Stat Papers (2019) 60:89–104
https://doi.org/10.1007/s00362-016-0828-x

REGULAR ARTICLE

Application of the full Bayesian significance test to model selection under informative sampling

A. Sikov[1] · **J. M. Stern**[1]

Received: 23 October 2015 / Revised: 13 June 2016 / Published online: 8 September 2016

Abstract Adopting likelihood based methods of inference in the case of informative sampling often presents a number of difficulties, particularly, if the parametric form of the model that describes the sample selection mechanism is unknown, and thus requires application of some model selection approach. These difficulties generally arise either due to complexity of the model holding in the sample, or due to identifiability problems. As a remedy we propose alternative approach to model selection and estimation in the case of informative sampling. Our approach is based on weighted estimation equations, where the contribution to the estimation equation from each observation is weighted by the inverse probability of being selected. We show how weighted estimation equations can be incorporated in a Bayesian analysis, and how the full Bayesian significance test can be implemented as a model selection tool. We illustrate the efficiency of the proposed methodology by a simulation study.

Keywords Informative sampling · Design variables · Inclusion probability · Bayesian significance measures · Horvitz–Thompson estimator · Population distribution · Sample distribution

1 Introduction

Survey sampling distinguishes the cases of non-informative and informative sampling. In the case of informative sampling, the sampling scheme is explicitly or implicitly associated with the variable under investigation. Consequently, the distribution of this

✉ A. Sikov
anna.sikov@mail.huji.ac.il

[1] Institute of Mathematics and Statistics, University of Sao Paulo (IME-USP), Rua do Matao, 1010, Cidade Universitaria, São Paulo CEP: 05508-090, Brazil

variable holding in the sample may be different from the distribution holding in the population. In many practical situations the analyst faces a problem of estimation and identification of the model, holding in the whole population based only on the information contained in a sample drawn from this population. It is generally known that if the sample was selected using a simple random sampling mechanism, then the model holding in the population coincides with the model holding in the sample. However, in survey selection practice it is often the case that the sample selection process employs a more complex mechanism. In this case approximation of the population model by a sample model may lead to biased inference. This often occurs when the sample selection probabilities are proportional to a design variable, which is associated with the variable under investigation (the outcome variable). For example, in business surveys, the inclusion probabilities of sample units (companies) are often proportional to company size, measured by the number of employees. Therefore, if the outcome variable is business income, which is positively correlated with a design variable (company size), this type of sample selection is informative in the sense that large companies with higher income might be overrepresented in the sample, whereas companies with smaller income might be underrepresented. This type of sampling mechanism is widely used in practice, and is referred to as Probability Proportional to Size (PPS) sampling. In our research we assume that probability of each population unit to be selected into the sample is proportional to some design variable, which is associated with an outcome variable. We consider an often practical situation, such as public use data, where the only design information, available to the analyst is the vector of the sample selection probabilities for the sampled units (see Pfeffermann et al. (1998) for discussion). In this case the sample selection mechanism is typically unknown to the analyst, which can considerably complicate estimation of the population model. There exist various approaches to handling informative sampling, however they generally focus on estimation of unknown parameters, leaving the problem of model selection almost unaddressed. Many of such procedures use classical inference methods, such as maximum likelihood estimation based on the approximation of the model holding for the sample measurements, see Pfeffermann et al. (1998), Pfeffermann and Sverchkov (2003) and Pfeffermann and Sverchkov (1999). Other well known procedures use weighting methods, which use either reciprocals of sampling probabilities, as in Binder (1983), Pfeffermann (1993), Pfeffermann (1996), Skinner et al. (1989), or modified sampling probabilities specifically designed for variance reduction, see Beaumont (2008), Kim and Skinner (2013) and Pfeffermann and Sverchkov (1999). In this research, we focus on the problem of model selection strategies to guide the choice of models describing the sampling process. As we shell see in Sect. 2, applicability of the classical approach to model selection may be questionable in the case of informative sampling due to complexity of the resulting models and possibly problems of identifiability. As in many other applications (see for example Ahmadi and Doostparast (2006), Kim et al. (2011), Miazhynskaia and Dorffner (2006)), the problems that arise when using the maximum likelihood framework can be resolved through the use of Bayesian paradigm and MCMC techniques. In this article we develop an approach within the Bayesian framework, which can be used as a model selection tool to choose an appropriate statistical model, via hypothesis testing, using the Full Bayesian Significance Test (FBST) (see Pereira and Stern (1999)). Assum-

ing that the sample inclusion probabilities are observed for all the sampled units, we specify weighted estimation equations and show how they can be incorporated into Bayesian paradigm. Our approach is implemented with the aid of Markov chain Monte Carlo techniques. We further apply the FBST, which is based on the Bayesian measure of evidence favoring the null hypothesis (see Sect. 3).

2 Informative sampling

In this section we provide a brief description of the problem of informative sampling and the principle methods to handle it. Since these methods generally adopt a frequentist approach, suppose first that the parameters, indexing the postulated models are fixed.

Let Y_i denote the value of the outcome variable Y, associated with unit i, belonging to a sample S drawn from a finite population U, and let X_i and V_i denote the vectors of auxiliary variables, associated with unit i. Suppose that the population values of Y_i are independent realizations from a *population* distribution with probability density function $f_p(Y_i|x_i, v_i; \theta)$, where θ is an unknown (vector) parameter, and that the sample selection process is independent between the units. The *sample* distribution $f_s(Y_i|x_i, v_i; \widetilde{\theta})$ with the vector of unknown parameters $\widetilde{\theta}$, is regarded as a *conditional* distribution, given the fact that the unit i has already been selected. Therefore, denoting by I_i the sampling indicator, which takes the value 1 if unit i was selected to the sample and 0 otherwise, we obtain $f_s(Y_i|x_i, v_i; \widetilde{\theta}) = f_p(Y_i|x_i, v_i, I_i = 1; \widetilde{\theta})$.

Following Pfeffermann et al. (1998), the distribution holding in the sample can be expressed as

$$f_s(Y_i|x_i, v_i; \widetilde{\theta}) = f_p(Y_i|x_i, v_i, I_i = 1; \widetilde{\theta}) = f_p(Y_i|x_i, v_i; \theta)\frac{P(I_i = 1|y_i, x_i, v_i; \gamma)}{P(I_i = 1|x_i, v_i; \theta, \gamma)},$$

where γ denotes the vector of unknown parameters, indexing the model for sample selection probabilities and

$$P(I_i = 1|x_i, v_i; \theta, \gamma) = \int f_p(y_i|x_i, v_i; \theta)P(I_i = 1|y_i, x_i, v_i; \gamma)dy_i.$$

Obviously, $\widetilde{\theta} = (\theta, \gamma)$. Note that if a selection probability of a unit i does not depend on the outcome variable (the sample selection process is not informative), then $f_s(Y_i|x_i, v_i; \theta) = f_p(Y_i|x_i, v_i; \theta)$.

In what follows we assume that the population model contains the covariates, x_i, and the model for the sample selection probabilities contains the covariates v_i, so that the sample distribution can be rewritten as

$$f_s(Y_i|x_i, v_i; \theta, \gamma) = f_p(Y_i|x_i; \theta)\frac{P(I_i = 1|y_i, v_i; \gamma)}{P(I_i = 1|x_i, v_i; \theta, \gamma)} \quad (1)$$

Now, let π_i be the sample inclusion probability of a unit i. Noting that

$$P(I_i = 1|y_i, v_i; \gamma) = \int P(I_i = 1|y_i, v_i, \pi_i) f(\pi_i|y_i, v_i; \gamma) d\pi_i = E_p(\pi_i|y_i, v_i; \gamma),$$

the relationship (1) can be also written in the form:

$$f_s(Y_i|x_i, v_i; \theta, \gamma) = f_p(Y_i|x_i; \theta) \frac{E_p(\pi_i|y_i, v_i; \gamma)}{E_p(\pi_i|x_i, v_i; \theta, \gamma)} \tag{2}$$

The representation (2) is quite general and is not restricted to any specific sample selection mechanism. In Sect. 4, we consider the case where the sample selection mechanism involves some design variable Z_i which depends on the outcome variable Y_i and the covariates V_i, such that the probability of a unit i to be selected into the sample is proportional to the value of Z_i.

Note that, in order to derive the distribution holding in the sample, it is sufficient to assume a parametric form for the population distribution, $f_p(Y_i|x_i; \theta)$ and for the expectations of inclusion probabilities $E_p(\pi_i|y_i, v_i; \gamma)$.

Example Let the population distribution be normal, $Y_i|x_i \sim N(\beta_0 + x_i^t\beta, \sigma^2)$, with $\theta = (\beta^t, \sigma^2)^t$ and $E_p(\pi_i|y_i, v_i) \propto \exp(A_1 y_i + g(v_i))$ for some function $g(v)$. Then it is easy to show that $E_p(\pi_i|v_i, x_i) \propto \exp(g(v_i) + A_1 x_i^t\beta + \frac{A_1^2\sigma^2}{2})$, and substituting this expression into (2), after some simple algebra we obtain that the sample distribution is $Y_i|x_i, v_i \sim N((\beta_0 + A_1\sigma^2) + x_i^t\beta, \sigma^2)$.

Following the results obtained in Pfeffermann et al. (1998), which state that, under certain regularity conditions, the sample measurements are asymptotically independent as the population size, N tends to infinity, the sample likelihood can be approximated as

$$L_{Samp} \approx \prod_{i=1}^{n} f_p(Y_i|x_i; \theta) \frac{E_p(\pi_i|y_i, v_i; \gamma)}{E_p(\pi_i|x_i, v_i; \theta, \gamma)}, \tag{3}$$

where the unknown parameters θ and γ index, respectively, the model, holding in the population, and the model underlying sample selection mechanism. The authors show that asymptotic independence holds for various commonly used sampling schemes, including the PPS scheme.

As noted by Pfeffermann et al. (1998), the functional form of the expectations, $E_p(\pi_i|y_i, v_i; \gamma)$ is not necessarily known, and therefore, must be approximated. The authors propose two possible approximations:

$$E_p(\pi_i|y_i, v_i) \approx C_1 \sum_{j=0}^{J} A_j y_i^j + h(v_i) \tag{4}$$

and

$$E_p(\pi_i|y_i, v_i) \approx C_2 \exp\left(\sum_{j=0}^{J} A_j y_i^j + h(v_i)\right), \tag{5}$$

where $h(v_i) = \sum_{p=1}^{m} \sum_{k=1}^{K(p)} B_{kp} v_{ip}^k$, $\{A_j\}$ and $\{B_{kp}\}$ are unknown parameters and C_1 and C_2 are some normalizing constants. However, if the number of parameters indexing the resulting sample model is large, estimation of unknown parameters based on the sample likelihood (3) with the approximations (4) or (5) may lead to complex computations and consequently, unstable estimates which may limit the use of this approach in practical applications. In addition, under approximation (5), the resulting sample distribution may be non-identifiable (as in the example above). In order to avoid identification problems and to facilitate computations, the authors propose to split the estimation process into two steps, where in the first step the parameters A_j and B_{kp} are estimated from the observed inclusion probabilities π_i , and in the second step the parameters, indexing the population model are estimated from the likelihood (3), with the parameters A_j and B_{kp} substituted by their estimates. Apart from solving identifiability and computational problems, this approach utilizes additional information, contained in the vector of inclusion probabilities. The disadvantage of this approach is that it is not directly helpful for choosing a model for the underlying sample selection mechanism. For example, in Eqs. (4) and (5), it is not clear how to determine the polynomial degree, J.

The problem of identifying the form of $E_p(\pi_i|y_i, v_i)$ was partially resolved by Pfeffermann and Sverchkov (1999). The authors propose an approach to test whether the sample selection mechanism is informative, based on the following identity (see Skinner (1994)),

$$E_p(Y_i|x_i, v_i) = \frac{E_s(w_i Y_i|x_i, v_i)}{E_s(w_i|x_i, v_i)}, \tag{6}$$

where the index s implies that the expectations are calculated under the sample distribution, and $w_i = \pi_i^{-1}$ denote the sampling weights. Assuming a linear regression model in the population, and denoting by $\epsilon_i = y_i - E_p(Y_i|x_i)$ the regression residuals, associated with the unit i, one can test the hypothesis of the form $E_p(\epsilon_i^k) = E_s(\epsilon_i^k)$, $k = 1, 2, \ldots$, which by (6) is equivalent to testing that $Corr_s(\epsilon_i^k, w_i) = 0, k = 1, 2, \ldots$, where $Corr_s$ denotes correlation under the sample distribution. The authors point out that it generally suffices to test the first 2–3 correlations. Obviously, this approach can only be useful if the question of the main interest is whether the sample selection mechanism is informative, and therefore, it can not be utilized as a basis for model selection. There exist a few other approaches to test whether the sampling mechanism is informative. All these approaches are generally based on the difference between the estimators of the regression coefficients under the assumed model and the model under ignorable sampling design (see Pfeffermann (1993) for discussion).

Application of the Bayesian approach and the FBST to handle the model selection problem seems a promising solution. We propose to choose a suitable model among

a collection of viable competitors by carrying out pairwise comparisons between the candidate models using hypothesis testing. At each step we compare between nested models by computing the Bayesian measure of evidence favoring the model under the null hypothesis. In particular, this approach can be applied in order to define the polynomial degree, J in the approximations (4) and (5). In Sect. 3 we provide a brief description of the FBST. For a comprehensive review of popular model selection methods in the Bayesian framework, and model comparison criteria, see Miazhynskaia and Dorffner (2006) and Cancho et al. (2012).

In principle, we can opt for a fully Bayesian analysis, based on the sample likelihood (3), however this may require tailored programming to perform the necessary computations. The other potential limitation is that identifiability problems can be encountered when using the approximation (5). In addition, our experience shows that, in order to apply a Bayesian approach, an informative prior distribution for the parameters indexing the sample selection model may be required. For example if approximation (4) is used, it could prove to be impossible to obtain proper mixing and convergence of the MCMC algorithm unless an informative prior distribution is specified. An example of a successful implementation of a Bayesian approach based on the sample likelihood is demonstrated in Pfeffermann et al. (2006). In that application the authors consider a multi-level modeling under informative multi-stage sampling, where the sample model is defined by a hierarchical model, holding in the population, and the first- and lower-level sample selection probabilities. In Sect. 4, we propose an alternative approach, which, on one hand, allows application of Bayesian techniques and the FBST, but on the other hand does not use the sample likelihood defined in (3), thus avoiding the problems mentioned above.

3 The full Bayesian significance test

As previously stated, implementation of the Bayesian approach permits application of the FBST for solving the problem of model selection via hypothesis testing. This can be carried out by determining whether the fitted model contains non-significant parameters. In our application this reduces to testing nested models, where the more complex model is tested versus the model under the null hypothesis, **H**, obtained by setting the coefficients of some group of variables to zero. Therefore, the FBST can be used as a tool for selecting the model which best fits the data within the class of nested models. The FBST is based on the measure of evidence in favor of hypothesis **H**, given the observed data, and is defined as follows.

Let us consider a standard parametric statistical model, i.e., for an integer n, $\theta \in \Theta \subseteq \Re^n$ is the parameter, $g(\theta)$ a prior probability density over Θ, x is the observation (a scalar or a vector), and $L_x(\theta)$ is the likelihood generated by the data x. After the data x have been observed, the sole relevant entity for the evaluation of the Bayesian evidence value ev, is the posterior probability (density) for θ given x, denoted by

$$g_x(\theta) = g(\theta|x) \propto g(\theta)L_x(\theta) \tag{7}$$

We are restricted to the case where the posterior probability distribution over Θ is absolutely continuous, that is $g_x(\theta)$ is a density over Θ. We are focusing on testing of

sharp hypothesis, $H : \theta \in \Theta_H \subset \Theta$. Let $\widetilde{s} = \sup_{\mathbf{H}} g_x(\theta)$ and $T = \{\theta \in \Theta : g_x(\theta) > \widetilde{s}\}$. The Bayesian evidence value agains **H** is defined as the posterior probability of the tangential set, i.e.,

$$\overline{ev} = P(\theta \in T|x) = \int_T g_x(\theta)d\theta \tag{8}$$

The evidence value supporting **H** is $1 - \overline{ev}$, not to be interpreted as an evidence against the alternative hypothesis A. The FBST rejects **H** whenever ev is small, with asymptotic levels prescribed in Pereira et al. (2008).

4 Proposed approach

4.1 Weighted estimation equations and Bayesian model

In what follows we consider a probability proportional to size sampling scheme where the sample selection probabilities, π_i are proportional to the value of the design variable Z_i, which depends on the outcomes Y_i and the covariates V_i. As discussed in Sect. 2, simultaneous estimation of the parameters θ and γ based on the sample likelihood (3), results in laborious computations and unstable estimates. Therefore, following Pfeffermann et al. (1998) we propose to estimate θ and γ separately. In order to define the weighted estimation equations, suppose that the model parameters are fixed. We first consider estimation of the parameters γ indexing a sample selection model. Suppose that

$$E(\pi_i|v_i, y_i; \gamma) = \gamma_v^t v_i + \gamma_{q+1} y_i, \quad i = 1, \ldots, n\ , \tag{9}$$

where $dim(v_i) = q$, $\gamma = (\gamma_v^t, \gamma_{q+1})^t$, $\gamma_v = (\gamma_1, \ldots, \gamma_q)$. Although the defined sample selection model is a special case of model (4), where $J = 1$ and $h()$ is a linear function of the covariates, our approach can be easily extended to the cases where $J > 1$ and $h()$ is a polynomial of an order m, for some $m > 1$.

Let $W(\gamma)$ define the population sum of squares of the regression residuals of the model for the sample selection probabilities,

$$W(\gamma) = \sum_{i=1}^{N} (\pi_i - (\gamma_v^t v_i + \gamma_y y_i))^2, \tag{10}$$

and by $\widetilde{W}(\gamma)$ the Horvitz–Thompson (H–T 1952) estimator of $W(\gamma)$, based on the observed sample S.

$$\widetilde{W}(\gamma) = \sum_{i \in S} \frac{(\pi_i - (\gamma_v^t v_i + \gamma_y y_i))^2}{\pi_i}. \tag{11}$$

Taking the derivatives of (11) with respect to γ yields the weighted estimation equations (originally introduced by Binder (1983)).

Let

$$\widetilde{J}_l = \frac{\partial \widetilde{W}(\gamma)}{\partial \gamma_l} = \sum_{i \in S} \frac{(\pi_i - (\gamma_v^t v_i + \gamma_y y_i))v_{li}}{\pi_i}, \quad l = 1, \ldots, q+1,$$

where $v_{l+1i} = y_i$. Note that the equations specified above, can be alternatively written as

$$\widetilde{J}_l = \frac{\partial \widetilde{W}(\gamma)}{\partial \gamma_l} = \sum_{i=1}^{N} \frac{(\pi_i - (\gamma_v^t v_i + \gamma_y y_i))v_{li}}{\pi_i} I_i, \quad l = 1, \ldots, q+1,$$

where I_i denotes the sampling indicator. Note also that in the specified equations the observed values of the variables Y and V, and of the inclusion probabilities are held fixed, and the only source of randomness is expressed by the sampling indicators $I_1, \ldots, I_N$, which only take the values 0 and 1.

Let

$$\widetilde{J} = (\widetilde{J}_1, \ldots, \widetilde{J}_q, \widetilde{J}_{q+1})^t = \left(\frac{\partial \widetilde{W}(\gamma)}{\partial \gamma_1}, \ldots, \frac{\partial \widetilde{W}(\gamma)}{\partial \gamma_q}, \frac{\partial \widetilde{W}(\gamma)}{\partial \gamma_{q+1}} \right)^t, \tag{12}$$

and

$$J = (J_1, \ldots, J_q, J_{q+1})^t = \left(\frac{\partial W(\gamma)}{\partial \gamma_1}, \ldots, \frac{\partial W(\gamma)}{\partial \gamma_q}, \frac{\partial W(\gamma)}{\partial \gamma_{q+1}} \right)^t. \tag{13}$$

Note that

$$E(\widetilde{J}_l | \pi, \mathbf{y}, \mathbf{v}) = \sum_{i=1}^{N} \frac{(\pi_i - (\gamma_v^t v_i + \gamma_y y_i))v_{li}}{\pi_i} E(I_i) = J_l = 0_{(q+1)\times 1},$$

where $0_{(q+1)\times 1}$ denotes a vector of zeros of dimension $(q+1)$ and π, $\mathbf{y}$, $\mathbf{v}$ denote the vector of inclusion probabilities and the vectors of the population realizations of the variables Y and V, respectively. Thus, the following equations can be obtained:

$$\widetilde{J} = 0_{(q+1)\times 1} + \nu, \tag{14}$$

where $\nu = (\nu_1, \nu_2, \ldots, \nu_{q+1})^t$ is a $(q+1)$-variate random variable. Implication of Eq. (14) is that even were the value of γ known, it is unlikely that the components of $\widetilde{J}$ would be equal to zero for any selected sample S, due to sampling variability, although we expect some of them to be close to zero.

Suppose now, that the vector of unknown parameters γ is random. Since the components of the vector $\widetilde{J}$ are defined in terms of sums of random variables, we assume

that given γ, the vector of random variables ν can be approximated by a $(q+1)$-variate normal distribution, that is,

$$\nu|\gamma, \pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs}, \approx N(0_{(q+1)\times 1}, \Sigma(\gamma)), \tag{15}$$

where π_{obs}, $\mathbf{y}_{obs}$ and $\mathbf{v}_{obs}$ denote the observed parts of the vectors π, $\mathbf{y}$ and $\mathbf{v}$, respectively. A well known result of classical sampling theory states (see Cochran (1977)) that the components of the variance matrix $\Sigma(\gamma)$ can be derived as follows:

$$\left[\Sigma(\gamma)\right]_{k,l} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\widetilde{e_i}\widetilde{e_j} v_{ki} v_{lj}}{\pi_i \pi_j}(\pi_{ij} - \pi_i \pi_j), \tag{16}$$

where $\widetilde{e_i} = \pi_i - (\gamma_v v_i + \gamma_y y_i)$.

Then, based on the sample measurements, the variance matrix in (16) can be estimated as

$$\left[\hat{\Sigma}(\gamma)\right]_{k,l} = \sum_{i=1}^{n}\sum_{j=1}^{n} \widetilde{e_i}\widetilde{e_j} v_{ki} v_{lj} \left(\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}}\right). \tag{17}$$

Note that calculation of the components of the variance matrix $\hat{\Sigma}(\gamma)$ requires knowledge of the joint inclusion probabilities π_{kl} for the units k and l, where $k, l = 1, \ldots, n$. These probabilities are generally unavailable, however, they can be obtained using the Hajek approximation of π_{ij}, proposed in Hajek (1964), which is applicable in the case of PPS sampling method (see also Berger (2004) for discussion),

$$\pi_i \pi_j - \pi_{ij} \approx \pi_i \pi_j (1 - \pi_i)(1 - \pi_j) d^{-1},$$

where $d = \sum_{i=1}^{N} \pi_i (1 - \pi_i)$.

Therefore, (16) can be rewritten as

$$\left[\Sigma(\gamma)\right]_{k,l} \approx -d^{-1} \sum_{i=1}^{N} \widetilde{e_i} v_{ki} (1 - \pi_i) \sum_{j=1}^{N} \widetilde{e_j} v_{lj} (1 - \pi_j), \tag{18}$$

which can be estimated as

$$\left[\hat{\Sigma}(\gamma)\right]_{k,l} = -\hat{d}^{-1} \sum_{i=1}^{n} \frac{\widetilde{e_i} v_{ki}}{\pi_i} (1 - \pi_i) \sum_{j=1}^{n} \frac{\widetilde{e_j} v_{lj}}{\pi_j} (1 - \pi_j), \tag{19}$$

where $\hat{d} = \sum_{i=1}^{n} (1 - \pi_i)$

In order to apply a Bayesian approach we use the model (15) and a diffuse normal prior centered around zero on γ. It should be noted that a diffuse normal prior is the typical example of the so-called just proper prior, which is proper but is very close to being a flat prior (Congdon (2007)).

Then we obtain that

$$f(\gamma|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs}) \propto \phi_{0_{(q+1)\times 1}, \hat{\Sigma}(\gamma)}(\widetilde{J}|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs}; \gamma)h(\gamma), \tag{20}$$

where $\phi_{0_{(q+1)\times 1}, \hat{\Sigma}(\gamma)}$ denotes a $(q+1)$-variate normal density function with the expectation vector, $0_{(q+1)\times 1}$ and the variance matrix $\hat{\Sigma}(\gamma)$, and $h(\gamma)$ denotes the prior distribution on γ.

It is important to emphasize, however, that the proposed method does not constitute a canonical Bayesian approach, in the sense that it is not based on an observed data likelihood. In the proposed method the role of observations plays the vector $\widetilde{J}$, the components of which depend on the parameters γ. However, as we have seen, given γ, the distribution of $\widetilde{J}$ can be approximated by a multivariate normal distribution, thus allowing for incorporation of $\widetilde{J}$ in the Bayesian formalism. Obviously, the proposed method is applicable to the situations where the sample selection mechanism follows model (5).

The approach described above can also be applied to estimation of parameters indexing the population model. For example, if the population model is given by $Y_i|x_i \sim N(x_i\beta, \sigma^2)$, we can define $U(\beta) = \sum_{i=1}^{N}(y_i - x_i\beta)^2$ and the corresponding Horvitz–Thompson estimators,

$$\widetilde{U}(\beta) = \sum_{i=1}^{n} \frac{(y_i - x_i\beta)^2}{\pi_i}. \tag{21}$$

If the population model belongs to the family of GLM with a canonical link function,

$$f_p(Y_i|x_i;\theta) = exp\left\{a(\phi)\left[y_i\sum_{k=0}^{h}\beta_k x_{ki} - g\left(\sum_{k=0}^{h}\beta_k x_{ki}\right) + d(y_i)\right] + \eta(\phi, y_i)\right\},$$

where x_i is of dimension $(h+1)$, $\theta = (\beta^t, \phi)$ defines the set of unknown parameters and $g(.)$, $a(.)$, $d(.)$ and $\eta(.)$ are known real functions with $g()$ strictly increasing and differentiable, then Eq. (21) can be rewritten as

$$\widetilde{U}(\beta) = \sum_{i=1}^{n} \frac{(y_i - g'(\sum_{k=0}^{h}\beta_k x_{ki}))^2}{\pi_i}.$$

Then, application of the same reasoning as in the case of estimation of γ, yields a normal approximation, as in (15), where the corresponding matrix of variances is calculated and approximated similarly to (16) and (19).

4.2 Application of the FBST

In this section we consider a simple case of hypothesis testing under the sampling model defined by (9), $H : \gamma_{q+1} = 0$, where rejection of H implies that the sample

selection mechanism is not informative. Extension of the proposed approach to the case of hypotheses of the form $\widetilde{H} : \widetilde{\gamma} = 0$, where $Span(\gamma \setminus \widetilde{\gamma}) \subseteq Span(\gamma)$ for more complex sample selection models, is straightforward.

Denote by $\hat{\gamma}^{post}$ the vector of random draws from the posterior distribution $f(\gamma|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs})$, obtained using the MCMC technique, in the case of the full model, and let $\dim(\hat{\gamma}^{post}) = K \times (q+1)$. Let γ_0 be the vector of unknown parameters, indexing the sampling selection model under $\mathbf{H}$. Then, the mode of the posterior distribution of γ under H is defined as

$$\hat{\gamma}_0 = \text{argmax}_{\gamma \in \mathbf{H}} \phi_{0_{(q+1)\times 1), \hat{\Sigma}(\gamma)}}(\widetilde{J}|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs}; \gamma)h(\gamma) \tag{22}$$

Now let

$$T = \left\{\gamma : f(\gamma|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs}) > f(\hat{\gamma}_0|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs})\right\} \tag{23}$$

It follows then (see Lauretto et al. (2003), Pereira and Stern (1999), Pereira et al. (2008), Rifo and Bernardini (2011)) that derivation of the value $\overline{ev}$, requires computation of the probability $P(T|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs})$. We propose to estimate this probability using the posterior draws $\hat{\gamma}^{post}$ as follows.

$$P(T|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs}) \approx \frac{1}{K}\sum_{k=1}^{K} I_{\left\{\frac{f(\gamma_k|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs})}{f(\hat{\gamma}_0|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs})} > 1\right\}}, \tag{24}$$

where

$$\begin{aligned} &\frac{f(\gamma_k|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs})}{f(\hat{\gamma}_0|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs})} \\ &= \frac{\phi_{0_{((q+1)\times 1), \hat{\Sigma}(\gamma_k)}}(\widetilde{J}|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs}; \gamma_k)h(\gamma_k)}{\phi_{0_{(q\times 1), \hat{\Sigma}(\hat{\gamma}_0)}}(\widetilde{J}|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs}; \hat{\gamma}_0)h(\hat{\gamma}_0)} \rho(\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs}) \end{aligned} \tag{25}$$

and

$$\rho(\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs}) = \frac{\int \phi_{\mathbf{0}_{((q+1)\times 1), \hat{\Sigma}(\gamma)}}(\widetilde{J}|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs}; \gamma)h(\gamma)d\gamma}{\int \phi_{\mathbf{0}_{(q\times 1), \hat{\Sigma}(\gamma_0)}}(\widetilde{J}|\pi_{obs}, \mathbf{y}_{obs}, \mathbf{v}_{obs}; \gamma_0)h(\gamma_0)d\gamma_0} \tag{26}$$

Therefore, application of the FBST requires carrying out the two following steps:

1. A maximization step, defined in (22).
2. Computation of the ratio of the integrals, defined in (26).

The first step constitutes a standard maximization problem, which can usually be carried out by application of the Newton–Raphson algorithm, which generally does not require laborious computations. The second step is usually much more computationally demanding. In order to facilitate the computation of the integrals involved, the MCMC techniques can be applied, as described by Stern and Zacks (2003). In that

work, the authors also show how to determine the required MCMC run length for attaining the desired precision of the evidence value.

5 Simulation study

In order to test the approach described in this article, we performed a simulation study consisting of three experiments. For these experiments, we generated $M = 500$ populations of size N, ($N = 500, 5000$) and for each generated population we selected one sample of size $n = 50$, where the units were randomly selected with inclusion probabilities proportional to the values of the design variable Z, defined below. Thus we consider two cases: in the first case the sampling fraction was equal to 0.1, while in the second case it was equal to 0.01. It should be noted that in practice, sampling fraction generally varies from one study to another. The choice of these sampling fractions was made in order to mimic the behavior of the tested approaches under two different scenarios.

In order to generate the data, we used the population and sample selection models, defined below.

$$Y_i = 3.5 + 0.8x_i - 0.1v_i + \epsilon_i, \quad i = 1, \ldots, N \tag{27}$$

where $\epsilon_i \sim N(0, 1.5)$ and the auxiliary variables X and V were generated from $Gamma(1, 1)$ and $Poisson(3)$ distributions, respectively. The true model for the design variable Z is defined as:

$$Z_i = 4 + 2.5v_i + 0.15y_i^2 + \nu_i, \tag{28}$$

where $\nu_i \sim N(0, 2.5)$.

The main objective of the experiments was to identify the model holding for $E(\pi_i|v_i, y_i)$ by testing the following hypotheses (recall that $\pi_i \propto Z_i$):

1. $H_1 : E(\pi_i|v_i, y_i) = \gamma_0 + \gamma_1 v_i$ **vs** $E(\pi_i|v_i, y_i) = \gamma_0 + \gamma_1 v_i + \gamma_2 y_i$
2. $H_2 : E(\pi_i|v_i, y_i) = \gamma_0 + \gamma_1 v_i + \gamma_2 y_i$ **vs** $E(\pi_i|v_i, y_i) = \gamma_0 + \gamma_1 v_i + \gamma_2 y_i + \gamma_3 y_i^2$
3. $H_3 : E(\pi_i|v_i, y_i) = \gamma_0 + \gamma_1 v_i$ **vs** $E(\pi_i|v_i, y_i) = \gamma_0 + \gamma_1 v_i + \gamma_2 y_i^2$

For each experiment, $j = 1, 2, 3$ and each sample $i, i = 1, \ldots, M$ we computed the Bayesian evidence value in favor of $\mathbf{H}_j$, ev_{ji}, using a diffuse normal prior, as presented in Sect. 4.2. As mentioned above, the FBST rejects $\mathbf{H}$ whenever ev is small. In order to define a rejection region, we considered the asymptotic distribution of the evidence value under $\mathbf{H}$, provided by Pereira et al. (2008). Besides the FBST, we applied the Likelihood Ratio test (LR), based on the model defined in (15), and the approach based on (6) (see Pfeffermann and Sverchkov (1999) for details), which, as previously mentioned, can only be implemented to testing informativeness of the sampling selection mechanism and is not applicable if both tested models are informative. Therefore, this test was not applied in Experiment 2. It must be emphasized that the classical LR test, based on the likelihood (3), is generally difficult to implement due to its computational complexity and potentially unstable estimators (see Sect. 3).

Table 1 Proportions of samples, where **H** was rejected (N = 500)

Significance level	Experiment 1			Experiment 2		Experiment 3		
	FBST	LR	Pfeffermann and Sverchkov (1999)	FBST	LR	FBST	LR	Pfeffermann and Sverchkov (1999)
0.010	0.921	0.908	0.910	0	0.116	0.955	0.928	0.938
0.025	0.966	0.940	0.955	0.007	0.176	0.980	0.952	0.964
0.050	0.977	0.960	0.977	0.062	0.284	0.980	0.964	0.964
0.100	0.989	0.984	0.977	0.205	0.396	0.995	0.986	0.992

Table 2 Proportions of samples, where **H** was rejected (N = 5000)

Significance level	Experiment 1			Experiment 2		Experiment 3		
	FBST	LR	Pfeffermann and Sverchkov (1999)	FBST	LR	FBST	LR	Pfeffermann and Sverchkov (1999)
0.010	0.768	0.734	0.710	0	0.100	0.800	0.778	0.790
0.025	0.812	0.788	0.786	0.004	0.152	0.842	0.810	0.830
0.050	0.860	0.818	0.832	0.050	0.242	0.884	0.832	0.868
0.100	0.934	0.888	0.908	0.176	0.340	0.944	0.906	0.926

Tables 1 and 2 summarize the proportions of samples for which the hypothesis **H** was rejected by each competitor method, under various significance levels, for $N = 500$ and for $N = 5000$ correspondingly. In these tables, higher empirical rejection levels indicate a lower Type II error rate, the probability of accepting the hypothesis when it is false.

The results indicate that for this simulation study the FBST presents good power properties for the first and third experiments, outperforming the alternative methods. As expected, all approaches showed lower power in the case of $N = 5000$. The high probability of rejection of **H** in the first experiment implies that our method succeeds in revealing that the sampling mechanism is indeed informative. Since it is impossible to test the model $E(\pi_i|v_i, y_i) = \gamma_0+\gamma_1 v_i+\gamma_2 y_i$ versus $E(\pi_i|v_i, y_i) = \gamma_0 + \gamma_1 v_i + \gamma_2 y_i^2$ (see Sect. 3), we propose to compare the evidence values obtained in the first and the third experiments. The power in the third experiment was higher, as we tested a non-informative sample selection model against the correct model (recall that the true model for the sampling selection process included the quadratic term of the value of outcome variable). For the second experiment, neither model is correct. The results indicate that, in this case, FBST and LR tend to not reject the model with a smaller number of coefficients.

In order to validate the significance levels of the competitor methods, we carried out an additional experiment (Experiment 4) with the same hypothesis **H** tested in the Experiment 1, but generating the design variable Z under the model:

$$Z_i = 4 + 2.5v_i + \nu_i, \tag{29}$$

Table 3 Nominal and empirical significance levels under **H**

Nominal levels	Empirical levels		
	FBST	LR	Pfeffermann and Sverchkov (1999)
0.025	0.014	0.032	0.033
0.050	0.037	0.056	0.044
0.100	0.074	0.096	0.083
0.250	0.200	0.228	0.200
0.500	0.510	0.464	0.510
0.750	0.749	0.732	0.778
0.900	0.900	0.896	0.883
0.950	0.946	0.936	0.939

where $v_i \sim N(0, 2.5)$.

Table 3 presents the nominal and empirical significance levels yielded in Experiment 4, in the case of $N = 500$. The results for $N = 5000$ are in general very similar. Notice that, in this experiment, samples are generated in accordance to the hypothesis to be tested, and, therefore, we should expect that the simulated significance levels (proportions of samples where the hypothesis **H** is rejected) should be close to the nominal significance level. Denoting by $\overline{ev}_{4,k}$ the evidence value against **H**, obtained from the sample k, $k = 1, \ldots, 500$, the empirical significance level for the FBST, corresponding to the nominal level α_j was computed as $\frac{1}{500} \sum_{k=1}^{500} I(\overline{ev}_{4,k} < \overline{ev}^{cr}_{(\alpha_j)})$, where $\overline{ev}^{cr}_{(\alpha_j)}$ is the critical value for nominal level α_j, based on the asymptotic distribution of $\overline{ev}$, provided by Pereira et al. (2008). The empirical levels for two other tests were calculated in a similar way, where the evidence values were substituted by the corresponding test statistics, and the critical values for nominal levels α_j were computed using the parametric bootstrap procedure. In general, the empirical significance levels of the methods are close to the corresponding nominal values, thus validating the use of all the discussed methods.

6 Concluding remarks

In this article we consider a problem of model selection via hypotheses testing under informative sampling, and apply the FBST, in order to test different model forms. We consider the case where the sample inclusion probabilities are known, and are utilized for estimation of model parameters and application of the FBST. The method can be recommended for analysis of survey data, which ordinarily contain inclusion probabilities (for example, files released for public use). As we mentioned previously, our approach can be applied, if the sample units are selected with probabilities proportional to some design variable (for example business or household surveys). In this case the proposed approach does not require knowledge of the joint inclusion probabilities, which are usually unknown. The results of the empirical study illustrate that

the proposed method has good power properties, however, as is any new approach, there is always need for extensions. Application of the proposed method requires a prior specification of the form of the sampling selection mechanism, which is usually unknown to the analyst. The most interesting and important question in this respect is whether different parametric forms of the sampling mechanism can be compared (for example, (4) vs. (5)). Another interesting question is whether the results are sensitive to the choice of the prior. We would also like to note that the proposed method can be applied to the problem of nonresponse in household surveys. As noted by Saärndal and Swensson (1987), nonresponse can be viewed as the result of a self selection process with usually unknown response probabilities. On the other hand in many household surveys the main reason for nonresponse is "not at home", where the larger households has larger probabilities to respond (to be selected in the sample of the respondents). In this case design information (household size) is generally observed for all the responding units, whereas the probabilities to respond are unavailable. We are investigating these problems, and we hope to be able to report our results in the near future.

Acknowledgements The authors are grateful for the support of IME-USP, the Institute of Mathematics and Statistics of the University of São Paulo; FAPESP - the State of São Paulo Research Foundation (grant CEPID 2013/07375-0 and 2013/17746-5); and CNPq - the Brazilian National Counsel of Technological and Scientific Development (grant PQ 301206/2011-2). Finally, the authors are grateful for the advice of colleagues and anonymous referees used to improve this work.

References

Ahmadi J, Doostparast M (2006) Bayesian estimation and prediction for some life distributions based on record values. Stat Pap 47:373–392

Beaumont JP (2008) A new approach to weighting and inference in sample surveys. Biometrika 95(3):539–553

Berger YG (2004) A simple variance estimator for unequal probability sampling without replacement. J Appl Stat 31:305–315

Binder DA (1983) On the variances of asymptotically normal estimators from complex surveys. Int Stat Rev 51(3):279–292

Cancho VG, Castro M, Rodrigues J (2012) A Bayesian analysis of the Conway–Maxwell–Poisson cure rate model. Stat Pap 53:165–176

Cochran WJ (1977) Sampling techniques. Wiley, Chichester

Congdon P (2007) Bayesian statistical modelling, 2nd edn. Wiley, Chichester

Hajek J (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. Ann Math Stat 35:1491–1523

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47:663–685

Kim C, Jung J, Chung Y (2011) Bayesian estimation for the exponentiated Weibull model under type II progressive censoring. Stat Pap 52:53–70

Kim JK, Skinner CJ (2013) Weighting in survey analysis under informative sampling. Biometrika 100(2):385–398

Lauretto M, Pereira CAB, Stern JM, Zacks S (2003) Full Bayesian signicance test applied to multivariate normal structure models. Braz J Prob Stat 17:147–168

Miazhynskaia T, Dorffner D (2006) A comparison of Bayesian model selection based on MCMC with an application to GARCH-type models. Stat Pap 47:525–549

Pereira CAB, Stern JM (1999) Evidence and credibility: full Bayesian signicance test for precise hypotheses. Entropy 1:69–80

Pereira CAB, Stern JM, Wechsler S (2008) Can a significance test be genuinely Bayesian? Bayesian Anal 3(1):79–100

Pfeffermann D (1993) The role of sampling weights when modeling survey data. Int Stat Rev 61(2):317–337
Pfeffermann D (1996) The use of sampling weights for sampling data analysis. Stat Methods Med Res 5:239–261
Pfeffermann D (2011) Modelling of complex survey data: why is it a problem? how should we approach it? Surv Methodol 37(2):115–136
Pfeffermann D, Sverchkov M (1999) Parametric and semi-parametric estimation of regression models fitted to survey data. Sankhya 61:166–186
Pfeffermann D, Sverchkov M (2003) Fitting generalized linear models under informative sampling. In: Skinner C, Chambers R (eds) Analysis of survey data. Wiley, New York, pp 175–196
Pfeffermann D, Krieger AM, Rinott Y (1998) Parametric distributions of complex survey data under informative probability sampling. Stat Sin 8:1087–1114
Pfeffermann D, Moura FAS, Silva PLM (2006) Multi-level modelling under informative sampling. Biometrica 93:943–959
Rifo LLR, de Bernardini DF (2011) Full Bayesian significance test for extremal distributions. J Appl Stat 38:851–863
Saärndal CE, Swensson B (1987) A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. Int Stat Rev 55:279–294
Skinner CJ (1994) Sample models and weights. In Proceedings of the section on survey research methods, American Statistical Association, pp 133–142
Skinner CJ, Holt D, Smith TMF (1989) Analysis of Complex Surveys. Wiley, Chichester
Stern JM, Zacks S (2003) Testing the independence of poisson variates under the holgate bivariate distribution: the power of a new evidence test. Stat Probab Lett 66:313–320