Multivariate Hierarchical Normal Modelling under Informative Sampling

Pedro Luis do Nascimento Silva[1]
Fernando Antonio da Silva Moura[2]
[1] Escola Nacional de Ciências Estatísticas - IBGE, Rio de Janeiro, Brasil
[2] Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil
Corresponding author: Pedro Silva – Email: pedronsilva@gmail.com

In this paper a model-dependent approach for multivariate hierarchical normal modelling that accounts for informative probability sampling of first and second level population units is developed. The approach involves extracting the hierarchical model holding for the sample data given the selected sample as a function of the corresponding population model and the sample selection probabilities, and then fitting the resulting sample model using Bayesian methods. This approach was developed earlier for univariate responses (Pfeffermann, Moura & Silva, 2006). An application of the approach is presented for modelling jointly Mathematics and Portuguese Language proficiency scores obtained from a Brazilian evaluation study of basic education conducted by the Brazilian National Institute of Education Research (INEP). The scores stem from applying Item Response Theory models to test results from the 'Prova Brasil 2009' study. A two-level multivariate hierarchical normal model is fitted, where the students and schools are the (first level) units and the groups (second level units) respectively. The analysis is restricted to the students from the 8th grade in elementary schools from the municipality of Rio de Janeiro. Simulation is also carried out in order to assess the frequentist properties of the approach.

Keywords: Credibility interval; Markov Chain Monte Carlo; Probability weighting; Educational assessment; sample model; multivariate normal.

1. Introduction

Multilevel or hierarchical models are frequently used to model socioeconomic data in a variety of contexts. In many applications modeling is carried out with data obtained from complex sample surveys, possibly with informative sample designs and/or response. Sample designs and response mechanisms are said to be informative when the model holding for the observed sample is not identical to the model holding for the population which the sample aims to describe.

When the sample design or response mechanism is informative, 'naïve' fitting of the multilevel model to the sample data without accounting for the design and/or response mechanism will produce inadequate inference for the model holding for the population (Pfeffermann, Moura & Silva, 2006), in the sense that it may yield biased estimates for the model parameters and for the mean square errors of the parameter estimates.

To tackle this issue, several approaches were developed which attempt to incorporate survey design weights in the estimation of the multilevel population model parameters – see for example Pfeffermann et al (1998), Kovacevic & Rai (2003), Grilli & Pratesi (2004), Asparouhov (2006), and Rabe-Hesketh & Skrondal (2006). Following an alternative path, Pfeffermann, Moura & Silva (2006) proposed using the 'sample model' approach for multi-level modelling under informative multi-stage sampling. Their approach first extracts the hierarchical model holding for the sample data given the selected sample, henceforth called the 'sample model', as a function of the corresponding population model and the first- and lower-level sample selection probabilities. It then fits the 'sample model' using Bayesian methods.

The general 'sample model' approach was first considered by Krieger & Pfeffermann (1997) for testing of population distribution functions, by Pfeffermann & Sverchkov (1999, 2003) for the fitting of linear and generalised linear population

regression models, and Sverchkov & Pfeffermann (2003) for the prediction of finite population totals. Pfeffermann & Sverchkov (2007) used this approach to obtain small area estimates under informative sampling. In the small area estimation context, Verret, Hidiroglou & Rao(2010) used the survey weights as an additional auxiliary variable when fitting a model to the sample and/or in the estimation of means and MSEs using the pseudo-EBLUP approach proposed by You and Rao (2002).

All of the above mentioned approaches attempt to estimate the parameters of the population model accounting for the sample selection and possibly also response probabilities. None of the above papers dealt with the case when the proposed population multilevel model is used to explain a multivariate response. Multivariate responses emerge in the context of educational assessment studies where students are assessed using two or more proficiency tests, each test providing a proficiency score. Such scores are typically correlated even after conditioning on covariates.

In this paper the sample model approach is applied for multivariate hierarchical normal modeling of sample data accounting for informative probability sampling of first and second level population units. The approach consists of first extracting the hierarchical model holding for the sample data given the selected sample as a function of the corresponding population model and the sample selection probabilities, and then fitting the resulting sample model using Bayesian methods. This approach evolved from similar models developed earlier for univariate responses.

2. Population and Sample Models

Let **y** denote a vector of response variables of interest, and **x** a vector of auxiliary variables. We consider the following multivariate multilevel (or hierarchical) model given by:

$$y_{ijk} = \mathbf{x'}_{ijk}\boldsymbol{\beta}_k + \nu_{ik} + \varepsilon_{ijk} \qquad \textbf{(1)}$$

where $\mathbf{y}_{ijk}$ denotes the value of the *k*-th response variable for unit *j* in group *i*, $\mathbf{x}_{ijk}$ is the vector of auxiliary variables used to explain the *k*-th response variable for unit *j* in group *i*, $\boldsymbol{\beta}_k$ denotes the vector of regression coefficients for the *k*-th response variable, ν_{ik} is the common random intercept for the *k*-th response variable for all units in group *i*, and ε_{ijk} is the unit level error for the *k*-th response variable for unit *j* in group *i*, $k=1,\ldots,p$, $j=1,\ldots,M_i$, $i=1,\ldots,N$.

We consider also that the random intercepts are given by:

$$\nu_{ik} = \mathbf{z'}_{ik}\boldsymbol{\gamma}_k + \eta_{ik} \qquad \textbf{(2)}$$

where $\mathbf{z}_{ik}$ is a vector of group level auxiliary variables used to explain the intercept for the *k*-th response variable for units in group *i*, $\boldsymbol{\gamma}_k$ is the vector of regression coefficients for the intercept of the *k*-th response variable, and η_{ik} is the group level random error for the intercept of the *k*-th response variable.

Here we assume that the vectors $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ip})'$ are independent and identically (IID) distributed as $MN(\mathbf{0};\ \boldsymbol{\Omega})$, where *MN* denotes the multivariate normal distribution. We also assume that the vectors $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \ldots, \varepsilon_{ijp})'$ are IID $MN(\mathbf{0};\ \boldsymbol{\Sigma})$, and are also independent of the $\boldsymbol{\eta}_i$.

The model defined by (1) and (2) plus the independence and distribution assumptions is called here the 'multivariate random intercept regression model' (MRIRM). It contains as unknown hyper-parameters the vectors of coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta'}_1, \ldots, \boldsymbol{\beta'}_p)'$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma'}_1, \ldots, \boldsymbol{\gamma'}_p)'$ and the variance matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$.

We further assume that the sample data are obtained by a two-stage sampling scheme. In the first stage, $n<N$ groups are selected with inclusion probabilities π_i that may be correlated with the random effects $\boldsymbol{\eta}_i$. In the second stage, m_i second-level units are sampled from group *i* selected in the first stage. Second stage sampling is carried out with conditional inclusion probabilities $\pi_{j|i} = pr(j \in s_i | i \in s)$, where s_i denotes the set of units sampled within group *i* and *s* denotes the set of groups selected

in the first stage. These conditional probabilities may be correlated with the outcomes $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijp})'$ even after conditioning on the regressors $\mathbf{x}_{ij}$.

To obtain the **sample model** corresponding to the MRIRM defined by (1) and (2), we first consider a model to incorporate informative sampling of the groups (first–level units). Hence we assume that the groups are sampled with probability proportional to size (PPS), with the sizes M_i satisfying:

$$[\log(M_i) \,|\boldsymbol{\nu_i}, \sigma_M^2\,] \sim N(\boldsymbol{t'}_i\boldsymbol{\theta} + \boldsymbol{\nu'_i}\boldsymbol{\alpha}; \sigma_M^2) \qquad \textbf{(3)}$$

where $\boldsymbol{t}_i$ is a q-vector of group level regressors for the group sizes, $\boldsymbol{\theta}$ is a q-vector of regression coefficients for the group sizes, $\boldsymbol{\nu_i} = (\nu_{i1}, \ldots, \nu_{ip})'$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$ is a vector of regression coefficients for the random intercepts in the model for the group sizes.

Note that equation (3) becomes also part of the population model. Following Pfeffermann et al (2006), the sample model defining the conditional distribution of the $\boldsymbol{\nu}_i$ given inclusion in the sample is:

$$f_s(\boldsymbol{\nu}_i \,|\boldsymbol{\mu}_i, \boldsymbol{\Omega}, \boldsymbol{\alpha}) = \frac{f_p(\boldsymbol{\nu}_i \,|\boldsymbol{\mu}_i, \boldsymbol{\Omega}) \mathrm{E}_p(\pi_i \,|\boldsymbol{\nu}_i)}{\mathrm{E}_p(\pi_i)} \qquad (4)$$

$$f_p(\boldsymbol{\nu}_i \,|\boldsymbol{\mu}_i, \boldsymbol{\Omega}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Omega}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\nu}_i - \boldsymbol{\mu}_i)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\nu}_i - \boldsymbol{\mu}_i)\right) \qquad (5)$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ip})'$, with $\mu_{ik} = \mathbf{z}'_{ik}\boldsymbol{\gamma}_k$ for k=1,…,p.

It is not difficult to show that:

$$\frac{\mathrm{E}_p(\pi_i \,|\boldsymbol{\nu}_i)}{\mathrm{E}_p(\pi_i)} \approx \exp\left((\boldsymbol{\nu}_i - \boldsymbol{\mu}_i)'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\,\boldsymbol{\Omega}\boldsymbol{\alpha})\right) \qquad (6)$$

Substituting (5) and (6) into (4), we have after some algebra:

$$f_s(\boldsymbol{\nu}_i \,|\boldsymbol{\mu}_i, \boldsymbol{\Omega}, \boldsymbol{\alpha}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Omega}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\nu}_i - \boldsymbol{\mu}_i - \boldsymbol{\Omega}\boldsymbol{\alpha})'\boldsymbol{\Omega}^{-1}(\boldsymbol{\nu}_i - \boldsymbol{\mu}_i - \boldsymbol{\Omega}\boldsymbol{\alpha})\right) \quad (7)$$

Therefore, the density distribution function in (7) is Multivariate Normal with mean $\boldsymbol{\mu}_i^s = \boldsymbol{\mu}_i + \boldsymbol{\Omega}\boldsymbol{\alpha}$ and variance-covariance matrix $\boldsymbol{\Omega}_s = \boldsymbol{\Omega}$. We also have $f_s(\log(M_i) \,|\boldsymbol{\nu}_i) = \mathrm{N}(\mathbf{t}'_i\boldsymbol{\theta} + \boldsymbol{\nu}'_i\boldsymbol{\alpha} + \sigma_M^2, \sigma_M^2)$.

For the special case with p=2, developed in the application, we have:

$$\boldsymbol{\mu}_i^s = \begin{pmatrix} \mu_{i1}^s \\ \mu_{i2}^s \\ \end{pmatrix} = \begin{pmatrix} \mathbf{z}'_{i1}\boldsymbol{\gamma}_1 + \alpha_1\sigma_{\nu 1}^2 + \alpha_2\sigma_{\nu 12} \\ \mathbf{z}'_{i2}\boldsymbol{\gamma}_2 + \alpha_2\sigma_{\nu 2}^2 + \alpha_1\sigma_{\nu 12} \\ \end{pmatrix}, \text{ where } \boldsymbol{\Omega} = \begin{bmatrix} \sigma_{\nu 1}^2 & \sigma_{\nu 12} \\ \sigma_{\nu 12} & \sigma_{\nu 2}^2 \end{bmatrix}.$$

Suppose that, within the selected groups, the units are sampled by disproportionate stratified sampling, with the stratum membership O_{ij} depended on the vector of outcome values $\mathrm{y_{ij}}$. For example, let $p_{ij} = b_0 + \mathbf{y}'_{ij}\mathbf{b}_1 + \varphi_{ij}$ with $\varphi_{ij} \sim \mathrm{N}(0; \sigma_\varphi^2)$ and independently distributed. Let $O_{ij} = 1$ (corresponding to stratum 1), if $p_{ij} < c_1$, and $O_{ij} = h$, if $c_{h\text{-}1} < p_{ij} < c_h$, for $h = 2, \ldots, H$.

The sample model defining the conditional distribution of $\mathbf{y}_{ij}$ given inclusion in the sample is:

$$f_s(\mathbf{y}_{ij} \mid \mathbf{x},\boldsymbol{\beta},\boldsymbol{\nu}_i,\boldsymbol{\Sigma},\psi) = \frac{f_p(\mathbf{y}_{ij} \mid \mathbf{x}_{ij},\boldsymbol{\beta},\boldsymbol{\nu}_i,\boldsymbol{\Sigma})\mathrm{E}_p(\pi_{j|i} \mid \mathbf{y}_{ij},\mathbf{x}_{ij},\boldsymbol{\beta},\boldsymbol{\nu}_i,\boldsymbol{\Sigma},\psi)}{\mathrm{E}_p(\pi_{j|i} \mid \mathbf{x}_{ij},\boldsymbol{\beta},\boldsymbol{\nu}_i,\boldsymbol{\Sigma},\psi)} \qquad (8)$$

where $\psi = (b_0, \mathbf{b}_1', \sigma_\varphi^2)$.

It can be shown that:

$$\begin{aligned}\mathrm{E}_p(\pi_{j|i} \mid \mathbf{y}_{ij},\mathbf{x}_{ij},\boldsymbol{\beta},\boldsymbol{\nu}_i,\boldsymbol{\Sigma},\psi) &= \sum_{h=1}^{H} \mathrm{f}_h^i \Pr(\mathrm{O}_{ij} = h \mid \mathbf{y}_{ij},\mathbf{x}_{ij},\boldsymbol{\beta},\boldsymbol{\nu}_i,\boldsymbol{\Sigma},\psi) \\ &= \mathrm{f}_1^i \mathrm{A}_1(\mathbf{y}_{ij}) + \sum_{h=2}^{H-1} \mathrm{f}_h^i (\mathrm{A}_h(\mathbf{y}_{ij}) - \mathrm{A}_{h-1}(\mathbf{y}_{ij})) + \mathrm{f}_H^i (1 - \mathrm{A}_H(\mathbf{y}_{ij}))\end{aligned} \qquad (9)$$

$$\begin{aligned}\mathrm{E}_p(\pi_{j|i} \mid \mathbf{x}_{ij},\boldsymbol{\beta},\boldsymbol{\nu}_i,\boldsymbol{\Sigma},\psi) &= \sum_{h=1}^{H} \mathrm{f}_h^i \Pr(\mathrm{O}_{ij} = h \mid \mathbf{x}_{ij},\boldsymbol{\beta},\boldsymbol{\nu}_i,\boldsymbol{\Sigma},\psi) \\ &= \mathrm{f}_1^i \mathrm{B}_1(\mathbf{y}_{ij}) + \sum_{h=2}^{H-1} \mathrm{f}_h^i (\mathrm{B}_h(\mathbf{y}_{ij}) - \mathrm{B}_{h-1}(\mathbf{y}_{ij})) + \mathrm{f}_H^i (1 - \mathrm{B}_H(\mathbf{y}_{ij}))\end{aligned} \qquad (10)$$

where f_h^i is the sampling fraction in stratum h, for h=1,…,H, of group i;

$$\mathrm{A}_h(\mathbf{y}_{ij}) = \Pr(\mathrm{p}_{ij} < \mathrm{c}_h \mid \mathbf{y}_{ij}) = \Phi\left(\frac{\mathrm{c}_h - \mathrm{b}_0 - \mathbf{b}_1'\mathbf{y}_{ij}}{\sigma_\varphi}\right);$$

$$\mathrm{B}_h(\mathbf{y}_{ij}) = \Phi\left(\frac{\mathrm{c}_h - \mathrm{b}_0 - \mathbf{b}_1'\,\boldsymbol{\mu}(\mathbf{y}_{ij})}{\sigma_\varphi\sqrt{\left(1 + \mathbf{b}_1'\boldsymbol{\Sigma}\mathrm{b}_1\right)}}\right), \quad \boldsymbol{\mu}(\mathbf{y}_{ij}) = (\mu(y_{ij1}),...,\mu(y_{ij1}))', \quad \text{with}$$

$\mu(y_{ijk}) = \mathbf{x}_{ijk}'\boldsymbol{\beta}_k + \boldsymbol{\nu}_{ik}$, for k=1,…,p.

3. Illustration

Multilevel models have been frequently used in educational assessment studies to model the proficiency scores of students as a function of student and class or school regressors. Most of these studies model a single proficiency score as the response, using the univariate multilevel model.

Here an application of the MRIRM is presented for modelling jointly Mathematics and Portuguese Language proficiency scores obtained from a Brazilian evaluation study of basic education conducted by the Brazilian National Institute of Education Research (INEP). The scores stem from applying Item Response Theory models to test results from the 'Prova Brasil 2009' study. A two-level multivariate hierarchical normal model is fitted, where the students and schools are the (first level) units and the groups (second level units) respectively. The analysis is restricted to the performance of students from the 8th grade in elementary schools from Rio de Janeiro municipality. There were 34.867 students with complete proficiency and predictor variables for this grade in the city, distributed in 440 schools.

'Prova Brasil 2009' obtained Mathematics (y_1) and Portuguese Language (y_2) proficiency scores for all students in the 8th grade in elementary schools who attended the exam. Therefore it is not a sample survey. However, to illustrate the potential effects of informative sampling when fitting the model, we considered an informative

sampling design that samples 50 schools using a PPS design, with the number of tested pupils in the school as the size measure, and then sampled 10 students within each school by simple random sampling, thus leading to samples of 500 students. The design may therefore be informative at the school (first stage) level.

We fitted the target model to the population data, where 47 main effect predictors were considered for the initial model fitting. We then performed model selection by backward elimination, and retained a model with only 8 predictors. Samples of 500 students were then selected using the specified design, and the same main effects model with 8 predictors is fitted to each sample. Simulation estimates of the Bias and Root Mean Square Error of the population model parameters are computed from deviations of the sample model estimates to the estimates obtained fitting the model to the full population dataset.

4. References

Asparouhov, T. (2006). General Multi-Level Modeling with Sampling Weights. *Communications in Statistics - Theory and Methods*, v. 35, n. 3, p. 439-460.

Grilli, L.; Pratesi, M. (2004) Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. Survey Methodology, v. 30, n. 1, p. 93-103.

Kovačević, M. S.; Rai, S. N. (2003). A Pseudo Maximum Likelihood Approach to Multilevel Modelling of Survey Data. *Communications in Statistics - Theory and Methods*, v. 32, n. 1, p. 103-121.

Krieger, A. M.; Pfeffermann, D. (1997). Testing of distribution functions from complex surveys. *Journal of Official Statistics*, v. 13, n. 2, p. 123-142.

Pfeffermann, D.; Moura, F. A. S.; Silva, P. L. N. Multi-level modelling under informative sampling. *Biometrika*, v. 93, n. 4, p. 943-959, 2006.

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

Pfeffermann, D.; Sverchkov, M. (2003) Fitting Generalized Linear Models under Informative Sampling. In Chambers, R. L. and Skinner, C. J., *Analysis of Survey Data*, John Wiley & Sons, Ltd.

Pfeffermann, D., & Sverchkov, M. (2007). Small area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, Volume 102, Issue 480, 1427-1439.

Rabe-Hesketh, S.; Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, v. 169, n. 4, p. 805-827.

Sverchkov, M.; Pfeffermann, D. (2003). Prediction of finite population totals based on the sample distribution. Southampton: Methodology Working Paper M03/06.

Verret, F.; Hidiroglou, M.A. & Rao, J.N.K (2010). Small Area Estimation Under Informative Sampling. Proceedings of the Survey Methods Section, SSC Annual Meeting, May 2010.

You Y. and Rao, J.N.K. (2002). A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights. *The Canadian Journal of Statistics*, Vol. 30, No. 3, 431-439.