the design-based approach, the survey outcomes Y are treated as

fixed and inference is based on the distribution of the sample inclusion indicator I, a binary variable that indicates whether a unit is included in the survey sample (Cochran, 2007). In the model-based approach, a regression model is specified to model Y and inference is based on joint distribution of Y and I. Inference can be based on the distribution of Y alone, as long as I is independent of Y conditional on the design variables (Rubin, 1976).

Estimation of finite population quantiles is intimately tied to estimation of finite population distribution functions (Dorfman, 2009). From a design-based perspective, Kuk (1988) compared three estimators of distribution functions, including the Hájek estimator, the Horvitz-Thompson estimator, and the complementary proportion estimator; and gave theoretical reasoning for preferring the Hájek or complementary proportion estimators to the Horvitz-Thompson estimator. The Hájek estimator, which is design-consistent and approximately design-unbiased, is considered the "customary design-based estimator" and is usually the estimator against which other estimators of cumulative distribution functions are compared (Dorfman, 2009). Taking a model-based approach, Chambers and Dunstan (1986) (CD) estimated the distribution function utilizing auxiliary information and specifying a linear super-population model with heterogeneous variance through the origin. Dorfman and Hall (1993) modified the CD estimator by replacing the linear model with nonparametric regression model. To estimate the distribution of the CD and the nonparametric CD estimators, Lombardıa, González-Manteiga, and Prada-Sánchez (2003, 2004) proposed bootstrap methods in which bootstrap populations were constructed by sampling the empirical distribution of the re-centered residuals from the fitted super-population model. Kuk and Welsh (2001) further modified the CD estimator with robust estimation technique to

handle outliers and model misspecification, which fine-tunes for departure from the model assumed by estimating the conditional distributions of the residuals as a function of the auxiliary variable after fitting the working model to the sample. Rao, Kovar, and Mantel (1990) (RKM) proposed a design-based alternative to the model-based CD estimator and constructed model-assisted difference and ratio estimators of distribution functions with reference to the CD's linear working model through the origin. Based on asymptotic variances, Wang and Dorfman (1996) constructed a weighted average of the CD and RKM estimators, with weights derived to achieve minimal (asymptotic) mean square error of the resulting estimator. Kuk (1993) proposed a method combining the known distribution of the auxiliary variable with a kernel estimate of the conditional distribution of the survey variable given the auxiliary variable. Chambers, Dorfman, and Wehrly (1993) proposed a robust model-based estimator via nonparametric kernel smoothing. Chen, Elliott, and Little (2012) (probit-BPSP) proposed a Bayesian penalized spline model-based estimator that first estimates cumulative distribution functions at selected survey outcome values by fitting a series of probit penalized spline regression models on the inclusion probabilities, and then smooths the estimated cumulative distribution functions using a monotonic smooth cubic regression model.

Quantile estimators can be obtained by inverting the estimators of finite population distribution functions with monotonicity property. For example, the CD estimator and the probit-BPSP estimator of finite population distribution function can be inverted to obtain quantile estimators. Quantiles can also be estimated directly without requiring estimation of distribution functions. Rao, Kovar, and Mantel (1990) proposed simple ratio and difference

estimators of quantiles. Using Bayesian model-based approach, Chen, Elliott, and Little (2012) develop a Bayesian two-moment penalized spline predictive (B2PSP) estimator that predicts the values of non-sampled units based on a normal distribution, with mean and variance both modeled with penalized splines on the inclusion probabilities. The B2PSP estimator is more efficient than the Hájeck estimator, the CD estimator, and the RKM's ratio and difference estimators using simulation studies on artificially generated data. The B2PSP estimator is also more robust to model misspecification than the CD estimator when the conditional normality assumption is reasonable. However, the B2PSP estimator is potentially biased when the normality assumption is violated. Although in practice, transformation can be applied before modeling, data can still be skewed after transformation, and thus it is of great interest to develop more flexible methods for modeling skewed data.

In this paper, we consider inference of finite population quantiles in probability proportional to size (PPS) sampling. In PPS sampling design, information of a size variable is available for all units in the finite population at design stage and a sample of units are drawn with probabilities of selection proportional to the values of the size variable. Such design improves efficiency in estimating finite population quantities when variance of the survey outcome increases with size of the unit, as units of larger size are selected with higher probabilities. We propose two Bayesian model-based predictive estimators of finite population quantiles, assuming skew-normal distribution for the survey outcome of interest given the probability of selection. We assume that, at analysis stage, unit level information of the size variable is available for all units in the finite population.

2.2 Methods

2.2.1 Notation

Let Y denote a continuous survey outcome of interest in a finite population U of size $N < \infty$, with values $\{y_i\}_{i=1}^N$. The finite-population α-quantile is defined as:

$$Q(\alpha) = \inf\left\{t; N^{-1}\sum_{i=1}^{N}\Delta(t-y_i) \geq \alpha\right\},$$

where $\Delta(u) = \mathbf{1}\{u > 0\}$. Note that

$$F_N(t) = N^{-1}\sum_{i=1}^{N}\Delta(t-y_i), \quad -\infty < t < \infty,$$

is known as finite population distribution function.

In PPS sampling, let X_i be the size variable for unit i. A PPS sample $s \subset U$ of size n is selected with the probability of selection $\pi_i = nX_i/\sum_{j=1}^{N} X_j$, $i = 1, \ldots, N$. Let I_i be the sample inclusion indicator for unit i with 1 for the sampled units and 0 for the non-sampled units. The Hájek estimator of the cumulative distribution function is defined as

$$\widehat{F}_{\mathrm{HA}}(t) = \frac{\sum_{i=1}^{N} I_i\pi_i^{-1}\Delta(t-y_i)}{\sum_{i=1}^{N} I_i\pi_i^{-1}} = \frac{\sum_{i\in s}\pi_i^{-1}\Delta(t-y_i)}{\sum_{i\in s}\pi_i^{-1}}, \tag{2.1}$$

with the estimated α-quantile of Y defined as $\widehat{Q}_{\mathrm{HA}}(\alpha) = \inf\left\{t; \widehat{F}_{\mathrm{HA}}(t) \geq \alpha\right\}$.

For model-based approach, the survey outcomes are partitioned into those of the units in the selected sample and those of the non-sampled units $\boldsymbol{Y} = (\boldsymbol{Y}_s, \boldsymbol{Y}_{ns})$. A regression model is first fitted using data in the sample $\boldsymbol{Y}_s$. The unobserved survey outcomes of the non-sampled units $\boldsymbol{Y}_{ns}$ are then predicted using the fitted regression model. Therefore, the model-based estimators, by plugging in the predicted survey outcomes for the non-sampled

units, naturally take the following form

$$\widehat{Q}(\alpha) = \inf\left\{t; N^{-1}\left(\sum_{i\in s}\Delta(t-y_i) + \sum_{j\in ns}\Delta(t-\hat{y}_j)\right) \geq \alpha\right\},$$

where $\hat{y}_j$ is a prediction for the jth non-sampled unit based on the fitted regression model. The model can be fitted in a frequentist (Royall, 1971) or Bayesian (Ericson, 1969) setting.

2.2.2 Bayesian Model-Based Inference

We consider the fully Bayesian approach as we find it a natural setting to implement predictive inference. The Bayesian model-based approach posits a probability model for the data $p(\boldsymbol{Y}|\boldsymbol{\pi}, \boldsymbol{\theta})$ with prior distributions on parameters $\boldsymbol{\theta}$, and focuses on prediction of the non-sampled units $\boldsymbol{Y}_{ns}$ relevant to the quantity $Q(\alpha)$ of interest (Little, 2004). In PPS sampling, we have $p(\boldsymbol{I}|\boldsymbol{Y}, \boldsymbol{\pi}) = p(\boldsymbol{I}|\boldsymbol{\pi})$, so that given $\boldsymbol{\pi}$, the sampling mechanism is ignorable (Gelman et al., 2014; Rubin, 1983), and it follows

$$p(\boldsymbol{Y}_{ns}|\boldsymbol{Y}_s, \boldsymbol{\pi}, \boldsymbol{I}) = p(\boldsymbol{Y}_{ns}|\boldsymbol{Y}_s, \boldsymbol{\pi}) = \int p(\boldsymbol{Y}_{ns}|\boldsymbol{y}_s, \boldsymbol{\pi}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}_s, \boldsymbol{\pi})\, d\boldsymbol{\theta}.$$

The posterior distributions of the finite population quantities are simulated by generating a large number of draws based on the Markov chain Monte Carlo (MCMC) simulation from the posterior predictive distributions. For each iteration of the MCMC simulation, indexed by $r = 1, \ldots, R$, the algorithm is as follows:

1. Draw $\boldsymbol{\theta}^{(r)} \sim p(\boldsymbol{\theta}|\boldsymbol{y}_s, \boldsymbol{\pi})$
2. Generate $\hat{\boldsymbol{y}}_{ns}^{(r)} \sim p(\boldsymbol{Y}_{ns}|\boldsymbol{y}_s, \boldsymbol{\pi}, \boldsymbol{\theta}^{(r)})$
3. Compute $\widehat{Q}(\alpha)^{(r)} = \inf\left\{t; N^{-1}\left(\sum_{i\in s}\Delta(t-y_i) + \sum_{j\in ns}\Delta(t-\hat{y}_j^{(r)})\right) \geq \alpha\right\}$

The point estimates of population quantiles are obtained using the median of the draws of $\widehat{Q}(\alpha)$ across R iterations, and the 95% credible intervals are formed by splitting the tail area equally between the upper and lower endpoints of the MCMC simulations or using the highest probability density method.

Under the Bayesian modeling framework, we propose two model-based predictive estimators by assuming the survey outcome to follow a skew-normal distribution, SkewNorm(ξ, ω^2, α), given the probability of selection. The location parameter ξ and scale parameter ω^2 are modeled as functions of the probability of selection π and the slant parameter α is used to catch the skewness in the data (Azzalini, 2013). We describe next the two proposed estimators in details.

2.2.2.1 Skew-Normal Bayesian P-Spline Predictive Approach (SN-BPSP)

The skew-normal penalized spline predictive approach models the location parameter using a penalized spline and the scale parameter as a polynomial function of the probability of selection, which leads to the following model specification.

$$\begin{aligned} Y_i|\pi_i, \boldsymbol{\beta}, \boldsymbol{b}, \sigma^2, \alpha, \gamma &\overset{\text{ind.}}{\sim} \text{SkewNorm}(SPL(\pi_i, \boldsymbol{m}), \omega_i^2, \alpha), \\ \omega_i^2 &= (\alpha^2+1)\sigma^2\pi_i^{2\gamma}, \\ SPL(\pi_i, \boldsymbol{m}) &= \beta_0 + \sum_{l=1}^{p}\beta_l\pi_i^l + \sum_{k=1}^{K} b_k(\pi_i - m_k)_+^p, \\ \boldsymbol{b} = (b_1, \ldots, b_K)^T|\tau_b^2 &\sim N(\boldsymbol{0}, \tau_b^2\boldsymbol{I}_K), \end{aligned} \tag{2.2}$$

where the constants $\boldsymbol{m} = (m_1, m_2, \ldots, m_K)^T$ are K pre-selected fixed knots, and $(\pi_i - m_1)_+^p, \ldots, (\pi_i - m_K)_+^p$ are truncated polynomial spline basis functions of degree p with $(u)_+^p = \{u \times I(u > 0)\}^p$ for $u \in \mathbb{R}$ and $p = 1, 2$ or 3 for linear, quadratic or cubic splines. Penalty is

imposed on the coefficients $\boldsymbol{b}$ by specifying a normal distribution, with the amount of penalty controlled by the variance component τ_b^2. In hierarchical Bayesian approach, a hyper prior is specified on τ_b and the amount of penalization is automatically determined in the posterior inference procedure.

The skewed-normal distribution has the following hierarchical representation that is crucial to the posterior distribution derivations for Bayesian inference.

Proposition 1. *If* $W \sim N(0,1)\mathbf{1}\{w > 0\}$ *and* $Y|W = w \sim N(\xi + \alpha\sigma w, \sigma^2)$, *then* $Y \sim SkewNorm(\xi, \omega^2, \alpha)$ *with probability density function*

$$f(y|\xi, \omega^2, \alpha) = \frac{2}{\omega}\phi\left(\frac{y-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{y-\xi}{\omega}\right)\right),$$

where $\omega^2 = (\alpha^2+1)\sigma^2$. *In other words, the skew-normal distribution* $SkewNorm(\xi, \omega^2, \alpha)$ *can be simulated using the algorithm* $Y = \alpha\sigma|Z_1| + (\xi + \sigma Z_2)$, *with* $Z_1, Z_2 \overset{i.i.d.}{\sim} N(0,1)$ *and* $\omega^2 = (\alpha^2+1)\sigma^2$.

Proof. See Appendix A. □

Using the hierarchical representation, the distribution specification in line 1 of model (2.2) can be rewritten as

$$Y_i|\pi_i, \boldsymbol{\beta}, \boldsymbol{b}, \sigma^2, \alpha, \gamma, W_i = w_i \overset{\text{ind.}}{\sim} N(SPL(\pi_i, \boldsymbol{m}) + \alpha\sigma\pi_i^{\gamma} w_i, \sigma^2\pi_i^{2\gamma}),$$

where $W_i \overset{\text{i.i.d.}}{\sim} N(0,1)\mathbf{1}\{w_i > 0\}$. We use a uniform prior $U(-2,+2)$ for the order of polynomial function γ, a weakly informative prior distribution $N(0, \varphi^2 = (10^3)^2)$ for each of the polynomial coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$, an improper uniform distribution $U(0, +\infty)$ for τ_b and σ, and a half-normal distribution $N(0, \psi^2 = 10^2)^+$ for the slant parameter α.

The full conditionals of the posteriors are included in Appendix A, which can be used to develop posterior simulation scheme. Improper uniform prior is considered here for purpose of posterior derivations. Half-Cauchy priors suggested in Gelman (2006) are also considered with $\tau_b \sim \text{Cauchy}(0,1)^+$ in the simulation studies, and implemented using probabilistic programming language Stan (Carpenter et al., 2017). Stan obtains samples from the posterior distribution using the no-U-turn sampler, a variant of Hamiltonian Monte Carlo (Hoffman and Gelman, 2014). The Stan script is included in Appendix A.

2.2.2.2 Skew-Normal Bayesian Two-Moment P-Spline Predictive Approach (SN-B2PSP)

The skew-normal Bayesian two-moment penalized spline predictive estimator takes a more flexible approach by modeling both the location and scale parameters of the skew-normal distribution as penalized spline functions of the probability of selection, which leads to the following model specification.

$$
\begin{aligned}
Y_i|\pi_i, \boldsymbol{\beta}, \boldsymbol{b}, \sigma_i^2, \alpha &\overset{\text{ind.}}{\sim} \text{SkewNorm}(SPL_1(\pi_i, \boldsymbol{m}), \omega_i^2, \alpha), \\
\omega_i^2 &= (\alpha^2+1)\sigma_i^2, \\
\sigma_i^2|\pi_i, \boldsymbol{\lambda}, \boldsymbol{\nu}, \sigma_A^2 &\overset{\text{ind.}}{\sim} \text{LogNorm}(SPL_2(\pi_i, \boldsymbol{m}), \sigma_A^2), \\
SPL_1(\pi_i, \boldsymbol{m}) &= \beta_0 + \sum_{l=1}^{p} \beta_l \pi_i^l + \sum_{k=1}^{K} b_k (\pi_i - m_k)_+^p, \\
SPL_2(\pi_i, \boldsymbol{m}) &= \lambda_0 + \sum_{l=1}^{q} \lambda_l \pi_i^l + \sum_{k=1}^{K} \nu_k (\pi_i - m_k)_+^q, \\
\boldsymbol{b}|\tau_b^2 \sim N(\boldsymbol{0}, \tau_b^2 \boldsymbol{I}_K), &\boldsymbol{\nu}|\tau_\nu^2 \sim N(\boldsymbol{0}, \tau_\nu^2 \boldsymbol{I}_K),
\end{aligned}
\tag{2.3}
$$

where the constants $m_1, \ldots, m_K$ are K selected fixed knots shared by both splines. Without loss of generality, number of knots and locations of knots can be different between the two

splines if suggested by data. Model (2.3) is an extension of the B2PSP model in Chen, Elliott, and Little (2012) which assumes a normal distribution.

Similarly, to derive the posterior distributions, we consider a weakly informative prior distribution $N(0, \varphi^2 = (10^3)^2)$ for each of the polynomial coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ and $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_q)^T$, an improper uniform distribution $U(0, +\infty)$ for the hierarchical variance components τ_b and τ_ν, and a half-normal distribution $N(0, \psi^2 = 10^2)^+$ for the slant parameter α. The full conditionals of the posteriors can be found in Appendix B2. The model can also be easily implemented in Stan (Carpenter et al., 2017). The Stan script is included in Appendix A.

2.2.3 Transformations on Outcomes and Selection Probabilities

In practice, transformations can be applied to the survey outcome and the probability of selection to achieve better model fit and, consequently, better predictive accuracy. If the conditional distribution $Y|\pi$ is skewed with a heavy tail, natural logarithm (log) or square root transformation on Y could be considered to reduce skeweness, after which the shape of the transformed distribution can be better modeled by the skew-normal distribution. If the values of π are not equally spread out over the range, logit or square root transformation again can be applied to reduce skewness and sparseness within certain ranges and improve model fit. To select appropriate transformations on Y and π, we use Bayesian Pareto smoothed importance sampling leave-one-out (PSIS-LOO) cross-validation (Vehtari, Gelman, and Gabry, 2017), which estimates expected log pointwise predictive density (elpd)

as a measure of out-of-sample predictive fit

$$\widehat{\text{elpd}}_{\text{PSIS-LOO}} = \sum_{i=1}^{n} \log \left(\frac{\sum_{r=1}^{R} w_i^{(r)} p_Y(y_i|\theta^{(r)})}{\sum_{r=1}^{R} w_i^{(r)}} \right), \tag{2.4}$$

where $p_Y(y|\theta)$ denotes the distribution of Y, $r = 1, \ldots, R$ labels the posterior draws and $w_i^{(r)}$ are regularized importance weights smoothed by fitting a generalized Pareto distribution to the tail distribution of the raw importance weights. Higher value of elpd indicates better out-of-sample predictive accuracy. Therefore, the transformation corresponding to the highest $\widehat{\text{elpd}}_{\text{PSIS-LOO}}$ is preferred. The algorithm for computing $\widehat{\text{elpd}}_{\text{PSIS-LOO}}$ based on a fitted model is implemented in the `loo` package in `R` (Vehtari, Gabry, Yao, and Gelman, 2018).

When transformation is applied to survey variable Y, the Bayesian PSIS-LOO cross-validation estimate needs to be modified based on the transformed variable $Z = Z(Y)$, with the following derivation:

$$\begin{aligned}
\widehat{\text{elpd}}_{\text{PSIS-LOO}} &= \sum_{i=1}^{n} \log \left(\frac{\sum_{r=1}^{R} w_i^{(r)} p_Y(y_i|\theta^{(r)})}{\sum_{r=1}^{R} w_i^{(r)}} \right) \\
&= \sum_{i=1}^{n} \log \left(\frac{\sum_{r=1}^{R} w_i^{(r)} p_Z(z_i|\theta^{(r)}) \left|\frac{dz_i}{dy_i}\right|}{\sum_{r=1}^{R} w_i^{(r)}} \right) \\
&= \sum_{i=1}^{n} \log \left|\frac{dz_i}{dy_i}\right| \left(\frac{\sum_{r=1}^{R} w_i^{(r)} p_Z(z_i|\theta^{(r)})}{\sum_{r=1}^{R} w_i^{(r)}} \right) \\
&= \sum_{i=1}^{n} \log \left|\frac{dz_i}{dy_i}\right| + \sum_{i=1}^{n} \log \left(\frac{\sum_{r=1}^{R} w_i^{(r)} p_Z(z_i|\theta^{(r)})}{\sum_{r=1}^{R} w_i^{(r)}} \right).
\end{aligned} \tag{2.5}$$

Note that the above first term is easily computed based on the specific transformation and the second term is the PSIS-LOO estimate based on the transformed variable Z that can be implemented using the `loo` package in `R`.

We considered no transformation, log or square root transformation on Y, and no transformation, logit or square root transformation on π. To compute $\widehat{\text{elpd}}_{\text{PSIS-LOO}}$, we used (2.4)

for the models without transformation on Y and (2.5) for the models with log or square root transformation on Y. We chose the transformations that lead to the largest $\widehat{\text{elpd}}_{\text{PSIS-LOO}}$ in the simulation studies and application.

2.3 Simulation Studies

Simulation studies were conducted to evaluate the performance of the two skew-normal model-based predictive estimators in estimating finite population quantiles. The two model-based estimators were compared with the conventional weighted quantile estimator $\hat{Q}_{\text{HA}}(\alpha) = \inf\{t, \hat{F}_{\text{HA}}(t) \geq \alpha\}$ (henceforth HA) obtained by inverting the Hájek (1971) estimator of distribution function defined in (2.1).

2.3.1 Simulation Design

Four artificial populations of size $N = 2,000$ were simulated with size variable X generated from a skewed Gamma distribution with shape parameter $k = 1.5$ and rate parameter $1/\theta = .001$. With sampling rate of 10%, systematic PPS samples of size $n = 200$ were drawn with the probability of selection $\pi_i = nX_i / \sum_{j=1}^{N} X_j$ for unit i, $i = 1, \ldots, N$. The survey outcome Y was generated using the following conditional distributions:

(a) Skew-Normal distribution with location parameter positively associated with probability of selection

$Y_i|X_i \overset{\text{ind.}}{\sim} \text{SkewNorm}(\xi = 150 + 100\pi_i, \sigma^2 = 12^2\pi_i^{2\times 0.8}, \alpha = 4)$

(b) Gamma distribution with constant mean

$Y_i|X_i \overset{\text{ind.}}{\sim} \text{Gamma}(k = 0.3 \log X_i, \theta = 500/\log X_i)$, with $\text{E}(Y_i|X_i) = k\theta = 150$

(c) Gamma distribution with mean positively associated with size variable

$$Y_i|X_i \overset{\text{ind.}}{\sim} \text{Gamma}(k = 0.3 \log X_i, \theta = 500\sqrt{X_i})$$

(d) Log-Normal distribution with location parameter positively associated with probability of selection

$$Y_i|X_i \overset{\text{ind.}}{\sim} \text{LogNorm}(\mu = -2 + 3 \log 10 + 5\pi_i, \sigma^2 = 0.8^2 \pi_i^{2\times 0.1})$$

Setting (a) was designed to examine whether the SN-BPSP works well when the underlying model is true and how much efficiency the SN-B2PSP loses by assuming a more complex model structure for the scale parameter. Settings (b)-(d) were used to assess whether the skew-normal models, combined with transformations, can adequately model other types of commonly seen skewed distribution other than skew-normal distribution. Figure 2.1 displays the scatter plots of Y against π for the four generated artificial populations, each with red diamonds denoting a selected PPS sample. Except for setting (b), where the conditional expectation of Y is not associated with π (NULL), the other three settings were constructed such that Y is positively associated with π (Positive).

For each of the simulation settings, 500 systematic PPS samples were drawn from the population, with population units permuted before sampling. The three estimators were compared in estimating quantiles at levels $\alpha = .05, .10, .25, .50, .75, .90, .95$. Empirical bias, root mean squared error (RMSE), average widths and non-coverage rates of 95% probability intervals were calculated. For Bayesian model-based approach, credible intervals were computed using equal tail quantiles from posterior predictive distributions. For HA weighted approach, confidence intervals were constructed using the variance estimation of Woodruff (1952) implemented in `survey` package in `R` (Lumley, 2016). Except for setting (a), where

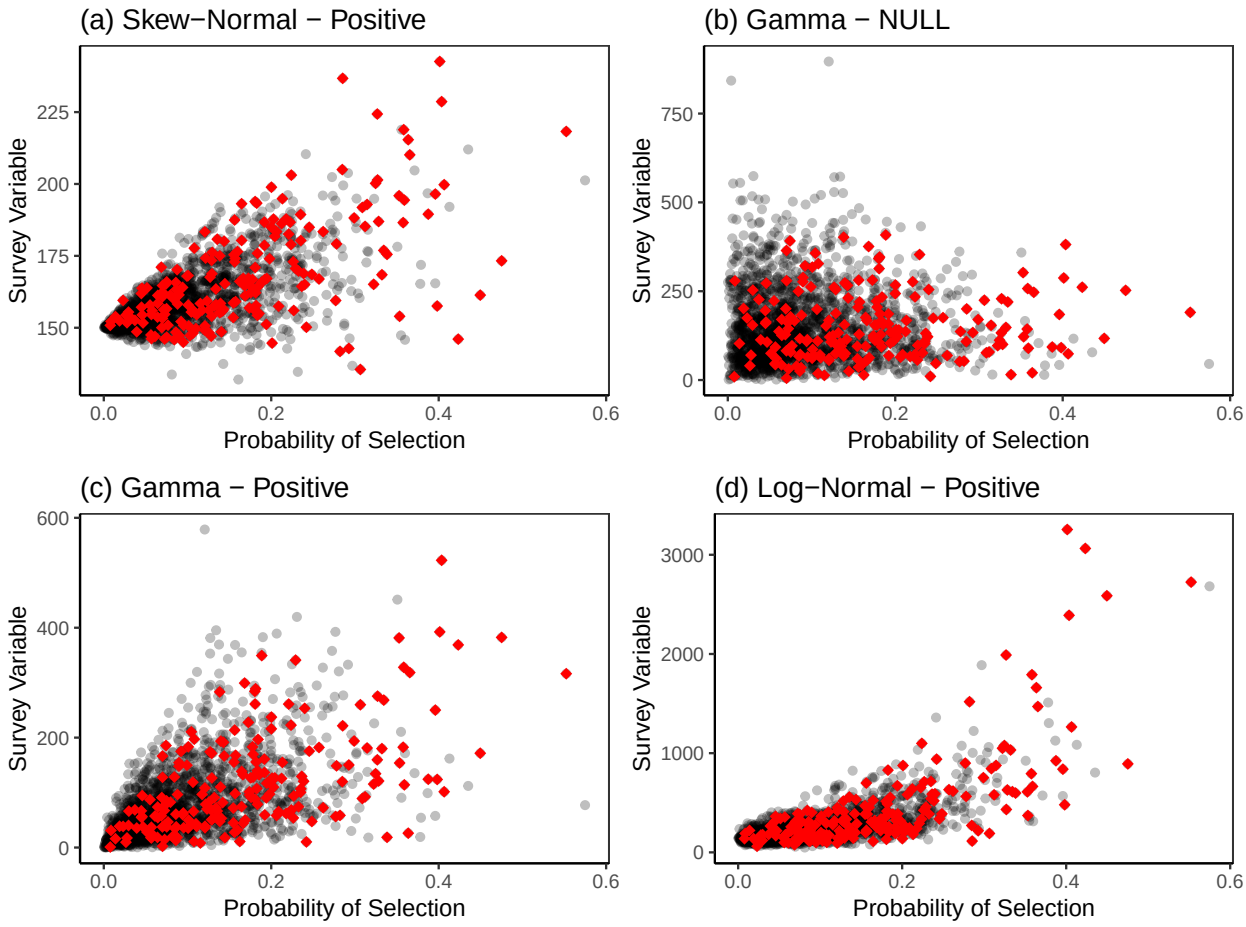

Figure 2.1: Scatter plots of survey outcome against probability of selection for the four artificially generated populations of size $N = 2,000$, each with red diamonds denoting a selected PPS sample of size $n = 200$.

skew-normal distribution is the true underlying distribution for generating the survey population, the Bayesian PSIS-LOO cross-validation suggests that, after squared root transformation on both survey outcome and probability of selection, the Bayesian models achieve better out-of-sample predictive fit. Therefore, squared root transformation was applied to both variables before fitting the Bayesian predictive models.

2.3.2 Results

Table 2.1 summarizes empirical bias and RMSE of the three estimators. The two Bayesian model-based estimators are more efficient than the HA quantile estimator at all quantile levels, with lower RMSE across the four scenarios. The two Bayesian model-based estimators lead to similar or larger bias compared to the HA quantile estimator, with the largest bias in estimating upper quantiles ($\alpha = .90, .95$) in scenarios (b) and (d). The two Bayesian model-based estimators perform similarly in general. In scenario (a), where the model assumed in SN-BPSP is the true underlying model, the SN-BPSP estimator yields slightly smaller bias and RMSE than the SN-B2PSP estimator. This indicates that the SN-B2PSP does not lose much efficiency by specifying a more complex model than the true model. In contrast, the SN-B2PSP estimator leads to smaller bias and RMSE than the SN-BPSP in scenarios (b) and (d), especially in estimating the upper quantiles. This suggests the benefit of a more flexible model of SN-B2PSP when the data have a more complex underlying distribution than skew-normal.

Table 2.2 shows average widths and non-coverage rates of the 95% credible/confidence intervals (CIs). The HA weighted approach has close to nominal level non-coverage rate of 5% in estimating middle and upper quantiles but poor coverage for lower quantiles in

Table 2.1: Empirical bias and RMSE of the proposed Bayesian model-based quantile estimators and the HA quantile estimator with 500 PPS samples of size $n = 200$ from the four artificial populations of size $N = 2,000$.

Truth		Bias			RMSE		
α	Q_α	SN-B2PSP	SN-BPSP	HA	SN-B2PSP	SN-BPSP	HA
Skew-Normal - Positive							
.05	148.0	−0.4	0.1	−0.1	1.0	0.8	0.9
.10	149.7	−0.3	0.0	−0.2	0.8	0.7	0.8
.25	152.3	0.3	−0.1	0.0	0.6	0.6	0.9
.50	156.7	0.4	0.1	0.0	0.7	0.5	1.1
.75	164.6	−0.1	0.0	−0.1	0.7	0.7	1.4
.90	175.0	−0.3	−0.1	0.0	1.1	1.0	2.1
.95	183.3	−0.5	−0.4	−0.1	1.4	1.3	1.9
Gamma Distribution - NULL							
.05	28.1	2.1	1.8	0.4	5.5	5.3	7.8
.10	41.5	3.0	2.7	0.7	6.2	5.9	9.1
.25	73.0	3.6	3.5	0.7	7.4	7.2	10.0
.50	125.8	1.9	2.3	1.1	9.0	9.1	12.8
.75	201.9	0.0	1.5	1.1	12.3	12.8	20.7
.90	288.3	3.1	6.3	1.4	19.3	21.5	32.7
.95	353.7	3.7	8.1	−4.9	26.5	29.9	40.2
Gamma Distribution - Positive							
.05	8.2	1.6	1.4	1.0	2.9	3.0	4.1
.10	13.3	2.6	2.1	1.0	3.8	3.6	5.4
.25	29.3	2.3	1.3	0.7	3.9	3.5	6.4
.50	58.6	2.2	1.1	1.0	4.4	4.1	7.8
.75	107.6	0.4	−0.2	−0.4	5.1	5.3	9.6
.90	171.6	0.2	0.4	−1.0	7.1	7.4	11.6
.95	225.4	−2.1	−1.8	−2.5	10.8	11.3	18.8
Log Normal - Positive							
.05	105.0	−3.3	−1.1	−1.9	9.3	9.1	10.7
.10	118.3	−2.0	−1.0	−0.1	8.5	9.2	10.1
.25	149.9	−1.8	−3.9	−0.4	7.9	9.5	10.1
.50	202.8	0.4	−3.4	0.1	8.1	8.9	14.0
.75	301.4	−0.3	2.1	−0.7	10.5	10.7	18.9
.90	449.4	15.1	30.3	6.1	23.4	35.1	37.8
.95	620.1	4.1	22.7	−3.5	25.8	33.8	44.0

all four scenarios. Both Bayesian model-based approaches lead to close to nominal level coverage rate in estimating all levels of quantiles in scenarios (a)–(c), whereas in scenario (d), the more flexible SN-B2PSP has better credible interval coverage than the SN-BPSP. The Bayesian model-based approaches also yield shorter intervals of 95% CIs than the HA approach in most of the scenarios.

2.4 Application to NDATSS

The National Drug Abuse Treatment System Survey (NDATSS) is a panel survey of substance abuse treatment programs in the United States (D'Aunno, Friedmann, Chen, and Wilson, 2015). The population considered here is defined by $N = 475$ substance abuse treatment programs surveyed in the 2016 wave of NDATSS with complete information on both number of staff and number of active clients receiving treatment at the unit. We conducted two sets of analysis to compare the two Bayesian model-based estimators with the HA estimator. We first treated number of staff as size variable, drew a systematic PPS sample of size $n = 50$, and estimated population quantiles of total number of active clients at $\alpha = .10, .25, .50, .75, .90$. We then conducted a simulation study by repeating the above procedure 500 times.

For the first set of analysis using a single PPS sample, we performed model checking using posterior predictive p-values (Gelman et al., 2014, chapter 6) based on two test quantities, including (a) $T_1(\boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and (b) $T_2(\boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \xi_i}{\omega_i}\right)^2$. The two test quantities catch different aspects of the data, with $T_1(\cdot)$ measuring the variability of the survey outcome and $T_2(\cdot)$ measuring the discrepancy between the survey outcome

Table 2.2: Average widths (AIW) and non-coverage rates of the 95% probability intervals for the proposed Bayesian model-based quantile estimators and the HA estimator with 500 PPS samples of size $n = 200$ from the four artificial populations of size $N = 2,000$.

Truth		95% AIW			non-coverage rate (%)		
α	Q_α	SN-B2PSP	SN-BPSP	HA	SN-B2PSP	SN-BPSP	HA
Skew-Normal - Positive							
.05	148.0	3.6	2.8	3.8	6.6	6.4	16.6
.10	149.7	3.0	2.4	3.3	6.6	7.8	14.8
.25	152.3	2.5	2.0	3.4	7.6	7.4	9.0
.50	156.7	2.5	2.1	4.4	9.2	4.2	4.6
.75	164.6	2.9	2.7	5.5	4.0	4.6	4.4
.90	175.0	4.3	4.1	8.5	4.6	2.4	4.4
.95	183.3	5.6	5.3	8.3	5.0	3.6	3.6
Gamma Distribution - NULL							
.05	28.1	21.5	21.8	23.5	4.4	4.2	14.4
.10	41.5	22.4	22.5	27.9	5.8	5.6	7.6
.25	73.0	25.8	25.7	35.4	8.0	6.6	6.8
.50	125.8	35.0	35.2	49.4	4.6	5.0	4.2
.75	201.9	53.5	54.8	79.5	3.6	4.6	4.4
.90	288.3	89.8	94.3	125.3	3.6	5.4	4.8
.95	353.7	128.7	137.0	161.8	3.0	4.0	5.4
Gamma Distribution - Positive							
.05	8.2	11.1	10.6	10.1	4.2	5.0	23.4
.10	13.3	12.2	11.6	15.3	7.6	7.6	14.2
.25	29.3	13.6	12.9	23.2	6.4	5.0	7.6
.50	58.6	17.1	16.4	29.0	5.8	5.6	7.4
.75	107.6	23.1	22.7	37.3	3.4	4.4	6.4
.90	171.6	35.2	34.6	52.2	3.4	3.6	4.4
.95	225.4	51.0	50.8	76.1	2.2	3.0	3.8
Log Normal - Positive							
.05	105.0	34.9	28.6	33.1	9.0	13.4	14.8
.10	118.3	31.8	27.0	36.0	7.4	15.4	11.2
.25	149.9	28.7	25.7	38.2	6.6	18.2	7.2
.50	202.8	31.9	29.1	54.4	7.0	10.2	5.6
.75	301.4	43.3	42.3	78.5	4.8	5.8	5.4
.90	449.4	77.0	80.4	145.9	7.6	30.4	4.2
.95	620.1	117.2	120.7	200.6	2.6	8.6	4.4

and fitted parametric distribution. In each MCMC iteration r, the realized test quantities $T_i(\boldsymbol{y}, \boldsymbol{\xi}^{(r)}, \boldsymbol{\omega}^{(r)})$ under the observed data and predictive test quantities $T_i(\tilde{\boldsymbol{y}}^{(r)}, \boldsymbol{\xi}^{(r)}, \boldsymbol{\omega}^{(r)})$ under the simulated data can be computed, with $\tilde{\boldsymbol{y}}^{(r)}$ drawn from the posterior predictive distribution. The Bayesian posterior predictive p-value is defined as the probability that the predictive test quantity is greater than the realized test quantity, evaluated over the posterior distribution. The Bayesian p-value measures the discrepancy between the observed data and the posterior predictive distribution in the aspect characterized by $T(\cdot)$. A Bayesian p-value close to 0.5 indicates good fit while a Bayesian p-value near 0 or 1 indicates that the observed pattern would be unlikely to happen if the model were true and, therefore, lack of fit. Both of the two Bayesian predictive models yielded a posterior predictive p-value close to 0.5, indicating adequate model fit. The posterior predictive plots and p-values for the SN-BPSP model are displayed in Figure 2.2. Since the number of active clients are known for all the units in our defined population, we also compared the distributions of predicted and actual number of active clients for the non-sampled units.

2.4.1 Quantile Estimation of Number of Active Clients Using a Single PPS Sample

Figure 2.3 displays number of active clients against probability of selection with and without squared root transformation on both variables. The black open circles and red dots represent units in the population and a PPS sample, respectively. The scatter plots show that the data points are more equally spread out with respect to both axes with squared root transformation on both variables. Such transformation was also suggested by the Bayesian PSIS-LOO

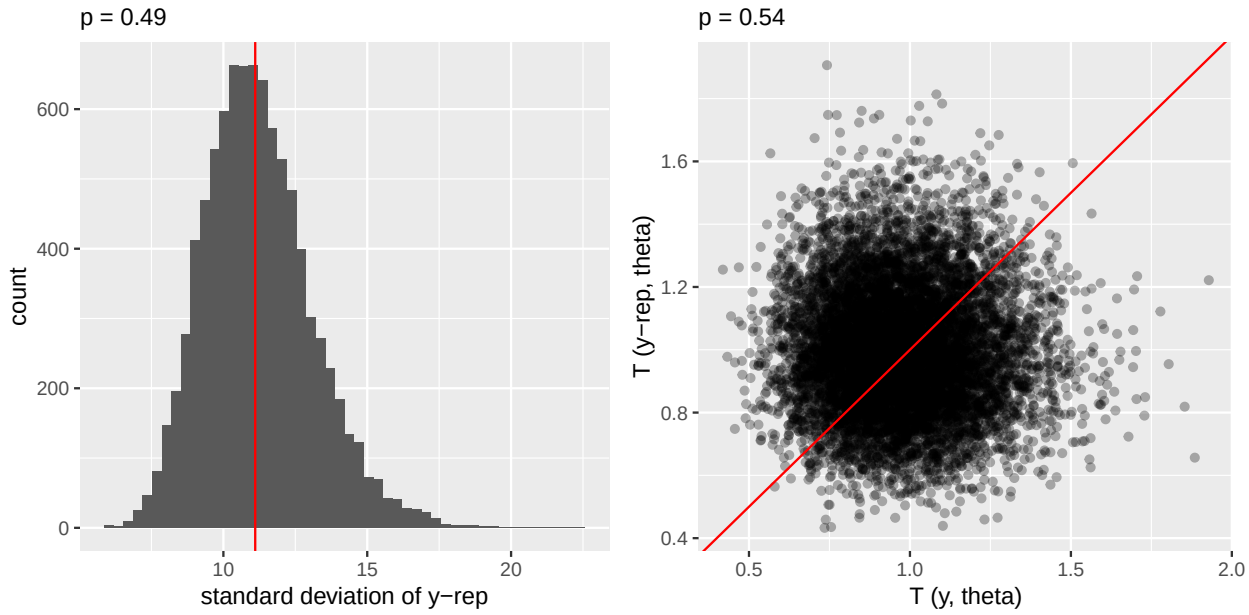

Figure 2.2: Realized v.s. posterior predictive distributions for the two test statistics and corresponding posterior predictive p-values for the SN-BPSP model with a PPS sample from the NDATSS population: (a) Sample standard deviation (vertical line) compared to 9000 simulations from the posterior predictive distribution of sample standard deviation. (b) Scatter plot showing the test statistic $T_2(\boldsymbol{y}, \boldsymbol{\xi}, \boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\xi_i}{\omega_i}\right)^2$ with $T_2(\tilde{\boldsymbol{y}}^{(r)}, \boldsymbol{\xi}^{(r)}, \boldsymbol{\omega}^{(r)})$ in the vertical axis and $T_2(\boldsymbol{y}, \boldsymbol{\xi}^{(r)}, \boldsymbol{\omega}^{(r)})$ in the horizontal axis based on 9000 simulations from the posterior distribution of $(\boldsymbol{\xi}, \boldsymbol{\omega}, \tilde{\boldsymbol{y}})$.

cross-validation in comparison to no transformation and natural logarithm transformation.

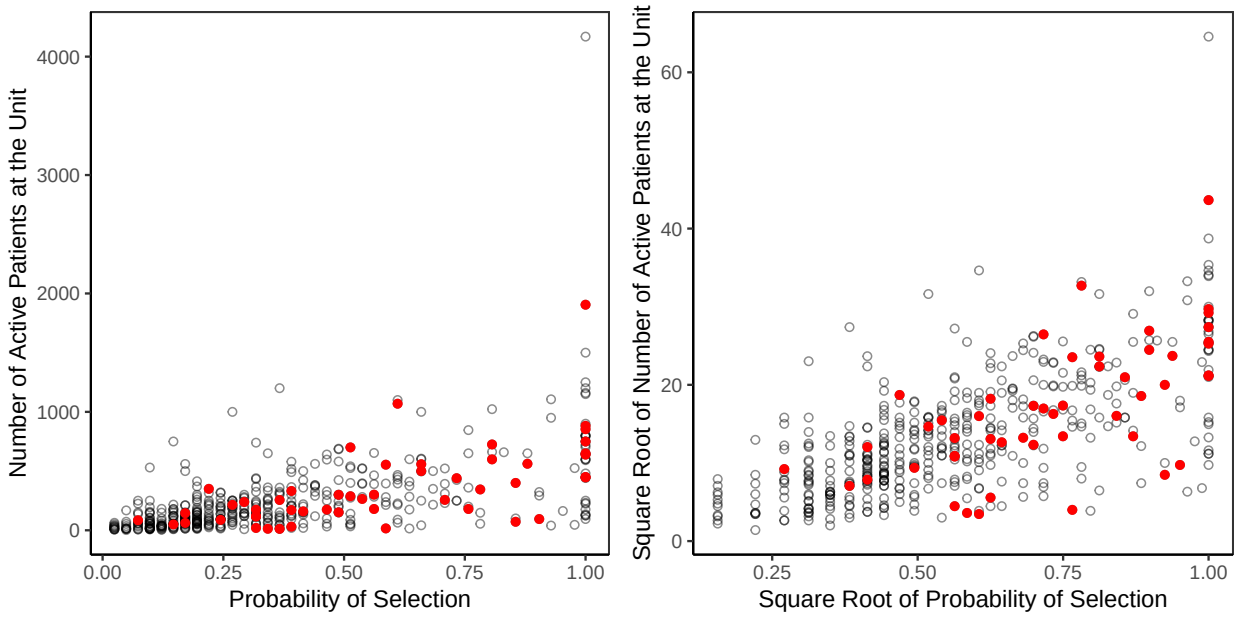

Figure 2.3: Scatter plots of number of active clients against probability of selection for NDATSS population of size $N = 475$ with and without squared root transformation and a PPS sample of size $n = 50$ in red dot.

Figure 2.4 shows the point estimates and 95% CIs of the HA estimator and the two skew-normal model-based estimators with squared root transformation applied to both Y and π in estimating population 10th, 25th, 50th, 75th, and 90th percentiles. The known true population quantiles are denoted using a solid horizontal line in each quantile plot. In estimating the population 10th percentile, the two model-based methods yield closer to the true quantile estimates than the HA method; while in estimating the other four quantiles, the two model-based methods yield shorter 95% CIs than the HA method.

Figure 2.5 compares the distribution of the predicted number of active clients to the distribution of actual number of active clients for the non-sampled units. Figure 2.5(a)

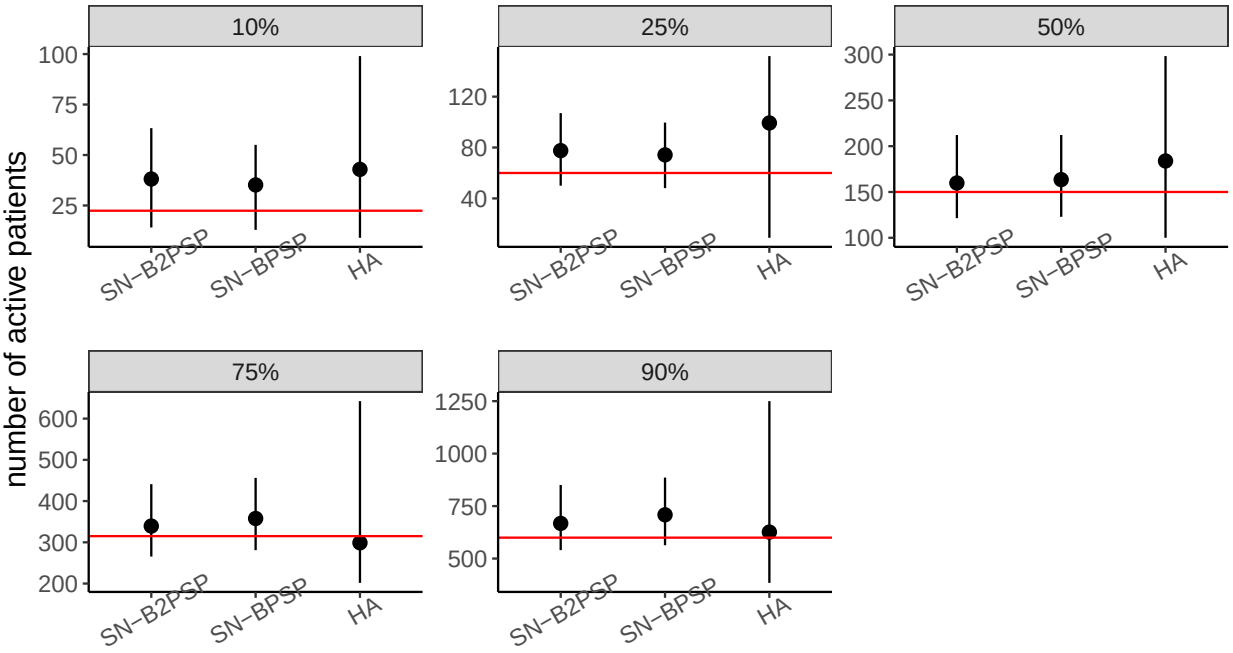

Figure 2.4: Point estimates and 95% probability intervals for quantiles of number of active patients at various quantile levels using a PPS sample of size $n = 50$ from the NDATSS population.

plots the densities of predicted number of active clients from 10 randomly selected MCMC iterations based on the SN-BPSP model and the density of actual number of active clients for the non-sampled units. Figure 2.5(b) plots predicted number of active clients from one MCMC iteration based on the SN-BPSP model versus the actual number against the probability of selection. Both plots suggest good model predictions for the non-sampled units using the SN-BPSP.

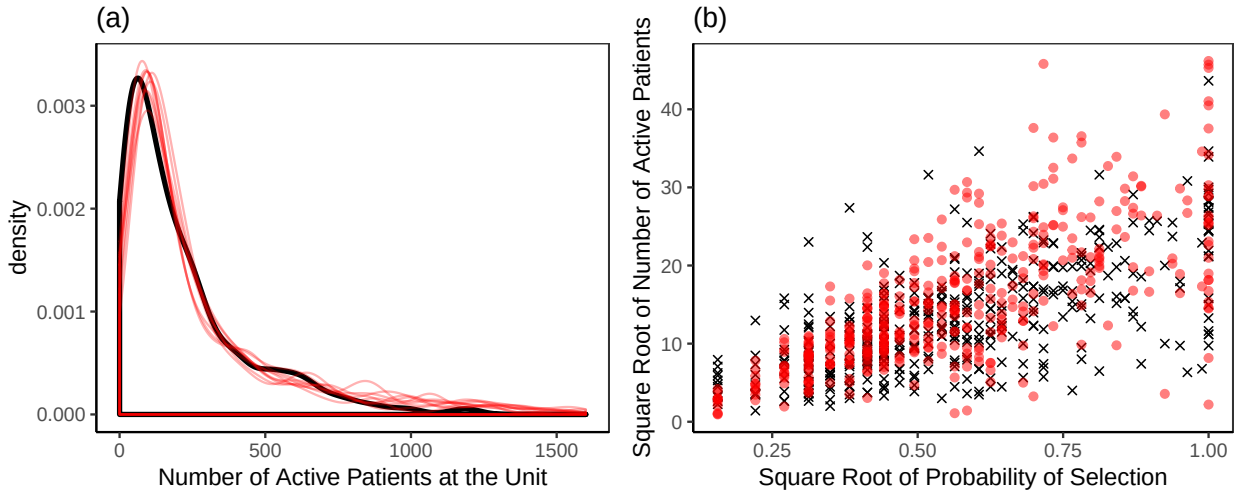

Figure 2.5: (a) Density plot of predicted number of active clients (in red) from 10 MCMC iterations based on SN-BPSP vs actual number of active clients (in thick black) for non-sampled units (b) scatter plot of predicted number of active clients (in red dots) vs actual number of active clients (in black crosses) against probability of selection with square root transformation

2.4.2 Repeated Simulation Studies on Quantile Estimation of Number of Active Clients

We estimated the quantiles of total number of active patients by repeatedly drawing 500 systematic PPS samples. Table 2.3 summarizes empirical bias, RMSE, average widths and non-coverage rates of 95% CIs, comparing the SN-BPSP, SN-B2PSP, and HA quantile estimates to the true population quantiles of $N = 475$ units. The simulation suggests that the two model-based estimators yield smaller RMSE, shorter 95% probability intervals, and closer to the nominal level coverage rates than the HA estimator. These findings are consistent with those conveyed by simulation study with artificially generated population data.

2.5 Discussion

Skewed data commonly arise in sample surveys. Estimation of population quantiles is of greater interest than population means for skewed data. Although weighted estimators of population quantiles are widely used in survey practice, they can be inefficient and have poor confidence coverage in small-to-moderate-sized samples. Model-based approaches can improve efficiency of survey estimates when the model is correctly specified. Previous literature on model-based methods mostly assumes a normal or log-normal distribution for the survey variables. Although transformations could be applied on the survey variables, in many scenarios skewness is still present after transformation. When the normality assumption is violated, the model-based estimators that rely on normality assumption can be biased. Therefore, development of more flexible modeling techniques for handling skewed data is of great interest. We propose two model-based predictive estimators for estimating

Table 2.3: Emprical bias, empirical RMSE, average 95% probability interval widths (AIW) and non-coverage rates for the proposed Bayesian model-based methods and the HA method with 500 PPS samples of size $n = 50$ from NDATSS population of size $N = 475$.

Truth	Method	Bias	RMSE	AIW	non-coverage rate (%)
	SN-B2PSP	6.2	14.3	49.6	4.6
$Q_{.10} = 22.4$	SN-BPSP	5.5	14.5	46.8	6.6
	HA	6.8	18.2	63.0	14.2
	SN-B2PSP	6.5	18.5	69.6	4.2
$Q_{.25} = 60.0$	SN-BPSP	4.3	17.8	65.3	5.2
	HA	4.7	28.1	154.8	1.4
	SN-B2PSP	1.0	26.4	103.9	2.8
$Q_{.50} = 150.0$	SN-BPSP	−1.4	26.3	98.6	5.4
	HA	3.7	42.5	319.2	0.4
	SN-B2PSP	3.8	43.2	176.7	5.2
$Q_{.75} = 315.0$	SN-BPSP	5.7	43.8	169.1	6.4
	HA	3.0	71.1	582.2	0.2
	SN-B2PSP	−1.2	65.2	315.2	3.0
$Q_{.90} = 599.6$	SN-BPSP	5.6	67.9	290.2	4.0
	HA	−16.2	92.8	1483.2	0.6

finite population quantiles with skewed survey data in the setting of probability proportional to size sampling. We assume a conditional skew-normal distribution for the survey variables given probability of selection and model the location and scale parameters of the skew-normal distribution as functions of the probability of selection. To allow a flexible association between the survey variable and probability of selection, the first method models the location parameter with a penalized spline and the scale parameter with a polynomial function, while the second method models both the location and scale parameters with penalized splines.

Simulations using both artificially generated population data and a real establishment survey suggest that the two skew-normal model-based quantile estimators outperform the weighted quantile estimator. Combined with transformations selected using the Bayesian PSIS-LOO cross-validation, our proposed skew-normal models can be used to handle various skewed data with distributions including but not limited to Gamma, log-normal, and skew-normal, in obtaining more efficient estimates of population quantiles than the weighted method. By using a fully Bayesian approach, the variance and 95% credible interval of the model-based estimators can be easily calculated from the posterior predictive distributions. The two model-based estimators yield shorter 95% credible intervals than the weighted estimator with variance estimated using the Woodruff's method. In estimating the lower tail regions of the population distribution where data is typically sparse, Woodruff's method tends to yield confidence intervals with lower coverage rates, whereas the two model-based estimators have closer to the nominal level coverage rates. By modeling the scale parameter as a second spline on the probability of selection, the SN-B2PSP estimator is more robust than the SN-BPSP estimator, yielding smaller RMSE and a closer to the nominal level cover-

age rate when a more complex model than the underlying model for the SN-BPSP is needed. On the other hand, the SN-B2PSP estimator does not lose much efficiency due to overfitting when a simpler model is required.

Although the methods are proposed in the context of PPS sampling, with the assumption that unit-level information of size variable is available for all units in the population, the methods can be naturally extended to handle skewed data in more general settings. In a PPS sampling where only aggregated (instead of unit-level) information of size variable for non-sampled units is available, Bayesian Bootstrap can be used to reconstruct unit-level information of size variable for the non-sampled units (Zangeneh and Little, 2015). In two-stage cluster sampling with PPS used in selecting primary sampling units, the proposed skew-normal model can be extended to include cluster specific random intercepts and unit-level covariates that are associated with both the survey variable and the sample inclusion indicator (Yuan and Little, 2007). The skew-normal models can also be used for small area estimation, where data tend to be sparse and normality assumption may not hold.

Bayesian methods are often criticized for being computationally intensive. However, with the availability of high performance computing clusters, computation would not be a major concern anymore. Using the compute cluster in the Department of Systems Biology at Columbia University that consists of 6,384 CPU cores, 23TB of RAM and 148 NVIDIA GPUs providing an additional 75,776 CUDA cores, the SN-BPSP method took 1.6 minutes and the SN-B2PSP method took 11 minutes to obtain the estimates and 95% CIs in a single sample in the NDATSS application, and the 500 replicates of simulation took less than 2 hours with 50 parallel computing tasks. Moreover, both of the two skew-normal model-

based estimators as well as their extensions can be easily implemented in the probabilistic programming language Stan (Carpenter et al., 2017). The Stan language is user-friendly, intuitive and easy to use. The users only need to specify a Bayesian statistical model and the priors for the parameters. Stan implements gradient-based MCMC algorithms for Bayesian inference. It can be assessed through the R software environment or RStudio Cloud on web. This is appealing to survey practitioners. Finally, although skew-normal distribution can be used to model the skewness in various skewed data, it is not intended for multimodal data due to mixture of distributions. More flexible models such as mixture normal or mixture skew-normal models can be used for such data.

Chapter 3

Inference from Non-Random Samples Using Bayesian Machine Learning

3.1 Introduction

Inference about a target population based on sample data relies on the assumption that the sample is representative. However, simple random samples are often not available in real data problems. Therefore, there is a need to generalize inference from the available non-random sample to the target population of interest. For example, randomized controlled trials (RCTs) are considered a gold standard to estimate treatment effects, but the measured effects can only be formally generalized to the participants within the trial. Recent evidence has indicated that subjects in an RCT can be much different from patients in routine practice. Such concern among clinicians about the external validity of RCTs has led to the underuse of effective treatments (Rothwell, 2005). This highlights the importance of generalizing treatment effect of RCTs to a definable patient population.

Survey sampling is a field that specifically deals with inference on populations with non-random samples, which can be viewed as a special case of generalizing inference. Probability samples collected via probability surveys have historically proven effective. However, such data comes with considerable cost, both time and budget. In the past several decades, large scale probability surveys have suffered increasingly high non-response rates, besides the rising costs. The probability surveys with low response rates are often non-representative, which challenges the validity of survey inference. In the meanwhile, recent development of information technology makes it increasingly convenient and cost-effective to collect large numbers of samples with detailed information via online surveys and opt-in panels. Such samples are highly non-representative due to selection bias. Classical weighting methods in survey literature such as post-stratification (Valliant, 1993) and raking (Deming and Stephan, 1940) can improve representativeness of survey samples when a small number of discrete auxiliary variables about populations are available for survey adjustments. However, such weighting methods can yield highly variable estimates of population quantities in the presence of extreme weights. Alternatively, model-based methods can be used. Wang, Rothschild, Goel, and Gelman (2015) demonstrates, through election forecast with non-representative voter intention polls on the Xbox gaming platform, that multilevel regression and post-stratification (MRP) can be used to generate accurate survey estimates from non-representative samples. Their estimates are in line with the forecasts from leading poll analyst. MRP is very appealing when statistical adjustment are made using a small number of discrete auxiliary variables.

In recent years, population data of high volume, variety, and velocity has become increas-

ing available, with examples including administrative data or electronic medical records. Such data contained detailed individual level information with high-dimensionality and can be used to generalize inference of non-random samples to their target populations. Although post-stratification, raking, and MRP methods can improve representativeness in the presence of a small number of discrete auxiliary variables, they are infeasible to be applied in high-dimensional settings. With high-dimensional auxiliary variables, Bayesian machine learning techniques have been shown to be effective in improving statistical inference in missing data and causal inference. Specially, Hill (2011) shows that Bayesian additive regression trees (BART) produces more accurate estimates of average treatment effects compared to propensity score matching, propensity-weighted estimators, and regression adjustment when the response surface is nonlinear and not parallel between treatment and control groups. Tan, Flannagan, and Elliott (2019) demonstrate, in the presence of missing data, that BART reduces bias and root mean square error of the doubly robust estimators when both propensity and mean models were misspecified. Inspired by these works, we propose Bayesian machine learning model-based methods and extensions for estimating population means using non-random samples. The proposed methods can be applied not only in the context of survey inference but also in more general settings, such as RCTs and epidemiological observational studies. We evaluate the proposed methods using simulation studies and demonstrate their applications in a mental health survey of Ohio Army National Guard service members and a non-random sample from an observational study using electronic medical records of COVID-19 patients.

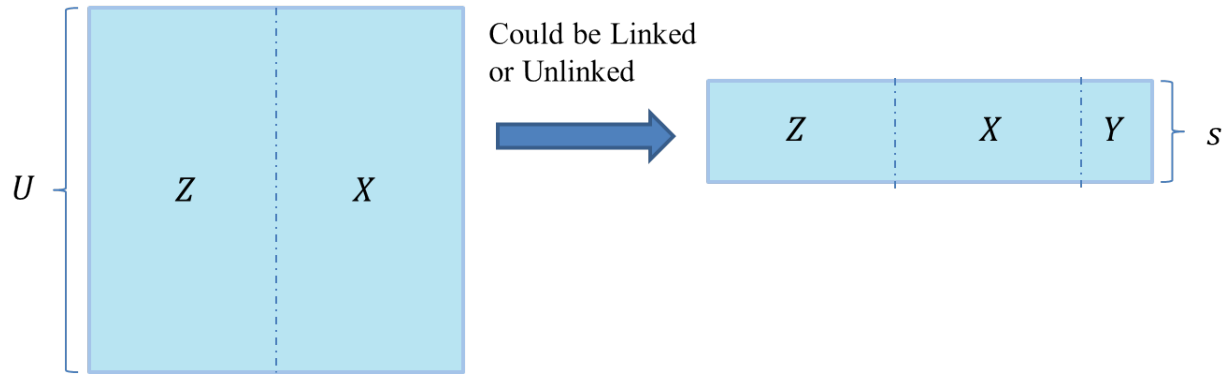

Figure 3.1: Population U and non-random sample s with shared discrete auxiliary variables $\boldsymbol{Z}$ and continuous auxiliary variables $\boldsymbol{X}$ as well as outcome Y measured only in s

3.2 Methods

3.2.1 Notation and Background

Let U be the finite population of size N and s be a non-random sample of size n from the population. In the sample s, information on the outcome of interest Y, discrete auxiliary variables $\boldsymbol{Z}$ and continuous auxiliary variables $\boldsymbol{X}$ were collected. In addition, data from the population U (e.g. census, administrative data, or electronic medical records) is also available with the the same set of auxiliary variables $\boldsymbol{Z}$ and $\boldsymbol{X}$ measured for all units in the population. Figure 3.1 illustrates the scenario under consideration, with population data on the left and the sample data on the right. Without loss of generality, we consider a continuous variable of interest Y with the estimand of interest being the finite population mean $Q(Y) = \frac{1}{N}\sum_{i \in U} Y_i$.

When the dimensions of $\boldsymbol{Z}$ and $\boldsymbol{X}$ are small, post-stratification, raking, and MRP can be applied by first discretizing the continuous auxiliary variables $\boldsymbol{X}$ as $\boldsymbol{X}^*$ using quantiles.

Using the joint distribution of discrete auxiliary variables $(\boldsymbol{Z}, \boldsymbol{X}^*)$, *post-stratification* partitions the population into J disjoint post-strata with $U = \bigcup_{j=1}^{J} U_j$ of size N_j and the sample into subsamples with $s = \bigcup_{j=1}^{J} s_j$ of size n_j for the jth post-stratum, correspondingly. With respect to the post-strata, the finite population mean can be rewritten as

$$Q(Y) = \frac{1}{N} \sum_{j=1}^{J} \sum_{i \in U_j} Y_i = \frac{1}{N} \sum_{j=1}^{J} N_j \theta_j,$$

where $\theta_j = \frac{1}{N_j} \sum_{i \in U_j} Y_i$ is subpopulation mean of post-stratum U_j. With the assumption that the sample units in each post-stratum are representative of population units in that post-stratum, the post-strata means are estimated using corresponding subsample means $\hat{\theta}_j = \bar{y}_j = \frac{1}{n_j} \sum_{i \in s_j} y_i$. Naturally, the post-stratification (PS) estimator takes the form

$$\widehat{Q}_{\text{PS}} = \frac{1}{N} \sum_{j=1}^{J} N_j \bar{y}_j = \frac{1}{N} \sum_{i \in s} w_i y_i, \tag{3.1}$$

where $w_i = N_j / n_j$ for $i \in U_j$ is the post-stratification weight assigned to sample unit i in post-stratum j which is inverse proportional to the sampling fraction n_j / N_j. The post-stratification estimator could be numerically unstable when such partition results in small cells in the sample, in other words, small n_j and large weights w_j.

Alternatively, *raking* generates weights w_i to match successively the marginal (rather than the joint) distributions of $(\boldsymbol{Z}, \boldsymbol{X}^*)$ via iterative proportional fitting. The raking weighted estimator takes the form

$$\widehat{Q}_{\text{R}} = \frac{1}{N} \sum_{i \in s} w_i y_i, \tag{3.2}$$

with w_j denoting raking weights. Raking weights could be highly variable, so the resulting weighted estimators could be inefficient. Also, raking may have convergence issues as the number of auxiliary variables increases.

Gelman (2007) reviews a model-based perspective on the PS estimator. In the model-based approach, a regression model is specified to model the conditional distribution of outcome given the discrete auxiliary variables $p(Y|\boldsymbol{Z}, \boldsymbol{X}^*)$. Define stratum-specific means $\theta_j = \mathrm{E}(Y_i|\boldsymbol{Z}_i, \boldsymbol{X}_i^*)$, $i \in U_j$. And estimating $\hat{\theta}_j = \widehat{\mathrm{E}}(Y_i|\boldsymbol{Z}_i, \boldsymbol{X}_i^*)$ based on the fitted model leads to the *regression and post-stratification* (RP) estimator

$$\widehat{Q}_{\mathrm{RP}} = \frac{1}{N}\sum_{j=1}^{J} N_j \hat{\theta}_j. \tag{3.3}$$

As a special case, specifying a saturated regression model (including all possible interactions terms) allows J post-stratum specific means and the least square estimators $\hat{\theta}_j = \bar{y}_j = \frac{1}{n_j}\sum_{i\in s_j} y_i$. As a result, $\widehat{Q}_{\mathrm{RP}} = \widehat{Q}_{\mathrm{PS}}$.

From the model-based perspective, the problem of unstable estimates due to small cells in post-stratification can be viewed as a model fitting problem due to model complexity. Such perspective motivates using alternative modeling techniques to improve estimation. Instead of using classical saturated regression models, *multilevel regression and post-stratification* (MRP) utilizes hierarchical regression models to achieve stable estimates. Both main effects and interaction terms could be specified as multilevel random effects so that information across post-strata can be partially pooled in the model fitting procedure (Gelman and Little, 1997). MRP improves efficiency in the population mean estimation than post-stratification and raking when data are sparse in some post-strata.

Still, it is challenging to perform MRP in high-dimensional setting, especially in the presence of a large number of noise variables not associated with Y, because a parametric form needs to be specified for the multilevel regression. Also, continuous auxiliary variables need to be discretized before modeling.

The model-based RP approach can also be viewed as a prediction approach and the RP estimator in (3.3) can be rewritten as

$$\widehat{Q}_{\text{RP}} = \frac{1}{N}\sum_{j=1}^{J} N_j \hat{\theta}_j = \frac{1}{N}\sum_{j=1}^{J}\sum_{i\in U_j} \hat{\theta}_j = \frac{1}{N}\sum_{j=1}^{J}\sum_{i\in U_j} \widehat{\text{E}}(Y_i|\boldsymbol{Z}_i, \boldsymbol{X}_i^*) = \frac{1}{N}\sum_{i\in U} \widehat{\text{E}}(Y_i|\boldsymbol{Z}_i, \boldsymbol{X}_i^*),$$

where $\widehat{\text{E}}(Y_i|\boldsymbol{Z}_i, \boldsymbol{X}_i^*)$ is predictive value of Y_i based on model $p(Y|\boldsymbol{Z}, \boldsymbol{X}^*)$. Such perspective motivates the use of modern statistical techniques for generalization of inference via valid predictions of the outcomes in the population. Specifically, the classical regression models in *regression and post-stratification* can be replaced by any regularized prediction methods that achieve stable estimates while including high-dimensional covariates. Such models also allows modeling the continuous $\boldsymbol{X}$ directly.

3.2.2 New Approach: Regularized Prediction

Tree-based methods are appealing techniques for handling high-dimensional problems. Sum-of-trees ensembles achieve high prediction accuracy and better approximate the functional forms of continuous variables, with each single tree regularized to obtain stable predictions and achieve bias variance trade-off. Taking a model-based predictive perspective, we extend the RP approach to high-dimensional setting by replacing parametric regression models with regularized additive regression trees. We consider the Bayesian modeling framework, as it is natural to implement predictive inference and straightforward for quantification of uncertainty.

3.2.2.1 BART and Soft BART Prediction

In the current setting, the conditional distribution of a continuous outcome given the high-dimensional auxiliary variables $p(Y|\boldsymbol{Z}, \boldsymbol{X})$ can be modeled using Bayesian additive regression trees (BART) or soft Bayesian additive regression trees (SBART) (Chipman, George, and McCulloch, 2010; Linero and Yang, 2018).

For continuous outcomes, BART and SBART assume Gaussian noise and model the location parameter using a non-parametric sum-of-trees structure, allowing both discrete and continuous auxiliary variables

$$Y = G(\boldsymbol{Z}, \boldsymbol{X}) + \epsilon = \sum_{m=1}^{M} g(\boldsymbol{Z}, \boldsymbol{X}; T_m, \boldsymbol{\mu}_m) + \epsilon, \quad \epsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \tag{3.4}$$

where M is fixed number of trees in the sum-of-trees structure, T_m is the m-th binary tree with $\boldsymbol{\mu}_m$ being the parameters associated with the terminal nodes, and $g(\cdot)$ is the function assigning $\boldsymbol{\mu}_m$ according to $(\boldsymbol{Z}, \boldsymbol{X})$. The sum-of-trees structure naturally handles high-dimensional auxiliary variables without specifying a parametric form, accounting for categorical variables, continuous variables and possible interactions. In the Bayesian framework, quantification of uncertainty is naturally characterized by the posterior and posterior predictive distributions.

In BART, $g(\cdot)$ is a deterministic function and the potential effect of continuous predictors, either linear or nonlinear, is approximated by step functions generated by cutting the continuous predictor at various splitting points in different trees. Regularization priors are specified on $p(T_m)$, $p(\boldsymbol{\mu}_m|T_m)$, $p(\sigma^2)$ such that each single tree T_m is a weak learner. Such specification aims at preventing the individual tree effects from unduly influential and achieving stable predictions, with automatic default specifications facilitating easy imple-

mentation. For $p(T_m)$, the prior is specified by three aspects: (i) the probability that a node is nonterminal, (ii) the distribution on the splitting variable assignments at each interior node, and (iii) the distribution on the splitting rule assignment in each interior node, conditional on the splitting variable. For $p(\boldsymbol{\mu}_m|T_m)$ and $p(\sigma^2)$, conjugate normal distributions and inverse chi-square distributions are specified. In practice, cross validation could be applied to determine the number of trees M and the hyperparameters in the regularization priors. Chipman et al. (2010) introduce default prior specification that puts most probability on tree of sizes 2 and 3 but allows many more terminal nodes if the data demands. According to their experience, as M increases, starting with $M = 1$, the predictive performance improves dramatically until at some point it levels off and then begins to degrade very slowly for large values of M. Therefore, it's important to avoid M being too small.

In SBART, $g(\cdot)$ associates the values of covariates with a probabilistic (instead of deterministic as in BART) path down the tree, with certain probability going left at each node. With such modification, a particular set of values of $(\boldsymbol{Z}, \boldsymbol{X})$ is associated with a certain terminal node with certain probability, obtained by averaging over all possible paths. Unlike hard decision trees in BART where each terminal node is constrained to influence the regression function locally, the soft decision trees in soft BART allow each terminal node to impose a global effect on the function. This global effect of local terminal nodes enables the soft decision trees to borrow information adaptively across different covariate regions. Sparsity-inducing priors are specified to achieve a balance between sparse and non-sparse settings. Linero and Yang (2018) develop default prior specification with $M = 50$ which performs universally well in all the 10 benchmark datasets considered in the paper. Cross

validation could be applied for parameter tuning, but it only substantially improves performance in one dataset where tuning M is required to attain optimal performance. Therefore, in practice, it is sufficient to only tune M to reduce computational burden.

The BART and SBART prediction estimators of finite population mean, $\widehat{Q}_{\text{BART}}$ and $\widehat{Q}_{\text{SBART}}$, are obtained with the following steps.

Step 1 Model $p(Y|\boldsymbol{Z}, \boldsymbol{X})$ using BART or soft BART, $Y = G(\boldsymbol{Z}, \boldsymbol{X}) + \epsilon, \epsilon \sim N(0, \sigma^2)$ with corresponding Bayesian priors.

Step 2 Obtain posterior distributions of $Q(Y) = \frac{1}{N}\sum_{i\in U} y_i$ using Markov chain Monte Carlo (MCMC) simulations. Specifically, in MCMC iteration t,

1. draw $G^{(t)}, \sigma^{(t)}|Y_{i\in s}, \boldsymbol{Z}_{i\in U}, \boldsymbol{X}_{i\in U}$
2. compute $\tilde{\theta}_i^{(t)} = G^{(t)}(\boldsymbol{Z}_i, \boldsymbol{X}_i)$ for $i \in U$
3. obtain $\widehat{Q}^{(t)}_{\text{(S)BART}} = \frac{1}{N}\left[\sum_{i\in U}\tilde{\theta}_i^{(t)} + \left(\sum_{i\in s} y_i - \sum_{i\in s}\tilde{\theta}_i^{(t)}\right)\right]$, using the observed y_i in the sample and the predicted values for the population units that are not in the sample.

Step 3 Obtain $\widehat{Q}_{\text{(S)BART}}$: point estimates using (posterior) median of $\widehat{Q}^{(t)}_{\text{(S)BART}}$ with credible intervals constructed using quantiles splitting the tails of posterior distribution equally.

In some cases, inference on subpopulation means are also of interest, which can be obtained via modification of item 3 in Step 2, restricting the average to predictions and observed outcomes in the corresponding subpopulation $\Omega \subset U$ and subsamples $s \cap \Omega$, $\widehat{Q}^{(t)}_{\Omega,\text{(S)BART}} = \frac{1}{N_\Omega}\left[\sum_{i\in\Omega}\tilde{\theta}_i^{(t)} + \left(\sum_{i\in s\cap\Omega} y_i - \sum_{i\in s\cap\Omega}\tilde{\theta}_i^{(t)}\right)\right]$.

3.2.2.2 BART and Soft BART Propensity Prediction

In the missing data literature, Little and An (2004) proposed including logit-transformed response propensity score as covariates using splines in the imputation models. This response propensity prediction method yields robust estimates of sample means when the imputation model is misspecified. Tan, Flannagan, and Elliott (2019) extended the method of Little and An (2004) by using BART to fit both the imputation model and the response propensity model. They show that adding BART-estimated propensity score in the BART imputation model reduces bias and RMSE and improves confidence interval coverage rates in the mean estimation.

Inspired by this, we extend the BART and SBART prediction with a two-step approach. First, we estimate sample inclusion propensity using a propensity model. If the sample data are linked to the population data, we code the sample inclusion indicators $I = 1$ for the units in the sample and $I = 0$ for the rest of the units in the population. The propensity score $\hat{\pi}$ can then be estimated via modeling $p(I|\boldsymbol{Z}, \boldsymbol{X})$ using probit Bayesian additive regression trees (Chipman et al., 2010). If the sample data is unlinked to the population data, we round the continuous $\boldsymbol{X}$ to $[\boldsymbol{X}]$ at a certain precision level and identify K categories with unique values of $(\boldsymbol{Z}, [\boldsymbol{X}])$. Within each category $k = 1, \ldots, K$, the number of units in the population N_k and that in the sample n_k can be counted. Once the counts (N_k, n_k) are created for each category, the propensity score $\hat{\pi}$ for the units to be included in the sample, given $(\boldsymbol{Z}, [\boldsymbol{X}])$, can be obtained via models for binomial outcomes. Next, we model $p(Y|\boldsymbol{Z}, \boldsymbol{X}, \hat{\pi})$ by additionally including $\hat{\pi}$ as a covariate in BART or SBART model with the rest of the steps being the same as Section 3.2.2.1. The detailed steps of obtaining the BART

propensity (BART-P) prediction estimator $\widehat{Q}_{\text{BART-P}}$ and the SBART propensity (SBART-P) prediction estimator $\widehat{Q}_{\text{SBART-P}}$ are outlined as follows.

Step 1 Model $p(I|\boldsymbol{Z}, \boldsymbol{X})$ with probit BART and estimate $\hat{\pi}$ using posterior mean

Step 2 Obtain the (S)BART-P prediction estimator for finite population mean

- model $p(Y|\boldsymbol{Z}, \boldsymbol{X}, \hat{\pi})$ using (S)BART, $Y = G(\boldsymbol{Z}, \boldsymbol{X}, \hat{\pi}) + \epsilon, \epsilon \sim N(0, \sigma^2)$
- estimate $\tilde{\theta}_i^{(t)} = G^{(t)}(\boldsymbol{Z}_i, \boldsymbol{X}_i, \hat{\pi}_i)$
- $\widehat{Q}^{(t)}_{\text{(S)BART-P}} = \frac{1}{N}\left[\sum_{i\in U} \tilde{\theta}_i^{(t)} + \left(\sum_{i\in s} y_i - \sum_{i\in s} \tilde{\theta}_i^{(t)}\right)\right]$
- $\widehat{Q}_{\text{(S)BART-P}}$: point estimates using (posterior) median of $\widehat{Q}^{(t)}_{\text{(S)BART-P}}$ with credible intervals constructed using quantiles splitting the tails of posterior distribution equally.

BART-P and SBART-P prediction methods are expected to be doubly robust. More specifically, as long as either of the mean model for the outcome or the propensity model is correctly specified, a consistent estimator of the population mean is obtained.

3.3 Simulation Studies

3.3.1 Simulation Design

Artificial populations with size $N = 3,000$ were simulated. For each unit i in the population, a total number of p binary auxiliary variables and r continuous variables were generated. The p binary variables $\{Z_{il}\}_{l=1,\dots p}$ were obtained with $Z_{il} = I(W_{il} < U_l)$, where $\{W_{il}\} \overset{\text{i.i.d}}{\sim} \text{N}(0, 1)$ and $U_l \overset{\text{i.i.d}}{\sim} \text{U}(-.4, .4)$, so that $\Pr(Z_{il} = 1)$ falls in the range $(.34, .66)$, $l = 1, \dots, p$. The r

continuous $\{X_{il}\}_{l=1,...r}$ were generated independently from $\mathrm{U}(0,1)$. Samples of size $n = 600$ were drawn from the populations with inclusion probability $\pi = \Pr(I = 1|\boldsymbol{Z}, \boldsymbol{X})$ as a function of the auxiliary variables $\boldsymbol{X}$ and $\boldsymbol{Z}$. We considered the following four simulation scenarios:

S1 Low-dimensional auxiliary variables ($p = 3, r = 1$) with higher inclusion propensity at the lower tail of X_1. The outcomes $\{Y_i\}_{i=1,...,N}$ were generated using an additive model: $Y = 26.81 - Z_1 - 2Z_2 - 3.5Z_3 - 25(X_1 - .75)^2 + \epsilon, \epsilon \sim \mathrm{N}(0, 3^2)$, and the samples were selected with $\pi \propto \mathrm{logit}^{-1}[-13.66 + .5Z_1 + Z_2 + 1.75Z_3 + 12.5(X_1 - .75)^2]$. Consequently, units with values of X_1 falling between 0.5 and 1 were under-sampled.

S2 High-dimensional auxiliary variables ($p = 30, r = 10$) with higher inclusion propensity at the lower tail of X_1. Same Y and π models as S1, but add noise auxiliary variables $\{Z_l\}_{l=4,...,30}$ and $\{X_l\}_{l=2,...,10}$ that are not associated with Y or π.

S3 High-dimensional auxiliary variables ($p = 30, r = 10$) with lower inclusion propensity at the lower tail of X_1. Same as S2, but change the signs of the coefficients in the model for π to introduce selection bias in the opposite direction: $\pi \propto \mathrm{logit}^{-1}[4.01 - .5Z_1 - Z_2 - 1.75Z_3 - 12.5(X_1 - .75)^2]$. Consequently, units with small values of X_1 were under-sampled, especially among those with $X_1 \leq 0.25$.

S4 High-dimensional auxiliary variables ($p = 30, r = 10$) with interaction and different relevant continuous predictors for Y and π. The outcomes $\{Y_i\}_{i=1,...,N}$ were generated using $Y = 36.81 - Z_1 - 2Z_2 - 3.5Z_3 - 10Z_1Z_2 - 9(X_1 - .75)^2 - 16Z_3(X_1 - .75)^2 + \epsilon, \epsilon \sim \mathrm{N}(0, 3^2)$, with samples selected using $\pi \propto \mathrm{logit}^{-1}[3.27 - .5Z_1 - Z_2 - 1.75Z_3 -$

$2Z_1Z_2 - 4(X_3 - .75)^2 - 3Z_3(X_3 - .75)^2 - (X_5 - .75)^2]$. Units at the tails of X_3 and X_5 were under-sampled,but X_3 and X_5 were not associated with Y.

Figure 3.2(a)-(b) show the scatter plots of Y against X_1, the continuous variable that is associated with both Y and π, of the simulated population overlaid with a selected sample in scenarios S1-S3. Population units with lower values of X_1 were more likely to be selected into samples in scenarios S1/S2 but less likely to be selected in scenario S3.

Scenario S4 was designed to assess whether tree-based methods handle interactions well and how they perform when the continuous variables that are associated with undersampling are not associated with outcome. Figure 3.2(c) visualizes population with a selected sample in scenario S4, using scatter plots of Y against X_1, the continuous variable related to Y but not π, and of Y against X_3, the continuous variables related to π but not Y. The plot on the left shows a positive association between Y and X_1 but units with different values of X_1 are equally likely to be included in the sample; while the plot on the right shows no association between Y and X_3 but units at the lower tail of X_3 are less likely to be included in the sample.

For each scenario, 500 replicates of simulation were conducted, with point and interval estimates of finite population mean computed for each. The Bayesian tree-based methods used all available auxiliary variables, as it is unknown which variables are involved in the true data generating process in practice. For scenario S1 with low-dimensional auxiliary variables, the tree-based methods were also compared to the PS and raking estimators using all four available variables with X_1 discretized using tertiles in PS and using quintiles in raking. Raw estimates were also calculated using sample means.

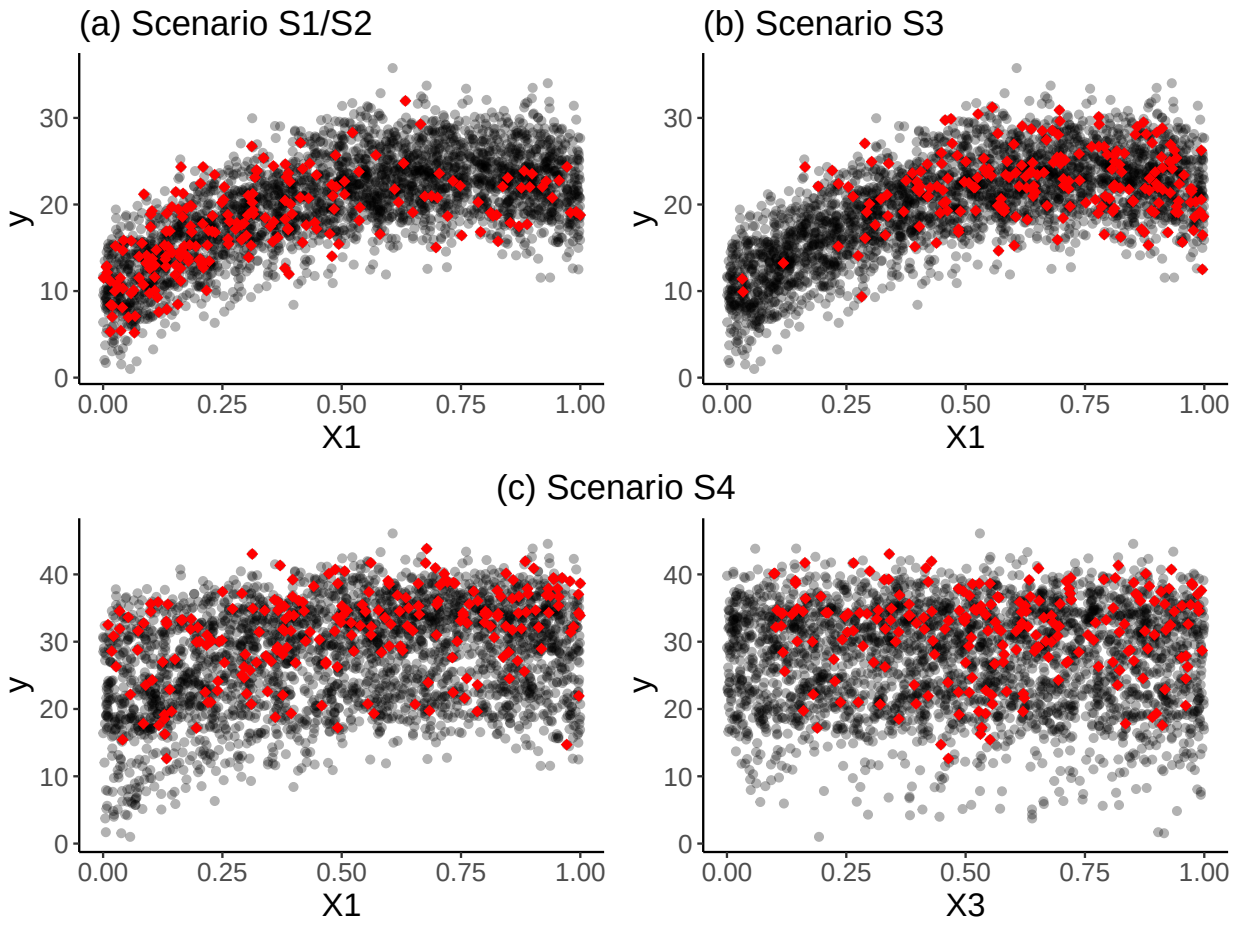

Figure 3.2: Scatterplots of outcomes Y versus continuous auxiliary variables of units in the population (in black dots) and a selected sample (in red diamonds) for (a) Scenario S1/S2 (b) Scenario S3 (c) Scenario S4

3.3.2 Simulation Results

The performance of point estimates are evaluated with empirical bias and empirical root mean squared error (RMSE), summarised in Table 3.1. Scenarios S1-S3 share the same outcome model, the same outcome values $\{Y_i\}_{i=1,\ldots,N}$ and the same ground truth for the finite population mean defined as $Q = \frac{1}{N}\sum_{i=1}^{N} Y_i$. The empirical coverage rates and average widths of 80% and 95% probability intervals are visualized in Figure 3.3. The raw estimates ignoring selection bias are off the chart, leading to confidence intervals with 0% coverage rates, therefore, not shown in Figure 3.3. For BART and BART-P, $M = 50, 100, 200$ trees were explored, with $M = 50$ trees performing the best and reported. For SBART and SBART-P, the default specification suggested by Linero and Yang (2018) was used with $M = 50$.

In scenario S1, where the weighting methods are feasible, raking is less biased as well as more efficient than post-stratification (PS). This is because raking maintains more information from the continuous variable X_1 by discretizing X_1 using quintiles as compared to tertiles in PS, and raking implicitly assumes an additive propensity model while PS assumes an interaction model. Both PS and raking generate confidence intervals with coverage rates lower than the nominal levels, with raking yielding shorter intervals but higher coverage rates. BART and SBART both outperform the weighting methods via utilizing the continuous form of X_1, generating credible intervals with coverage rates close to the nominal levels. BART and SBART perform similarly as all auxiliary variables are relevant in this low-dimensional setting. Including propensity score in BART and SBART leads to a small bias reduction which is offset by efficiency loss, indicated by slightly higher RMSE and

Table 3.1: Simulation results - empirical bias and RMSE of various methods in estimating population means, from 500 simulation replicates, for each simulation setting

Method	S1		S2		S3		S4	
	$Q = 19.88$						$Q = 27.74$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
raw	−2.99	2.99	−2.99	2.99	2.43	2.43	3.13	3.14
PS*	−0.37	0.43						
raking**	−0.16	0.22						
BART	−0.08	0.17	−0.17	0.22	0.37	0.43	0.07	0.17
BART-P	−0.06	0.18	−0.12	0.20	0.30	0.38	0.06	0.17
SBART	−0.08	0.17	−0.10	0.19	0.24	0.32	0.04	0.16
SBART-P	−0.07	0.18	−0.10	0.19	0.24	0.32	0.04	0.16

Note 1: *PS is based on Z_1, Z_2, Z_3 and X_1 discretized using tertiles; **Raking is based on Z_1, Z_2, Z_3 and X_1 discretized using quintiles.

Note 2: The standard errors of empirical bias from 500 simulation replicates are $< 7.5 \times 10^{-3}$ for all methods

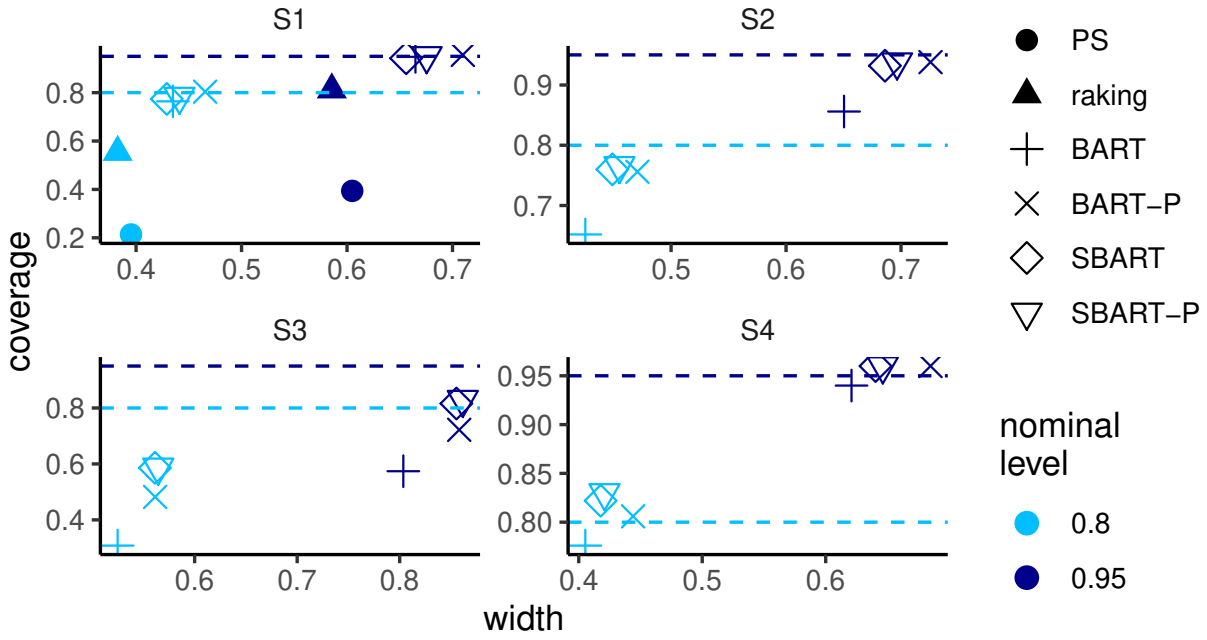

Figure 3.3: Simulation results - empirical coverage rates of 80% and 95% probability intervals (with the horizontal dashed lines denoting the nominal levels) against average probability interval widths, from 500 simulation replicates, for each simulation setting

slightly wider credible intervals.

Scenario S2 differs from scenario S1 by adding irrelevant auxiliary variables. PS and raking are not feasible due to high-dimensionality. Units with X_1 falling between 0.5 and 1.0 have lower selection probabilities than those with X_1 in between 0 and 0.5 as shown in Figure 3.2(a). SBART outperforms BART with lower bias, lower RMSE and better credible interval coverage. Including propensity score in BART reduces bias and RMSE and fixes credible interval coverage. However, such improvement is not obvious for SBART. BART-P, SBART and SBART-P all yields valid credible intervals but not BART, with SBART having the shortest intervals.

Scenario S3 differs from scenario S2 in the direction of selection bias. Moreover, because π was negatively associated with $(X_1 - .75)^2$, the units in the lower tail of X_1 (e.g. $X_1 < .25$ in Figure 3.2(b)) have even smaller inclusion probabilities than the units in the upper tail of X_1 in scenario S2. Consequently, there are sparse data in the lower tail of X_1. In this setting, neither BART nor SBART performs well with large bias and RMSE, although SBART yields smaller bias and RMSE than BART. The empirical coverage rates for both BART and SBART are lower than the nominal levels due to bias in the estimation. By including propensity score, BART-P improves credible interval coverage as well as bias and RMSE than BART, but does not fix the undercoverage issue. Again, such improvement is not obvious for SBART.

In scenario S4, both BART and SBART performs well with small bias and RMSE and close to nominal level coverage rate. SBART yields slightly smaller bias, smaller RMSE, and better coverage rate than BART. Including propensity score slightly reduces bias and

improves coverage rate in BART. Although there are sparse data in the lower tails of X_3 and X_5, these two X variables are not associated with Y and thus such biased selection did not yield poor performance of the tree-based methods like in Scenario S3.

In all the scenarios considered here, SBART outperforms other competing methods and is, therefore, recommended. However, it should be used with caution, as it still does not perform well when selection bias results in sparse data at the tails of continuous auxiliary variables associated with the outcome.

3.3.3 Comparison of BART and SBART Prediction

We took a further investigation to compare the performance of BART and SBART in scenario S3, where neither BART nor SBART performs well with BART performing worse than SBART. We consider two random samples from the population. For sample I, data is sparse at the lower tail of X_1, while, for sample II, no data is available at the lower tail, $X_1 < .2$. The top panels I(a) and II(a) of Figure 3.4 shows the population in block dots, with sample I and II in red, respectively. For a closer examination of the data at the lower tail of X_1, we restrict to a subset with $Z_2 = Z_3 = 0$ and focus on the lower tail with $X_1 < 0.3$. The middle panels I(b) and II(b) show the population and corresponding sample data in this restricted subgroup. Finally, the bottom panels I(c) and II(c) plot the population units of Y in this subgroup (gray points) overlaid by the posterior means of the location parameters, $G(\boldsymbol{Z}, \boldsymbol{X})$, of each population units estimated using BART and SBART as shown using red pluses and blue crosses, respectively. Panel I(c) shows that both BART and SBART fit the data well within the region $X_1 > .2$ where sample data are available. However, the SBART fit the data much better than BART when $X_1 < 0.2$ where very sparse data are available.

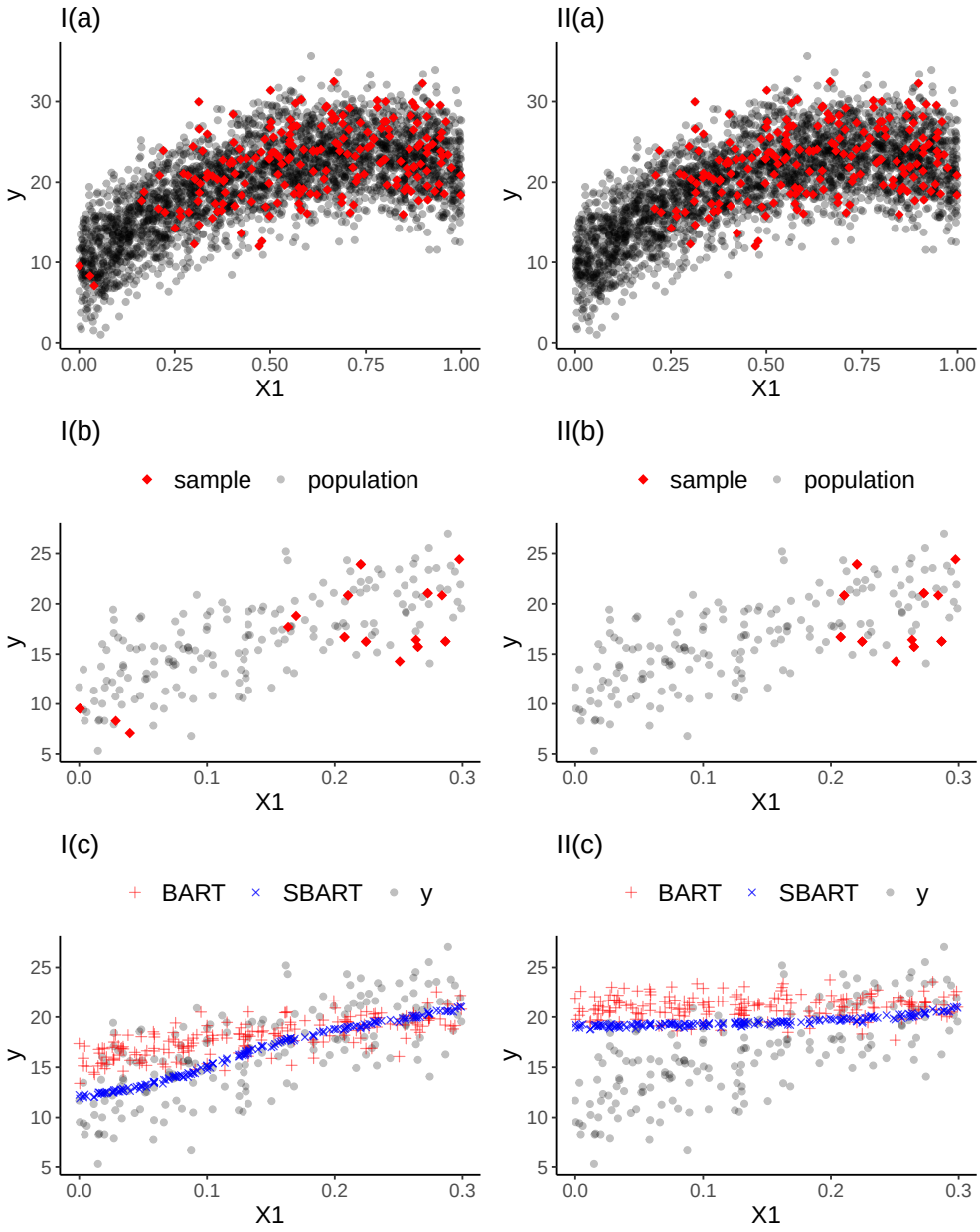

Figure 3.4: Two selected samples I and II from the population in Scenario S3: (a) Scatter plots of Y versus X_1 with the population in gray dots and a selected sample in red diamonds (b) Scatter plots of Y versus X_1, restricted to $Z_2 = Z_3 = 0$ and $X_1 < .3$ (c) Scatter plots of Y versus X_1 in the subpopulation, overlapped with posterior means of $G(\boldsymbol{Z}, \boldsymbol{X})$ estimated from the BART and SBART models based on the whole sample.

The estimated posterior means of the location parameters based on SBART are also less noisy, due to the sparsity-induced priors that tend to exclude the noise auxiliary variables in model fitting. Panel II(c) shows that both models fail to produce valid predictions in the region $X_1 < .2$ where there is no sample data available. In the simulation study, about 5% of the 500 simulated samples do not include units with $X_1 < .1$ and one third include fewer than 10 units with $X_1 < .2$.

3.4 Applied Examples

We demonstrate the application of the proposed methods using real data from two different studies. The first application example deals with a mental health survey assessing psychiatric disorders among the Ohio Army National Guard (OHARNG) service members. The second application is in a clinic setting where it is of interest to generalize inference on COVID-19 patients when clinical outcomes are only available in a subset of patients.

3.4.1 Ohio Army National Guard Survey of Mental Health

The Ohio Army National Guard (OHARNG) Mental Health Initiative is a population-based observational survey study for estimating the prevalence and identifying correlates of mental illness and health service utilization among the OHARNG service members. The study population of the baseline survey is defined as all $N = 12570$ soldiers who served in the OHARNG between June 2008 and February 2009. A survey sample with $n = 2562$ service members was selected. In this analysis, we are interested in estimating the mean trauma score among the OHARNG service members using the selected sample, with potential selection

bias due to under-coverage of sampling frame and non-response. Auxiliary information is available at individual level for the entire study population, including age (17-24 yr, 25-34 yr, 35+ yr), sex, race (Whites, Black, Other), rank (enlisted, officer), marital status (married, non-married), and years of service (in years). We apply the proposed trees-based methods to correct the discrepancy between the sample and population utilizing the five categorical and one continuous auxiliary variables. For BART-P and SBART-P, the propensity models were built using probit BART.

Before modeling, $\log(y+1)$ transformation was applied to trauma scores to reduce right skewness such that the normality assumption in BART and SBART is reasonable. Distributions of the only continuous variable, years of service, in the sample and population were checked to avoid prediction failure due to sparse data at the tails (see Figure S1 in Appendix B). After fitting the models, we performed model checking using posterior predictive graphics checking (Gelman et al., 2014, chapter 6) based on the following test quantities, including (a) $T_1(\boldsymbol{y}) = \bar{y}$, (b) $T_2(\boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, and (c) $T_3(\boldsymbol{y}, G, \sigma) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\theta_i}{\sigma}\right)^2$, where $\theta_i = G(\boldsymbol{z}_i, \boldsymbol{x}_i)$. The test quantities catch different aspects of the data, with $T_1(\cdot)$ and $T_2(\cdot)$ measuring the location and variability of the survey outcome while $T_3(\cdot)$ measuring the discrepancy between the survey outcome and fitted distribution. In each MCMC iteration t, the realized test quantities $T_i(\boldsymbol{y}, G^{(t)}, \sigma^{(t)})$ under the observed data and predictive test quantities $T_i(\tilde{\boldsymbol{y}}^{(t)}, G^{(t)}, \sigma^{(t)})$ under the simulated data were computed and compared, with $\tilde{\boldsymbol{y}}^{(t)}$ drawn from the posterior predictive distribution. For each quantity $T_i(\cdot)$, a Bayesian posterior predictive p-value can also be computed, which is defined as the probability that the predictive test quantity is greater than the realized test quantity, evaluated over the

posterior distribution. The Bayesian p-value measures the discrepancy between the observed data and the posterior predictive distribution in the aspect characterized by $T(\cdot)$. A Bayesian p-value close to 0.5 indicates good fit while a Bayesian p-value near 0 or 1 indicates that the observed pattern would be unlikely to happen if the model were true and, therefore, lack of fit. Figure S2 in Appendix B shows the posterior predictive graphics checking and corresponding p-values for SBART. All four Bayesian methods, BART, BART-P, SBART and SBART-P, yielded fitted models with posterior predictive p-values close to 0.5, indicating adequate model fit.

We compare the results of proposed Bayesian methods with the raw estimates in estimating the mean trauma score on the log scale, with point estimates and 95% probability intervals visualized in Figure 3.5(a). The Bayesian methods yields lower estimates for mean trauma score compared to raw estimates without adjustment. BART and SBART yields similar results, as this is a low-dimensional setting with one continuous auxiliary variable and the benefit of soft decision trees is not so obvious. Including propensity scores do not lead to much change in the estimates. We recommend reporting the estimates using SBART in this analysis.

3.4.2 New York City COVID-19 Study

The COVID-19 is a global pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2). The first positive case was confirmed in New York City on March 1, 2020. The city had more cases than any country other than the United States by May 2020. The urgent need for therapeutic agents has resulted in repurposing and redeployment of experimental agents. Hydroxychloroquine, combined with azithromycin, was widely ad-

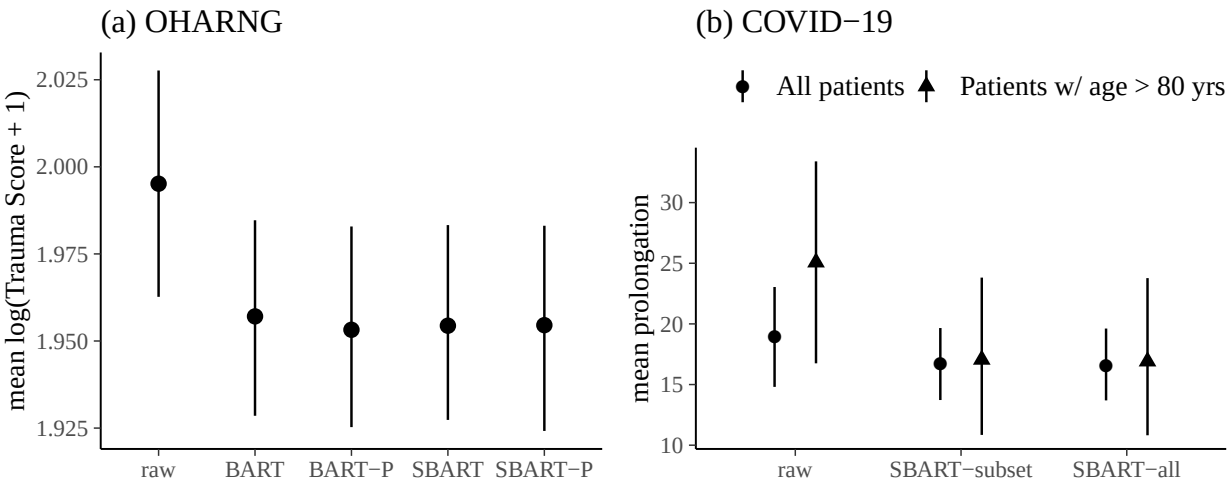

Figure 3.5: (a) Point estimates and 95% probability intervals of mean log(trauma score + 1) among soldiers who served in the OHARNG between June 2008 and February 2009 (b) Point estimates and 95% probability intervals of mean prolongation among all patients and patients with age $\geq$ 80 years old, comparing raw sample means, SBART with baseline QTc and treatment (SBART-subset), and SBART with all auxiliary variables (SBART-all).

ministered to patients with COVID-19 without robust evidence supporting its use, with the U.S. Food and Drug Administration (FDA) issuing an emergency use authorization (EUA) to allow doctors to begin treating patients with hydroxychloroquine in hospitalized settings outside clinical trials on March 28, 2020. Such EUA was later revoked as of June 15, 2020, with a randomized clinical trial in hospitalized patients showing no benefits and reports of serious heart rhythm problems along with other safety issues. Both hydroxychloroquine and azithromycin are characterized as definite QTc prolongers that increase risk of sudden cardiac deaths. Between March 1st, 2020 through May 1st, 2020, there were 470 patients admitted to Columbia University Irving Medical Center, treated with hydroxychloroquine (H+) or hydroxychloroquine combined with azithromycin (A+H+). All patients have baseline ECG measurements of QTc while, due to lack of personal protective equipment, only 244 of them have ECG QTc measurements on Day 2 of medication. We are interested in estimating the mean QTc prolongation, defined as difference in QTc measures between day 2 and day 0, of all the 470 COVID-19 patients who received H+ or A+H+ treatments. However, the QTc prolongation was only measured among the 244 patients who had QTc measurements at day 2. To improve the estimation, we also collected the data of these 470 patients on their demographic characteristics and relevant biomarkers from electronic medical records.

Exploratory analysis indicates a strong negative association between prolongation and baseline QTc measurement and that patients with higher baseline QTc are less likely to have ECG QTc measurement on Day 2, demonstrated in Figure S3 in the Appendix B. Other auxiliary variables include treatment (H+, A+H+), demographic characteristics, including gender, age (in years), race (white, black, other), BMI (log scale), along with 7 biomarkers.

We used the recommended method SBART for two estimands of interest: (i) the mean QTc prolongation among all 470 patients, and (ii) the mean QTc prolongation among the 87 (out of 470) patients who were over 80 years old. We compared two SBART models, with the first model only including baseline QTc and treatment (SBART-subset) and the second model including all covariates (SBART-all).

As is visualized in Figure 3.5(b), SBART yields lower estimates of mean QTc prolongation compared to the raw estimates ignoring selection bias, for both estimands of interest. For estimand (i), including baseline QTc and treatment in SBART leads to obvious drop in the mean prolongation estimates from a raw estimate of 18.9 (95% CI : $14.8, 23.0$) milliseconds to 16.7 (95% CI : $13.7, 19.7$) milliseconds. Additionally adding other auxiliary variables does not lead to further obvious change in the estimates. As is visualized in Figure 3.5(b), SBART yields lower estimates of mean QTc prolongation compared to the raw estimates ignoring selection bias, for both estimands of interest. For estimand (i), including baseline QTc and treatment in SBART leads to obvious drop in the mean prolongation estimates from a raw estimate of 18.9 (95% CI : $14.8, 23.0$) milliseconds to 16.7 (95% CI : $13.7, 19.7$) milliseconds. Additionally adding other auxiliary variables does not lead to further obvious change in the estimates.

3.5 Discussion

We consider generalization of inference on a descriptive estimand from a non-random sample to a target population in data-rich settings where high dimensional auxiliary information is available in both the sample and population, with survey inference being a special case.

Existing methods such as post-stratification and raking are challenging or infeasible to be performed due to high-dimensionality and the need to discretize continuous auxiliary variables before applying such methods leads to loss of information. To address such issues, we propose a regularized prediction approach by modeling the conditional distribution of the outcomes given the high-dimensional auxiliary variables using Bayesian machine learning techniques. In this paper, we specifically consider Bayesian additive regression trees (BART) and soft BART which handles both discrete and continuous auxiliary variables as well as potential interactions. Besides the auxiliary variables, we also consider modified methods that estimates the propensity score for a unit to be included in the sample and also include the estimated propensity score as a covariate in the BART and soft BART model.

Artificial data simulation studies demonstrate that the Bayesian additive-trees-based methods outperform post-stratification (PS) and raking in low-dimensional settings where PS and raking are feasible, as the regularized additive trees better utilize information in the continuous auxiliary variables and avoid model overspecification. The Bayesian additive-trees-based methods also yield valid inference in high-dimensional settings when PS and raking are not feasible, as long as selection bias does not result in sparse data points at the tails of relevant continuous auxiliary variables. In high-dimensional setting with sparse signals, SBART, with soft decision trees and sparsity-inducing priors, is less biased and more efficient than BART. In challenging settings where the additive-trees-based methods underperform, including propensity score in BART could reduce bias and improve credible interval coverage while such benefit is not obvious for SBART. Therefore, the soft BART prediction method is recommended for generalization of inference with high-dimensional

auxiliary variables. The soft BART better utilizes information in the continuous auxiliary variables and more effectively regularize the effect of irrelevant noise auxiliary variable. As is demonstrated in the OHARNG mental health study and the COVID-19 study, the proposed methods could be applied in both survey and more general settings, with estimands being overall population as well as subpopulation quantities.

The Bayesian additive-trees-based methods need to be used with caution. More specifically, both BART and SBART prediction fail when selection bias results in very sparse data point at the tails of the continuous covariates. Such prediction failure cannot be fixed via robust methods involving propensity scores. Therefore, for important continuous variables associated with the outcomes, the range and distribution in the sample and population needs to be checked before using the methods. In some cases, transformation on such auxiliary variables could be applied to reduce sparsity at the tails.

Although BART and SBART are considered in this paper, the regularized prediction approach is general and any Bayesian machine learning techniques that achieve valid predictions could be applied.

In real world applications, missing data could arise in high-dimensional data-rich settings. Depending on the missing proportion, single or multiple imputation could be used to impute the missing values before applying the proposed methods. Imputation can be performed using machine learning techniques such as random forests or (soft) BART, depending on the available computational resource. In future work, it would be interesting to assess the impact of missing auxiliary information under various missing data mechanism, including potential bias and imputation uncertainty.

Acknowledgement

We thank Dr. Sandro Galea for sharing the data on the OHARNG service members, Drs. Elaine Wan and Marc Waase for sharing the data on COVID-19 patients admitted at Columbia University Irving Medical Center.

Chapter 4

Survey Design for Multilevel Regression and Post-Stratification

4.1 Introduction

Multilevel regression and post-stratification (MRP) has been widely applied in survey analysis, for both probability and non-probability samples. The technique was originally developed by Gelman and Little (1997) and subsequently expanded by Park, Gelman, and Bafumi (2004, 2006) to estimate state-level public opinions. Wang, Rothschild, Goel, and Gelman (2015) demonstrates, through 2012 US presidential election forecast with non-representative voter intention polls on the Xbox gaming platform, that multilevel regression and post-stratification (MRP) can be used to generate accurate survey estimates from non-representative samples. Besides political science, MRP has also been used in the field of epidemiology by Downes et al. (2018).

In spite of the rich literature on survey analysis using MRP, literature on survey design

for MRP is scarce. Empirical studies using simulation could be considered in the design stage, but it is computationally intensive. Therefore, it is of interest to develop theoretical results.

In this chapter, we consider survey design for MRP. We propose a close form formula to calculate margin of errors given the design parameters and validate the theoretical results using simulation studies. We demonstrate the use of the formula in two survey design scenarios, online panels using quota sampling and telephone surveys with fixed total sample sizes.

4.2 Methods

4.2.1 An Overview of Multilevel Regression and Post-Stratification

When there is discrepancy between a survey sample and the target population of interest, post-stratification corrects for the known differences by partitioning the population into a series of disjoint subpopulations (post-strata) such that the subsamples in the corresponding post-strata are representative. Consider a finite population of size N with R categorical auxiliary variables $\{Z_r\}_{r=1,\ldots,R}$, with Z_r having J_r levels. The joint distribution of the auxiliary variables generates a total number of $J = \prod_{r=1}^{R} J_r$ cells, labeled as $j = 1, \ldots, J$. Therefore, the finite population U can be partitioned into J disjoint post-strata $U = \bigcup_{j=1}^{J} U_j$ of size N_j such that $N = \sum_{j=1}^{J} N_j$. Without loss of generality, we consider a continuous survey variable of interest Y with the estimand of interest being the the finite population mean $Q(Y) = \frac{1}{N} \sum_{i \in U} Y_i$. With respect to the post-strata, the finite population mean can be

rewritten as

$$Q(Y) = \frac{1}{N}\sum_{j=1}^{J}\sum_{i\in U_j} Y_i = \frac{1}{N}\sum_{j=1}^{J} N_j\theta_j,$$

where $\theta_j = \frac{1}{N_j}\sum_{i\in U_j} Y_i$ is subpopulation mean of post-stratum U_j. With valid estimators $\hat{\theta}_j$ for the post-stratum means, the post-stratification estimator takes the form

$$\widehat{Q}_{\text{PS}} = \frac{1}{N}\sum_{j=1}^{J} N_j\hat{\theta}_j.$$

Multilevel regression and post-stratification specifies a multilevel regression model to model the conditional distribution of the survey outcome given the auxiliary variables $p(Y|\boldsymbol{Z})$ and predicts the post-stratum means based on the model fitted with sample survey data. For a continuous outcome, with a sample of size n, a normal model with multilevel varying intercept can be specified and fitted $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma_y^2\boldsymbol{I})$, where $\boldsymbol{X}_{n\times k}$ is a design matrix of binary indicators created from the auxiliary variables, with prior $N(\boldsymbol{0}, \Sigma_{\boldsymbol{\beta}})$ on the regression coefficients $\boldsymbol{\beta}_{k\times 1}$.

4.2.2 An Illustration Example Using OHARNG

As a illustration example, we consider a study population defined as all $N = 12570$ soldiers who served in the Ohio Army National Guard (OHARNG) between June 2008 and February 2009, with two auxiliary variables gender and age. Table 4.1 lists the $J = 2 \times 4 = 8$ post-stratification cells defined by gender with 2 levels and age with 4 levels. Most service members are male and most are less than 55 years old. Independent estimation of the mean response of a survey outcome of interest in post-stratum $j = 8$ could be difficult, if a sample do not include enough females who are ≥ 55 years old.

Table 4.1: Definition of post-strata and corresponding post-strata sizes for OHARNG

Gender	Age, in years	label j	N_j
Male	$17-24$	1	3269
	$25-34$	2	3292
	$35-44$	3	1949
	≥ 55	4	783
Female	$17-24$	5	774
	$25-34$	6	454
	$35-44$	7	194
	≥ 55	8	63

A multilevel model with the following term could be specified to partially pool information across post-strata:

- A constant term μ_0
- An indicator for female μ_1
- 4 indicators for age categories $\alpha_1, \ldots, \alpha_4$
- 8 indicators for gender $\times$ age interactions $\gamma_1, \ldots, \gamma_8$.

The model has $k = 1+1+4+8 = 14$ coefficients $\boldsymbol{\beta} = (\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\gamma})'$, and a prior precision matrix could have the form $\Sigma_{\boldsymbol{\beta}}^{-1} = \text{Diag}(0, 0, \sigma_\alpha^{-2}, \ldots, \sigma_\alpha^{-2}, \sigma_\gamma^{-2}, \ldots, \sigma_\gamma^{-2})$, with the parameters σ_α and σ_γ estimated from the data. A fully Bayesian approach would also involve prior specification for $\pi(\boldsymbol{\mu})$, $\pi(\sigma_\alpha)$, $\pi(\sigma_\gamma)$ and $\pi(\sigma_y)$. To set up the design matrix, a vector $\boldsymbol{x}_j$ can be defined correspondingly to represent each post-stratum j, with a matrix $\boldsymbol{X}^{\text{pop}}$ presenting all cells in

the population as follows:

$$\boldsymbol{X}^{\text{pop}} = \begin{array}{c} \\ \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \boldsymbol{x}_3 \\ \boldsymbol{x}_4 \\ \boldsymbol{x}_5 \\ \boldsymbol{x}_6 \\ \boldsymbol{x}_7 \\ \boldsymbol{x}_8 \end{array} \begin{array}{c} \begin{array}{cccccccccccccc} \mu_0 & \mu_1 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 & \gamma_6 & \gamma_7 & \gamma_8 \end{array} \\ \begin{pmatrix} 1 & & 1 & & & & 1 & & & & & & & \\ 1 & & & 1 & & & & 1 & & & & & & \\ 1 & & & & 1 & & & & 1 & & & & & \\ 1 & & & & & 1 & & & & 1 & & & & \\ 1 & 1 & 1 & & & & & & & & 1 & & & \\ 1 & 1 & & 1 & & & & & & & & 1 & & \\ 1 & 1 & & & 1 & & & & & & & & 1 & \\ 1 & 1 & & & & 1 & & & & & & & & 1 \end{pmatrix} \end{array}.$$

The ith row of the design matrix $\boldsymbol{X}$ takes value $\boldsymbol{x}_j$ if unit i in the sample belongs to post-stratum j. Based on the fitted model, the post-strata means can be obtained $\hat{\boldsymbol{\theta}} = \boldsymbol{X}^{\text{pop}}\hat{\boldsymbol{\beta}}$, and the multilevel regression and post-stratification estimator can be generally written

$$\widehat{Q}_{\text{MRP}} = \left(\frac{N_1}{N}, \ldots, \frac{N_J}{N}\right) \boldsymbol{X}^{\text{pop}}\hat{\boldsymbol{\beta}}.$$

Subgroup estimators are readily available by restricting the weighted average to the corresponding subgroup. In the OHARNG example, the finite subpopulation mean among female service members is estimated as

$$\widehat{Q}^{\text{Female}}_{\text{MRP}} = \left(0, 0, 0, 0, \frac{N_5}{\sum_{j=5}^{8} N_j}, \frac{N_6}{\sum_{j=5}^{8} N_j}, \frac{N_7}{\sum_{j=5}^{8} N_j}, \frac{N_8}{\sum_{j=5}^{8} N_j}\right) \boldsymbol{X}^{\text{pop}}\hat{\boldsymbol{\beta}}.$$

4.2.3 Calculating Margin of Error (MOE)

Conditional on the variance parameters $\Sigma_{\boldsymbol{\beta}}^{-1}$, σ_y, the posterior distribution of regression coefficients $\boldsymbol{\beta}$ is multivariate normal with the following closed form

$$\begin{aligned}\pi(\boldsymbol{\beta}|\boldsymbol{y}, \Sigma_{\boldsymbol{\beta}}^{-1}, \sigma_y) &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\sigma_y^2\boldsymbol{I}_{n\times n})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}'\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\right\} \\ &\sim N\left(\left(\frac{1}{\sigma_y^2}\boldsymbol{X}'\boldsymbol{X} + \Sigma_{\boldsymbol{\beta}}^{-1}\right)^{-1}\frac{1}{\sigma_y^2}\boldsymbol{X}'\boldsymbol{y}, \left(\frac{1}{\sigma_y^2}\boldsymbol{X}'\boldsymbol{X} + \Sigma_{\boldsymbol{\beta}}^{-1}\right)^{-1}\right),\end{aligned}$$

with $\text{Var}(\boldsymbol{\beta}|\boldsymbol{y}, \Sigma_{\boldsymbol{\beta}}^{-1}, \sigma_y) = \left(\frac{1}{\sigma_y^2}\boldsymbol{X}'\boldsymbol{X} + \Sigma_{\boldsymbol{\beta}}^{-1}\right)^{-1}$. Note that the design matrix $\boldsymbol{X}$ is a $n \times k$ matrix with n_j rows (out of n) taking values $\boldsymbol{x}_j$ and knowing the cell counts n_j in the J post-strata is sufficient to compute $\boldsymbol{X}'\boldsymbol{X}$.

In the survey design stage, margin of errors can be calculated based on a specific multilevel model, with information on the design parameters $\sigma_y, \Sigma_{\boldsymbol{\beta}}^{-1}, \{n_j\}_{j=1}^J$, using the following steps:

Step 1 Calculate posterior variances of the J post-strata means:

$$\text{Var}(\boldsymbol{\theta}|\boldsymbol{y}, \Sigma_{\boldsymbol{\beta}}^{-1}, \sigma_y) = \boldsymbol{X}^{\text{pop}}\text{Var}(\boldsymbol{\beta}|\boldsymbol{y}, \Sigma_{\boldsymbol{\beta}}^{-1}, \sigma_y)\boldsymbol{X}^{\text{pop}'}.$$

Step 2 Calculate posterior variance of the population mean:

$\text{Var}(Q|\boldsymbol{y}, \Sigma_{\boldsymbol{\beta}}^{-1}, \sigma_y) = \boldsymbol{C}\text{Var}(\boldsymbol{\theta}|\boldsymbol{y}, \Sigma_{\boldsymbol{\beta}}^{-1}, \sigma_y)\boldsymbol{C}'$, where $\boldsymbol{C} = \left(\frac{N_1}{N}, \ldots, \frac{N_J}{N}\right)$. As is demonstrated in Section 4.2.2, other estimands such as subgroup means can be calculated by modifying $\boldsymbol{C}$ accordingly.

Step 3 Obtain corresponding margin of error: $\text{MOE} = 2\sqrt{\text{Var}(Q|\boldsymbol{y}, \Sigma_{\boldsymbol{\beta}}^{-1}, \sigma_y)}$

4.3 Validation Using Simulation Studies

Simulation studies were conducted as a validation procedure, comparing the margins of errors (MOE) calculated using the procedure in Section 4.2.3 with empirical results.

4.3.1 Simulation Design

The simulation studies are designed with the OHARNG population introduced in Section 4.2.2. We consider two discrete auxiliary variables, gender (male, female) and age group with 4 levels, resulting in $J = 2 \times 4 = 8$ post-strata, with definition and distribution listed in Table 4.1. We generate survey outcomes Y based on the following multilevel model

$$Y = \mu_0 + \mu_1 I(\text{female}) + \sum_{l=1}^{4} \alpha_l I(\text{age group } l) + \sum_{j=1}^{8} \gamma_j I(\text{post-stratum } j) + \epsilon, \tag{4.1}$$

where $\alpha_l \overset{\text{i.i.d}}{\sim} N(0, \sigma_\alpha^2)$, $\gamma_j \overset{\text{i.i.d}}{\sim} N(0, \sigma_\gamma^2)$, $\epsilon \sim N(0, \sigma_y^2)$. In the simulation, we use $\mu_0 = 40$, $\mu_1 = 5$, $\sigma_\alpha = 10$, $\sigma_\gamma = 2.5$, $\sigma_y = 10$. To obtain empirical margin of errors (MOE), we repeatedly draw samples from the population and compute 95% credible intervals for various estimands of interests. Samples of various sample sizes were considered, with various sample allocation plans, (a) equal sample sizes for all post-strata, (b) sample sizes proportionate to post-strata sizes, (c) equal sample sizes for post-strata within gender, and (d) sample sizes proportionate to post-strata sizes within gender. The total sample sizes and sample sizes by post-stratification cells are summarized in Table 4.2.

For each scenario, 500 samples were drawn and 95% credible intervals were computed for post-strata means, the overall population mean and subpopulation means by gender ($8 + 1 + 2 = 11$ estimands). For each estimand, empirical MOE is obtained as half of

Table 4.2: Total sample sizes and sample sizes by post-stratification cells for all simulation scenarios

sampling plan	n	$\{n_j\}_{j=1}^{8}$
(a)	200	$\{25, 25, 25, 25, 25, 25, 25, 25\}$
	300	$\{38, 38, 38, 38, 38, 38, 38, 38\}$
	400	$\{50, 50, 50, 50, 50, 50, 50, 50\}$
(b)	205	$\{61, 62, 37, 15, 15, 9, 4, 2\}$
	303	$\{91, 92, 55, 22, 22, 13, 6, 2\}$
	405	$\{122, 123, 73, 30, 29, 17, 8, 3\}$
	804	$\{243, 245, 145, 59, 58, 34, 15, 5\}$
(c)	300	$\{50, 50, 50, 50, 25, 25, 25, 25\}$
	400	$\{75, 75, 75, 75, 25, 25, 25, 25\}$
	500	$\{100, 100, 100, 100, 25, 25, 25, 25\}$
	800	$\{175, 175, 175, 175, 25, 25, 25, 25\}$
	200	$\{10, 10, 10, 10, 40, 40, 40, 40\}$
(d)	205	$\{8, 8, 5, 2, 94, 56, 24, 8\}$
	205	$\{36, 36, 21, 9, 53, 31, 14, 5\}$
	305	$\{36, 36, 21, 9, 105, 62, 27, 9\}$

the average lengths of the 500 credible intervals. The credible intervals were generated using MRP with the model defined in Equation 4.1. The models were fitted using the `brms` package in `R 3.6.0`, the default option of which specifies non-central t-distributions for $\pi(\boldsymbol{\mu})$ and half t-distributions for $\pi(\sigma_\alpha)$, $\pi(\sigma_\gamma)$ and $\pi(\sigma_y)$, with hyperparameters adaptive to data.

Correspondingly, theoretical MOEs were calculated for all 11 estimands using the procedure in Section 4.2.3. Naive MOEs ignoring partial pooling were also calculated for the 8 post-strata means using $2\frac{\sigma_y}{\sqrt{n_j}}$, for comparison purpose.

4.3.2 Simulation Results

The theoretical margin of errors (MOE) were compared with the empirical MOEs from the simulation studies. Figure 4.1(a) displays the theoretical MOEs against the empirical MOEs with a scatterplot. The MOEs calculated with the naive approach are plotted in circles while the MOEs accounting for partial pooling are plotted in crosses. For post-strata with small cell counts n_j in the sample (reflected via large values of $2\frac{\sigma_y}{\sqrt{n_j}}$), the naive approach leads to overestimation while the approach accounting for partial pooling yields results similar to empirical MOEs. The theoretical MOEs using approach accounting for partial pooling also align with empirical results for overall and subpopulation means, as is shown in Figure 4.1(b).

4.4 Application Scenarios

We consider survey design with MRP for two different application scenarios, online panels that utilize quota sampling and telephone surveys with a fixed budget. For both scenarios,

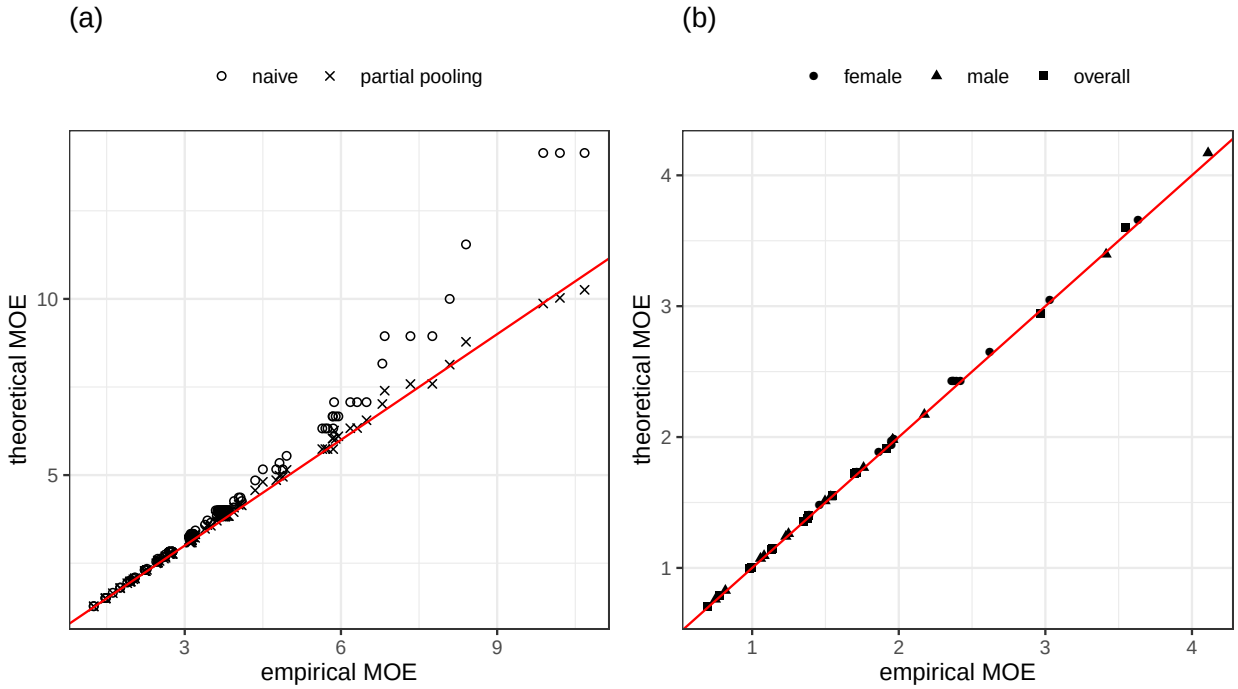

Figure 4.1: Simulation results for all scenarios - scatter plots of (a) theoretical MOEs using naive method (in circles) and approach accounting for partial pooling (in crosses) vs empirical MOEs from simulations for all post-strata (b) theoretical MOEs using approach accounting for partial pooling vs empirical MOEs from simulations by estimands

we use the illustration example of OHARNG population in Section 4.2.2 and demonstrate how to calculate MOE if we plan to use the multilevel regression model defined by Equation 4.1 for analysis. We assume that information on the variance parameters is available at the design stage, with $\sigma_\alpha = 10$, $\sigma_\gamma = 2.5$.

4.4.1 Online Panels Using Quota Sampling

Survey data collection using online panels is increasingly popular due to its cost-effectiveness. online panels enables access to large and diverse samples in a short time period, take less time to get the data ready for analysis and are easy to replicate with the standardized data collection process. In spite of the advantages, online panels tend to have low response rate. As a result, many users of online panels utilize a quota sampling approach by targeting respondents with certain demographic and other characteristics to mitigate non-coverage and non-response. Similar to stratified sampling, quota sampling first partitions the population into mutually exclusive subgroups and enables including a pre-specified number of units in each subgroup.

Suppose we need to design a survey using online panels for the OHARNG population in Section 4.2.2. Quota sampling enables including a pre-specified number of solders in all the cells defined by the combinations of gender and age categories. We consider four different sample allocation plans, (a) equal sample sizes for all post-strata, (b) sample sizes proportionate to post-strata sizes, (c) fixed total sample size for females at 200 and equal sample sizes for post-strata within gender, and (d) fixed total sample size for females at 200 and sample sizes proportionate to post-strata sizes within gender. We calculate theoretical MOEs for overall mean and mean among females. The MOEs for both estimands are plotted

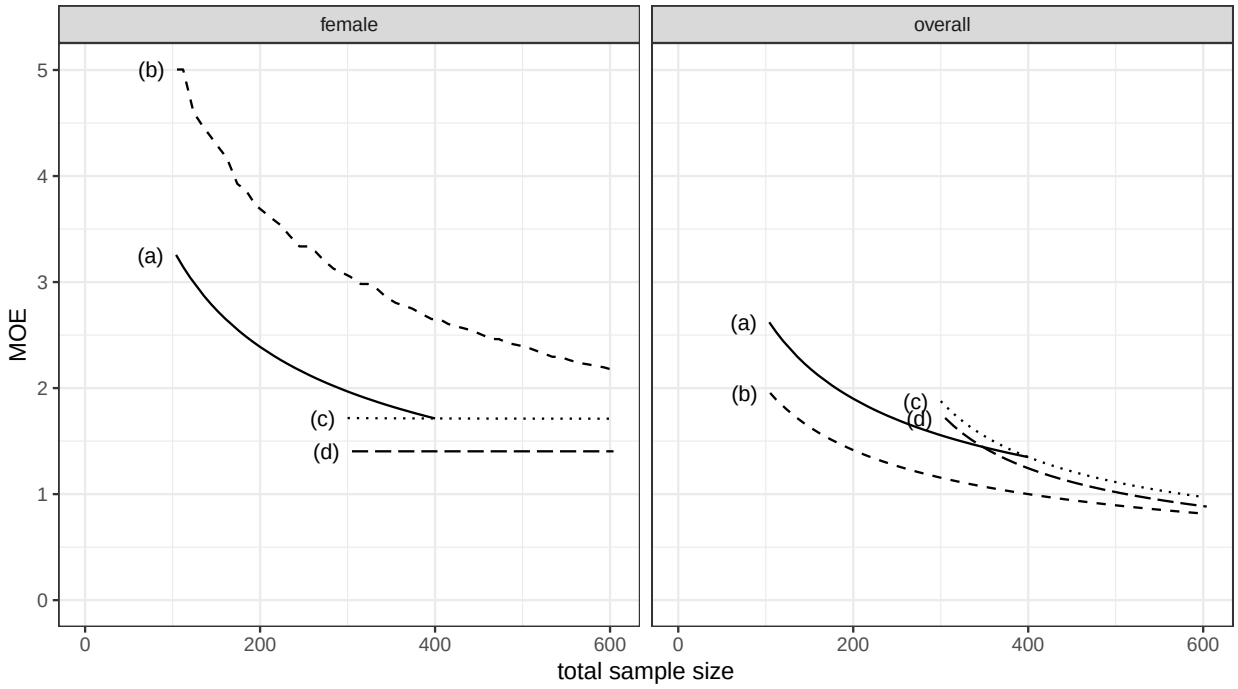

Figure 4.2: Quota sampling - theoretical MOEs accounting for partial pooling as total sample size increases for different sample allocation plan: (a) equal sample sizes for all post-strata, (b) sample sizes proportionate to post-strata sizes, (c) fixed total sample size for females at 200 and equal sample sizes for post-strata within gender, and (d) fixed total sample size for females at 200 and sample sizes proportionate to post-strata sizes within gender

in Figure 4.2.

Sample allocation plan (a) with sample sizes proportional to post-strata sizes yields lowest MOE for overall population mean but highest MOE for subgroup mean for females. Total sample size for plan (b) with equal sample sizes is bounded by the smallest post-strata size, as there are only 63 people in post-stratum $j = 8$ and it might not be even feasible to exhaust the post-stratum in practice. Plans (c)(d) fix total sample size for females at 200 and allow sample size increase while maintaining the allocation ratios within both genders. MOEs for

overall mean decreases as total sample size increases for all allocation plans while MOEs for subgroup mean among females remains similar for plans (c)(d), as total sample sizes for females are fixed and effect of partial pooling is not obvious when the subgroup sample size is fixed at 200. In practice, the survey could be designed with a upper bound of MOE for the estimand within females which determines the fixed total sample sizes among females $\sum_{j=5}^{8} n_j$ before the total sample size n is set with a target MOE for the overall estimand.

4.4.2 Telephone Surveys with Fixed Total Sample Sizes

Telephone surveys are typically designed with a certain budget limit which often results in a fixed total sample size. Different from online panels using quota sampling, telephone surveys typically do not enables including a pre-specified number of respondents from subgroups defined by demographics and certain characteristics. Therefore, assumptions on inclusion probabilities in the subgroups are needed in order to calculate expected cell counts in the design stage.

Suppose we need to design a telephone survey for the OHARNG population in Section 4.2.2. The service members are selected using equal probability of selection method (EPSEM). With assumptions on response rate in the post-stratification cells p_j, the sample sizes in each cell n_j follow a multinomial distribution with cell probabilities $\pi_j = N_j p_j / \sum_l N_l p_l$, with $\mathrm{E}(n_j) = n\pi_j$, for $j = 1, \ldots, J$. As a numerical example, we assume $\boldsymbol{p} = (.3, .4, .5, .6, .55, .6, .65, .7)$, resulting in $\boldsymbol{\pi} = (0.213, 0.286, 0.211, 0.102, 0.092, 0.059, 0.027, 0.010)$. The theoretical MOEs can be calculated for various total sample sizes by plugging in $n\pi_j$ as the cell counts n_j. Figure 4.3 display the theoretical MOE for subgroup mean among females and overall mean for various total sample sizes. Due to the small cell probabilities

π_j for female cells ($j = 5, \ldots, 8$), a small increase in total sample size does not necessarily results in increase in total sample size of females, leading to slightly non-smoothness in the curve.

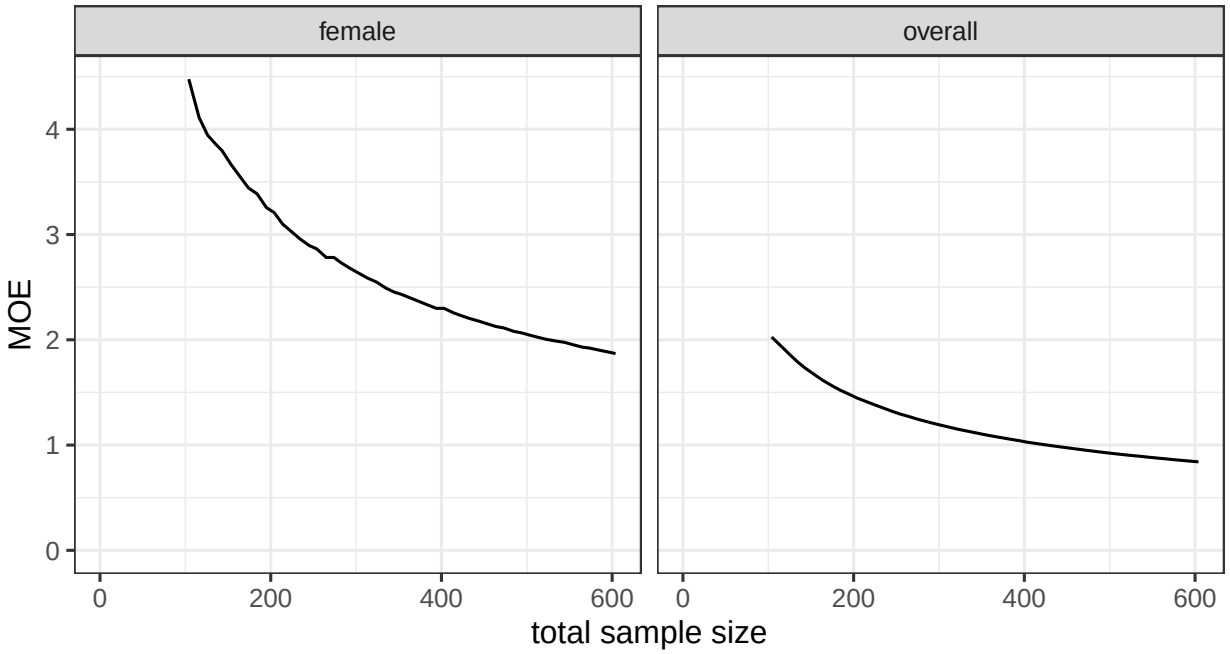

Figure 4.3: Telephone survey - theoretical MOEs accounting for partial pooling, calculated with expected cell counts as total sample increases, for subgroup mean among females and overall mean

In practice, the actual sample sizes could differ from the expected cell counts. For the cells with small π_j, the actual cell count could be very small, regardless of the total sample size. Therefore, we investigate the robustness of the results when small cell counts occurs by setting the smallest cell $n_8 = 3$ for all total sample sizes considered. Also, the information on the variance parameters used in the design stage could differ from the actual values in the population. As sensitivity analysis, we explore the following scenarios where σ_α and σ_γ

are overestimated or underestimated in the design stage.

Table 4.3: Value sets of the variance parameters for sensitivity analysis

scenario	variance parameters	note
(a)	$\sigma_\alpha = 10$, $\sigma_\gamma = 2.5$	assumed as true values
(b)	$\sigma_\alpha = 15$, $\sigma_\gamma = 5$	overestimate both
(c)	$\sigma_\alpha = 15$, $\sigma_\gamma = 2.5$	overestimate σ_α
(d)	$\sigma_\alpha = 10$, $\sigma_\gamma = 5$	overestimate σ_γ
(e)	$\sigma_\alpha = 5$, $\sigma_\gamma = 1$	underestimate both
(f)	$\sigma_\alpha = 5$, $\sigma_\gamma = 2.5$	underestimate σ_α
(g)	$\sigma_\alpha = 10$, $\sigma_\gamma = 1$	underestimate σ_γ

Figure 4.4 display the MOE for the post-stratum mean with the smallest cell probability $\pi_8 = .010$, calculated by setting the cell count $n_8 = 3$ as total sample size increases, using various parameter sets, faceted by σ_α. Overestimation of variance parameters leads to underestimation of effect of partial pooling and overestimation of MOE for the cell mean. The value of MOE is sensitive to the specification of σ_γ but less sensitive to σ_α. MOEs for the subgroup mean of female and overall mean are not sensitive to n_8 or the specification of the variance parameters and, therefore, results are not shown here. In practice, value sets (b)(e) could be used to provide upper and lower bound for the MOEs.

4.5 Discussion

We consider survey design for multilevel regression and post-stratification (MRP). Existing literature mainly focus on analysis of survey data using MRP. This is the among the first papers from a design perspective.

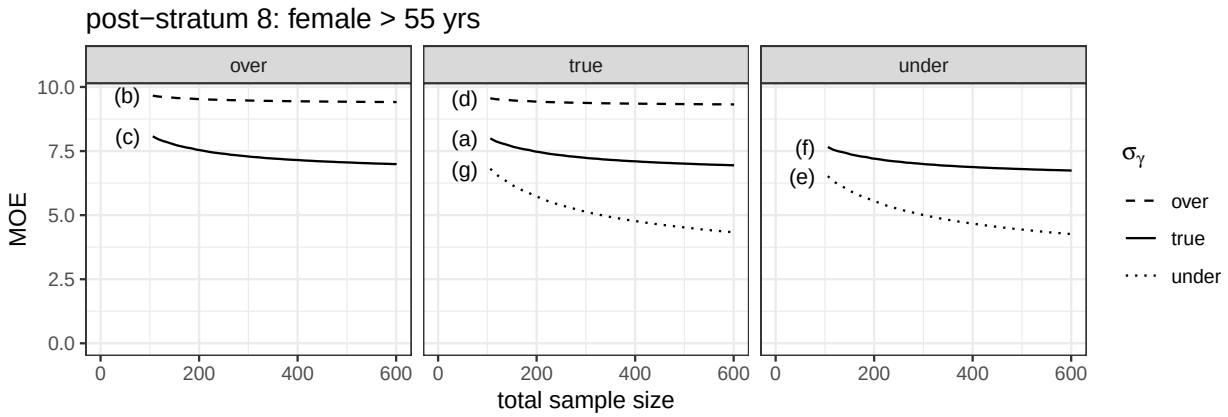

Figure 4.4: Sensitivity analysis setting $n_8 = 3$ - theoretical MOE for post-stratum 8 as total sample size increases, using various value sets of variance parameters listed in Table 4.3, faceted by overestimated, true, underestimated σ_α, with various types of line for overestimated, true, underestimated σ_γ

We propose a closed form formula to calculate theoretical margin of errors for various estimands based on the variance parameters in the multilevel regression model and sample sizes in the post-strata. We validate the theoretical MOEs via comparisons with the empirical MOEs in simulations studies covering various sample allocation plans. The validation procedure indicates that the theoretical MOEs based on the formula aligns with the empirical results for various estimands.

We demonstrate the application of the formula in two different survey design scenarios, online panels that utilize quota sampling and telephone surveys with a fixed budget.

The method could be potentially extended to binary outcome Y via introducing a normally distributed latent variable Y^* such that $Y = \mathbf{1}\{Y > 0\}$.

Chapter 5

Conclusion

This dissertation addresses three problems in Bayesian design and analysis for sampling. Flexible Bayesian models are utilized to incorporate auxiliary information about the target population of interest, resulting in robust and more efficient inference.

In Chapter 2, we consider quantile estimation for skewed survey data in PPS sampling, where the values of a size variable is available for all population units. While the design-based weighted method only utilizes information in the survey sample, the proposed model-based methods incorporate probabilities of selection for all population units. Combined with transformation, the skew-normal distribution, with location and scale parameters modeled with penalized spline functions, is flexible enough to handle skewed data from various distributions and robust against model misspecification. By incorporating more information in the analysis, the model-based methods are more efficient than the design-based weighted method. The Bayesian model-based methods also demonstrate advantages in small sample scenarios, as the design-based methods typically are based on asymptotic properties.

In Chapter 3, we consider inference from non-random samples in data-rich settings where

high-dimensional auxiliary information is available both in the sample and the target population, with survey inference being a special case. The proposed regularized prediction approach using Bayesian machine learning predicts the outcomes in the population using a large number of auxiliary variables such that the ignorability assumption is reasonable. The machine learning models naturally accommodate discrete variables, nonlinear effect of continuous variables and possible interactions. In terms of model fitting, regularization handles noise variables and achieves stable predictions while the Bayesian approach is straightforward for quantification of uncertainty. The method using soft Bayesian additive regression trees outperforms existing methods even in low dimensional settings, as it better utilize information in the continuous auxiliary variables.

In Chapter 4, we consider survey design for multilevel regression and post-stratification (MRP). When there is discrepancy between a survey sample and the target population of interest, post-stratification corrects for the known differences by partitioning the population into a series of disjoint subpopulations (post-strata) such that the subsamples in the corresponding post-strata are representative. However, post-stratification could yield unstable estimates in the presence of sparse cells in some post-strata. MRP specifies a multilevel regression model to partially pool information across post-strata and achieves stable estimates. MRP can also be viewed from a regularized prediction perspective. Regularization is imposed via multilevel varying intercepts to achieve stable predictions for the outcomes in the post-strata. If a hierarchical Bayesian approach is used, with hyper priors specified, the amount the regularization is determined by the data in the model fitting procedure, without parameter tuning. We propose a closed form formula to calculate margin of er-

rors (MOEs) accounting for partial pooling and regularization when naive methods ignoring partial pooling and regularization overestimates uncertainty of estimates in small cells.

Classical statistical methods are mainly design-based, where quantification of uncertainty relies on the distribution of sample inclusion indicator and asymptotic properties. The survey outcomes are treated as fixed, therefore, limiting the potential to borrow strength from the recent development of modern statistical models. Model-based survey inference specifies statistical distributions on the survey outcomes, which handles both probability and non-probability samples and naturally allows improving survey inference with modern statistical techniques via incorporating auxiliary information in data-rich settings. Model-based methods are subject to violation of model assumptions and model misspecification. However, modern statistical models are flexible enough to yield robust inference. In Chapter 2, when the size variable or the selection probability is the only auxiliary variable available for conditional modeling, the normality assumption could be violated. However, the skew-normal penalized spline regression models are flexible enough to adequately model various skewed distributions. In Chapter 3, the regularized prediction approach accommodate high-dimensional auxiliary variables such that the ignorability assumption is reasonable and the sum-of-tree ensembles allow flexible functional forms of the predictors. In Chapter 4, multilevel models are able to capture intrinsic structures of multilevel data. Considering sample designs from a model-based perspective is important. In terms of statistical inference and model fitting, the Bayesian approach allows fitting complex statistical models and is straightforward for generating probability intervals. And the hierarchical Bayesian approach avoids parameter tuning in imposing regularization, seen in Chapter 2 and Chapter 4. In terms

of implementation, recent advancement in the probabilistic programming languages such as `Stan` and related user-friendly software has enables practitioners to easily fit a variety of Bayesian models, making Bayesian modeling accessible to wider user community. In conclusion, flexible Bayesian models are powerful tools in design and analysis for sampling, straightforward for quantification of uncertainty and yielding efficient inference.

Bibliography

Azzalini, A. (2013). *The skew-normal and related families*, volume 3. Cambridge University Press.

Breidt, F., Claeskens, G., and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika* **92,** 831–846.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics* pages 1026–1053.

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science* **32,** 190–205.

Breidt, F. J., Opsomer, J. D., Johnson, A. A., and Ranalli, M. G. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology* **33,** 35.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* **76,**.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97,** 465–480.

BIBLIOGRAPHY

Chambers, R. L., Dorfman, A. H., and Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association* **88,** 268–277.

Chambers, R. L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika* **73,** 597–604.

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

Chen, Q., Elliott, M. R., and Little, R. J. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey methodology* **36,** 23.

Chen, Q., Elliott, M. R., and Little, R. J. A. (2012). Bayesian inference for finite population quantiles from unequal probability samples. *Survey Methodology* **38,** 203–214.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics* **4,** 266–298.

Cochran, W. G. (2007). *Sampling techniques*. John Wiley & Sons.

D'Aunno, T., Friedmann, P. D., Chen, Q., and Wilson, D. M. (2015). Integration of substance abuse treatment organizations into accountable care organizations: Results from a national survey. *Journal of health politics, policy and law* **40,** 797–819.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled

frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics* **11,** 427–444.

Dorfman, A. H. (2009). Inference on distribution functions and quantiles. In *Handbook of Statistics*, volume 29, pages 371–395. Elsevier.

Dorfman, A. H. and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics* pages 1452–1475.

Downes, M., Gurrin, L. C., English, D. R., Pirkis, J., Currier, D., Spittal, M. J., et al. (2018). Multilevel regression and poststratification: A modeling approach to estimating population quantities from highly selected survey samples. *American journal of epidemiology* **187,** 1780–1790.

Ericson, W. A. (1969). Subjective bayesian models in sampling finite populations. *Journal of the Royal Statistical Society. Series B (Methodological)* **31,** 195–233.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis* **1,** 515–534.

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*. CRC press.

Gelman, A. and Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* **23,** 127–135.

BIBLIOGRAPHY

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association* **88,** 881–889.

Hájek, J. (1971). Comment on a paper by d. basu. *Foundations of Statistical Inference* **236,**.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20,** 217–240.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research* **15,** 1593–1623.

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

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition* **2,** 813–830.

Kuk, A. Y. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika* **75,** 97–103.

Kuk, A. Y. (1993). A kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika* **80,** 385–392.

Kuk, A. Y. and Welsh, A. (2001). Robust estimation for finite populations based on a working model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63,** 277–292.

Linero, A. R. and Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80,** 1087–1110.

Little, R. and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica* **14,** 949–968.

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

Little, R. J. (2004). To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association* **99,** 546–556.

Lohr, S. (2009). *Sampling: design and analysis*. Nelson Education.

Lombardıa, M., González-Manteiga, W., and Prada-Sánchez, J. (2003). Bootstrapping the chambers–dunstan estimate of a finite population distribution function. *Journal of Statistical Planning and Inference* **116,** 367–388.

Lombardıa, M., González-Manteiga, W., and Prada-Sánchez, J. (2004). Bootstrapping the dorfman–hall–chambers–dunstan estimator of a finite population distribution function. *Nonparametric Statistics* **16,** 63–90.

Lumley, T. (2016). survey: analysis of complex survey samples. R package version 3.32.

McConville, K. and Breidt, F. (2013). Survey design asymptotics for the model-assisted penalised spline regression estimator. *Journal of Nonparametric Statistics* **25,** 745–763.

Mcconville, K. S., Jay Breidt, F., Lee, T. C., and Moisen, G. G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology* **5,** 131–158.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83,** 1023–1032.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association* **100,** 1429–1442.

Park, D. K., Gelman, A., and Bafumi, J. (2004). Bayesian multilevel estimation with post-stratification: State-level estimates from national polls. *Political Analysis* pages 375–385.

Park, D. K., Gelman, A., and Bafumi, J. (2006). State-level opinions from national surveys: Poststratification using multilevel logistic regression. *Public opinion in state politics* pages 209–28.

Rao, J., Kovar, J., and Mantel, H. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* pages 365–375.

Rao, J., Wu, C., and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology* **18,** 209–217.

Rao, J. N. and Wu, C. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association* **83,** 231–241.

Rothwell, P. M. (2005). External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet* **365,** 82–93.

Royall, R. M. (1971). Linear regression models in finite population sampling theory. *Foundations of statistical inference* pages 259–279.

Royall, R. M. (1992). The model based (prediction) approach to finite population sampling theory. *Lecture Notes-Monograph Series* **17,** 225–240.

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

Rubin, D. B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys: Comment. *Journal of the American Statistical Association* **78,** 803–805.

Särdnal, C., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling.* Springer-Verlag.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **5,** 127.

Si, Y., Pillai, N. S., and Gelman, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis* **10,** 605–625.

Smith, T. (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society. Series A (General)* pages 183–204.

Smith, T. (1994). Sample surveys 1975-1990; an age of reconciliation? *International Statistical Review/Revue Internationale de Statistique* **62,** 5–19.

Tan, Y. V., Flannagan, C. A., and Elliott, M. R. (2019). "robust-squared" imputation models using bart. *Journal of Survey Statistics and Methodology* **7,** 465–497.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 267–288.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association* **88,** 89–96.

Vehtari, A., Gabry, J., Yao, Y., and Gelman, A. (2018). loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.0.0.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing* **27,** 1413–1432.

Wang, S. and Dorfman, A. H. (1996). A new estimator for the finite population distribution function. *Biometrika* **83,** 639–652.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting* **31,** 980–991.

Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association* **47,** 635–646.

Yuan, Y. and Little, R. J. (2007). Model-based estimates of the finite population mean for two-stage cluster samples with unit non-response. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **56,** 79–97.

Zangeneh, S. Z. and Little, R. J. (2015). Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. *Journal of Survey Statistics and Methodology* **3,** 162–192.

Zheng, H. and Little, R. J. (2003). Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *Journal of Official Statistics* **19,** 99.

Appendix A

Appendices to Chapter 2

A.1 Proof of Proposition

Proposition 1. *If* $W \sim N(0,1)\mathbf{1}\{w > 0\}$ *and* $Y|W = w \sim N(\xi + \alpha\sigma w, \sigma^2)$*, then* $Y \sim SkewNorm(\xi, \omega^2, \alpha)$ *with probability density function*

$$f(y|\xi, \omega^2, \alpha) = \frac{2}{\omega}\phi\left(\frac{y-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{y-\xi}{\omega}\right)\right),$$

where $\omega^2 = (\alpha^2+1)\sigma^2$*. In other words, the skew-normal distribution* $SkewNorm(\xi, \omega^2, \alpha)$ *can be simulated using the algorithm* $Y = \alpha\sigma|Z_1| + (\xi + \sigma Z_2)$*, with* $Z_1, Z_2 \overset{i.i.d.}{\sim} N(0,1)$ *and* $\omega^2 = (\alpha^2+1)\sigma^2$*.*

Proof. Let $W \sim N(0,1)\mathbf{1}\{w>0\}$ and $Y|W=w \sim N(\xi+\alpha\sigma w, \sigma^2)$.

$$
\begin{aligned}
f(y|\xi,\sigma^2,\alpha) =& \int_{\mathbb{R}} f(y,w|\xi,\sigma^2,\alpha)\,dw = \int_{\mathbb{R}} f(y|w,\xi,\sigma^2,\alpha)f(w)\,dw \\
=& \int_{\mathbb{R}} \frac{1}{\sigma}\phi\left(\frac{y-(\xi+\alpha\sigma w)}{\sigma}\right) \times 2\phi(w)\mathbf{1}\{w>0\}\,dw \\
=& \int_0^{+\infty} \frac{1}{\sigma}(2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{[(y-\xi)-\alpha\sigma w]^2}{2\sigma^2}\right\} 2(2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{w^2}{2}\right\} dw \\
=& \frac{2}{\sigma}(2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{(y-\xi)^2}{2\sigma^2}\right\} \\
& \int_0^{+\infty} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{(\alpha^2+1)\sigma^2w^2-2\alpha\sigma w(y-\xi)}{2\sigma^2}\right\} dw \\
=& \frac{2}{\sigma}(2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{(y-\xi)^2}{2(\alpha^2+1)\sigma^2}\right\} \int_0^{+\infty} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{[w-\frac{\alpha}{\alpha^2+1}\frac{(y-\xi)}{\sigma}]^2}{2(\alpha^2+1)^{-1}}\right\} dw \\
=& \frac{2}{\omega}\phi\left(\frac{y-\xi}{\omega}\right) Pr(W>0), \text{ where } W \sim N\left(\frac{\alpha}{\alpha^2+1}\frac{(y-\xi)}{\sigma}, \frac{1}{\alpha^2+1}\right) \\
=& \frac{2}{\omega}\phi\left(\frac{y-\xi}{\omega}\right) Pr(Z>-\frac{\alpha(y-\xi)}{\omega}), \text{ where } Z \sim N(0,1) \\
=& \frac{2}{\omega}\phi\left(\frac{y-\xi}{\omega}\right) \Phi\left(\alpha\left(\frac{y-\xi}{\omega}\right)\right)
\end{aligned}
$$

□

A.2 Posterior Simulation Scheme

A.2.1 SN-BPSP

Here we present detailed posterior simulation steps for the skew-normal Bayesian penalized spline predictive approach (SN-BPSP) defined in (2.2). Using the hierarchical representation

in Proposition 1, model (2.2) can be rewritten as

$$Y_i|\pi_i, \boldsymbol{\beta}, \boldsymbol{b}, \sigma^2, \alpha, \gamma, W_i \overset{\text{ind.}}{\sim} N(SPL(\pi_i, \boldsymbol{m}) + \alpha\sigma\pi_i^{\gamma} W_i, \sigma^2\pi_i^{2\gamma}),$$

$$W_i \overset{\text{i.i.d.}}{\sim} N(0,1)\mathbf{1}\{w_i > 0\},$$

$$SPL(\pi_i, \boldsymbol{m}) = \beta_0 + \sum_{l=1}^{p} \beta_l \pi_i^l + \sum_{k=1}^{K} b_k (\pi_i - m_k)_+^p,$$

$$\boldsymbol{b} = (b_1, \ldots, b_K)^T | \tau_b^2 \sim N(\mathbf{0}, \tau_b^2 \boldsymbol{I}_K),$$

Assume the following prior distributions for derivation purpose

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \varphi^2 \boldsymbol{I}_{p+1} = (10^3)^2 \boldsymbol{I}_{p+1})$$

$$\tau_b, \sigma \sim U(0, +\infty), \gamma \sim U(-2, 2)$$

$$\alpha \sim N(0, \psi^2 = 10^2)\mathbf{1}\{\alpha > 0\}.$$

Denote by

$$\boldsymbol{\Pi} = \begin{bmatrix} \pi_1^{2\gamma} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \pi_n^{2\gamma} \end{bmatrix}_{n\times n} \quad \boldsymbol{X} = \begin{bmatrix} 1 & \pi_1 & \ldots & \pi_1^p \\ \vdots & \ddots & & \vdots \\ 1 & \pi_n & \ldots & \pi_n^p \end{bmatrix}_{n\times(p+1)}$$

$$\boldsymbol{Z} = \begin{bmatrix} (\pi_1 - m_1)_+^p & \cdots & (\pi_1 - m_K)_+^p \\ \vdots & \ddots & \vdots \\ (\pi_n - m_1)_+^p & \cdots & (\pi_n - m_K)_+^p \end{bmatrix}_{n\times K}$$

and define $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b}$, the full conditionals of the posteriors for (2.2) are detailed as belows:

1. $[\boldsymbol{\beta}] \sim N_{p+1}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{X}^T (\sigma^2\boldsymbol{\Pi})^{-1} (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{b} - \alpha\sigma\boldsymbol{\Pi}^{\frac{1}{2}}\boldsymbol{w}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left[\boldsymbol{X}^T (\sigma^2\boldsymbol{\Pi})^{-1}\boldsymbol{X} + (\varphi^2 \boldsymbol{I}_{p+1})^{-1}\right]^{-1}$
2. $[\boldsymbol{b}] \sim N_K(\boldsymbol{\Sigma}_{\boldsymbol{b}} \boldsymbol{Z}^T (\sigma^2\boldsymbol{\Pi})^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \alpha\sigma\boldsymbol{\Pi}^{\frac{1}{2}}\boldsymbol{w}), \boldsymbol{\Sigma}_{\boldsymbol{b}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{b}} = \left[\boldsymbol{Z}^T (\sigma^2\boldsymbol{\Pi})^{-1}\boldsymbol{Z} + (\tau_b^2 \boldsymbol{I}_K)^{-1}\right]^{-1}$

3. $[\boldsymbol{w}] \sim N_n(\boldsymbol{\Sigma}_{\boldsymbol{w}}\alpha(\sigma^2\boldsymbol{\Pi})^{-\frac{1}{2}}\boldsymbol{e}, \boldsymbol{\Sigma}_{\boldsymbol{w}}) \prod_{i=1}^{n} \mathbf{1}\{w_i > 0\}$, where $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \left[\alpha^2\boldsymbol{I}_n + \boldsymbol{I}_n^{-1}\right]^{-1}$

4. $[\alpha] \sim N(\sigma_\alpha^2\boldsymbol{w}^T(\sigma^2\boldsymbol{\Pi})^{-\frac{1}{2}}\boldsymbol{e}, \sigma_\alpha^2)\mathbf{1}\{\alpha > 0\}$, where $\sigma_\alpha^2 = \left[\boldsymbol{w}^T\boldsymbol{w} + (\psi^2)^{-1}\right]^{-1}$

5. $[\tau_b^2] \sim IG(\frac{K-1}{2}, \frac{1}{2}\boldsymbol{b}^T\boldsymbol{I}_K^{-1}\boldsymbol{b})$

6. $[\gamma] \propto (\prod_i \pi_i)^{-\gamma} \exp\left\{-\frac{1}{2}\left[\boldsymbol{e}^T(\sigma^2\boldsymbol{\Pi})^{-1}\boldsymbol{e} - 2\alpha\boldsymbol{w}^T(\sigma^2\boldsymbol{\Pi})^{-\frac{1}{2}}\boldsymbol{e}\right]\right\} I_{(-2,+2)}(\gamma)$

7. $[\sigma] \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\left[\boldsymbol{e}^T(\sigma^2\boldsymbol{\Pi})^{-1}\boldsymbol{e} - 2\alpha\boldsymbol{w}^T(\sigma^2\boldsymbol{\Pi})^{-\frac{1}{2}}\boldsymbol{e}\right]\right\} I_{(0,+\infty)}(\sigma)$

All the above full conditionals have an explicit form except for 6 and 7. For 6 and 7, the Metropolis-Hastings algorithm is used with a normal proposal distribution centered at the current value and a small variance.

A.2.2 SN-B2PSP

Here we present detailed posterior simulation steps for the skew-normal Bayesian two-moment penalized spline predictive approach (SN-B2PSP) defined in (2.3). Using the hierarchical representation in Proposition 1, model (2.3) can be rewritten as

$$Y_i|\pi_i, \boldsymbol{\beta}, \boldsymbol{b}, \sigma_i^2, \alpha, W_i \overset{\text{ind.}}{\sim} N(SPL_1(\pi_i, \boldsymbol{m}) + \alpha\sigma_i W_i, \sigma_i^2),$$

$$W_i \overset{\text{i.i.d.}}{\sim} N(0,1)\mathbf{1}\{w_i > 0\},$$

$$\sigma_i^2|\pi_i, \boldsymbol{\lambda}, \boldsymbol{\nu}, \sigma_A^2 \overset{\text{ind.}}{\sim} \text{LogNorm}(SPL_2(\pi_i, \boldsymbol{m}), \sigma_A^2),$$

$$SPL_1(\pi_i, \boldsymbol{m}) = \beta_0 + \sum_{l=1}^{p}\beta_l\pi_i^l + \sum_{k=1}^{K} b_k(\pi_i - m_k)_+^p,$$

$$SPL_2(\pi_i, \boldsymbol{m}) = \lambda_0 + \sum_{l=1}^{q}\lambda_l\pi_i^l + \sum_{k=1}^{K} \nu_k(\pi_i - m_k)_+^q,$$

$$\boldsymbol{b}|\tau_b^2 \sim N(\mathbf{0}, \tau_b^2\boldsymbol{I}_K), \boldsymbol{\nu}|\tau_\nu^2 \sim N(\mathbf{0}, \tau_\nu^2\boldsymbol{I}_K).$$

Assume the following prior distributions for derivation purpose

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \varphi^2 \boldsymbol{I}_{p+1} = (10^3)^2 \boldsymbol{I}_{p+1}), \boldsymbol{\lambda} \sim N(\mathbf{0}, \varphi^2 \boldsymbol{I}_{q+1} = (10^3)^2 \boldsymbol{I}_{q+1})$$

$$\tau_b, \tau_\nu \sim U(0, +\infty)$$

$$\alpha \sim N(0, \psi^2 = 10^2)\mathbf{1}\{\alpha > 0\}.$$

Denote by

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{bmatrix}_{n \times n} \qquad \boldsymbol{L} = (\log \sigma_1^2, \ldots, \log \sigma_n^2)^T$$

$$\boldsymbol{X}^{(1)} = \begin{bmatrix} 1 & \pi_1 & \ldots & \pi_1^p \\ \vdots & \ddots & & \vdots \\ 1 & \pi_n & \ldots & \pi_n^p \end{bmatrix}_{n \times (p+1)} \qquad \boldsymbol{Z}^{(1)} = \begin{bmatrix} (\pi_1 - m_1)_+^p & \cdots & (\pi_1 - m_K)_+^p \\ \vdots & \ddots & \vdots \\ (\pi_n - m_1)_+^p & \cdots & (\pi_n - m_K)_+^p \end{bmatrix}_{n \times K}$$

$$\boldsymbol{X}^{(2)} = \begin{bmatrix} 1 & \pi_1 & \ldots & \pi_1^q \\ \vdots & \ddots & & \vdots \\ 1 & \pi_n & \ldots & \pi_n^q \end{bmatrix}_{n \times (q+1)} \qquad \boldsymbol{Z}^{(2)} = \begin{bmatrix} (\pi_1 - m_1)_+^q & \cdots & (\pi_1 - m_K)_+^q \\ \vdots & \ddots & \vdots \\ (\pi_n - m_1)_+^q & \cdots & (\pi_n - m_K)_+^q \end{bmatrix}_{n \times K},$$

and define $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}^{(1)}\boldsymbol{\beta} - \boldsymbol{Z}^{(1)}\boldsymbol{b}$, $\boldsymbol{r} = \boldsymbol{L} - \boldsymbol{X}^{(2)}\boldsymbol{\lambda} - \boldsymbol{Z}^{(2)}\boldsymbol{\nu}$, the full conditionals of the posteriors for model (2.3) are detailed as below:

1. $[\boldsymbol{\beta}] \sim N_{p+1}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{X}^{(1)T} \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{Z}^{(1)}\boldsymbol{b} - \alpha \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{w}), \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = [\boldsymbol{X}^{(1)T} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}^{(1)} + (\varphi^2 \boldsymbol{I}_{p+1})^{-1}]^{-1}$

2. $[\boldsymbol{b}] \sim N_K(\boldsymbol{\Sigma}_{\boldsymbol{b}} \boldsymbol{Z}^{(1)T} \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}^{(1)}\boldsymbol{\beta} - \alpha \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{w}), \boldsymbol{\Sigma}_{\boldsymbol{b}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{b}} = [\boldsymbol{Z}^{(1)T} \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}^{(1)} + (\tau_b^2 \boldsymbol{I}_K)^{-1}]^{-1}$

3. $[\boldsymbol{\lambda}] \sim N_{q+1}(\boldsymbol{\Sigma}_{\boldsymbol{\lambda}} \boldsymbol{X}^{(2)T} (\sigma_A^2 \boldsymbol{I}_n)^{-1}(\boldsymbol{L} - \boldsymbol{Z}^{(2)}\boldsymbol{\nu}), \boldsymbol{\Sigma}_{\boldsymbol{\lambda}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{\lambda}} = [\boldsymbol{X}^{(2)T} (\sigma_A^2 \boldsymbol{I}_n)^{-1} \boldsymbol{X}^{(2)} + (\varphi^2 \boldsymbol{I}_{q+1})^{-1}]^{-1}$

4. $[\boldsymbol{\nu}] \sim N_K(\boldsymbol{\Sigma_\nu} \boldsymbol{Z}^{(2)T}(\sigma_A^2 \boldsymbol{I}_n)^{-1}(\boldsymbol{L} - \boldsymbol{X}^{(2)}\boldsymbol{\lambda}), \boldsymbol{\Sigma_\nu})$, where $\boldsymbol{\Sigma_\nu} = [\boldsymbol{Z}^{(2)T}(\sigma_A^2 \boldsymbol{I}_n)^{-1}\boldsymbol{Z}^{(2)} + (\tau_\nu^2 \boldsymbol{I}_K)^{-1}]^{-1}$

5. $[\alpha] \sim N(\sigma_\alpha^2 \boldsymbol{w}^T \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{e}, \sigma_\alpha^2)\mathbf{1}\{\alpha > 0\}$, where $\sigma_\alpha^2 = (\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{\psi^2})^{-1}$

6. $[\boldsymbol{w}] \sim N_n(\boldsymbol{\Sigma_w}\alpha\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{e}, \boldsymbol{\Sigma_w}) \prod_{i=1}^n \mathbf{1}\{w_i > 0\}$, where $\boldsymbol{\Sigma_w} = (\alpha^2 \boldsymbol{I}_n + \boldsymbol{I}_n^{-1})^{-1}$

7. $[\tau_b^2] \sim IG(\frac{K-1}{2}, \frac{1}{2}\boldsymbol{b}^T \boldsymbol{I}_K^{-1}\boldsymbol{b})$

8. $[\tau_\nu^2] \sim IG(\frac{K-1}{2}, \frac{1}{2}\boldsymbol{\nu}^T \boldsymbol{I}_K^{-1}\boldsymbol{\nu})$

9. $[\sigma_i] \propto \exp\left\{-\frac{1}{2\sigma_i^2}(\boldsymbol{e} - \alpha\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{w})^T(\boldsymbol{e} - \alpha\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{w})\right\} \frac{1}{\sigma_i^3} \exp\left\{-\frac{1}{2\sigma_A^2}\boldsymbol{r}^T\boldsymbol{r}\right\} \prod_{i=1}^n I_{(0,+\infty)}(\sigma_i)$

A.3 Stan Scripts

A.3.1 SN-BPSP

```
data {

  int<lower=1> n1;  // # of sampled units
  int<lower=1> n2;  // # of non-sample units

  int<lower=1> p;
  int<lower=1> K;  // # of truncated polynomial bases

  vector[n1] y;  // survey variable of units in the sample

  row_vector[p] X[n1];
  row_vector[K] Z[n1];  // truncated polynomials
```

```
  row_vector[p] predX[n2];
  row_vector[K] predZ[n2];  // truncated polynomials

  real<lower=0> Pi[n1];
  real<lower=0> predPi[n2];

}

parameters {

  vector[p] beta;
  vector[K] b;
  real alpha;
  real<lower=0> sigma;
  real gamma;
  real<lower=0> sigmab;

}

transformed parameters {

  real xi[n1];
  real<lower=0> omega[n1];

  for (i in 1:n1) {
    xi[i] = X[i] * beta + Z[i] * b;
    omega[i] = sqrt( (pow(alpha, 2) + 1) ) * pow(Pi[i], gamma) * sigma;
  }
}
```

```
model {

  for (i in 1:n1) {
    y[i] ~ skew_normal(xi[i], omega[i], alpha);
  }

  for (l in 1:p) {
    beta[l] ~ normal(0, 1e3);
  }

  for (l in 1:K) {
    b[l] ~ normal(0, sigmab);
  }

  sigmab ~ cauchy(0, 1);
  sigma ~ cauchy(0, 1);

  alpha ~ normal(0, 10) T[0, ];
}
```

A.3.2 SN-B2PSP

```
data {

  int<lower=1> n1;  // # of sampled units
  int<lower=1> n2;  // # of non-sample units

  int<lower=1> p;
```

```
  int<lower=1> K;  // # of truncated polynomial bases

  vector[n1] y;  // survey variable of units in the sample

  row_vector[p] X[n1];
  row_vector[K] Z[n1];  // truncated polynomials

  row_vector[p] predX[n2];
  row_vector[K] predZ[n2];  // truncated polynomials
}

parameters {

  vector[p] beta;
  vector[K] b;
  real alpha;
  vector[p] lambda;
  vector[K] nu;
  real<lower=0> sigma[n1];
  real<lower=0> sigmab;
  real<lower=0> sigmanu;

}

transformed parameters {

  real xi[n1];
  real<lower=0> omega[n1];
  real SPL2[n1];
```

```
  for (i in 1:n1) {
    xi[i] = X[i] * beta + Z[i] * b;
    omega[i] = sqrt( (pow(alpha, 2) + 1) ) * sigma[i];
    SPL2[i] = X[i] * lambda + Z[i] * nu;
  }

}

model {

  for (i in 1:n1) {
    y[i] ~ skew_normal(xi[i], omega[i], alpha);
    sigma[i] ~ lognormal(SPL2[i], 0.1);
  }

  for (l in 1:p) {
    beta[l] ~ normal(0, 1e3);
    lambda[l] ~ normal(0, 1e3);
  }

  for (l in 1:K) {
    b[l] ~ normal(0, sigmab);
    nu[l] ~ normal(0, sigmanu);
  }

  sigmab ~ cauchy(0, 1);
  sigmanu ~ cauchy(0, 1);
```

```
  alpha ~ normal(0, 10) T[0, ];
}
```

Appendix B

Appendices to Chapter 3

B.1 Figures

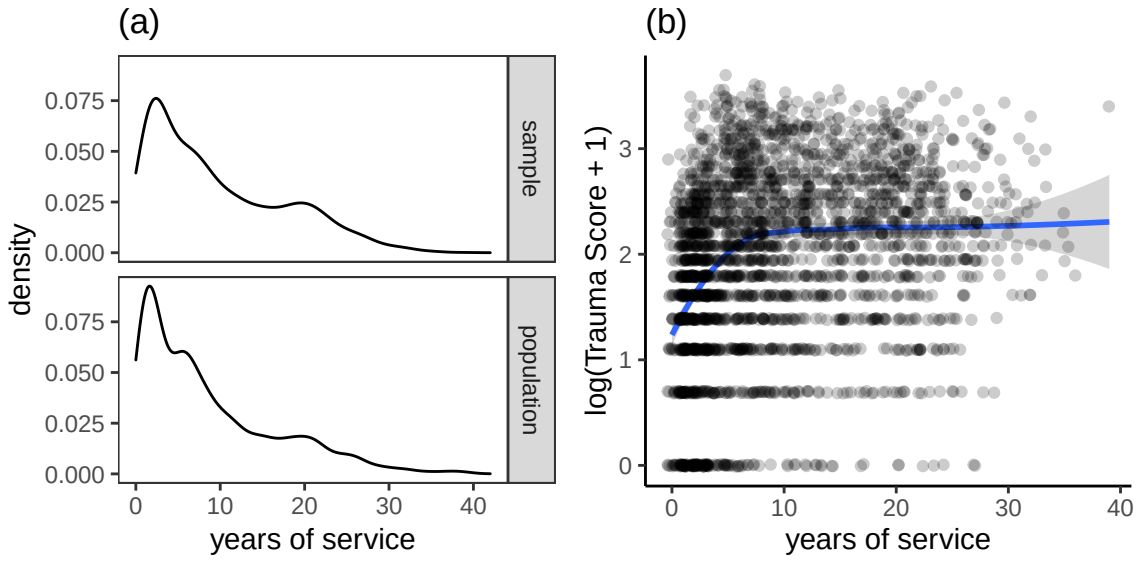

Figure S1: (a) Density plots of years of service among the OHARNG soldiers in the sample and population (b) Scatterplot of log(trauma score + 1) vs years of service among OHARNG soldiers in the sample from with a locally estimated scatterplot smoothing (LOESS) surve

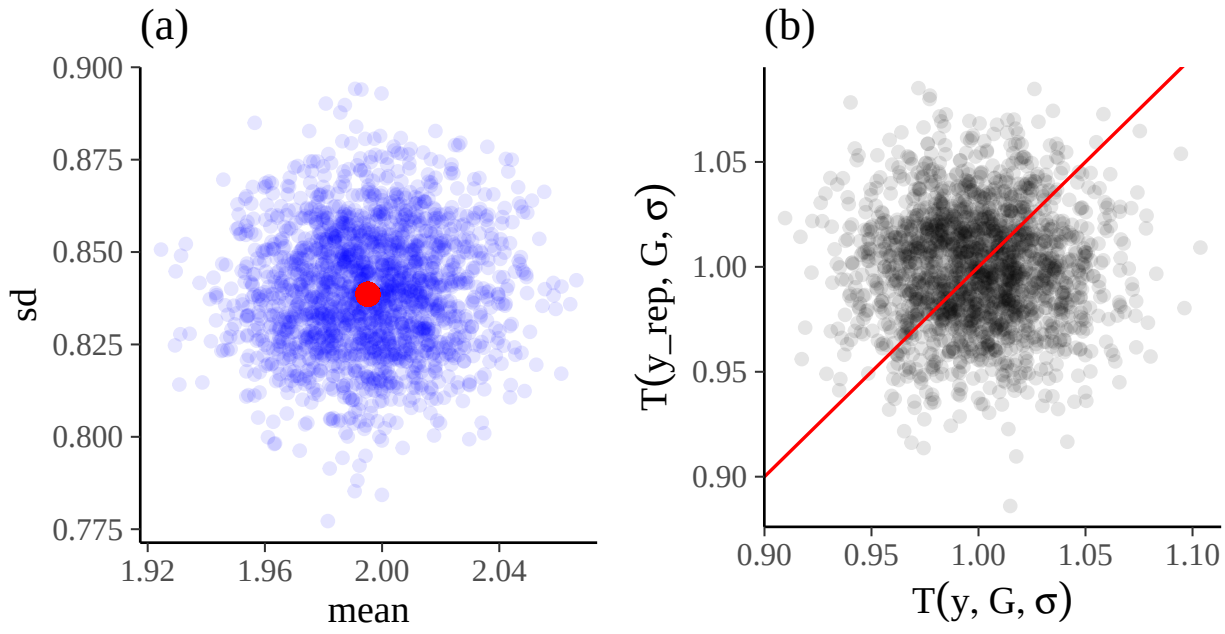

Figure S2: Realized versus posterior predictive distributions for the test quantities (a) $(T_1(\boldsymbol{y}), T_2(\boldsymbol{y})) = (\text{mean}, \text{sd}) = (\bar{y}, \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2)$ and (b) $T_3(\boldsymbol{y}, G, \sigma) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \theta_i}{\sigma}\right)^2$ with $\theta_i = G(\boldsymbol{z}_i, \boldsymbol{x}_i)$. The observed quantity $(T_1(\boldsymbol{y}), T_2(\boldsymbol{y}))$ is at the center of the cloud of the predictive quantities and the observed quantity $T_3(\boldsymbol{y}, G, \sigma)$ has about half the chance to be below the 45 degree line. The Bayesian posterior predictive p-values for $T_1(\cdot)$, $T_2(\cdot)$ and $T_3(\cdot)$ are $p_1 = .50$, $p_2 = .51$ and $p_3 = .50$, respectively.

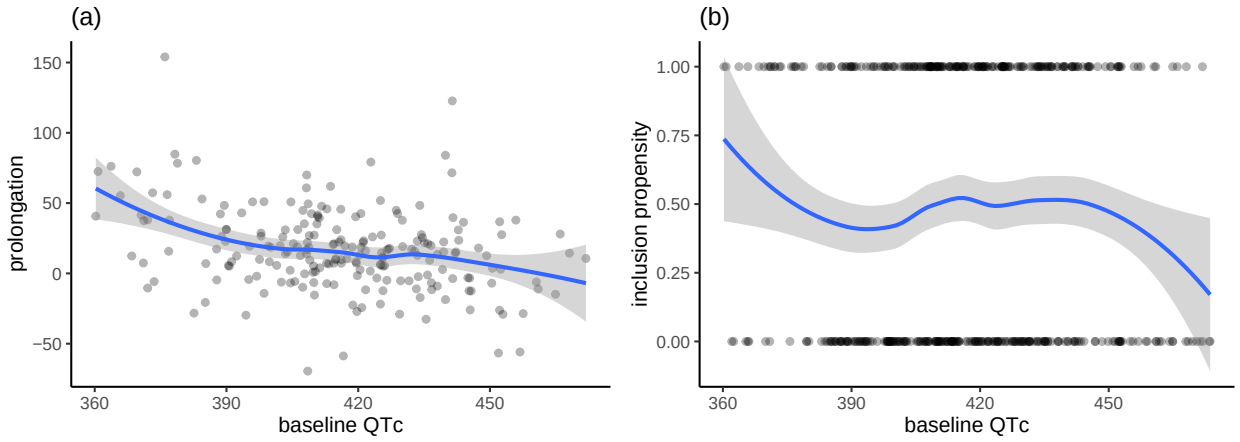

Figure S3: (a) Scatterplot of prolongation vs baseline QTc with a LOESS curve (b) Inclusion propensity vs baseline QTC using LEOSS among COVID patients admitted at CUIMC