Statistical Methodology 11 (2013) 1–21

Contents lists available at SciVerse ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Bayesian predictive inference of a finite population proportion under selection bias

Balgobin Nandram [a,*], Dilli Bhatta [a], Dhiman Bhadra [a], Gang Shen [b]

[a] Department of Mathematical Sciences, Worcester Polytechnic Institute 100, Institute Road, Worcester, MA 01609, United States
[b] Department of Statistics, North Dakota State University, 1360 Bolley Drive, Fargo, ND 58108, United States

ARTICLE INFO

Article history:
Received 3 December 2011
Received in revised form
9 May 2012
Accepted 27 August 2012

Keywords:
Accept–reject algorithm
Binary responses
Monte Carlo methods
Nonignorable selection model
Survey weights
Selection not at random

ABSTRACT

We show how to infer about a finite population proportion using data from a possibly biased sample. In the absence of any selection bias or survey weights, a simple ignorable selection model, which assumes that the binary responses are independent and identically distributed Bernoulli random variables, is not unreasonable. However, this ignorable selection model is inappropriate when there is a selection bias in the sample. We assume that the survey weights (or their reciprocals which we call 'selection' probabilities) are available, but there is no simple relation between the binary responses and the selection probabilities. To capture the selection bias, we assume that there is some correlation between the binary responses and the selection probabilities (e.g., there may be a somewhat higher/lower proportion of positive responses among the sampled units than among the nonsampled units). We use a Bayesian nonignorable selection model to accommodate the selection mechanism. We use Markov chain Monte Carlo methods to fit the nonignorable selection model. We illustrate our method using numerical examples obtained from NHIS 1995 data.

1. Introduction

A serious concern of many government agencies is selection bias in survey data. In many complex surveys, individuals are sampled with differential probabilities of selection, and these are incorporated in the survey weights. For continuous responses, many researchers have worked on the problem of selection bias, but for discrete data, very little has been done. We look at a problem in which a

* Corresponding author.
E-mail addresses: balnan@wpi.edu (B. Nandram), drb122@wpi.edu (D. Bhatta), dbhadra@wpi.edu (D. Bhadra), gang.shen@ndsu.edu (G. Shen).

1572-3127/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2012.08.003

biased sample is taken from a finite population where the proportion of positive responses among the sampled units may be different from the proportion of positive responses among the nonsampled units. We assume that the survey weights can at least partially explain this difference. Our main target of inference is the finite population proportion when there is likely to be a selection bias. Specifically, we consider binary data as a special case of discrete data since not much work has been done in this area. We note the work of Malec et al. [8] and Nandram [9] for binary data. We work with the model of Malec et al. [8], which is appropriate for binary data when there is a selection bias. We found two problematic issues with their model and we discuss them in detail in this paper.

There are two ways to model selection bias. First, one can adjust the sample part of a population model. Once the parameters are estimated from the biased sample, the entire population is predicted using the population model whose parameters are obtained from the biased sample; see [9,8,12]. Second, one can utilize an explicit relation between the survey weights and the response variable. This approach for continuous data is well known [17,21,20,18]. Another approach is provided by Chambers et al. [1], who assume that the selection probabilities are related to the continuous responses; Nandram et al. [13] developed this approach further in a business application. Other approaches to selection bias include the work of Chen et al. [2], who also use penalized spline to obtain a Bayesian predictive inference for PPS sampling; see also [22] for another approach to include survey weights. These approaches are difficult to use because they require some information about the nonsample selection probabilities, and in fact, this is not the problem of interest. Nevertheless, the literature on selection bias shows convincingly the need to incorporate the selection probabilities in a sensible manner.

Pfeffermann et al. [17] consider problems similar to the one investigated in this paper, where they assume that the first-order selection probabilities are related to the response variables and these probabilities are known only for the sampled units. To make inference for the superpopulation parameters they derive marginal likelihoods using weighted distributions in the spirit of Patil and Rao [16]. However, to obtain the joint likelihood they have to use asymptotic arguments to justify combining the marginal likelihoods. Moreover, their methodology permits inference only for the superpopulation parameters. In their framework, extension to inference for finite population parameters is difficult; see [7,19] for related work. Sverchkov and Pfeffermann [21] define the sample and sample-complement distributions as two separate weighted distributions (see [16]) for developing design consistent predictors of the finite population total. Further development of this work was given recently within the small-area context [20] with informative sampling (i.e., selection bias) of areas and within selected areas. Opsomer et al. [15] discussed non-parametric small-area estimation using penalized spline regression; the selection probabilities can be incorporated in a similar manner albeit the selection probabilities are available only for the sample. Again, these works are for continuous response, and they are not directly applicable to our situation. In this paper, we analyze binary data from a single area, and we assume that nonsampled selection probabilities are not available.

When one includes the selection probabilities in a model, there are two possible choices, an ignorable or nonignorable selection model. In an ignorable selection model the response variable is not related to the selection mechanism, but in a nonignorable selection model the response is related to the selection mechanism, at least partially. For example, for binary data there may be a higher/lower proportion of positive responses among the sampled values than among the nonsampled values. To account for this discrepancy, one can allow the response binary variable to be correlated with the survey weights or their reciprocals. We use the latter approach in this paper. The notions of the selection mechanism are similar to those for missing data mechanisms, such as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). We can coin analogous notions such as selection completely at random (SCAR), selection at random (SAR) and selection not at random (SNAR). We have a nonignorable selection model for the SNAR mechanism, and this is different from the ignorable selection model for the SCAR or SAR mechanism; see [12].

For an ignorable selection model, one possibility is to have the binary response variable, y, not dependent on the selection probability associated with it. So the ignorable selection model is

simply $p(y \mid p)$. That is, $y_1, \ldots, y_N \mid p \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ and in a Bayesian analysis one assigns a prior distribution to p. In fact, any parametric model, in which the y_i are not correlated with the survey weights, suffices. However, we have a biased sample of size n from the finite population.

To incorporate selection bias into the ignorable selection model, Malec et al. [8] use a hierarchical Bayesian model to estimate a finite population proportion when there are binary data. Difficulty in including the selection probabilities directly in the model forces them to make an ad hoc adjustment to the likelihood function and use a Bayes empirical Bayes (i.e., not a full Bayesian) approach. However, without their adjustment, Nandram and Choi [12] have incorporated selection probabilities into a nonignorable nonresponse model to analyze continuous body mass index data using a full Bayesian analysis. For binary data, we start with the model of Malec et al. [8], which includes an unspecified relation between the selection probabilities and a binary characteristic. Henceforth, we will use the term, MDC, to refer to Malec et al. [8].

MDC imagines a sampled unit representing itself and $N^* - 1$ other units (not sampled) in the population, and the model selection indicators are $\delta_j,\ j = 1, \ldots, N^*$; here N^* is unknown. So that $\delta_1 = 1$ for the unit actually sampled, and $\delta_2 = \cdots = \delta_{N^*} = 0$ for the imaginary units this single sampled unit represents. Letting $\pi_u^*,\ u = 1, \ldots, U$, be specified values, the pertinent part of the MDC model is

$$\delta_j \mid N^*, \pi_j \overset{\text{ind}}{\sim} \text{Bernoulli}(\pi_j), \quad j = 1, \ldots, N^*, \tag{1}$$

$$P(\pi_j = \pi_u^* \mid \underset{\sim}{\theta}, y) = \theta_{uy}, \quad y = 0, 1,\ u = 1, \ldots, U, \tag{2}$$

$$P(Y_j = y \mid p) = p^y(1-p)^{1-y}, \quad y = 0, 1,\ 0 \le p \le 1,\ j = 1, \ldots, N^*. \tag{3}$$

A priori MDC took $P(N^*) = 1,\ N^* \ge 1$ (in practice, $N^* \gg 1$). This is a uniform prior on $1, 2, \ldots$ which allows easy integration but it is improper. We review the development of MDC in Appendix A, where we provide a complete proof and show that

$$p(Y = y, \pi = \pi_u^* \mid \underset{\sim}{\theta}, p) = \frac{\pi_u^* \theta_{uy} P(Y = y \mid p)}{\sum_{u=1}^{U} \pi_u^* \sum_{y=0}^{1} \theta_{uy} P(Y = y \mid p)}, \quad y = 0, 1,\ u = 1, \ldots, U; \tag{4}$$

MDC did not present this form. (We drop subscript j from Y because (4) is the likelihood of a single observation.) In (4) there is also a conditioning on $\delta_j = 1$ which we have dropped from the notation; see Appendix A. Once p is estimated, we can draw the entire finite population values, $y_1, \ldots, y_N$, via surrogate sampling; see [9,12].

However, there are two possible problems with this model. First, the θ_{uy} are only weakly identified. Second, the parameters θ_{uy} are never known, and in a Bayesian framework these must also be stochastic. In this paper, in a single attempt, we show how to solve these two problems (weak identifiability and stochastic parameters) for a biased sample drawn from a binary population using information from the survey weights (or selection probabilities).

It is easy to see why the θ_{uy} might be weakly identified. Note that, since $\sum_{u=1}^{U} \theta_{uy} = 1, y = 0, 1$, the θ_{uy} are not invariant to scale (i.e., scale is not a problem). However, it is possible that the θ_{uy} are invariant to location. For example, adding α to θ_{uy}, where $\alpha + \theta_{uy} < 1$ and $\alpha < \theta_{u'y}$, and subtracting α from $\theta_{u'y}$ does not change $\sum_{u=1}^{U} \theta_{uy} = 1, y = 0, 1$. This will lead to long-range dependence in a Gibbs sampler (i.e., poor mixing) with a uniform prior; see [4]. This problem can be corrected by putting a proper informative prior on θ_{uy}.

A related problem occurs. When $\theta_{u0} \approx \theta_{u1}, u = 1, \ldots, U$, there will be difficulty in estimation. In fact, when the θ_{uy} do not depend on y, the term $\pi_u^* \theta_{uy} / \sum_{u=1}^{U} \pi_u^* \theta_{uy}$ in (4) becomes independent of $P(Y = y \mid p)$ and the selection probabilities do not matter. Therefore, at least some of the θ_{u0} must be sufficiently different from the θ_{u1}. If the sampling scheme is close to simple random sampling (i.e., very little selection bias), there will be difficulties in model fitting.

In this paper, we consider the problem of making inference about a finite population proportion when a possibly biased sample is available from it. Specifically we show how to fix the two problems of MDC by putting a proper informative prior on θ_{uy}. The survey weights help to adjust for the bias.

In Section 2 we show how to adjust a standard ignorable selection model to incorporate the survey weights. We also show how to perform the computations. In Section 3 we provide a numerical example on severe activity limitation in the 1995 National Health Interview Survey (NHIS 1995). Section 4 has concluding remarks. Further explanations are given in the appendices.

2. Methodology

We consider a finite population of N units, and we view this finite population as a random sample from a superpopulation. However, the sample from the finite population can be biased. That is, a probability sample of size n is taken with selection probabilities $\pi_i, i = 1, \ldots, N$. The selection probabilities are observed only for the sampled values. These selection probabilities are adjusted by the design scientists because of various reasons such as nonresponse and different weights from various sources.

Let $y_1, \ldots, y_N$ denote the finite population values and a sample S of size n is taken from the population; also let $\bar{S}$ denote the set of nonsampled values. Let the sampled values be $y_1, \ldots, y_n$. Let $P = \sum_{i=1}^{N} y_i/N$ denote the finite population proportion and $\hat{p} = \sum_{i=1}^{n} y_i/n$ denote the corresponding sample proportion. While p is the parameter of the superpopulation, P is the analogous parameter in finite population. Clearly, $\hat{p}$ can be a biased estimator of p. Typically, the sample selection probabilities of the sampled units are known. In design-based survey analysis, P is a fixed unknown quantity, but clearly in Bayesian inference P is a random variable which is to be predicted. Our main interest is to predict P when a biased sample is available from the superpopulation.

We describe an ignorable selection model and a nonignorable selection model. We also show how to fit these models and how to make inference about the finite population proportion. Under the nonignorable selection model, inference about the finite population proportion is obtained using surrogate samples.

2.1. Ignorable selection model

A standard ignorable selection model for the binary variables y_i, $i = 1, \ldots, N$, is

$$y_i \mid p \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p) \quad \text{and} \quad p \sim \text{Uniform}(0, 1).$$

Then, letting $s = \sum_{i \in S} y_i$, $p \mid s \overset{\text{ind}}{\sim} \text{Beta}(s + 1, n - s + 1)$ with $E(p|s) = (s + 1)/(n + 2)$, $\text{var}(p|s) = (s + 1)(n - s + 1)/(n + 2)^2(n + 3)$ and the posterior modal estimator of p is $\hat{p} = \frac{s}{n}$. It is also easy to show that the $100(1 - \alpha)\%$ highest posterior density (HPD) interval of p exists. Let $F(\cdot)$ denote the cdf of the $Beta(s + 1, n + 1)$ random variable. If the mode is at 0, the HPD interval will be $(0, F^{-1}(1 - \alpha))$. If the mode is at 1, then the HPD interval will be $(F^{-1}(\alpha), 1)$. If the mode is not on the boundary (i.e., 0 or 1), then the HPD interval will be $(a, F^{-1}\{1 - \alpha + F(a)\})$, where $0 < a < 1$ is the solution to

$$\left[F^{-1}\{(1 - \alpha) + F(a)\}\right]^s \left[1 - F^{-1}\{(1 - \alpha) + F(a)\}\right]^{(n-s)} = a^s(1 - a)^{n-s}$$

which can be easily computed using the bisection method.

Bayesian predictive inference of P is performed in the following manner. Letting $\underset{\sim}{y}_s$ denote the vector of sampled values and $\underset{\sim}{y}_{\bar{s}}$ denote the vector of nonsampled values. Then, letting $f = n/N$ denote the sampling fraction, we have

$$P = f\hat{p} + (1 - f)\bar{Y}_{\bar{s}} \quad \text{with } \bar{Y}_{\bar{s}} = T/(N - n), T = \sum_{i \in \bar{S}} y_i,$$

where $\hat{p}$ is the sample proportion and

$$T \mid p \sim \text{Binomial}(N - n, p).$$

Thus, inference is straightforward (e.g., Rao–Blackwellized estimators of the posterior density of P can be obtained).

2.2. Nonignorable selection model

We assume that the sample selection probabilities $(\pi_1, \ldots, \pi_n)$ have support over the set $\pi_u^*,\ u = 1, \ldots, U$. That is, $\pi_i,\ i = 1, \ldots, n$, have a histogram where the midpoints of the categories are the π_u^*. Throughout these π_u^* are assumed known and the π_i are assumed to be random quantities. The distribution of the selection probabilities, given the binary response y_i, is

$$\Pr(\pi_i = \pi_u^* \mid \underset{\sim}{\theta}, y_i = y) = \theta_{uy}, \quad u = 1, \ldots, U,\ y = 0, 1,\ i = 1, \ldots, n$$

and

$$y_i \mid p \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p), \quad i = 1, \ldots, N.$$

Note that p is the proportion of ones in the entire superpopulation. It is worth noting that if $\theta_{u0} = \theta_{u1}, u = 1, \ldots, U$, the model cannot accommodate selection bias. We will discuss this issue later.

Using (4), it is easy to show that

$$P(Y = y \mid \pi = \pi_u^*, \underset{\sim}{\theta}, p) = \frac{\theta_{uy} p^y (1-p)^{1-y}}{\sum_y \theta_{uy} p^y (1-p)^{1-y}} \tag{5}$$

and

$$P(Y = y | \underset{\sim}{\theta}, p) = \frac{\sum_u \pi_u^* \theta_{uy} p^y (1-p)^{1-y}}{\sum_y \sum_u \pi_u^* \theta_{uy} p^y (1-p)^{1-y}}. \tag{6}$$

The sampled data actually come from the probability mass function in (6) and the entire population is described by $P(Y = y \mid p) = p^y (1-p)^{1-y}, y = 0, 1$, thereby showing how the selection bias enters into the model. Note again that if θ_{uy} does not depend on y, these two probability mass functions are exactly the same. Thus, the key parameters in the selection bias are the θ_{uy}. Thus, one has to be cautious in model fitting.

Now, in (5),

$$P(Y = y | \pi = \pi^*, \underset{\sim}{\theta}, p) = \left[\frac{\theta_{uy}}{\sum_y \theta_{uy} p^y (1-p)^{1-y}} \right] p^y (1-p)^{1-y}.$$

Here $P(Y = y | \pi = \pi^*, \underset{\sim}{\theta}, p)$ is a weighted distribution with weights $w = \frac{\theta_{uy}}{\sum_y \theta_{uy} p^y (1-p)^{1-y}}$ and $P(Y = y \mid p)$ is distribution of the population without selection bias (i.e., under the ignorable selection model). Next we discuss how these weights affect the original distribution. Now $P(Y = 1 | \pi = \pi^*, p = 1) = 1$ and $P(Y = 0 | \pi = \pi^*, p = 0) = 1$. That is, if p is near the boundary at 0 or 1, there is little selection bias. However, $P(Y = 1 | \pi = \pi^*, p = 1/2, \underset{\sim}{\theta}) = \theta_{u1}/(\theta_{u0} + \theta_{u1})$, and this is not equal to a half and it does depend on $\underset{\sim}{\theta}$. Thus, when p is around the middle of (0, 1), our model will pick up the selection bias.

Since the sampling units are independent, using (4), the joint density of the entire sample is

$$P(\underset{\sim}{y}, \underset{\sim}{\pi} \mid \underset{\sim}{\theta}, p) = \frac{\prod_{u=1}^{U} (\pi_u^* \theta_{u0})^{g_{u0}} \prod_{u=1}^{U} (\pi_u^* \theta_{u1})^{g_{u1}}}{\left[p \sum_u \pi_u^* \theta_{u1} + (1-p) \sum_u \pi_u^* \theta_{u0} \right]^n} p^s (1-p)^{(n-s)}, \tag{7}$$

where $s = \sum_{i=1}^n y_i$, g_{u0} is the cell count for category u for $y = 0$ and g_{u1} is the cell counts for category u for $y = 1$. Note that $\sum_{u=1}^U g_{u0} = n - s$, $\sum_{u=1}^U g_{u1} = s$ and $\sum_{u=1}^U (g_{u0} + g_{u1}) = n$. This likelihood includes the selection bias.

The parameters $\underset{\sim}{\theta}_0$ and $\underset{\sim}{\theta}_1$ are weakly identified. This can lead to poor performance of a Markov chain Monte Carlo algorithm (e.g., Gibbs sampler). Thus, it is sensible to minimize the effect of the

nonidentifiability as much as possible. A priori we assume that p, $\underset{\sim}{\theta}_0$ and $\underset{\sim}{\theta}_1$ are independent, and we take

$$p \sim \text{Uniform}(0, 1)$$
$$\underset{\sim}{\theta}_0 \mid \tau \sim \text{Dirichlet}(\underset{\sim}{\theta}_0^{(0)}\tau) \quad \text{and} \quad \underset{\sim}{\theta}_1 \mid \tau \sim \text{Dirichlet}(\underset{\sim}{\theta}_1^{(0)}\tau),$$

where $\underset{\sim}{\theta}_0^{(0)}$ and $\underset{\sim}{\theta}_0^{(0)}$ are to be specified. Finally,

$$p(\tau) = \frac{1}{(1+\tau)^2}, \tau \geq 0.$$

This latter prior, also called a shrinkage (reference, objective) prior, is used to avoid the difficulties associated with improper priors of the form $p(\tau) \propto 1/\tau$ (e.g., see [5] for a discussion of such improper priors and Natarajan and Kass [14] for a discussion of shrinkage prior). Therefore, the joint prior density of p, $\underset{\sim}{\theta}_0$, $\underset{\sim}{\theta}_1$, τ is

$$\pi(p, \underset{\sim}{\theta}_0, \underset{\sim}{\theta}_1, \tau) = \frac{\prod_{u=1}^{U} \theta_{u0}^{\theta_{u0}^{(0)}\tau-1}}{D(\underset{\sim}{\theta}_0^{(0)}\tau)} \frac{\prod_{u=1}^{U} \theta_{u1}^{\theta_{u1}^{(0)}\tau-1}}{D(\underset{\sim}{\theta}_1^{(0)}\tau)} \frac{1}{(1+\tau)^2}. \tag{8}$$

Next, using Bayes' theorem, the joint posterior density of p, $\underset{\sim}{\theta}_1$, $\underset{\sim}{\theta}_0$, τ given the data, $\underset{\sim}{\pi}$, $\underset{\sim}{y}$, is

$$\pi(p, \underset{\sim}{\theta}_1, \underset{\sim}{\theta}_0, \tau \mid \underset{\sim}{\pi}, \underset{\sim}{y}) \propto \frac{\prod_{u=1}^{U} (\pi_u^* \theta_{u0})^{g_{u0}} \prod_{u=1}^{U} (\pi_u^* \theta_{u1})^{g_{u1}}}{\left[p \sum_u \pi_u^* \theta_{u1} + (1-p) \sum_u \pi_u^* \theta_{u0}\right]^n} p^s (1-p)^{n-s}$$
$$\times \frac{\prod_{u=1}^{U} \theta_{u0}^{\theta_{u0}^{(0)}\tau-1}}{D(\underset{\sim}{\theta}_0^{(0)}\tau)} \frac{\prod_{u=1}^{U} \theta_{u1}^{\theta_{u1}^{(0)}\tau-1}}{D(\underset{\sim}{\theta}_1^{(0)}\tau)} \frac{1}{(1+\tau)^2}, \tag{9}$$

where $\sum_{u=1}^{U} g_{u0} = n - s$ and $\sum_{u=1}^{U} g_{u1} = s$. Henceforth, we will drop $\underset{\sim}{\pi}$ from the conditioning.

Letting

$$a_y = \sum_u \pi_u^* \theta_{uy}, \quad y = 0, 1,$$

the conditional posterior density of p is

$$\pi(p \mid \underset{\sim}{\theta}_1, \underset{\sim}{\theta}_0, \tau, \underset{\sim}{y}) \propto \frac{1}{[a_1 p + a_0(1-p)]^n} p^s (1-p)^{n-s}$$

which is a weighted beta distribution [16]. Note that $\pi(p \mid \underset{\sim}{\theta}_1, \underset{\sim}{\theta}_0, \tau, \underset{\sim}{y})$ does not depend on τ so that sometimes we will drop τ from the conditioning. It is also true that the conditional posterior distribution of p is unimodal. To see this, let $\Delta(p) = -n \ln[a_1 p + a_0(1-p)] + s \ln(p) + (n-s) \ln(1-p)$. Then,

$$\Delta'(p) = -\frac{n(a_1 - a_0)}{a_1 p + a_0(1-p)} + \frac{s}{p} - \frac{n-s}{1-p}$$

and

$$\Delta''(p) = \frac{n(a_1 - a_0)^2}{[a_1 p + a_0(1-p)]^2} - \frac{s}{p^2} - \frac{n-s}{(1-p)^2}.$$

Solving $\Delta'(\tilde{p}) = 0$, we have

$$\tilde{p} = \frac{a_0 n}{s(a_0 - a_1) + a_1 n} \hat{p}$$

as a solution. It is also easy to show that $\Delta''(\tilde{p}) < 0$, so that $\tilde{p}$ is the unique mode. This is useful because it is relatively much easier to study a unimodal density such as $\pi(p \mid \underset{\sim}{\theta}_1, \underset{\sim}{\theta}_0, \tau, \underset{\sim}{y})$ using sampling-based analysis. However, one can see that $\pi(p \mid \underset{\sim}{\theta}_1, \underset{\sim}{\theta}_0, \tau, \underset{\sim}{y})$ is not logconcave (i.e., strongly unimodal).

It is interesting to look at $\tilde{p}$ momentarily. When $a_0 = a_1$, $\tilde{p} = \hat{p}$ and there is no selection bias. This means that if $a_0 = a_1$, $\sum_u \pi_u^*(\theta_{u0} - \theta_{u1}) = 0$, the weighted average of the deviations $\theta_{u0} - \theta_{u1} = 0$ with weight π_u^*, there is no selection bias. Therefore, it is not required that $\theta_{u0} = \theta_{u1}$ for every u to have no selection bias. Let

$$\nu = \frac{a_0 n}{\{s(a_0 - a_1) + a_1 n\}} = \frac{a_0 n}{\{a_0 n + (n - s)(a_1 - a_0)\}}.$$

Thus, if $a_0 \geq a_1$ then $\nu \geq 1$ and so $\tilde{p} \geq \hat{p}$. Similarly if $a_0 < a_1$ then $\nu < 1$ and so $\tilde{p} < \hat{p}$. This explains how the model accounts for the selection bias as the estimates of p under the ignorable selection model and nonignorable selection model are different.

2.3. Computation

It is convenient to use a sampling based method to make inference about p. Using the joint posterior density in (9) and assuming the π_u^* are fixed and known, to perform the Gibbs sampler, we need the conditional posterior densities, given by

$$\tilde{g}_1(\underset{\sim}{\theta}_0 | \underset{\sim}{\theta}_1, p, \tau, \underset{\sim}{y}) \propto \frac{\prod_{u=1}^{U} (\pi_u^* \theta_{u0})^{g_{u0}}}{[a_1 p + a_0(1-p)]^n} \prod_{u=1}^{U} \theta_{u0}^{\theta_{u0}\tau - 1},$$

$$\tilde{g}_2(\underset{\sim}{\theta}_1 | \underset{\sim}{\theta}_0, p, \tau, \underset{\sim}{y}) \propto \frac{\prod_{u=1}^{U} (\pi_u^* \theta_{u1})^{g_{u1}}}{[a_1 p + a_0(1-p)]^n} \prod_{u=1}^{U} \theta_{u1}^{\theta_{u1}\tau - 1},$$

$$\tilde{g}_3(p | \underset{\sim}{\theta}_0, \underset{\sim}{\theta}_1, \tau, \underset{\sim}{y}) \propto \frac{1}{[a_1 p + a_0(1-p)]^n} p^s (1-p)^{n-s},$$

and

$$\tilde{g}_4(\tau \mid \underset{\sim}{\theta}_0, \underset{\sim}{\theta}_1, p, \underset{\sim}{y}) \propto \left(\frac{\Gamma(\tau)}{1+\tau}\right)^2 \prod_{u=1}^{U} \left\{ \frac{\theta_{u0}^{\theta_{u0}\tau - 1}}{\Gamma(\theta_{u0}\tau)} \frac{\theta_{u1}^{\theta_{u1}\tau - 1}}{\Gamma(\theta_{u1}\tau)} \right\}.$$

However, because of the accept–reject method we use for drawing p, this Gibbs sampler is a bit slow.

We use an alternative procedure that avoids the accept–reject algorithm within the Gibbs sampler. Rather we use the accept–reject algorithm in an output analysis of the Gibbs sampler. This procedure accelerates the Gibbs sampler; see [11] for a related procedure to accelerate the Gibbs sampler. To proceed, we first transform p to q in (9) keeping all other variables untransformed. It is easy to see that the quantity, which is affected by the transformation, is $H(p, \underset{\sim}{\theta})$, where

$$H(p, \underset{\sim}{\theta}) = \frac{p^s (1-p)^{n-s}}{[p a_1 + (1-p) a_0]^n} = \frac{a_1^s p^s \{a_0(1-p)\}^{n-s}}{a_1^s a_0^{n-s} \{p a_1 + (1-p) a_0\}^n}.$$

We make a one-to-one transformation from p to q via

$$q = \frac{a_1 p}{a_1 p + a_0(1-p)}.$$

Note that retransforming to p, we have

$$p = \frac{a_0 q}{a_0 q + a_1(1-q)}.$$

The Jacobian of the transformation is $a_0a_1/(qa_0+(1-q)a_1)^2$. Then,

$$\begin{aligned} H(q,\underset{\sim}{\theta}) &= \frac{q^{s-1}(1-q)^{n-s-1}}{a_0^{n-s}a_1^s}\left(\frac{a_0}{a_0q+a_1(1-q)}\right)\left(\frac{a_1}{a_0q+a_1(1-q)}\right)\\ &= \frac{q^{s-1}(1-q)^{n-s-1}}{a_0^{n-s}a_1^s}\left(\frac{(1-q)a_1}{a_0q+a_1(1-q)}\right)\left(\frac{qa_0}{a_0q+a_1(1-q)}\right)\\ &= q^{s-1}(1-q)^{n-s-1}\frac{\Delta_{\underset{\sim}{\theta}}(q)(1-\Delta_{\underset{\sim}{\theta}}(q))}{\left(\sum_{u=1}^{U}\pi_u^*\theta_{u0}\right)^{n-s}\left(\sum_{u=1}^{U}\pi_u^*\theta_{u1}\right)^{s}}, \end{aligned}$$

where $\Delta_{\underset{\sim}{\theta}}(q)=\frac{(1-q)a_1}{a_0q+a_1(1-q)}$ and $\underset{\sim}{\theta}=(\underset{\sim}{\theta}_0,\underset{\sim}{\theta}_1)$.

Therefore, the joint posterior density of $q,\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0,\tau$ given the data $\underset{\sim}{y}$, is

$$\begin{aligned} \pi(q,\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0,\tau\mid\underset{\sim}{y}) \propto\ & q^{s-1}(1-q)^{n-s-1}\Delta_{\underset{\sim}{\theta}}(q)\{1-\Delta_{\underset{\sim}{\theta}}(q)\}\\ & \times\frac{\prod_{u=1}^{U}(\pi_u^*\theta_{u0})^{g_{u0}}\prod_{u=1}^{U}(\pi_u^*\theta_{u1})^{g_{u1}}}{\left(\sum_{u=1}^{U}\pi_u^*\theta_{u0}\right)^{n-s}\left(\sum_{u=1}^{U}\pi^*\theta_{u1}\right)^{s}}\\ & \times\frac{\prod_{u=1}^{U}\theta_{u0}^{\theta_{u0}^{(0)}\tau-1}}{D(\underset{\sim}{\theta}_0^{(0)}\tau)}\frac{\prod_{u=1}^{U}\theta_{u1}^{\theta_{u1}^{(0)}\tau-1}}{D(\underset{\sim}{\theta}_1^{(0)}\tau)}\frac{1}{(1+\tau)^2}. \end{aligned} \tag{10}$$

To accelerate the Gibbs sampler, we integrate out q from (10). The integrated posterior density is

$$\begin{aligned} \pi(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0,\tau\mid\underset{\sim}{y}) \propto\ & I(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0;\underset{\sim}{y})\frac{\prod_{u=1}^{U}(\pi_u^*\theta_{u0})^{g_{u0}}\prod_{u=1}^{U}(\pi_u^*\theta_{u1})^{g_{u1}}}{\left(\sum_{u=1}^{U}\pi_u^*\theta_{u0}\right)^{n-s}\left(\sum_{u=1}^{U}\pi^*\theta_{u1}\right)^{s}}\\ & \times\frac{\prod_{u=1}^{U}\theta_{u0}^{\theta_{u0}^{(0)}\tau-1}}{D(\underset{\sim}{\theta}_0^{(0)}\tau)}\frac{\prod_{u=1}^{U}\theta_{u1}^{\theta_{u1}^{(0)}\tau-1}}{D(\underset{\sim}{\theta}_1^{(0)}\tau)}\frac{1}{(1+\tau)^2} \end{aligned} \tag{11}$$

where

$$I(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0;\underset{\sim}{y})=\int_0^1\frac{q^{s-1}(1-q)^{n-s-1}}{B(s,n-s)}\Delta_{\underset{\sim}{\theta}}(q)\{1-\Delta_{\underset{\sim}{\theta}}(q)\}dq.$$

Clearly, $0\le I(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0;\underset{\sim}{y})\le 1$, and is therefore well defined. Note also that $I(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0;\underset{\sim}{y})$ does not depend on τ.

It is easy to compute the one-dimensional integral, $I(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0;\underset{\sim}{y})$. Let $F(t)$ denote the cumulative distribution function of $T\sim\text{Beta}(s,n-s)$. We divide $[0,1]$ into 100 sub-intervals with end points $b_j=0.01j, j=0,\ldots,100$. Let $c_j=(b_{j-1}+b_j)/2, j=1,\ldots,100$. Then, an efficient estimator of $I(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0;\underset{\sim}{y})$ is

$$\hat{I}(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0;\underset{\sim}{y})=\sum_{j=1}^{100}\Delta_{\underset{\sim}{\theta}}(c_j)(1-\Delta_{\underset{\sim}{\theta}}(c_j))\{F(b_j)-F(b_{j-1})\}.$$

This runs very quickly.

Thus, the conditional posterior densities needed to execute the integrated Gibbs sampler are

$$g_1(\underset{\sim}{\theta}_0|\underset{\sim}{\theta}_1,\tau,\underset{\sim}{y}) \propto \frac{\prod_{u=1}^{U}(\pi_u^*\theta_{u0})^{g_{u0}}}{\left(\sum_{u=1}^{U}\pi_u^*\theta_{u0}\right)^{n-s}}\left[\prod_{u=1}^{U}\theta_{u0}^{\theta_{u0}^{(0)}\tau-1}\right]I(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0;\underset{\sim}{y}), \tag{12}$$

$$g_2(\underset{\sim}{\theta}_1|\underset{\sim}{\theta}_0,\tau,\underset{\sim}{y}) \propto \frac{\prod_{u=1}^{U}(\pi_u^*\theta_{u1})^{g_{u1}}}{\left(\sum_{u=1}^{U}\pi^*\theta_{u1}\right)^{s}}\left[\prod_{u=1}^{U}\theta_{u1}^{\theta_{u1}^{(0)}\tau-1}\right]I(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0;\underset{\sim}{y}), \tag{13}$$

and transforming τ to $\phi = \tau/(1+\tau)$,

$$g_3(\phi \mid \underset{\sim}{\theta}_0,\underset{\sim}{\theta}_1,\underset{\sim}{y}) \propto \left[\{\Gamma(\tau)\}^2\prod_{u=1}^{U}\left\{\frac{\theta_{u0}^{\theta_{u0}^{(0)}\tau-1}}{\Gamma(\theta_{u0}^{(0)}\tau)}\frac{\theta_{u1}^{\theta_{u1}^{(0)}\tau-1}}{\Gamma(\theta_{u1}^{(0)}\tau)}\right\}\right]_{\tau=\phi/(1-\phi)}, \quad 0<\phi<1. \tag{14}$$

The integrated Gibbs sampler runs by drawing a sample from (12)–(14), each in turn, and iterating the procedure. Samples from these conditional posterior densities are obtained using a grid method. Note that, because of the limited amount of data, τ is expected to be small; so that drawing ϕ using a grid method is efficient.

Note $I(\underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0;\underset{\sim}{y})$ has to be computed at each step of the Gibbs sampler. However, we need to compute the $F(b_j)$ just once. This speeds up the procedure. Also, note that when $\underset{\sim}{\theta}_0$ or $\underset{\sim}{\theta}_1$ is drawn, the algorithm is performed component-wise. For example, when draws are made from the conditional posterior distribution of θ_{u0}, $u = 1,\ldots,U-1$, the grid is performed not on the entire interval [0, 1], but on a much shorter interval $[0, \sum_{u'=1,u'\neq u}^{U-1}\theta_{u'0}]$.

The conditional posterior density of q is

$$\pi(q \mid \underset{\sim}{\theta}_0,\underset{\sim}{\theta}_1,y) \propto q^{s-1}(1-q)^{n-s-1}\Delta_{\underset{\sim}{\theta}}(q)\{1-\Delta_{\underset{\sim}{\theta}}(q)\}.$$

Sampling from $\pi(q \mid \underset{\sim}{\theta}_0,\underset{\sim}{\theta}_1,y)$ can be done using the *simple* accept–reject algorithm which we now discuss. We use the candidate generating density $q \sim \text{Beta}(s, n-s)$ which we denote by $\pi_a(q \mid \underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0,\underset{\sim}{y})$. Then,

$$\frac{\pi(q \mid \underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0,\underset{\sim}{y})}{\pi_a(q \mid \underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0,\underset{\sim}{y})} = \Delta_{\underset{\sim}{\theta}}(q)\{1-\Delta_{\underset{\sim}{\theta}}(q)\} \leq \frac{1}{4}.$$

Hence, to run the accept–reject algorithm, we need the acceptance probability,

$$\frac{\pi(q \mid \underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0,\underset{\sim}{y})}{\pi_a(q \mid \underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0,\underset{\sim}{y})}\Bigg/\sup_{0<q<1}\frac{\pi(q \mid \underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0,\underset{\sim}{y})}{\pi_a(q \mid \underset{\sim}{\theta}_1,\underset{\sim}{\theta}_0,\underset{\sim}{y})} = 4\Delta_{\underset{\sim}{\theta}}(q)\{1-\Delta_{\underset{\sim}{\theta}}(q)\}.$$

Then, we draw $q \sim \text{Beta}(s, n-s)$ and accept it with probability $4\Delta_{\underset{\sim}{\theta}}(q)\{1-\Delta_{\underset{\sim}{\theta}}(q)\}$. It is worth noting that we cannot use *adaptive* rejection sampling (e.g., [6]) because while the conditional posterior density of q is unimodal, it is not logconcave (required for adaptive rejection sampling). That is, the *adaptive* rejection sampling is a special case of the accept–reject sampling which we just described.

Once p is estimated, we draw the entire finite population values, $y_1,\ldots,y_N$, independently from Bernoulli(p); Nandram [9] and Nandram and Choi [12] have done this in a similar manner. Here, we simply need $\sum_{i=1}^{N} y_i \mid p \sim \text{Binomial}(N,p)$. So really we have corrected the observed biased sample and replaced it by a surrogate sample for every p that we obtained from the nonignorable selection model. Let $\pi(p \mid \underset{\sim}{y}_s,\underset{\sim}{\pi})$ denote the posterior density of p. Note again that p is the proportion of ones in the entire superpopulation (i.e., the selection bias has been removed). Thus,

$$\Pi(P \mid \underset{\sim}{y}_s,\underset{\sim}{\pi}) = \int \pi(P \mid p)\pi(p \mid \underset{\sim}{y}_s,\underset{\sim}{\pi})dp.$$

Once samples are obtained from the posterior density of p, it is now easy to take a census of the entire population using the composition method. To every sample of p, obtained from the Gibbs sampler, a sample of P is obtained by drawing $\sum_{i=1}^{N} y_i$ from the Binomial distribution (one does not need to draw each Bernoulli random variable) and divide the result by N. Thus, we have obtained a Rao–Blackwellized estimator of the posterior density of P.

For our examples we have run the Gibbs sampler in exactly the same manner. We used the first 1000 iterates as a 'burn in', and we took every tenth thereafter to get 1000 iterates which we use to infer about the finite population proportion. We use the trace plots and the autocorrelation coefficient among the iterates to monitor convergence of the Gibbs sampler.

Finally, we note that the highest posterior density (HPD) interval of the finite population proportions are obtained using the algorithm of Chen and Shao [3], which requires unimodality of the posterior densities. The algorithm, based on the bisection method which we described earlier for the ignorable selection model, gives answers which are very close to the nonparametric method of Chen and Shao [3].

3. Numerical examples

The data we used for this study comes from the 1995 National Health Interview Survey (NHIS 95). The National Health Interview Survey is an important source of information on the health of the US population. We construct twelve examples from these data to allow us to compare different patterns of selection bias. Although our data analyses are important to the National Center for Health Statistics, the specific purpose of our study is to show differences between the ignorable selection model and our nonignorable selection model.

One of the variables in NHIS is activity limitation, which is a major health problem among adults, with chronic conditions in the United States. For adults, age 30–80 years, we study severe activity limitation (SAL), where $y = 1$ if an adult has SAL and $y = 0$ otherwise. In the original data, there are seven levels of education, and we have recoded it into three levels (pre-college: L, college: M and post-college: H). Sex has two levels (male: M and female: F). Race also has two levels (white: W and nonwhite: B). Each combination of education, sex and race is considered as a domain; therefore, the dataset is divided into twelve domains. We will call these domains LMW, LMB, LFW, LFB, MMW, MMB, MFW, MFB, HMW, HMB, HFW, HFB (e.g., LMW: white males with pre-college education, LMB: black males with pre-college education, etc.). Thus, there are twelve examples corresponding to data from these twelve domains which are analyzed independently.

Since the NHIS 95 uses a multistage sampling design to draw samples from the US population, it is necessary to use an adult's survey weight for accurate analysis of the data. The survey weight for each sampled adult is the product of four components. These are the probability of selection, household adjustment within segment, first-stage ratio adjustment, and post-stratification by age-sex-race-ethnicity. In addition to the design and ratio adjustments included in a person's basic weight, a person's weight is further modified depending on the variable selected, the length of the recall period, and the period of time for which the estimate is to be made. The combined weight is used as our sampling weight. Nandram et al. [10] has an extensive discussion of the survey design and data collection. They also have a more detailed discussion of these data with a different emphasis. We will consider the reciprocal of a survey weight as the 'selection' probability of each adult.

We order the selection probabilities from smallest to largest $\pi_{(1)}, \ldots, \pi_{(n)}$. Let the quintiles be $t_1 = \pi_{(0.20n)}$, $t_2 = \pi_{(0.40n)}$, $t_3 = \pi_{(0.60n)}$, $t_4 = \pi_{(0.80n)}$ and let $t_0 = \pi_{(1)}$ and $t_5 = \pi_{(n)}$. We define $\pi_u^* = (t_{u-1} + t_u)/2$, $u = 1, \ldots, 5$ (i.e., the midpoint). Note that θ_{uy} is the proportion of units in the interval (t_{u-1}, t_u) conditional on $y = 0, 1$. If the θ_{u0} are considerably different from θ_{u1}, there is strong evidence that the sampled values are biased.

The data are shown in Table 1. The selection probabilities on the *average* are mostly similar for $y = 0$ and $y = 1$. The sampling fractions are very small but the sample sizes within the domains are large. Because the sample size of each domain is very large, there is enough data. Unfortunately large sample sizes do not matter when there is selection bias. Thus, the simultaneous analysis of these domains as in small-area estimation is not appropriate, and therefore it is not the focus of this paper.

Table 1
Summaries of the key features of the data on severe activity limitation (SAL) and the selection probabilities.

Domain	n	s	$\hat{p}$	f	avg0	avg1	p-value
LMW	1738	267	0.154	0.303	0.402	0.363	0.000
LMB	305	82	0.269	0.275	0.317	0.311	0.927
LFW	1997	268	0.134	0.325	0.452	0.419	0.095
LFB	400	77	0.193	0.286	0.336	0.355	0.483
MMW	7595	571	0.075	0.227	0.273	0.267	0.218
MMB	1507	184	0.122	0.250	0.284	0.310	0.083
MFW	8918	538	0.060	0.233	0.285	0.277	0.546
MFB	2004	179	0.089	0.279	0.310	0.322	0.334
HMW	7555	228	0.030	0.213	0.243	0.249	0.383
HMB	1236	52	0.042	0.236	0.269	0.283	0.134
HFW	7794	275	0.035	0.219	0.254	0.242	0.101
HFB	1682	65	0.039	0.257	0.292	0.301	0.438

Note: Domains are formed by crossing education (pre-college: L, college: M and post-college: H), sex (male: M, female: F) and race (white: W, black: B). Here n is the total sample size, s is the number of adults with SAL, and $\hat{p} = s/n; f = n/N$ is the sampling fraction; avg0 is the average of the selection probabilities for $y = 0$ (SAL, no) and avg1 is the average of the selection probabilities for $y = 1$ (SAL, yes) (f, avg0, avg1 must be multiplied by 10^{-3}); p-value corresponds to that of a chi-squared test of equality of $\theta_{u0} = \theta_{u1},\ u = 1, \ldots, 5$.

The proportion of individuals with SAL is relatively large for low level education. We have compared the counts in the two sets of bins from the histograms of the selection probabilities for $y = 0, 1$. In fact, this is a test of independence in a 2×5 categorical table, and we use a chi-squared test of $\theta_{u0} = \theta_{u1}, u = 1, \ldots, U$ ($U = 5$). The p-values are presented in the last column of Table 1. As is evident from the p-value, selection bias should matter mostly in Domain LMW and perhaps in Domains LFW, MMB and HFW. However, to see the differences clearer, we have plotted the sampling distributions (obtained using kernel density estimation) of the selection probabilities in Fig. 1. We have compared the distributions of the two responses ($y = 0$ and $y = 1$) by domain. There are important differences especially in the tails of the distributions. It is worth noting that there are significant differences between the two responses in many domains (e.g., LMW, LFW, LFB, MMB, MFB, HMW and HWB).

To specify $\underset{\sim}{\theta}_y^{(0)}$ in the prior distributions, $\underset{\sim}{\theta}_y \overset{\text{ind}}{\sim} \text{Dirichlet}(\underset{\sim}{\theta}_y^{(0)})$, we take $\underset{\sim}{\theta}_y^{(0)} = \hat{\underset{\sim}{\theta}}_y$, the maximum likelihood estimator of $\underset{\sim}{\theta}_y$, $y = 0, 1$. Following MDC, we show how to obtain the MLE in Appendix B.

In Table 2, we compare summaries of the finite population proportion under the ignorable selection model (IGM) and nonignorable selection model (NIGM) for the 10 domains. We compare the posterior means (PM), posterior standard deviations (PSD) and 95% highest posterior density (HPD) intervals. We note that numerical standard errors of the PMs, obtained by the batch means method, are smaller than 0.001. The effect of the selection bias is seen because the PMs under the NIGM are smaller, and for some domains much smaller, than under the IGM for every domain. It is also interesting that for all domains, except for LMB, the PSDs under NIGM are slightly smaller than under IGM, thereby leading to slightly shorter HPD intervals. But what is remarkable is that all the HPD intervals for NIG, except the one for domain LMB, are to the left of those for IGM and without any overlap. Thus, these selection probabilities have a substantial effect.

In Fig. 2 we have shown plots of the empirical posterior densities of the finite population proportions. (These are obtained using the Rosenbaltt–Parzen kernel density estimator.) These plots show that the proportions are all approximately normally distributed, and this is expected because of the large sample sizes within the domains. More importantly, these plots show that the selection bias is important because, except for the domain LMB, the posterior distributions for the nonignorable selection model are mostly to the left of those for the ignorable selection model.

It is provocative to investigate these results if the population size is much smaller (relative to the sample size) than the original population. So we decide to subsample the data from each domain to get smaller sample and population sizes. For each domain, we take a simple random sample of 20% of the adults and we reduce the population size to 20 times the sample size (i.e., 5% sampling). It becomes necessary to adjust the survey weights to reflect the population sizes. Then we fit both the ignorable and nonignorable selection models to the new data. The posterior summaries are shown

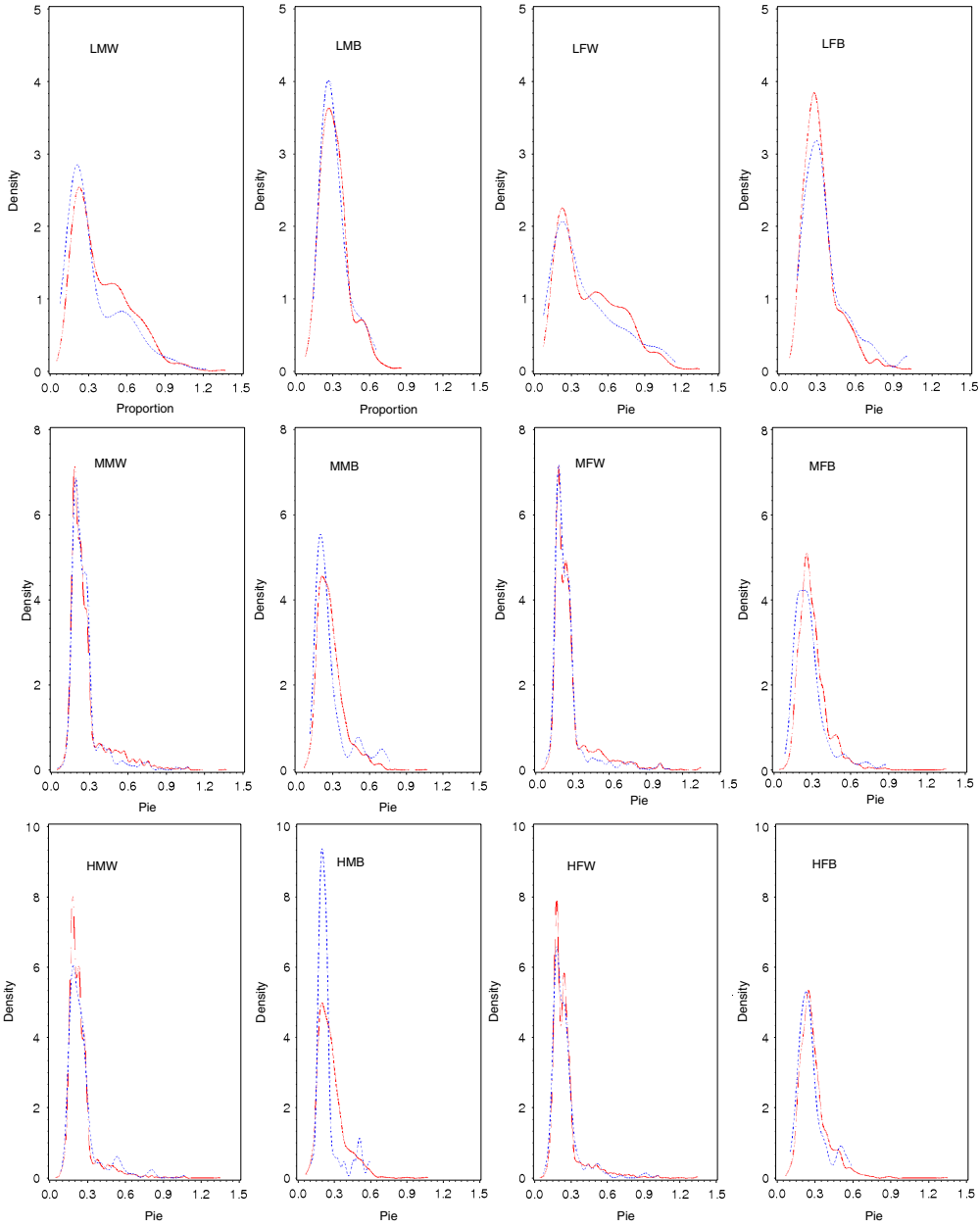

Fig. 1. Plots of the sampling densities of the selection probabilities by domains (dotted: SAL; solid: no SAL). The values on the horizontal axes must be multiplied by 0.0001 to get the selection probabilities.

in Table 3. Again the PMs under IGM are larger (for some domains, more than two times as much) than those under NIGM, and except for domains LMW and LMB, the PSDs under NIGM are smaller than those under IGM. Also, the 95% HPD intervals under NIGM are to the left of those from the IGM, with overlaps in a few domains. Thus, the results are similar to the NHIS 95 data, but there are much sharper distinctions between the two models.

Table 2
Comparison of the ignorable and nonignorable selection models using summaries of the posterior distributions of the finite population proportion by domain.

Domain	Ignorable			Nonignorable		
	PM	PSD	CI	PM	PSD	CI
LMW	0.154	0.009	(0.136, 0.172)	0.083	0.007	(0.070, 0.098)
LMB	0.269	0.025	(0.219, 0.319)	0.192	0.028	(0.143, 0.250)
LFW	0.134	0.008	(0.119, 0.149)	0.068	0.005	(0.059, 0.078)
LFB	0.192	0.020	(0.155, 0.231)	0.113	0.014	(0.089, 0.144)
MMW	0.075	0.003	(0.069, 0.081)	0.024	0.001	(0.022, 0.026)
MMB	0.122	0.008	(0.106, 0.139)	0.060	0.005	(0.052, 0.070)
MFW	0.060	0.002	(0.055, 0.065)	0.019	0.001	(0.017, 0.020)
MFB	0.089	0.006	(0.077, 0.102)	0.043	0.003	(0.037, 0.050)
HMW	0.030	0.002	(0.027, 0.034)	0.009	0.001	(0.008, 0.011)
HMB	0.042	0.006	(0.032, 0.053)	0.019	0.003	(0.014, 0.025)
HFW	0.035	0.002	(0.031, 0.040)	0.010	0.001	(0.009, 0.011)
HFB	0.038	0.005	(0.030, 0.048)	0.016	0.002	(0.013, 0.020)

Note: Domains are formed by crossing education (pre-college: L, college: M and post-college: H), sex (male: M, female: F) and race (white: W, black: B). Here PM is the posterior mean, PSD is the posterior standard deviation and CI is the 95% HPD interval. The numerical standard errors of the PMs are generally much smaller than 0.001 for most domains.

Table 3
Comparison of the ignorable and nonignorable selection models using summaries of the posterior distributions of the finite population proportion with reduced sample size and population size by domain.

Domain	Ignorable			Nonignorable		
	PM	PSD	CI	PM	PSD	CI
LMW	0.212	0.020	(0.175, 0.254)	0.111	0.036	(0.067, 0.208)
LMB	0.315	0.055	(0.213, 0.434)	0.219	0.058	(0.120, 0.342)
LFW	0.167	0.017	(0.136, 0.202)	0.080	0.014	(0.057, 0.108)
LFB	0.142	0.033	(0.087, 0.216)	0.066	0.025	(0.028, 0.121)
MMW	0.089	0.006	(0.077, 0.102)	0.024	0.002	(0.019, 0.029)
MMB	0.176	0.020	(0.138, 0.217)	0.092	0.015	(0.065, 0.121)
MFW	0.082	0.006	(0.072, 0.095)	0.023	0.002	(0.019, 0.028)
MFB	0.106	0.013	(0.082, 0.133)	0.047	0.008	(0.032, 0.063)
HMW	0.037	0.004	(0.028, 0.045)	0.010	0.002	(0.007, 0.013)
HMB	0.045	0.012	(0.024, 0.070)	0.019	0.007	(0.009, 0.034)
HFW	0.046	0.005	(0.037, 0.056)	0.011	0.002	(0.008, 0.015)
HFB	0.061	0.012	(0.040, 0.085)	0.030	0.007	(0.017, 0.046)

Note: Domains are formed by crossing education (pre-college: L, college: M and post-college: H), sex (male: M, female: F) and race (white: W, black: B). Here PM is the posterior mean, PSD is the posterior standard deviation and CI is the 95% HPD interval. The numerical standard errors of the PMs are generally much smaller than 0.002 for most domains. The sample sizes are 20% of the NHIS 95 and the population sizes are 20 times the sample sizes.

In Fig. 3 we have shown plots of the posterior densities of the finite population proportions using these reduced sample sizes and population sizes for the domains. Many of the posterior densities overlap, with a few still not overlapping. Thus, there is clear difference for smaller sample sizes. It appears that when the sample sizes are smaller, the effect of selection bias is smaller. One possible reason for this is that the selection probabilities for adults with SAL and those without may be similar.

We also study sensitivity to the specification of the hyperparameters $\underset{\sim}{\theta}_y^{(0)}$, $y = 0, 1$. Our specifications are $\underset{\sim}{\theta}_y^{(0)} = \hat{\underset{\sim}{\theta}}_y$, $y = 0, 1$, where $\hat{\underset{\sim}{\theta}}_y$ are the MLEs of $\underset{\sim}{\theta}_y$. Our alternative for a sensitivity analysis is $\underset{\sim}{\theta}_y^{(0)} = (1, \ldots, 1)'$, $y = 0, 1$, which we call a 'uniform' prior; here $\tau = 5$. Thus, for the MLE prior, we have $\underset{\sim}{\theta}_y \mid \tau \overset{\text{ind}}{\sim} \text{Dirichlet}(\underset{\sim}{\theta}_y\tau)$ and for the uniform prior we have $\underset{\sim}{\theta}_y \mid \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(1, \ldots, 1)$. (Note that there is no unknown τ in the uniform prior.)

In Table 4, we compare inference of the finite population proportion from the MLE prior and the uniform prior for the original samples and population sizes. The PMs, PSDs and 95% HPD intervals are virtually the same. Thus, we consider using the reduced sample sizes and population sizes for the sensitivity analysis as well. In Table 5, we compare inference of the finite population proportion

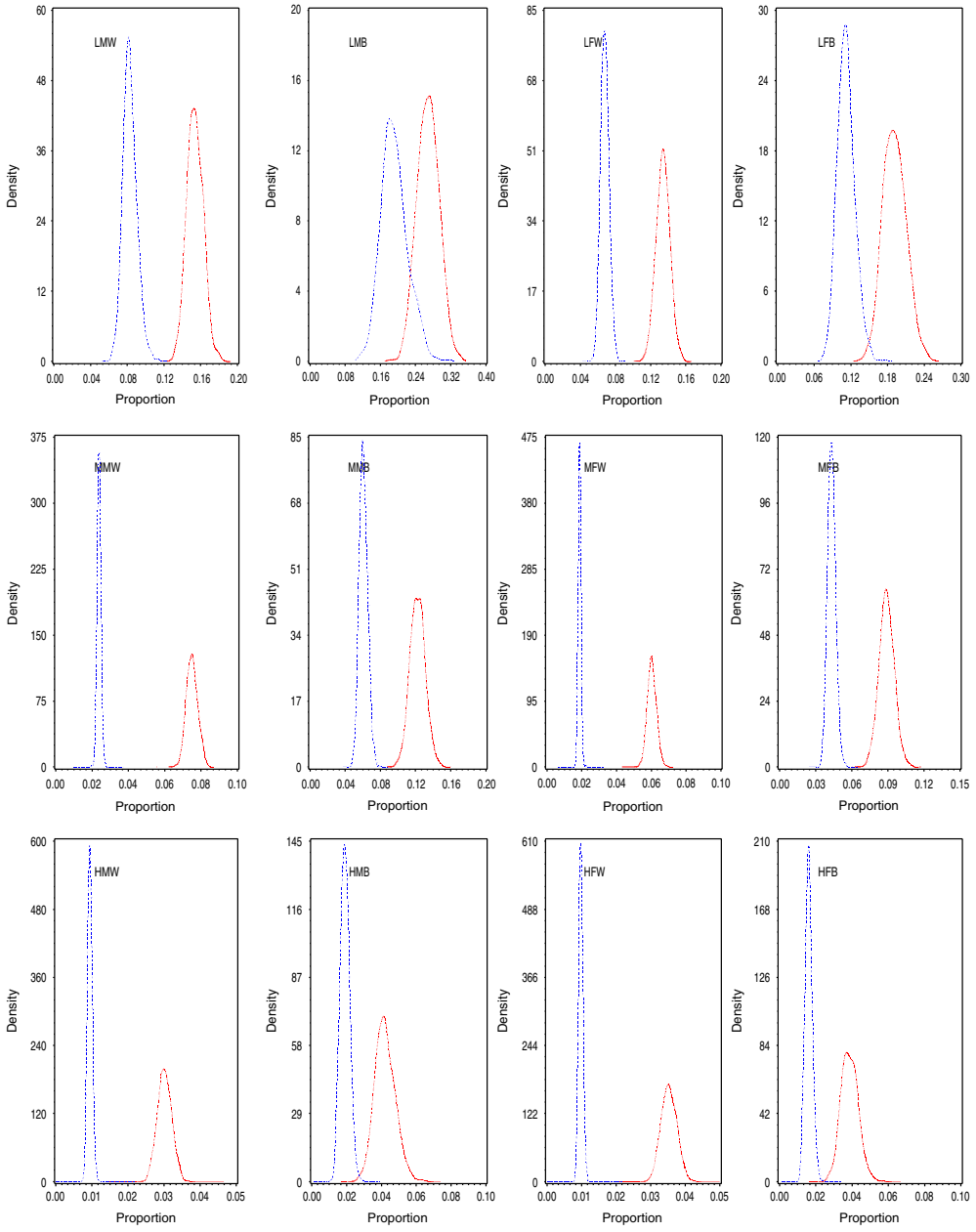

Fig. 2. Plots of the empirical posterior densities of the finite population proportions by domains (dotted: nonignorable selection model; solid: ignorable selection model).

from the MLE prior and the uniform prior for the reduced samples and population sizes. Now there are some differences, and these can be mainly seen in domain LMW; here the posterior means of the finite population proportion are 0.144 vs. 0.145, the posterior standard deviations are 0.034 vs. 0.040 and the 95% HPD intervals are (0.102, 0.238) vs. (0.098, 0.261), but clearly these are small. We compare the posterior densities of the finite population proportions in Fig. 4, where we can hardly see any

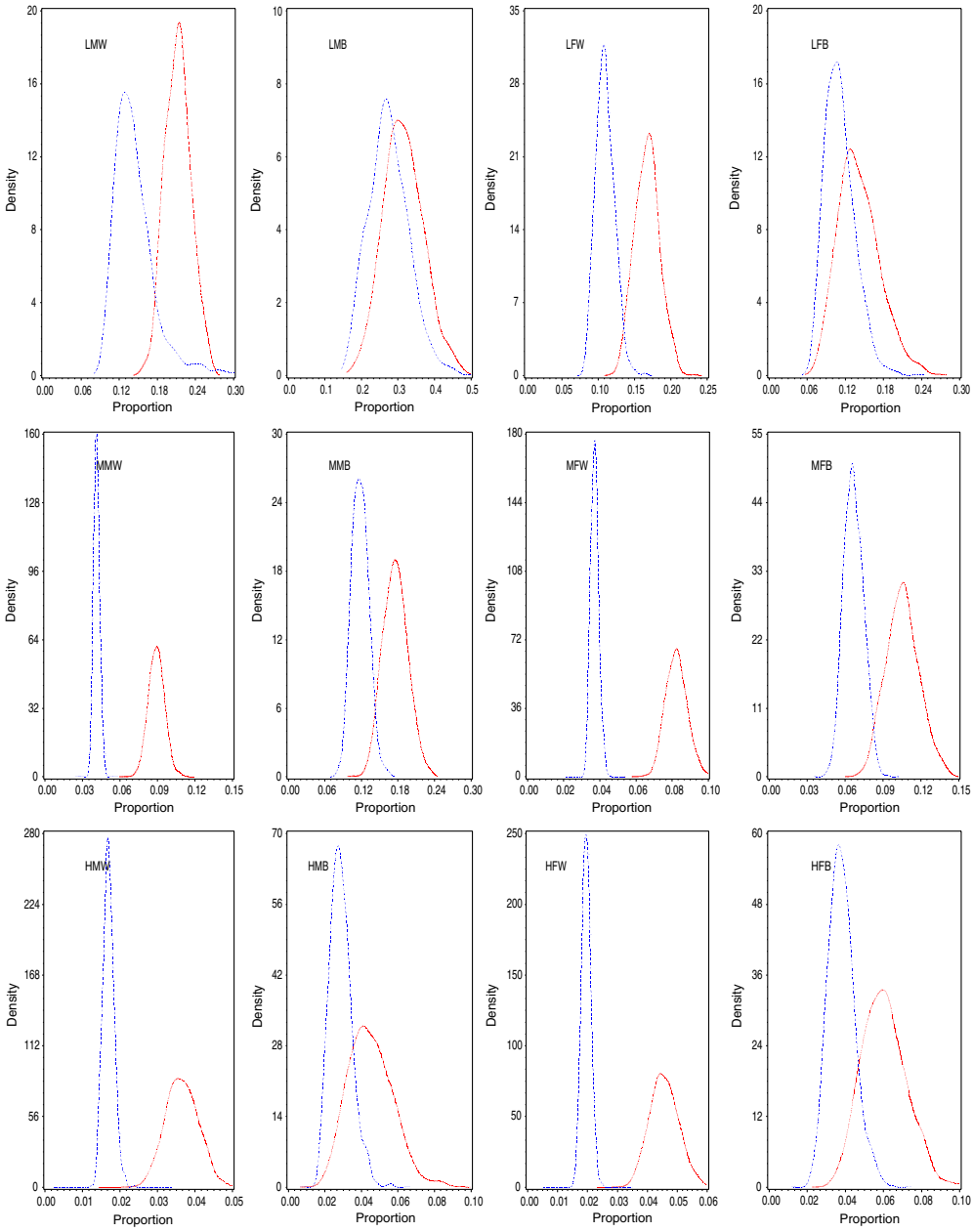

Fig. 3. Plots of the empirical posterior densities of the finite population proportions by domains (dotted: nonignorable selection model; solid: ignorable selection model) for the reduced sample size.

difference in these posterior densities by domain. Thus, a priori we can take $\underset{\sim}{\theta}_y \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(1, \ldots, 1)$ and, indeed, it is informative that we do not really need to bother about specifying $\underset{\sim}{\theta}_y^{(0)}$, $y = 0, 1$ at the MLEs.

Our examples show that it is important to include a component for the selection bias into a model, otherwise there is likely to be misleading estimates of the finite population proportion.

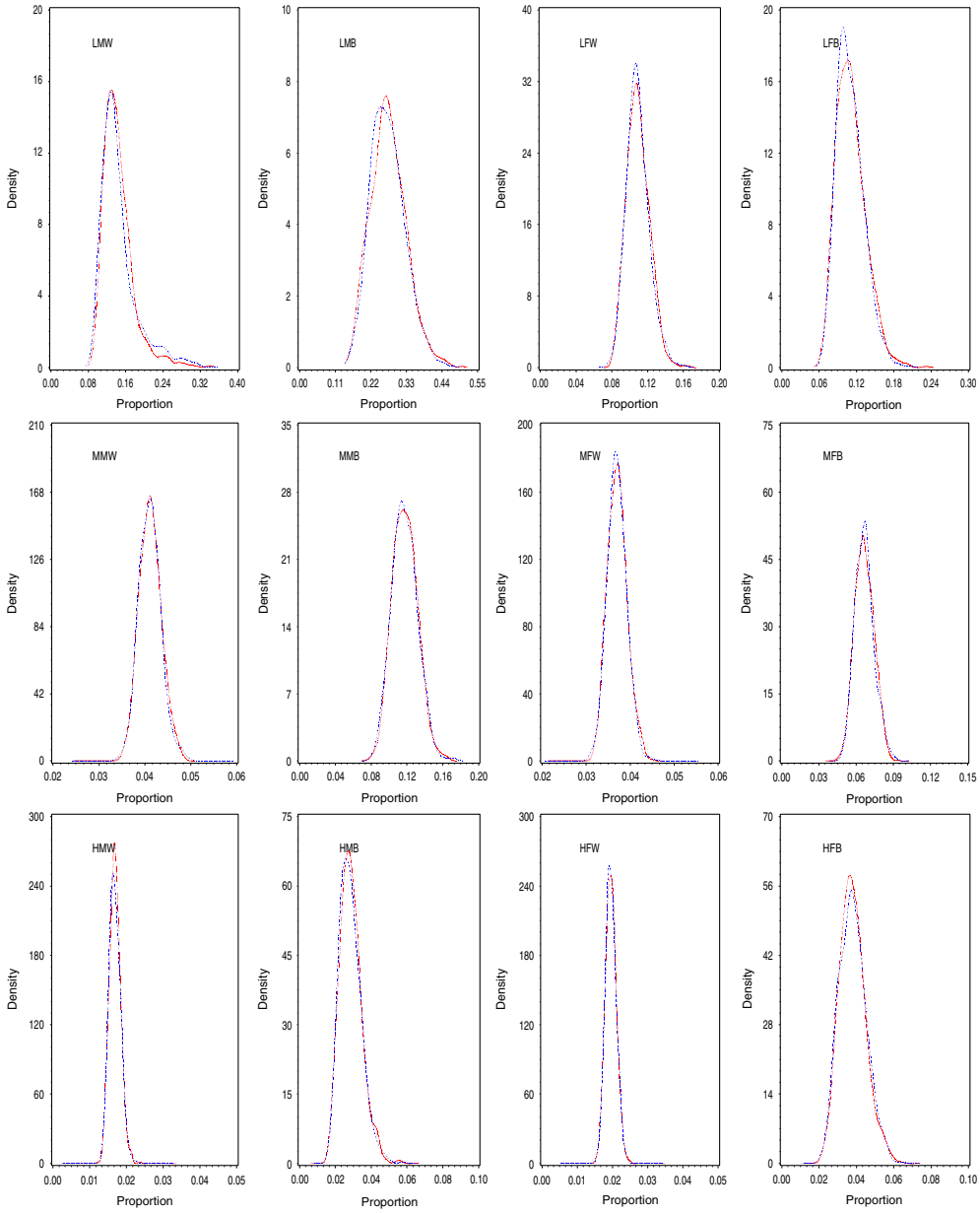

Fig. 4. Plots of the empirical posterior densities of the finite population proportions under the nonignorable selection model by domains (solid: MLE; dotted: Uniform) for the reduced sample size.

4. Concluding remarks

We have constructed a hierarchical Bayesian model to accommodate selection bias when inference is required for a finite population proportion. We have extended the work of Malec et al. [8] in a direction different from that of Nandram and Choi [12]. We have shown how to accommodate a mild nonidentifiability in the model and have provided a full Bayesian analysis when some

Table 4
Sensitivity of posterior inference about the finite population proportion to the prior specification of the nonignorable selection model.

Domain	MLE			Uniform		
	PM	PSD	CI	PM	PSD	CI
LMW	0.083	0.007	(0.070, 0.098)	0.082	0.007	(0.070, 0.098)
LMB	0.192	0.028	(0.143, 0.250)	0.189	0.028	(0.140, 0.247)
LFW	0.068	0.005	(0.059, 0.078)	0.068	0.005	(0.059, 0.078)
LFB	0.113	0.014	(0.089, 0.144)	0.112	0.014	(0.087, 0.142)
MMW	0.024	0.001	(0.022, 0.026)	0.024	0.001	(0.022, 0.026)
MMB	0.060	0.005	(0.052, 0.070)	0.060	0.005	(0.052, 0.070)
MFW	0.019	0.001	(0.017, 0.020)	0.019	0.001	(0.017, 0.020)
MFB	0.043	0.003	(0.037, 0.050)	0.043	0.003	(0.037, 0.050)
HMW	0.009	0.001	(0.008, 0.011)	0.009	0.001	(0.008, 0.011)
HMB	0.019	0.003	(0.014, 0.025)	0.019	0.003	(0.014, 0.025)
HFW	0.010	0.001	(0.009, 0.011)	0.010	0.001	(0.009, 0.011)
HFB	0.016	0.002	(0.013, 0.020)	0.016	0.002	(0.013, 0.020)

Note: Domains are formed by crossing education (pre-college: L, college: M and post-college: H), sex (male: M, female: F) and race (white: W, black: B). Here PM is the posterior mean, PSD is the posterior standard deviation and CI is the 95% HPD interval. The numerical standard errors of the PMs are generally much smaller than 0.002 for most domains. This is a comparison of posterior inference about the finite population proportion using two different prior distributions, $\underset{\sim}{\theta}_y \mid \tau \overset{\text{ind}}{\sim} \text{Dirichlet}(\underset{\sim}{\theta}_y\tau)$, $y = 0, 1$ (MLE) and $\underset{\sim}{\theta}_y \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}\{(1, \ldots, 5)\}$, $y = 0, 1$ (uniform).

Table 5
Sensitivity of posterior inference about the finite population proportion to the prior specification of the nonignorable selection model with reduced sample size and population size.

Domain	MLE			Uniform		
	PM	PSD	CI	PM	PSD	CI
LMW	0.144	0.034	(0.102, 0.238)	0.145	0.040	(0.098, 0.261)
LMB	0.275	0.055	(0.180, 0.391)	0.272	0.052	(0.180, 0.389)
LFW	0.109	0.013	(0.088, 0.135)	0.109	0.013	(0.087, 0.136)
LFB	0.111	0.024	(0.075, 0.163)	0.110	0.022	(0.075, 0.158)
MMW	0.041	0.002	(0.037, 0.046)	0.041	0.002	(0.037, 0.046)
MMB	0.118	0.014	(0.093, 0.146)	0.117	0.014	(0.091, 0.145)
MFW	0.037	0.002	(0.033, 0.042)	0.037	0.002	(0.033, 0.041)
MFB	0.067	0.008	(0.053, 0.082)	0.067	0.008	(0.053, 0.083)
HMW	0.017	0.001	(0.014, 0.020)	0.017	0.002	(0.014, 0.020)
HMB	0.029	0.006	(0.018, 0.043)	0.028	0.006	(0.019, 0.042)
HFW	0.020	0.002	(0.017, 0.023)	0.020	0.002	(0.017, 0.023)
HFB	0.038	0.007	(0.026, 0.054)	0.038	0.007	(0.026, 0.053)

Note: Domains are formed by crossing education (pre-college: L, college: M and post-college: H), sex (male: M, female: F) and race (white: W, black: B). Here PM is the posterior mean, PSD is the posterior standard deviation and CI is the 95% HPD interval. The numerical standard errors of the PMs are generally much smaller than 0.002 for most domains. The sample sizes are 20% of the NHIS 95 and the population sizes are 20 times the sample sizes. This is a comparison of posterior inference about the finite population proportion using two different prior distributions, $\underset{\sim}{\theta}_y \mid \tau \overset{\text{ind}}{\sim} \text{Dirichlet}(\underset{\sim}{\theta}_y\tau)$, $y = 0, 1$ (MLE) and $\underset{\sim}{\theta}_y \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}\{(1, \ldots, 5)\}$, $y = 0, 1$ (uniform).

parameters are specified. We have also shown how to perform the basic Gibbs sampler to provide a Rao–Blackwellized estimator density estimator for the finite population proportion. We have avoided the Metropolis–Hastings sampler in this application because it can perform poorly when there are nonidentifiable parameters in a model. Our nonignorable selection model appears to accommodate the selection mechanism reasonably well. Moreover, all parameters can be stochastic, and we do not need to specify the hyper parameters $\theta_{uy}^{(0)}$.

There are several additional problems which can be solved using our approach. First, we can think about another approach to deal with nonidentifiability. One way to do this is to assume that π comes from a parametric distribution but still maintaining the π_u^*, $u = 1, \ldots, U$. Let (t_{u-1}, t_u), $u = 1, \ldots, U$, denote the bins formed by the histogram of the π_i, the selection probabilities. One can take π_i to have

a common beta distribution truncated in each of the U intervals. So that $\pi_i \mid y \overset{\text{ind}}{\sim} \text{Beta}\{\mu_y\tau, (1-\mu_y)\tau\}$, $y = 0, 1$ and θ_{uy} can be calculated from the beta distribution. This will reduce the complete set of the θ_{uy} parameters to just three parameters, μ_0, μ_1 and τ, because θ_{u0} is a function of μ_0 and τ, and θ_{u1} is a function of μ_1 and τ. This will also allow us to increase the number of bins in the histogram.

A second problem of great interest is on the generalization of our framework to the context of small-area estimation. Now, interest is on the finite population mean of the rth area. Here, our model can take a similar form with different superpopulation proportions in different domains. That is, for ℓ areas

$$p_r \overset{\text{i.i.d.}}{\sim} \text{Beta}\{\mu\tau, (1-\mu)\tau\}, \quad r = 1, \ldots, \ell,$$
$$p(\mu, \tau) = \frac{1}{(1+\tau)^2}, \quad \tau \geq 0.$$

Of course, another avenue is to generalize this work to multinomial data. (A Ph.D. student is currently working on this project.) Clearly, there are numerous situations in which useful work can be done (e.g., misclassification, nonresponse and other survey designs such as cluster sampling where there is correlation within groups of data).

Appendix A. Review of a key result of Malec et al. [8]

Malec et al. [8] models the biased selection mechanism assuming that each response is a sample of size 1 from a stratum of an unknown size. Here, we have a biased sample, $y_1, \ldots, y_n$, from a finite population of size N. Thus, y_i is a sample of size 1 from a stratum of size N_i which is unknown. Without loss of generality, we assume that each sampled unit from each imaginary stratum is the first unit (which itself is representative of the whole stratum). Then, for stratum i, y_{i1} is the observed binary response and the nonsampled units are $y_{i2}, \ldots, y_{i,N_i}$. Here we give an update of the derivation in MDC and we give a faster derivation as well.

Let δ_{ij} represent a Bernoulli variable for the inclusion of the jth unit in the sample from the ith stratum. The selection probability for each of these is π_{ij} so that $\delta_{i1} = 1$ with probability π_{i1} and that $\delta_{ij} = 0$, $j = 2, \ldots, N_i$. For the rest of the Appendix, we drop the subscript i. That is, the first unit is a sample of size 1 from a stratum of size N^* (not to be confused with the population size N). Then, the probability distribution associated with the first individual in the sample of size 1 conditional on the stratum size N^* is

$$\Pr(\delta_1 = 1, y_1, \pi_1, \{\delta_k = 0, y_k, \pi_k\}_{k=2,\ldots,N^*} \mid N^*) = P(\delta_1 = 1|y_1, \pi_1)P(\pi_1|y_1)P(y_1) \times \prod_{k=2}^{N^*} P(\delta_k = 0|y_k, \pi_k)P(\pi_k|y_k)P(y_k),$$

where it is assumed that given N^*, δ_1, y_1, π_1 are independent of δ_k, y_k, π_k, $k = 2, \ldots, N^*$. This may be reasonable since N^* is assumed to be much bigger than 1. Then,

$$\Pr(\delta_1 = 1, y_1, \pi_1, \{\delta_k = 0, y_k, \pi_k\}_{k=2,\ldots,N^*} \mid N^*) = \pi_1 P(\pi_1|y_1)P(y_1) \prod_{k=2}^{N^*} (1-\pi_k)P(\pi_k|y_k)P(y_k), \tag{A.1}$$

where, for simplicity, $P(\delta_k = 1|y_k, \pi_k) = \pi_k = P(\delta_k = 1|\pi_k)$.

By summing over the unobserved components in (A.1), and assuming that the π_i can take one of the known values, π_u^*, $u = 1, \ldots, U$, we get

$$\Pr(\delta_1 = 1, y_1, \pi_1, \{\delta_k = 0\}_{k=2,\ldots,N^*}|N^*) = \pi_1 P(\pi_1|y_1)P(y_1) \times \sum_{y_2}\sum_{\pi_2}\cdots\sum_{y_{N^*}}\sum_{\pi_{N^*}}\{(1-\pi_2)P(\pi_2|y_2)P(y_2)\ldots(1-\pi_{N^*})P(\pi_{N^*}|y_{N^*})P(y_{N^*})\}$$

$$= \pi_1 P(\pi_1|y_1)P(y_1)\left[1 - \sum_y \sum_u \pi_u^* P(\pi_u^*|y)P(y)\right]^{N^*-1}. \tag{A.2}$$

Now, assuming a noninformative prior for N^* with $P(N^*) = 1$, $N^* \geq 1$,

$$\begin{aligned}
&\Pr(\delta_1 = 1, y_1, \pi_1, \{\delta_k = 0\}_{k=2,\ldots,N^*}, N^*) \\
&\quad \propto P(\delta_1, y_1, \pi_1, \{\delta_k = 0\}_{k\neq 1} | N^*) \\
&\quad = \pi_1 P(\pi_1|y_1)P(y_1)\left[1 - \sum_y \sum_u \pi_u^* P(\pi_u^*|y)P(y)\right]^{N^*-1}.
\end{aligned}$$

Since N^* is unknown, we integrate it out. This is accommodated using the formula for the sum of a geometric series with $\sum\sum \pi_u^* P(\pi_u^*|y)P(y) < 1$. We get

$$\begin{aligned}
&\Pr(\delta_1 = 1, y_1 = y, \pi_1 = \pi_u^*, \{\delta_k = 0\}_{k\geq 2}) \\
&\quad \propto \pi_u^* P(\pi_u^*|y)P(y) \sum_{N^*=1}^{\infty}\left[1 - \sum_y \sum_u \pi_u^* P(\pi_u^*|y)P(y)\right]^{N^*-1} = \frac{\pi_u^* P(\pi_u^*|y)P(y)}{\sum_y \sum_u \pi_u^* P(\pi_u^*|y)P(y)}.
\end{aligned}$$

Thus, we have

$$\Pr\left(\delta_1 = 1, y_1 = y, \pi_1 = \pi_u^*, \{\delta_k = 0\}_{k\geq 2} \mid \underset{\sim}{\theta}, p\right) \propto \frac{\pi_u^* \theta_{uy} P(y_1 = y \mid p)}{\sum_y \sum_u \pi_u^* \theta_{uy} P(y_1 = y \mid p)}. \tag{A.3}$$

Finally, the joint conditional probability mass function of π_1 and y_1 is

$$\begin{aligned}
&\Pr(y_1 = y, \pi_1 = \pi_u^* | \underset{\sim}{\theta}, p, \delta_1 = 1, \{\delta_k = 0\}_{k\geq 2}) \\
&\quad \propto \frac{P\left(y_1 = y, \pi_1 = \pi_u^*, \delta_1 = 1, \{\delta_k = 0 \mid \underset{\sim}{\theta}, p\}_{k\geq 2}\right)}{P\left(\delta_1 = 1, \{\delta_k = 0\}_{k\geq 2} \mid \underset{\sim}{\theta}, p\right)}.
\end{aligned}$$

It is now easy to show that

$$\Pr(y_1 = y, \pi_1 = \pi_u^* | \underset{\sim}{\theta}, p, \delta_1 = 1, \{\delta_k = 0\}_{k\geq 2}) = \frac{\pi_u^* \theta_{uy} P(Y_1 = y)}{\sum_y \sum_u \pi_u^* \theta_{uy} P(Y_1 = y)}, \tag{A.4}$$

where an equality sign replaces the proportionality sign. For convenience, we will drop the conditioning on $(\delta_1 = 1, \{\delta_k = 0\}_{k\geq 2})$, although it holds.

It is also interesting that we can develop (A.4) much faster than was done by MDC or our new updated derivation here. Again, let δ_i, π_i, y_i denote the selection indicator, the selection probability and the binary response of the ith unit in the population. Essentially, MDC postulated that the (δ_i, π_i, y_i) are independent and identically distributed with

$$\begin{aligned}
&P(\delta_i = \delta, \pi_i = \pi_u^*, y_i = y \mid \underset{\sim}{\theta}, p) \\
&\quad = P_1(\delta_i = \delta \mid \pi_i = \pi_u^*) P_2(\pi_i = \pi_u^* \mid y_i = y, \underset{\sim}{\theta}) P_3(y_i = y \mid p) \\
&\quad = (\pi_u^*)^{\delta}(1 - \pi_u^*)^{1-\delta}\theta_{uy} p^y (1-p)^{1-y}, \quad \delta = 0, 1, \pi = \pi_u^*,\ u = 1, \ldots,\ U, y = 0, 1.
\end{aligned}$$

Thus, there is a joint probability mass function for the selection indicator and the response indicator. Therefore, the model that MDC specified is a nonignorable selection model (i.e., MDC assumed that the selection mechanism is SNAR). Now because there are no data when $\delta = 0$ (i.e., π and y are both unobserved), MDC used the conditional probability mass function

$$P(\pi_i = \pi_u^*, y_i = y \mid \delta_i = 1, \underset{\sim}{\theta}, p) = \frac{\pi_u^* \theta_{uy} p^y (1-p)^{1-y}}{\sum_{y=0}^{1}\sum_{u=1}^{U} \pi_u^* \theta_{uy} p^y (1-p)^{1-y}},$$

exactly the probability mass function in (A.4). In our work we drop the notation $\delta_i = 1$.

Appendix B. Specifications of the hyperparameters $\theta_{uy}^{(0)}$

For $y = 0, 1$, the maximum likelihood estimators, $\hat{\theta}_{uy}$, are used to specify $\theta_{uy}^{(0)}$, $y = 0, 1$. Note that Malec et al. [8] used the maximum likelihood estimators to obtain a Bayes empirical Bayes analysis.

Using (4), we get $P(\pi_1 = \pi_u^* | Y_1 = y, \underset{\sim}{\theta}) = \pi_u^* \theta_{uy} / \sum_u \pi_u^* \theta_{uy}$. Assuming that the π_i are independent and identically distributed, the likelihood function for each y is

$$\mathrm{L}(\underset{\sim}{\theta}_y) = \prod_u \left(\frac{\pi_u^* \theta_{uy}}{\sum_u \pi_u^* \theta_{uy}} \right)^{g_{uy}}, \quad y = 0, 1. \tag{B.1}$$

This likelihood function corresponds to a multinomial mass function with

$$(g_{1y}, \ldots, g_{Uy})' \mid \underset{\sim}{\theta}_y \sim \text{Multinomial}\left(\sum_{u=1}^{U} g_{uy}, (v_1, \ldots, v_U)' \right), \quad y = 0, 1,$$

where $v_u = \pi_u^* \theta_{uy} / \sum_u \pi_u^* \theta_{uy}$ and we are conditioning on $\sum_{u=1}^{U} g_{uy}$. Therefore, the MLE of $\pi_u^* \theta_{uy} / \sum_u \pi_u^* \theta_{uy}$ is $g_{uy} / \sum_{u=1}^{U} g_{uy}$, $u = 1, \ldots, U$, and by the invariance principle, the MLEs of the θ_{uy} are given by

$$\frac{\pi_u^* \hat{\theta}_{uy}}{\sum_u \pi_u^* \hat{\theta}_{uy}} = \frac{g_{uy}}{\sum_u g_{uy}}. \tag{B.2}$$

In (B.2), $\pi_u^* \theta_{uy} \propto g_{uy}$, $u = 1, \ldots, U$. Therefore, $\pi_u^* \theta_{uy} = k g_{uy}$, where $k = 1 / \sum_u (g_{uy} / \pi_u^*)$. Thus, the MLE of θ_{uy} is $\hat{\theta}_{uy} = (g_{uy} / \pi_u^*) / \sum_u (g_{uy} / \pi_u^*)$, and we specify $\theta_{uy}^{(0)}$ as

$$\theta_{uy}^{(0)} = \frac{(g_{uy} / \pi_u^*)}{\sum_u (g_{uy} / \pi_u^*)}, \quad u = 1, \ldots, U, y = 0, 1.$$

While the MLEs are obtained, other issues will arise when the θ_{uy} are stochastic.

References

[1] R. Chambers, A. Dorfman, S. Wang, Limited information likelihood analysis of survey data, Journal of the Royal Statistical Society. Series B (Statistical Methodology) 60 (1998) 397–411.
[2] Q. Chen, M.R. Elliott, R.J.A. Little, Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling, Survey Methodology 36 (2010) 23–24.
[3] M.-H. Chen, Q.-M. Shao, Monte Carlo estimation of Bayesian credible and HPD intervals, Journal of Computational and Graphical Statistics 8 (1999) 69–92.
[4] A.E. Gelfand, A.F.M. Smith, Sampling-based approaches to calculating marginal densities, Journal of the American Statistical Association 85 (1990) 398–409.
[5] A. Gelman, Prior distribution for variance parameters in hierarchical models, Bayesian Analysis 1 (2006) 515–533.
[6] W. Gilks, P. Wild, Adaptive rejection sampling for Gibbs sampling, Applied Statistics 41 (2) (1992) 337–338.
[7] A. Krieger, D. Pfeffermann, Maximum likelihood estimation from complex sample surveys, Survey Methodology 18 (1992) 225–239.
[8] D. Malec, W.W. Davis, X. Cao, Model-based small area estimates of overweight prevalence using sample selection adjustment, Statistics in Medicine 18 (1999) 3189–3200.
[9] B. Nandram, Bayesian predictive inference under informative sampling via surrogate samples, in: S.K. Upadhyay, Umesh Singh, Dipak K. Dey (Eds.), Bayesian Statistics and Its Applications, Anamaya, New Delhi, 2007, pp. 356–374 (Chapter 25).
[10] B. Nandram, Y. Bai, J.W. Choi, Hierarchical Bayesian models for assessing possible changes in prevalence of activity limitation, Advances and Applications in Statistical Sciences 6 (5) (2011) 285–311.
[11] B. Nandram, M.-H. Chen, Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence, Journal of Statistical Computation and Simulation 54 (1996) 129–144.
[12] B. Nandram, J.W. Choi, A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection, Journal of the American Statistical Association 105 (2010) 120–135.
[13] B. Nandram, J.W. Choi, G. Shen, C. Burgos, Bayesian predictive inference under informative sampling and transformation, Applied Stochastic Models in Business and Industry 22 (2006) 559–572.

[14] R. Natarajan, R.E. Kass, Reference Bayesian methods for generalized linear mixed models, Journal of the American Statistical Association 95 (2000) 227–237.
[15] J.D. Opsomer, G. Glaeskens, M.G. Ranalli, G. Kauermann, F.J. Breidt, Non-parametric small area estimation using penalized spline regression, Journal of the Royal Statistical Society. Series B (Statistical Methodology) 70 (2008) 265–286.
[16] G.P. Patil, C.R. Rao, Weighted distributions and size-biased sampling with applications to wildlife populations and human families, Biometrics 34 (1978) 179–189.
[17] D. Pfeffermann, A.M. Krieger, Y. Rinott, Parametric distributions of complex survey data under informative probability sampling, Statistica Sinica 8 (1998) 1087–1114.
[18] D. Pfeffermann, C.J. Skinner, D.J. Holmes, H. Goldstein, J. Rasbash, Weighting for unequal selection probabilities in multilevel models, Journal of the Royal Statistical Society. Series B (Statistical Methodology) 60 (1998) 23–40. (with discussion).
[19] D. Pfeffermann, M. Sverchkov, Parametric and semi-parametric estimation of regression models fitted to survey data, Sankhyā Series B 61 (1999) 1–21.
[20] D. Pfeffermann, M. Sverchkov, Small-area estimation under informative probability sampling of areas and within selected areas, Journal of the American Statistical Association 102 (2007) 1427–1439.
[21] M. Sverchkov, D. Pfeffermann, Prediction of finite population totals based on the sample distribution, Survey Methodology 30 (2004) 79–92.
[22] M. Torabi, J.N.K. Rao, Mean squared error estimators of small area means using survey weights, Canadian Journal of Statistics 38 (2010) 598–608.