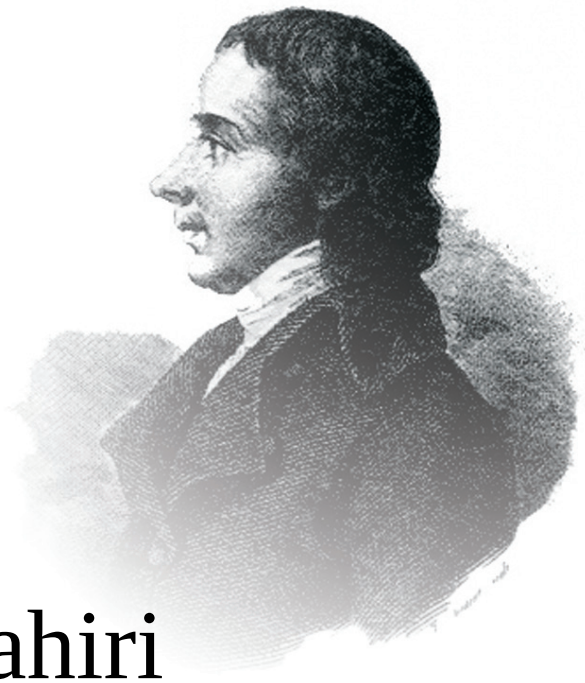
Rapporto n. 4 2007
dmsia unibg.it



E. Fabrizi and P. Lahiri
A design-based approximation to the BIC in finite population sampling

Serie Ricerca

Dipartimento
di Matematica, Statistica,
Informatica e Applicazioni
"Lorenzo Mascheroni"

UNIVERSITÀ DEGLI STUDI DI BERGAMO

A design-based approximation to the BIC in finite population sampling

BY

ENRICO FABRIZI,

DMSIA, University of Bergamo,

Via dei Caniana 2, 24127 Bergamo, Italy,

enrico.fabrizi@unibg.it

PARTHA LAHIRI,

JPSM, University of Maryland,

1218 LeFrak Hall, College Park MD 20742, U.S.A.,

plahiri@survey.umd.edu

Key words: Bayes factor; Model selection; Hypothesis testing; Pseudo-maximum likelihood; Cluster sampling.

ABSTRACT. In this article, various issues related to the implementation of the usual Bayesian Information Criterion (BIC) are critically examined in the context of modelling for finite populations. A suitable design-based approximation to the BIC is proposed in order to avoid the derivation of the exact likelihood of the sample which is often very complex in a finite population sampling. The approximation is justified using a theoretical argument and a Monte Carlo simulation study.

1. Introduction

Survey researchers frequently use statistical models. However, such models, known as a superpopulation models (Deming and Stephen, 1941), are generally used to describe the finite populations of interest and has been used earlier for evaluation, sampling design development and making inferences on either the relevant superpopulation or the finite population parameters. In the analytic use of survey data (Deming 1953) where the main goal is to address various scientific questions, inference for the superpopulation parameters is more important than that for the finite population parameters. Graubard and Korn (2002) discussed the importance of inference for superpopulation parameters using survey data and cited a number of practical examples such as the estimation of superpopulation means, linear regression and logistic regression coefficients using complex survey data from the U.S. National Health Interview Survey, the third National Health and Nutrition Examination Survey and the 1986 National Hospital Discharge Survey.

Selection of a model among different plausible models has received considerable attention in the statistical literature. The Institute of Mathematical Statistics (IMS) monograph on model selection edited by Lahiri (2002) contains four long review articles that critically examine various classical and Bayesian approaches to model selection. For further important developments in the Bayesian literature on the

subject see also Spiegelhalter *et al*. (2002). The impact of the superpopulation model misspecification has been studied by Holt et al. (1980), Hansen et al. (1983), among others. However, to our knowledge the related issue of model selection, especially the well-known likelihood based methods such as the BIC has received little attention in the survey research literature.

If the full specification of the superpopulation likelihood for the finite population is possible, there is conceptionally no problem in deriving the likelihood for the complex sample and so an extension of the BIC to finite population sampling in such a situation is quite straightforward. Finite populations studied in many social and economic surveys are in general very complex and heterogeneous. For this reason some survey researchers feel "that most statistical models in finite population inference are either wrong or (at best) incomplete" (Kott, 1989). Also, the public-use file for the sample may not contain all the relevant information required to specify the complete likelihood of the sample. In this paper, we discuss a simple approach to approximate the BIC for the analysis of complex survey data that avoids the complete specification of the sample likelihood.

In Section 2, we review the Bayes factor and its relation to the BIC for hypothetical infinite population. In Section 3, we critically examine two possible ways to adapt the BIC in the context of the finite population sampling. The first approach consists in finding a formula for the BIC based on the superpopulation likelihood for the finite population and then estimating this finite population BIC. We argue that this model selection criterion does not even work for a simple hypothesis testing problem, a special case of model selection, with data collected by a simple random sampling with replacement. This approach makes the disagreement between the data and the null hypothesis look more than it really is. The second approach is the BIC based on the sample likelihood. This certainly provides us a meaningful model selection criterion. However, quite often we do not have the complete information in the public-use file to specify the complete likelihood for the sample. In Section 3, we discuss the impact of model misspecification on the BIC based on the sample likelihood.

In Section 4, we propose a new model selection criterion which is essentially the Wald's statistics based on the weighted survey estimator of the superpopulation parameter of interest and its randomization-based variance estimator. Our model selection criterion is robust and can be used, for example, to test the significance of a regression coefficient with unspecified distribution for the error term using complex survey data. We show that under certain regularity conditions, the new model selection criterion is indeed an approximation to the BIC for a large sample. In Section 5, we verify the regularity conditions for two commonly used sampling designs. We provide results from a Monte Carlo simulation studies in Section 6. Our simulation results demonstrate the good performance of the new criterion in a complex situation involving clustered binary data with unknown intra-cluster correlation and in a regression problem with heteroskedastic errors and unequal selection probabilities.

2. The Bayes Factor and the BIC

The Bayesians frequently use the Bayes factor (BF) in hypothesis testing and model selection problems. To illustrate the BF, let $y_s = (y_1,...,y_n)$ be an independent and identically distributed (*iid*) sample from a distribution belonging to a family of probability distributions parameterized by (β,θ) with $\dim(\beta,\theta) = m$ and $\dim(\beta) = m_0$. Consider the following hypothesis testing problem:

$$M_0 : \theta = \theta_0 \qquad \text{versus} \qquad M : \theta \in \mathbb{R}^{m-m_0} \tag{1}$$

The BF is defined to be the ratio of the *aposteriori* and the *apriori* odds in favour of the larger model *M*:

$$BF = \frac{prob(M \mid y_s)}{prob(M_0 \mid y_s)} \Big/ \frac{prob(M)}{prob(M_0)} = \frac{\int p(y_s \mid \beta,\theta)\pi(\beta,\theta)\,d\beta d\theta}{\int p(y_s \mid \beta,\theta_0)\pi_0(\beta)\,d\beta}. \tag{2}$$

The calculation of the BF requires a full specification of the prior distributions for the parameters in both M_0 and *M*. In many applications rules for "objectively" selecting priors have been proposed (see Berger and Pericchi 2001). Alternatively, one can use a suitable approximation to the logarithm of the BF. One popular approximation is the Bayes Information Criterion (Schwartz, 1978) given by:

$$S = \lambda - \frac{m - m_0}{2} \log n ,$$

where $\lambda = \ell\left(\hat{\beta}, \hat{\theta}\right) - \ell_0\left(\tilde{\beta}, \theta_0\right)$, the logarithm of the likelihood ratio, and $\ell\left(\hat{\beta}, \hat{\theta}\right)$, $\ell_0\left(\tilde{\beta}, \theta_0\right)$ denote the log-likelihood under *M* and M_0 evaluated at the corresponding maximum likelihood estimators $(\hat{\beta}, \hat{\theta})$ and $\left(\tilde{\beta}, \theta_0\right)$ of (β, θ) respectively.

The statistic *S* is based on the Laplace approximation to the integrals appearing in the numerator and denominator of (2). See Kass and Wassermann (1995) for details. The quality of the approximation *S* to the logarithm of the Bayes Factor depends on the prior distributions of the unknown parameters under M_0 and *M*. In general, it is rather crude since it neglects terms up to a constant order. Nonetheless, Kass and Wassermann (1995) showed that for a suitable choice of the prior distributions (e.g., unit information prior)

$$S = \log BF + O_p\left(n^{-1/2}\right).$$

Moreover, the Bayes Information Criterion is also popular among the frequentists because it incorporates a penalized deviance criterion. It should also be stressed that *S* is a consistent model selection method, i.e., if one of the hypotheses (models) being tested is true, the BIC selects the true hypothesis with probability 1 as the sample size tends to infinity. For the problem (1), *S* goes to $+\infty$ $\left(-\infty\right)$ with probability one if *M* (M_0) is true.

In a hypothesis testing problem, *S* can be compared against the scale of evidence introduced by Jeffreys (1961) as an alternative to the frequentist scale of evidence

introduced by R.A. Fisher in the 1920s. For a discussion and comparison between the Jeffreys' and Fisher's scales of evidence, see Efron and Gous (2001).

3. Two possible approaches to adapt bic to the finite population sampling

For this section and the rest of the paper, we need some notations. Let $U = \{1, \cdots, N\}$ denote the units of a finite population of known size *N*. Let $y_U = (y_1, ..., y_N)$, where y_i is the value of a characteristic of interest for the *i*th unit of the finite population $(i = 1, \cdots, N)$. Let $p(s)$ be the probability of drawing a particular sample *s* from the universe of all possible samples *S*. Thus, $p(s) \geq 0$ and $\sum_{s \in S} p(s) = 1$. Let $d_s = \{d_i : i \in s\}$, where d_i contains all possible design and other auxiliary information on the unit $i \in s$. For example, d_i may contain information on the label and sampling weight w_i for the unit $i \in s$. The sampling weight w_i is defined as the inverse of the inclusion probability for the unit *i* and represents a certain number of units in the finite population. Define $y_s = \{y_i : i \in s\}$ and $z_s = [d_s, y_s]$.

In the following two subsections, we shall discuss two possible approaches to extend the BIC to select model for the superpopulation for y_U and point out possible difficulties in implementing them in complex survey data analyses.

3.1. *An estimator of the finite population BIC*

Let the observations y_i $(i = 1, \cdots, N)$ of the finite population be generated randomly from $N(\theta, 1)$. Consider the following hypothesis testing problem, a special case of model selection:

$$M_0 : \theta = 0 \qquad M_1 : \theta \neq 0. \tag{3}$$

If all units of the finite population were observed, then it is easy to see that the BIC based on all the observations in the finite population is given by

$$S_{POP}(y_U) = \frac{N}{2} \bar{y}_U^2 - \frac{1}{2} \log N. \tag{4}$$

We call $S_{POP}(y_U)$ the finite population BIC. Of course, we cannot use $S_{POP}(y_U)$ since $\bar{y}_U$ is unknown. Let $\hat{\bar{y}}_U$ be a design consistent estimator of $\bar{y}_U$. An estimator is design consistent estimator of the corresponding finite population parameter if the estimator approaches in the probability induced by the sampling design to the true finite population parameter as $n \to \infty$. Replacing $\bar{y}_U$ by a design-consistent estimator $\hat{\bar{y}}_U$, the following naïve model selection criterion is obtained:

$$S_{Plugin}(z_s) = \frac{N}{2}\hat{\bar{y}}_U^2 - \frac{1}{2}\log N. \tag{5}$$

Remark: We observe that, since $n \le N$, the limit $n \to \infty$ makes sense only in a setting in which the population size N is also allowed to increase. We assume a mathematical definition of the limit for $n \to \infty$ which is consistent with most literature on inference in finite population sampling. A description of this framework may be found in Isaki and Fuller (1982).

We note that the simple plug-in approach as described above does not work even for a simple random sampling with replacement. Under this sampling design, when N is very large compared to n (the sample size), one would expect a reasonable finite population sampling implementation of S to be very close to the following standard BIC S_{IID} obtained under the assumption of independently and identically distributed observations from a normal population:

$$S_{IID}(y_s) = \frac{n\bar{y}_s^2}{2} - \frac{1}{2}\log n. \tag{6}$$

This is a reasonable expectation since in this case simple random sampling from a finite population can be regarded as a random sample from the assumed hypothetical superpopulation.

But, if we replace $\bar{y}_U$ in (5) by the usual design-consistent estimator $\bar{y}_s$, we obtain:

$$S_{Plugin}(z_s) - S_{IID}(y_s) = \frac{(N-n)}{2}\bar{y}_s^2 - \frac{1}{2}\ln\left(\frac{N}{n}\right). \tag{7}$$

This difference tends to 0 when $n \to N$ but, for n fixed, it diverges to infinity as $N \to \infty$ and not to 0 as we would like. This implies that for N large enough, (5) provides stronger evidence against M_0 than (6) does. The reason is that (5) approximates the S we would have obtained if all the units in the finite population were observed and thereby making the disagreement between the data and the null hypothesis look more than it really is.

3.2. *The BIC based on the exact likelihood for the sample*

Like in the standard BIC calculation for a hypothetical infinite population, this approach is also based on the sample likelihood. However, we must obtain the sample likelihood using the superpopulation model for the finite population and the sampling design used. For an informative sampling, this is quite complicated since under such sampling the sample likelihood could be very different from the finite population likelihood. Even for non-informative sampling, spcification of a reasonable sample likelihooh could be a formidable task for a variety of reasons. Survey populations usually have complex structures and misspecification of the assumed model is quite likely (see Kott 1991). Also, quite often the analyst may not have all the necessary information in the public-use file which makes modeling difficult. We now illustrate this point through a simple example.

Let the observations in the finite population be normally distributed with common mean θ. We assume that the observations within the same cluster are equally correlated, the common intra-cluster correlation being τ. Furthermore, observations from two different clusters are assumed to be uncorrelated. We consider the same testing problem on the overall population mean θ as in (3).

For the finite population described in the previous paragraph, a cluster sampling is often employed. Suppose we have a finite population of size N divided into M clusters each of size N_c. A sample of m clusters is selected by simple random sampling (with replacement) and all the units of the sampled cluster are selected. Thus, $n = mN_c$. In this case a suitable model for y_s is given by

$$y_{ij} = \theta + \alpha_j + e_{ij},$$

where α_j and e_{ij}'s are all uncorrelated with $V(\alpha_j) = \tau$ <1and $V(e_{ij}) = (1-\tau)$ for $j = 1,...,m \quad i = 1,...,N_c$. Note that marginally $V(y_i) = 1$ so we are consistent with the model assumed in the *iid* case. This leads to $\bar{y}_s$ as the maximum likelihood estimator of θ and to the following BIC:

$$S(z_s) = \frac{1}{2}\frac{n\bar{y}_s^2}{\{1+(N_c-1)\tau\}} - \frac{1}{2}\log n. \qquad (8)$$

We note that

$$S_{IID}(y_s) - S(z_s) = \frac{n\bar{y}_s^2}{2}\left\{\frac{(N_c-1)\tau}{1+(N_c-1)\tau}\right\},$$

where $S_{IID}(y_s)$, given in (6), is the appropriate *BIC* when there is no clustering of the population units. The error increases with τ. In other words, if we neglect the clustering of the population units, we shall reject the null hypothesis more often than we really should.

Unfortunately, unlike the previous example the likelihood for the sample may be very complicated and in some cases it may be even impossible to write down. To this end, reconsider the same hypothesis testing problem of (3) based on a probability proportional to size with replacement sampling in which the size variable *X* is positively correlated with the target variable *Y*. One can consider a model for $f(y_s \mid d_s = x_s)$. However, we are interested in testing a hypothesis for the superpopulation mean θ that characterizes the marginal distribution of *Y* and not the mean conditional on *X*. Since the sampling design is not simple random sampling and larger values of *X* are more likely to be observed, we need to obtain a marginal likelihood for y_s by integrating out x_s:

$$f(y_s) = \int f(y_s \mid x_s)\, f(x_s)\, dx_s.$$

This is certainly not as simple as the previous example. Actually, when analyzing data from complex surveys we observe a sample from $Y \mid d_s$. A researcher may not

be interested in $f(y \mid d_s)$ but may be interested in an appropriate marginal model - one that averages out some of the population features incorporated in the sampling design. For instance, one may be interested in testing hypothesis about the overall population mean, ignoring possible differences among the means of different subgroups of the population. In general, some degree of aggregation in modelling may be necessary in complex surveys from a finite population (see Holt, 1989).

In any case, the calculation of the BIC based on the sample likelihood requires that we use all information needed to specify a suitable model for $f(y_s \mid d_s)$. This may not be the case in many applications. It is typical that the analyst may not be provided with all the information about the sample design but only with the survey weights, defined as the inverse of the inclusion probabilities and adjusted for post-stratification and non-response.

4. A Robust Design-Based Approximation to the BIC.

Let y_U be a realization from an underlying superpopulation distribution characterized by a parameter θ. We are interested in testing $M_0 : \theta = \theta_0$ vs $M_a : \theta \neq \theta_0$. In this case the *BIC* is given by $S = \lambda - \frac{1}{2}\log n$, where $\lambda = \ell(\hat{\theta}) - \ell(\theta_0)$ is the logarithm of the likelihood ratio.

As noted in the previous section, it is often difficult or even impossible to obtain an exact expression of the sample likelihood due to a complex population structure. In this section, we shall consider a design-based approximation to the *S*. The approximation essentially involves an estimator of θ using the following method and its design consistent variance estimator.

Let $U(y_U, \theta) = 0$ be an estimating equation for θ. The solution $T(y_U)$ of the equation $U(y_U, T(y_U)) = 0$ is known as the corresponding descriptive population quantity (CDPQ) of θ. We can estimate $T(y_U)$ by a design-based estimator $\hat{T}(z_s)$.

For example, $\hat{T}(z_s)$ could be obtained using the pseudo-maximum likelihood approach. A thorough discussions of this class of methods can be found in the book of Sarndal, Swensson and Wretman (1992).

We propose the following model selection criterion:

$$S_{DB} = \frac{1}{2} W_{DB} - \frac{1}{2} \log n, \tag{9}$$

where $W_{DB} = \left\{ \hat{V}_D \left(\hat{T}(z_s) \right) \right\}^{-1} \left(\hat{T}(z_s) - \theta_0 \right)^2$ and $\hat{V}_D \left(\hat{T}(z_s) \right)$ is a design-consistent estimator of $V_D \left(\hat{T}(z_s) \right)$, the variance of $\hat{T}(z_s)$ under the randomization distribution.

In sample surveys, the design effect is defined as $Deff = V_D \left(\hat{T}(z_s) \right) \left\{ V_{SRS} \left(\hat{T}(y_s) \right) \right\}^{-1}$, where $V_{SRS} \left(\hat{T}(y_s) \right)$ is the randomization variance of the un-weighted estimator $\hat{T}(y_s)$ of $T(y_U)$ under a simple random sampling of size *n*. The design effect corrects the variability of the survey estimator for the complexities in the survey design. For, example, both weighting and clustering usually increase the variability of a survey estimator and in such cases the design effect helps inflating the simple random sampling variance. See Lynn and Gabler (2005). The effective sample size is defined as $n^* = n / Deff$ and it can be interpreted as the sample size of a simple random sampling that is as efficient as the corresponding sample size for a complex survey. Usually, n^* is smaller than n.

We obtain a different model selection criterion, say $S_{DB}(n^*)$, when we replace *n* by n^* in (9). However, we note that the order of $\log(Deff)$ is often small compared to $\ell(\hat{\theta}) - \ell(\theta_0)$. Thus, asymptotically $S_{DB}(n^*) \cong S_{DB}(n)$ in most cases since $\log(n^*) = \log(n) - \log(Deff)$.

The following proposition shows that S_{DB} approximates *S* well, the error of approximation being lower than the one used to approximate $\log BF$ by S.

Proposition 1: Assume the following regularity conditions**:**

(i) $\left(\hat{\theta}-\theta_0\right)=O_\xi\left(n^{-1/2}\right)$ under model M_0, where $O_\xi\left(n^{-1/2}\right)$ denotes a stochastic order with respect to the superpopulation distribution ξ;

(ii) the log-likelihood function, i.e., $\ell(\theta)$ is twice differentiable with $-\ell''\left(\hat{\theta}\right)=I\left(\theta_0\right)+O_\xi\left(n^{-1/2}\right)$, where $I(\theta_0)=-E\left\{\frac{\partial^2 l(z,\theta)}{\partial\theta^2}\right\}_{\theta=\theta_0}$ is the Fisher information matrix evaluated at θ_0;

(iii) $\hat{T}\left(z_s\right)=\hat{\theta}+o_{D\xi}\left(n^{-1/2}\right)$, where $o_{D\xi}\left(n^{-1/2}\right)$ denotes a stochastic order with respect to the compound model/randomization distribution $D\xi$;

Moreover assume that

(iv) $\hat{V}_D\left(\hat{T}\left(z_s\right)\right)=\left\{I\left(\theta_0\right)\right\}^{-1}+o_{D\xi}\left(n^{-1}\right)$.

Under regularity conditions (i)-(iii) and assumption (iv) we have

$S-S_{DB}=o_{D\xi}\left(n^{-1/2}\right)$.

Proof. Using the Taylor series expansion of $\ell\left(\hat{\theta}\right)$ around θ_0, we have

$$\lambda=\ell\left(\hat{\theta}\right)-\ell\left(\theta_0\right)=-\frac{1}{2}\ell''\left(\theta_0\right)\left(\hat{\theta}-\theta_0\right)^2+o_\xi[\left(\hat{\theta}-\theta_0\right)^2]$$

so that regularity conditions (i) and (ii) imply

$$\lambda=-\frac{1}{2}\ell''\left(\theta_0\right)\left(\hat{\theta}-\theta_0\right)^2+o_\xi\left(n^{-1/2}\right)=\frac{1}{2}I\left(\theta_0\right)\left(\hat{\theta}-\theta_0\right)^2+o_\xi\left(n^{-1/2}\right).$$

Now using regularity conditions (iii) and (iv), we have

$$W_{DB}=I\left(\theta_0\right)\left(\hat{\theta}-\theta_0\right)^2+o_{D\xi}\left(n^{-1}\right).$$

The theorem now follows from the fact that $S_{DB}=\frac{1}{2}W_{DB}-\frac{1}{2}\log n$ and

$S=\lambda-\frac{1}{2}\log n$.

Remark: We note that the regularity conditions of Kass and Wassermann (1995), given in their section 2, are analogous to our assumptions i) and ii). Thus, we can conclude that under i) and ii) and unit information priors

$$\log BF = S_{DB} + O_{pD\xi}\left(n^{-1/2}\right).$$

5. Two examples

In this section, we shall verify the regularity conditions needed to prove Theorem 1 for two well-known sampling designs and the associated superpopulation models.

Example 1: One stage cluster sampling and the associated one-way random effects model (as in Skinner 1989, p. 37).

Consider a clustered finite population described by the following superpopulation model

$$y_{ij} = \theta + \alpha_j + e_{ij},$$

where α_j and e_{ij} are uncorrelated with $V\left(\alpha_j\right) = \tau\sigma_0^2$ and $V\left(e_{ij}\right) = \left(1-\tau\right)\sigma_0^2$, $j = 1,...,M$, $i = 1,...,N_c$. Note that τ can be interpreted as the intra-cluster correlation coefficient.

Suppose we are interested in testing $M_0 : \theta = \theta_0$ vs $M_a : \theta = \theta_0$ based on a one-stage cluster sample in which *m* clusters are selected by a simple random sample without replacement.

For the one-way random effects model, we have $\hat{\theta} = \overline{y}_s$. Condition (i) is a standard property of the maximum likelihood estimator in regular problems. In order to verify condition (ii), note that $I\left(\theta_0\right) = \dfrac{n}{\left[1+\left(N_c-1\right)\tau\right]\sigma_0^2}$, (see Searle, Casella and McCullogh, p. 80) and the fact that the log-likelihood function is a quadratic form with $-\ell\left(\hat{\theta}\right)$ free from θ and y_s. Under the sampling design, $\hat{T}\left(z_s\right) = \hat{\theta}$ so condition (iii) is trivially verified.

Turning to condition (iv), we note that under the cluster sampling design: $T\left(y_U\right) = \overline{y}_U$, $\hat{T}\left(z_s\right) = \overline{y}_s$. Thus, $\hat{\theta} = T\left(y_U\right)$ and

$$\hat{V}_D(\bar{y}_s) = \frac{N-n}{N}\frac{\left[1+(N_c-1)\tau\right]}{n}s_y^2,$$

where $s_y^2 = (n-1)^{-1}\sum_{i\in s}(y_i - \bar{y}_s)^2$. Condition (iv) can now be verified by showing that $E\left\{\hat{V}_D(\bar{y}_s)\right\} = \left\{I(\theta_0)\right\}^{-1} + o\left(n^{-1}\right)$ and $V\left\{\hat{V}_D(\bar{y}_s)\right\} = o\left(n^{-2}\right)$, where *E* and *V* denote the expectation and the variance with respect to both the sampling design and the model.

Example 2. Two-stage sampling and the associated one-way random effects model

We consider the same one-way random effects model and the same testing problem for a two-stage sampling where *m* primary stage units (psu) are first selected by a simple random sample without replacement and then n_c secondary stage units are randomly selected from each sampled psu. In this case, it can be shown that $\hat{\theta} = \bar{y}_s$ and $I(\theta_0) = \frac{n}{\left[1+(n_c-1)\tau\right]\sigma_0^2}$.

Verification of conditions (i)-(iii) is similar to that of the one-stage cluster sampling case. To verify condition (iv), we first note that $T(y_U) = \bar{y}_U$, $\hat{T}(z) = \bar{y}_s$, and

$$\hat{V}_D(\bar{y}_s) = \frac{N-n}{Nn}s_{yt}^2 + \frac{1}{N}\left(\frac{N_0}{n_0} - 1\right)s_{ye}^2,$$

where $s_{yt}^2 = \frac{n_c}{m-1}\sum_{j=1}^{m}(\bar{y}_j - \bar{y}_s)^2$ and $s_{ye}^2 = \frac{n_c}{m(n_0-1)}\sum_{j=1}^{m}\sum_{i=1}^{n_c}(y_{ij} - \bar{y}_j)$ (see Cochran, 1977, Theorem 10.2). Noting that $s_{ty}^2 = \frac{n-1}{m-1}\frac{1}{n_c}\left[1+(n_c-1)\tau\right]s_y^2$ we have

$$\hat{V}_D(\bar{y}_s) \simeq \frac{N-n}{Nn}\left[1+(n_c-1)\tau\right]s_y^2 + \frac{1}{N}\left(\frac{N_c}{n_c} - 1\right)s_{ye}^2.$$

Verification of condition (iv) is now similar to that of Example 1.

6. Monte Carlo simulation

As mentioned in the introduction, the main advantage of our proposed model selection criterion S_{DB} is that it can be applied even when the exact *BIC* (S_E) cannot be obtained because of the unavailability of the exact sample likelihood. However, it is important to understand its performance when the sample likelihood can be fully specified so we can compare with the exact BIC, the gold standard. In this section, we achieve this goal using a Monte Carlo simulation. In order to understand the role of the sampling design, we include a naïve BIC (S_N), a BIC that ignores the sampling design, in our simulation study.

6.1 Effect of Clustering

We consider two artificial finite populations, each consisting of $M = 200$ clusters of equal size $N_c = 10$. Thus, the size of each finite population is $N = 2000$. We generate the finite populations using the following model:

$$\begin{aligned} y_{ij} \mid \pi_j &\overset{ind}{\sim} Ber\left(\pi_j\right); \\ \pi_j &\overset{ind}{\sim} Beta\left(\frac{\mu}{\gamma}, \frac{1-\mu}{\gamma}\right), \end{aligned} \tag{10}$$

$i = 1, \cdots, N_c$; $j = 1, \cdots, M$. Note that the above model implies that the common marginal proportion and the common intra-cluster correlation are μ and $\rho = \gamma\left(\gamma + 1\right)^{-1}$, respectively. The parameter μ affects the skewness of the finite population. The parameter γ indicates the level of the intra-cluster correlation. Both the populations are fairly skewed ($\mu = 0.25$). The second population is more clustered ($\gamma = 0.\overline{3}$ or ρ=0.25) than the first population ($\gamma = 1$ or ρ=0.5.)

As far as the sampling design is concerned, we assume a simple random sampling (with replacement) of clusters and consider two different sample sizes: $n = 30$ and $n = 60$ (i.e., a sample of 3 and 6 clusters).

In summary, we consider four different settings characterized by the values summarized in Table 1.

Table 1: Different settings

Setting	ρ	*(n,m)*
1	0.5	$(30, 3)$
2	0.5	(60,6)
3	0.25	$(30, 3)$
4	0.25	$(60, 6)$

Let us consider the following hypothesis testing problem:

$$M_0 : \mu = 0.25 \qquad M_1 : \mu \neq 0.25.$$

If we completely ignore the clustering of the observations, we can specify a binomial likelihood and compute our maximum likelihood estimate of μ as $\hat{\mu} = n^{-1} y$, where *y* is the number of ones observed in the sample. The S statistic based on this incorrect likelihood is referred to as S_N. On the contrary, if we consider the clustered population model given by (10), we can specify the exact Beta-Binomial likelihood for the parameter vector (μ, γ). In this case, the maximum likelihood estimate $(\hat{\mu}, \hat{\gamma})$ cannot be obtained in a closed form, but can be computed using a numerical method (see Griffiths, 1973 for details). The *S* statistic based on this exact likelihood at the sample level is referred to as S_E. The performances of S_N and S_{DB} are compared with S_E.

In order to summarize the evidence provided by the various statistics in favour or against the null hypothesis, we consider the logarithm of the scale of evidence proposed by Jeffreys (1961) and the same cut-off point of 1.1. Values lower than 1.1 are supposed to provide "positive" evidence in favour of the model suggested by the null hypothesis.

The entries in Table 2 represent the percentage of samples with statistics lower than 1.1 over 1000 simulated samples, each drawn independently according to the

sampling design described above. Clearly, the effect of clustering on S_N is very severe for all the three cases, the acceptance rates being considerably lower than those using our gold standard S_E. The difference between S_N and S_E increases with the increase of the intra-cluster correlation. The increase in the sample size contributes very little in resolving the difference Our approximation S_{DB} tracks S_E very well even for this non-normal situation and for a moderate sample size. Needless to say, both S_{DB} and S_E are not affected by the variation of the intra-cluster correlation.

Table 2: Percentage of S statistics lower than 1.1 under M_0

	S_N	S_E	S_{DB}
Setting 1	57	97	87
Setting 2	51	96	93
Setting 4	75	99	86
Setting 5	77	99	93

We also compare the behaviour of the three procedures under a few selected alternatives: $\mu_{ALT1} = 0.5$, $\mu_{ALT2} = 0.6$, $\mu_{ALT3} = 0.75$ $\mu_{ALT4} = 0.9$.

The entries of Table 2 and Table 3 have similar interpretations.

Table 3: Percentage of the various S statistics lower than 1.1 under different null hypotheses

		S_{Naive}	S_E	S_{DB}
Null Hypothesis 1: $\mu_{ALT1} = 0.5$	Setting 1	64	75	71
	Setting 2	54	63	67
	Setting 3	37	64	69
	Setting 4	23	49	58
Null Hypothesis 2: $\mu_{ALT2} = 0.6$	Setting 1	51	59	65
	Setting 2	35	49	49
	Setting 3	23	51	51
	Setting 4	7	39	41
Null Hypothesis 3	Setting 1	18	35	45
	Setting 2	8	31	27

$\mu_{ALT3} = 0.75$	Setting 3	4	27	4
	Setting 4	0	9	12
Null Hypothesis 4: $\mu_{ALT4} = 0.9$	Setting 1	4	7	26
	Setting 2	0	1	5
	Setting 3	0	6	17
	Setting 4	0	0	2

Under all null hypotheses considered, S_{DB} and S_E perform quite closely. They both seem to be rather conservative in rejecting the null hypothesis compared to S_N. We stress that since S_N underestimates the variability in the data it overestimates the evidence against the model suggested by the null hypothesis. Settings 1,2,3 correspond to high intracluster correlation coefficients that reduce the effective sample sizes substantially. For this reason both S_{DB} and S_E have problems in finding positive evidence against the wrong model when it is very close to the true one (e.g., null hypothesis 1). This effect is somewhat weaker in settings 4 and 5 that correspond to relatively lower intra-cluster correlation coefficients.

6.2 Effect of weighting and model misspecification

We shall now consider a situation when we have a *non-epsum* sampling and study the effect of weighting. In the same example, we shall also study the effect of model misspecification. To this end, we a finite population generated using the following superpopulation model:

$$\begin{aligned} y_i &= \beta x_i + \varepsilon_i \\ \varepsilon_i &\sim \left[0, \sigma^2 x_i^2\right] \quad 1 \le i \le N \end{aligned} \tag{11}$$

with $\beta = 1,\ \sigma^2 = 1,\ N = 100,000$. Population size is set to be large in order to make the effect of replacement negligible with all sample sizes considered in this simulation.

We are interested in testing

$$H_0 : \beta = \beta_0 \quad vs \quad H_1 : \beta \neq \beta_0 \tag{12}$$

based on a sample of size *n* drawn using a probability proportional to size (PPS) sampling with replacement, in which the size variable is given by $x_i^2 \quad 1 \le i \le N$. We consider different sample sizes ranging from $n = 30$ to $n = 500$.

In this simulation study, we consider the following model selection criteria:

(a) The BIC based on the correct specification of the model:

$$S_{MB} = \frac{n}{\hat{\sigma}^2}\left(\hat{\beta}_{MB} - \beta_0\right) - \frac{1}{2}\log n,$$

where $\hat{\beta}_{MB} = n^{-1}\sum_{i \in s}\frac{y_i}{x_i}$, and $\hat{\sigma}^2 = n^{-1}\sum_{i \in s}\frac{\left(y_i - \hat{\beta}_{MB}x_i\right)^2}{x_i^2}$.

(b) The BIC based on the iid model:

$$S_{iid} = \frac{n}{\tilde{\sigma}^2}\left(\hat{\beta}_{iid} - \beta_0\right) - \frac{1}{2}\log n,$$

where $\hat{\beta}_{iid} = \frac{\sum_{i \in s} x_i y_i}{\sum_{i \in s} x_i^2}$ and $\tilde{\sigma}^2 = n^{-1}\sum_{i \in s}\left(y_i - \hat{\beta}_{iid}x_i\right)^2$.

(c) The proposed design-based BIC:

$$S_{DB} = \frac{\left(\hat{B}_{DB} - \beta_0\right)}{v\left(\hat{B}_{DB}\right)} - \frac{1}{2}\log n,$$

where $\hat{B}_{DB} = \left(X_s^T W_s X_s\right)^{-1} X_s^T W_s y_s = \left(\sum_{i \in s} y_i\right)\left(\sum_{i \in s} x_i\right)^{-1}$.

where $v\left(\hat{B}_{DB}\right)$ is a design-consistent estimator of randomization variance $V_{DB}\left(\hat{B}_{DB}\right)$ obtained using either the linearization method of Fuller (1975) or the jackknife. It can be shown that that under certain regularity conditions: $E_{MB}\left[v\left(\hat{B}_{DB}\right)\right] = V_{MB}\left(\hat{B}_{DB}\right) + o\left(n^{-1}\right)$ (E_{MB}, V_{MB} being moments with respect to (11)). Kott (1991) showed that the result holds for a fairly general class of

sampling designs that include the stratified multi-stage sampling (with unequal probability of selection at all stages).

We are interested in assessing the model-based properties of the proposed design-based model selection criterion. Thus, for each Monte Carlo replicate we generate a finite population from model (11), draw a sample from it using the PPS design and then compute the three model selection criteria.

We consider two different null hypotheses: $\beta_0 = 1$ (i.e. the "null hypothesis" is exactly true) and $\beta_0 = 1.3$ (i.e. the "null hypothesis" is false). In both cases the null model is compared against an unrestricted alternative. As in previous simulations, the cutoff point of 1.1 is used to discriminate between the two models, i.e. the models corresponding to the "non rejection" and "rejection" of the null hypothesis. For a given model selection criterion, if the statistic is greater than 1.1., we reject the null hypothesis. Thus, for $\beta_0 = 1$, the smaller the rejection rate the better. The opposite is true when $\beta_0 = 1.3$. In evaluating different model selection criteria, one must compare the rejection rates for both $\beta_0 = 1$ and $\beta_0 = 1.3$ *simultaneously*. This is because the rejection rate for $\beta_0 = 1$ can be small and even zero for some nonsensical model selection criterion (e.g., for a model selection statistics which is a constant less than 1.1 will always produce a rejection rate of zero).

Table 4: Simulation results for $\beta_0 = 1$. Rejection rate over R=1,000 MC replicates

n	S_{MB}	S_{DB}	S_{iid}
30	0.120	0.114	0.033
100	0.070	0.068	0.022
200	0.056	0.051	0.010
300	0.046	0.040	0.008
500	0.026	0.021	0.004

Table 4 suggests that S_{iid} has the lowest rejection rate. But, we explained earlier, this table alone is not conclusive – we need to analyze Table 4 and Table 5 together. Note

that S_{iid} has a small rejection rate simply because the variance estimator is severely positively biased, leading to a very conservative model selection criterion. A similar argument can be used to explain why the rejection rate for S_{DB} is smaller than that for S_{MB}. The model-based estimator is more efficient than the design-based since its variance is approximately 30% lower on the average (see Table 6) . For this reason S_{DB} appears to be more conservative than S_{MB}.

Table 5. Simulation results for $\beta_0 = 1.3$. Rejection rate over R=1,000 MC replicates

n	S_{MB}	S_{DB}	S_{iid}
30	0.53	0.42	0.14
100	0.87	0.74	0.26
200	0.99	0.93	0.42
300	1	0.99	0.59
500	1	1	0.76

Table 5 reports the rejection rates for the three model selection criteria when $\beta_0 = 1.3$. In this case, S_{MB} and S_{DB} have both rejection rates converging to 1 with the first being faster than the second. This is consistent with the fact that the model-based estimation based on a correctly specified model is optimal, and thus more efficient than the design-based alternative. The effect of model misspecification on the model-based BIC, i.e. S_{iid}, is now transparent – this model selection criterion has an extremely low convergence. In fact, neglecting heteroscedasticity (and not correcting by weights) leads to overestimation of σ^2.

It is interesting to note that in both cases $v\left(\hat{B}_{DB}\right)$, the variance estimate based on the linearization method of Fuller (1975), is approximately unbiased for the model variance of $\hat{B}_{DB}$. This is consistent with the theory. The performances of variance estimators are described in Table 6:

Table 6: MC comparison of $\hat{\beta}_{MB}$, $\hat{B}_{DB}$, $\hat{V}_{MB}\left(\hat{\beta}_{MB}\right)$, $v\left(\hat{B}_{DB}\right)$, *1,000 replicates*

n	$\frac{V_{MC}\left(\hat{\beta}_{MB}\right)}{V_{MC}\left(\hat{B}_{DB}\right)}$	$\frac{E_{MC}\left[\hat{V}_{MB}\left(\hat{\beta}_{MB}\right)\right]}{E_{MC}\left[v\left(\hat{B}_{DB}\right)\right]}$	$\frac{E_{MC}\left[\hat{V}_{MB}\left(\hat{\beta}_{MB}\right)\right]}{V_{MC}\left(\hat{\beta}_{MB}\right)}-1$	$\frac{E_{MC}\left[v\left(\hat{B}_{DB}\right)\right]}{V_{MC}\left(\hat{B}_{DB}\right)}-1$
30	0.74	0.71	-0. 072	-0.018
100	0.66	0.67	0.005	0.049
200	0.67	0.67	-0.005	-0.011
300	0.72	0.69	-0.025	0.012
500	0.71	0.70	-0.006	-0.001

Note that in Table 6 $E_{MC}(\,.\,)$, $V_{MC}(\,.\,)$ are the moments obtained from the empirical distribution of Monte Carlo replicates and $\hat{V}_{MB}\left(\hat{\beta}_{MB}\right)=\hat{\sigma}^2\left(X^T\Omega^{-1}X\right)$, $\Omega=diag_{1\le i\le n}\left(x_i^2\right)$ is the usual model-based variance of $\hat{\beta}_{MB}$ under the correctly specified model. Second column describes the relative variance of model-based and design-based estimators: the model-based one is more efficient, as expected. Third column is the ratio between their variances as estimated by the method described above. The second and the third columns are providing consistent results. Fourth and fifth columns provide the relative biases of the model-based and design-based variance estimators – it is clear that $v\left(\hat{B}_{DB}\right)$ is performing very well in terms of model-unbiasedeness criterion.

7. Concluding remarks

We have presented a robust approximation to the BIC that can be used with complex survey data. Our method is expected to be useful in situations where it is *not* possible to obtain the *exact* likelihood for the sample since our proposed method merely requires an estimator of the superpopulation parameter with good design-based properties (e.g., pseudo-maximum likelihood) and its design consistent variance estimator. Thus, this paper fills in an important research gap in the analytic use of survey data.

References

Berger J.O. & Pericchi L.R. (2001) Objective bayesian methods for model selection: introduction and comparison. In *Model selection*, P. Lahiri ed., Institute of Mathematical Statistics, Lecture Notes – Monograph series, Vol. 38, 135-193

Chambers R. & Skinner C.J. (2003) *Analysis of survey data*. John Wiley and Sons, New York.

Cochran W.G. (1977) *Sampling techniques*. John Wiley and Sons, New York.

Deming, W.E. & Stephen, F.F. (1941) On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36, 45-49.

Deming W.E. (1953) On the distinction between enumerative and analytic surveys. *Journal of the American Statistical Association*, 48, 244-255.

Efron B. & Gous A. (2001) Scales of evidence for model Selection: Fisher versus Jeffreys. In *Model selection*, P. Lahiri ed., Institute of Mathematical Statistics, Lecture Notes – Monograph series, Vol. 38, 208-246.

Hansen M. H., Madow W.G. & Tepping B.J. (1983) An evaluation of model-dependent and probability-sampling inference in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Griffiths D.A. (1973) Maximum likelihood estimation for the Beta-Binomial eistribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, 29, 637-648.

Graubard B.I. & Korn, E.L. (2002) Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17, 73-96.

Holt D., Smith T.M.F. & Winter P.D. (1980) Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society, Ser. A*, 143, 474-487.

Holt D. (1989) Introduction to Part C. In *Analysis of complex surveys*. C.J. Skinner, D. Holt, T.M.F. Smith eds., 209-220, John Wiley, Chichester.

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

Jeffreys H. (1961) *Theory of probability* 3rd edition. Oxford University Press, Oxford

Lahiri P. (2001) *Model Selection*. Institute of Mathematical Statistics, Lecture Notes – Monograph series, Vol. 38

Lynn P. & Gabler S. (2005) Approximations to b^* in the prediction of design effects due to clustering, *Survey Methodology*, 31, 101-105.

Kass R.E. & Wassermann L. (1995) A reference test for nested hypotheses and its relationship to the Schwartz criterion. *Journal of the Americal Statistical Association*, 90, 928-934

Kott P.S. (1989) Robust small domain estimation using random effects modelling. *Survey Methodology* 15, 3-12.

Kott P.S. (1991) A model-based look at linear regression with survey data. *The American Statistician*, 45, no 2, 107-112.

Sarndal C. E., Swensson B. & Wretman J. (1992) *Model assisted survey sampling*. Springer Verlag, New York

Searle S.R., Casella G. & McCullogh C.E. (1996) *Variance components*. John Wiley and Sons, New York.

Schwartz G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464

Skinner C.J. (1989) Introduction to Part A. In *Analysis of complex surveys*, C.J. Skinner, D. Holt, T.M.F. Smith eds., 23-58, John Wiley & Sons, Chichester.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Ser. B*, 64, 583-639.

Redazione

Dipartimento di Matematica, Statistica, Informatica ed Applicazioni
Università degli Studi di Bergamo
Via dei Caniana, 2
24127 Bergamo
Tel. 0039-035-2052536
Fax 0039-035-2052549

La Redazione ottempera agli obblighi previsti dall'art. 1 del D.L.L. 31.8.1945, n. 660 e successive modifiche

Stampato nel 2007
presso la Cooperativa
Studium Bergomense a r.l.
di Bergamo