Multi-level Modelling Under Informative Probability Sampling

Danny Pfeffermann
Hebrew University, Israel
Fernando Moura
Federal University of Rio de Janeiro, Brazil
Pedro Nascimento Silva, IBGE, Brazil

1 Introduction

Multi-level models are frequently used in the human and biological sciences for the modelling of hierarchically clustered populations. Classical theory underlying the use of these models assumes implicitly either that all the clusters at all levels are represented in the sample or that they are selected by simple random sampling. This assumption may not hold in a typical sample survey under which the clusters and/or the final sampling units are often selected with unequal selection probabilities. When the selection probabilities are related to the values of the response variable even after conditioning on the model covariates, the sampling process becomes *informative* and the model holding for the sample data is 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' proficiencies with schools as the second level units and pupils as first level units, and suppose that the schools are selected to the sample with probabilities proportional to their sizes. If the size of the school is related to the school's achievements level, say, the large schools are mostly in poor areas with low achievements, and the size of the school is not included among the model covariates, the sample of schools will tend to contain large schools with low achievements and hence no longer represent the population schools.

In a recent article, Pfeffermann *et al.* (1998) propose *probability weighting* procedures of first and second level units that adjust for the effect of informative sampling on the estimation of two-level models. The authors develop also appropriate variance estimators. The use of these procedures is justified based on asymptotic arguments but they are shown to perform well in a simulation study even with moderate sample sizes. Nonetheless, the use of the sample weights for bias correction has three important limitations:

1- the weights are designed so as to protect against the randomisation bias over all possible sample selections. As well known, the use of this kind of weights often inflates the standard errors of point estimators.
2- Inference is restricted primarily to point estimation. Probabilistic statements require asymptotic normality assumptions. The exact distribution of weighted point estimators is generally not traceable.
3- The use of the sampling weights does not permit to condition on the selected sample of clusters (second and higher level units) or values of the model explanatory variables.

The purpose of this article is to propose a *model dependent* approach for multi-level model analysis that accounts for informative sampling. The idea behind the proposed approach is to extract the hierarchical model holding for the sample data as a function of the population model and the first order sample inclusion probabilities. The resulting *sample model* is then analysed using classical theory. Evidently, if the sample model is correctly specified, the use of this approach overcomes the limitations mentioned with respect to probability weighting. We restrict for convenience to a two-level model and apply the full Bayesian paradigm by use of Markov Chain Monte Carlo (MCMC) integration, but the approach can be extended to higher level models and different inference procedures.

In Section 2 we define the population model and sampling design and extract the corresponding sample model. Section 3 discusses estimation details and Section 4 describes a simulation experiment designed to study the performance of the proposed approach and compare it to the practice of ignoring the sampling design. The simulation study follows closely the Brazilian Basic Education Evaluation study in terms of sampling design, survey variables and sample sizes. We conclude with a brief outline of future work in Section 5.

2 Population Model, Sampling Design and Corresponding Sample Model

2.1 Population Model

Consider the following two-level hierarchical model:

First level: $y_{ij} \mid \beta_{0i} = \beta_{0i} + x_{ij}'\beta + e_{ij}\,;\, e_{ij} \sim N(0, \sigma_e^2) \quad j = 1..M_i$ (1)

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

This model is often referred to in the literature as the *random intercept regression model* and it contains as unknown parameters the vector coefficients β and γ and the first and second level variances σ_e^2 and σ_u^2. Here the intercepts are modelled as linear functions of known regressor variables z_i. In the simulation experiment the outcome y_{ij} is the test score of pupil j in school i, x_{ij} defines the sex, age and parents' education of the same

pupil and z_i consists of two dummy variables defining the geographical region to which school i belongs. The second level random effects u_i account for the variation of the intercept terms not explained by the variables z_i. In what follows we refer for convenience to the first level units as 'pupils' and the second level units as 'schools'.

2.2 Sampling Design

We assume two stage sampling. In the first stage $n < N$ schools are selected with probabilities $p_i = Pr(i \in s)$ that may be correlated with the random effects u_i. In the second stage m_i pupils are sampled from school i selected in the first stage with probabilities $p_{j|i} = Pr(j \in s_i \mid i \in s)$ that may be correlated with the residuals e_{ij}. In the description of the simulation study we elaborate on the sampling process in more detail.

2.3 The Sample Model

Following Pfeffermann *et al.* (1998), the sample distributions of the first and second level units are:

$$f_s(y_{ij} \mid x_{ij}, q_i) = f(y_{ij} \mid x_{ij}, q_i, j \in s_i) = E_p(p_{j|i} \mid y_{ij}, x_{ij}, q_i)\, f_p(y_{ij} \mid x_{ij}, q_i) / E_p(p_{j|i} \mid x_{ij}, q_i) \quad (3)$$

$$f_s(b_{0i} \mid z_i, l) = f(b_{0i} \mid z_i, l, i \in s) = E_p(p_i \mid b_{0i}, z_i, l) f_p(b_{0i} \mid z_i, l) / E_p(p_i \mid z_i, l) \quad (4)$$

where $q_i = (b_{0i}, b', s_e^2)$, $l = (g', s_u^2)$ and $f_p(\cdot)$ and $f_s(\cdot)$ are the population and sample distributions respectively.

The expectations in (3) and (4) can be modelled based on knowledge of the sampling process and the sample data. See Pfeffermann and Sverchkov (1999) for discussion and examples. The corresponding expressions for the cases considered in the simulation study are presented in Section 4.

3 Estimation via Markov Chain Monte Carlo Integration (MCMC)

The population model defined by (1) and (2) has a Bayesian Hierarchical structure with b, g, s_e^2 and s_u^2 as hyper-parameters. The MCMC algorithm consists of sampling alternately from the conditional distribution of each of the unknown quantities, given the data and the remaining quantities. Our proposed adjustment for the effect of informative sampling consists of replacing the population densities $f_p(y_{ij} \mid x_{ij}, q_i)$ and $f_p(b_{0i} \mid z_i, l)$ defined by (1) and (2) by their sample counterparts defined by (3) and (4). In the simulation study we assign non-informative priors to the hyper-parameters as follows: $b \sim U_4(-\infty, +\infty)$, $g \sim U_3(-\infty, +\infty)$, $t_e = 1/s_e^2 \sim Gamma(0.01, 0.01)$ and $t_u = 1/s_u^2 \sim Gamma(0.01, 0.01)$. The notation $U_p(-\infty, +\infty)$ signifies that the p components of the vector random variable have independent uniform distributions over the real line. The MCMC computations have been implemented by use of Version 1.3 of the WinBUGS program (Spiegelhalter *et al.*, 2000).

4 Monte-Carlo Simulation Experiment

4.1 Generation of Population Values

The purpose of the present simulation experiment is to compare the performance of the proposed approach with the "standard" alternative of ignoring the sampling process in the estimation process. We plan to compare the approach with the weighting procedures proposed by Pfeffermann *et al.* (1998) in the near future. The model and sampling design underlying this experiment mimic a real data set obtained from the Basic Education Evaluation study (hereafter the BEE study) carried out in 1996 for the municipality of Rio de Janeiro in Brazil.

The experiment was carried out by generating 500 populations and selecting four samples from each population, with each sample selected by a different sampling design. The number of schools in each population was *N = 392* (as in the real data). The populations were generated in 5 steps:

Step 1- Generate school random intercept terms as $\hat{a}_{0i} = 86.914 - 6.782 \times \mathbf{Region1}_i - 13.772 \times \mathbf{Region2}_i + u_i$ with $u_i \sim N(0, 11.496^2)$ independently between schools. The variables $Region1_i$ and $Region2_i$ are dummy variables defining three geographical regions. The schools were classified to the three regions in the same proportions as in the BEE study (the first 65 schools in region 1, the next 213 schools in region 2 and the last 114 schools in region 3).

Step 2 – Generate school sizes M_i as $M_i = INT[\ a_i \times \exp(r \times \hat{a}_{0i})\ ]$ where r=0.0506 and the a_i vary according to region; $a_1 = 2.20$, $a_2 = 3.11$ and $a_3 = 4.42$. The parameter values were chosen so that the resulting school sizes behave similarly to the BEE data.

Step 3 – Set explanatory variables values (x_{ij}) for the M_i students in school $i, i = 1...N$ by selecting at random (with replacement) M_i vector values of explanatory variables from the corresponding BEE data in the region containing school i. The explanatory variables are dummy variables defining *Sex* (1 for females), *Age1* (1 for ages 15-16), *Age2* (1 for ages 17 and older) and Parents education (1 for pupils with at least one parent with university education).

Step 4 – Generate proficiency scores for student j in school i using the model,

$y_{ij} = \hat{a}_{0i} - 10.936 \times Sex_{ij} - 16.022 \times Age1_{ij} - 36.458 \times Age2_{ij} - 7.161 \times Parents_{ij} + \hat{a}_{ij}$, where $\hat{a}_{ij} \sim N(0;\ 963.049)$.

In the BEE study pupils were asked to grade the difficulty of the proficiency test. Corresponding grades were generated in the simulation experiment as follows:

Step 5 – Generate evaluation scores $e_{ij} = b_0 + b_1 y_{ij} + \hat{\imath}_{ij}$, where $b_0 = 1.60$, $b_1 = 0.006$ and $x_{ij} \sim N(0, 0.23^2)$. These parameter values satisfy, $Corr\ (y_{ij}, e_{ij}) = 0.6$. Set grades $O_{ij} = 1$ *if* $e_{ij} < 1.76$, $O_{ij} = 2$ *if* $1.76 < e_{ij} < 2.17$; and $O_{ij} = 3$ *if* $2.17 < e_{ij}$. The cut-off values were chosen so that the proportion of the three grades in the simulated populations matches the proportions in the BEE data.

4.2 Sampling Schemes

We consider two different methods for the selection of schools and two different methods for the sampling of pupils within the selected schools, defining a total of 4 different two-stage sampling schemes. Schools were selected using a) simple random sampling without replacement (SRSWR) and b) probability proportional to size (PPS), using Sampford method. The second method is informative as the sizes M_i depend on the intercepts β_{0i}. Students within the selected schools were sampled by a) SRSWR and b) disproportionate stratified sampling with the strata defined by the grades o_{ij} (three strata, see Stage 5). The second method is informative as the grades depend on the evaluation scores e_{ij}, which in turn depend on the proficiency scores y_{ij}. One sample of 40 schools and 10 pupils from each selected school was drawn from each population under each of the 4 sampling schemes. For the stratified sample selection we sampled 4 pupils from the first 2 strata *(O_{ij}=1,2)* and 2 pupils from the third stratum *(O_{ij}=3)*.

4.3 Sample models under the sampling schemes considered

The sample models for general two-stage sampling schemes are defined by (3) and (4). The expectations defining these models under the informative sampling schemes considered in the present simulation experiment are as follows:

$$E_p(p_{ji} \mid y_{ij}, x_{ij}, q_i) = \sum_{k=1}^{3} q_k^i P(O_{ij} = k \mid y_{ij}, x_{ij}, q_i) = q_1^i A_1(y_{ij}) + q_2^i \{A_2(y_{ij}) - A_1(y_{ij})\} + q_3^i \{1 - A_2(y_{ij})\} \quad (5.1)$$

where $q_k^i = \Pr(j \in s_i \mid o_{ij} = k)$ is the sampling fraction in school i of stratum k, $A_k(y_{ij}) = \Pr(e_{ij} < c_k \mid y_{ij}, x_{ij}, q_i) = \Phi[(c_k - b_0 - b_1 y_{ij}) / s_x]$ and $c_1 = 1.76,\ c_2 = 2.17$ are the cutoff values defining the grades O_{ij} (see Step 5). Notice that by definition of the evaluation scores e_{ij} (Step 5), $s_x^2 = b_1^2 s_e^2 (1 - r^2) / r^2$ where $r = Corr\ (y_{ij}, e_{ij})$. (r =0.6 in the set up of present experiment),

$$E_p(p_{ji} \mid x_{ij}, q_i) = \sum_{k=1}^{3} q_k^i P(O_{ij} = k \mid x_{ij}, q_i) = q_1^i B_1(m(y_{ij})) + q_2^i \{B_2(m(y_{ij})) - B_1(m(y_{ij}))\} + q_3^i \{1 - B_2(m(y_{ij}))\} \quad (5.2)$$

where $B_k(m(y_{ij})) = \Pr(e_{ij} < c_k \mid x_{ij}, q_i) = \Phi[\sqrt{1 - r^2}(c_k - b_0 - b_1 m(y_{ij}) / s_x]$; $m(y_{ij}) = \beta_{0i} + x_{ij}' \beta = E_p(y_{ij} \mid x_{ij}, q_i)$,

$E_p(p_i \mid b_{oi}, z_i, l) \cong a_i \exp[r b_{0i}]$; $E_p(p_i \mid z_i, l) \cong a_i \exp[r z_i \gamma + r^2 s_u^2 / 2)]$. (6)

The two approximations in (6) assume that $p_i = nM_i / N\bar{M} \cong nM_i / E_p(M_i)$ where $\bar{M} = \sum_{i=1}^{N} M_i / N$ is the population mean of the school sizes.

4.4 Simulation results

The results reported in this section are based on 500 replications. Table 1 shows the empirical absolute relative bias and the *p*-values of the conventional *t*-tests of bias for the various parameter estimates, as obtained under the four sampling schemes when postulating the population distribution (ignoring the method of sampling), and

when accounting for the correct sample distribution. (For the case of SRSWR of schools and SRSWR of pupils there is only one set of measures.) The results in Table 1 suggest that informative sampling of pupils within the schools (by use of the disproportionate stratified sampling) has a much stronger biasing effect than informative selection of schools (use of PPS sampling), with particularly large biases when estimating the between schools variance s_u^2 and the two region coefficients g_1 and g_2 . The use of the sample distribution reduces these biases very drastically but in the case of the between schools variance a statistically significant relative bias of about 24% is still present. We need to explore further the explanation for this bias. The only other appreciable relative bias under the sample distribution is found when estimating g_1 under SRSWR of schools and stratified sampling of pupils (relative bias = 10.5%), but this bias is non-significant. There are a few other significant biases under the sample distribution (e.g., in the estimation of s_e^2), but the corresponding relative biases are small. These kinds of biases are not unusual when applying the MCMC approach.

Table 1. Percent Absolute Bias (PAB) and P-Values (P-V) of Tests of Bias when Ignoring the Sampling Process and when Employing the Sample Distribution Under Four Sampling Schemes

Selection of Schools	**Non Informative**						**Informative**							
Sample of pupils	**Non Infor.**		**Informative**				**Non Informative**				**Informative**			
Dist. for Inference	**Pop. Dist.**		**Pop. Dist.**		**Sam. Dist.**		**Pop. Dist.**		**Sam. Dist.**		**Pop. Dist.**		**Sam. Dist.**	
Empirical Measures	**PAB**	**P-V**	**PAB**	**P-V**	**PAB**	**P-V**	**PAB**	**P-V**	**PAB**	**P-V**	**PAB**	**P-V**	**PAB**	**P-V**
γ_0= 86.91	0.4	0.35	4.3	0.00	0.2	0.79	6.6	0.00	0.4	0.34	1.3	0.00	0.7	0.15
β_1=-10.94	1.1	0.40	0.5	0.71	2.4	0.14	0.8	0.56	1.1	0.42	0.8	0.59	3.7	0.04
β_2=-16.02	0.9	0.40	3.2	0.00	0.5	0.68	0.7	0.46	0.7	0.50	3.1	0.00	2.0	0.13
β_3=-36.46	0.6	0.33	1.8	0.01	0.7	0.27	0.4	0.53	1.0	0.10	2.0	0.00	0.7	0.36
β_4=-7.16	4.3	0.04	1.8	0.39	0.7	0.73	3.7	0.07	2.1	0.30	1.9	0.35	2.7	0.22
γ_1= -6.78	3.3	0.55	20.4	0.00	10.5	0.28	7.1	0.11	7.8	0.08	18.7	0.00	1.0	0.82
γ_2= -13.77	2.3	0.43	27.3	0.00	5.0	0.30	4.5	0.06	4.9	0.05	26.9	0.00	0.7	0.77
σ^2_u=132.16	1.0	0.62	80.5	0.00	22.9	0.00	4.0	0.05	4.1	0.05	79.6	0.00	24.2	0.00
σ^2_ε=963.05	1.4	0.00	11.3	0.00	2.6	0.00	1.6	0.00	0.9	0.01	10.7	0.00	3.1	0.00

5 Future work

The results obtained so far are encouraging in terms of the use of the sample distribution but it should be noted that the sample model (3)-(6) used for this experiment is correct. In practice the relationships between the sample selection probabilities and the model dependent variables need to be identified from the sample, see Pfeffermann and Sverchkov (1999) for discussion and examples. We plan to study the performance of the estimators obtained when identifying the models holding for the sample selection probabilities from the sample data, and also compare the results with the results obtained by the weighting procedures proposed in Pfeffermann *et al.* (1998).

REFERENCES

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998). 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. (1998). 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.

Spiegelhalter, D., Thomas, A., and Best, N.G. (2000). Bayesian Inference using Gibbs Sampling. WinBUGS version. 1.3, User manual. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, U.K.

RÉSUMÉ

Cette communication propose une approche Bayesienne, dépendant d' un modèle, pour traiter les modèles multiniveaux sous un processus de sondage informatif. L' approche se base sur la développement du modèle multiniveau qui tient pour l' échantillon comme une fonction du modèle de la population et d' employer l' algorithme MCMC pour une estimation Bayesienne. Des résultats empiriques démontrent l' efficacité de l' approche proposée.