Biometrika (2018), **105**, 1, *pp.* 91–102
Printed in Great Britain

doi: 10.1093/biomet/asx073
Advance Access publication 18 December 2017

Approximate Bayesian inference under informative sampling

By Z. WANG, J. K. KIM

Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.
wangzl@iastate.edu jkim@iastate.edu

AND S. YANG

Department of Statistics, North Carolina State University, 2311 Stinson Drive,
Campus Box 8203, Raleigh, North Carolina 27695, U.S.A.
syang24@ncsu.edu

Summary

Statistical inference with complex survey data is challenging because the sampling design can be informative, and ignoring it can produce misleading results. Current methods of Bayesian inference under complex sampling assume that the sampling design is noninformative for the specified model. In this paper, we propose a Bayesian approach which uses the sampling distribution of a summary statistic to derive the posterior distribution of the parameters of interest. Asymptotic properties of the method are investigated. It is directly applicable to combining information from two independent surveys and to calibration estimation in survey sampling. A simulation study confirms that it can provide valid estimation under informative sampling. We apply it to a measurement error problem using data from the Korean Longitudinal Study of Aging.

Some key words: Analytic inference; Complex sampling; Predictive inference; Pseudolikelihood.

1. Introduction

Survey data are frequently collected under complex sampling designs, which are often informative; that is, the distribution of the sample is different from that of the population (Sugden & Smith, 1984). In analytic inference, where the parameters of interest are those in the superpopulation model that the finite population follows, survey weights are incorporated to obtain valid inference under informative sampling. Under a parametric superpopulation model, the pseudo maximum likelihood method produces consistent estimators of parameters in the model. Korn & Graubard (1995), Skinner & Coker (1996), Chambers & Skinner (2003) and Fuller (2009) have given comprehensive overviews of this topic. Breidt et al. (2013) discussed testing for informativeness of the sampling designs.

Superpopulation modelling is also important in developing predictive inference in survey sampling (Firth & Bennett, 1998; Valliant et al., 2000). By modelling and predicting the nonsampled part of the finite population, inference about the finite population quantities can be improved. Addressing the uncertainty associated with estimated model parameters in the predictive approach is an important problem in survey sampling.

Bayesian approaches are widely used to handle complex problems. In a Bayesian approach, a sample distribution is specified for the data, a prior distribution is specified for parameters, and inferences are based on the posterior distribution for the parameters. However, when applied to survey sampling, Bayesian methods implicitly assume that the sampling design is noninformative. Ignoring the sampling design can cause selection bias (Pfeffermann & Sverchkov, 1999), which raises the question of how to take sampling information into account in the Bayesian approach. Little (2004, 2012) advocated using frequentist methods to specify an analysis model, and using Bayesian methods for inference under this model, leading to the so-called calibrated Bayesian approach. In this approach, the posterior means and credible intervals are calibrated to design-based estimates and confidence intervals.

In this paper, we develop a new Bayesian method for informative sampling, which requires only a straightforward form of the data model and whose implementation involves simple computations. To achieve this we specify the data model as the sampling distribution of a summary statistic, where the design features are incorporated using design-based point and variance estimators. This differs from the usual Bayesian approach, and we do not specify the full sample-data likelihood, which is infeasible under informative sampling. Prior knowledge of the population parameters can be incorporated into the prior distribution. For example, if we know some population parameter values from other sources, we can use informative priors concentrated at the known values. This flexibility can be used to incorporate several sources of information sequentially; see § 4 for examples. Furthermore, the proposed method is calibrated Bayesian in the sense that it leads to valid frequentist confidence intervals under noninformative priors.

2. Basic idea

Suppose that the finite population $Y_N = (y_1, \ldots, y_N)$ is a random sample of size N from a superpopulation model ζ with density $f(y;\theta)$, where $\theta \in \Theta$ and Θ is the parameter space. Assume that the first n elements are sampled using a probability sampling design. Let Y_N be partitioned into two parts, $Y_N = (Y_n, Y_{N-n})$, where $Y_n = (y_1, \ldots, y_n)$ and $Y_{N-n} = (y_{n+1}, \ldots, y_N)$ are the sampled and nonsampled parts of Y_N, respectively. The finite population quantity of interest is $Q_N = Q(Y_N)$, such as $Q(Y_N) = N^{-1}\sum_{i=1}^N y_i$. Classical Bayesian inference assumes that the sampling mechanism is noninformative under the model ζ and generates posterior draws of Q_N as described in the following steps.

Step 1. Generate

$$\theta^* \sim p(\theta \mid Y_n) \propto L(\theta; Y_n)\pi(\theta), \tag{1}$$

where $p(\theta \mid Y_n)$ is the posterior distribution of θ given Y_n, $L(\theta; Y_n)$ is the likelihood function for θ, and $\pi(\theta)$ is a prior distribution. Throughout the paper, for simplicity we omit the dependence of $\pi(\theta)$ on hyperparameters. Under noninformative sampling, $L(\theta; Y_n) = \prod_{i=1}^n f(y_i;\theta)$.

Step 2. Generate y_i^* from $f(y_i;\theta^*)$ $(i = n+1, \ldots, N)$, where θ^* is generated in Step 1.

Step 3. A posterior value of Q_N given Y_n is computed as $Q_N^* = Q(Y_n, Y_{N-n}^*)$, where $Y_{N-n}^* = (y_{n+1}^*, \ldots, y_N^*)$.

This procedure is not directly applicable when the sampling mechanism is informative. Specifically, we cannot directly compute the likelihood function $L(\theta; Y_n)$ in (1) from the population density $f(y;\theta)$. In this case, $L(\theta; Y_n)$ is different from $\prod_{i=1}^n f(y_i;\theta)$; see Pfeffermann et al. (1998) and Pfeffermann & Sverchkov (1999). To circumvent this difficulty, one often augments the model by incorporating the design information so that the sampling design becomes non-informative under the augmented model; however, design information is not always available. Furthermore, correct specification of the augmented model can be challenging, and may not be of interest to the subject matter experts who use the survey data.

We consider an alternative that does not use the likelihood $L(\theta; Y_n)$ in (1). Let $\hat{\theta}$ be a direct estimator of θ which satisfies

$$\hat{\theta} = \theta + o_{\mathrm{p}}(1), \tag{2}$$

where the reference distribution is the joint distribution of the superpopulation model ζ and the sampling mechanism. We then use the sampling distribution of $\hat{\theta}$ to obtain an approximate posterior distribution; that is, instead of using (1), we use

$$\theta^* \sim p(\theta \mid \hat{\theta}) \propto g_1(\hat{\theta} \mid \theta)\pi(\theta), \tag{3}$$

where $p(\theta \mid \hat{\theta})$ is the posterior distribution of θ given $\hat{\theta}$, and $g_1(\hat{\theta} \mid \theta)$ is the sampling distribution of $\hat{\theta}$. The proposed Bayesian method is valid in the sense that the coverage properties hold for the resulting posterior credible set (Monahan & Boos, 1992) when there exists a large-sample approximation of the sampling distribution of $\hat{\theta}$ given θ (Fuller, 2009).

One way to obtain $\hat{\theta}$ satisfying (2) is to solve the following pseudo-score equation for θ:

$$\hat{S}(\theta) = \sum_{i=1}^{n} w_i S(\theta; y_i) = 0, \tag{4}$$

where w_i is the sampling weight for element i and $S(\theta; y) = \partial \log f(y; \theta)/\partial\theta$ is the score function of θ. The solution to (4), called the pseudo maximum likelihood estimator of θ, is consistent under both informative and noninformative sampling; see, for example, Godambe & Thompson (1986), Pfeffermann (1993), Korn & Graubard (1999, Ch. 3), Chambers & Skinner (2003, Ch. 2) and Fuller (2009, § 6.5).

Under regularity conditions (Fuller, 2009, § 1.3.2), it can be shown that

$$\hat{V}(\hat{\theta})^{-1/2}(\hat{\theta} - \theta) \to \mathcal{N}(0, I) \tag{5}$$

in distribution as $n \to \infty$, where $\mathcal{N}(0, I)$ is the normal distribution with mean zero and an identity covariance matrix, and $\hat{V}(\hat{\theta})$ is a design-consistent variance estimator of $\hat{\theta}$, with the reference distribution being the joint distribution of the superpopulation model ζ and the sampling mechanism. Thus, for a large sample, the normal distribution of $\hat{\theta}$ in (5) can be used to approximate the sampling distribution $g_1(\hat{\theta} \mid \theta)$ in (3). In general, the posterior mean of θ, conditional on the pseudo maximum likelihood estimator $\hat{\theta}$, may be less efficient than that obtained from a full Bayesian approach, conditional on full sample data. However, the latter is not tractable under informative sampling.

Once posterior values of θ are generated from (3), Steps 2 and 3 in the classical Bayesian method can be used with modification. That is, we obtain $y_i^* \sim f(y_i; \theta^*)$ for $i = 1, \ldots, N$ and compute $Q_N^* = Q(Y_N^*)$ as a posterior value of Q_N, where θ^* is generated from the posterior distribution of θ and $Y_N^* = (y_1^*, \ldots, y_N^*)$. We generate synthetic values for the whole population, similar in spirit to the synthetic population approach considered by Dong et al. (2014). Bayesian inference for Q_N can be obtained from the posterior distribution of Q_N^*, but the result is conservative unless the sampling fraction is negligible.

Remark 1. In (3), the sampling variance of $\hat{\theta}$ is consistently estimated by the standard sandwich formula (Binder, 1983). Instead of using (3), we can consider

$$\theta^* \sim p\{\theta \mid \hat{S}(\theta)\} \propto g_2\{\hat{S}(\theta) \mid \theta\}\pi(\theta), \tag{6}$$

where $p\{\theta \mid \hat{S}(\theta)\}$ is the posterior density of θ conditional on $\hat{S}(\theta)$, and $g_2\{\hat{S}(\theta) \mid \theta\}$ is the sampling distribution of the pseudo-score function $\hat{S}(\theta)$ in (4). Assuming that $\hat{S}(\theta)$ follows an asymptotic normal distribution, we can approximate $g_2\{\hat{S}(\theta) \mid \theta\}$ by

$$\hat{g}_2\{\hat{S}(\theta) \mid \theta\} = \exp\left\{-\frac{1}{2}\hat{S}(\theta)^{\mathrm{T}}\hat{V}_{ss}(\theta)^{-1}\hat{S}(\theta) - \frac{1}{2}\log\left|(2\pi)\hat{V}_{ss}(\theta)\right|\right\},$$

where $\hat{V}_{ss}(\theta)$ is a design-consistent estimator of the sampling variance of $\hat{S}(\theta)$. The density $\hat{g}_2\{\hat{S}(\theta) \mid \theta\}$ is easier to compute than $g_1(\hat{\theta} \mid \theta)$, since it does not involve Taylor linearization. However, the posterior distribution of θ in (6) is typically not log-concave. In this case, we can use the adaptive rejection Metropolis sampling algorithm within the Gibbs sampler (Gilks et al., 1995) to generate posterior values for θ. The detailed procedure is provided in the Supplementary Material.

3. SEMIPARAMETRIC BAYESIAN METHOD

The method proposed in § 2 is based on the assumption of a parametric superpopulation model. We now extend it to a general class of parameters without assuming a parametric model. Suppose that we are interested in estimating a finite population quantity θ, satisfying $\sum_{i=1}^{N} U(\theta; y_i) = 0$ where $U(\theta; y)$ is an estimating function for θ. For example, if $U(\theta; y) = I(y < 1) - \theta$, then $\theta = N^{-1}\sum_{i=1}^{N} I(y_i < 1)$, where $I(y < 1) = 1$ if $y < 1$ and 0 otherwise.

A consistent estimator $\hat{\theta}$ can be obtained by solving the following estimating equation for θ:

$$\hat{U}(\theta) = \sum_{i=1}^{n} w_i U(\theta; y_i) = 0. \tag{7}$$

We assume that the solution to (7) exists uniquely almost surely, and that a design-consistent estimator of var$\{\hat{U}(\theta)\}$ is available for a fixed θ. If $w_i = 1/\pi_i$, where π_i is the first-order inclusion probability of element i, the Horvitz–Thompson variance estimator (Horvitz & Thompson, 1952) is

$$\hat{V}\{\hat{U}(\theta)\} = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}\pi_i\pi_j} U(\theta; y_i)U(\theta; y_j)^{\mathrm{T}},$$

where π_{ij} is the second-order inclusion probability of the ith and jth elements.

Under certain regularity conditions, the estimating function $\hat{U}(\theta)$ in (7) has an asymptotic normal distribution; that is, $\hat{V}\{\hat{U}(\theta)\}^{-1/2}\hat{U}(\theta) \to \mathcal{N}(0, I)$ in distribution in the neighbourhood of θ_0 as $n \to \infty$ and $N \to \infty$, where θ_0 is the limit of the finite population quantity as $N \to \infty$. Therefore, we can construct a pseudolikelihood function of θ as

$$g\{\hat{U}(\theta) \mid \theta\} = \exp\left[-\frac{1}{2}\hat{U}(\theta)^{\mathrm{T}}\hat{V}\{\hat{U}(\theta)\}^{-1}\hat{U}(\theta) - \frac{1}{2}\log\left|(2\pi)\hat{V}\{\hat{U}(\theta)\}\right|\right].$$

With this pseudolikelihood function, the approximate posterior distribution of θ becomes

$$p\{\theta \mid \hat{U}(\theta)\} \propto g\{\hat{U}(\theta) \mid \theta\}\pi(\theta). \tag{8}$$

To generate posterior draws from (8), we use the following two-step method.

Step 1. Generate η^* from a normal distribution with mean zero and covariance matrix $\hat{V}\{\hat{U}(\theta)\}$ evaluated at $\theta = \hat{\theta}$.

Step 2. Given η^* generated from Step 1, obtain θ^* by solving $\hat{U}(\theta) = \eta^*$ for θ.

This Bayesian method is semiparametric and is attractive for two reasons. First, unlike the direct posterior distribution in (3), the sampling distribution of $\hat{U}(\theta)$, namely $g\{\hat{U}(\theta) \mid \theta\}$ in (8), does not involve Taylor linearization. Thus, it is more straightforward to generate samples from (8) than from (3). Second, the proposed method does not require parametric model assumptions for the superpopulation model; only the sampling distribution of a consistent estimator of θ is needed, similar to the Bayesian generalized method of moments of Yin (2009).

Remark 2. If the model is over-identified in the sense that the dimension of $U(\theta)$ is larger than that of θ, we may not obtain a unique solution to $\hat{U}(\theta) = \eta^*$, so we cannot directly apply the above two-step method for generating posterior values. In this case, we generate posterior values directly from (8), which may involve the adaptive rejection Metropolis sampling algorithm discussed in Remark 1.

The following theorem presents an asymptotic property of the proposed method.

Theorem 1. *Under the regularity conditions described in the Supplementary Material, conditional on the full sample data,*

$$\sup_{\theta\in\Theta} \left| p\{\theta \mid \hat{U}(\theta)\} - \phi_{\hat{\theta},\hat{V}(\hat{\theta})}(\theta) \right| \to 0 \tag{9}$$

as $n \to \infty$ almost surely, where Θ is the feasible region for θ, $p\{\theta \mid \hat{U}(\theta)\}$ is given in (8)*, $\phi_{\mu,\Sigma}(x)$ is the normal density function with mean μ and covariance matrix Σ, $\hat{\theta}$ is a consistent estimator obtained by solving* (7)*, and $\hat{V}(\hat{\theta})$ is a consistent estimator of* $\mathrm{var}(\hat{\theta})$.

Theorem 1 is a special case of the Bernstein–von Mises theorem (van der Vaart, 2000, § 10.2) in survey sampling, and its proof is given in the Supplementary Material. According to Theorem 1, the credible region for θ constructed from the posterior distribution is asymptotically equivalent to the frequentist one obtained based on the normal distribution $\phi_{\hat{\theta},\hat{V}(\hat{\theta})}(\theta)$. Therefore, the Bayesian inference is calibrated to the corresponding frequentist inference asymptotically. The consistency of the Bayesian estimator follows directly from (9) because $\hat{V}(\hat{\theta})$ converges to zero in probability.

4. Examples

4·1. *Estimating the population mean under informative sampling*

Suppose that a probability sample consists of n elements of (x_i, y_i) selected from a finite population of size N. Let $\theta = (\theta_1, \theta_2)$ be a finite population quantity, where $\theta_1 = N^{-1}\sum_{i=1}^{N} x_i$

and $\theta_2 = N^{-1}\sum_{i=1}^{N} y_i$. In particular, we are interested in estimating θ_2. Let $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ be the design-based estimator of θ, and let

$$\hat{V} = \begin{pmatrix} \hat{V}_{xx} & \hat{V}_{xy} \\ \hat{V}_{xy}^{\mathrm{T}} & \hat{V}_{yy} \end{pmatrix}$$

be the corresponding design-consistent variance estimator. In the proposed Bayesian framework, the pseudolikelihood of θ is the sampling distribution of the sample estimator $\hat{\theta}$, i.e., $g(\hat{\theta} \mid \theta)$, which is approximated by a normal distribution with mean θ and covariance matrix $\hat{V}$. Let the prior distribution $\pi(\theta)$ be normal with mean $\mu = (\mu_x, \mu_y)$ and covariance matrix

$$\Lambda = \begin{pmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{xy}^{\mathrm{T}} & \Lambda_{yy} \end{pmatrix}.$$

Then the posterior distribution of θ is normal with mean $\Gamma(\hat{V}^{-1}\hat{\theta} + \Lambda^{-1}\mu)$ and covariance matrix $\Gamma = (\hat{V}^{-1} + \Lambda^{-1})^{-1}$. To derive the posterior distribution of θ_2 given $\hat{\theta}$, we consider two cases. In Case I, where θ_1 and θ_2 are unknown, we have no prior information on the population parameters, so we consider a flat prior with the diagonal elements of Λ diverging to ∞ and with other hyperparameters taking any finite values. Then we can show that the posterior distribution $p(\theta_2 \mid \hat{\theta})$ is normal with mean $\hat{\theta}_2$ and variance $\hat{V}_{yy}$. In Case II, where θ_1 is known, we consider a degenerate prior for θ_1 with a point mass on $\mu_x = \theta_1$ and a flat prior for θ_2; that is, we can set $\mu_x = \theta_1$, $\Lambda_{xx} = 0$ and $\Lambda_{xy} = 0$, and let the diagonal elements of Λ_{yy} diverge to ∞ and other hyperparameters take any finite values. Then we can show that $p(\theta_2 \mid \hat{\theta})$ is a normal distribution with mean $\hat{\theta}_2 - \hat{V}_{xy}^{\mathrm{T}}\hat{V}_{xx}^{-1}(\hat{\theta}_1 - \theta_1)$ and variance $\hat{V}_{yy} - \hat{V}_{xy}^{\mathrm{T}}\hat{V}_{xx}^{-1}\hat{V}_{xy}$. The posterior mean of θ_2 given $(\hat{\theta}_1, \hat{\theta}_2)$ is equal to the optimal regression estimator of θ_2 (Rao, 1994). Therefore the proposed Bayesian inference is calibrated to the frequentist inference asymptotically.

4·2. *Estimating the population mean by combining information from two surveys*

How to combine information from several sources from the same population is an important research problem in survey sampling (Zieschang, 1990; Singh et al., 2005; Merkouris, 2010; Kim & Rao, 2012). Continuing with the set-up in § 4·1, suppose that samples A and B are independently selected from the same finite population. The variable x_i is observed in sample A and (x_i, y_i) is observed in sample B. We now use our method to combine information from the two samples to make inference about θ_2. Let $\hat{\theta}_{1,\mathrm{A}}$ and $\hat{V}_{xx,\mathrm{A}}$ be design-consistent estimators of θ_1 and its variance, respectively, obtained from the observations in sample A. The sampling density of θ_1, denoted by $g(\hat{\theta}_{1,\mathrm{A}} \mid \theta_1)$, is approximated by a normal density with mean θ_1 and covariance matrix $\hat{V}_{xx,\mathrm{A}}$. Because there is no prior information on the population parameter, we specify a flat prior for θ_1. By Bayes' rule, the posterior density $p(\theta_1 \mid \hat{\theta}_{1,\mathrm{A}})$ is normal with mean $\hat{\theta}_{1,\mathrm{A}}$ and variance $\hat{V}_{xx,\mathrm{A}}$. From sample A, we update our knowledge of θ_1, and we will incorporate this information as the prior distribution in the analysis of sample B. Based on sample B, let $\hat{\theta}_{\mathrm{B}} = (\hat{\theta}_{1,\mathrm{B}}, \hat{\theta}_{2,\mathrm{B}})$ be a design-consistent estimator of θ, and let

$$\hat{V} = \begin{pmatrix} \hat{V}_{xx,\mathrm{B}} & \hat{V}_{xy,\mathrm{B}} \\ \hat{V}_{xy,\mathrm{B}}^{\mathrm{T}} & \hat{V}_{yy,\mathrm{B}} \end{pmatrix}$$

be the corresponding design-consistent variance estimator of $\hat{\theta}_{\mathrm{B}}$.

The pseudolikelihood of θ based on sample B is the sampling distribution of the sample estimator $\hat{\theta}_{\mathrm{B}}$, i.e., $g(\hat{\theta}_{\mathrm{B}} \mid \theta)$, which is approximately normal with mean θ and covariance matrix $\hat{V}$. The prior distribution is informative, updated by the knowledge from sample A. Specifically, the prior distribution of θ_1 is normal with mean $\hat{\theta}_{1,\mathrm{A}}$ and variance $\hat{V}_{xx,\mathrm{A}}$, and the prior distribution of θ_2 is flat. After some algebra, the posterior distribution $p(\theta_2 \mid \hat{\theta}_{1,\mathrm{A}}, \hat{\theta}_{1,\mathrm{B}}, \hat{\theta}_{2,\mathrm{B}})$ can be shown to be normal with mean

$$\mu_p = \hat{\theta}_{2,\mathrm{B}} + \hat{V}^{\mathrm{T}}_{xy,\mathrm{B}} \hat{V}^{-1}_{xx,\mathrm{B}} (\hat{V}^{-1}_{xx,\mathrm{A}} + \hat{V}^{-1}_{xx,\mathrm{B}})^{-1} \hat{V}^{-1}_{xx,\mathrm{A}} (\hat{\theta}_{1,\mathrm{A}} - \hat{\theta}_{1,\mathrm{B}}) \tag{10}$$

and covariance matrix

$$\Sigma_p = \{\hat{V}^{-1}_{yy\cdot x,\mathrm{B}} - \hat{V}^{-1}_{yy\cdot x,\mathrm{B}} \hat{V}^{\mathrm{T}}_{xy,\mathrm{B}} \hat{V}^{-1}_{xx,\mathrm{B}} (\hat{V}^{-1}_{xx\cdot y,\mathrm{B}} + \hat{V}^{-1}_{xx,\mathrm{A}})^{-1} \hat{V}^{-1}_{xx,\mathrm{B}} \hat{V}_{xy,\mathrm{B}} \hat{V}^{-1}_{yy\cdot x,\mathrm{B}}\}^{-1},$$

where $\hat{V}_{xx\cdot y,\mathrm{B}} = \hat{V}_{xx,\mathrm{B}} - \hat{V}_{xy,\mathrm{B}} \hat{V}^{-1}_{yy,\mathrm{B}} \hat{V}^{\mathrm{T}}_{xy,\mathrm{B}}$ and $\hat{V}_{yy\cdot x,\mathrm{B}} = \hat{V}_{yy,\mathrm{B}} - \hat{V}^{\mathrm{T}}_{xy,\mathrm{B}} \hat{V}^{-1}_{xx,\mathrm{B}} \hat{V}_{xy,\mathrm{B}}$. The posterior mean (10) is equal to the generalized least squares estimator (3.3.46) in Fuller (2009). It also agrees with the optimal regression estimator (3.27) of Hidiroglou (2001), so the proposed Bayesian inference is calibrated to the frequentist inference.

The merits of our approach are that the inference can be made directly based on posterior draws and the computation is greatly simplified.

5. Simulation study

To evaluate the finite-sample performance of the proposed Bayesian method, we conducted a simulation study. The finite population $\{(x_i, y_i) : i = 1, \ldots, N\}$ is generated from

$$x_i \sim \mathcal{N}(\mu_x, \sigma_x^2), \tag{11}$$

$$y_i \mid x_i \sim \mathrm{Ber}(p_i), \tag{12}$$

where $N = 5000$, $(\mu_x, \sigma_x) = (5, 1)$, $\mathrm{logit}\, p_i = -4 + x_i$ and $\mathrm{logit}\, x = \log(x) - \log(1 - x)$. In addition, we generate z_i by $z_i \mid (x_i, y_i) = \min\{\log(1 + |x_i + y_i + e_i|), 2{\cdot}5\}$, with $e_i \sim \mathrm{Ex}(2)$. Let $\theta_1 = N^{-1} \sum_{i=1}^N x_i$ and $\theta_2 = N^{-1} \sum_{i=1}^N y_i$ be two finite population quantities. The goal is to make inference for θ_2.

We select a sample of size n from each finite population, using probability-proportional-to-size sampling with size measure z_i, and we consider $n = 50$ and 500.

To incorporate information from x_i, we use the estimation equation

$$\hat{U}(\theta) = \sum_{i=1}^{n} \frac{1}{\pi_i} \{(x_i, y_i) - \theta\} = 0, \tag{13}$$

where $\theta = (\theta_1, \theta_2)$ and $\pi_i = n z_i / (\sum_{k=1}^N z_k)$. The solution $\hat{\theta}$ to (13) is the Hájek estimator (Fuller, 2009, § 1.3.3) of θ. The design variance of $\hat{U}(\theta)$ can be estimated by $\hat{V}\{\hat{U}(\theta)\} = \{n/(n-1)\} \sum_{i=1}^n (u_i - \bar{u})^{\otimes 2}$, where $u_i = \{(x_i, y_i) - \hat{\theta}\}/\pi_i$, $\bar{u} = \sum_{i=1}^n u_i / n$, and $B^{\otimes 2} = BB^{\mathrm{T}}$.

In this simulation, we consider two cases.

Case I. Only the sample observations (x_i, y_i) $(i = 1, \ldots, n)$ are available.

Case II. In addition to the sample observations, the finite population mean θ_1 is available.

We compare the following four methods to make inference for θ_2.

Method 1: the frequentist method. For Case I we use the Hájek estimator, that is, the one that solves (13). The regression estimator using θ_1 is used for Case II. For both cases, the 95% confidence interval is obtained as $\hat{\theta}_2 \pm 1{\cdot}96\hat{V}(\hat{\theta}_2)$, where $\hat{V}(\hat{\theta}_2)$ is a design-consistent variance estimator of $\hat{\theta}_2$.

Method 2: the proposed semiparametric Bayesian method presented in § 3. To be specific,

$$\theta^* \sim p\{\theta \mid \hat{U}(\theta)\} \propto g\{\hat{U}(\theta) \mid \theta\}\pi(\theta),$$

where $g\{\hat{U}(\theta) \mid \theta\}$ is approximated by a normal density function with mean zero and covariance $\hat{V}\{\hat{U}(\theta)\}$. The priors discussed in § 4·1 are used for both cases, and the adaptive rejection Metropolis sampling algorithm within the Gibbs sampler is employed in the implementation.

Method 3: the parametric Bayesian method. In this approach we use the parametric model for $f(x, y; \eta) = f(x; \mu_x, \sigma_x)f(y \mid x; \beta_0, \beta_1)$, where $\eta = (\mu_x, \sigma_x, \beta_0, \beta_1)$, and $f(x; \mu_x, \sigma_x)$ and $f(y \mid x; \beta_0, \beta_1)$ are the density functions for (11) and (12). A flat prior for η is used for Case I, and a degenerate prior $\pi(\eta) \propto I_{\{\theta_2\}}(\mu_x)$ is used for Case II, where $I_{\{\theta_2\}}(\mu_x) = 1$ if $\mu_x = \theta_2$ and 0 otherwise. Posterior values of η are generated from (3), using the asymptotic normality of the pseudo maximum likelihood estimator of η. Once η^* is obtained, we generate (x_i^*, y_i^*) from $f(x, y; \eta^*)$ and then compute $\theta_2^* = N^{-1}\sum_{i=1}^{N} y_i^*$.

Method 4: the Bayesian penalized spline predictive method of Chen et al. (2010). The suggested spline model is

$$\Phi^{-1}\{E(y_i \mid \alpha, b, \pi_i)\} = \alpha_0 + \alpha_1\pi_i + \sum_{l=1}^{15} b_l(\pi_i - k_l)_+ \quad (i = 1, \ldots, N),$$

where $\Phi^{-1}(\cdot)$ is the standard normal inverse cumulative distribution function, $\alpha = (\alpha_0, \alpha_1)$, $b = (b_1, \ldots, b_{15})$ with $b_l \sim \mathcal{N}(0, \tau^2)$ for $l = 1, \ldots, 15$, the constants k_l are preselected knots, and $(x)_+ = x$ if $x \geqslant 0$ and 0 otherwise. In this simulation, the fixed knots are chosen such that $m_\pi < k_1 < \cdots < k_{15} < M_\pi$, and they are equally spaced, where m_π and M_π are the minimum and maximum values of the first-order inclusion probabilities in the sample. Flat priors for α and τ^2 are used. This method requires that the first-order inclusion probability be available throughout the finite population, which is not always the case in practice.

The root mean squared error of the point estimators and the coverage rate for 95% credible interval estimates are computed for each case from 2000 Monte Carlo samples. Table 1 shows that the proposed method performs similarly to the frequentist approach when the sample size is large. In terms of root mean squared error, the Bayesian penalized spline predictive method is most efficient, as it is based on unweighted analysis using first-order inclusion probabilities for the finite population. However, the coverage rates for this method are 94·3% and 93·9% for $n = 50$ and $n = 500$, respectively. Our proposed methods show better coverage properties when the sample size is large. The parametric Bayesian method provides conservative inference for θ_2. The point estimator is more efficient in Case II than in Case I, because it uses extra information on θ_1. Unfortunately, the current Bayesian penalized spline predictive method cannot incorporate the external information on θ_1 in the Bayesian model.

Table 1. *Root mean squared error and coverage rate obtained from Monte Carlo simulations using a frequentist approach, the semiparametric and parametric Bayesian methods, and the Bayesian penalized spline predictive method*

		Case I		Case II	
Sample size	Method	RMSE (×100)	CR (%)	RMSE (×100)	CR (%)
	Frequentist	6·60	94·1	6·12	94·6
$n = 50$	Semiparametric Bayesian	6·60	93·8	6·11	93·5
	Parametric Bayesian	6·62	95·3	6·18	95·0
	BPSP	5·80	94·3	N/A	N/A
	Frequentist	2·11	95·3	1·94	95·1
$n = 500$	Semiparametric Bayesian	2·11	95·2	1·95	95·2
	Parametric Bayesian	2·10	96·0	1·94	96·1
	BPSP	1·80	93·9	N/A	N/A

RMSE, root mean squared error; CR, coverage rate; BPSP, Bayesian penalized spline predictive method; N/A, not available.

6. Application

We present an analysis of data from the 2006 Korean Longitudinal Study of Aging, conducted by the Korean Labor Institute for South Korean citizens aged 45 or over. Details about the study can be found at http://www.kli.re.kr/klosa/en/about/introduce.jsp. The data include a primary sample of size $n_1 = 9842$ and a validation sample randomly selected from the primary sample, with sample size $n_2 = 505$. For the primary sample, self-reported body mass index z, age x_2, and a binary outcome y indicating hypertension status are obtained. For the validation sample, in addition to z, x_2 and y, a physical measure of body mass index, x_1, is reported.

The goal is to investigate the relationship between hypertension and two factors, body mass index and age. Specifically, we are interested in estimating the parameters in the following logistic model for y:

$$\text{logit}\,\text{pr}(y = 1 \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Since we do not observe x_1 in the primary sample, we can use its surrogate z by incorporating the Berkson (1950) error model

$$x_1 = \alpha_0 + \alpha_1 z + e, \quad e \sim \mathcal{N}(0, \sigma^2). \tag{14}$$

Figure 1 displays the physical measurements and self-reported values of body mass index, and it shows that (14) is reasonable.

The surrogate assumption implies $f(y \mid x_1, x_2, z) = f(y \mid x_1, x_2)$ and, therefore,

$$f(x_1 \mid y, x_2, z) \propto f_1(y \mid x_1, x_2; \theta_1) f_2(x_1 \mid z; \theta_2),$$

where $\theta_1 = (\beta_0, \beta_1, \beta_2)$ and $\theta_2 = (\alpha_0, \alpha_1, \sigma^2)$.

To apply the proposed Bayesian method, we use a flat prior for $\theta = (\theta_1, \theta_2)$. Gibbs sampling is used to obtain posterior draws for θ and X_1 in the primary survey as follows.

Step 1. Given $\theta^* = (\theta_1^*, \theta_2^*)$, x_{1i}^* for the primary sample is generated from

$$f(x_1 \mid y_i, x_{2i}, z_i; \theta^*) \propto f_1(y_i \mid x_1, x_{2i}; \theta_1^*) f_2(x_1 \mid z_i; \theta_2^*)$$

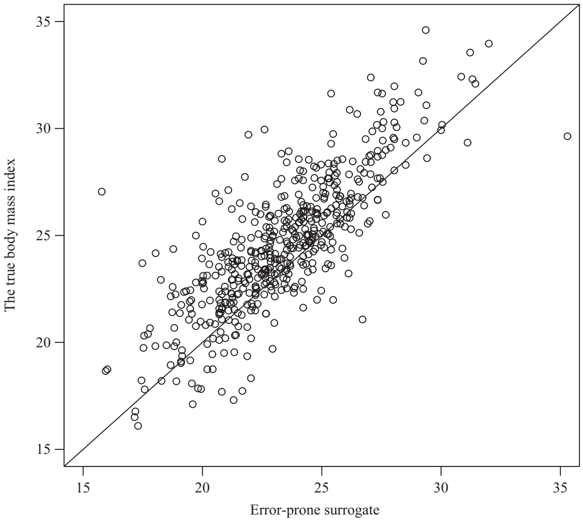

Fig. 1. Relationship between the true body mass index x_1 and the error-prone surrogate z, with the solid line representing $x_1 = z$.

Table 2. *Summary of data analysis using the proposed Bayesian method and fractional imputation*

Method		β_0	β_1	β_2
Proposed Bayesian method	Mean	$-9{\cdot}64$	0·193	0·062
	Standard error	0·35	0·012	0·002
Fractional imputation	Mean	$-9{\cdot}75$	0·194	0·063
	Standard error	0·41	0·014	0·002

by the one-step Metropolis–Hasting algorithm.

Step 2. Based on x_{1i}^*, we can generate θ_1^* from the posterior distribution of θ_1 using the estimating equation

$$\sum_{i \in V} S_1(\theta_1; y_i, x_{1i}, x_{2i}) + \sum_{i \in V^{\mathrm{c}}} S_1(\theta_1; y_i, x_{1i}^*, x_{2i}) = 0,$$

where V is the set of indices for the validation sample, V^{c} is the set of indices in the primary sample but not in the validation sample, and $S_1(\theta_1; y, x_1, x_2) = \partial \log f_1(y \mid x_1, x_2; \theta_1)/\partial\theta_1$. Similarly, we can generate θ_2^* from the posterior distribution of θ_2 using the estimating equation $\sum_{i \in V} w_i S_2(\theta_2; x_{1i}, z_i) = 0$, where $S_2(\theta_2; x_1, z) = \partial \log f_2(x_1 \mid z; \theta_2)/\partial\theta_2$. The estimating equation for θ_2 is constructed based on the validation sample only. The posterior distribution is derived from (3) using the asymptotic normality of $\hat{\theta}$.

The above two steps are repeated $M = 1000$ times to obtain θ^*. Based on the posterior draws of θ_1, we compute the posterior mean and the posterior standard error.

The results are presented in Table 2. For comparison, we include the estimates from the fractional imputation method of Kim et al. (2016). The estimates are similar, and body mass index is found to be significantly associated with hypertension. The proposed Bayesian method is easier to implement than the other method, demonstrating its promise in applications.

Acknowledgement

We are grateful to three referees and the associate editor for comments that have helped to improve this paper. The research of the second author was partially supported by the U.S. National Science Foundation.

Supplementary material

Supplementary material available at *Biometrika* online includes a brief description of the adaptive rejection Metropolis sampling algorithm within the Gibbs sampler, the proof of Theorem 1 and an additional example for calibration estimation.

References

Berkson, J. (1950). Are there two regressions? *J. Am. Statist. Assoc.* **45**, 164–80.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.* **51**, 279–92.

Breidt, F. J., Opsomer, J. D., Herndon, W., Cao, R. & Francisco-Fernandez, M. (2013). Testing for informativeness in analytic inference from complex surveys. In *Proc. 59th ISI World Statist. Congr.*, vol. 87. The Hague, The Netherlands: International Statistical Institute, pp. 889–93.

Chambers, R. & Skinner, C. J. E. (2003). *Analysis of Survey Data*. Chichester: Wiley.

Chen, Q., Elliott, M. R. & Little, R. J. A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Surv. Methodol.* **36**, 23–34.

Dong, Q., Elliott, M. R. & Raghunathan, T. E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Surv. Methodol.* **40**, 29–46.

Firth, D. & Bennett, K. (1998). Robust models in probability sampling. *J. R. Statist. Soc.* B **60**, 3–21.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, New Jersey: Wiley.

Gilks, W. R., Best, N. & Tan, K. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Appl. Statist.* **44**, 455–72.

Godambe, V. P. & Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *Int. Statist. Rev.* **54**, 127–38.

Hidiroglou, M. (2001). Double sampling. *Surv. Methodol.* **27**, 143–54.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **47**, 663–85.

Kim, J. K., Berg, E. & Park, T. (2016). Statistical matching using fractional imputation. *Surv. Methodol.* **42**, 19–40.

Kim, J. K. & Rao, J. N. K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika* **99**, 85–100.

Korn, E. L. & Graubard, B. I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *Am. Statistician* **49**, 291–5.

Korn, E. L. & Graubard, B. I. (1999). *Analysis of Health Surveys*. New York: Wiley.

Little, R. J. A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *J. Am. Statist. Assoc.* **99**, 546–56.

Little, R. J. A. (2012). Calibrated Bayes, an alternative inferential paradigm for official statistics. *J. Offic. Statist.* **28**, 309–34.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *J. R. Statist. Soc.* B **72**, 27–48.

Monahan, J. F. & Boos, D. D. (1992). Proper likelihoods for Bayesian analysis. *Biometrika* **79**, 271–8.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Int. Statist. Rev.* **61**, 317–37.

Pfeffermann, D., Krieger, A. & Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statist. Sinica* **8**, 1087–114.

Pfeffermann, D. & Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhyā* B **61**, 166–86.

Rao, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *J. Offic. Statist.* **10**, 153–65.

Singh, K., Xie, M. & Strawderman, W. E. (2005). Combining information from independent sources through confidence distributions. *Ann. Statist.* **33**, 159–83.

Skinner, C. J. & Coker, O. (1996). Regression analysis for complex survey data with missing values of a covariate. *J. R. Statist. Soc.* A **159**, 265–74.

Sugden, R. & Smith, T. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika* **71**, 495–506.

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

van der Vaart, A. W. (2000). *Asymptotic Statistics*. New York: Cambridge University Press.

Yin, G. (2009). Bayesian generalized method of moments. *Bayesian Anal.* **4**, 191–208.

Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *J. Am. Statist. Assoc.* **85**, 986–1001.

[*Received on* 24 *January* 2017. *Editorial decision on* 3 *November* 2017]