A PSEUDO MAXIMUM LIKELIHOOD APPROACH TO INFERENCE ON HIERARCHICALLY STRUCTURED DATA

Milorad S. Kovačević and Shesh N. Rai, Statistics Canada
Shesh Rai, Statistics Canada, 15-E, R.H.Coats Bldg., Ottawa, ON K1A 0T6

Key Words: Linear multi-level model; Two-stage sampling; Variance estimation; Weighting

1. INTRODUCTION

Populations studied in social research, public health, environmental or educational research are usually hierarchical with easily recognizable levels and nested structures. Different types of variables are available at different levels. For example, at the group level usually there are group identifiers, aggregates of lower level unit variables (means, totals, counts, percentages, etc.), and the global variables for the groups. Some data may come from a survey, some, especially for higher level units, may come from a census or administrative files. Variables that are available as aggregates at the group level may not be available at the unit level, etc.

By disaggregation of all higher order variables to the individual level one can ignore the hierarchical structure and analyze data using simpler statistical techniques assuming independence of the observations. On the other hand, if all the individual level variables are aggregated to the higher level, one can analyze data on the higher level. In the first scenario, if the data structure is hierarchical, the observations within the groups are correlated; and therefore, the assumption of independence of observations is untenable. In the second scenario, important information is lost, and an interpretation of the results of aggregate analysis at the individual level is usually fallacious. Thus, aggregating and disaggregating may not be completely satisfactory for the analysis of hierarchically structured data.

The appropriate modelling combines the different levels of the hierarchical data in the form of hierarchical models. The main interest is to model the relationships at the unit level taking into account the impact of higher level units on these relationships. For an excellent presentation of hierarchical models, known also as multilevel statistical models, the reader is referred to Bryk and Raudenbush (1992) and Goldstien (1995).

A motivating example considers the data from Cycle 1 (1994-1995) of the Canadian National Longitudinal Survey of Children and Youth (NLSCY) – an initiative to develop a national database on the characteristics and life experiences of children and youth in Canada. The target population is children aged 0-11 years living in households across Canada. Children were identified using a stratified, multistage probability sample design based on area frames in which dwellings (residences) are the ultimate sampling units. As a consequence, the data set is inherently hierarchical: children are nested within families and families are nested within geographical areas or places. In a multilevel study of the neighbourhood influences on children behavior, Boyle and Lipman (1998) considered at the individual level the following dependent variables: score measures of conduct problems, hyperactivity and emotional problems, then the independent variables: age, sex, and school attendance. At the family level (level-2) the independent variables are family type and a variety of socio-economic measures for families. At the geographic level (level-3) the independent variables are taken from the 1996 Census such as the percentage of families led by one parent, the percentage of families below the poverty line, urban/rural type of the area, etc.

When data come from surveys the estimation of the model parameters has to take into account the sampling design used for selecting the respondents. Recently, Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998) addressed the problem of weighting in the multilevel models using the probability weighted iterative generalised least squares method.

The goal of this paper is to show how to incorporate the design information into the inference about the model parameters when modelling a finite hierarchical population. A method that we are proposing relies on ideas of pseudo maximum likelihood estimation (Gourieroux, Monfort, Trognon, 1984) to provide the finite population estimating equations which are then estimated using an available hierarchical (multi-stage) sample. These estimated equations lead to the consistent estimates of the model parameters under very general conditions as in Binder (1983). The proposed method seems to be simpler than the probability weighted iterative generalised least squares method considered by Pfeffermann et al. (1998). Some other sampling considerations are also tackled in the paper: how to approximate the weights for units at different levels in hierarchy when only a limited information on design is available, and how to provide the weights for the higher level units when they were not the design units.

The second section contains the basic theory of hierarchical linear modelling. Section 3 shows how the model parameters can be defined as finite population parameters. A proposed method for estimation of the variance is given in this section. In section 4 the finite population parameters defined in section 3 are estimated using data from a complex survey. A small simulation study was used to empirically confirm the consistency of the resulting estimates under several realistic scenarios. Section 5 deals with issues of necessity and availability of the weights for different model levels. Section 6 contains a summary of the proposed method.

2. A TYPICAL MULTI-LEVEL MODEL

We begin this section with a description of a simple linear two-level model which can be specified with two equations. The first one is a "within-group" equation and is designed to describe the relationship between unit-level dependent variables and the unit-level covariates within each group. Some or all of the parameters of the "within-group" equation are viewed as varying randomly across the group-level population. Then in the second equation, "between-group" equation, these parameters are modelled as dependent variables in a model with the group-level variables as covariates.

Let y_{gi} be the value of a dependent variable for individual i ($i = 1,...,N_g$) in group g ($g = 1,...,G$), and let there be $P+Q$ independent variables, x_{pgi}, z_{qgi}, $p = 1,...,P$ and $q = 1,...,Q$ that describe an individual. Then, the unit level within-group regression equation is

$$y_{gi} = b_{0g} + \sum_p \beta_p x_{pgi} + \sum_q b_{qg} z_{qgi} + e_{gi} \qquad (1)$$

where β_p are fixed regression coefficients, b_{qg} are within-group regression coefficients that vary across the groups, and e_{gi} are the random disturbances independent from b_{qg}. A more convenient matrix expression is

$$\boldsymbol{y}_g = \boldsymbol{X}_g \boldsymbol{\beta} + \boldsymbol{Z}_g \boldsymbol{b}_g + \boldsymbol{e}_g \qquad (2)$$

Here, $\boldsymbol{y}_g$ is $N_g \times 1$ vector of dependent variable, the parameter vectors are column vectors, and the covariates are given as matrices, $N_g \times P$ and $N_g \times (Q+1)$, respectively. The random intercept b_{0g} is a part of the random vector $\boldsymbol{b}_g$ assuming that the first column of the $\boldsymbol{Z}_g$ is a vector of 1's, **1**.

The group level regression equation relates the random within-group coefficients, b_{qg} to group-level characteristics, u_{rg}, $r = 1,...,R$ and $g = 1,...,G$:

$$b_{qg} = \gamma_{q0} + \sum_{r=1}^{R} \gamma_{qr} u_{rg} + d_{qg}, \quad q = 0,...,Q \qquad (3)$$

The d_{qg} are group level disturbances independent from e_{gi} and represent the contribution of each group that remains unexplained by model (3). Written in a matrix form, equation (3) is

$$\boldsymbol{b}_g = \boldsymbol{F}_g \gamma + \boldsymbol{d}_g \qquad (4)$$

where $\boldsymbol{b}_g$ is a Q+1 by 1 vector, $\boldsymbol{F}_g$ is a Q+1 by R(Q+1) matrix obtained as a direct product $\boldsymbol{u}_g \otimes \boldsymbol{I}_{Q+1}$, $\boldsymbol{u}_g$ is a row vector of length R, γ is a R(Q+1) vector of the unknown but fixed parameters, and $\boldsymbol{d}_g$ is a vector of group random effects.

The standard assumptions about the disturbances apply at both levels: $E(\boldsymbol{e}_g) = \boldsymbol{0}$, $E(\boldsymbol{d}_g) = \boldsymbol{0}$, i.e., the disturbances are centered at 0, the within group variability is expressed by $\sigma^2_{(1)}$ and is constant across the population of groups, and the variance of $\boldsymbol{d}_g$ is captured by $\Sigma_{(2)}$, the $(Q+1)$ by $(Q+1)$ covariance matrix at the group level, and the level disturbances are not correlated with each other.

If there is no covariate available other than a group identifier, model ((1), (3)) reduces to one-way ANOVA model with random effects:

$$y_{gi} = b_{0g} + e_{gi} \qquad (6)$$

$$b_{0g} = \gamma_{00} + d_g \qquad (7)$$

or written together

$$y_{gi} = \gamma_{00} + d_g + e_{gi} \qquad (8)$$

Here $\gamma_{00}$is an unknown fixed grand mean, d_g is a *g-th* group effect ~ (0, $\sigma^2_{(2)}$) , and e_{gi} is an individual effect ~(0, $\sigma^2_{(1)}$).

The generalization of two-level model ((1), (3)) to fit a multi-level hierarchy is straight-forward.

In the motivating example the family level is critical for estimation of the residual parameters due to a small number of children per family, frequently only one. Because of that it is reasonable to express the family level variables as the individual characteristics with an extra variable introduced to indicate if there are other individuals in population that share the same family characteristics. Ignoring completely the family level, the family clustering effect may cause some of the coefficients to appear more significant than they actually are.

3. FINITE POPULATION ESTIMATING EQUATIONS FOR MODEL PARAMETERS

In this section we define the model parameters as functions of the finite population data.

Equations (2) and (4) are written jointly so that a two-level model is expressed by one equation

$$\begin{aligned} \boldsymbol{y}_g &= \boldsymbol{X}_g \boldsymbol{\beta} + \boldsymbol{Z}_g \boldsymbol{F}_g \boldsymbol{\gamma} + \boldsymbol{Z}_g \boldsymbol{d}_g + \boldsymbol{e}_g \\ &= (\boldsymbol{X}_g \mid \boldsymbol{Z}_g \boldsymbol{F}_g) \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} + \boldsymbol{Z}_g \boldsymbol{d}_g + \boldsymbol{e}_g \\ &= \mathbf{H}_g \boldsymbol{\eta} + \boldsymbol{a}_g \end{aligned} \tag{12}$$

where $\mathbf{H}_g = (\boldsymbol{X}_g \mid \boldsymbol{Z}_g \boldsymbol{F}_g)$ is a known N_g by $P+(Q+1)R$ matrix of observed covariates and their products at both levels, $\boldsymbol{\eta}$ is a $P+(Q+1)R$ vector of unknown fixed effects, and $\boldsymbol{a}_g$ is an N_g by 1 vector of random effects with $a_{gi} = z_{gi} \boldsymbol{d}_g + e_{gi}$. Here z_{gi} represents a row vector of values of z variables for the ith individual in the gth group. Evidently, $E(\boldsymbol{a}_g) = \mathbf{0}$, and

$$Var(\boldsymbol{a}_g) = \boldsymbol{V}_g = \boldsymbol{Z}_g \boldsymbol{\Sigma}_{(2)} \boldsymbol{Z}_g^{\prime} + \sigma_{(1)}^2 \boldsymbol{I}_g. \tag{13}$$

We assume that there is a single parameter $\sigma_{(1)}^2$ that describes the variability between level 1 units, and that there are (Q+1)(Q+2)/2 unknown parameters in the covariance matrix $\boldsymbol{\Sigma}_{(2)}$.

Stacking of the G vectors $\boldsymbol{y}_g$ into a block vector $\mathbf{y}^{\prime} = [y_1^{\prime}, \dots, y_G^{\prime}]$, then creating a block matrix $\mathbf{H} = [\boldsymbol{H}_1^{\prime} \mid \dots \mid \boldsymbol{H}_G^{\prime}]^{\prime}$, and stacking of the G vectors $\boldsymbol{a}_g$ into a block vector $\mathbf{a}^{\prime} = [\boldsymbol{a}_1^{\prime}, \dots, \boldsymbol{a}_G^{\prime}]$, equation (12) can be written for all levels jointly as

$$\mathbf{y} = \mathbf{H} \boldsymbol{\eta} + \mathbf{a} \tag{14}$$

where $\mathbf{a}$ is an N by 1 vector of random errors, assumed to be centered at $\mathbf{0}$ and with a covariance matrix $\mathbf{V} = Var(\mathbf{a})$. While matrix $\mathbf{H}$ represents total information available on covariates in the population, matrix $\mathbf{V}$ represents the complete correlation structure of the hierarchical population under study. For the population of groups it is reasonable to assume that $\mathbf{V}$ is a block diagonal matrix with the blocks defined by (13), and $Cov(\boldsymbol{a}_g, \boldsymbol{a}_{g^{\prime}}) = 0$, for $g \neq g^{\prime}$.

The unknown finite population parameters $\boldsymbol{\eta}$, $\boldsymbol{\Sigma}_{(2)}$ and $\sigma_{(1)}^2$ have to be expressed as functions of finite population data. Assuming that $\mathbf{V}$ is known, using the method of generalized least squares (GLS), the unknown $\boldsymbol{\eta}$ can be expressed as:

$$\begin{aligned} \hat{\boldsymbol{\eta}}_{GLS} &= [\mathbf{H}^{\prime} \mathbf{V}^{-1} \mathbf{H}]^{-1} \mathbf{H}^{\prime} \mathbf{V}^{-1} \mathbf{y} \\ &= \left(\sum_g \boldsymbol{H}_g^{\prime} \boldsymbol{V}_g^{-1} \boldsymbol{H}_g \right)^{-1} \sum_g \boldsymbol{H}_g^{\prime} \boldsymbol{V}_g^{-1} \boldsymbol{y}_g \end{aligned} \tag{15}$$

The covariance matrix of $\hat{\boldsymbol{\eta}}_{GLS}$ is given by

$$\begin{aligned} Var(\hat{\boldsymbol{\eta}}_{GLS}) &= [\mathbf{H}^{\prime} \mathbf{V}^{-1} \mathbf{H}]^{-1} \\ &= \left(\sum_g \boldsymbol{H}_g^{\prime} \boldsymbol{V}_g^{-1} \boldsymbol{H}_g \right)^{-1} \end{aligned} \tag{16}$$

Estimator (15) coincides with the maximum likelihood (ML) estimators under the assumption of normality of the vector $\mathbf{y}$, $\mathbf{y} \sim MVN(\mathbf{H}\boldsymbol{\eta}, \mathbf{V})$, and assuming that $\mathbf{V}$ is a known block-diagonal matrix.

Since $\mathbf{V}$ is not known and has to be estimated, a procedure like the iterative generalised least squares where one iterates between estimating $\boldsymbol{\eta}$ and $\mathbf{V}$ until a convergence criteria is met, is usually used. The problem with such method is in computational intensity due to the number of parameters that need to be estimated in an iterative procedure. A good review of the method and its application is given in Goldstein (1995). Also a weighted version of the method is examined by Pfeffermann et al (1998).

Here we suggest a pseudo maximum likelihood estimation of $\boldsymbol{\eta}$ by replacing parameter $\mathbf{V}$ by its estimate $\tilde{\mathbf{V}}$ (obtained elsewhere) in the likelihood equation and then solving it, so that

$$\begin{aligned} \hat{\boldsymbol{\eta}}_{PML} &= [\mathbf{H}^{\prime} \tilde{\mathbf{V}}^{-1} \mathbf{H}]^{-1} \mathbf{H}^{\prime} \tilde{\mathbf{V}}^{-1} \mathbf{y} \\ &= \left(\sum_g \boldsymbol{H}_g^{\prime} \tilde{\boldsymbol{V}}_g^{-1} \boldsymbol{H}_g \right)^{-1} \sum_g \boldsymbol{H}_g^{\prime} \tilde{\boldsymbol{V}}_g^{-1} \boldsymbol{y}_g \end{aligned} \tag{17}$$

with the corresponding covariance

$$\begin{aligned} Var(\hat{\boldsymbol{\eta}}_{PML}) &= [\mathbf{H}^{\prime} \tilde{\mathbf{V}}^{-1} \mathbf{H}]^{-1} \mathbf{H}^{\prime} \tilde{\mathbf{V}}^{-1} \mathbf{V} \tilde{\mathbf{V}}^{-1} \mathbf{H} [\mathbf{H}^{\prime} \tilde{\mathbf{V}}^{-1} \mathbf{H}]^{-1} \\ &= \left(\sum_g \boldsymbol{H}_g^{\prime} \tilde{\boldsymbol{V}}_g^{-1} \boldsymbol{H}_g \right)^{-1} \left(\sum_g \boldsymbol{H}_g^{\prime} \tilde{\boldsymbol{V}}_g^{-1} \boldsymbol{V}_g \tilde{\boldsymbol{V}}_g^{-1} \boldsymbol{H}_g \right) \cdot \\ &\quad \left(\sum_g \boldsymbol{H}_g^{\prime} \tilde{\boldsymbol{V}}_g^{-1} \boldsymbol{H}_g \right)^{-1} \end{aligned} \tag{18}$$

Proposed Estimation of V

Equation (12) can be rewritten in a way that combines fixed and random parameters in the same vector

$$\begin{aligned} \boldsymbol{y}_g &= \boldsymbol{H}_g \boldsymbol{\eta} + \boldsymbol{Z}_g \boldsymbol{d}_g + \boldsymbol{e}_g. \\ &= (\boldsymbol{H}_g \mid \boldsymbol{Z}_g) \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{d}_g \end{pmatrix} + \boldsymbol{e}_g \\ &= \boldsymbol{K}_g \boldsymbol{\xi}_g + \boldsymbol{e}_g \end{aligned} \tag{19}$$

The unknown vector $\xi_g = (\eta' \mid d'_g)'$ is random since one of its parts, d_g, varies across the groups. However the size of the vector remains fixed $P+(Q+1)R+(Q+1)$ over all groups.

Note that assuming that ξ_g is fixed, its ML estimate is

$$\hat{\xi}_g = \left(K'_g K_g\right)^{-1} K'_g y_g,\ g=1,\ldots, G, \qquad (20)$$

due to a constant variance $Var(e_g) = \sigma^2_{(1)} I_g$.

The variance V_g, given by (13), can be expressed as

$$\begin{aligned} V_g &= Var(y_g) \\ &= E_{\xi_g}\, Var(y_g \mid \xi_g) + Var_{\xi_g}\, E(y_g \mid \xi_g) \end{aligned} \qquad (21)$$

where $Var(y_g \mid \xi_g) = \sigma^2_{(1)} I_g$ and $E(y_g \mid \xi_g) = K_g \xi_g$

Then V_g can be unbiasedly estimated as

$$\tilde{V}_g = \hat{\sigma}^2_{(1)} I_g + K_g\, V\hat{a}r(\hat{\xi}_g)\, K'_g \qquad (22)$$

where

$$\hat{\sigma}^2_{(1)} = \frac{1}{G} \sum_g \mathbf{1}'_g (y_g - K_g \hat{\xi}_g)(y_g - K_g \hat{\xi}_g)' \mathbf{1}_g / (N_g - 1) \qquad (23)$$

and

$$V\hat{a}r(\hat{\xi}_g) = \frac{1}{G-1} \sum_g (\hat{\xi}_g - \frac{1}{G} \sum_g \hat{\xi}_g)\,(\hat{\xi}_g - \frac{1}{G} \sum_g \hat{\xi}_g)' \qquad (24)$$

Note that $K_g\, V\hat{a}r(\hat{\xi}_g)\, K'_g$ reduces to $Z_g\, V\hat{a}r(\hat{d}_g)\, Z'_g$, where $V\hat{a}r(\hat{d}_g)$ is obtained from (24) using the appropriate part of $\hat{\xi}_g$ vector.

4. ESTIMATION BASED ON COMPLEX SAMPLES

If we observe the complete populations of individuals and groups the estimates (17) and (22) are the finite population values of the model parameters. The variance (18) can be treated as a finite population parameter as well. Having only observed a sample taken from the finite population we need to estimate these parameters. Here we present the estimation based on a complex sample.

Without loss of generality, we assume a simple scenario where the sampling design hierarchy is the same as the model hierarchy, meaning that the groups (level 2 units) are the primary sampling units and that the individuals (level 1 units) are the second stage units.

Let a sample of m out of G groups be selected, and let from gth selected group a sample of n_g out of N_g individuals be selected. Also, we assume that the final individual weight w_{gi} is a product of the known components: the group weight w_g and the conditional individual weight $w_{i|g}$, thus $w_{gi} = w_g\ w_{i|g}$. The weights satisfy the usual unbiasedness criteria:

$$E\left(\sum_{g=1}^{m} \sum_{i=1}^{n_g} w_{gi} \right) = N,\ E\left(\sum_{g=1}^{m} w_g \right) = G, \text{and } E\left(\sum_{i=1}^{n_g} w_{i|g} \right) = N_g \qquad (25)$$

Let $W_{\cdot|g}$ be a diagonal matrix of order $n_g \times n_g$ with the conditional weights $w_{i|g}$ on the diagonal. Then the sample based estimate of the vector $\hat{\xi}_g$ (20) is

$$\breve{\xi}_g = \left(K'_g W_{\cdot|g} K_g\right)^{-1} K'_g W_{\cdot|g}\, y_g, \qquad (26)$$

where K_g is a known matrix of size $n_g \times [P+(Q+1)R+(Q+1)]$ and y_g is a vector of size n_g. To estimate the variance component $\hat{\sigma}^2_1$, given by (23), which has the form of the population mean of the values $\mathbf{1}'_g (y_g - K_g \hat{\xi})(y_g - K_g \hat{\xi})' \mathbf{1}_g / (N_g - 1)$, we use the sample mean

$$\breve{\sigma}^2_1 = \frac{1}{\sum_g w_g} \sum_{g=1}^{m} \frac{w_g}{(\sum_i w_{i|g} - 1)} \mathbf{1}'_g (y_g - K_g \breve{\xi}_g)\,(y_g - K_g \breve{\xi}_g)' \mathbf{1}_g \qquad (27)$$

Variance (24) is estimated by appropriate weighting as

$$V\breve{a}r(\breve{\xi}_g) = \frac{1}{\sum w_g - 1} \sum_g w_g (\breve{\xi}_g - \breve{\xi})\,(\breve{\xi}_g - \breve{\xi})' \qquad (28)$$

where $\breve{\xi} = \sum_g w_g \breve{\xi}_g \ / \ \sum_g w_g$.

The matrix of the random components is estimated by

$$\breve{V}_g = \breve{\sigma}^2_1 I_g + K_g\, V\breve{a}r(\breve{\xi}_g)\, K'_g \qquad (29)$$

The finite population parameter (17) is estimated by

$$\breve{\eta} = \left(\sum_g w_g H'_g \breve{V}^{-1}_g H_g \right)^{-1} \sum_g w_g H'_g \breve{V}^{-1}_g y_g, \qquad (30)$$

and its variance is simply estimated by

$$\breve{V}(\breve{\eta}) = \left(\sum_g w_g H'_g \breve{V}^{-1}_g H_g \right)^{-1} \left(\sum_g w^2_g H'_g \breve{V}^{-1}_g H'_g \right). \left(\sum_g w_g H'_g \breve{V}^{-1}_g H_g \right)^{-1}$$

(31)

Simulation Study

In this subsection, we present the results of a limited simulation study on the finite sample properties of the proposed method for inference about the model parameters. We consider the following simple model for simulation study

$$y_{gi} = b_{0g} + e_{gi}$$

where $$b_{0g} = \eta + d_g.$$

The first-level error distribution is assumed to be normal with mean zero and standard deviation $\sigma_1 = 1$. The error associated with level-2 is also assumed to be normally distributed with mean zero and five known values of standard deviation σ_2 such that the variance ratioσ_1^2/σ_2^2 takes values $\{0.1, 0.2, 0.5, 1, 2\}$. Their values reflect the respective within and between groups variabilities. Without loss of generality, we assigned 0 value to η in generating y_{gi} . For three values of sample sizes for groups and three ranges of sample sizes in each group, the data are generated. For each set of distributions and each combination of number of groups and group sample sizes, we generate 1000 independent sample runs of y_{gi}. The mean and variance of y_{gi}, η, $\sigma^2 = \sigma_1^2 + \sigma_2^2$, are estimated from each generated sample. The Monte Carlo standard errors of $\hat{\eta}$ and $\hat{\sigma}^2$ are calculated based on these 1000 samples. Results are summarized in Table 1.

The simulation study confirmed that the applied estimation resulted in a negligible relative bias for both $\hat{\eta}$ and $\hat{\sigma}^2$, and for all combinations of sample sizes and ratios of "between" and "within" variances. It is also evident that stability of the variance estimate $\hat{\sigma}^2$ depends mostly on the number of groups in the sample and very little on the size of the subsample of level-1 units. This is a typical situation in a survey. Stability of the variance estimate also depends on the ratio σ_1^2/σ_2^2. The smaller variability between groups in comparison with the within variability leads to a more stable variance estimate. The most important finding is that the sizes of the group subsamples n_g have very small impact on the stability of the variance estimates.

5. SOME SAMPLING CONSIDERATIONS

Design and model hierarchies are the same

Analysts have usually access to the final weights w_{gi}, $g=1,2,\cdots,m$; $i \in s_g$ where m is the number of groups (PSU's) selected from total G groups and s_g is the collection of units (n_g) selected from the gth group . Also they know the total number of groups (PSU's) G, the number of selected groups m, and the number of selected individuals from the selected groups n_g, $n = \Sigma n_g$. Usually, the group weights (w_g) and the conditional weights ($w_{i|g}$) are not readily available to analysts, but are needed for analysis. Based on the available information mentioned above, one can approximate the weights $w_{i|g}$ and $w_g$by $\hat{w}_{i|g}$and$\hat{w}_g$, respectively, so that

$$\sum_{g=1}^{m} \hat{w}_g \approx G, \; \sum_{i=1}^{n_g} \hat{w}_{i|g} \approx N_g \text{ and } \hat{w}_g \, \hat{w}_{i|g} \approx w_{gi} \quad (32)$$

Details on this will be given in a subsequent paper.

Design and model hierarchies are different

So far we assumed that the sampling design hierarchy is the same as the model hierarchy meaning that the level 2 units are the primary sampling units and the level 1 are the second stage units. See Fig.1a. When the multilevel structure of the model is different from the hierarchy used in sampling we suggest a conditional "retrospective sampling" approach. Conditioning is done according to the realized sample sizes. The retrospective sampling that we are considering using the ideas of Neuhaus and Jewell (1990), makes the selection of a model group dependent on the realization of the sample obtained by the applied sampling design. Consequently, the retrospective probability of selecting the model group becomes the function of the probability of selection of the design groups. Details will be given in a subsequent paper.

6. SUMMARY

In this paper we showed how to model a hierarchical data set coming from a finite population.

When population is hierarchical it can hardly be seen as an iid sample from the universe due to the intraclass correlations found within the groups and because of between groups variability. Consequently when finite population parameters are defined as ML estimates, the covariance structure of the finite population has to be accounted for, and since it is unknown it has to be estimated using the same data.

Here we used the method of the pseudo ML to define the finite population parameters of the hierarchical model. It is pseudo because we used an estimate of the variance obtained outside of the ML estimation process. The resulting estimates have ML estimates properties since the variance is estimated unbiasedly, meaning that the finite population parameters are well defined. For a given sample from the finite population we showed how to obtain the consistent estimates and calculate their standard errors. A small simulation study showed that even small subsamples from the groups give the stable variance estimates. Also a problem of obtaining appropriate weights for the different levels of the hierarchy is pointed out.

REFERENCES

Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review,* 51, 279-292.

Boyle, M.H. and Lipman, E.L. (1998). *Do Places Matter? A Multilevel Analysis of Geographical Variations in Child Behavior in Canada.* Human ResourcesDevelopment Canada. Applied Research Branch Working Paper, **W-98-16E**

Bryk, A. S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Sage Publications, Newbury Park.

Gourieroux, C., Monfort, A., Trognon, A. (1984). Pseudo Maximum Likelihood Methods: Theory. *Econometrica, Vol.* 52, 681-700.

Goldstein, H. (1995). *Multilevel Statistical Models.* Second edition. Edward Arnold, London.

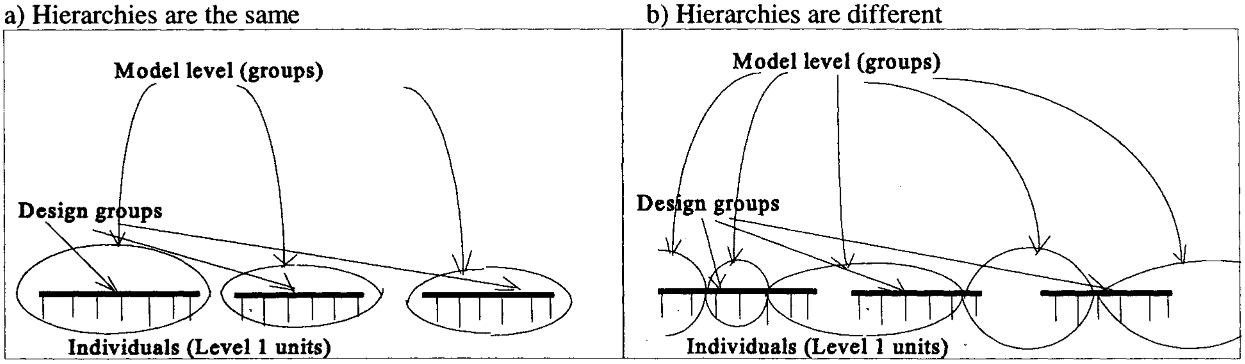
Neuhaus, J.M. and Jewell, N.P. (1990). The effect of retrospective sampling on binary regression models for clustered data. *Biometrics*, 46, 977-900.

Pfefferman, D., Skinner, C.J., Holmes, D.J., Goldstein, H., and Rasbash J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models. *Journal of Royal Statistical Society*, B, 60, 23-40.

Table 1. Results of the simulation study averaged over 1000 simulations and multiplied by 100. Standard errors are the Monte-Carlo standard errors.

		m=5				m=50				m=200			
n_g	σ_1^2/σ_2^2	$\hat{\eta}$	$se(\hat{\eta})$	$\hat{\sigma}^2$	$se(\hat{\sigma}^2)$	$\hat{\eta}$	$se(\hat{\eta})$	$\hat{\sigma}^2$	$se(\hat{\sigma}^2)$	$\hat{\eta}$	$se(\hat{\eta})$	$\hat{\sigma}^2$	$se(\hat{\sigma}^2)$
All large [50-100]	0.1	3.0	4.5	1110.4	21.7	0.3	1.4	1108.0	6.4	-0.6	0.7	1100.8	3.1
	0.2	1.7	3.2	602.4	11.7	0.8	1.0	599.3	3.2	0.2	0.5	602.0	1.6
	0.5	-1.8	2.0	301.7	4.4	0.2	0.6	301.6	1.3	0.3	0.3	301.6	0.6
	1	2.8	1.5	203.8	2.3	-0.1	0.4	202.3	0.7	-0.2	0.2	202.2	0.3
	2	2.2	1.0	150.3	1.2	0.3	0.3	151.2	0.3	0.3	0.2	151.4	0.2
Some small [5-100]	0.1	8.0	5.0	1087.0	22.0	0.3	1.6	1105.9	6.4	-0.7	0.8	1103.6	3.2
	0.2	1.0	4.0	599.0	12.0	1.1	1.1	608.6	3.2	-0.6	0.6	603.2	1.6
	0.5	2.0	2.0	298.0	5.0	0.1	0.7	303.2	1.3	0.7	0.4	304.6	0.6
	1	0.0	2.0	202.0	2.0	0.0	0.5	202.2	0.7	0.2	0.3	202.5	0.3
	2	0.0	1.0	154.0	1.0	0.1	0.4	152.2	0.3	0.1	0.2	152.9	0.2
All small [5-10]	0.1	2.9	4.6	1137.9	23.3	2.4	1.5	1117.7	6.6	-1.6	0.7	1114.8	3.3
	0.2	3.1	3.2	619.9	12.1	0.1	1.0	618.4	3.3	0.0	0.5	612.2	1.6
	0.5	-3.5	2.0	307.2	4.6	0.3	0.7	310.9	1.4	-0.5	0.3	313.8	0.7
	1	0.1	1.5	210.4	2.6	0.2	0.5	213.3	0.8	-0.1	0.2	212.8	0.4
	2	0.8	1.1	161.3	1.6	-0.1	0.4	162.0	0.5	0.0	0.2	163.6	0.2

Figure 1. A two-level model of a two-stage sample


a) Hierarchies are the same

b) Hierarchies are different