Multi-level Modelling Under Informative Sampling

Danny Pfeffermann, Fernando Moura, Pedro Nascimento Silva

Abstract

We consider a model dependent approach for multi-level modelling that accounts for informative probability sampling, and compare it with the use of probability weighting as proposed by Pfeffermann et al. (1998a). The new modelling approach consists of first extracting the hierarchical model holding for the sample data as a function of the corresponding population model and the first and higher level sample selection probabilities, and then fitting the resulting sample model using Bayesian methods. An important implication of the use of this approach is that the sample selection probabilities feature in the analysis as additional outcome values that strengthen the estimators. A simulation experiment is carried out in order to study and compare the performance of the two approaches. The simulation study indicates that both approaches perform generally equally well in terms of point estimation, but the model dependent approach yields confidence (credibility) intervals with better coverage properties. A robustness simulation study is performed, which allows to assess the impact of misspecification of the models assumed for the sample selection probabilities under informative sampling schemes.

S^{3}RI Methodology Working Paper M04/09

MULTI-LEVEL MODELLING UNDER INFORMATIVE SAMPLING

By DANNY PFEFFERMANN

Department of Statistics, Hebrew University, Jerusalem, 91905 Israel,
Southampton Statistical Sciences Research Institute, Southampton SO17 1BJ, UK

msdanny@mscc.huji.ac.il

FERNANDO MOURA

Department of Statistical methods, Federal University of Rio de Janeiro, 21945-970, Brazil (*)

fmoura@im.ufrj.br

PEDRO NASCIMENTO SILVA

Escola Nacional de Ciencias Estatisticas, Rio de Janeiro, 20231-050, Brazil

pedrosilva@ibge.gov.br

SUMMARY

We consider a model dependent approach for multi-level modelling that accounts for informative probability sampling, and compare it with the use of probability weighting as proposed by Pfeffermann *et al*. (1998a). The new modelling approach consists of first extracting the hierarchical model holding for the sample data as a function of the corresponding population model and the first and higher level sample selection probabilities, and then fitting the resulting sample model using Bayesian methods. An important implication of the use of this approach is that the sample selection probabilities feature in the analysis as additional outcome values that strengthen the estimators. A simulation experiment is carried out in order to study and compare the performance of the two approaches. The simulation study indicates that both approaches perform generally equally well in terms of point estimation, but the model dependent approach yields confidence (credibility) intervals with better coverage properties. A robustness simulation study is performed, which allows to assess the impact of misspecification of the models assumed for the sample selection probabilities under informative sampling schemes.

Some key words: Confidence (Credibility) intervals, MCMC, Probability weighting, Small area estimation, Sample distribution

(*)- Most of the research underlying this article was carried out while the second author visited Southampton Statistical Sciences Research Institute. The work of the second author was partially funded by a research grant from CNPq in Brazil.

1. INTRODUCTION

Multi-level (mixed linear) models are frequently used in the social and medical sciences for modelling hierarchically clustered populations. Classical theory underlying the use of these models assumes implicitly that either all the clusters at all the levels are represented in the sample, or that they are sampled completely at random. This assumption may not hold in a typical sample survey where the clusters and/or the final sampling units are often sampled with unequal selection probabilities. When the sampling probabilities are related to the values of the outcome variable even when conditioning on the model covariates, the sampling process becomes *informative* and the model holding for the sample data is then different from the population model. Ignoring the sampling process in such cases may yield biased point estimators and distort the analysis.

As an example, consider an education study of pupils' proficiency with schools as the second level units and pupils as first level units, and suppose that the schools are sampled with probabilities proportional to their sizes. Under this (commonly used) sampling scheme the sample of schools will tend to contain mostly large schools, and if the size of the school is related to the pupils' proficiency but the size is not included among the model covariates, the schools in the sample will not represent correctly the schools in the population and therefore follow a different model. A situation where the size of the school is related to the pupils' proficiency is when the larger schools are mostly located in poor areas with low proficiency.

As implicitly suggested by this example, a possible way of handling the problem of informative sampling is by including among the model covariates all the design variables that define the selection probabilities at the various levels. However, this paradigm is often not practical. First, not all the design variables used for the sample selection may be known or accessible to the analyst, or that there may be too many of them, making the fitting and validation of such models formidable. Second, by including the design variables among the model covariates, the resulting model may no longer be of scientific interest. This is not necessarily a problem when the fitting of the model is for prediction purposes, but is clearly not acceptable when the purpose of the analysis is to study the structural relationship between the outcome variable and covariates of interest.

In order to deal with the effects of informative sampling, Pfeffermann *et al.* (1998a) proposed probability-weighting of first and second level units that control the bias of the parameter estimators under the randomization (repeated sampling) distribution. The authors developed also appropriate variance estimators. The use of this approach is justified based on

asymptotic arguments but it was shown to perform well in a simulation study also with moderate sample sizes. Nonetheless, the use of the sampling weights (inverse of the sample inclusion probabilities) for bias correction has four important limitations:

1- The variances of the weighted estimators are generally larger than the variances of the corresponding unweighted estimators.
2- Inference is restricted primarily to point estimation. Probabilistic statements require asymptotic normality assumptions. The exact distribution of weighted point estimators is generally unknown.
3- The use of the sampling weights does not permit in general to condition on the selected sample of clusters (second and higher level units), or values of the model covariates.
4- It is not clear how to predict with this approach second and higher level random effects under informative sampling; for example, how to predict the mean school proficiency for schools not represented in the sample. Notice that under informative sampling, the schools not represented in the sample also form an 'informative sample' that behaves differently from the schools in the population. We mention in this respect that multi-level models are in common use for Small Area Estimation problems, where the prediction of the higher level (area) means is the primary objective of the model fitting.

In this article we consider a model dependent approach for multi-level modelling under informative sampling and compare it to the use of probability weighting. The idea behind the use of the modelling approach is to first extract the hierarchical model holding for the sample data as a function of the corresponding population model and the conditional expectations of the first and higher level sample selection probabilities given the observed data and the model random effects, and then fit the sample model using Bayesian methods. An important implication of the use of this approach is that the sample selection probabilities feature into the analysis as additional outcome values that strengthen the estimators. Evidently, if the sample model is specified correctly, the use of this approach overcomes the limitations underlying the use of probability weighting mentioned above. However, as illustrated and discussed later, misspecification of the models assumed for the sample selection probabilities may bias some of the model parameter estimators. (A similar problem is shown to underlie the use of probability weighting when the sample selection probabilities are unknown, like as in nonresponse.)

We consider for convenience a two-level model and apply the full Bayesian paradigm by use of Markov Chain Monte Carlo (MCMC) simulations, but the approach can be extended to higher level models and different inference procedures. The empirical study is restricted to simulated data from known models, which enables us to study the bias of the various estimators and the performance of the corresponding confidence (credibility) intervals.

In Section 2 we define the population model and extract the corresponding sample model for a general class of sampling designs underlying the present study. Section 3 outlines the probability weighting approach proposed in Pfeffermann *et al.* (1998a). Section 4 describes the simulation experiment designed for studying and comparing the performance of the two approaches, and develops the corresponding sample model. Section 5 describes the various steps in the application of the MCMC algorithm for fitting the sample model. The results of the simulation study are presented and discussed in Section 6. Section 7 presents the results of a robustness simulation study carried out for assessing the performance of the two approaches when the sampling schemes are informative but the models assumed for the sample selection probabilities are misspecified. We conclude in section 8 by summarizing the main conclusions from the present study.

2. POPULATION MODEL, SAMPLING DESIGN AND SAMPLE MODEL

2.1 *Population Model*

In this article we consider the following two-level hierarchical model:

$$\underline{\textit{First level}}: \qquad y_{ij} \mid \beta_{0i} = \beta_{0i} + x_{ij}'\beta + \varepsilon_{ij} \; ; \; \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \; j = 1...M_i \tag{1}$$

$$\underline{\textit{Second level}}: \qquad \beta_{0i} = z_i'\gamma + u_i \; ; \; u_i \sim N(0, \sigma_u^2), \; i = 1...N \tag{2}$$

This model is often referred to in the literature as the random intercept regression model, and it contains as unknown hyper-parameters the vectors of coefficients β, γ, and the first and second level variances σ_ε^2 and σ_u^2. Note that the intercepts are modelled as linear functions of known regressor values, z_i. In the simulation experiment described in Section 4 we refer to the outcome y_{ij} as the *test score* of pupil j in school i, x_{ij} defines the *sex*, *age* and *parents' education* of that pupil and z_i consists of two dummy variables defining

geographical regions. The second level random effect u_i accounts for the variation of the intercept β_{0i}, not explained by the regressors z_i.

2.2 *Sampling Design*

We assume a two stage sampling process. In the first stage, $n < N$ second level units (say, *schools*) are selected with probabilities $\pi_i = \Pr(i \in s)$ that may be correlated with the random effects u_i. In the second stage, m_i first level units (say, *pupils*) are sampled from second level unit i selected in the first stage with probabilities $\pi_{j|i} = \Pr(j \in s_i \mid i \in s)$ that may be correlated with the residuals ε_{ij}. Notice that the sampling of the first level units may correspond to a response process, in which case the probabilities $\pi_{j|i}$ define the (unknown) response probabilities. As seen below, the case of unknown response probabilities is handled under the model dependent approach by modelling the conditional expectation of these probabilities given the observed data (see also Section 7). In Section 4 we elaborate on the sampling procedures used for the simulation study.

2.3 *The Sample model*

In what follows we denote by $\theta_i = (\beta_{0i}, \beta', \sigma_\varepsilon^2)$ and $\lambda = (\gamma', \sigma_u^2)$ the respective first and second level parameters of the population model. Following Pfeffermann *et al.* (1998b), the corresponding two-level sample model is,

$$f_{s_i}(y_{ij} \mid x_{ij}, \theta_i) = f(y_{ij} \mid x_{ij}, \theta_i, j \in s_i) = \frac{E_p(\pi_{j|i} \mid y_{ij}, x_{ij}, \theta_i) f_p(y_{ij} \mid x_{ij}, \theta_i)}{E_p(\pi_{j|i} \mid x_{ij}, \theta_i)} \tag{3}$$

$$f_s(\beta_{0i} \mid z_i, \lambda) = f(\beta_{0i} \mid z_i, \lambda, i \in s) = \frac{E_p(\pi_i \mid \beta_{0i}, z_i, \lambda) f_p(\beta_{0i} \mid z_i, \lambda)}{E_p(\pi_i \mid z_i, \lambda)} \tag{4}$$

where s_i defines the first level sample from second level unit i, s defines the second level sample and $f_p(\cdot)$ and $f_s(\cdot)$ are the population and sample distributions with expectations $E_p(\cdot)$ and $E_s(\cdot)$ respectively.

The sample model defined by (3) and (4) depends on the population model and the (conditional) expectations of the first order sample selection probabilities of first and second level units. The expectations featuring in the two equations can be modelled based on knowledge of the sampling process and the sample data; see Pfeffermann and Sverchkov

(1999, 2003) for discussion and examples. The corresponding expressions under the sampling schemes considered for the simulation study of this article are presented in Section 4.

3. PROBABILITY WEIGHTED (DESIGN BASED) APPROACH

In this section we describe briefly the weighting procedure developed by Pfeffermann *et al.* (1998a). We restrict for convenience to the population model defined by (1) and (2) that contains a single second level random effect β_{0i}. Suppose first that all the population units are surveyed and denote by $y_{pi} = (y_{i1}...y_{iM_i})$' and $T_{pi} = (t_{i1}...t_{iM_i})$' the data measured for second level unit i of size M_i, where $t_{ij} = (x_{ij}', z_i')$'. Denoting also $e_{pi} = (e_{i1}...e_{iM_i})$', where $e_{ij} = (\varepsilon_{ij} + u_i)$, $V_i = J\sigma_u^2 + I\sigma_\varepsilon^2$, where J and I define the unit matrix and the identity matrix of order M_i respectively and $\delta' = (\beta', \gamma')$, the 'census model' defined by (1) and (2) can be written alternatively as,

$$y_{pi} = T_{pi}\delta + e_{pi}\,;\;\; e_{pi} \sim N(0, V_i)\,,\; i = 1...N\,,\; E(e_{pi}e_{pk}') = 0 \;\text{ for }\; i \neq k \tag{5}$$

A commonly used procedure for estimating the vector coefficients δ and the variances $\varphi = (\sigma_u^2, \sigma_\varepsilon^2)$' is the iterative generalized least squares (IGLS) algorithm, developed by Goldstein (1986). The algorithm consists of iterating between the estimation of δ for 'given' φ, and the estimation of φ for 'given' δ, with the 'given' values defined by the estimates obtained on the previous iteration. The two sets of estimators are the corresponding generalized least square estimators (considering the 'given' values as 'true'), where the observed values of the dependent variable for the estimation of δ are the vectors $\{y_{pi},\, i = 1...N\}$, and the 'observed' values of the dependent variable for the estimation of φ on the rth iteration are the elements of the matrices $\{D_i(\hat{\delta}^{(r)}) = (y_i - T_i\hat{\delta}^{(r)})(y_i - T_i\hat{\delta}^{(r)})',\, i = 1...N\}$, written as a vector. Notice that for known δ, $D_i(\delta)$ has expectation V_i and that for normal error terms (u_i, ε_{ij}), the variances and covariances of the elements of $D_i(\delta)$ are known functions of φ; see Anderson (1973) for the corresponding expressions. The rth iteration of the IGLS yields therefore the estimators,

$$\hat{\delta}^{(r)} = [Q^{(r)}]^{-1} y^{(r)} \quad ; \quad \hat{\varphi}^{(r)} = [R^{(r)}]^{-1} d^{(r)}\,,\; r = 1,2,... \tag{6}$$

with appropriate definitions of the matrices $Q^{(r)}, R^{(r)}$ and the vectors $y^{(r)}, d^{(r)}$; see Pfeffermann *et al.* (1998a) for details. The iterations can be started by estimating δ by

ordinary least squares (OLS), so that the estimators $\hat{\varphi}^{(r)}$ use the estimators $\hat{\delta}^{(r)}$, whereas the estimators $\hat{\delta}^{(r)}$ use the estimators $\hat{\varphi}^{(r-1)}$, $r = 1,2,...$ Under some regularity conditions the IGLS estimators converge to the 'census' maximum likelihood estimators (MLE) $(\hat{\delta}_C, \hat{\varphi}_C)$ as $r \to \infty$. (The census MLE are the MLE based on all the population data).

Suppose now that data are available for only a probability sample of second and first level units as defined in Section 2.2. The weighting procedure developed by Pfeffermann *et al.* (1998a) consists of writing the matrices $Q^{(r)}, R^{(r)}$ and the vectors $y^{(r)}, d^{(r)}$ in iteration *r* as sums over second level units i and first level units j, and replacing each population sum by the corresponding weighted sum of the sample units, with the weights defined by the inverse of the respective selection probabilities. Denoting the second level weights by $w_i = 1/\pi_i$, and the first level weights by $w_{j|i} = 1/\pi_{j|i}$, and using the generic notations $\aleph_i$ and $\aleph_{ij}$ to define second and first level expressions respectively, the procedure consists of replacing,

$$\sum_{i=1}^{N} \aleph_i \leftrightarrow \sum_{i=1}^{n} w_i \aleph_i \quad ; \quad \sum_{j=1}^{M_i} \aleph_{ij} \leftrightarrow \sum_{j=1}^{m_i} w_{j|i} \aleph_{ij} \tag{7}$$

Notice that each second level expression $\aleph_i$ is again a sum of the form $\sum_{j=1}^{M_i} \aleph_{ij}$. It is shown in Pfeffermann *et al.* (1998a) that under standard conditions the estimators $\hat{\delta}_{PW}$ and $\hat{\varphi}_{PW}$ obtained at the end of the iterations are consistent for the census estimators $(\hat{\delta}_C, \hat{\varphi}_C)$ under the randomization (repeated sampling) distribution, with the latter estimators being consistent for (δ, ϕ) under the model. The consistency of $\hat{\varphi}_{PW}$ requires that both *n* and m_i tend to infinity, but appropriate scaling of the weights $w_{j|i}$ controls the bias of these estimators for small m_i. A simple scaling method used in the simulation experiment of the present article is to replace $w_{j|i}$ by $w_{j|i} / (\sum_{j=1}^{m_i} w_{j|i} / m_i)$.

4. MONTE-CARLO SIMULATION EXPERIMENT

The purpose of the simulation experiment is to study the performance of the model dependent approach introduced in Section 2.2 (see details below), and compare it with the weighting procedure described in Section 3. The sampling design and the explanatory variables values underlying this experiment were taken from the 'Basic Education Evaluation study' carried out in 1996 for the municipality of Rio de Janeiro in Brazil (hereafter the BEE

study). The target outcome values in that study were the proficiency scores of *M=14,831* pupils, learning in *N=392* schools, located in 3 different regions. In what follows we use '*schools*' to define the second level units and '*pupils*' to define the first level units. The simulation experiment consists of generating 400 populations from the model defined by (1) and (2) and selecting four samples from each population using four different sampling schemes. The various stages of the simulation experiment are described in Sections 4.1 to 4.4.

4.1 *Generation of Population Values*

The population values were generated in 5 steps:

Step 1 **-** Generate school random intercept terms from the model (Equation 2),

$\beta_{0i} = \gamma_0 + \gamma_1 \text{Region1}_i + \gamma_2 \text{Region2}_i + u_i = z_i'\gamma + u_i; \quad u_i \sim N(0, \sigma_u^2)$, $i = 1...392$,

independently between schools. The numerical values of the γ - coefficients and σ_u^2 are listed in the tables of Section 6. The variables $Region1_i$ *and* $Region2_i$ are dummy variables defining three school regions. The number of schools in the three regions is approximately the same.

Step 2- Generate school sizes M_i from the lognormal distribution, $\log(M_i) \sim N[\alpha' z_i + \alpha_3 \beta_{0i}, \sigma_M^2]$, with z_i defined as above and $\alpha_0 = 9.25$, $\alpha_1 = 0.31$, $\alpha_2 = 0.62$, $\alpha_3 = -0.045$ and $\sigma_M^2 = 0.050$. The use of these parameter values yields school sizes with a similar distribution to the sizes of the schools in the BEE study.

Step 3 - Set explanatory variables values x_{ij} for the M_i students in school i by sampling at random with replacement M_i vectors of explanatory variables from the corresponding BEE data in the region containing that school. The explanatory variables are dummy variables defining *Sex* (1 for females), *Age*1 (1 for age 15-16), *Age*2 (1 for age 17 and older) and *Parents education* (1 for pupils with at least one parent having an academic degree).

Step 4 - Generate proficiency score for student j of school i using the model (Equation 1),

$y_{ij} = \beta_{0i} + \beta_1 Sex_{ij} + \beta_2 Age1_{ij} + \beta_3 Age2_{ij} + \beta_4 Parents_{ij} + \varepsilon_{ij} = \beta_{0i} + x_{ij}'\beta + \varepsilon_{ij}$; $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$

The numerical values of the β - coefficients and σ_ε^2 are listed in the tables of Section 6.

In order to allow for informative sampling (response) of pupils, we stratified the pupils within each school into 3 strata based on propensity scores generated as follows:

Step 5 – Generate propensity scores $p_{ij} = b_0 + b_1 y_{ij} + \zeta_{ij}$; $\zeta_{ij} \sim N(0, \sigma_\zeta^2)$ where $b_0 = 1.60$, $b_1 = 0.0056$ and $\sigma_\zeta^2 = 0.034$. Strata membership has been assigned by setting $O_{ij} = 1$ ($stratum 1$) if $p_{ij} < c_1$, $O_{ij} = 2$ ($stratum 2$) if $c_1 \le p_{ij} < c_2$, $O_{ij} = 3$ ($stratum 3$) if $p_{ij} \ge c_2$, with $c_1 = 1.76$ and $c_2 = 2.16$.

4.2 *Sampling Schemes*

We consider two different methods for the sampling of schools and two different methods for the sampling (response) of pupils within the selected schools, defining a total of 4 different two-stage sampling schemes. Schools were selected using either *Method* A1- simple random sampling without replacement (SRSWOR) or *Method* A2- probability proportional to size (PPS), using Sampford (1967) method. Note that Method A2 is informative since the sizes M_i depend on the intercepts β_{0i} (Step 2). Students within the selected schools were sampled either by *Method* B1- SRSWOR or *Method* B2- disproportionate stratified sampling with the strata defined by the strata membership indicators O_{ij} (three strata, Step 5). Method B2 is informative since the strata indicators are defined based on the propensity scores p_{ij}, which depend on the proficiency scores y_{ij} (Step 5). One sample of 50 schools and 12 pupils from each selected school was drawn from each population using each of the 4 sampling schemes. For the stratified sample selection (Method B2) we sampled 3 pupils from Stratum 1, 4 pupils from Stratum 2 and 5 pupils from Stratum 3, yielding mean sample selection probabilities (over schools) of about 0.08 in stratum 1, 0.04 in stratum 2 and 0.14 in stratum 3.

4.3. *Sample models under sampling methods A2 and B2*

The sample models under general two-stage sampling schemes are defined by (3) and (4). The expectation in the numerator of the first level model (Equation 3), under the sampling method B2 is,

$$E_p(\pi_{j|i} \mid y_{ij}, x_{ij}, \theta_i) = \Pr(j \in s_i \mid y_{ij}, x_{ij}, \theta_i) = \sum_{k=1}^{3} q_k^i \Pr(O_{ij} = k \mid y_{ij}, x_{ij}, \theta_i)$$

$$= q_1^i A_1(y_{ij}) + q_2^i [A_2(y_{ij}) - A_1(y_{ij})] + q_3^i [1 - A_2(y_{ij})] \quad (8)$$

where s_i denotes as before the sample of pupils in school i, $q_k^i = \Pr(j \in s_i \mid O_{ij} = k)$ is the sampling fraction in stratum k of school i, k=1,2,3 and $A_k(y_{ij}) = \Pr(p_{ij} < c_k \mid y_{ij}, x_{ij}, \theta_i)$ $= \Phi[(c_k - b_0 - b_1 y_{ij}) / \sigma_\zeta]$, where Φ defines the cumulative normal distribution and c_1 and c_2 are the cut-off values defining the strata membership (Step 5 in Section 4.1).

The expectation in the denominator of (3) is obtained by following similar steps using the conditional density $f(p_{ij} \mid x_{ij}, \theta_i)$. We find,

$$E_p(\pi_{j|i} \mid x_{ij}, \theta_i) = \sum_{k=1}^{3} q_k^i \Pr(O_{ij} = k \mid x_{ij}, \theta_i)$$

$$= q_1^i B_1[\mu(y_{ij})] + q_2^i \{B_2[\mu(y_{ij})] - B_1[\mu(y_{ij})]\} + q_3^i \{1 - B_2[\mu(y_{ij})]\}, \quad (9)$$

$$\mu(y_{ij}) = \beta_{0i} + x_{ij}'\beta = E_p(y_{ij} \mid x_{ij}, \theta_i) \; ; \; B_k[\mu(y_{ij})] = \Pr(p_{ij} < c_k \mid x_{ij}, \theta_i) = \Phi\left(\frac{c_k - b_0 - b_1 \mu(y_{ij})}{\sqrt{\sigma_\zeta^2 + b_1^2 \sigma_\varepsilon^2}} \right).$$

The expectations featuring in the numerator and the denominator of the second level sample model (Equation 4), under the sampling method A2 are,

$$E_p(\pi_i \mid \beta_{0i}, z_i, \lambda) \cong C^* \exp[\alpha' z_i + \alpha_3 \beta_{0i} + \frac{\sigma_M^2}{2}] \quad (10)$$

$$E_p(\pi_i \mid z_i, \lambda) \cong C^* \exp[\alpha' z_i + \alpha_3 z_i \gamma + \frac{\alpha_3^2 \sigma_u^2 + \sigma_M^2}{2}], \quad (11)$$

using familiar properties of the lognormal distribution used to generate the school sizes and the approximation, $\pi_i = n \frac{M_i}{N\overline{M}} \cong C^* M_i$, where $\overline{M} = \frac{1}{N} \sum_{i=1}^{N} M_i$ is the population mean of the school sizes. Notice that the constant C^* cancels out in the numerator and denominator of (4).

5. ESTIMATION OF MODEL PARAMETERS BY MARKOV CHAIN MONTE CARLO (MCMC) SIMULATION

The MCMC algorithm consists of sampling alternately from the conditional posterior distribution of each of the unknown parameters, given the data and the remaining parameters. We used for the present study the version 1.4 of the WinBUGS program, (Spiegelhalter *et al.* 2003), generating 5000 samples from each posterior distribution after discarding the first

15000 values as 'burn in'. Let $Y=\{y_{ij};\, i=1...n,\, j=1...m_i\}$, $O=\{O_{ij};\, i=1...n,\, j=1...m_i\}$ and $M=\{M_i;\, i=1...n\}$ denote respectively the observed y-values, the corresponding strata membership indicators and the school sizes in the sample. The observed data consist therefore of the triple $D_{obs}=(Y,O,M)$.

In what follows we use the transformations $\eta_k=(c_k-b_0)/\sigma_\zeta$, $k=1,2$, $\eta_3=-(b_1/\sigma_\zeta)$, such that the probabilities $A_k(y_{ij})$ and $B_k[\mu(y_{ij})]$ in (8) and (9) can be written as, $A_k(y_{ij})=\Phi(\eta_k+\eta_3 y_{ij})$ and $B_k[\mu(y_{ij})]=\Phi\left(\dfrac{\eta_k+\eta_3\mu(y_{ij})}{\sqrt{1+\eta_3^2\sigma_\varepsilon^2}}\right)$. Notice that the η-coefficients are considered as unknown parameters in the estimation process, implying that the cut-off values c_k, the variance σ_ζ^2 and the b-coefficients used to define the sampling strata in Step 5 are also considered as unknown.

The joint distribution of the observations D_{obs}, the parameters $(\beta_{0i},\beta,\sigma_u^2,\sigma_\varepsilon^2)$ indexing the population model and the additional parameters (η,α,σ_M^2) indexing the sample model can be written as,

$$
\begin{aligned}
&f_s(D_{obs},\{\beta_{0i}\},\beta,\eta,\alpha,\sigma_u^2,\sigma_\varepsilon^2,\sigma_M^2)\\
&=\prod_{i=1}^{n}\prod_{j=1}^{n} f_s(y_{ij}\mid x_{ij},\beta_{0i},\beta,\eta,\sigma_\varepsilon^2)\,\Pr_s(O_{ij}\mid y_{ij},\eta)\,f_s(\beta_{0i}\mid z_i,\gamma,\alpha,\sigma_u^2)\\
&\times f_s(M_i\mid z_i,\beta_{0i},\alpha,\sigma_M^2).p(\beta).p(\gamma).p(\eta).p(\alpha).p(\sigma_u^2).p(\sigma_\varepsilon^2).p(\sigma_M^2) \qquad (12)
\end{aligned}
$$

where $\Pr_s(O_{ij}=k\mid y_{ij},\eta)=\Pr(O_{ij}=k\mid y_{ij},\eta,\, j\in s_i)=\dfrac{q_{ki}\Pr(O_{ij}=k\mid y_{ij},\eta)}{\sum_{l=1}^{3} q_{ki}\Pr(O_{ij}=l\mid y_{ij},\eta)}$, $k=1,2,3$.

The sample distribution $f_s(y_{ij}\mid x_{ij},\beta_{0i},\beta,\eta,\sigma_\varepsilon^2)$ is defined by (3), with the expectations appearing in the numerator and the denominator defined by (8) and (9) as functions of the unknown η-coefficients defined above. The sample distribution $f_s(\beta_{0i}\mid z_i,\gamma,\alpha,\sigma_u^2)$ is defined by (4), with the expectations appearing in the numerator and the denominator defined by (10) and (11). Here again, the α-coefficients and the variance σ_M^2 are additional unknown parameters. The probabilities $\Pr(O_{ij}=k\mid y_{ij},x_{ij},\theta_i)$ are defined in (8), and after the transformation of the probabilities $A_k(y_{ij})$ defined above, they take the simple form,

$$\Pr(O_{ij}=1\mid y_{ij},\eta)=A_1(y_{ij}),\ \Pr(O_{ij}=2\mid y_{ij},\eta)=A_2(y_{ij})-A_1(y_{ij}),$$

$$\Pr(O_{ij}=3\mid y_{ij},\eta)=1-A_2(y_{ij}) \tag{13}$$

The sample distribution of the school sizes is obtained similarly to (4) as,

$$f_s(M_i\mid z_i,\beta_{0i},\alpha,\sigma_M^2)=\frac{E_p(\pi_i\mid M_i,z_i,\beta_{0i},\alpha,\sigma_M^2)f_p(M_i\mid z_i,\beta_{0i},\alpha,\sigma_M^2)}{E_p(\pi_i\mid z_i,\beta_{0i},\alpha,\sigma_M^2)}. \tag{14}$$

By Pfeffermann *et al.* (1998b), the conditional density in (14) under the sampling method A2 is lognormal;

$$f_s[\log(M_i)\mid z_i,\beta_{0i},\alpha,\sigma_M^2]=N[\alpha' z_i+\alpha_3\beta_{0i}+\sigma_M^2,\sigma_M^2)]. \tag{15}$$

Note that the only difference between the population distribution and the sample distribution is in this case the addition of the term σ_M^2 to the mean.

The conditional posterior distributions of the various parameters given the data and the remaining parameter values, required for the application of the MCMC simulation, are obtained from the joint distribution $f_s(D_{obs},\{\beta_{0i}\},\beta,\eta,\alpha,\sigma_u^2,\sigma_\varepsilon^2,\sigma_M^2)$ defined in (12). Following are the posterior distribution of $\{\beta_{0i}\}$ and each of the vector coefficients and variances, using the generic notation 'Rest' to denote the data and the remaining parameters. The notation $p(\cdot)$ is used to denote the corresponding prior distributions defined below.

$$f(\beta_{0i}\mid \text{Rest})\propto\prod_{j=1}^{m_i}f_s(y_{ij}\mid x_{ij},\beta_{0i},\beta,\eta,\sigma_\varepsilon^2)f_s(\mathsf{M}_i\mid z_i,\beta_{0i},\alpha,\sigma_M^2)f_s(\beta_{0i}\mid z_i,\gamma,\alpha,\sigma_u^2),\ i=1..n \tag{16a}$$

$$f(\beta\mid \text{Rest})\propto\prod_{i=1}^{n}\prod_{j=1}^{m_i}f_s(y_{ij}\mid x_{ij},\beta_{0i},\beta,\eta,\sigma_\varepsilon^2)p(\beta) \tag{16b}$$

$$f(\gamma\mid \text{Rest})\propto\prod_{i=1}^{n}f_s(\beta_{0i}\mid z_i,\gamma,\alpha,\sigma_u^2)p(\gamma) \tag{16c}$$

$$f(\alpha\mid \text{Rest})\propto\prod_{i=1}^{n}f_s(\mathsf{M}_i\mid z_i,\beta_{0i},\alpha,\sigma_M^2)p(\alpha) \tag{16d}$$

$$f(\eta\mid \text{Rest})\propto\prod_{i=1}^{n}\prod_{j=1}^{m_i}f_s(y_{ij}\mid x_{ij},\beta_{0i},\beta,\eta,\sigma_\varepsilon^2)\Pr_s(O_{ij}\mid y_{ij},\eta)p(\eta) \tag{16e}$$

$$f(\sigma_\varepsilon^2\mid \text{Rest})\propto\prod_{i=1}^{n}\prod_{j=1}^{m_i}f_s(y_{ij}\mid x_{ij},\beta_{0i},\beta,\eta,\sigma_\varepsilon^2)p(\sigma_\varepsilon^2) \tag{16f}$$

$$f(\sigma_u^2\mid \text{Rest})\propto\prod_{i=1}^{n}f_s(\beta_{0i}\mid z_i,\gamma,\alpha,\sigma_u^2)p(\sigma_u^2) \tag{16g}$$

$$f(\sigma_M^2\mid \text{Rest})\propto\prod_{i=1}^{n}f_s(\mathsf{M}_i\mid z_i,\beta_{0i},\alpha,\sigma_M^2)p(\sigma_M^2) \tag{16h}$$

The prior distributions used in the present study are,

$$p(\beta) = N[0, 10^6 \mathrm{I}_4],\ p(\gamma) = N[0, 10^6 \mathrm{I}_3],\ p(\alpha) = N[0, 10^3 \mathrm{I}_4],\ p(\eta) = N[0, 10^3 \mathrm{I}_3]$$

$$p(\sigma_\varepsilon) = U(0, 10^3),\ p(\sigma_M) = U(0, 10),\ p(\sigma_u) = U(1, 10^3); \tag{17}$$

where $U(a, b)$ defines the Uniform distribution with parameters a and b. As can be seen, all the prior distributions are very 'flat'. See the Comment at the end of this section regarding the choice of the prior distribution for σ_u^2.

6. SIMULATION RESULTS

The results of the simulation study are summarized in Tables 1-3. They are based on 500 replications, each consisting of generating a new population and selecting one sample of 50 schools and 12 pupils from each selected school by each of the four sampling methods described in Section 4.2. Table 1 shows the results obtained when ignoring the sample selection schemes and fitting the population model. These results serve as benchmarks for assessing the performance of the two approaches considered in this article for dealing with the effects of informative sampling. The *p*-values (P-V) in the table refer to the conventional *t*-tests of bias, with the standard deviation (SD) of the mean estimates computed as $(1/\sqrt{500})$ times the empirical SD of the estimates over the 500 replications. The parameter estimates in a given replication are the empirical means of 5000 observations drawn from the posterior distribution of each of the parameters under the population model (ignoring the sample selection schemes), after discarding the first 15000 values as 'burn in'.

Insert Table 1 about here

The results in Table 1 illustrate the kind of biases that can be encountered when ignoring an informative sample selection scheme. In the present experiment, informative sampling (response) of pupils within the schools (Method B2) has a much stronger biasing effect than informative selection of schools (Method A2). In particular, very large relative biases are obtained for the estimators of the 'between schools' variance, σ_u^2, and the two 'region coefficients' γ_1 and γ_2. The p-values for the significance of the bias show that all the estimators are biased under informative sampling (response) of pupils (Method B2). Statistically significant biases are obtained also for the estimators of the intercept γ_0 and the

variances σ_u^2 and σ_ε^2 under non-informative sampling (response) of pupils but informative selection of schools, but as mentioned above, the biases are much smaller in this case. When both the selection of schools and the sampling (response) of pupils are noninformative, all the percent relative biases are very small and nonsignificant, except for the estimator of σ_u^2. We discuss the problem with the estimation of σ_u^2 in the comment at the end of this section.

Table 2 shows the percent relative biases obtained under the two approaches that account for the sample selection and discussed in Sections 3 (probability weighting), and in Section 4 (use of the sample model). For the case of noninformative selection of schools (Method A1) and the use of the sample model, we show the results obtained when including among the model equations the equation defining the population distribution of the school sizes (Step2, Section 4.1), and also the results obtained without that equation. We refer to the first model as the 'full' sample model (FSM) and to the second model as simply the sample model (SM). Notice in this respect that the school selection probabilities, π_i, are proportional to the school sizes M_i, with the latter providing additional information on the random intercepts β_{0i}, and hence on the vector coefficient γ and the variance σ_u^2 indexing the distribution of the intercepts, irrespective of the sampling scheme used to select the schools. This additional information is 'automatically' accounted for in the case of informative selection of schools (method A2) via the corresponding sample model (Equation 16a).

Insert Table 2 about here

The biases in Table 2 are seen to be generally much smaller than the biases in Table 1, particularly under Method B2, but large (and statistically significant) biases still persist in the estimation of σ_u^2 under both methods. However, the biases are in this case much smaller when using the sample model compared to the use of probability weighting. All the other biases are in most cases statistically insignificant (the p-values are not shown in the table), except for the biases in the estimation of σ_ε^2 and γ_o that are occasionally significant. The latter biases, however, are small. As expected, the use of the full sample model (FSM) that includes the model holding for the school sizes reduces some of the biases obtained under non-informative selection of schools with the use of the sample model (SM) that ignores this relationship, particularly in the estimation of σ_u^2.

Table 3 shows the empirical percentage coverage of nominal 95% confidence (credibility) intervals (C.I.) for the model parameters, as obtained by ignoring the sampling process (IG) and by use of probability weighting (PW) and the sample models. The PW C.I. are the conventional C.I. obtained by approximating the distribution of the point estimators by the normal distribution. The randomization variances have been estimated using the sandwich-estimators developed in Pfeffermann *et al.* (1998a). When using the sample models or when ignoring the sampling process (assuming the population model), the C.I. have been constructed based on the 2.5% and 97.5% quantiles of the corresponding empirical posterior distributions.

Insert Table 3 about here

As becomes evident from Table 3, the use of either PW or the sample model yields in general acceptable coverage percentages, except in the case of the 'between schools' variance, σ_u^2, where the PW C.I. perform badly under all 4 sampling schemes. The sample models C.I.'s for σ_u^2 on the other hand perform generally well, despite the relatively large biases of the corresponding point estimators (see Table 2). The use of the sample model outperforms PW also under informative selection of schools and noninformative sampling (response) of pupils. The bad performance of the PW C.I. in the case of σ_u^2 suggests that the use of the conventional confidence intervals for this parameter is not justified with the sample sizes considered in this study. (See also the comment below). Ignoring the informative sampling schemes is seen to deteriorate the performance of the corresponding C.I. very severely under informative sampling (response) of pupils (Method B2). An unexpected result for which we don't have a clear explanation is that under informative sampling of pupils, the use of the sample model without the equation for the school sizes yields better C.I. for the γ-coefficients than the use of the full sample model. Notice, however, that the full sample model C.I. for σ_u^2 performs somewhat better than the sample model C.I., which is consistent with the results of Table 2 regarding the biases of the corresponding point estimators. We computed also for each of the methods the means and standard deviations of the lengths of the C.I. over the 500 replications (not shown), and none of the approaches dominates the other in this regard.

Comment: We mentioned above the large biases in the estimation of σ_u^2 with the use probability weighting under all four sampling schemes, and to a lesser extent with the use of the sample models via the posterior mean. The first point to be made in this respect is that unlike the estimation of the 'within school' variance σ_ε^2, that uses all the 50×12 individual (pupils) observations, the 'effective' sample size for the estimation of σ_u^2 is 50, the number of selected schools (see below for the effect of increasing the number of selected schools).

As regards the use of the sample model, another interesting (but seemingly not new) phenomenon encountered in our study is that the bias in estimating σ_u^2 depends also on the choice of the corresponding 'noninformative prior' distribution. In this article we followed the recommendation made in Gelman (2004) and used a noninformative Uniform prior distribution for the *standard deviation* σ_u (and the two other standard deviations), see Equation (17). (The use of WinBUGS requires setting a finite upper bound for the uniform distribution. We set the upper bound in the case of σ_M to be 10, since σ_M^2 is the variance of the log of the school sizes.) As discussed in Gelman (2004), the use of this prior with more than 2 second level units (schools) guarantees a proper posterior density, and it has other desirable properties. (Gelman considers a simple special case of the population model defined by (1) and (2), but points out that similar arguments in favour of the use of a uniform prior distribution for σ_u apply under more complicated models.).

Browne and Draper (2001) likewise noticed the strong dependency of the behaviour of the posterior mean of σ_u^2 on the choice of the prior distribution under a two-level model that is similar to the population model defined by (1) and (2). (In that paper the authors compare Bayesian and likelihood-based inference methods but they do not consider informative sampling). The authors found that the bias largely disappears when using a noninformative Inverse Gamma prior for σ_u^2 and estimating the variance by the posterior *median*, and that similar bias reductions are obtained when using a non-informative Uniform prior for σ_u^2 and estimating the variance by the posterior *mode*. They conclude that there is a clear trade-off between the choice of the prior distribution and the choice of the point estimator of this variance.

Another reason for the problems in estimating σ_u^2 in our case is the large ratio $(\sigma_\varepsilon^2/\sigma_u^2)=7.3$ between the two variances. In a simulation study with a very simple multi-level model, Kovacevic and Rai (2003) found that "the larger this ratio", the larger is the

relative bias of the estimator of σ_u^2, which seems very reasonable. Evidently, the effect of the magnitude of this ratio depends on the sample sizes of the first and second level units.

In order to study the effect of the number of schools on the behaviour of the estimators of σ_u^2, we repeated the runs for the case of noninformative sampling schemes at both levels (Methods A1 and B1), increasing the number of selected schools from 50 to 80. The percent relative biases obtained in this case are -7.1% under PW, 3.7% under the sample model and 2.2% under the full sample model. The corresponding biases for the case of 50 schools are (Table 2), -10.6%, 6.8% and 4.3% respectively.

7. ROBUSTNESS SIMULATION STUDY

The purpose of the analysis in this section is to study the robustness of the two approaches to possible misspecification of the models assumed for the sample selection probabilities under informative sampling schemes. To this end we changed the first and second level selection probabilities and then repeated the simulation study, assuming the original sample models defined by (8)-(11). Specifically, we generated the school sizes from a truncated non-central t-distribution with 3 degrees of freedom instead of the lognormal distribution defined in Step 2 of Section 4.1, and sampled the pupils within the selected schools with probabilities proportional to a size variable (PPS), instead of the stratified sampling scheme B2 defined in Section 4.2.

The model used for generating the school sizes is,

$$M_i \sim t_{(3)}(\lambda_0 + \lambda_1 \mathrm{Re}\,gion1_i + \lambda_2 \mathrm{Re}\,gion2_i + \lambda_3 \beta_{0i},\ \sigma_M^2) \qquad (18)$$

with $\lambda_0 = 3258.1,\ \lambda_1 = 3052.0,\ \lambda_2 = 2839.9$, $\lambda_3 = -30.3$ and $\sigma_{M^*}^2 = 1225$. School sizes smaller than 50 were set as 50 and sizes larger than 1200 were set as 1200, such that the range of the sizes is similar to the range in the simulation study of Section 6. Figure 1 shows the histogram of the actual sample of the school sizes under the sampling scheme A2 when the school sizes are generated by the t-distribution in (18). Figure 2 shows the histogram of the predicted sizes under the misspecified lognormal distribution assumed for the sample school sizes (Equation 15; the sizes in both figures are in the log scale). As becomes evident, the two distributions are very different, implying a bad fit of the misspecified model to the actual sizes.

Pupils within the selected schools were sampled with probabilities proportional to size, with the size defined as $s_{ij} = \exp(p_{ij})$ where the p_{ij}'s are the propensity scores defined under Step 5 in Section 4.1. Here again we can compare the actual sample selection probabilities with the misspecified selection probabilities obtained by assuming the stratified sampling scheme B2. The averages of the actual selection probabilities in the three strata (over the sampled clusters in 50 samples) are, 0.036, 0.047 and 0.055 respectively. Assuming the stratified sampling scheme B2, the 3 averages are 0.016, 0.018 and 0.025 respectively. It follows that the actual selection probabilities are much more variable than the probabilities assumed under the model.

The application of the robustness study in the case of probability weighting requires an explanation. As described in Section 3, the only information needed for the use of probability weighting are the first and second level selection probabilities, so that the performance of this approach does not depend on the models assumed for these probabilities. However, when the selection probabilities of the pupils within the schools are unknown, like in the case of nonresponse, these probabilities need to be estimated from the sample data. The present experiment allows therefore studying the effect of wrongly estimating these probabilities. (Assuming equal response probabilities within 'imputation classes' defined by the three strata, whereas the true response probabilities are proportional to the sizes s_{ij}.) For the selection of schools, we used the correct inclusion probabilities ($\pi_i \prec M_i$) with the M_i's generated by (18).

The results of the robustness study are exhibited in tables 4 and 5, which are analogous to Tables 2 and 3. Notice that the noninformative sampling schemes (Methods A1 and B1) were assumed to be known, so that the models assumed for the sampling probabilities in these cases are correct. The present robustness study is restricted therefore to situations where an informative sampling scheme at either level is misspecified.

Insert Tables 4 and 5 about here

The conclusions emerging from this study can be summarized as follows: Misspecification of the models assumed for the sampling probabilities has no biasing effect on the estimation of the coefficients $\beta_1 ... \beta_4$ and the variance σ_ε^2, indexing the first level model. Similarly, the confidence intervals computed for these parameters (Table 5) perform generally well under both approaches. In fact, the confidence intervals computed for these

parameters under the assumption of ignorable sampling schemes also perform very well, and the percent relative biases of the corresponding point estimators (not shown) are in all the cases less than 3.2%. This outcome is very different from the results obtained under the selection schemes underlying the results in Table 1, where large biases were obtained when ignoring sample selection within the schools.

The results of the robustness simulation study are generally satisfactory also with respect to the estimation of the second level model coefficients. Except for the case of informative sampling of schools (Method A2) and informative sampling of pupils (Method B2) where the estimator of γ_2 has bias of about 8% under probability weighting, the biases of the estimators of γ_0 and γ_2 are in all the other cases less than 4.5%. The biases of the estimators of γ_1 are somewhat larger, but not much larger than the biases obtained under noninformative sampling of schools and pupils (Methods A1 and B1). The confidence intervals computed under the two approaches perform similarly in the case of γ_0, with undercoverage of up to 11 percentage points compared to the 95% nominal level, but ignoring the sample selection schemes in this case results in more severe undercoverage. The use of probability weighting yields confidence intervals for γ_1 and γ_2 with undercoverage of up to 7 percentage points, but the use of the sample models credibility intervals yields undercoverage of at most 2.5 percentage points under all four sampling schemes.

The two approaches are not robust with regard to the estimation of σ_u^2 under the model misspecifications considered in the present study. Although the estimation of this variance is already problematic under correct model specification and even under noninformative sampling schemes (see the discussion in the comment at the end of Section 6), the biases under the misspecified models are even larger, particularly under informative selection of the schools when using the sample model, and under informative sampling of pupils (but noninformative sampling of schools), when applying the probability weighting approach. The percent undercoverage of the confidence intervals for σ_u^2 are likewise higher in the case of the misspecified model, particularly with the use of probability weighting under informative selection of schools. Recalling that in the application of probability weighting we used the correct school selection probabilities, the empirical coverage of 71.% in the case of informative selection of schools (Method A2) and noninformative sampling of pupils (Method B1) seems odd, but notice that a severe undercoverage was already observed for the same sampling schemes in Table 3 (78.6% coverage). As in the case of the γ-coefficients,

the credibility intervals for σ_u^2 computed by use of the sample models perform generally better than the probability weighted confidence intervals, but an undercoverage of 14 percentage points is observed for the case where both sampling schemes are informative.

In summary, both probability weighting and the use of the sample model seem to be equally robust with regard to point estimation of the model coefficients, but the use of the sample models yields confidence (credibility) intervals with somewhat better coverage properties even under the misspecified models for the sample selection probabilities. The two approaches fail to yield reliable point estimators and confidence intervals for σ_u^2, but as mentioned before, the estimation of this variance is known to be problematic even under the classical (population) multi-level model with no sampling effects.

8. FURTHER REMARKS AND CONCLUSIONS

An important message reinforced in the present study is that ignoring an informative sample selection scheme and fitting the population model may yield large biases of point estimators that distort the analysis. We describe and compare two approaches to control the bias. The first approach uses probability weighting to obtain approximately unbiased and consistent estimators for the corresponding census estimators under the randomization (repeated sampling) distribution. The census estimators are the hypothetical estimators computed from all the population values, and with large populations they can be expected to be sufficiently close to the true model parameters. The second approach attempts to identify the parametric model holding for the sample data as a function of the population model and the first order sample selection probabilities, and then fits the sample model to the sample data. The two approaches have been shown in the simulation experiment to remove the bias of all the point estimators except in the case of the 'between school' variance σ_u^2, where with a small number of schools the use of probability weighting produces large biases under all the sampling schemes considered, including the non-informative scheme where both the selection of schools and the sampling of pupils within the selected schools is by simple random sampling. The use of the sample model likewise produces biased estimators for this variance with a small number of schools, and as discussed in Section 6, the bias depends also in this case on the choice of the corresponding prior distribution.

Probability weighting has two important advantages over the use of the sample model. First and foremost, it does not require any additional assumptions beyond the specification of

the population model, although the validation of the model under this approach is an open problem. The second advantage of this approach is that it is very simple and requires minimal computation resources, including the estimation of the variances of the point estimators. The use of this approach has, however, some serious limitations already discussed in the introduction. In particular, it requires large sample normality assumptions for the computation of confidence intervals.

The use of the sample model is more flexible and with the specification of appropriate prior distributions, it allows simulating from the posterior distribution of the target parameters. This advantage of the use of the sample model is demonstrated in Tables 3 and 5 where we compare the percentage coverage of confidence intervals produced by the two approaches. Inference based on the sample model requires, however, the specification of the conditional expectations of the sample selection probabilities at the various levels of the model hierarchy, given the values of the corresponding dependent and independent variables. As illustrated in the present study, these expectations may depend on a large number of unknown parameters that need to be estimated along with the population parameters. Application of this approach with the aid of MCMC simulations is computation intensive and as discussed in Section 6, with a small number of second level units, the performance of the variance estimators is rather erratic and may depend on the specification of the prior distributions, even when restricting to 'non-informative' priors. Nonetheless, with 'correct' specification of the sample model the use of this approach overcomes the inference limitations of probability weighting noted in the introduction.

The robustness study carried out in this article suggests that even quite drastic misspecification of the models assumed for the sample selection probabilities may only have a modest effect on the estimation of the model coefficients and the performance of the corresponding confidence intervals, but large biasing effects are observed when estimating the 'between school' variance σ_u^2. With a small number of second level units, the estimation of this variance is known to be problematic even under the standard multi-level model with no biasing sampling effects, but the biases obtained under the misspecified model indicate the need for powerful diagnostic procedures for the identification of the models underlying the sample selection schemes. As discussed in the Introduction, the use of the sample model is inevitable under informative selection of the second level units, when the objective of the analysis is the prediction of characteristics of these units, like in small area estimation problems.

REFERENCES

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance structures with linear structure. *Annals of Statistics*, **1**, 135-141.

Browne, W. J., and Draper, D. (2001). A comparison of Bayesian and likelihood-based methods for fitting multilevel models (unpublished manuscript).

Gelman, A. (2004). Prior distributions for variance parameters in Hierarchical models (unpublished manuscript).

Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, **73**, 43-56.

Kovacevic, M. S., and Rai, S. N. (2003). A pseudo maximum likelihood approach to multilevel modelling of survey data. *Communications in Statistics, Theory and Methods*, **32**, 103-121.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998a). Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society, Series B, **60**, 23-40.

Pfeffermann, D., Krieger, A.M., and Rinott, Y. (1998b). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, **8**, 1087-1114.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya*, Series B, **61**, 166-186.

Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, **54**, 499-513.

Spiegelhalter, D., Thomas, A., and Best, N.G. (2003). Bayesian Inference using Gibbs Sampling. WinBUGS version 1.4, User manual. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, U.K.

Table 1. *Percent relative bias (PRB) and p-values (P-V) of tests of bias when ignoring the sampling process.*

	Selection of Schools							
	Non Informative, Method A1				Informative, Method A2			
Selection of Students	Non Informative Method B1		Informative Method B2		Non Informative Method B1		Informative Method B2	
Parameter	PRB	P-V	PRB	P-V	PRB	P-V	PRB	P-V
γ_0= 86.9	0.5	3.1	9.2	0.0	-7.0	0.0	4.1	0.0
γ_1= -6.8	-4.1	22.2	27.3	0.0	0.9	79.0	27.1	0.0
γ_2= -13.8	-3.4	4.1	36.5	0.0	-0.7	67.6	37.2	0.0
β_1= -10.9	-1.6	15.2	-13.7	0.0	0.7	52.8	-12.5	0.0
β_2= -16.0	0.2	76.3	-13.9	0.0	0.9	27.3	-12.2	0.0
β_3= -36.5	0.3	51.2	-16.1	0.0	0.3	57.0	-15.9	0.0
β_4= -7.2	-1.0	57.8	-15.8	0.0	-0.1	93.8	-14.4	0.0
σ_u^2=132.2	6.8	0.0	-80.9	0.0	6.6	0.0	-79.9	0.0
σ_ε^2=963.0	0.4	11.2	18.9	0.0	1.0	0.0	20.1	0.0

Table 2. *Percent relative bias (PRB) when accounting for the sampling process by use of probability weighting (PW), the sample model (SM), and the full sample model (FSM).*

	Selection of Schools									
	Non Informative, Method A1						Informative, Method A2			
Selection of Students	Non Informative Method B1			Informative Method B2			Non Informative Method B1		Informative Method B2	
Parameter	PW	SM	FSM	PW	SM	FSM	PW	SM	PW	SM
γ_0= 86.9	0.5	0.5	0.2	1.0	1.2	0.6	-1.2	-0.3	-0.3	0.2
γ_1= -6.8	-4.1	-4.1	-4.2	-0.4	-1.0	-1.4	4.7	-0.3	1.9	-3.0
γ_2= -13.8	-3.5	-3.4	-3.1	-1.3	-1.7	-2.7	-0.5	-1.5	-0.9	-1.4
β_1= -10.9	-1.5	-1.6	-0.3	-2.3	-2.1	-0.3	2.3	1.9	-1.1	0.2
β_2= -16.0	0.4	0.2	1.1	-2.5	-2.3	-0.9	1.4	1.6	-0.6	0.6
β_3=-36.5	0.4	0.3	0.6	-2.7	-2.3	-0.7	0.7	0.5	-2.0	0.0
β_4= -7.2	-0.8	-1.0	0.4	-3.4	-3.3	-1.3	-0.5	1.4	-4.3	-0.2
σ_u^2 =132.2	-10.6	6.8	4.3	-20.0	-14.7	-8.1	-15.1	4.8	-21.5	-6.5
σ_ε^2 =963.0	-1.0	0.4	0.4	-0.2	-0.7	0.0	-0.6	0.9	0.1	0.8

Table 3. *Percent coverage of nominal 95% confidence intervals when ignoring the sampling process (IG), under probability weighting (PW), the sample model (SM) and the full sample model (FSM).*

	Selection of Schools												
	Non Informative, Method A1							Informative, Method A2					
Selection of Students	Non Informative Method B1			Informative Method B2				Non Informative Method B1			Informative Method B2		
Parameter	PW	SM =IG	FSM =IG	IG	PW	SM	FSM	IG	PW	SM	IG	PW	SM
γ_0= 86.9	92.8	93.8	92.6	41.4	91.6	93.0	89.4	71.2	89.4	92.8	83.2	92.0	89.8
γ_1= -6.8	95.2	96.2	95.0	91.2	94.4	95.2	91.2	94.6	91.2	94.6	93.0	90.0	92.0
γ_2= -13.8	94.0	94.8	93.6	74.0	94.2	95.2	94.2	94.2	90.8	93.6	72.8	91.0	91.2
β_1= -10.9	95.2	95.2	94.6	89.8	92.8	92.2	90.8	93.8	92.2	93.6	92.6	94.0	94.2
β_2= -16.0	95.4	96.2	95.6	91.0	95.4	96.0	93.6	95.8	92.0	95.4	91.6	95.8	94.6
β_3= -36.5	93.4	94.4	94.2	73.8	96.0	93.6	93.6	95.4	95.2	95.2	76.8	92.6	94.4
β_4= -7.2	93.6	95.0	95.4	95.8	96.6	95.6	95.4	95.2	91.4	96.0	91.2	92.0	92.0
σ_u^2 =132.2	87.8	94.8	95.0	8.6	79.2	92.4	94.2	94.8	78.6	94.8	9.4	72.0	92.2
σ_ε^2 =963.0	94.0	94.8	94.8	16.6	94.0	95.4	95.8	96.6	93.8	97.2	12.2	93.4	97.0

Table 4. *Percent relative bias (PRB) when accounting for the sampling process by use of probability weighting (PW), the sample model (SM), and the full sample model (FSM); Robustness study.*

	Selection of Schools									
	Non Informative, Method A1						Informative, Method A2			
Selection of Students	Non Informative Method B1			Informative Method B2			Non Informative Method B1		Informative Method B2	
Parameter	PW	SM	FSM	PW	SM	FSM	PW	SM	PW	SM
γ_0= 86.9	0.0	0.0	-0.3	4.2	4.0	3.6	-0.4	-1.1	3.3	-2.6
γ_1= -6.8	4.6	4.9	3.8	-1.4	2.0	3.8	-4.5	-5.9	-5.3	0.0
γ_2= -13.8	1.1	1.2	1.3	-4.5	1.0	0.7	-2.6	-3.5	-8.2	-2.2
β_1= -10.9	-1.0	-1.0	0.1	-1.3	-3.6	-2.4	0.9	2.1	-1.3	-1.4
β_2= -16.0	-0.1	-0.3	0.7	0.6	-1.0	0.5	-1.6	-0.4	-0.3	-0.5
β_3= -36.5	0.2	0.1	0.4	0.4	-1.6	-0.4	-1.1	-0.5	0.3	-0.1
β_4= -7.2	-0.2	-0.3	0.2	1.9	0.2	2.0	-2.8	-0.9	3.3	3.5
σ_u^2=132.2	-10.8	6.6	4.1	-9.1	-15.4	-8.7	-18.7	15.1	-22.0	-29.1
σ_ε^2=963.0	-0.7	0.7	0.7	-1.5	4.3	2.6	-1.0	-0.5	-1.5	2.2

Table 5. *Percent coverage of nominal 95% confidence intervals when ignoring the sampling process (IG), under probability weighting (PW), the sample model (SM) and the full sample model (FSM); Robustness study.*

	Selection of Schools												
	Non Informative, Method A1							Informative, Method A2					
Selection of Students	Non Informative Method B1			Informative Method B2				Non Informative Method B1			Informative Method B2		
Parameter	PW	SM = IG	FSM =IG	IG	PW	SM	FSM	IG	PW	SM	IG	PW	SM
γ_0= 86.9	94.2	95.0	93.2	79.2	86.4	85.8	83.6	77.0	89.2	91.4	94.8	87.0	87.8
γ_1= -6.8	94.2	96.0	94.8	95.6	94.6	95.0	92.4	94.6	90.2	92.8	95.4	89.4	93.4
γ_2= -13.8	92.0	93.2	92.2	95.6	95.0	94.8	93.4	95.4	88.8	94.4	96.2	88.0	93.2
β_1= -10.9	93.8	94.6	94.8	94.6	93.4	95.2	94.4	96.2	93.4	95.4	94.4	93.4	93.6
β_2= -16.0	94.2	95.4	95.8	93.8	92.2	94.0	91.0	94.2	93.6	94.6	94.8	93.2	93.6
β_3= -36.5	95.2	95.6	96.0	95.8	94.2	94.8	95.8	95.8	94.8	96.0	93.8	91.2	93.0
β_4= -7.2	94.8	95.6	95.8	95.8	95.4	95.0	94.8	94.8	92.8	93.6	95.0	92.8	94.4
σ_u^2 =132.2	84.4	93.8	93.4	94.8	87.2	90.0	92.8	90.4	71.4	87.8	89.0	67.2	81.2
σ_ε^2 =963.0	93.8	94.8	94.4	94.6	91.6	91.0	93.0	94.8	91.8	94.8	95.4	91.4	95.4

Figure 1. *Histogram of actual sample of school sizes when the population school sizes are generated from the truncated t-distribution. Sampling scheme* A2

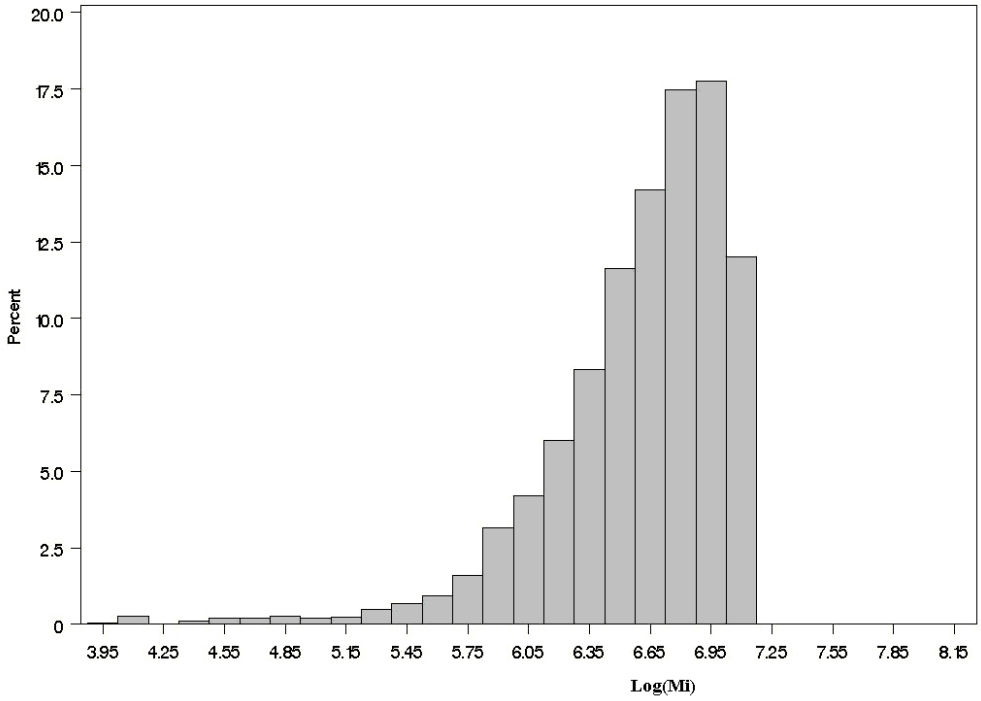

Figure 2. *Histogram of predicted sample of school sizes under lognormal assumption when the population school sizes are generated from the t-distribution. Sampling scheme* A2