BAYESIAN PREDICTIVE INFERENCE FOR THREE TOPICS IN SURVEY SAMPLES

by

Qixuan Chen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2009

Doctoral Committee:

Associated Professor Michael R. Elliott, Co-Chair
Professor Roderick J.A. Little, Co-Chair
Professor Emeritus David H. Garabrant
Professor James M. Lepkowski

To My Parents

Acknowledgements

This research was supported in part by The Dow Chemical Company through an unrestricted grant to the University of Michigan Dioxin Exposure Study. I appreciate the helpful advices and useful comments from my advisors Professors Roderick Little, Michael Elliott, James Lepkowski, and David Garabrant. I also thank the referees at Survey Methodology and Epidemiology for their comments that led to valuable improvements through the research.

Table of Contents

List of Figures

List of Tables

Chapter I

INTRODUCTION

Population-based health studies in a community or in the nation require a well-planned survey design. Sampling designs usually involve multistage cluster sampling and unequal selection probabilities. A statistical analysis that ignores the complex design feature could lead to biased estimates and invalid inferences of the population quantities. To make valid inferences for the target population, survey design information needs to be incorporated in the statistical analysis. This thesis is concerned with two major topics of finite population inference. In Chapter II and III, I focus on the model-based inference of finite population proportions and quantiles in unequal probability sampling. In Chapter IV and V, I study two specific problems of incomplete data analysis from complex survey samples.

In descriptive survey inference, the target population consists of all the units in the population, from which the sample is drawn. Let *N* be the number of units in the population and let $Y = \{Y_1, Y_2, \ldots, Y_N\}$ be the outcome variable of interest. Let $I = \{I_1, I_2, \ldots, I_N\}$ be the indicator of inclusion in the sample, with $I_i = 1$ if Y_i is observed and $I_i = 0$ if Y_i is not observed. The indicator for inclusion is fully observed for all the units in the population. Assume that the probabilities of selection $\pi = \{\pi_1, \pi_2, \ldots, \pi_N\}$ are known for all the units in the population before a sample is drawn. Let *s* be an unequal probability random sample and *n* be the number of units in the

sample. The survey outcome Y can be partitioned into two parts, where Y_{obs} denote the sample values with $I = 1$ and $Y_{non-obs}$ denote the non-sampled values corresponding to $I = 0$.

There are two inferential paradigms in the estimation of population values for survey samples. They are design-based and model-based inferences.

In the design-based framework, the population is regarded as fixed whereas the inference is based on the random distribution of the sample inclusion indicator. The random distribution refers to the distribution of the estimates that result from all possible samples under the sampling design. Horvitz and Thompson (1952) formulated three classes of linear estimators for population totals in unequal probability sampling and derived the most commonly used design-based Horvitz-Thompson estimator: $\hat{T} = \sum_{i \in s} y_i / \pi_i$. When the population total of an auxiliary variable *X* associated with *Y* is known and *X* is observed in the sample, a more efficient estimator is the ratio estimator $\left(\sum_{i \in s} y_i / \pi_i \Big/ \sum_{i \in s} x_i / \pi_i\right) \sum_{i=1}^{N} x_i$. These design-based estimators are design consistent (Isaki and Fuller 1982) and provide reliable inferences in large samples by assuming asymptotic normality. However, the design-based estimators are potentially very inefficient. Moreover, the small sample distribution of the design-based estimators is generally unknown. The variance estimation is not clear and sometimes cumbersome because it requires second-order selection probabilities.

An alternative approach is model-based inference. The model-based approach assumes that the finite population has itself been generated from a super population. If the super population can be specified with some distribution, the values in the non-sampled units can be related to the values in the sample via the assumed super population

distribution. For the population total, the model-based estimator is defined as $\hat{T} = \sum_{i \in s} Y_i + \sum_{j \in s} \hat{Y}_j$, where $\hat{Y}_j$ is the predictor of Y_j from some regression model. When some auxiliary or design variables are available, the model-based estimator, by employing the relationship between the survey outcome and these auxiliary or design variables, can improve the efficiency compared to the design-based estimators (Smith 1976). In general, the variance of the design-based estimators is larger than the model-based estimators under a given model. The increase in variance is higher the smaller the sample size and the larger the variability of the weights (Pfeffermann 1993, 1996). Model-based inference is a predictive statistical inference. In contrast to the design-based estimators, the inference of the model-based estimators is based on the distribution of *Y* instead of the randomness in the sampling process. Therefore, the variance and confidence interval are more straightforward to calculate in model-based inference.

When the sampling design is defined by a set of design variables other than the response variable and the design variables are included in the model, the design mechanism is ignorable. Rubin (1983) showed that when the probability sampling mechanism satisfied $\Pr(I \mid Y, X) = \Pr(I \mid \pi)$, where *X* are the design variables, the selection probabilities provide a complete summary of the data used to make sampling decisions. Thus, conditioning on π the sampling design is ignorable and modelers should focus primarily on the proper specification of the conditional distribution of the outcome variable given the selection probabilities.

The model-based prediction estimator is efficient under the assumed model, but is potentially subject to very large bias when the underlying model is misspecified. Parametric models are not adequate to approximate all continuous functions. Recently,

efforts have been made to develop nonparametric model-based inference of descriptive population quantities. In Kuk (1993), a kernel-based estimator was proposed, which combines the known distribution of the auxiliary variable with a kernel estimate of the conditional distribution of the survey variable given the value of the auxiliary variable. Chambers, Dorfman, and Wehrly (1993) also gave a model-based estimator using kernel smoothing to estimate distribution functions. Zheng and Little (2003, 2005) estimated the finite population total using a nonparametric regression on a penalized spline (p-spline) of the selection probabilities for continuous survey data.

In Chapter II and III, I extend the penalized spline model-based estimator for population totals proposed by Zheng and Little (2003) to estimate population proportions and quantiles using penalized splines on the selection probabilities. I consider the setting of one-stage unequal probability sampling without clusters or strata, ignoring the problems of nonresponse and post-stratification.

In Chapter II, I give a Bayesian p-spline predictive (BPSP) estimator of proportions that is suitable for a binary outcome. I adopt a Bayesian approach to inference for this model, since Bayesian methods often yield better inference for small sample problems and are conveniently implemented for our proposed model via the Gibbs' sampler (Gelman, Carlin, Stern, and Rubin, 2004). Simulation studies show that the BPSP estimator is more efficient, and its 95% credible interval provides better coverage with shorter average width than the sample-weighted and the generalized regression estimators, especially when the population proportion is close to zero or one for small samples. Compared to parametric model-based predictive estimators, the BPSP estimators achieve robustness to model misspecification and influential observations in the sample.

In Chapter III, I develop two robust Bayesian model-based estimators of finite population quantiles for continuous survey variables in unequal probability sampling. The first method is to estimate cumulative distribution functions of the continuous survey variable using the BPSP approach. The finite population quantiles are then obtained by inverting the distribution function, but heavy computation is involved in this procedure. Therefore, I consider the second method that posits a smoothly-varying relationship between the continuous survey variable and the probability of selection by modeling both the mean function and the variance function nonparamtrically. Simulation studies show that both methods yield smaller root mean squared errors than the sample-weighted estimator. With spare data included in the sample, the 95% credible intervals of the two new methods have closer to the nominal level confidence coverage than the sample-weighted estimator.

The second part of the thesis focuses on incomplete data analysis with complex survey data. In survey studies, item nonresponse happens when particular items are missing for an individual. If the amount of item nonresponse is nontrivial or if a small amount of nonresponse occurs in several variables in different individuals, the default strategy of eliminating all incomplete cases from the analysis wastes data that is costly to collect, and can lead to problematic inference for the target population. Item nonresponse is often handled by multiple imputation (Rubin 1987; Little and Rubin 2002). Multiple imputation refers to a procedure of replacing each missing value by multiple imputed values. In Chapters IV and V, I discuss two statistical issues in using multiple imputation, which are motivated by the University of Michigan Dioxin Exposure Study (UMDES).

The UMDES was designed to assess exposures to polychlorinated dibenzo-p-dioxins in the adult population of Midland and Saginaw Counties, Michigan, USA. (Garabrant et al, 2009) The sampling used a stratified two-stage area probability selection of housing units and a third stage of selection of an eligible person within each sample housing unit. (Lepkowski et al. 2006) As is typical in surveys, item nonresponse was encountered in the UMDES and was handled by multiple imputation. By measuring factors that reflect potential exposure to dioxins, the study sought to determine factors that explain variation in serum dioxin concentrations. However, within the framework of multiple imputation, the complete-data variable selection methods often yielded a different selected model from each imputed data set, which posed difficulties in pooling the variable selection results across multiple imputed data sets. Some authors suggest including in a model variables that were selected (for example using stepwise selection) in at least 3 out of 5 (60%) of the imputed data sets (Heymans et al. 2007; Wood et al. 2005; Brand 1999). I refer to this variable selection strategy as the "select then combine" (SC) method. For simplicity, researchers also perform stepwise variable selection with any one of the multiply imputed data sets. I call this approach the "single imputation" (SI) method. In Chapter IV, I develop a "combine then select" (CS) method which calculated combined *p*-values using the multiple imputation combining rule and then selected variables based on the combined *p*-values in each step of the selection. The CS method leads to a single selected variable set. With the CS method, uncertainty due to missing data is taken into account in the variable selection process, which leads to a parsimonious model.

The limit of detection (LOD) issue has posed formidable limitations to the estimation of serum dioxin concentrations in the general population. The LOD is defined as the

concentration of analyte which gives a signal equal to a laboratory blank (obtained when no analyte is present) plus three times the standard deviation of the blank. (Keith et al. 1983) The LOD represents the level below which we cannot be confident whether or not the analyte is actually present. Conventional approaches of imputing the values below the LOD as 0, LOD, LOD/2, or LOD/$\sqrt{2}$ depend on the blood sample volume and the LOD levels of the measurement methods and may lead to biased estimates of serum TCDD concentration, especially in the scenario of high proportion of data below the LOD and high LOD levels. (Hornung and Reed 1990) In Chapter V, I employ a proper multiple imputation approach to impute the serum dioxin concentrations for those below the LOD in the combined data of the NHANES 2003-2004 (n=719) and the UMDES reference population (n=251). For each imputation, a bootstrap sample is generated from the combined data, and a weighted left-censored linear regression model, assuming a lognormal distribution, is fitted on the bootstrap sample with important predictors of serum dioxin concentrations. For those subjects having values below the LOD, the natural logarithm transformed imputed values are drawn from a normal distribution with mean and variance estimated from the left-censored regression model, with left truncation at their corresponding natural logarithm LOD. The above procedure is repeated five times to generate five imputed data sets. The multiply imputed complete data is then used to predict the age- and gender- specific percentiles of serum dioxin concentrations using survey-weighted quantile regression models among the non-Hispanic white population.

Chapter II

BAYESIAN PENALIZED SPLINE MODEL-BASED INFERENCE FOR FINITE POPULATION PROPORTION IN UNEQUAL PROBABILITY SAMPLING

II.1. Introduction

Unequal probability sampling designs are commonly employed in data collection by science and government. Perhaps the simplest unequal probability design is stratified sampling, which samples units from different strata with different selection probabilities. Another important form of unequal probability sampling is probability-proportional-to-size (pps) sampling, in which the selection probability is proportional to the value of a size variable measured for all population units.

An unequal probability sampling design such as pps sampling is often used for efficient estimation of population means of continuous variables, for which the variance increases with size of sample unit. However, inferences about discrete variables are often also of interest in a multipurpose survey (e.g. Lehtonen and Veijanen 1998, Lehtonen, Särndal and Veijanen 2005). In this paper, we focus on methods of inference for finite population proportions from unequal probability sampling designs, based on an auxiliary variable measured for all the units in the population. We use pps sampling as a specific design to illustrate and assess our methods.

The selection probabilities play important and somewhat different roles in design-based and model-based inference from unequal probability survey samples (Smith 1976,

1994; Kish 1995; Little 2004). In design-based inference, survey variables are fixed, and inference is based on the distribution of the sample inclusion indicators; the standard design-based approaches to estimation such as the Horvitz-Thompson (HT) estimator (1952) and its extensions weight sampled units by the inverse of their selection probabilities. These estimators are design consistent (Isaki and Fuller 1982) and provide reliable inferences in large samples without the need for modeling assumptions. However, these estimators are potentially very inefficient, as illustrated in Basu's (1971) famous elephant example. Also, variance estimation is cumbersome because it requires second-order selection probabilities. Corresponding confidence intervals are based on asymptotic theory, and may deviate from nominal levels for moderate or small sample sizes.

Model-based inference predicts values of survey variables in the non-sampled units by including the selection probabilities as covariates in the prediction model (Little 2004). Model-based prediction estimators are consistent and efficient under the assumed model, but are potentially subject to very large bias when the underlying model is misspecified. This limitation motivates the development of flexible statistical models that are more robust to model misspecification. For continuous survey data, Zheng and Little (2003) estimated the finite population total using a nonparametric regression on a penalized spline (p-spline) of the selection probabilities. We propose here Bayesian p-spline prediction (BPSP) estimators that are suitable for a binary, as opposed to continuous, outcome. We adopt a Bayesian approach to inference for this model, since Bayesian methods often yield better inference for small sample problems, and are conveniently implemented for our proposed model via the Gibbs' sampler. In this approach, auxiliary

variables other than the selection probability can also be included in the model, but the selection probability is singled out since modeling of this variable is prone to model misspecification.

We compare the performance of BPSP estimators with Hájek (Horvitz-Thompson-type) estimators and with generalized regression (GR) estimators for a binary outcome proposed by Lehtonen and Veijanen (1998). The GR approach is a popular model-assisted modification of the design-based estimators that combines predictions from a model with design-weighted model residuals (Montanari 1998), to yield estimates that are approximately design unbiased.

Zheng and Little (2003; 2005) compared HT, penalized spline prediction, and GR estimates of the total of a continuous survey variable by simulation. They found that p-spline model-based estimators had better root mean squared error than the other methods, and with jackknife standard errors providing superior confidence coverage to HT or GR inferences. We conduct similar comparisons for inference about a population proportion for a binary outcome, and show similar advantages for our BPSP estimator over the design-based and GR alternatives.

II.2. Design-based estimator

Suppose that we have a finite population consisting of N identifiable units. Let Y be the binary survey variable of interest and $p = N^{-1}\sum_{i=1}^{N} Y_i$ be the proportion of the population for which $Y = 1$. Let π_i denote the probability of inclusion for unit i, which is assumed to be known for all units in the finite population before a sample is drawn.

An unequal probability random sample *s* with elements $y_1,...y_n$ is then drawn from the finite population according to the selection probabilities $\pi_1,...,\pi_N$. The design-based Hájek estimator in the discussion of Basu (1971) is defined as

$$\hat{p}_H = \frac{\sum_{i \in s} y_i / \pi_i}{\sum_{i \in s} 1/\pi_i}. \quad (1)$$

The variance for $\hat{p}_H$ can be estimated via linearization of the Yates-Grundy estimator (1953) of totals,

$$\hat{V}_{YG}(\hat{p}_H) = \left(\sum_{k \in s} 1/\pi_k \right)^{-2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i - \hat{p}_H}{\pi_i} - \frac{y_j - \hat{p}_H}{\pi_j} \right)^2 .$$

The Yates-Grundy variance estimator requires pairwise selection probabilities. When the pairwise selection probabilities are not available, as in our simulations, the approximate formula proposed by Hartley and Rao (1962),

$$\pi_{ij} \approx \frac{n-1}{n} \pi_i \pi_j + \frac{n-1}{n^2} \left(\pi_i^2 \pi_j + \pi_i \pi_j^2 \right) - \frac{n-1}{n^3} \pi_i \pi_j \sum_{k=1}^{N} \pi_k^2 ,$$

has frequently been used. An approximate $1-\alpha$ level confidence interval for the population proportion $\hat{p}_H$ is then obtained based on the normal approximation.

II.3. Bayesian p-spline predictive (BPSP) estimator

Royall (1970) argued for the use of models for finite-population descriptive inferences by predicting the unobserved values based on models, since model-based inferences should be more efficient than design-based inferences. To model the relationship between the binary outcome *Y* and the continuous selection probability π, we need to fit a binary regression of *Y* on π. Parametric binary regressions, such as the

linear or quadratic logistic or probit model, may not be adequate in fitting the data. One solution for this problem of inflexibility is to fit a binary regression on a spline of π by adding some knots. However, too many knots may result in roughness of model fit. One way to overcome this problem is to retain all of the knots but to constrain their influence, by fitting a binary penalized spline (p-spline) regression model.

Common methods for modeling a binary outcome are logistic and probit regressions, and they generally give similar results. We choose to adopt probit models in our study for computational convenience. The probit regression model for binary outcomes has an underlying truncated normal regression structure on latent continuous data. If the latent continuous data are known, the parameters in binary p-spline regression models can be estimated using standard approaches for Gaussian p-spline regression models. In a Bayesian context, the posterior distribution of parameters in the probit p-spline model can be computed using Gibbs sampling (Albert and Chib 1993; Ruppert, Wand, and Carroll 2003, chapter 16). In contrast, the logistic p-spline regression model requires a more complicated computation procedure such as the Metropolis-Hastings algorithm. The computational advantage makes the probit link function more desirable than the logit link function in Bayesian binary p-spline regression models.

There are various types of p-splines. When applying p-splines, we need to make choices on the degree and knot locations, and the basis functions used to present the model. We choose to use the truncated polynomial p-splines because they are simple and intuitive. More numerically stable estimators can be obtained using B-splines via orthogonalizing the truncated power bases (Eilers and Marx 1996). The probit truncated

polynomial p-spline regression model has a generalized linear mixed model representation,

$$\Phi^{-1}\left(E\left(y_i \mid \beta, b, \pi_i\right)\right) = \beta_0 + \sum_{k=1}^{p} \beta_k \pi_i^k + \sum_{l=1}^{m} b_l \left(\pi_i - k_l\right)_+^p \tag{2}$$

$$b_l \sim N\left(0, \tau^2\right)$$

$$l = 1,..., m\,; i = 1,..., n\,,$$

where $\Phi^{-1}(\cdot)$ denote the inverse CDF of a standard normal distribution, and the constants $k_1 < ... < k_m$ are *m* selected fixed knots. A function such as $\left(\pi_i - k\right)_+^p$ is called a truncated polynomial spline basis function with power *p*, where $\left(u\right)_+^p$ is equal to $\left\{u \times I\left(u \geq 0\right)\right\}^p$ for any real number *u*. Since the truncated polynomial spline basis function has *p*-1 continuous derivatives, higher values of *p* lead to smoother spline functions. By specifying a normal distribution for *b*, the influence of the *m* knots is constrained in Model (2), which is equivalent to smoothing the splines via the penalized likelihood.

The parameters in Model (2) can be estimated using generalized linear mixed model methods. An alternative Bayesian approach that simplifies computation is to assume weak prior and hyperprior distributions and use Gibbs sampling to obtain draws from the posterior distributions of the parameters as follow: the probit regression model for binary responses has an underlying normal regression structure on latent continuous data; if the latent data are known, the posterior distribution of the parameters can be computed using standard results for normal regression models; and given the posterior distribution of the parameters, the latent continuous data can be simulated from a suitable truncated normal distribution. (Ruppert, Wand, and Carroll 2003, p. 290) The detailed algorithm of Gibbs sampling is in the Appendix. In addition, the Bayesian inference for p-spline regression

can also been implemented using WinBUGS, the standard Bayesian analysis software (Crainiceanu, Ruppert, and Wand 2005).

The posterior distribution of the population proportion is simulated by generating a large number D of draws of the form $p^{(d)} = N^{-1}\left(\sum_{i\in s} y_i + \sum_{j\notin s} \hat{y}_j^{(d)}\right)$, where $\hat{y}_j^{(d)}$ is a draw from the posterior predictive distribution of the j^{th} non-sampled unit of the binary outcome. The average of these draws simulates the Bayesian p-spline predictive (BPSP) estimator of the finite population proportion, and is denoted as $\hat{p}_{BPSP}$. The Bayesian analog of a $100\times(1-\alpha)\%$ confidence interval for the population proportion is a $100\times(1-\alpha)\%$ credible interval, which can be formed in a number of different ways. We split the tail area α equally between the upper and lower endpoints in the simulations.

Firth and Bennett (1998) showed that any parametric logistic regression model containing an intercept term and the inverse of selection probabilities as a covariate, fitted by ordinary, unweighted maximum likelihood, was "internally bias calibrated" (IBC) for population proportions, and thus yields design consistency. This property is also true for logistic truncated polynomial p-spline regression models on the inverse of selection probabilities, fitted via penalized likelihood. With the probit link function used instead of the logit link function and fitted via Markov chain Monte Carlo algorithm instead of maximum penalized likelihood, the BPSP estimator may no longer have the IBC property. However, the similarity between the probit model and the logistic model implies that the predictive estimator based on a probit p-spline regression model is approximately design-consistent. We believe that obtaining efficient estimates with close to nominal confidence coverage in finite samples is more important than exact design consistency.

II.4. Generalized regression (GR) estimator

For the estimation of class frequencies of a discrete response variable, Lehtonen and Veijanen (1998) proposed a GR estimator $\hat{t}$ of the total, which combines the predicted values $\hat{u}_i = \hat{\Pr}(Y_i = 1 \mid \pi_i)$ based on a suitable model and the HT estimator for the residuals $r_i = y_i - \hat{u}_i$ of the sampled units,

$$\hat{t} = \sum_{i=1}^{N} \hat{u}_i + \sum_{i \in s} r_i / \pi_i \tag{3}$$

The GR estimator in Equation (3) is then used in constructing an estimator for population proportions by dividing by the known population size *N* (Duchesne 2003),

$$\hat{p}_{GR\_1} = \frac{1}{N}\left(\sum_{i=1}^{N} \hat{u}_i + \sum_{i \in s} r_i / \pi_i \right). \tag{4}$$

We also consider here another version of the GR estimator for the estimation of finite population proportions, in which the denominator of the bias calibration term for the residuals r_i is the estimated population size $\sum_{i \in s} 1/\pi_i$,

$$\hat{p}_{GR\_2} = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i + \left(\sum_{i \in s} r_i / \pi_i \right) \left(\sum_{i \in s} 1/\pi_i \right)^{-1} \tag{5}$$

For the variance estimate of (4), we use the variance estimator of the estimated total of a discrete response variable, given by Lehtonen and Veijanen (1998), divided by N^2. For the variance estimate of (5), we apply the Taylor linearization technique (Särndal, Swensson, and Wertman 1992, p.182).

$$\hat{V}\left(\hat{p}_{GR\_1}\right) = \frac{1}{N^2} \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{r_k}{\pi_k} \frac{r_l}{\pi_l},$$

$$\hat{V}\left(\hat{p}_{GR\_2}\right)=\left(\sum_{i\in s}1/\pi_i\right)^{-2}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\frac{e_k}{\pi_k}\frac{e_l}{\pi_l},$$

where $e_k=r_k-\left(\sum_{i\in s}r_i/\pi_i\right)\left(\sum_{i\in s}1/\pi_i\right)^{-1}$. These variance estimators also require pairwise selection probabilities, which can be approximated by the method of Hartley and Rao (1962).

However, the Hartley and Rao approximation may lead to bias in the variance estimator. Thus, we also consider the jackknife method for variance estimation (Shao and Wu 1989). The sample is stratified into *n/G* strata each of size *G* with similar values of selection probabilities, and the *G* subgroups are then constructed by selecting one element at a time from each stratum without replacement (Zheng and Little 2005). Let $\hat{p}_{(g)}$ be the same GR estimators in (4) and (5) calculated from the reduced sample without the elements in the g^{th} subgroup, and let $\bar{p}$ be the average of the *G* estimators based on the *G* reduced samples. The jackknife variance estimator of $\hat{p}_{GR}$ is

$$\hat{V}_{jackknife}\left(\hat{p}_{GR}\right)=\frac{G-1}{G}\sum_{g=1}^{G}\left(\hat{p}_{(g)}-\bar{p}\right)^2.$$

A design-weighted logistic regression model on other covariates was used as the assisting model to predict $\hat{u}_i$ in the GR estimators for binary outcomes (Lehtonen and Veijanen 1998; Lehtonen, Särndal, and Veijanen 2005). Since our interest here is in comparisons of GR estimators with the BPSP estimator, we apply the estimators (4) and (5) with parametric probit models and probit p-spline models, as described in detail in Section 5. For the GR estimator using a linear probit model as the assisting model, we apply the selection probability as a covariate as well as a weight in our simulations.

II.5. Simulation study

II.5.1 Design of the simulation study

Simulation studies are conducted to study the performance of the BPSP estimator compared with the Hájek estimator, the GR estimators, and the parametric model-based predictive estimators for a variety of populations in pps sampling. We present the simulation results for the following six estimators:

a) HK, the Hájek estimator in equation (1).

b) LR, predictive estimator with the maximum likelihood predictions from the linear logistic regression model containing a constant term and the reciprocal selection probability as the covariate. LR has the IBC property, and hence is design-consistent. LR is exactly the same as its GR estimator in equation (4).

c) PR, predictive estimator with predictions from the Bayesian linear probit model containing an intercept term and the selection probability as the covariate.

d) PR_GR, the GR estimator in equation (5), where $\hat{u}_i$ is the prediction for unit *i* with unknown parameters replaced by weighted maximum likelihood estimates from the probit model with a constant term and the selection probability as the covariate.

e) BPSP, the BPSP estimator with $p = 1$ and inverse-gamma prior distribution for τ^2 and using 15 knots.

f) BPSP_GR, the GR estimator in equation (5), where $\hat{u}_i$ is the posterior mean of $\Pr(Y_i = 1 \mid \pi_i)$ from the BPSP model.

We only report the simulation results based on the linear splines for the BPSP estimator, since simulations not shown here suggest that linear splines perform as well as quadratic splines or cubic splines in all the simulation scenarios. We choose two fixed numbers of knots (15 or 30), and place knots at evenly spaced sample percentiles. The choices of knots work well and a number of 15 knots is good enough to catch the curvatures in our simulations. In addition, the GR estimators in (4) perform similarly to the estimators in (5); some differences between these estimators emerge in the real application in Section 6, leading us to prefer (5) over (4).

We simulated two artificial populations of size 2,000 with sampling rates of 5% and 10%, where the size variable takes the consecutive integer values 71, 72, ..., 2,070. The selection probabilities in the population were then calculated as proportional to the size variable, with the maximum value about 30 times the minimum value.

Continuous data Z were generated from normal distributions with mean structure $f(\pi)$ and constant error variance 0.04. Two different mean structures $f(\pi)$ were simulated: a linearly increasing function (LINUP) $f(\pi_i) = k_1\pi_i$ and an exponential function (EXP) $f(\pi_i) = \exp(-4.64 + k_2\pi_i)$. To make the range of *Z* similar across different mean structures, k_1 takes values of 3 and 6, and k_2 takes values of 26 and 52, when the sampling rate is 10% and 5%, respectively. Figure II.1 plots the two populations. Binary outcomes $Y = \{Y_1, Y_2, Y_3\}$ were then created by using the superpopulation 10th, 50th, and 90th percentiles of *Z* as cut-off values. For instance, Y_1 is equal to one if *Z* is less than or equal to its superpopulation 10th percentile, otherwise Y_1 is equal to zero. The target of inference here is the population proportion with *Y* equal to one.

In each simulation replicate, a finite population was generated before a sample was drawn. A pps sample was then drawn systematically from a randomly ordered list of the finite population. For each population and sample size combination, 1,000 replicates were obtained and the six estimators were compared in terms of bias, root mean squared error (RMSE), and the non-coverage rate of the 95% confidence /credible interval. Simulation results are presented in Tables II.1 through II.3.

II.5.2 Simulation results

Figure II.2 shows the posterior means of $\Pr(Y_i = 1 \mid \pi_i)$ and 95% credible intervals based on the Bayesian probit linear p-spline model for a random pps sample from the EXP case. The upper left plot is the scatter plot of the continuous variable *Z* in a pps sample, with three horizontal parallel lines superimposed, representing the superpopulation 10^{th}, 50^{th}, and 90^{th} percentiles, respectively. In the upper right plot, the binary variable *Y*, defined as 1 if *Z* is less than or equal to the superpopulation 10^{th} percentile, are plotted with black circles, and the superpopulation $\Pr(Y_i = 1 \mid \pi_i)$ are plotted with a solid black curve. The solid grey curve and two dashed grey curves are the posterior means of $\Pr(Y_i = 1 \mid \pi_i)$ and 95% credible intervals based on the Bayesian probit linear p-spline regression model. The other two plots are similar to the upper right plot, but with superpopulation 50^{th} and 90^{th} percentiles as cut-off values in defining *Y*. These plots show that the true probabilities of $Y = 1$ fall within the 95% credible intervals, and are close to the posterior means of $\Pr(Y_i = 1 \mid \pi_i)$. We conclude that the Bayesian probit p-spline regression model fits well for the binary outcomes in the nonlinear case.

Table II.1 shows the empirical bias ($\times 10^4$) for the six estimators in the two populations. Overall the design-based estimators are less biased than the model-based estimators. In the LINUP case, the linear probit regression model is correctly specified, so that the empirical bias of the PR estimators are similar to the empirical bias of the BPSP estimator; while in the EXP case, a nonlinear probit regression is needed to fit the data, and thus the PR estimator is more biased than the BPSP estimator when the true population proportions are 0.1 and 0.5. However, we do not observe this severe bias in the LR estimator because of its IBC property. Compared to the model-based PR and BPSP estimators, the PR_GR and BPSP_GR estimator reduce the bias by adding the bias calibration term.

Table II.2 shows the empirical root mean squared error ($\times 10^3$) for the six estimators. The BPSP estimator has the smallest empirical root mean squared errors in all cases. The PR estimator performs as well as the BPSP estimator in the LINUP case, but is less efficient than the BPSP estimator in the EXP case. To protect again model misspecification, the GR estimators lose some efficiency compared to their corresponding model-based predictive estimators.

Table II.3 shows the noncoverage probability ($\times 10^2$) of 95% confidence/credible intervals. For the LR, PR_GR, and BPSP_GR estimators, we use both the linearization (V1) and the jackknife resampling (V2) methods to calculate the variances of estimators. Overall, the confidence coverage of credible interval for the BPSP estimator is closer to the nominal level than the other five estimators, especially when the population proportion is close to zero or one or when few observations are selected into sample in the tails. Specifically, the BPSP estimator achieves more improvement in coverage when

p is close to zero in both the LINUP and EXP cases, since little data are included in the sample from the lower tail of the two populations. Of interest is the rather poor confidence coverage of the HK, LR, PR_GR, and BPSP_GR confidence intervals, even when the sample size is considerable; the coverages are consistently below nominal levels. Note that the improved coverage of the BPSP estimator is achieved with intervals that are narrower on average than those of the HK, LR, PR_GR, and BPSP_GR estimators.

The choice of prior and hyperprior distributions in mixed models can have a big effect on inferences. We used the prior distribution $\beta_i \propto N(0,10^6)$ for the fixed effects parameters, β_i. In our simulations, we report results based on a proper inverse-gamma prior distribution for τ^2, namely $\tau^2 \propto IG(0.1,0.1)$. To assess sensitivity to the choice of prior distributions, we also computed results using $\tau^2 \propto IG(0.01,0.01)$ and $\tau^2 \propto IG(0.001,0.001)$, as well as an improper uniform prior distribution on τ (Gelman 2006). These different priors had little impact on posterior inference of the proportion of interest.

II.6. Example of tax auditing

We now compare the BPSP estimator with alternative methods on a real population involving income tax auditing data (Compumine 2007). The data set consists of 3,119 Swedish income tax returns for persons who during the year sold mutual funds managed in a foreign country. The outcome of interest *Y* is whether the income tax return is incorrect (coded as 1 for incorrect, and 0 for correct), and it is measured for all observations in this data set. We treated the 3,119 income tax returns as a finite

population here, so that the true population proportion of incorrect income tax returns is 0.517. Since the amount of the realized profit or loss is an important feature for determining whether a taxpayer reports his return of income of the sale of a foreign fund correctly, it was chosen as the size variable used in drawing pps samples. If the amount of the realised profit is negative (a negative number represents loss), it was assigned a value of 1 Swedish Krona, the minimum amount of the positive profits, where negative values are not allowed in the size variable. One thousand repeated systematic pps samples of size 300 and 600 were drawn without replacement from randomly ordered population lists. The 78 and 241 returns with largest profits were included with certainty into the samples of size 300 and 600 respectively.

Figure II.3 shows that the probability of selection has a highly right-skewed distribution for the population even after excluding the observations with selection probability of 1. We applied the same six estimators as in the simulation study with 30 knots on the pps samples, and compared their performances in terms of bias, RMSE, and average width and noncoverage rate of the 95% confidence/credible interval. For the BPSP estimator, a fixed number of 30 knots are placed at evenly spaced sample percentiles of the selection probabilities. For the GR estimators, neither the linearization nor the jackknife variance estimator has predominantly better performance than the other, we present the inference based on the linearization variance estimator for simple calculation. We report the GR estimators based on both equations (4) and (5). The results are displayed in Table II.4.

Table II.4 shows that the BPSP estimator has slightly increased bias but smaller RMSE, shorter average width and closer to the nominal level credible interval than the

design-based estimators. Results not shown here indicate that the BPSP estimator with a uniform prior distribution has slightly better performance than that with inverse-gamma prior distribution with respect to bias, RMSE, and coverage rate, because there are more fluctuations in the data and the uniform prior allows the fitted function to have more flexibility. The BPSP_GR estimator is less biased, but achieves less efficiency and worse coverage rate than the BPSP estimator. The predictive estimator using the probit linear regression model as prediction model performs poorly here since the model is misspecified, but its GR estimator does reduce bias and RMSE and improve coverage rate. The BPSP_GR estimator based on equation (4) performs very poorly in terms of RMSE compared to the estimator in equation (5), because a situation similar to that in Basu's (1971) circus elephant example occurs, where one or more observations having very low selection probabilities are selected into the sample and hence receive large weights. However, the PR_GR estimator in equation (4) performs as well as that in equation (5) with predictions obtained from the weighted maximum likelihood estimates, where selection probability is used as a covariate as well as the sample weights. Overall, the GR estimator in equation (5) is more desirable than that in equation (4). As the sample size increases from 300 to 600, the noncoverage probability of the 95% credible interval of the BPSP estimator approaches the nominal level quickly from 14% to 5%, but the coverages are consistently below the nominal level for the other estimators.

Compared to the parametric model-based predictive estimators, the BPSP estimator is robust not only to model misspecification, but also to the influential observations in the sample. To demonstrate the robustness to the influential observations, we compare the changes in the model fitting using probit p-spline models, linear probit model, and

quadratic probit model based on the pps sample only in Figure II.4, and based on the pps sample as well as the observations with selection probabilities of 1 in Figure II.5. In each figure, the population is stratified by the 100 quantiles of the probabilities of selection, and the true probabilities of $Y = 1$ are calculated and plotted with a black dot for each stratum. The grey curves are the posterior means of $\Pr(Y_i = 1 \mid \pi_i)$ from 10 random pps samples using 3000-iterate Gibbs sampler and linear spline in the left plot, using linear probit regression in the middle plot, and using quadratic probit regression in the right plot. Figure II.4 shows that the probit p-spline regression model is more flexible in catching the pattern among the observations than the parametric models. From Figure II.4 to Figure II.5, the posterior means of $\Pr(Y_i = 1 \mid \pi_i)$ do not change except for those with very large selection probabilities using the p-spline model. However, the posterior means curves change dramatically using the parametric probit regressions, especially the quadratic probit regression. These comparisons indicate that probit p-spline regression model is less likely affected by influential observations than parametric regression models, and hence is a good choice of prediction model in the model-based inference.

II.7. Discussion

Bayesian inferences based on the p-spline model outperform the Hájek estimator, the GR estimator, and parametric model-based prediction estimators in our simulations. The BPSP estimators are more efficient than the Hájek and GR estimators, and despite slightly higher empirical bias, their 95% credible intervals provide better confidence coverage and shorter average interval width, especially when the population proportion is closer to zero or one and few data are selected into the sample in the tails. This suggests

the importance of current research in estimating finite population prevalence of rare events. Compared to parametric model-based predictive estimators, the BPSP estimator achieves robustness to model misspecification and influential observations in the sample by using a flexible model, without much loss of efficiency for the sample sizes considered.

The BPSP estimators are not sensitive to two choices of prior distributions of τ^2 considered here, though it appears from the tax auditing example that the uniform prior yields slightly smaller bias and RMSE, shorter 95% credible intervals, and better coverage when a nonlinear prediction model is needed. The tax auditing example also shows that in the GR estimator, an estimated population size using the sum of inverse selection probabilities is more desirable than the true population size when one or more observations with very low selection probability are included in the sample, since the GR estimator with denominator *N* has high variance and low efficiency in this case.

The design-based estimators and their 95% confidence intervals can provide valid inferences for population proportions when the sample is large. However, these asymptotic properties do not appear to hold when the sample size is moderate or small and the true population proportion to be estimated is close to 0 or 1. The BPSP approach can provide more valid inferences for small samples, although confidence coverage appears to be less than nominal when the sample size gets small, and lack of parsimony of the model is an issue.

The choice of variance estimator is problematic for some unequal probability designs for the design-based estimators, but the Bayesian p-spline prediction approach provides a simulation approximation of the full posterior distribution of the population proportion.

Extra work is not needed to estimate the variance or 95% credible interval for the BPSP estimator, as it can be obtained simultaneously with the point estimators. In Zheng and Little (2005), three variance estimators of the p-spline model-based estimator for finite population total in a pps sample were compared, including the model-based empirical Bayes variance estimator, the jackknife variance estimate, and the balanced repeated replication (BRR) variance estimate. The simulation studies showed that the jackknife method worked well, whereas the BRR method tended to yield conservative standard errors and the model-based empirical Bayes estimator was vulnerable to misspecification of the variance structure. In the present work, the $1-\alpha$ level credible interval for the BPSP estimator of population proportion is constructed by splitting α equally between the upper and lower endpoints of the posterior distribution of p. This pure Bayesian approach based on draws from the posterior distributions seemed to work well in our setting, and avoids the heavy computation associated with the jackknife and BRR method.

The BPSP estimator we propose here can be extended to include additional auxiliary covariates by adding linear terms for these variables. For domain estimation, an interaction term between the spline of selection probabilities and the domain indicator should also be modeled. Both the additive effects of auxiliary variables and the interaction between the domain indicator and selection probabilities can be represented in a mixed model (Ruppert, Wand, and Carroll 2003, p231) and estimated using Gibbs sampling or WinBUGS (Crainiceanu, Ruppert, and Wand 2005). The BPSP estimator for finite population proportions can also be extended to a more general case of a polychotomous response. The Gibbs sampling approach for the binary case can be generalized to the case of ordered categories, and can be applied to the unordered

categories with a latent multinomial distribution (Albert and Chib 1993). Another extension for the BPSP estimator is in the small area estimation, by combing small area random effects with the smooth spline on the selection probabilities (Opsomer et al., 2008). This extension will be the focus of future research.

Finally, one reviewer questioned whether the proposed approach can be applied in a multipurpose survey with many outcomes, since the modeling procedure does not provide a single set of weights and needs to be repeated for all variables of interest. It is true that our methods are more computationally intensive than existing approaches, but the BPSP method can be easily implemented with a Gibbs sampling algorithm or using WinBUGS, so computing is not a major obstacle. We point out that the simulations in the paper involved repeating the iterative Gibbs analysis 6000 times, so an equivalent level of computation on a single survey of comparable size would allow the implementation of the BPSP method for 6000 outcomes! These were done on a garden-variety laptop PC. While we do not advocate automatic use of any analytical method, design or model-based, our point is that computational complexity is no longer a major obstacle to applying these methods. We suggest that the statistical properties of a method are more important than computing time, given modern day computing resources.

Appendix

Algorithm of Gibbs sampling

Model (2) can also be written in the matrix form,

$$\Phi^{-1}\left(E\left(y_i \mid \beta, b, X, Z\right)\right) = \left(X\beta + Zb\right)_i,\ i = 1,...,n$$

$$\beta = \left(\beta_0, \beta_1, ..., \beta_p\right)^T,\ b = \left(b_1, ..., b_m\right) \sim N_m\left(0, \tau^2 I_m\right)$$

$$X = \begin{pmatrix} 1 & \pi_1 & ... & \pi_1^p \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ 1 & \pi_n & ... & \pi_n^p \end{pmatrix},\ Z = \begin{pmatrix} \left(\pi_1 - k_1\right)_+^p & ... & \left(\pi_1 - k_m\right)_+^p \\ . & ... & . \\ . & ... & . \\ . & ... & . \\ \left(\pi_n - k_1\right)_+^p & ... & \left(\pi_n - k_m\right)_+^p \end{pmatrix}.$$

The algorithm of Gibbs sampling for estimating the parameters in Model (2) is as follows:

a) The probit regression model for the binary outcome $y = \left[y_1, ..., y_n\right]^T$ corresponds to a normal regression model for a latent continuous data $y^* = \left[y_1^*, ..., y_n^*\right]^T$, which has a truncated multivariate normal distribution with mean $\left(X\beta + Zb\right)$ and identity covariance matrix (Albert and Chib 1993), and y_i is the indicator that $y_i^* > 0$. With some initial values of $\left(\beta, b\right)$, values of the latent continuous data y_i^* can be simulated.

b) Specifying a proper flat normal prior distribution $N\left(0, 10^6\right)$ on β and an inverse gamma distribution $IG\left(0.1, 0.1\right)$ on τ^2, the posterior distribution of $\left(\beta, b, \tau^2\right)$ given the simulated latent continuous data y^* is

$$\left(\beta, b\right) \mid \tau^2, y^* \sim MVN_{m+p+1}\left(\left(C^T C + D/\tau^2\right)^{-1} C^T y^*, \left(C^T C + D/\tau^2\right)^{-1}\right)$$

$$\tau^2 \mid \beta, b \sim IG\left(0.1 + m/2, 0.1 + \left\|b\right\|^2/2\right), \qquad (6)$$

where C=[X, Z] and D is a diagonal matrix with $p+1$ values of 10^{-6} followed by m ones on the diagonal. Gelman (2006) recommended a uniform prior distribution on τ, which results in the posterior distribution for τ^2 as

$$\tau^2 \mid \beta, b \sim IG\left((m-1)/2, \|b\|^2/2\right) \tag{7}$$

c) At iteration t, draws of $\left(\beta^{(t)}, b^{(t)}, \tau^{2(t)}\right)$ from the posterior distribution in equation (6) or (7) are used to generate new latent data $\hat{y}^{*(t)}$ conditional on observed binary variable y for the sample, and to obtain the posterior predicted values $\hat{y}^{(t)}$ for non-sample units. We then can obtain draws from the posterior distribution of the finite population proportion at iteration t as

$$\hat{p}^{(t)} = N^{-1}\left(\sum_{i \in s} y_i + \sum_{j \notin s} \hat{y}_j^{(t)}\right)$$

Table II.1 Empirical bias × 10^3 of six estimators

Population	n	True prop.	HK	LR	PR	PR_GR	BPSP	BPSP_GR
LINUP	100	0.10	-0.01	13.04	10.3	1.62	8.04	1.18
		0.25	-3.04	2.36	0.90	-2.09	-0.77	-2.45
		0.5	-3.99	-2.88	-4.28	-2.97	-5.21	-3.27
		0.75	-1.77	-0.61	-3.58	-0.75	-3.83	-0.88
		0.90	-0.40	0.34	-2.46	0.27	-2.88	0.08
	200	0.10	2.49	7.92	5.82	1.49	5.07	1.43
		0.25	2.60	1.85	0.71	0.22	0.15	0.05
		0.50	3.32	-0.13	-1.34	-0.06	-1.66	-0.17
		0.75	3.80	1.15	-0.30	1.24	-0.50	1.16
		0.90	1.63	0.35	-1.01	0.32	-1.23	0.33
EXP	100	0.10	1.19	18.05	25.79	4.67	16.99	3.91
		0.25	-1.16	6.59	27.72	2.76	12.65	0.55
		0.5	-4.00	-3.48	12.52	-1.61	-1.44	-3.39
		0.75	-2.01	-0.41	-2.98	0.01	-3.43	-0.52
		0.90	-1.27	-0.15	-0.98	-0.13	-1.01	-0.24
	200	0.10	3.11	11.03	22.05	3.48	13.38	2.73
		0.25	3.22	4.04	27.58	2.63	9.64	0.95
		0.50	3.77	-0.56	14.01	0.44	0.01	-0.72
		0.75	4.22	0.02	-2.86	0.07	-2.37	0.04
		0.90	2.30	0.12	-0.69	0.13	-0.68	0.02

Table II.2 Empirical RMSE × 10^3 of six estimators

Population	n	True prop.	HK	LR	PR	PR_GR	BPSP	BPSP_GR
LINUP	100	0.10	55	57	46	51	47	52
		0.25	71	62	54	59	55	60
		0.5	65	51	47	50	48	50
		0.75	46	36	36	36	36	36
		0.90	26	23	23	23	23	23
	200	0.10	39	41	32	36	32	36
		0.25	48	43	35	39	36	40
		0.50	46	36	33	34	33	35
		0.75	31	24	24	24	24	24
		0.90	18	15	16	15	16	15
EXP	100	0.10	51	60	54	52	52	52
		0.25	67	65	59	61	57	61
		0.5	66	56	43	53	47	52
		0.75	44	23	22	22	23	22
		0.90	24	12	12	12	12	12
	200	0.10	36	42	40	36	36	36
		0.25	47	49	45	43	41	43
		0.50	45	39	31	36	32	35
		0.75	29	15	14	15	15	14
		0.90	16	8	8	8	8	8

Table II.3 Average length of 95% CI × 10^2 of six estimators

Population	n	True prop.	HK	LR	PR		PR_GR		BPSP		BPSP_GR	
				V2	V3	V4	V1	V2	V3	V4	V1	V2
LINUP	100	0.10	19	21	16	15	16	20	15	15	17	21
		0.25	26	22	19	19	20	23	19	19	21	23
		0.5	25	19	18	18	19	20	18	18	19	20
		0.75	17	13	14	13	13	14	14	14	13	14
		0.90	10	8	9	9	9	9	9	9	9	9
	200	0.10	15	15	12	11	13	14	12	11	13	15
		0.25	19	16	14	14	15	16	14	14	15	16
		0.50	18	13	13	12	13	14	13	13	13	14
		0.75	12	9	9	9	9	10	9	9	9	10
		0.90	7	6	6	6	6	6	6	6	6	6
EXP	100	0.10	18	21	16	16	16	19	16	16	16	20
		0.25	25	23	18	18	21	23	19	19	21	25
		0.5	25	19	14	14	19	21	17	16	18	20
		0.75	17	8	8	8	8	9	8	8	8	9
		0.90	9	4	4	4	4	5	5	4	4	5
	200	0.10	13	15	12	12	12	14	12	12	13	14
		0.25	18	17	13	13	16	17	14	14	16	18
		0.5	18	13	10	9	14	14	12	12	13	14
		0.75	11	5	5	5	6	6	6	6	6	6
		0.90	6	3	3	3	3	3	3	3	3	3

* V1: variance estimator using linearization; V2: jackknife variance estimator
V3: posterior variance using equal tails; V4: highest probability density posterior variance

Table II.4 Noncoverage rate of 95% CI × 10^2 of six estimators

Population	n	True prop.	HK	LR	PR		PR_GR		BPSP		BPSP_GR	
				V2	V3	V4	V1	V2	V3	V4	V1	V2
LINUP	100	0.10	16.2	18.0	8.4	10.3	20.9	16.1	9.0	11.1	18.4	14.2
		0.25	9.1	12.5	7.4	9.4	12.9	10.2	7.3	8.6	12.2	9.6
		0.5	7.5	9.4	5.0	6.0	7.2	7.6	4.4	4.8	7.3	7.1
		0.75	7.6	11.8	5.5	5.2	7.3	9.9	5.5	5.5	7.8	8.8
		0.90	7.4	11.4	5.7	6.5	8.0	9.4	5.4	5.9	8.4	7.1
	200	0.10	10.8	12.6	6.4	8.1	13.9	10.9	6.2	7.8	12.6	9.4
		0.25	6.3	10.2	5.0	5.6	8.2	6.7	5.2	6.0	7.3	6.7
		0.50	5.5	8.3	5.5	6.4	6.2	5.9	5.1	5.7	6.0	5.5
		0.75	5.8	9.3	5.7	5.9	7.4	6.2	5.6	5.7	7.0	6.2
		0.90	6.0	8.4	4.4	5.4	6.1	4.4	4.7	5.0	6.3	5.5
EXP	100	0.10	15.0	18.1	10.5	11.4	19.4	14.8	9.2	11.7	18.4	14.4
		0.25	9.6	15.1	11.1	12.7	12.0	10.9	8.1	10.6	12.9	10.7
		0.5	7.4	13.5	12.2	13.3	9.0	11.4	8.9	9.2	10.2	8.4
		0.75	6.2	10.7	9.6	9.4	7.5	8.4	8.2	8.4	8.2	7.2
		0.90	6.1	10.5	7.9	8.7	9.9	7.6	7.0	8.1	9.8	7.2
	200	0.10	10.8	13.3	9.9	11.0	12.5	11.7	7.5	8.9	12.4	9.4
		0.25	6.4	13.6	14.8	15.9	9.7	8.7	7.3	7.9	11.2	7.8
		0.5	6.0	11.5	14.3	15.2	7.2	8.5	6.2	6.4	7.5	6.9
		0.75	5.4	7.4	5.6	6.1	5.1	4.2	4.5	4.5	5.2	3.9
		0.90	5.5	8.8	5.5	5.7	6.8	4.6	5.5	5.8	6.6	3.7

* V1: variance estimator using linearization; V2: jackknife variance estimator
V3: posterior variance using equal tails; V4: highest probability density posterior variance

Table II.5 Comparison of various estimators for empirical bias, root mean squared error, and average width and noncoverage rate of 95% CI, in the tax return example

Methods	bias*100		RMSE*100		average width*100		noncoverage*100	
	300	600	300	600	300	600	300	600
HK	-2.4	-1.8	12.4	10.2	36	29	14.1	10.2
LR	6.7	5.5	11.9	9.2	27	21	43.5	45.6
PR	-11.6	-10.1	12.4	10.6	18	14	69.8	83.4
PR_GR1	-1.2	-0.4	11.5	8.7	31	25	22.4	16.8
PR_GR2	-1.2	-0.3	11.5	8.8	33	26	16.1	11.4
BPSP	-6.8	-2.7	9.3	5.2	27	19	14.2	5.0
BPSP_GR1	-3.0	-0.5	102.6	56.9	77	57	14.4	9.2
BPSP_GR2	-0.7	0.2	12.0	10.1	34	26	15.9	12.8

* GR_1: GR estimators using equation (4);
GR_2: GR estimators using equation (5).

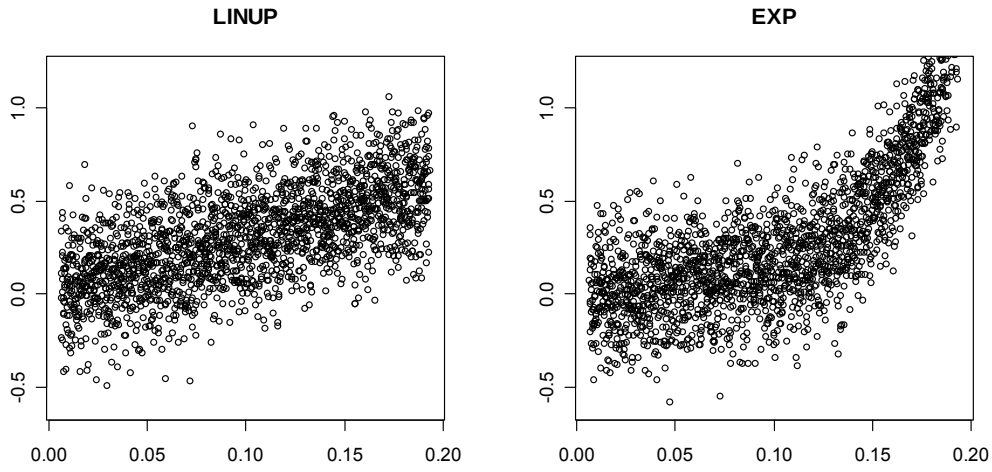

Figure II.1 Two simulated artificial populations (N=2,000) X-axis: selection probability; Y-axis: Z

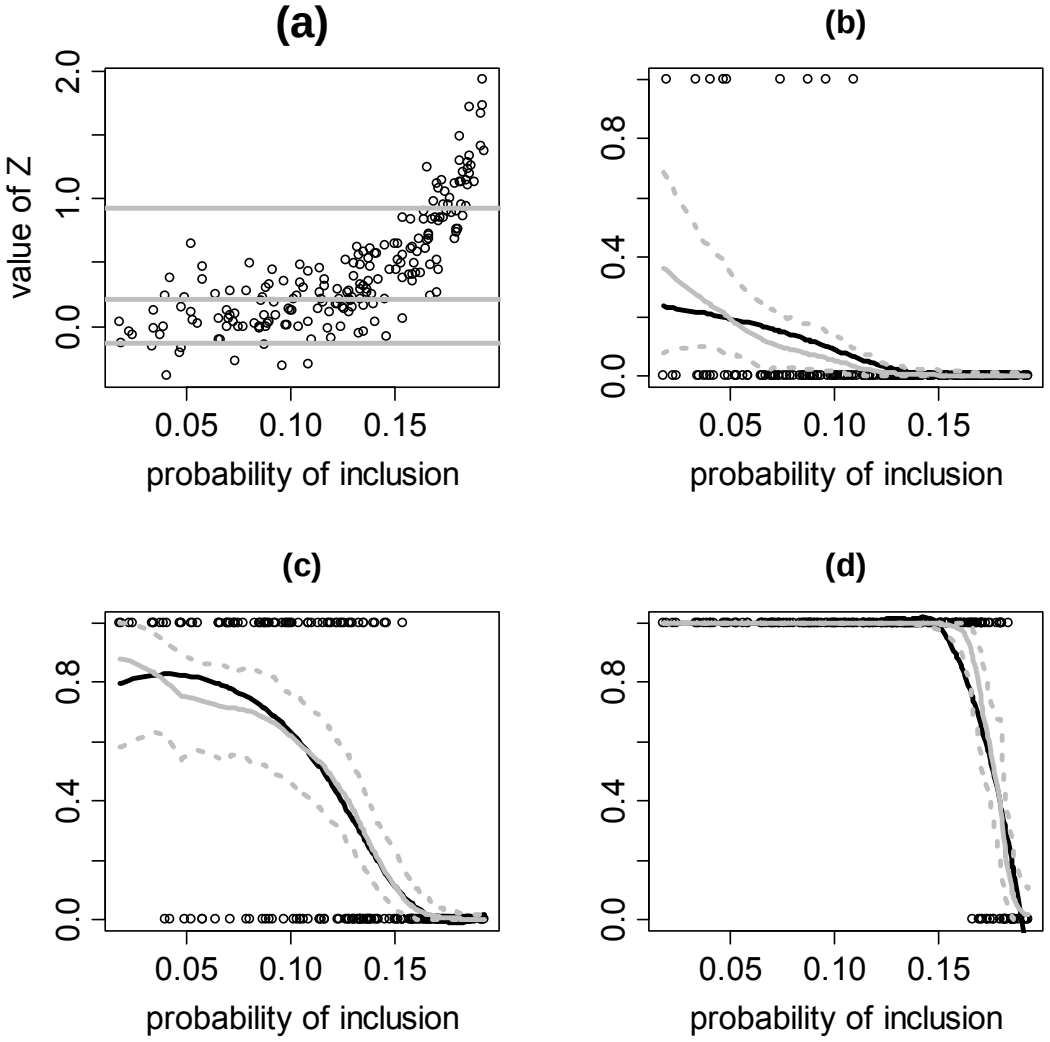

Figure II.2 A random pps sample from the EXP case (n=200, N=2000): (a) scatter plot of *Z*; the three grey lines are the superpopulation 10^{th}, 50^{th}, and 90^{th} percentiles, respectively. (b) black circles are observed units of binary survey variable *Y* in the sample, defined as $Y = I$ ($Z \leq 10^{th}$ percentile); the grey solid and dashed curves are posterior means of $Pr(Y_i=1|\pi_i)$ and 95% credible intervals, respectively, simulated based on a probit p-spline model on π; and the black curve is the superpopulation $Pr(Y_i=1|\pi_i)$. (c) similar to (b), but with $Y = I$ ($Z \leq 50^{th}$ percentile). (d) similar to (b), but with $Y = I$ ($Z \leq 90^{th}$ percentile).

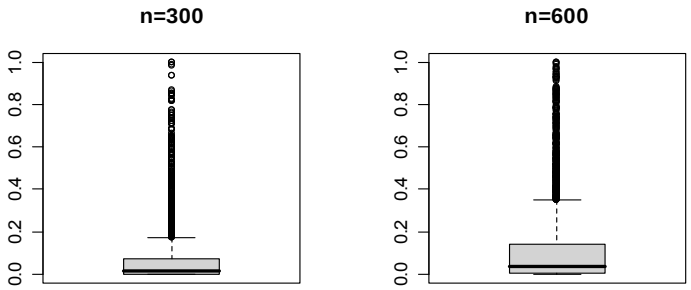

Figure II.3 Box plots of the probabilities of selection for two sample sizes in the tax auditing example

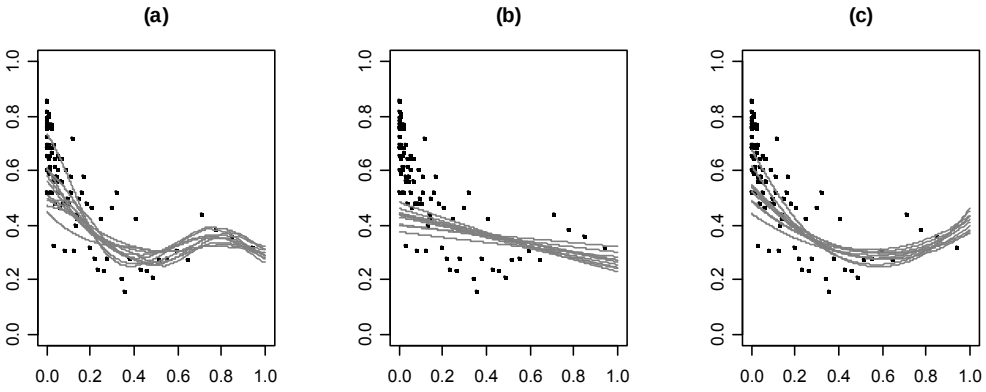

Figure II.4 Predictions based on pps samples only in the tax auditing example, *X*-axis: selection probabilities π, *Y*-axis: $P(Y=1|\pi)$; black dots are the true $P(Y=1|\pi)$ within each percentile of π; grey curves are ten realizations of the posterior means of $P(Y=1|\pi)$. The prediction models are (a) probit linear p-spline regression, (b) linear probit regression, (c) quadratic probit regression.

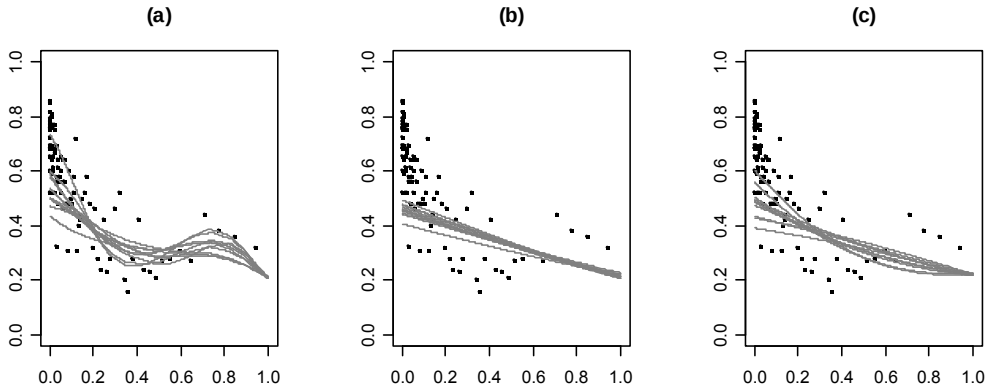

Figure II.5 Predictions based on the combined data of pps samples and the observations sampled with certainty in the tax auditing example, *X*-axis: selection probabilities π, *Y*-axis: $P(Y=1|\pi)$; black dots are the true $P(Y=1|\pi)$ within each percentile of π; grey curves are ten realizations of the posterior mean of $P(Y=1|\pi)$. The prediction models are (a) probit linear p-spline regression, (b) linear probit regression, (c) quadratic probit regression.

Chapter III

BAYESIAN INFERENCE OF FINITE POPULATION QUANTILES FROM UNEQUAL PROBABILITY SAMPLES

III.1. Introduction

In survey practice, it is often of interest to estimate quantiles for continuous finite population characteristics. The finite-population quantiles are often estimated using the sample-weighted quantiles in the traditional survey sample approach. These estimators are design-based estimators. In sample surveys it is common that the design variable or a correlated auxiliary variable is measured along with the survey variables of interest. When the design variable or the auxiliary variable is measured on the non-sampled units, this information should be incorporated to construct an estimator more efficient than the sample-weighted estimators.

The use of auxiliary information for estimating finite-population distribution functions has been extensively studied. Chambers and Dunstan (CD) (1986) proposed a model-based method which allows the use of auxiliary information to improve the estimation of finite-population distribution functions. Rao, Kovar, and Mantel (1990) proposed design-based ratio and difference estimators of finite-population distribution functions and demonstrated the advantage of the ratio and difference estimators over the CD estimator for large samples under model misspecification. Wang and Dorfman (1996) suggested a weighted average of the CD and the Rao's estimators. Kuk and Welsh (2001) modified the CD estimator to fine-tune for departure from the model assumed by

estimating the conditional distribution of residuals as a function of the auxiliary variable. In Kuk (1993), a kernel-based estimator was proposed, which combines the known distribution of the auxiliary variable with a kernel estimate of the conditional distribution of the survey variable given the value of the auxiliary variable. In Chambers, Dorfman, and Wehrly (1993), a model-based estimator using kernel smoothing was proposed to estimate distribution functions, with the estimator calibrated for its bias under the model in Chambers and Dunstan (1986). Wu and Sitter (2001) also proposed a model-calibration estimator of finite population distribution functions.

Compared to distribution functions, the research on finite population quantiles in using auxiliary information is limited. Chambers and Dunstan (1986) discussed the estimation of α -quantile by inverting the CD estimator of the distribution function. However, heavy computation is involved in this procedure, so that they did not compare the performances of this quantile estimator in their simulation study. Moreover, Chambers and Dunstan assumed a superpopulation model for Y that corresponds to a regression through the origin with some specified heteroscedastic errors, $Y_i = \beta x_i + v(x_i)U_i$. When the association between Y and the auxiliary variable is more complicated than this assumed model, the CD estimator can be very biased (Rao, Kovar, and Mantel 1990).

In this paper, we assume that the selection probabilities in an unequal probability sampling are known for all the units in the population. We develop two robust Bayesian model-based estimators of finite population quantiles by incorporating the selection probabilities at the estimation stage. The first method is to estimate the distribution functions evaluated at a number of sample values using Bayesian Penalized Spline

Predictive estimators (Chen, Elliott, and Little 2007). The finite population quantiles are then estimated by inverting the predictive distribution function. We also propose a Bayesian two moment p-spline predictive (B2PSP) estimator for quantiles, by predicting values of nonsampled units based on a normal model with the mean and the variance both modeled using penalized splines on the selection probabilities. We use a simulation study to compare the performance of these two new methods with the sample-weighted estimator.

III.2. Estimators of the quantiles

Suppose that there is a finite population consisting of *N* identifiable units. Let π_i denote the probability of selection for unit *i*, which is assumed to be known for all units in the finite population before a sample is drawn. Let *s* denote an unequal probability random sample with a sample size of *n*, which is drawn from the finite population according to the selection probabilities $\{\pi_1, \pi_2, \ldots, \pi_N\}$. Let *Y* denote the continuous survey variable, with $\{y_1, y_2, \ldots, y_n\}$ observed in the random sample *s*.

The finite-population α-quantile of *Y* is given by $\theta(\alpha) = \inf\left\{t; N^{-1}\sum_{i=1}^{N}\Delta(t - Y_i) \geq \alpha\right\}$.

The finite population α-quantile is often estimated using the sample weighted α-quantile $\hat{\theta}(\alpha)$. Woodruff (1952) proposed a method of calculating confidence limits for the sample weighted α-quantile: a pseudo-population is obtained by weighting each sample item by its proper weight; the standard deviation of the percentage of items less than the estimated α-quantile is estimated, and then the estimated standard deviation is added to and subtracted from α to construct the confidence limits for the percentage of

items less than the estimated α-quantile; and the values of the survey variable corresponding to the confidence limits of the percentage of items less than the estimated α-quantile are read-off the pseudo-population arrayed in order of size. Sitter and Wu (2001) showed that the Woodruff intervals perform very well even in the moderate to extreme tail regions of the distribution function. Alternative variance estimation for finite population quantiles was derived by Francisco and Fuller (1991) using a smoothed version of the large-sample test inversion.

III.2.1. Invert-CDF Bayesian model-based approach

In probability theory, a quantile function is the inverse of its cumulative distribution function. We can estimate the finite population quantiles by first building a continuous and strictly monotonic predictive finite population distribution function. The finite population distribution function evaluated at *t* is defined as $F(t) = N^{-1}\sum_{i=1}^{N}\Delta(t - y_i)$, where $\Delta(x) = 1$ when $x \geq 0$ and $\Delta(x) = 0$ elsewhere. By treating $\Delta(t - y)$ as a binary outcome variable, we can apply the methods of estimating finite population proportions to estimate this finite population distribution function.

Chen, Elliott, and Little (2007) provided a Bayesian penalized spline predictive (BPSP) estimator for finite population proportions in unequal probability sampling. They first fitted a probit penalized spline regression model (1) with *m* pre-selected fixed knots between a binary survey variable *z* and the selection probabilities in the sample:

$$\Phi^{-1}\left(E(z_i \mid \beta, b, \pi_i)\right) = \beta_0 + \sum_{k=1}^{p}\beta_k \pi_i^k + \sum_{l=1}^{m} b_l (\pi_i - k_l)_+^p \qquad (1)$$

$$b_l \sim N(0, \tau^2)$$

They then simulated the posterior predictive distribution of *z* among the non-sampled units. The posterior distribution of the finite population proportion is obtained by generating a large number of draws of the average of the observed sample units and the predictive non-sample units. They showed through simulation studies that the BPSP estimator is more efficient with confidence coverage closer to nominal levels than the sample-weighted estimators in estimating finite population proportions in unequal probability sampling.

We employ the BPSP approach *n* times to estimate $F(t)$, $t = \{y_1, y_2, \ldots, y_n\}$. The BPSP estimator of $F(t)$ is based on a pointwise argument which does not take into account the fact that we are estimating a whole distribution function. As a result, the monotonic property of a cumulative distribution function does not necessarily hold here. In addition, linear interpolation of the *n* estimated distribution functions may lead to a rough predictive cumulative distribution function. To overcome these two problems, we fit a smooth cubic regression curve on the *n* estimated distribution functions with monotonicity constraints (Wood 1994). We denote the estimated distribution function as $\hat{F}(t)$. We also fit another two monotonic smooth cubic regression curves on the upper and lower limits of the 95% credible intervals of these estimated distribution functions, denoted as $\hat{F}_U(t)$ and $\hat{F}_L(t)$. To reduce computation time, we only estimate the distribution functions for $k < n$ times on *k* pre-selected sample points in our simulation studies.

The basic principle behind the invert-CDF Bayesian approach can best be explained graphically in Figure III.1. Let us assume that a sample of size 100 is drawn from a finite population. We pick 20 observations from the sample and estimate their corresponding

distribution functions and associated 95% credible intervals using the BPSP estimator. In (a), we plot the BPSP estimates of these 20 points with black dots and the upper and lower limits of 95% CI with "-" signs and connect the upper and lower limits with solid lines. In (b), we add three monotonic smooth predictive curves using black solid curve for the point estimate and black dash curves for the upper and lower limits of 95% CI.

We can easily estimate the finite population α-quantile by inverting the predictive monotonic smooth cumulative distribution function. Let x denote the estimated α-quantile θ. The 95% CI of an estimated quantile is then calculated in two ways. The first method is the extension of the Woodruff's method to the Bayesian setting (1952). This can be illustrated graphically in Figure III.2(a). Let us draw two horizontal red dashed lines across the graph with the upper (A) and lower (B) limits of the BPSP estimator for $F(x)$ as the y-axis values. We read x_A and x_B from the x-axis which correspond to $\hat{F}(x_A) = A$ and $\hat{F}(x_B) = B$, the posterior means or medians of $F(x_A)$ and $F(x_B)$. According to the definition of the 95% credible interval for $F(x)$, the probability that $F(x)$ falls between A and B given the sample is exactly 0.95. If we take a $\hat{F}^{-1}(\cdot)$ transformation on $F(x)$, it is immediately apparent that the posterior probability that $\hat{F}^{-1}(F(x))$ falls between $\hat{F}^{-1}(A) = x_A$ and $\hat{F}^{-1}(B) = x_B$ is also 0.95 because of the monotonic property of $\hat{F}(\cdot)$. With $\theta = F^{-1}(\hat{F}(x))$ ($F(\theta) = \alpha$ and $\hat{F}(x) = \alpha$) and assuming that the posterior mean/median $\hat{F}(\cdot)$ is a good estimate of $F(\cdot)$, we can conclude that an approximate 95% credible interval for θ is within the limits x_A and x_B. The performance of the Woodruff's CI relies on how well $\hat{F}(\cdot)$ perform in estimating $F(\cdot)$.

The second method is illustrated in Figure III.2(b). We draw a horizontal line across the graph with α as the y-axis value. We read x'_A and x'_B from the x-axis such that $\hat{F}_L(x'_A) = \alpha$ and $\hat{F}_U(x'_B) = \alpha$, respectively. If the 95% credible interval of the distribution function $F(\cdot)$ is formed by equally splitting the tail area in the posterior distribution, the interval formed by x'_A and x'_B is another 95% credible interval of θ. The proof is as follows: Because α is the lower limit of the 95% credible interval of $F(x'_A)$, only 2.5 percents of the draws of $F(x'_A)$ in the posterior distribution are smaller than α. That is, the probability that $F^{-1}(\alpha)$ is greater than $F^{-1}(F(x'_A))$ is equal to 0.025 given the observed data, or θ is greater than x'_A with only a 2.5 percent chance. Similarly, because α is the upper limit of the 95% credible interval of $F(x'_B)$, we conclude that the probability that θ is smaller than x'_B given the data is 0.025. Therefore, there is 95% probability that θ is within x'_A and x'_B in the posterior distribution, given the sample.

This invert-CDF Bayesian model-based approach does not depend on any strong modeling assumption. It can be applied to normal or skewed distributions as long as the selection probabilities are known for all the units in the finite population. However, the limitation of this approach is the heavy computation associated with the estimation of the distribution functions. There is a trade-off between precision and computation time here. When we estimate the distribution functions *n* times at all sample units, we have best used the sample information but this requires intensive computation. As we pick *k* of the *n* sample units and estimate only the *k* distribution functions, we lose information, but need less computation time. A simulation study not shown here indicates that the distribution function curve estimated based on a well selected subset of the *k* sample units

is similar to the curve estimated based on all sample units but the computation time is significantly reduced.

III.2.2. Bayesian two-moment penalized spline predictive approach

We consider an alternative straightforward model-based estimator of the finite population quantiles defined as:

$$\hat{\theta}(\alpha) = \inf\left\{t; N^{-1}\left(\sum_{i\in s}\Delta(t-y_i)+\sum_{j\notin s}\Delta\left(t-\hat{y}_j\right)\right)\geq\alpha;\right\}, \tag{2}$$

where $\hat{y}_j$ is the predicted value of the j^{th} observation in the non-sample units based on some statistical model. For a continuous survey variable, the most commonly used prediction model is the normal linear regression model

$$Y_i \overset{ind}{\sim} N\left(\beta_0+\beta_1\pi_i, c_i\sigma^2\right). \tag{3}$$

Model (3) assumes a strong linear association between the survey variable and the selection probabilities. When this linear association is not true, Model (3) will lead to a very biased and inefficient estimator of θ. For the estimation of finite population totals, Zheng and Little (2003, 2005) replaced the parametric mean function on the selection probabilities with a penalized spline and assumed c_i was equal to π_i^{2k} with some known value of *k*. They showed by simulation that their model-based estimator of finite population totals outperforms the design-based estimators despite assuming the wrong variance structure.

However, correctly specifying the variance structure is as important as the mean structure for quantile estimation. Ignoring heteroscedasticity may lead to incorrect inferences. Therefore, we extend the penalized spline model in Zheng and Little (2003)

to a two-moment model by modeling both the mean and the variance nonparametrically. The two-moment penalized spline model can be written as (Ruppert, Wand, and Carroll 2003, p.264):

$$\begin{cases} Y_i \overset{ind}{\sim} N\left(SPL_1(\pi_i,k), \exp\left(SPL_2(\pi_i,k')\right)\right) \\ SPL_1(\pi_i,k) = \beta_0 + \sum_{k=1}^{p_1}\beta_k \pi_i^k + \sum_{l=1}^{m_1} b_l (\pi_i - k_l)_+^{p_1}, \quad b \overset{iid}{\sim} N\left(0,\tau_b^2\right) . \\ SPL_2(\pi_i,k') = \alpha_0 + \sum_{k=1}^{p_2}\alpha_k \pi_i^k + \sum_{l=1}^{m_2} \nu_l (\pi_i - k_l')_+^{p_2}, \quad \nu \overset{iid}{\sim} N\left(0,\tau_\nu^2\right) \end{cases} \tag{4}$$

In Model (4), the continuous survey variable is assumed to follow a normal distribution, given the selection probabilities. The mean is modeled as a first penalized spline (SPL_1) on the selection probabilities and the variance is modeled as an exponential of a second spline (SPL_2). The exponential transformation is applied on the second spline to guarantee the positive estimate of the variance. Without losing generality, we allow the different orders of the polynomial splines (p_1, p_2) and the variate number and location of knots (k, k') for the two splines.

Ruppert, Wand, and Carroll (2003) suggested an iterative frequentist approach to estimate the parameters in Model (4). They first assumed that SPL_2 was known and fit a linear mixed model to obtain the parameter estimates in SPL_1. They calculated the square of the difference between Y and SPL_1, which followed a Gamma distribution with the shape parameter as ½ and the scale parameter of $2SPL_2$. And then they fit a generalized linear mixed model of the squared difference to obtain the parameter estimates in SPL_2. They iterated the above procedures until the parameter estimates converged. This iterative frequentist approach is simple to implement. However, our goal here in using Model (4) is not to estimate the parameters but to obtain the predictive estimates of Y in

the non-sample units so that we can use Formula (2) to estimate the quantiles. In using the frequentist approach, we face the challenges of obtaining the predictive values of *Y* in the non-sampled units as well as of estimating the 95% CI for the quantiles.

In this paper, we extend the iterative frequentist approach to a *Gibbs-Bootstrap* iterative approach. At the beginning of the iteration, we get the initial values of the two splines as $\hat{f}^{(0)} = \hat{SPL}_1$ and $\hat{g}^{(0)} = \hat{SPL}_2$. For the *Gibbs* step, we pretend that $\exp\left(\hat{g}^{(0)}\right)$ is the actual variance, so that Model (4) is simplified as a first-moment model with the variance known. We then use Gibbs sampling to obtain a draw from the posterior predictive distributions of β and τ_b^2 and a draw of $\hat{f}^{(1)}$ by assuming the prior distribution $\beta \propto N\left(0,10^6\right)$ and the hyperprior distribution $\tau_b^2 \propto IG\left(10^{-6},10^{-6}\right)$. Given $\hat{f}^{(1)}$, we proceed to the *Bootstrap* step – we calculate the squared error $\hat{r}^2 = \left(y - \hat{f}^{(1)}\right)^2$. To properly account for the modeling uncertainty in this step, we draw a Bootstrap sample from the unequal probability random sample *s* (Holmberg 1998); we then fit a generalized smooth regression model of the squared errors in the Bootstrapped sample using a frequentist approach and obtain the fitted function $\hat{g}^{(1)}$. After the *Gibbs-Bootstrap* step, we obtain a draw of $\hat{y}_j^{(1)}$ for the non-sampled units from Model (4) by plugging in $\hat{f}^{(1)}$ and $\hat{g}^{(1)}$ and a draw of $\hat{\theta}^{(1)}$ defined in formula (2). We iterate the above steps to simulate the posterior distributions of the finite-population quantiles. The average or the median of the draws of $\hat{\theta}^{(\cdot)}$ simulates the Bayesian two-moment penalized spline prediction (B2PSP) estimator $\hat{\theta}_{B2PSP}$.

Instead of the Gibbs-Bootstrap approach, we could also use WinBUGS to implement Model (4) and obtain the posterior distribution of the quantiles defined in formula (2). However, Crainiceanu et al. (2007) stated that "Our implementation of the MCMC using multivariate Metropolis-Hastings steps proved to be unstable with poor mixing properties". They suggested adding error terms to the second spline to make computations feasible by replacing sampling from complex full conditionals by simple univariate Metropolis-Hastings steps. This idea can be expressed as

$$Y_i \overset{ind}{\sim} N\left(SPL_1(\pi_i,k),\sigma_\varepsilon^2(\pi_i)\right),\ \log\left(\sigma_\varepsilon^2(\pi_i)\right)\overset{iid}{\sim} N\left(SPL_2(\pi_i,k'),0.01\right).$$

In the simulation studies, we only show the results of the B2PSP estimators for quantiles using the Gibbs-Bootstrap iterative approach and use the other approaches in the sensitivity analysis.

III.3 Simulation study

III.3.1. Design of the simulation study

Simulation studies are conducted to compare the performance of the following three estimators of finite population quantiles using systematic probability-proportional-to-size (pps) sampling design:

(1) $\hat{\theta}(\alpha)$, sample-weighted estimator;

(2) $\hat{\theta}_{inv-CDF}(\alpha)$, invert-CDF Bayesian model-based estimator;

(3) $\hat{\theta}_{B2PSP}(\alpha)$, Bayesian two-moment penalized spline predictive estimator.

We simulate an artificial finite population (N = 2,000) from a super-population (M = 20,000). The size variable *x* in the super-population takes 20,000 consecutive integer

values from 710 to 20,709. The finite population is selected from the super-population using systematic pps sampling with the probability proportional to the inverse of the size variable. We consider the following six finite populations.

(1) LINUP + homogeneity: $Y_i \overset{iid}{\sim} N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$

(2) EXP + homogeneity: $Y_i \overset{iid}{\sim} N\left(\exp\left(\beta_0 + \beta_1 x_i\right), \sigma^2\right)$

(3) LINUP + heterogeneity: $Y_i \overset{iid}{\sim} N\left(\beta_0 + \beta_1 x_i, \sqrt{x_i}\sigma^2\right)$

(4) EXP + heterogeneity: $Y_i \overset{iid}{\sim} N\left(\exp\left(\beta_0 + \beta_1 x_i\right), \sqrt{x_i}\sigma^2\right)$

(5) LINUP + mixture model: $Z_i \sim Binomial\left(M, p = 0.7\right)$

$$Y_i \mid (Z_i = 1) \overset{iid}{\sim} N\left(\beta_0 + \beta_1 x_i, \sqrt{x_i}\sigma^2\right)$$

$$\log\left(Y_i \mid (Z_i = 0)\right) \overset{iid}{\sim} N\left(\beta_0 + \beta_1 x_i, \nu^2\right)$$

(6) EXP + mixture model: $Z_i \sim Binomial\left(M, p = 0.7\right)$

$$Y_i \mid (Z_i = 1) \overset{iid}{\sim} N\left(\exp\left(\beta_0 + \beta_1 x_i\right), \sqrt{x_i}\sigma^2\right)$$

$$\log\left(Y_i \mid (Z_i = 0)\right) \overset{iid}{\sim} N\left(\exp\left(\beta_0 + \beta_1 x_i\right), \nu^2\right).$$

We draw a sample of size 100 from each finite population using systematic pps sampling with *x* as the size variable, thus the selection probabilities are $\pi_i = nx_i \Big/ \sum_{j=1}^{N} x_j$. The scatter plots of *Y* versus π for these six populations are displayed in Figure III.3.

One thousand replicates of the simulations are obtained and the three estimators are compared in terms of empirical bias, root mean squared error (RMSE), average width and non-coverage rate of the 95% confidence/credible interval. In each replicate of

simulation, a finite-population is generated before a sample is drawn. For the 95% CI, we use Woodruff's method for the sample-weighted estimator, the two methods illustrated in Figure III.2 for the invert-CDF Bayesian model-based estimator, and the 95% posterior probability of the quantile with equal tails for the B2PSP method. We use cubic splines with 15 equally spaced knots for both the probit spline in the invert-CDF method and the two-moment linear spline in the B2PSP method. In using the invert-CDF method, we estimate the distribution functions upon 20 observations (the 3 smallest, the 3 largest, and the other 14 equally spaced points in the middle from the ordered sample) and apply logit transformation on the 20 estimated distribution functions to achieve better fit of the monotonic smooth spline.

III.3.2. Simulation results

Tables III.1 shows the simulation results for $\hat{\theta}(\alpha)$, $\hat{\theta}_{inv-CDF}(\alpha)$, and $\hat{\theta}_{B2PSP}(\alpha)$ in estimating the finite-population 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles when the survey variable follows a normal distribution with homogeneous errors. The empirical bias is similar among the three estimators. Both of the Bayesian model-based approaches yield smaller root mean squared errors and shorter average 95% CI width than the sample-weighted estimator, especially with the B2PSP estimator. The coverage rate of the 95% CI is similar among the three estimators, except when α is equal to 0.1. When α is equal to 0.1, the 95% CI of the B2PSP estimator has the shortest average width and the best coverage, while the sample-weighted estimator has serious under-coverage. This happens because the Woodruff method for the sample-weighted estimator is based on a

large sample assumption but here with pps sampling only a small fraction of data is sampled in the lower tail.

Table III.2 is similar to Table III.1 but studies the scenario of heteroscedastic errors. The three estimators are comparable in empirical bias. Compared to the sample-weighted estimator, both $\hat{\theta}_{inv-CDF}(\alpha)$ and $\hat{\theta}_{B2PSP}(\alpha)$ have smaller root mean squared errors. As α increases from 0.1 to 0.9, the gain in efficiency becomes more and more substantial using the two Bayesian approaches. For instance, when α is equal to 0.75 or 0.9, the root mean squared error for the sample-weighted estimator is about twice of that for the B2PSP estimator in both the LINUP and EXP cases. When α is equal to 0.1 or 0.25, the invert-CDF Bayesian approach yield the shortest average width and very close to nominal level confidence coverage, but the confidence coverage is too low using the sample-weighted estimator or too high using the B2PSP estimator. When α is equal to 0.5, 0.75, or 0.9, the confidence coverage of the sample-weighted estimator is close to the nominal level but the confidence interval is very wide. Both the two Bayesian approaches lead to shorter average 95% credible intervals, but the confidence coverage for the B2PSP is over-estimated.

The validity of the B2PSP estimator relies on the normality assumption of the survey outcome conditioning on selection probabilities. When the normality assumption does not hold, the B2PSP estimator may no longer have the advantages we see in Tables III.1 and III.2. The invert-CDF approach does not assume normality, so we are interested in comparing the sample-weighted and the invert-CDF Bayesian estimators in the scenario of non-normal data. Their comparisons are displayed in Table III.3. Overall, the empirical bias is similar between the two estimators, and the invert-CDF approach yields

smaller root mean squared errors. The 95% CI for the invert-CDF estimator using the method illustrated in Figure III.1(b) has similar or shorter width of confidence intervals and similar confidence coverage compared to the sample-weighted estimator, except when α is 0.1 or 0.9. When α is equal to 0.1, the sample-weighted estimator has low confidence coverage as we see in the other scenarios. When α is equal to 0.9, the confidence coverage for the invert-CDF approach is higher than the nominal level. The invert-CDF estimator using Woodruff's CI leads to the under-estimate of confidence coverage, as we discussed before the performance of the Woodruff's CI is dependent on the point estimate of the distribution functions here.

Finally, we consider the conditional behavior of estimates by studying the variation in the bias generated by each estimator as the sample mean for the selection probability increases. We use the technique in Royall & Cumberland (1981): the estimates from the 1,000 samples are ordered according to the sample mean of the selection probabilities and are split into 20 groups of 50 each, and then the empirical bias is calculated for each group. Figure III.4 displays the conditional bias of various estimators of the 90th percentile for the "EXP + homogeneity" case. Figure III.4 shows that there is a linear trend for the bias of $\hat{\theta}(\alpha)$ as the sample mean of the selection probabilities increases, while the grouped bias of $\hat{\theta}_{inv-CDF}(\alpha)$ and $\hat{\theta}_{B2PSP}(\alpha)$ is less affected by the sample mean of selection probabilities, although Table III.1 shows that the overall bias is similar among the three estimators. Similar findings are also seen in other scenarios.

III.4. Discussion

Sample-weighted estimators for finite population quantiles are widely used in survey practice. Although the sample-weighted estimators with Woodruff's confidence intervals can provide valid large-sample inferences, they may be inefficient and confidence coverages can be poor in small-to-moderate-sized samples. Model-based estimators can improve the efficiency of the estimates when the model is correctly specified. For the quantile estimation of a continuous survey variable, we can either estimate the model-based distribution functions then invert the distribution function to obtain quantiles or model the survey outcome on the selection probabilities directly.

We show by simulations that Bayesian inferences of finite population quantiles based on inverting model-based predictive distribution functions or the Bayesian two-moment penalized spline regression outperform the sample-weighted estimator. Though the two Bayesian model-based estimators and the sample-weighted estimator have comparable overall empirical bias, there is a linear trend in the variation of bias for the sample-weighted estimator as the sample mean of selection probabilities increases. Both new methods yield smaller root mean squared errors than the sample-weighted estimators. In some scenarios, the improvement in efficiency using the two Bayesian methods is very significant. When the normality assumption of the survey outcome given the selection probabilities is correct, the B2PSP estimator has smaller root squared mean errors than the invert-CDF approach. The Woodruff's method performs well when a large fraction of the data is selected from the finite population. However, when data from the population is sparse, the Woodruff's method tends to under-estimate the confidence coverage. On the other hand, the 95% CI calculated from the posterior distribution of the quantiles associated with the B2PSP estimator is more likely to have higher than the

nominal level confidence coverage. Interestingly, the conservative confidence intervals are associated with shorter intervals compared to the sample-weighted estimator in most of cases. The invert-CDF approach generally performs well for confidence coverage in most cases.

Although both the invert-CDF and the B2PSP estimators outperform the sample-weighted estimator, they have various advantages and disadvantage. The B2PSP estimator requires a strong normality assumption. It yields smaller root mean squared errors than the invert-CDF approach when the normality assumption is true. On the other hand, since the invert-CDF approach is based upon inverting the distribution functions, intensive computation is involved in the estimation of distribution functions but weaker assumptions are needed. In the future work, we intend to remove the normality assumption for the B2PSP estimator, so that it can be applied to data having any distribution of form.

Table III.1 Empirical bias × 10^2, root mean squared errors × 10^2, average width of 95% CI × 10^2, and non-coverage rate of 95% CI × 10^2 of $\theta(\alpha)$ for α = 0.1, 0.25, 0.5, 0.75, and 0.9: the homogeneous errors case.

	α = 0.1			α = 0.25			α = 0.5			α = 0.75			α = 0.9		
	$\hat{\theta}_W$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}_{B2PSP}$	$\hat{\theta}_W$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}_{B2PSP}$	$\hat{\theta}_W$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}_{B2PSP}$	$\hat{\theta}_W$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}_{B2PSP}$	$\hat{\theta}_W$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}_{B2PSP}$
LINUP															
Bias	-1.5	-0.9	0.7	-0.3	0.1	0.3	-0.3	-0.2	-0.2	-0.2	-0.3	-0.1	-0.6	-0.5	0.2
RMSE	7.7	6.8	5.4	5.7	5.0	4.2	4.8	3.9	3.5	4.4	3.3	3.0	4.2	3.2	2.7
Width of 95% CI	24.8	25.7^{+} $26.4^{\#}$	21.0	23.1	19.1^{+} $20.7^{\#}$	15.7	18.8	14.9^{+} $15.5^{\#}$	12.8	17.9	13.4^{+} $13.6^{\#}$	11.3	18.7	13.3^{+} $13.5^{\#}$	10.6
Noncov. Rate	11.9	8.6^{+} $7.5^{\#}$	5.6	6.0	6.6^{+} $4.5^{\#}$	5.9	4.2	6.3^{+} $3.7^{\#}$	5.6	4.1	7.1^{+} $4.1^{\#}$	5.6	3.9	8.4^{+} $4.3^{\#}$	5.9
EXP															
Bias	-0.9	-0.4	0.04	-0.3	0.4	-0.1	-0.2	0.4	-0.6	-0.2	0.5	-0.6	-0.8	0.8	0.7
RMSE	6.5	5.9	4.8	4.9	4.5	3.9	4.6	3.9	3.6	5.0	4.0	3.5	7.2	4.5	3.4
Width of 95% CI	23.1	22.8^{+} $24.4^{\#}$	19.6	19.9	17.5^{+} $18.8^{\#}$	14.6	17.5	15.2^{+} $16.0^{\#}$	12.6	21.0	16.0^{+} $17.1^{\#}$	13.2	40.2	18.7^{+} $23.0^{\#}$	13.4
Noncov. rate	10.6	8.5^{+} $5.8^{\#}$	5.5	6.4	5.9^{+} $3.7^{\#}$	6.0	4.7	6.9^{+} $3.8^{\#}$	5.8	4.0	7.0^{+} $3.4^{\#}$	6.6	4.0	7.0^{+} $4.2^{\#}$	6.4

* $\hat{\theta}_W$ is the sample-weighted estimator; $\hat{\theta}_{inv-CDF}$ is the invert-CDF Bayesian estimator; $\hat{\theta}_{B2PSP}$ is the B2PSP estimator.

+ Woodruff's CI method for the invert-CDF Bayesian approach illustrated in Figure III.2(a)

The second CI calculation method for the invert-CDF Bayesian approach illustrated in Figure III.2(b)

Table III.2 Empirical bias × 10^2, root mean squared errors × 10^2, average width of 95% CI × 10^2, and non-coverage rate of 95% CI × 10^2 of $\theta(\alpha)$ for α = 0.1, 0.25, 0.5, 0.75, and 0.9: the heterogeneous errors case.

	α = 0.1			α = 0.25			α = 0.5			α = 0.75			α = 0.9		
	$\hat{\theta}_W$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}_{B2PSP}$	$\hat{\theta}_W$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}_{B2PSP}$	$\hat{\theta}_W$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}_{B2PSP}$	$\hat{\theta}_W$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}_{B2PSP}$	$\hat{\theta}_W$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}_{B2PSP}$
LINUP															
Bias	-0.5	-0.4	-1.0	-0.1	0.3	-0.1	-0.01	0.7	0.8	-0.1	0.1	0.8	-0.4	-0.4	0.1
RMSE	3.1	2.9	3.0	2.8	2.4	2.5	3.3	2.5	2.3	4.5	2.7	2.4	5.1	3.5	3.0
Width of 95% CI	14.1	12.5^{+} $13.6^{\#}$	18.8	11.0	10.1^{+} $10.8^{\#}$	15.9	13.3	9.6^{+} $10.1^{\#}$	13.7	18.4	11.8^{+} $12.0^{\#}$	12.6	21.9	14.9^{+} $15.1^{\#}$	12.9
Noncov. Rate	13.8	6.2^{+} $5.2^{\#}$	1.5	6.9	5.4^{+} $3.4^{\#}$	0.2	4.1	5.1^{+} $4.3^{\#}$	0.5	5.0	4.0^{+} $3.1^{\#}$	2.2	4.2	4.9^{+} $3.5^{\#}$	3.8
EXP															
Bias	-0.3	-0.7	-1.7	-0.3	-0.1	-1.2	-0.2	0.6	-0.05	0.1	1.0	1.0	-0.2	-0.4	0.4
RMSE	3.0	3.0	3.2	2.6	2.3	2.7	2.6	2.3	2.1	4.1	3.0	2.5	8.4	4.4	3.5
Width of 95% CI	13.5	12.5^{+} $13.8^{\#}$	19.2	10.0	9.8^{+} $10.5^{\#}$	15.6	10.6	9.4^{+} $10.1^{\#}$	13.4	18.6	11.7^{+} $12.8^{\#}$	13.7	37.8	19.9^{+} $22.4^{\#}$	14.8
Noncov. rate	11.1	4.9^{+} $3.6^{\#}$	2.4	6.5	5.2^{+} $3.8^{\#}$	0.9	4.6	4.8^{+} $2.8^{\#}$	0.1	4.5	4.5^{+} $5.3^{\#}$	2.7	3.4	5.0^{+} $3.4^{\#}$	4.7

* $\hat{\theta}_W$ is the sample-weighted estimator; $\hat{\theta}_{inv-CDF}$ is the invert-CDF Bayesian estimator; $\hat{\theta}_{B2PSP}$ is the B2PSP estimator.

\+ Woodruff's CI method for the invert-CDF Bayesian approach illustrated in Figure III.2(a)

\# The second CI calculation method for the invert-CDF Bayesian approach illustrated in Figure III.2(b)

Table III.3 Empirical bias × 10^2, root mean squared errors × 10^2, average width of 95% CI × 10^2, and non-coverage rate of 95% CI × 10^2 of $\theta(\alpha)$ for α = 0.1, 0.25, 0.5, 0.75, and 0.9: the mixture model case.

	α = 0.1		α = 0.25		α = 0.5		α = 0.75		α = 0.9	
	$\hat{\theta}$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}$	$\hat{\theta}_{inv-CDF}$	$\hat{\theta}$	$\hat{\theta}_{inv-CDF}$
LINUP										
Bias	-0.4	-0.04	0.1	1.1	0.5	0.2	0.5	-0.4	0.4	-0.2
RMSE	3.6	3.5	4.0	3.6	6.5	5.1	10.7	9.7	22.3	19.9
Width of 95% CI	15.5	14.6^{+} $15.6^{\#}$	15.5	13.4^{+} $14.3^{\#}$	25.7	18.5^{+} $20.3^{\#}$	44.8	35.0^{+} $40.8^{\#}$	126.8	67.0^{+} $88.5^{\#}$
Noncov. Rate	14.2	7.3^{+} $4.6^{\#}$	7.2	9.0^{+} $4.8^{\#}$	5.2	10.9^{+} $7.1^{\#}$	5.4	11.9^{+} $6.3^{\#}$	5.0	17.0^{+} $9.6^{\#}$
EXP										
Bias	-0.5	-0.7	-0.1	0.7	0.7	1.6	-0.5	-0.4	-0.7	0.7
RMSE	3.4	3.3	3.4	3.3	6.7	5.7	5.2	4.5	6.5	5.2
Width of 95% CI	15.5	14.2^{+} $15.6^{\#}$	13.2	13.5^{+} $14.1^{\#}$	23.8	18.7^{+} $19.6^{\#}$	21.8	18.3^{+} $19.3^{\#}$	42.8	19.4^{+} $22.4^{\#}$
Noncov. Rate	9.9	4.0^{+} $3.6^{\#}$	6.9	5.9^{+} $3.7^{\#}$	5.0	11.7^{+} $6.3^{\#}$	4.5	8.4^{+} $4.6^{\#}$	2.7	7.5^{+} $3.8^{\#}$

* $\hat{\theta}_W$ is the sample-weighted estimator; $\hat{\theta}_{inv-CDF}$ is the invert-CDF Bayesian estimator

+ Woodruff's CI method for the invert-CDF Bayesian approach illustrated in Figure III.2(a)

The second CI calculation method for the invert-CDF Bayesian approach illustrated in Figure III.2(b)

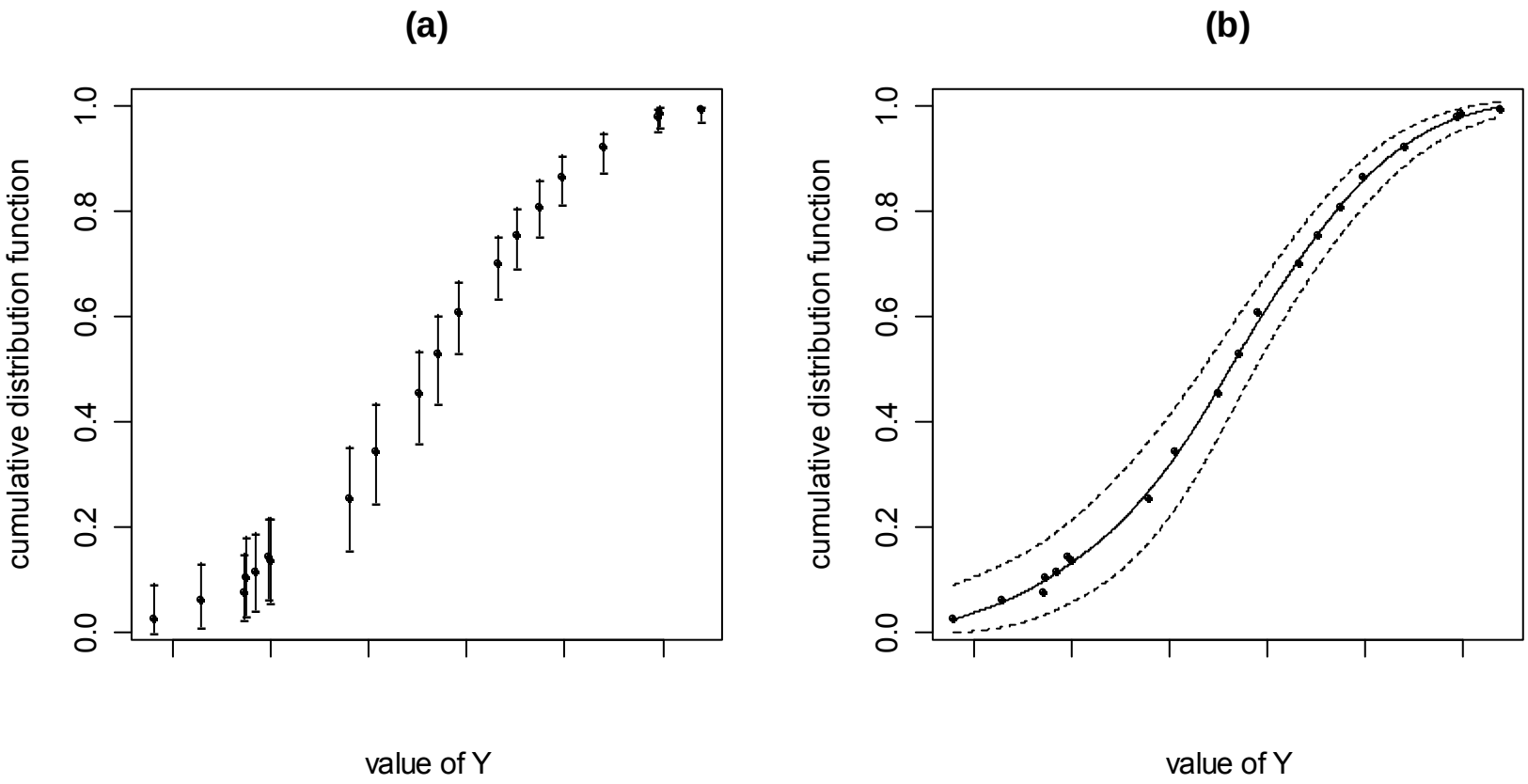

Figure III.1 Bayesian model-based approach in estimating finite population distribution functions illustrated using a sample of size 100 drawn from a finite population. (a) BPSP method is used to estimate the finite population distribution functions at 20 sample points; the dots denote BPSP estimators and the minus signs denote the upper and lower limits of the 95% CI. (b) Three monotonic smooth cubic regression models are fit on the BPSP estimators, upper limits, and lower limits; the solid curve is the predictive continuous distribution functions and the two dash curves are the 95% CI of the distribution functions.

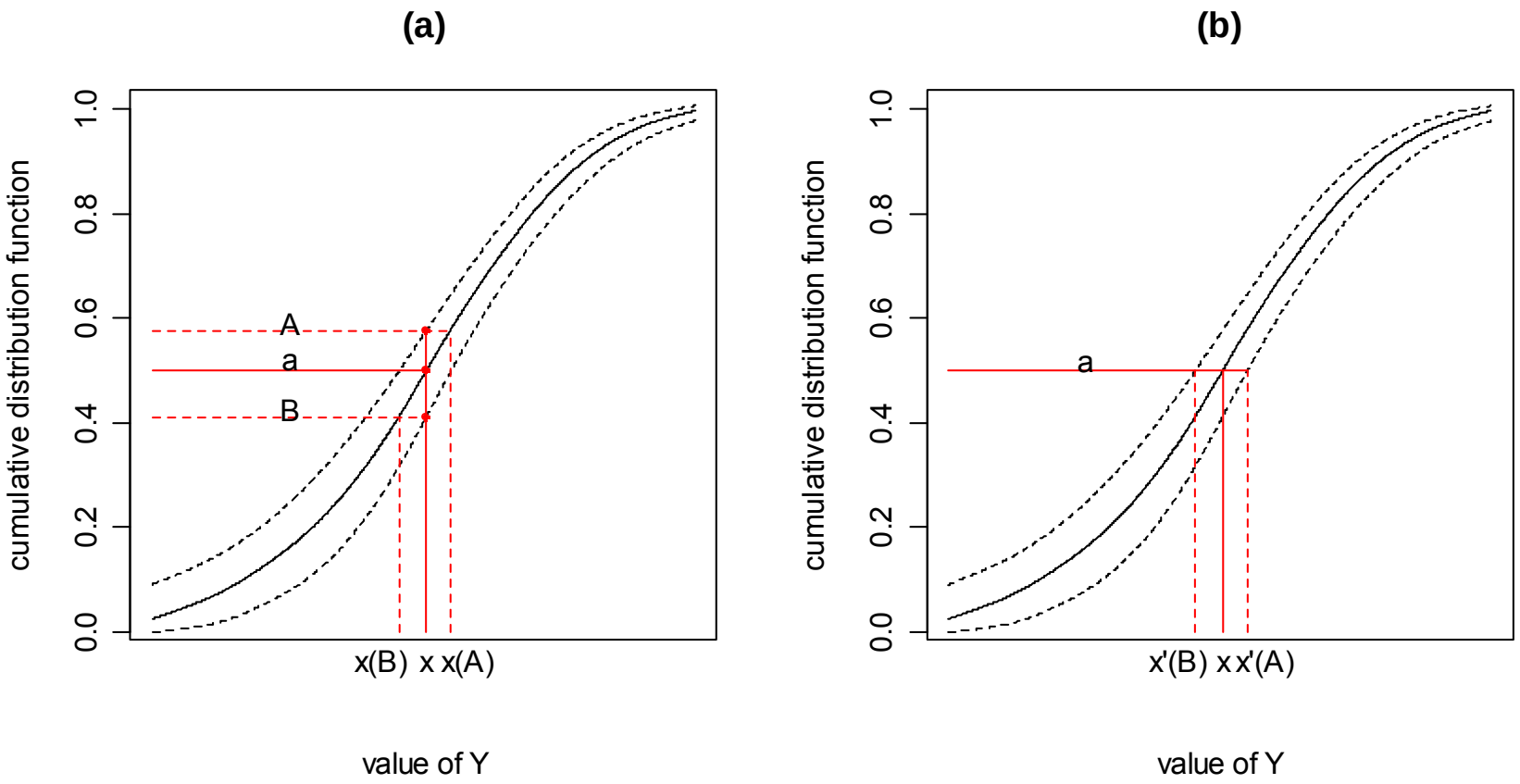

Figure III.2 Invert-CDF Bayesian model-based approach of estimating finite population quantiles and methods of calculating the 95% CI. (a) Extension of the Woodruff's confidence intervals; *x* is the estimated *α*-quantile, *A* and *B* are the upper and lower limits of the 95% CI of *F(x)*, and x_A and x_B are the upper and lower limits of 95% CI for the quantile. (b) Direct estimation from the upper and lower bound of the 95% CI of the distribution functions; x'_A and x'_B are the upper and lower limits of 95% CI for the quantile.

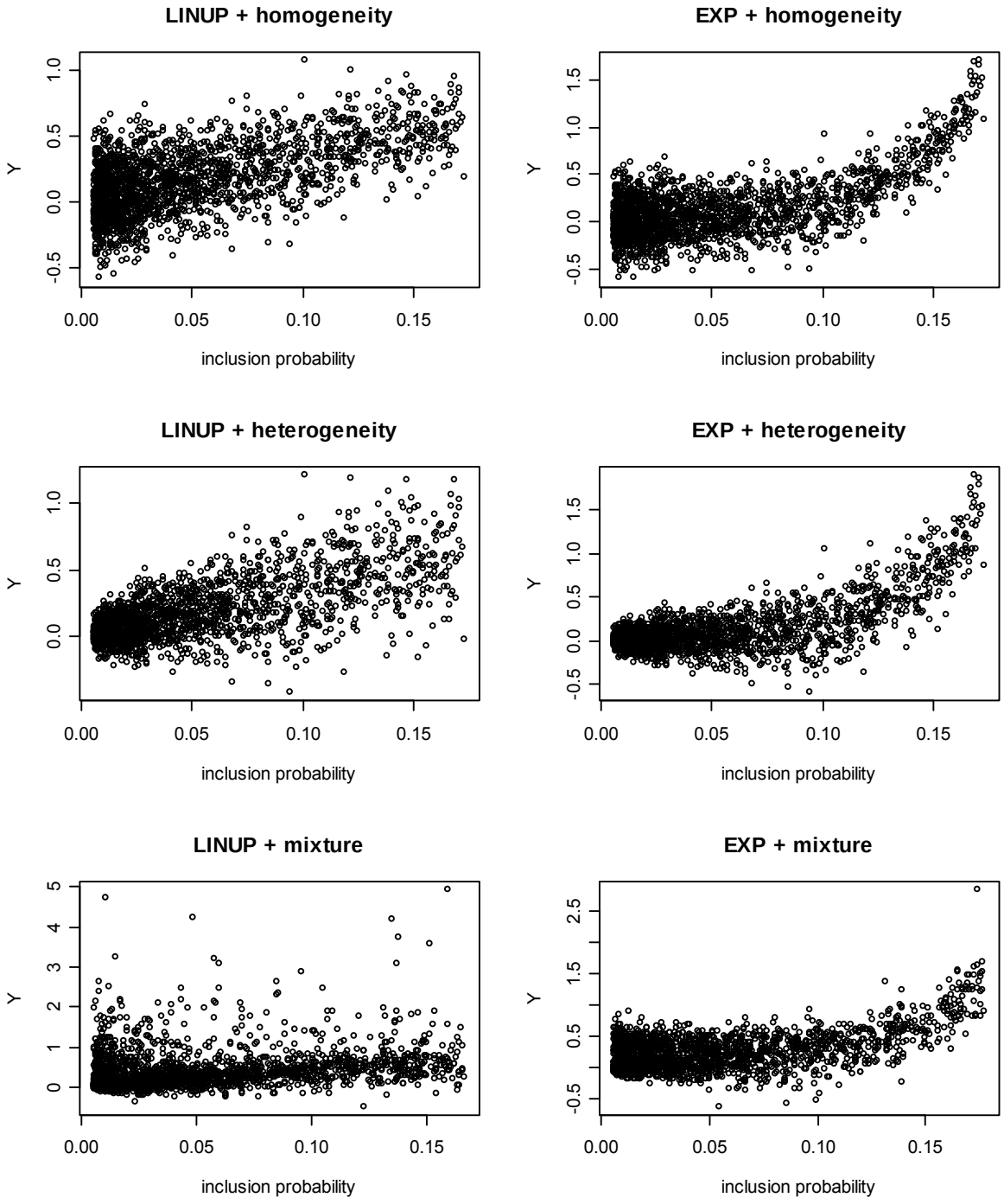

Figure III.3 Scatter plots of Y versus the selection probabilities among six artificial finite populations.

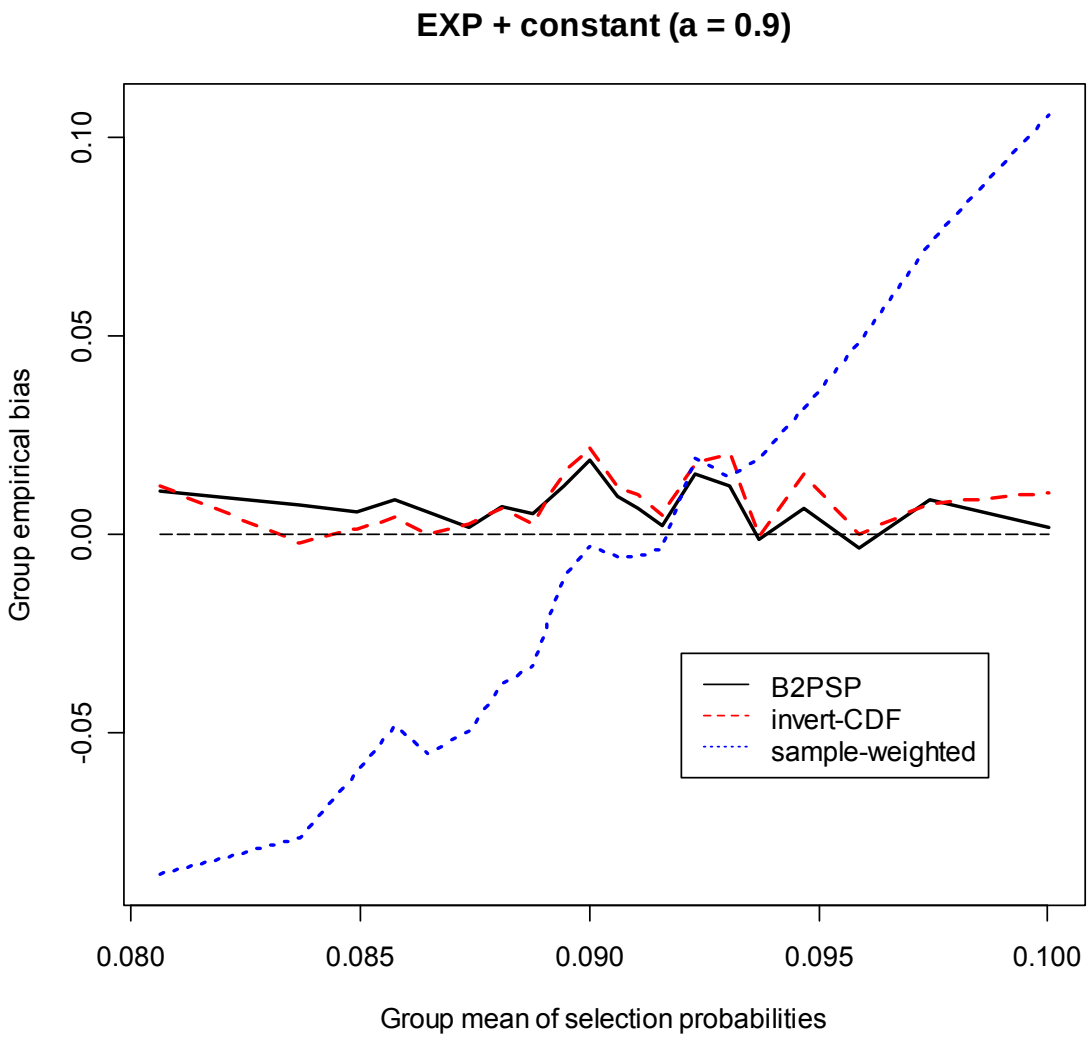

Figure III.4 Variation of empirical bias of the three estimators for 90th percentile from the "EXP + homogeneity" case

Chapter IV

STEPWISE VARIABLE SELECTION IN MULTIPLY IMPUTED DATA

IV.1 Introduction

In health survey studies, item nonresponse happens when particular items are missing for an individual. For example, a sensitive question regarding income may not be answered on a questionnaire. The missing values due to item nonresponse typically have a haphazard pattern. If the amount of item nonresponse is nontrivial or if a small amount of nonresponse occurs in several variables in different individuals, the default strategy of eliminating all incomplete cases from the analysis is wasteful of costly collected data, and can lead to problematic inference for the target population.

Item nonresponse is often handled by multiple imputation (Rubin 1987; Little and Rubin 2002), which refers to a procedure of 'filling in' missing data with plausible values $D > 1$ times to create multiple datasets. Commonly, $D = 5$. It has been used in, for example, the National Health and Nutrition Examination Survey (Schafer et al. 1993), the National Survey of Family Growth (Lepkowski et al. 2006), and National Health Interview Survey (Schenker et al. 2006). After multiple imputation, each imputed dataset is analyzed identically by a complete-data method. The multiple parameter estimates and standard errors are then combined using Rubin's multiple imputation combining rule (Rubin 1987) to account for both the within-imputation and between-imputation variations.

Besides the problem of item nonresponse, there are often many potential explanatory variables for an outcome of interest in large health surveys. Health scientists usually have some predictors of interest in mind, but they are also interested in knowing if there are any confounders or additional important explanatory variables beyond their hypotheses of interest. When the number of candidate explanatory variables is large, variable selection algorithms are used to identify statistically significant predictors.

While there are other more rigorous variable selection methods available, the methods of all possible subsets, backward elimination, and forward stepwise selection are the most commonly used variable selection methods for regression models in complete-data. The method of all possible subsets is only feasible with small number of potential predictors (less than or equal to 10). Backward elimination and forward stepwise selection methods are commonly used when investigators need to explore large numbers of potential explanatory variables, and are practical because standard statistical software packages allow their implementation. Backward elimination starts with all potential explanatory variables in the model and deletes a variable in each step. It is impossible to use backward elimination when the number of predictors is equal to or larger than the number of observations. Forward stepwise selection starts with null model and adds a variable in each step. It is the only feasible method for very large predictor pools. When selecting a set of important variables, statistical significance should not be the only criterion considered. Consideration should also be given to issues such as the form of the variables in the model, variable stability, confounding, collinearity, and clinic significance. Therefore, model checking must be performed for a statistical model determined by data-driven statistical selection methods.

However, there are currently no guidelines for the extension of these variable selection methods to the setting of multiply imputed data. Given the multiply imputed data, the method of all possible subsets can be implemented easily by applying Rubin's multiple imputation combining rule to all possible subsets of models. While backward or forward stepwise selection methods commonly produce different statistically significant variables for different imputed data sets; this poses difficulty in pooling the variable selection results across imputed data sets. That is, Rubin's multiple imputation combining rule cannot be directly applied in this situation. Some authors suggest including in a model variables that were selected in at least 3 out of 5 (60%) of the imputed data sets. (Heymans et al. 2007 and Brand 1999) In this paper we refer to this strategy as the select then combine (SC) method. For simplicity, researchers also perform variable selection with any one of the multiply imputed data sets. We call this approach the single imputation (SI) method. Both the SC and SI methods do not fully account for the imputation uncertainty. Alternatively, we develop and implement a combine then select (CS) method by applying Rubin's multiple imputation combining rule in each step of backward or forward stepwise selection (Wood et al. 2008). The CS method leads to a single set of selected variables. We focus on the CS forward stepwise selection method in this paper, since there is a large pool of candidate variables in our real-world example of a community-based dioxin exposure study collected by the University of Michigan. This dioxin exposure study data is used to demonstrate the CS forward stepwise selection method, and simulation studies are conducted to compare the various methods.

IV.2 Materials and methods

IV.2.1 Study design

The University of Michigan Dioxin Exposure Study (UMDES) was designed to assess exposures to polychlorinated dibenzo-*p*-dioxins, polychlorinated dibenzofurans, and dioxin-like polychlorinated biphenyls in the adult population of Midland and Saginaw Counties, Michigan, USA. By measuring factors that reflect potential exposure to dioxins through air, water, soil, food intake, occupations, and various recreational activities, the study sought to identify factors which explain variation in serum dioxin concentrations. Blood serum was analyzed for the World Health Organization 29 list of dioxin-like compounds (Van den Berg et al. 2005) using high resolution gas chromatography-mass spectrometry. Serum concentrations of dioxins were summarized into a single toxic equivalency (TEQ) value, which is the sum of the mass concentrations of the individual dioxin-like compounds multiplied by their toxic equivalent factors (Van den Berg et al. 2005). In total, 946 participants had serum TEQ measures.

IV.2.2 Potential explanatory variables

The following demographic and health variables were considered to be potentially associated with serum TEQ levels: age, gender, race, education, income, body mass index (BMI), BMI loss and gain in the past 12 months, smoking status, pregnancy, childbearing and months of breastfeeding for each child. In addition, the respondent recalled possible dioxin exposure pathways over their entire lifetime, including a full residential history (residence region and number of years living in the contaminated areas), property use (trash burning, pets entering the home, wearing shoes in the home, gardening activities,

using weed killers, raising crops or poultry, use of fireplaces and wood burning stoves, fire damage to the home, and flooding in contaminated areas), occupations with likely exposure to dioxins, military service in Vietnam, fishing, hunting, water sports in the contaminated areas, and consumption of meat, fish, game, eggs, milk, other dairy products, fruits, and vegetables from the contaminated areas or bought from stores (consumption frequency in the last 5 years and the number of years of consumption in the lifetime). According to our hypotheses, TEQ levels in soil and dust samples from the participant's home were of particular interest.

IV.2.3 Statistical analyses

As typically encountered in surveys, the presence of item missing values was encountered in the UMDES. By assuming missing at random (MAR) (Rubin 1987; Little and Rubin 2002), the item missing values in the survey questionnaire and the dust and soil samples were imputed five times using a sequential regression imputation procedure (Raghunathan et al. 2001) as implemented in IVEware (Raghunathan et al. 2008). A logarithm 10 transformation was taken on the serum dioxin concentrations. Forms of candidate explanatory variables and collinearity were investigated before variable selection. A simple, credible model on the serum TEQ levels was constructed with nine important variables based on our prior knowledge and hypotheses of interest. Once we had specified the starting model, we added complexities by an expanding search process using the CS, SC, and SI forward stepwise selection methods. There were a total of 114 continuous or categorical variables in the pool of candidate explanatory variables. All regression models were fit using the survey weights dictated by the study design.

However, the variable selection algorithms can be used in either survey weighted or unweighted regression settings.

The implementation of the CS forward stepwise selection is straightforward. For each data set completed by imputation, a survey-weighted linear regression model was fitted on a variable list including the nine forced-in variables and one variable at a time from the 114 potential variables. Let $\hat{\beta}_{ij}, W_{ij}$, $i = 1,\cdots,5$, $j = 1,\cdots,114$ be the estimated regression coefficient and its variance, respectively, for j^{th} variable in the i^{th} imputed data set. By applying the multiple imputation combining rule, the combined coefficient estimate for the j^{th} variable is $\overline{\beta}_j = \sum_{i=1}^{5} \hat{\beta}_{ij} / 5$. The variability associated with $\overline{\beta}_j$ is $T_j = \overline{W}_j + (1+1/5)B_j$, where $\overline{W}_j = \sum_{i=1}^{5} W_{ij} / 5$, and $B_j = \sum_{j=1}^{5} \left(\hat{\beta}_{ij} - \overline{\beta}_j\right)^2 / (5-1)$. The statistic test $\left(\beta_j - \overline{\beta}_j\right) \times T_j^{-1/2}$ follows a t distribution, with $v = (5-1)\left(1 + \left(\overline{W}_j / B_j\right)/(5+1)\right)^2$ degrees of freedom. Let p_j, $j = 1,\cdots,114$, denote the combined *P* value associated with the t-test for the j^{th} variable. The variable with the smallest p_j was selected into the model, so long as this p_j was not greater than 0.05, the entry significance level into the model.

The procedure successively re-fitted the linear regression models on a variable list including the forced-in variables, the variables entering the model in the previous steps, and one new variable at a time. At the same time, variables once entered could be dropped if they were no longer significant (we set the significance level for staying in the model as 0.05, though this significance level is usually set higher than the significance level for entry into the model) as other variables were added. The stepwise selection

procedure continued until the p_j values for all variables in the model were less than the significance level for staying and none of the variables not in the model satisfied the significance level for entry.

IV.2.4 Results

A SAS macro %MI_SREG_SW was designed to perform the CS stepwise variable selection in linear regression models for complex survey data. This SAS macro can be easily modified and extended to other SAS procedures, such as REG, GLM, and LOGISTIC.

The CS stepwise selection method identified 14 out of the 114 candidate explanatory variables as statistically significant predictors, not including the nine forced-in variables already in the model. The 23 variables accounted for 71.7% (R-square) of variation in serum TEQ concentrations. For comparison to the SI and SC methods, Table IV.1 lists side by side the statistically significant variables selected by using SI method based on each imputed dataset, SC method using a selection rule of significance in at least two or three of five imputations, and the CS method. Table IV.1 shows that the first 11 variables were selected into a model by using any of the three selection methods. In addition, the SI method selected another three, five, nine, six, and nine variables into a model by using imputations 1-5, respectively. As a result, the SC method selected 12 additional variables that appeared at least twice in the SI method, and five additional variables that appeared at least three times in the SI method. This shows that the SC method can lead to different numbers of variables being selected, based on an arbitrary selection rule.

For the CS method, variables 12-14 together with the first 11 common variables were selected into a model, where variables 12-14 were selected in different combinations in different imputations with the SI method. We performed regression model diagnostics which indicated that the significance of some of variables 15-24 (selected by the SI method) was due to a few influential observations. This demonstrated that the SI method was more likely to select unstable variables into the model compared to the CS method.

IV.3. Simulation study

We performed a series of simulation studies similar to those given by Yang et al. (2005) in a Bayesian variable selection. We compared the performance of the CS stepwise selection method with two existing methods: the SI method and the SC method (with variable entry requiring significance in at least three out of five simulations). We generated a complete dataset of size 500 with 10 variables $\{X_1, \quad X_2, \quad \cdots \quad X_{10}\}$ which followed a multivariate normal distribution. We considered two compound symmetric correlation matrices for the 10 variables, with an off-diagonal value of 0.1 for lower collinearity and of 0.5 for higher collinearity. The outcome, *Y*, was generated from a normal distribution with a mean function of $X_1 + 2X_4 + X_6 + 3X_7$, and a variance of 2.5. We studied two data missing mechanism: missing complete at random (MCAR) and MAR. Under the MCAR mechanism, where the missingness does not depend on the data values, 5% or 10% of observations were dropped independently for each variable X_1 to X_{10}, yielding 60% or 35% complete cases across the 10 variables. Alternatively, under the MAR mechanism, where missingness depends on the observed data but not on the missing components, X_1, X_2, X_3, X_4, and X_5 were fully observed and 10% or 20% of data

were deleted independently in each variable X_6, X_7, X_8, X_9, and X_{10}, producing comparative missing fractions as with the MCAR mechanism. The probability of missingness among X_6, X_7, X_8, X_9, and X_{10} was in a form of $logit(\mathrm{P}(X_k = \text{missing})) = \alpha_{0k} + \sum_{j=1}^{5} \alpha_{jk} X_j$, $k = 6,\ldots,10$, with parameters chosen to produce the 10% or 20% of missing data. We used sequential regression imputation to impute the missing data five times before the variable selection.

The performance of the three methods were compared by counting the number of incorrect variable selections *C*, including the number of variables that were not selected into model among the important variables $\{X_1, \quad X_4, \quad X_6, \quad X_7\}$ and the number of variables that were selected into model among the remaining six other variables. A total of 100 replicates were created, and the mean and standard deviation of *C* were compared among the SI, SC (with significance in at least 3 out of 5 imputations), and the CS methods.

Table IV.2 shows that the CS method produces fewer incorrectly selected variables than either the SI or SC methods, and the SC method performs better than the SI method in all cases. The latter result is because the SC has partially incorporated the imputation uncertainty, whereas the SI method does not. In addition, the rule of including variables that were statistically significant in three of five imputations is arbitrary, and led to more incorrectly selected variables than the CS method. The *C* values are higher when the missing fraction is higher and when the correlation coefficient ρ is large.

IV.4. Discussion

Researchers often face the challenge of performing variable selection with multiply-imputed datasets. The commonly used approaches have limitations and deficiencies that reduce their utility. Stepwise variable selection based on a single imputation fails to account for imputation uncertainty. A modification includes choosing only the variables that are selected repeatedly across the multiple imputations. However, the selection rule based on an arbitrarily chosen significance proportion (such as three of five imputations) can lead to different variables being selected if different significant proportions are chosen. We modified the stepwise variable selection method for complete-data to the setting of multiply imputed data by using the multiple-imputation combination rule to obtain combined inferences from multiple datasets in each step of variable selection and selecting variables based on the combined *P* values. This CS stepwise selection method has theoretical advantages because it accounts for imputation uncertainty and yields a single model. Furthermore, our approach, which is based on Rubin's multiple imputation combining rule, preserves the type I error, which the SI and SC methods do not. (Wood et al. 2008)

A simulation study in the setting of MCAR and MAR showed that the CS stepwise selection method is less likely to incorrectly select variables into the models compared to both the SI method and the SC method (with 60% as cutoff value). We did not include the setting of not missing at random (NMAR) in the simulation study, because it usually requires more complicated multiple imputation methods than the sequential regression imputation method, but the CS stepwise selection method is still applicable after the missing values are properly imputed.

The number of incorrectly selected variables included the number of important variables that were missing and the number of noise variables that were incorrectly included. When these two kinds of mistakes were separated, the incorrectly selected variables were actually all the noise variables that were incorrectly included in the simulation study. All the important variables were included by using all three methods, but more noise variables were included in the SI and SC methods than the CS method, because imputation uncertainty is not or only partially accounted for.

Although in this article we only describe the modified stepwise variable selection method in the setting of multiply imputed data, this combine then select variable selection algorithm can also be applied to forward and backward variable selection methods. We have not applied this method to the setting of simultaneous imputation and variable selection. There are many instances where these activities cannot be combined, as in the use of public datasets for example, where the imputation is performed before the data are released to the public. Extension of our approach to this setting would be valuable. Other approaches such as Bayesian variable selection for multiply imputed data have theoretical justification (Yang et al. 2005). However, there are currently no Bayesian variable selection methods available for multiply imputed complex survey data.

We do not suggest that automated methods are preferable over careful model building based on background knowledge. Automated methods are often used because there are too many possible variables and expansion terms to consider within a reasonable length of time (Rothman et al. 2008). Problems with falsely narrow confidence limits, noise variables gaining entry into models, and lack of logical justification for models must be considered when automated methods are used. (Wood et al. 2008; Rothman et al. 2008)

Table IV.1 Comparison of various selection methods in selecting statistically significant predictors of $\log_{10}$ serum TEQ levels in the UMDES example

Var #	Variable Name	Methods							
		SI					SC		CS
		1st.	2nd.	3rd.	4th.	5th.	≥ 2/5	≥ 3/5	
1	Living in Midland/Saginaw counties (yrs)	x	x	x	x	x	x	x	x
2	No. of pregnancies without children	x	x	x	x	x	x	x	x
3	BMI loss in last 12 months	x	x	x	x	x	x	x	x
4	Interaction term of BMI and gender	x	x	x	x	x	x	x	x
5	Interaction term of gender and age	x	x	x	x	x	x	x	x
6	Maximum soil dioxin level on property	x	x	x	x	x	x	x	x
7	No. of yrs living on a farm in 1940- 1959	x	x	x	x	x	x	x	x
8	No. of yrs using weed killers on the property in 1960-1979	x	x	x	x	x	x	x	x
9	No. of yrs living in a property ever damaged by a fire in 1960 -1979	x	x	x	x	x	x	x	x
10	No. of yrs working at Dow Chem. Co.	x	x	x	x	x	x	x	x
11	No. of meals of fish except walleye or perch in Saginaw River/Bay in last 5 yrs	x	x	x	x	x	x	x	x
12	No. of days doing water activities in or around the Tittabawassee River 1960 -1979	x		x	x	x	x	x	x
13	No. of days fishing in Saginaw River/Bay after 1980	x	x		x		x	x	x
14	No. of years eating fish after 1980	x	x				x		x
15	Household dust dioxin loading		x	x	x	x	x	x	
16	No. of years doing water activities in or around any other Michigan rivers or lakes in1960-1979			x	x	x	x	x	
17	No. of meals of sport-caught walleye or perch not from the contaminated area in last 5 yrs			x	x	x	x	x	
18	No. of days of fishing in Tittabawassee River in 1960-1979		x		x		x		
19	No. of meals of walleye or perch from Tittabawassee River in last 5 yrs.		x						
20	No. of meals of fish except walleye or perch from Tittabawassee River in last 5 yrs.			x		x	x		
21	No. of years living on a property ever damaged by a fire in 1940-1959			x		x	x		
22	No. of years of working in dioxin exposed jobs in 1960-1979			x		x	x		
23	No. of days hunting in the area of the Tittabawassee River after 1980			x		x	x		
24	No. of days fishing in the area of the Tittabawassee River after 1980			x		x	x		

Table IV.2 Comparison of various selection methods showing the average number of incorrectly selected variables C for two levels of correlation among predictors and varying types and levels of item missing data

	Methods		
	SI	SC	CS
Average number of incorrectly selected variables C (standard deviation)			
$\rho = 0.1$			
MCAR 5%	0.54 (0.64)	0.31 (0.51)	0.10 (0.33)
MCAR 10%	1.14 (0.77)	0.81 (0.69)	0.27 (0.45)
MAR 10%	0.54 (0.66)	0.20 (0.43)	0.04 (0.20)
MAR 20%	0.69 (0.66)	0.32 (0.55)	0.05 (0.22)
$\rho = 0.5$			
MCAR 5%	1.77 (0.78)	1.56 (0.76)	0.99 (0.59)
MCAR 10%	2.80 (0.77)	2.78 (0.92)	1.91 (0.64)
MAR 10%	1.73 (0.78)	1.36 (0.80)	0.82 (0.64)
MAR 20%	2.17 (0.71)	2.08 (0.91)	1.35 (0.61)

Chapter V

ESTIMATION OF BACKGROUND SERUM 2, 3, 7, 8 - TCDD CONCENTRATIONS USING QUANTILE REGRESSION IN THE MICHIGAN (UMDES) AND NHANES POPULATIONS

V.1. Introduction

The University of Michigan Dioxin Exposure Study (UMDES) was conducted in response to concern that people's body burdens of dioxins might be elevated in Midland and Saginaw counties, Michigan, because of environmental contamination from the Dow Chemical Company facilities in the City of Midland and sediments in the Tittabawassee River flood plain. To assess whether the concentrations of blood serum dioxins are elevated among residents in Midland/Saginaw, they need to be compared with the background concentrations of serum dioxins in other areas where there are no known, unusual sources of dioxin exposures. The most studied dioxin congener, 2, 3, 7, 8-tetrachlorodibenzo-*p*-dioxin (TCDD), is formed as an unintentional by-product of incomplete combustion and has been classified as a probable human carcinogen (Agency for Toxic Substances and Disease Registry 1998). Our goal in this paper is to estimate the mean and quantiles of serum TCDD levels in the general population.

Studies have shown that, among the general public, the concentrations of serum dioxins increase with age. (Patterson Jr. et. al 2004; Wittsiepe et al. 2000) These increases are most likely the result of higher levels of dioxins in the environment in the 1960's and 1970's than in recent years, the number of years of past exposure, and slower

elimination among older people. In addition, the difference in serum dioxin concentrations by sex may be due to differences in elimination between males and females. (Patterson Jr. et al. 2008) Therefore, the estimation of background concentrations of serum dioxins must be adjusted for these factors.

In exposure assessment, quantiles are sometimes of more interest than means, from a public health perspective. In the presence of a skewed distribution, quantiles can also catch important information that might be missed by measurements of central tendency and dispersion. Since age and sex are associated with serum TCDD concentrations, an age- and sex- specific quantile estimate among the reference population is of greater interest than a univariate quantile estimate. Quantile regression (Koenker 2005) is used to estimate and allow inferences about conditional quantile functions given covariates, similarly as linear regression is used to predict conditional means given covariates.

We used two referent populations for estimation of the quantiles of serum TCDD levels: the population of Jackson and Calhoun counties in Michigan and the 2003-2004 National Health and Nutrition Examination Survey (NHANES). (NCEH/CDC 2005) The Jackson/Calhoun sample was advantageous because it was representative of local Michigan residents. The NHANES sample was advantageous because it included large numbers of subjects and was representative of the U.S. general population. However, for the serum TCDD, about half the NHANES data were below the limit of detection (LOD). The LOD is defined as the concentration of analyte which gives a signal equal to a laboratory blank (obtained when no analyte is present) plus three times the standard deviation of the blank. (Keith et al. 1983) The LOD represents the level below which we cannot be confident whether or not the analyte is actually present. The high proportion of

TCDD samples below LOD in the NHANES data has resulted in the difficulty in estimating age- and sex- specific mean, median, and lower quantiles of serum TCDD concentration in general U.S. population based on the NHANES data only.

The present study applies linear and quantile regression methods in the setting of complex survey data to quantify the age- and sex- specific mean and quantiles estimates of the background serum TCDD concentrations in the general Michigan population and separate estimates in the general U.S. population. It also illustrates a multiple imputation approach for imputing values below the LOD. The LOD issue has posed formidable limitations to the estimation of serum TCDD levels (and levels of other environmental contaminants that are commonly measured near the limit of detection) in the general population. Conventional approaches of imputing the values below the LOD as 0, LOD, LOD/2, or LOD/$\sqrt{2}$ depend on the blood sample volume and the LOD levels of the measurement methods and may lead to biased estimates of serum TCDD concentration, especially in the scenario of high proportion of data above the LOD and high LOD levels. (Hornung and Reed 1990)

V.2. Materials and methods

V.2.1. Study population

Jackson and Calhoun counties, Michigan, are over 100 miles away from Midland, Michigan. The population of these counties was chosen as the reference population in the UMDES because it was similar to Midland and Saginaw counties in terms of demographics, urban/rural distribution, and percent employment in industry – except that there is no known unusual source of dioxins, such as the Dow Chemical Company. To be

eligible for participation in this study, the Jackson/Calhoun residents were required to be 18 years or older and to have lived in their current residence for at least five years. The sampling used a two-stage area probability selection of housing units in Jackson and Calhoun counties and a third stage of selection of an eligible person within each sample housing unit. (Lepkowski et al. 2006) Participants provided written, informed consent that had been approved by the University of Michigan Health Sciences Institutional Review Board. Participants who met Red Cross criteria for blood donation (no clotting disorders or blood thinner medications, no recent chemotherapy, weight of at least 110 pounds, etc.) were invited to provide an 80 milliliter (ml) sample of blood. In Jackson/Calhoun counties, the study cooperation rates (proportion of known eligible persons who provided data) were 82.2 percent in the interviewing stage and 78.4 percent in the blood collection stage. (Lepkowski et al. 2006) A total of 359 persons in Jackson and Calhoun counties completed the UMDES study questionnaire and among whom 251 gave blood samples in the summer of 2005. (UMDES 2008)

All serum TCDD analyses were performed by Vista Analytical Laboratory of El Dorado Hills, CA, using high resolution gas chromatography-mass spectrometry. To ensure the precision and accuracy of the serum results, Vista Analytical Labs first synchronized its serum analysis methods with the methods of the National Center for Environmental Health (NCEH) laboratories at the Centers for Disease Control and Prevention (CDC) before the start of UMDES fieldwork. Additionally, 20 serum quality assurance and quality control (QA/QC) samples were supplied by NCEH, blind analyzed by Vista Analytical Laboratory during fieldwork, and the results verified by NCEH labs. The mean lipid content of these samples was measured at 586 mg/dL (s.d. = 20) by Vista,

compared to 603 mg/dL (s.d. = 21) by NCEH. Vista's analytical results for TCDD concentration were within 2 standard deviations of the sample means determined by NCEH after repeated testing of these samples over time. Serum standard reference materials, supplied by the National Institute of Standards and Technology (NIST), and pooled serum samples were analyzed periodically (1 each per 40 samples) to verify the method performance. The serum TCDD concentration was divided by the total lipids and was reported in parts per trillion (ppt) or picograms/gram lipids. The lipids were determined by measurements of triglycerides and total cholesterol and then the total lipids were calculated using the Phillips method, as was done in the NCEH laboratory. (Philips et al. 1989)

One limitation of the Jackson/Calhoun data is that relatively few participants (n = 20) were older than 75 years, especially males (n = 3). As a result, estimating age- and sex-specific upper percentiles was problematic in this group. However, a substantial data set of serum TCDD concentrations in adults aged 18 to 85 years exists in the 2003-2004 NHANES. (NCEH/CDC 2005) The serum dioxin analyses in the NHANES were performed by NCEH. Since the methods for serum dioxin and lipid quantification are comparable between the NCEH labs and the Vista Analytical Lab and the results verified via blind sample introduction, the blood serum data of the UMDES and the NHANES can be combined with little or no expectation of bias. Since the population in Jackson/Calhoun counties was predominantly non-Hispanic whites (91 percent) and pregnant women were excluded, we examined information from the NHANES subsample of 719 non-Hispanic whites (excluding pregnant women) who had serum TCDD measures. There were 98 participants (40 males) older than 75 years who had serum

TCDD concentrations above the LOD in the NHANES data, and who could be useful in improving the age- and sex-specific percentile estimates among older people in Jackson/Calhoun.

V.2.2. Statistical analyses

The 719 observations from NHANES data were concatenated with the 251 observations from UMDES Jackson/Calhoun data, with an indicator for data source (1 for NHANES, 0 for UMDES). To be consistent with the NHANES dataset, participants who were older than 85 years in the Jackson/Calhoun dataset were recorded as being age 85 to preserve the anonymity of people participating in the study. Age and sex were fully observed for both samples. Other covariates potentially associated with serum TCDD concentration were body mass index (BMI), recent BMI change, cigarette smoking, income, education, and breast feeding history among women. (Garabrant et al. 2009) All of these covariates had less than 7.5 percent of data missing in both samples. They were imputed separately using a sequential regression imputation method in both samples before the combination. (Raghunathan et al. 2001) The survey sampling weights were standardized within each data source by dividing by their respective mean sampling weights and then multiplying by 100 in order to maintain the ratio of the data source sample sizes; this prevents the analysis results from being overwhelmed by the NHANES data due to its much larger sampling weights (each individual observation represents many more people in the population).

A multiple imputation technique was performed to impute the TCDD concentrations for those below the LOD in the combined data of Jackson/Calhoun and NHANES.

(Rubin 1987; Little and Rubin 2002) For each imputation, a bootstrap sample of 970 (n=251 for Jackson/Calhoun, n=719 for NHANES) observations was generated from the combined Jackson/Calhoun and NHANES data, and a survey weighted left-censored (Tobit) linear regression model, assuming a lognormal distribution, was fitted on the bootstrap sample with important covariates including data source, age, sex, BMI, BMI change in the past 12 months, pack-years of cigarette smoking, number of children breast fed (among women subjects), income, education, and the two-way interaction terms among age, sex, and data source. Then, for those subjects having values below the LOD, the natural logarithm transformed imputed values were drawn from a normal distribution with mean and variance estimated from the left-censored regression model, with left truncation at their corresponding natural logarithm LOD. The above procedure was repeated five times to generate five imputed data sets.

A natural logarithmic transformation was applied to the serum TCDD concentrations, because there is an approximately linear association between log (serum TCDD) and age. To estimate the age- and sex-specific serum TCDD measures, survey weighted mean, quartiles (25^{th} percentile, median, 75^{th} percentile) and 95^{th} percentile of quantile regression models of serum TCDD concentrations were fitted on age, sex, data source indicator, and their three two-way interaction terms. Age was centered at 50 years to facilitate interpretation of the intercept and to remove collinearity of age with its interaction terms with sex and data source. Since neither the interaction term between age and data source indicator nor the interaction term between sex and data source indicator was significant in any of the mean or quantile regression models, they were removed from all of the models. For the mean regression, we used the conventional

method for complex surveys by using the SURVEYREG procedure in SAS, version 9.1 (SAS Institute Inc., Cary, North Carolina). However, since there is no statistical software package currently available that provides correct standard error estimates for quantile regression in complex surveys, we corrected the estimates of standard errors of the regression coefficients using 1,000 bootstrap samples. (Efron and Tibshirani 1994) The estimates for each parameter from five imputed data sets were averaged to get the combined parameter estimate, and the variances were computed using standard multiple imputation combining rules that account for between and within imputation variances. (Rubin 1987)

We used the bootstrap method for stratified multistage samples in both the multiple imputation and the quantile regression. (Rust and Rao 1996) For a single replicate of bootstrap, for each stratum h, draw, from the n_h primary sampling units (PSUs) in the sample, a simple random sample with replacement of $m_h = (n_h\text{-}1)$ PSUs. Let $r_{hi}^{(t)}$ denote the number of times that PSU i from stratum h is included in replicate t and let w_{hij} denote the sample weight for unit j in the PSU i and stratum h, the bootstrap weights were calculated as $w_{hij}^{(t)} = w_{hij} \cdot \frac{n_h}{n_h - 1} \cdot r_{hi}^{(t)}$. The bootstrap weights were then used for statistical analysis in the bootstrap samples.

Predictions of conditional mean, quartiles, and 95th percentile of the serum TCDD concentrations were plotted in raw scale versus age. Each value below the LOD was plotted using the average of its imputed values in five imputed data. All p-values are based on two-sided hypothesis tests. The statistical analyses were carried out using SAS,

version 9.1 (SAS Institute Inc., Cary, North Carolina), and Figure V.1 was created by R version 2.6.1 (R Development Core Team, Vienna, Austria).

V.3. Results

Table V.1 presents characteristics of the 251 Jackson/Calhoun UMDES participants with the 719 NHANES non-Hispanic white participants shown for comparison. The proportion of serum TCDD values below the LOD was 48 percent in the NHANES data compared to 21 percent in the Jackson/Calhoun data. The median LOD levels among the samples below the LOD were 1.1 ppt in the NHANES data, but 0.5 ppt in the Jackson/Calhoun data. The differences in the proportion of serum TCDD values below the LOD were due in part to larger serum specimens analyzed in Jackson/Calhoun (20 ml) than in NHANES (5-10 ml). The two populations were similar in BMI, BMI change in the last 12 months, pack-years smoking, income, education, and number of children breast fed (among females). However, the Jackson/Calhoun population was slightly older (p-value = 0.05) and had a smaller proportion of males (p-value = 0.04) than the NHANES population.

Table V.2 shows results of the five regression models with parameter and standard error estimates. Age was a strong positive predictor in all the five regression models (p-*value* < 0.01), and the age and sex interaction term was also significant in the mean, 25th percentile, median, and 95th percentile regressions. For example, for each 10-year increase in age, the mean serum TCDD concentrations were estimated to be increased by 60 percent ($e^{0.047*10 \text{ years}} = 1.60$) among females and by 34 percent ($10^{(0.047-0.018)*10 \text{ years}} = 1.34$) among males. The data source variable was not significant in the mean, 25th

percentile, median, or 95th percentile regressions, but had marginally positive significant effects in the 75th percentile (*p-value* = 0.07). This indicates that the Jackson/Calhoun population is similar to the NHANES non-Hispanic white population in the age- and sex-specific serum TCDD concentration.

Any age- (between ages 18 and 85 years) and sex-specific predicted mean, quartiles, and 95th percentile of background serum TCDD concentrations can be obtained for Jackson/Calhoun and for the NHANES from the regression results in Table V.2. For example, the predicted 95th percentile of serum TCDD concentrations measured in parts per trillion equals $\exp^{(1.139 + 0.031 \times (\text{age-50}) + 0.063 \times \text{sex} - 0.015 \times \text{sex} \times (\text{age-50}) + 0.239 \times \text{source})}$. The predicted mean and three quartiles can be obtained similarly. Table V.3 displays these estimates for 50 year old men and 50 year old women from Jackson/Calhoun and the NHANES as examples. For a 50 year old man (woman) in Jackson/Calhoun, the mean, 25th percentile, median , 75th percentile, and 95th percentile serum TCDD concentrations are estimated to be 1.1 (1.3), 0.6 (0.8), 1.1 (1.4), 1.8 (2.1), and 3.3 (3.1) ppt, respectively; and for a 50-year old man (woman) in the NHANES, the mean, 25th percentile, median, 75th percentile, and 95th percentile serum TCDD concentrations are estimated to be 1.1 (1.3), 0.6 (0.8), 1.1 (1.4), 2.2 (2.5), and 4.2 (4.0) ppt, respectively.

Figure V.1 compares the predicted mean, three quartiles, and 95th percentile serum TCDD values over age by sex between the NHANES and the UMDES reference populations. A circle represents an observed serum TCDD concentration above the LOD, and an "x" is the average of the five imputations for those below the LOD. The plots show that for people older than 75, the NHANES data improved the estimates in the Jackson/Calhoun population, especially among males (age- and sex-specific upper

percentiles among people older than 75 could not be fitted using only the Jackson/Calhoun data). In addition, the background serum TCDD concentrations increased with age and increased more steeply with age in females than in males in both data sources. Moreover, the plots show that the 75th and 95th percentile regression models on age and sex were fitted based on serum TCDD measures that were above the LOD for both the NHANES and the Jackson/Calhoun data sets, while the 25th percentile, mean, and median among young adults were estimated primarily based on the imputed values in the NHANES and the observed values above the LOD in Jackson/Calhoun.

V.4. Discussion

This study shows that the serum TCDD concentrations in non-Hispanic whites increased with age, and the rates of increase in the mean, 25th percentile, median, and 95th percentiles over age were greater among females than males. This difference is probably the results of a longer TCDD half-life among females than males because of higher percent body fat in females and the peak level of TCDD in the environment in the 1960's and 1970's. (Mibrath et al. 2009) As a result, the overall mean and percentiles of the background concentrations depend on the distribution of age and sex in the reference population, and it is not valid to compare the overall mean or percentiles of serum TCDD concentrations between populations that have different age and sex structures. Therefore, it is important to quantify the background levels of serum TCDD concentration by age and sex. For example, in comparisons to residents in Midland/Saginaw, we compared the serum TCDD concentrations to the background concentrations of people of the same age and sex to see whether the serum TCDD concentrations were elevated. (Garabrant et al.

2007) Quantile regression generalizes a single quantile estimate of serum TCDD concentrations to continuous conditional quantile estimates given age and sex. These age- and sex-adjusted quantile estimates provide better quantification of quantiles than the traditional method of calculating the population quantiles without adjusting for age or of adjusting for a limited number of age groups or strata.

We expected to see similar results for the Jackson/Calhoun data and the NHANES data because they both represented general populations who were not exposed to any known, unusual sources of dioxins. The present study shows that the effects of age and sex on the serum TCDD concentrations were not significantly different between the Jackson/Calhoun and the NHANES populations, and that the Jackson/Calhoun population was not significantly different from the NHANES population in the age- and sex- specific 25^{th} percentile, mean, median, and 95^{th} percentile, but was slightly lower than the NHANES population in the 75^{th} percentile. This implies that the Jackson/Calhoun population is similar to the NHANES population in age- and sex- specific serum TCDD concentration, and thus is a valid reference population for serum TCDD concentration for other predominantly white populations in Michigan. The marginally higher levels of age- and sex- specific 75^{th} percentile in the NHANES than in the Jackson/Calhoun can be explained as slightly larger variation of serum TCDD concentrations in the U.S. population than in the two counties in Michigan. This could be due to more heterogeneity of TCDD exposures among regional U.S. populations. However, the geographic information is not available in the publicly released NHANES data set, so that the geographic variation in serum TCDD concentration can not be accounted for in the

models. With data source indicator in the model, we allow for the effect of the different data source to be incorporated into the model.

Values below the limit of detection are common in studies of dioxin-like compounds. Simple ways of handling values below the LOD include imputing them with 0, LOD, LOD/2, and LOD/$\sqrt{2}$. (Hornung and Reed 1990) However, these imputation methods do not account for imputation uncertainty, and they depend on the blood sample volume and the LOD levels of the measurement methods. (In other words, the same serum sample analyzed by two different methods having different LODs would be assigned different values.) For studies with a low proportion of data below the LOD and low LOD levels, the estimation of conditional percentiles is less affected by how the LODs are imputed, especially for upper percentiles. However, for environmental contaminants for which the concentrations are near the LOD, a substantial proportion of the analytic results will be below the LOD. Lower percentiles and sometimes even median estimates in such data are more sensitive to the imputation methods used.

In the present study, multiple imputations based on a left-censored regression model using the observed TCDD measures and the LOD levels of the non-detects were employed to impute the values below the LOD to obtain multiple complete data sets, so that complete-data statistical methods (such as quantile regression) can be implemented. In the multiple imputations, we assumed that the serum TCDD concentrations followed a lognormal distribution, because the lognormal assumption appeared reasonable for the Jackson/Calhoun data with 79 percent of the data that were observed (above LOD). By concatenating the Jackson/Calhoun data with the NHANES data, we improved the imputation for the values below the LOD in the NHANES data by incorporating the

observed serum TCDD measures in the Jackson/Calhoun data. At the same time, inclusion of the NHANES data enhanced the estimates of the upper percentiles of serum TCDD values among older people in the Jackson/Calhoun population. The multiple imputation with the combined dataset has improved the percentile estimation in both data sources. This method can be applied in other environmental and public health studies where LOD is an issue and multiple sources of data are available. This article also provides an important example on how to incorporate the complex survey design information in every detail of statistical analysis in a population-based study.

The potential limitation of the multiple imputation approach is the assumption of lognormality. Although the lognormal assumption can be replaced by other statistical distributions, such as Gamma distribution or Weibull distribution according to some prior information, some distribution-free methods for handling values below the LOD, such as Schisterman's method (Schisterman et al. 2006), are of great interest. In using the bootstrap method for stratified multistage samples, we have modified the sample weights with the bootstrap weights. However, the further weight modifications such as nonresponse and poststratification adjustments are not feasible here because of the limited information from the sub-sample of the non-Hispanic white population in the NHANES data. We combined the UMDES and the NHANES data by concatenating the two data sets directly and normalizing their sample weights. In future work, other methods for combining multiple data sources such as Bayesian hierarchical methods will also be considered. (Raghunathan et al. 2007) Finally, we imputed the small fraction of missing covariates before the multiple imputation for values below the LOD to simplify

the imputation procedure. In the future, we plan to work on multiple imputation methods that simultaneously impute the missing covariates and values below the LOD.

Table V.1 Comparison of LOD and population-based demographics between Jackson/Calhoun, Michigan, 2005 and NHANES 2003-2004 populations

	NHANES (n=719) [†]	Jackson/Calhoun (n=251)	*p*-value [§]
Proportion of below LOD (amt serum)	48% (5-10 ml)	21% (20 ml)	---
Median LOD levels (range) [‡]	1.1 (0.4-3.1) ppt	0.5 (0.3-3.2) ppt	---
Mean age (range)	47.0 (18-85) yrs	49.9 (18-85) yrs	0.051
Mean BMI change in the last 12 months (SE)	0.2 (0.1) kg/m^2	-0.1 (0.2) kg/m^2	0.194
Mean BMI (SE)	27.7 (0.3) kg/m^2	28.7 (0.5) kg/m^2	0.101
Mean pack-yrs smoking (SE)	11.3 (0.6)	12.5 (1.4)	0.440
Mean No. of children breast-fed among women (SE)	0.8 (0.1)	1.0 (0.1)	0.198
Mean income (SE)	$ 52,000 (2,000)	$ 56,000 (2,000)	0.220
Sex (proportion of males)	47.6%	38.1%	0.035
Education (proportion of ≥ High School)	86.6%	86.2%	0.897

[†] Non-Hispanic white adults (excluding pregnant women) having serum TCDD measures in 2003-2004 NHANES

[‡] The median LOD levels (among the observations below LOD)

[§] *p*-values using F tests to compare the population-based demographics between the NHANES and Jackson/Calhoun populations

Table V.2 Results of linear and quantile regressions of log (serum TCDD concentration) in the combined data of Jackson/Calhoun, Michigan, 2005 and NHANES 2003-2004

Factor	Mean[†]	Q_1[‡]	Median[‡]	Q_3[‡]	95th percentile[‡]
Intercept	0.232 (0.069) ***	-0.176 (0.149)	0.346 (0.058) ***	0.726 (0.075) ***	1.139 (0.119) ***
Age[††]	0.047 (0.003) ***	0.054 (0.005) ***	0.048 (0.003) ***	0.036 (0.002) ***	0.031 (0.005) ***
Sex[§]	-0.183 (0.082) **	-0.273 (0.136) *	-0.223 (0.103) **	-0.126 (0.087)	0.063 (0.108)
Age[††]×sex[§]	-0.018 (0.004) ***	-0.025 (0.007) ***	-0.015 (0.005) **	-0.006 (0.005)	-0.015 (0.007) **
Source[‡‡]	0.051 (0.084)	-0.094 (0.169)	0.015 (0.100)	0.190 (0.099) *	0.239 (0.151)

Results are reported as estimate (standard error) p-value; *** $p \leq 0.01$; ** $p \leq 0.05$; * $p \leq 0.1$.

[†] The mean model was obtained by fitting a linear regression for complex survey data using SURVEYREG procedure in SAS.

[‡] The percentile models were fitted using quantile regressions for complex survey data using bootstrap method to calculate the standard errors. (Q_1: 25th percentile; Q_3: 75th percentile)

[††] Age minus 50 (years)

[§] Sex (females = 0, males = 1)

[‡‡] Data source (Jackson/Calhoun = 0, NHANES = 1)

Table V.3 Predicted mean, quartiles, and 95th percentile for a 50-year old person by sex

Units = ppt (lipids)	Mean	Q_1‡	Median	Q_3‡	95th %tile
50 year old woman in Jackson/Calhoun	1.3	0.8	1.4	2.1	3.1
50 year old man in Jackson/Calhoun	1.1	0.6	1.1	1.8	3.3
50 year old woman in NHANES	1.3	0.8	1.4	2.5	4.0
50 year old man in NHANES	1.1	0.6	1.1	2.2	4.2

‡ Q_1: 25th percentile, Q_3: 75th percentile

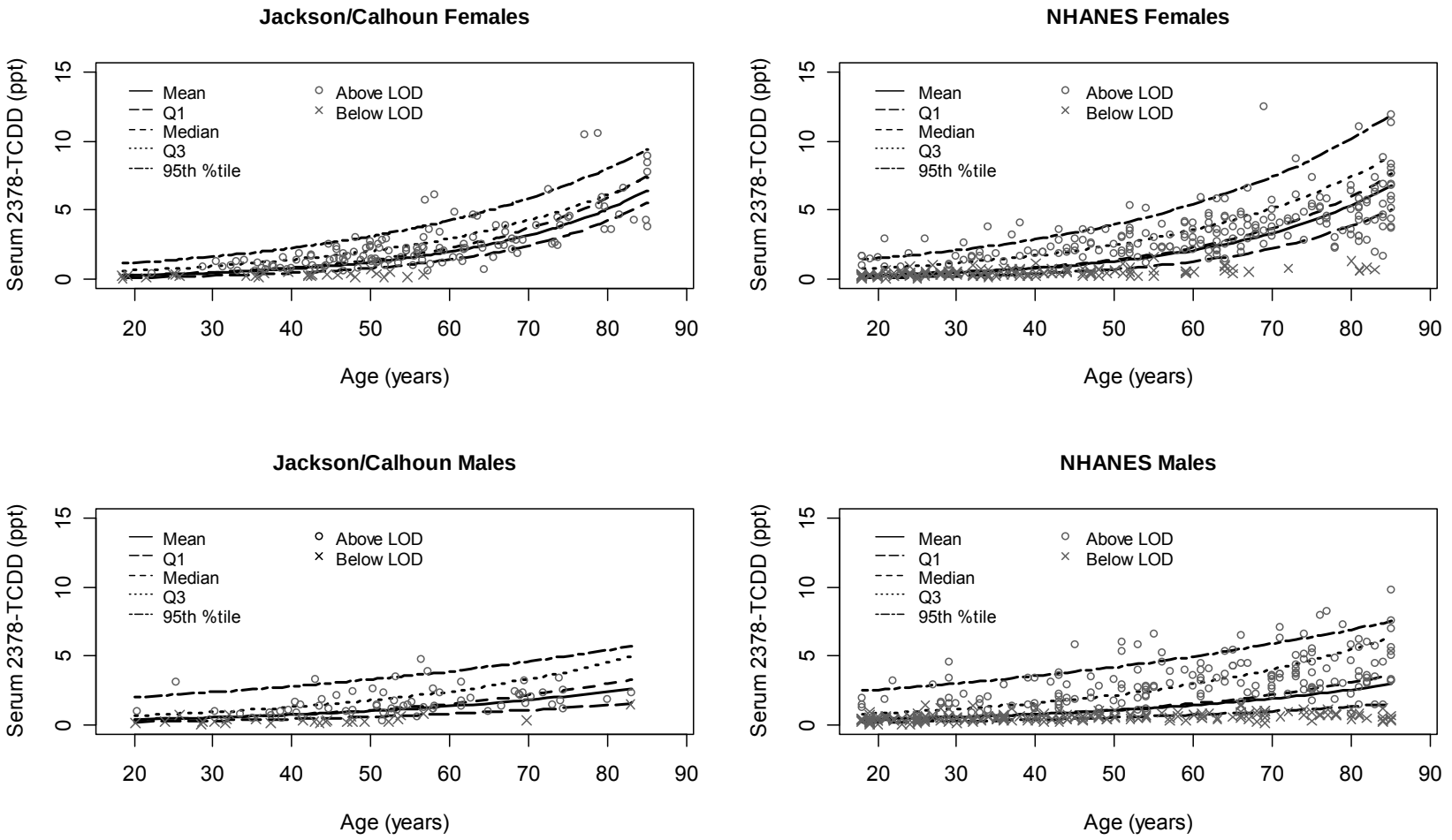

Figure V.1 Comparisons of predicted mean, quartiles, and 95th percentile of serum TCDD levels over age by sex between the Jackson/Calhoun, Michigan, 2005, and the NHANES 2003-2004 populations.

Chapter VI

CONCLUSION

The first part of the research shows that the Robust Bayesian model-based inference for finite population proportions and quantiles outperform the design-based estimators. By modeling the conditional distribution of the selection probabilities, the design mechanism is ignorable when the selection probability is a good summary of the design variables. This is often true in a one-stage unequal probability sampling design, such as sampling with strata or probability proportional to size sampling. By using the relationship between the survey outcome and the selection probabilities, the model-based estimators yield more efficient estimation of the population quantities than the design-based estimators. Compared to the parametric model-based estimators, the penalized spline regression model-based estimator automatically catches the potential nonlinear association between the survey outcome and the selection probabilities; this avoids the inefficiency because of model misspecification. When some sample units have very small selection probabilities, the penalized spline model-based estimators still perform well, while the design-based estimators are often very inefficient and the parametric model-based estimators are sensitive to these influential points in the regression model.

The Bayesian inference for a population summary of the survey outcome follows from the posterior predictive distribution of the values of the survey outcome in the non-sampled units given the values in the sample units and the selection probabilities.

By specifying noninformative prior distributions on the parameters in the super population distribution, the posterior distribution of the descriptive population quantities can be simulated by using MCMC simulations. The 95% credible intervals can then be easily calculated from the posterior distribution of the population quantities. In general, the CI calculated from the model-based estimators is shorter than the design-based estimators. The Bayesian penalized spline model-based predictive estimator for population proportions yields closer to the nominal level confidence coverage than the sample-weighted estimator or the generalized regression estimator. The Bayesian model-based predictive estimator for population quantiles based on a penalized spline regression model accounting for heteroscedastic errors often provide conservative but shorter credible interval than the design-based estimators. When sample size is small, the robust Bayesian model-based estimators have the most significant improvement over the design-based estimators in the confidence coverage, where the confidence intervals associated with the sample-weighted estimator leads to serious under-coverage because of the failure of the large sample assumption.

The robust Bayesian model-based inference can be extended to include additional auxiliary covariates by adding linear terms for these variables. In the future, I am going to study the robust Bayesian model-based inference for multistage sampling design with strata and clusters. The research on the variable selection method for multiply imputed data in Chapter IV provides a handy and useful approach for performing variable selection in multiple imputation. The "combine then select" (CS) method is an intuitive use of the multiple imputation combining rule in the variable selection procedure and it preserves the type I error. This CS method accounts for imputation uncertainty and is

less likely to incorrectly select variables into the model than competing stepwise selection methods currently used in epidemiological studies applied to multiply imputed data. Although in this article I only describe the modified stepwise variable selection method in the setting of multiply imputed data, I am interested in extending this to more advanced variable selection methods.

The multiple imputation method proposed in Chapter V is a promising method for imputing serum dioxin levels below the limit of detection (LOD). The traditional methods for dealing with the values below LOD include imputing them with 0, LOD, LOD/2, and LOD/$\sqrt{2}$, which do not account for imputation uncertainty, and they depend on the blood sample volume and the LOD levels of the measurement methods. The multiple imputation approach based on a left-censored regression model improves the imputation of the values below the LOD. The left-censored regression relates the unobserved serum dioxin levels below LOD with the observed serum dioxin concentrations above LOD by modeling the relationship between the observed serum dioxin concentration and the demographic variables. This method can be applied in other environmental and public health studies where LOD is an issue.

References

Agency for Toxic Substances and Disease Registry. (1998). *Toxicological Profile for Chlorinated Dibenzo-p-Dioxins*. Public Health Service, U.S. Department of Health and Human Services, Atlanta, GA.

Albert, J.H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of American Statistical Association,* **88**, 669-679.

Gelman A., Carlin J.B., Stern H.S., and Rubin D.B.. (2004). *Bayesian Data Analysis*. London: Chapman and Hall. Second edition.

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part 1, in *Foundations of Statistical Inference*, eds. V. P. Godambe and D. A. Sprott, Toronto: Holt, Rinehart and Winston, pp. 203–242.

Brand J.P.L. (1999). Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets, Enschede: Print Partners Ipskamp.

Chambers, R.L., Dorfman, A.H., and Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of 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., and Little, R.J.A. (2007). Bayesian penalized spline model-based estimation of the finite population proportion from probability-proportional-to-size samples. *Proceedings of the Joint Statistical Meetings,* 2969-2975.

Compumine. (2007). Re: analysis – Tax audit data mining. Feb. 2007. http://www.compumine.com/web/public/newsletter/20071/tax-audit-data-mining

Crainiceanu C.M, Ruppert D., Carroll R.J., Joshi A., and Goodner B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic error. *Journal of Computational and Graphical Statistics*, **16**, 265-288.

Crainiceanu C.M., Ruppert D., and Wand M. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, 14.

Duchesne, P. (2003). Estimation of a proportion with survey data. *Journal of Statistics Education*, **11**, 3.

Efron B., Tibshirani R., (1994) *An introduction to the bootstrap.* Chapman & Hall.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (width discussion). *Statistical Science*, **11**, 89-121.

Firth D. and Bennett K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, Series B*, **60**, 3-21.

Francisco, C.A. and Wayne, A.F. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, **19**, 454-469.

Garabrant D., Chen Q., Hong B., et al. (2007). Logistic regression models for high serum 2,3,7,8-TCDD concentrations in residents of Midland, Michigan, USA. *Organohalogen Compounds.* **69**, 2203-2206.

Garabrant, D., Franzblau, A., Lepkowski, J., et al. (2009). The University of Michigan Dioxin Exposure Study: Predictors of Human Serum Dioxin Concentrations in Midland and Saginaw, Michigan, *Environmental Health Perspectives*, **117**, 818-824.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **3**, 515-533.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, **33**, 350–374.

Heymans M.W., Buuren S.V., Knol D.L., et al., (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study, *BMC Medical Research Methodology,* **7**, 33.

Hornung R.W., Reed L.D. (1990) Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene,* **5**, 46-51.

Holmberg, A. (1998). A bootstrap approach to probability proportional-to-size sampling. *Proceedings of Section on Survey Research Methods, the Joint Statistical Meetings*. 378-383.

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

Isaki, C.T., Fuller, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, **77**, 89-96.

Keith L.H., Crummett W., Deegan J., et al. (1983) Principles of environmental analysis. *Analytical Chemistry,* **55**, 2210-2218.

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

Koenker R., (2005). *Quantile regression.* Econometric Society Monograph Series, Cambridge University Press.

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

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

Lehtonen, R., Särndal, C.-E., and Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, **7**, 649-673.

Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, **24**, 51-55.

Lepkowski J.M., Mosher W.D., Davis K.E., Groves R.M., Van Hoewyk J., and Willem J. (2006). *National Survey of Family Growth, Cycle 6: Sample design, imputation, and variance estimation.* National Center for Health Statistics. Vital Health Stat 2(142).

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

Little R.J.A. and Rubin D.B. (2002). *Statistical analyses with missing data.* New York: John Wiley.

Keith L.H., Crummett W., Deegan J., et al. (1983). Principles of environmental analysis. *Analytical Chemistry,* **55**, 2210-2218.

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

Lepkowski J., Olson K., Ward B., et al. (2006) Survey methodology in an environmental exposure study: methods, missing data, and inference. *Organohalogen Compounds,* **68**, 209-12.

Little R.J.A., Rubin D.B. (2002). *Statistical analyses with missing data.* New York: John Wiley.

Milbrath M.O., Wenger Y., Chang C.-W., et al. (2009). Apparent half-lives of dioxins, furans, and PCBs as a function of age, body fat, smoking status, and breastfeeding. *Environmental Health Perspectives,* **117**, 417-425

Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology,* **24**, 69-77.

National Center for Environmental Health. (2005). Third national report on human exposure to environmental chemicals. NCEH Pub No. 05-0570, 1-475. Atlanta, GA,

Department of Health and Human Services, Centers for Disease Control and Prevention.

Opsomer J.D., Claeskens G., Ranalli M.G., Kauermann G., and Breidt F.J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Series B*, **70**, 265-286.

Patterson D.G. Jr., Patterson D., Canady R., et al. (2004). Age specific dioxin TEQ reference range. *Organohalogen Compounds,* **66**, 2878-83.

Patterson D.G. Jr., Turner W.E., Caudill S.P., et al. (2008). Total TEQ reference range (PCDDs, PCDFs, cPCBs, mono-PCBs) for the US population 2001-2002. *Chemosphere,* **73**, 261-277.

Pfeffermann D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, **61**, 317-337.

Pfeffermann D. (1996). The use of sampling weights for survey data analysis. Statistical *Methods in Medical Research*, **5**, 239-261.

Philips D.L., Pirkle J.L., Burse V.W., Bernert Jr. J.T., Henderson LO, and Needham LL. (1989). Chlorinated hydrocarbon levels in human serum: effects of fasting and feeding. *Archives of Environmental Contamination and Toxicology,* **8**, 495-500.

Rao, J.N.K., Kovar, J.G., and Mantel, H.J. (1990). On estimating distribution function and quantile from survey data using auxiliary information. *Biometrika*, **77**, 365-375.

Raghunathan T.E., Lepkowski J.M., Van Hoewyk J., and Solenberger P. (2001). A Multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology.* **27:** 85-95.

Raghunathan T.E., Solenberger P.W., and Van Hoewyk J. IVEware: Imputation and Variance Estimation Software. (http://www.isr.umich.edu/src/smp/ive/). (Accessed September 17, 2008).

Raghunathan T.E., Xie D., Schenker N., et al. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of American Statistical Association,* **102**, 474-486.

Rothman KJ, Greenland S., and Lash TL. (2008). *Modern Epidemiology.* Lippincott Williams & Wilkins.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika,* **57**, 377-387.

Royall, R.M., and Cumberland, W.G. (1981). The finite-population linear regression estimator and estimators of its variance - an empirical study, *Journal of the American Statistical Association*, **76**, 924-930.

Rubin D.B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys: Comment. *Journal of the American Statistical Association*, **384**, 803-805.

Rubin D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression.* Cambridge, UK: Cambridge University Press.

Rust K.F. and Rao J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research,* 5: 283-310.

Särndal C.-E., Swensson B. and Wertman J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Schafer J.L., Khare M., and Ezzati-Rice T.M. (1993). Multiple imputation of missing data in NHANES III. *Proceedings of the Annual Research Conference*. 459–487. Washington, DC: Department of Commerce, Bureau of the Census.

Schenker N., Raghunathan T.E., Chiu P.-L., Makuc D.M., Zhang G., and Cohen A.J. Multiple imputation of missing income data in the National Health Interview Survey. (2006). *Journal of American Statistical Association.* **101:** 924-933.

Schisterman E.F., Vexler A., Whitcomb B.W., and Liu A. (2006). The limitations due to exposure limits for regression models. *American Journal of Epidemiology,* **163**, 374-383.

Shao J. and Wu C.F.J. (1989). A general theory for jackknife variance estimation. *Annuals of Statistics*, **17**, 1176-1197.

Sitter R.R. and Wu C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics and Probability Letters*, **52**, 353-358,

Smith, T.M.F. (1976). The foundations of survey sampling: A review (with discussion). *Journal of the Royal Statistical Society, Series A*, **139**, 183–204.

Smith, T.M.F. (1994). Sample surveys 1975–1990: An age of reconciliation? (with discussion). *International Statistical Review*, **62**, 5–34.

University of Michigan. (2008). University of Michigan Dioxin Exposure Study. Available at http://www.umdioxin.org. Assessed December 22, 2008

Van den Berg M., Birnbaum L.S., Denison M., et al. (2006). The 2005 World Health Organization reevaluation of human and Mammalian toxic equivalency factors for dioxins and dioxin-like compounds. *Toxicological Sciences*. **93:** 223-241.

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

Wittsiepe J., Schrey P., Ewers U., et al. (2000). Decrease of PCDD/F levels in human blood from Germany over the past ten years (1989-1998). *Chemosphere*, **40**, 1103-9.

Wood A.M., White I.R., Hillsdon M., et al. (2005). Comparison of imputation and modeling methods in the analysis of a physical activity trial with missing outcomes, *International Journal of Epidemiology*. **34**, 89-99.

Wood A.M., White I.R., and Royston P. (2008). How should variable selection be performed with multiply imputed data? *Statistics In Medicine.* **27:** 3227-3246.

Wood, S.N. (1994). Monotonic smoothing splines fitted by cross validation *SIAM Journal on Scientific Computing*, **15**, 1126-1133.

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

Wu, C. and Sitter, R.R. (2001). A model-calibration approach to using complex auxiliary information from survey data. *Journal of the American Statistical Association*, **96**, 185-193.

Yang X., Belin T.R. and Boscardin W.J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, **61:** 498-506.

Yates, F., and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, Series B,* **15**, 235-261.

Zheng, H. and Little, R.J.A. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, **19**, 99–117.

Zheng, H. and Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, **21**, 1-20.