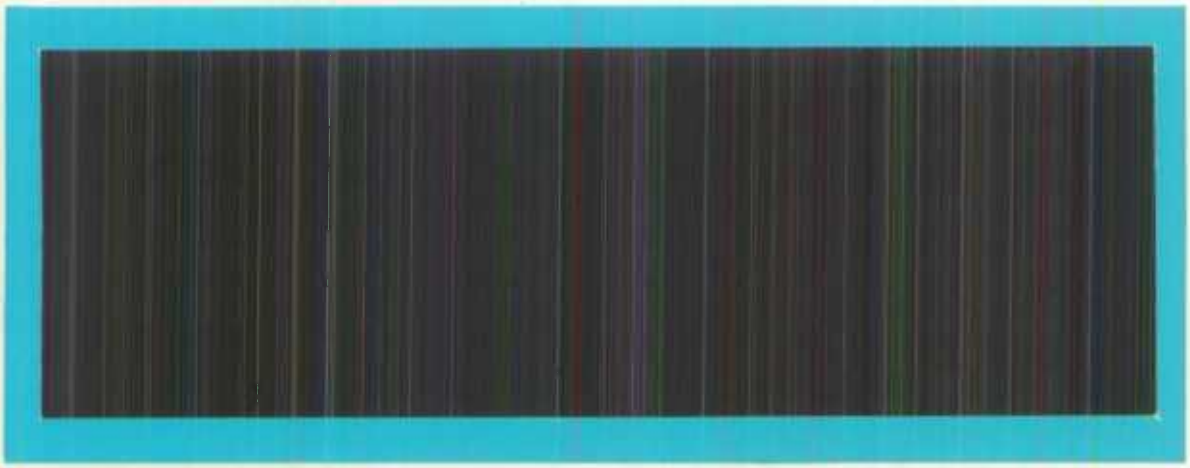
Statistics Canada Statistique Canada



Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Canada

WORKING PAPER
METHODOLOGY BRANCH

A PSEUDO MAXIMUM LIKELIHOOD APPROACH TO INFERENCE ABOUT HIERARCHICALLY STRUCTURED DATA

SSMD - 2000-004E

Milorad S. Kovacevic and Shesh N. Rai

April 2000

A PSEUDO MAXIMUM LIKELIHOOD APPROACH TO INFERENCE ABOUT HIERARCHICALLY STRUCTURED DATA

Milorad S. Kovacevic
and
Shesh N. Rai

Summary

An application of the pseudo maximum likelihood method is demonstrated on estimation for a mixed linear model fitted to the dependent observations coming from a hierarchical population. This approach provides a closed form solution for estimating the parameters of the mixed linear models which seems to be simpler than the iterative procedures such as iterative probability weighted least squares method of Pfeffermann *et al.* (1998) . We also discuss some issues relating to model and sample design hierarchies and their impact on estimation. A small simulation study showed that the proposed procedure is efficient even for small sample sizes at higher levels.

Key words: finite population sampling, hierarchical linear modelling, variance estimation

---

Milorad Kovacenic, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6, e-mail: kovamil@statcan.ca, and

Shesh Rai, Department of Biostatistics & Epidemiology, St. Jude Children's Research Hospital, 323 N. Lauderdale St., Memphis, TN 38105-2794.
E-mail: Shesh.Rai@stjude.org

UNE PSEUDO-APPROCHE PAR LE MAXIMUM DE VRAISEMBLANCE POUR L'INFERENCE POUR DES DONNÉES HIÉRARCHIQUES

Milorad S. Kovacevic
et
Shesh N. Rai

Résumé

On montre une application de la pseudo-méthode par le maximum de vraisemblance pour l'estimation d'un modèle linéaire mixte ajusté à des observations dépendantes issues d'une population structurée de façon hiérarchique. L'estimation des paramètres de modèles linéaires mixtes semble d'être plus simple avec cette approche qu'avec des procédures itératives telles que la méthode itérative des moindres carrés pondérés (dans laquelle les poids sont une fonction de probabilités) de Pfeffermann et coll. (1998) puisqu'elle fournit une solution qui n'exige pas d'itération. Nous discutons également de quelques problèmes liés aux hiérarchies du modèle et du plan d'échantillonage ainsi que de leur impact sur l'estimation. Une petite étude par simulation a montré que la procédure proposée est efficace mème quand les tailles d'échantillon des niveaux plus élevées sont petites.

Mots clés: échantionage dans une population finie, estimation de la variance, modelization linéaire hiérarchique

---

Milorad Kovacenic, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Ontario K1A 0T6, couriel: kovamil@statcan.ca, and

Shesh Rai, Department of Biostatistics & Epidemiology, St. Jude Children's Research Hospital, 323 N. Lauderdale St., Memphis, TN 38105-2794.
E-mail: Shesh.Rai@stjude.org

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. At the individual level there are usual types of variables that describe an individual: measurements on continuous scale, indicators of different subpopulations that the individual belongs to, ranks and categories. 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 variables that are available as aggregates at the group level may not be available at the unit level. Some data may come from a survey, some, especially for higher level units, may come from a census or administrative files.

By disaggregation of all higher order variables to the individual level one can ignore the hierarchical structure and analyze data 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 at 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 modeling 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, also known as multilevel 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 (level-1) 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 often called the census equations (Krieger and Pffefermann, 1992) 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 discussed 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. The penultimate section deals with issues of necessity and availability of the weights for different model levels. Section 6 contains some concluding remarks.

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 level-1 (within-group) equation and is designed to describe the relationship between level-1 dependent variables and the level-1 covariates, within each group. Some or all of the parameters of the level-1 equation are viewed as varying randomly across the level-2 unit (group) population. Then in the second equation, level-2 (between-group) equation, these parameters are modelled as dependent variables 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 $q = 1,...,Q$ representing the characteristics of the ith individual. Then, the level-1 (within group) regression equation is

$$y_{gi} = b_{0g} + \sum_{p} \beta_p x_{pgi} + \sum_{q} b_{qg} z_{qgi} + e_{gi}, \tag{1}$$

for $i = 1,...,N_g$ and $g = 1,...,G$, 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 of (1) is

$$\boldsymbol{y}_g = \boldsymbol{X}_g \boldsymbol{\beta} + \boldsymbol{Z}_g \boldsymbol{b}_g + \boldsymbol{e}_g, \tag{2}$$

for $g = 1,...,G$. 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 level-2 (between group) regression equation relates the random within-group coefficients, b_{qg} to group-level characteristics, u_{rg}, $r = 1,...,R$ and $g = 1,...,G$:

$$b_{qg} = \gamma_{q0} + \sum_{r=1}^{R} \gamma_{qr} u_{rg} + d_{qg}, \tag{3}$$

for $q = 0,...,Q$. Group level disturbances d_{qg} are independent from e_{gi} and represent the contributions of the groups to variability that remains unexplained by model (3). Written in a matrix form, equation (3) is

$$\boldsymbol{b}_g = \boldsymbol{F}_g \boldsymbol{\gamma} + \boldsymbol{d}_g \tag{4}$$

where $\boldsymbol{b}_g$ is a Q+1 by 1 vector, $\boldsymbol{F}_g$ is a Q+1 by (R+1)(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+1 whose first element is the constant 1, $\boldsymbol{\gamma}$ is a (R+1)(Q+1) vector of the unknown but fixed parameters, and $\boldsymbol{d}_g$ is a vector of length (Q+1) 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 $\boldsymbol{\Sigma}_{(2)}$, 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, the model ((1), (3)) reduces to one-way ANOVA model with random effects:

$$y_{gi} = b_{0g} + e_{gi} \tag{6}$$

$$b_{0g} = \gamma_{00} + d_g \tag{7}$$

or written together

$$y_{gi} = \gamma_{00} + d_g + e_{gi} \tag{8}$$

Here γ_{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 a two-level model to a model that fits 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. CENSUS 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

$$\begin{aligned} \boldsymbol{y}_g &= \boldsymbol{X}_g \boldsymbol{\beta} + \boldsymbol{Z}_g \boldsymbol{F}_g \boldsymbol{\gamma} + \boldsymbol{Z}_g \boldsymbol{d}_g + \boldsymbol{e}_g \\ &= (\boldsymbol{X}_g \mid \boldsymbol{Z}_g \boldsymbol{F}_g) \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} + \boldsymbol{Z}_g \boldsymbol{d}_g + \boldsymbol{e}_g \\ &= \mathbf{H}_g \boldsymbol{\eta} + \boldsymbol{Z}_g \boldsymbol{d}_g + \boldsymbol{e}_g \\ &= \mathbf{H}_g \boldsymbol{\eta} + \boldsymbol{a}_g \end{aligned} \tag{9}$$

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

$$\boldsymbol{V}_g = Var(\boldsymbol{y}_g) = Var(\boldsymbol{a}_g) = \boldsymbol{Z}_g \boldsymbol{\Sigma}_{(2)} \boldsymbol{Z}_g' + \sigma_{(1)}^2 \boldsymbol{I}_g. \tag{10}$$

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 y_g into a block vector $\mathbf{y}' = [y_1', \ldots y_G']$, then creating a block matrix $\mathbf{H} = [H_1' \mid \ldots \mid H_G']'$, and stacking of the G vectors a_g into a block vector $\mathbf{a}' = [a_1', \ldots, a_G']$, equation (9) can be written for all levels jointly as

$$\mathbf{y} = \mathbf{H}\,\boldsymbol{\eta} + \mathbf{a} \tag{11}$$

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 (10), and $Cov(a_g, a_{g'}) = 0$, for $g \neq g'$.

The unknown finite population parameters $\boldsymbol{\eta}$, $\Sigma_{(2)}$ and $\sigma^2_{(1)}$ can be expressed as functions of the finite population data by solving the corresponding census estimating equations. Assuming that $\mathbf{V}$ is known, using the method of generalized least squares (GLS), the finite population parameter $\boldsymbol{\eta}$ is expressed as:

$$\begin{aligned} \tilde{\boldsymbol{\eta}}_{GLS} &= [\mathbf{H}'\mathbf{V}^{-1}\mathbf{H}]^{-1}\,\mathbf{H}'\mathbf{V}^{-1}\mathbf{y} \\ &= \left(\sum_g H_g' V_g^{-1} H_g\right)^{-1} \sum_g H_g' V_g^{-1} y_g \end{aligned} \tag{12}$$

Here we use ~ to denote a finite population parameter obtained as a solution of the census estimating equations. Allowing for randomness in the generation of the finite population, $\tilde{\boldsymbol{\eta}}_{GLS}$ is random with the covariance matrix given by

$$\begin{aligned} Var\,(\tilde{\boldsymbol{\eta}}_{GLS}) &= [\mathbf{H}'\mathbf{V}^{-1}\mathbf{H}]^{-1} \\ &= \left(\sum_g \boldsymbol{H}_g' \boldsymbol{V}_g^{-1} \boldsymbol{H}_g \right)^{-1} \end{aligned} \tag{13}$$

Estimator (12) 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. Here *MVN* stands for the multivariate normal distribution.

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 criterion is met, is usually used. The problem with such a 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).

We suggest a pseudo maximum likelihood (PML) method to express the finite population parameter $\boldsymbol{\eta}$ by replacing parameter $\mathbf{V}$ in the likelihood equation with its estimate $\tilde{\mathbf{V}}$ and then solving the equation. Estimate $\tilde{\mathbf{V}}$ is obtained using some other appropriate method. A method for obtaining $\tilde{\mathbf{V}}$ appropriate for the hierarchical populations is suggested in section 3.1. The PML expression of the finite population parameter $\boldsymbol{\eta}$ is then

$$\tilde{\boldsymbol{\eta}}_{PML} = [\mathbf{H}'\tilde{\mathbf{V}}^{-1}\mathbf{H}]^{-1}\mathbf{H}'\tilde{\mathbf{V}}^{-1}\mathbf{y}$$
$$= \left(\sum_g H_g'\tilde{V}_g^{-1}H_g\right)^{-1}\sum_g H_g'\tilde{V}_g^{-1}y_g. \qquad (14)$$

The finite population parameter $\tilde{\boldsymbol{\eta}}_{PML}$ can be considered as random variable by seeing the finite population as a realization of a random model. In such a case $\tilde{\boldsymbol{\eta}}_{PML}$ has a corresponding covariance

$$Var(\tilde{\boldsymbol{\eta}}_{PML}) = [\mathbf{H}'\tilde{\mathbf{V}}^{-1}\mathbf{H}]^{-1}\mathbf{H}'\tilde{\mathbf{V}}^{-1}\mathbf{V}\tilde{\mathbf{V}}^{-1}\mathbf{H}[\mathbf{H}'\tilde{\mathbf{V}}^{-1}\mathbf{H}]^{-1}$$
$$= \left(\sum_g H_g'\tilde{V}_g^{-1}H_g\right)^{-1}\left(\sum_g H_g'\tilde{V}_g^{-1}V_g\tilde{V}_g^{-1}H_g\right)\left(\sum_g H_g'\tilde{V}_g^{-1}H_g\right)^{-1} \qquad (15)$$

3.1 Proposed Method for Obtaining $\tilde{\mathbf{V}}$

Equation (9) can be rewritten in a way that combines fixed and random parameters in the same vector

$$y_g = H_g\boldsymbol{\eta} + Z_g d_g + e_g$$
$$= (H_g \mid Z_g)\begin{pmatrix}\boldsymbol{\eta}\\ d_g\end{pmatrix} + e_g \qquad (16)$$
$$= K_g\xi_g + e_g$$

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

Assuming that ξ_g is fixed, its GLS estimate based on only y_g is given by

$$\tilde{\xi}_g = \left(K_g' K_g\right)^{-1} K_g' y_g, \tag{17}$$

for $g=1,\ldots, G$, since $Var(e_g) = \sigma^2_{(1)} I_g$. Furthermore, $\tilde{d}_g$ is the part of $\tilde{\xi}_g$ that corresponds to d_g. Estimator (17) coincides with the maximum likelihood (ML) estimators under the assumption of normality of the vector $\mathbf{y}$. Note that (17) ignores the fact that η is constant over all the groups.

The variance V_g, given by (10), can be reexpressed as

$$\begin{aligned} V_g &= Var(y_g) \\ &= E_{\xi_g} Var(y_g \mid \xi_g) + Var_{\xi_g} E(y_g \mid \xi_g) \\ &= \sigma^2_{(1)} I_g + K_g Var(\xi_g) K_g' \\ &= \sigma^2_{(1)} I_g + Z_g Var(d_g) Z_g' \end{aligned} \tag{18}$$

We approximate conservatively $Var(d_g)$ by $Var(\tilde{d}_g)$ and then V_g can be estimated as

$$\tilde{V}_g = \tilde{\sigma}^2_{(1)} I_g + Z_g \tilde{Var}(\tilde{d}_g) Z_g' \tag{19}$$

where

$$\tilde{\sigma}^2_{(1)} = \frac{1}{G}\sum_g (y_g - \boldsymbol{K}_g\tilde{\xi}_g)'(y_g - \boldsymbol{K}_g\tilde{\xi}_g)/(N_g - \upsilon_g) \tag{20}$$

and

$$\tilde{Var}(\tilde{\boldsymbol{d}}_g) = \frac{1}{G-1}\sum_g (\tilde{\boldsymbol{d}}_g - \frac{1}{G}\sum_g \tilde{\boldsymbol{d}}_g)(\tilde{\boldsymbol{d}}_g - \frac{1}{G}\sum_g \tilde{\boldsymbol{d}}_g)' \tag{21}$$

In equation (21) υ_g is the number of restrictions imposed by the 1st level model. Obviously, even when the group sizes N_g are small, (20) may provide the consistent estimate of $\sigma^2_{(1)}$. Note that (19) is a conservative estimate of (18). We consider a different estimator of (18) elsewhere.

4. ESTIMATION BASED ON A SAMPLE WITH THE COMPLEX DESIGN

If the complete populations of individuals and groups are observed the estimates (14) and (19) are the finite population values of the model parameters. The variance (15) 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 conditions:

$$E\left(\sum_{g=1}^{m}\sum_{i=1}^{n_g} w_{gi}\right) = N,\ E\left(\sum_{g=1}^{m} w_g\right) = G, \text{ and } E\left(\sum_{i=1}^{n_g} w_{i|g}\right) = N_g \tag{22}$$

Let $\boldsymbol{W}_{.|g}$ 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 $\tilde{\xi}_g$, given by (17), is

$$\hat{\xi}_g = \left(\boldsymbol{K}_g' \boldsymbol{W}_{.|g} \boldsymbol{K}_g\right)^{-1} \boldsymbol{K}_g' \boldsymbol{W}_{.|g} \boldsymbol{y}_g, \tag{23}$$

where $\boldsymbol{K}_g$ is a known matrix of size $n_g \times [P+R(Q+1)+(Q+1)]$ and $\boldsymbol{y}_g$ is a vector of size n_g. To estimate the variance component $\tilde{\sigma}^2_{(1)}$, given by (20), which has the form of the population mean

of the values $(y_g - K_g\tilde{\xi}_g)'(y_g - K_g\tilde{\xi}_g)/(N_g - \upsilon_g)$ over g, we use the sample mean

$$\hat{\sigma}^2_{(1)} = \frac{1}{\sum_g w_g}\sum_{g=1}^{m}\frac{w_g}{(\sum_i w_{i|g} - \upsilon_g)}(y_g - K_g\hat{\xi}_g)'(y_g - K_g\hat{\xi}_g) \tag{24}$$

Finite population variance (21) is estimated by the appropriate weighting as

$$\hat{Var}(\hat{d}_g) = \frac{1}{\sum w_g - 1}\sum_g w_g(\hat{d}_g - \bar{\hat{d}})(\hat{d}_g - \bar{\hat{d}})' \tag{25}$$

where $\bar{\hat{d}} = \sum_g w_g\hat{d}_g / \sum_g w_g$.

The matrix of the random components is estimated conservatively by

$$\hat{V}_g = \hat{\sigma}^2_{(1)} I_g + Z_g \hat{Var}(\hat{d}_g) Z_g' \tag{26}$$

The finite population parameter (14) is estimated by

$$\hat{\eta}_{PML} = \left(\sum_g w_g H_g' \hat{V}_g^{-1} H_g\right)^{-1} \sum_g w_g H_g' \hat{V}_g^{-1} y_g, \tag{27}$$

with the corresponding variance $V(\hat{\eta})$ estimated as

$$\hat{V}(\hat{\eta}) = \left(\sum_g w_g H_g' \hat{V}_g^{-1} H_g\right)^{-1}\left(\sum_g w_g^2 H_g' \hat{V}_g^{-1} H_g'\right)\left(\sum_g w_g H_g' \hat{V}_g^{-1} H_g\right)^{-1} \tag{28}$$

This estimate is similar to one obtained by Binder (1983) in the following way: The first and

the last term in (28) are the unbiased estimates of the first derivative of the score function (which is also a design unbiased estimate of information). The middle part is the estimate of the variance of the design based estimate of the score function under an assumption of iid for the sample of groups. Sometimes the difference between these two types of variances is ignored and (28) is reduced to $\left(\sum_g w_g H_g' \hat{V}_g^{-1} H_g \right)^{-1}$. To explain further, consider an estimating function

$$\hat{\psi}(\eta) = \sum_g w_g H_g' V_g^{-1} (y_g - H_g \eta)$$

with

$$E(\partial \hat{\psi}(\eta)/\partial \eta) \doteq \sum_g w_g H_g' V_g^{-1} H_g$$

and

$$Var(\hat{\psi}(\eta)) \doteq \sum_g w_g^2 H_g' V_g^{-1} H_g$$

Equation (28) is a sandwich type estimator based on this estimating function.

6. SOME SAMPLING CONSIDERATIONS

6.1 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 (PSUs) selected from a population of G groups, and s_g is the collection of n_g level-1 units selected from the gth group . The total number of groups (PSUs), G, the number of selected groups, m, the number of selected individuals from a selected group, n_g, and the total sample size, $n=\sum_{g=1}^{G} n_g$, are assumed known. Usually, the group weights, w_g, and the conditional weights, $w_{i|g}$ are not readily available to analysts, although they are needed for analyses. One needs to approximate the weights $w_{i|g}$ and w_g by $\hat{w}_{i|g}$ and $\hat{w}_g$ respectively, so that

$$\sum_{g=1}^{m} \hat{w}_g \approx G, \quad \sum_{i=1}^{n_g} \hat{w}_{i|g} \approx N_g \text{ and } \hat{w}_g\, \hat{w}_{i|g} \approx w_{gi} \tag{29}$$

This can be done iteratively using some of the known raking algorithms. In Table 2 some approximations of the group and the conditional weights for different combinations of available design information are suggested.

It may happen that a complete population of groups is available and the subsamples of individuals are taken from each group. In such a case w_g is equal to 1, and consequently $w_{i|g} = w_{gi}$. In such a case one may question if the effects of groups are random or fixed. From the model point of view, especially if there are many such groups, we consider their effects still to be random.

Table 2: Approximations of the weights for the two-level models

Sample design for groups, parameters	Approximated weights			$\hat{N}_g$
	$\hat{w}_g$	$\hat{w}_{i\|g}$	$\hat{w}_{gi}$	
SRS, N_g unknown	$\frac{G}{m}$	$w_{gi}\frac{m}{G}$	w_{gi}	$W_g\frac{m}{G}$
SRS, N_g known	$\frac{G}{m}$	$w_{gi}\frac{N_g}{\hat{N}_g}\frac{m}{G}$	$w_{gi}\frac{N_g}{\hat{N}_g}$	$W_g\frac{m}{G}$
PPS, N_g unknown	$\frac{G}{m}\frac{1}{W_g}\frac{1}{\hat{H}}$	$w_{gi}W_g\hat{H}\frac{m}{G}$	w_{gi}	$W_g\frac{m}{G}W_g\hat{H}$
PPS, N_g known	$\frac{G}{m}\frac{1}{N_g}\frac{1}{H}$	$w_{gi}N_gH\frac{m}{G}$	w_{gi}	$W_g\frac{m}{G}N_gH$

where $W_g=\sum_{i\in s_g} w_{gi}$, $H=\frac{1}{m}\sum_g\frac{1}{N_g}$, $\hat{H}=\frac{1}{m}\sum_g\frac{1}{W_g}$.

6.2 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 Figure 1a).

Suppose that the two hierarchies are not the same (see Figure 1b). A typical example is when children are selected by an area/household sample but analyzed using schools instead of area units. In this case the areas are called the design groups and the schools are the model groups. 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.

Again we assume that the individual final weights, w_{gi}, are available, as well as the number of sampled design groups, m, and the number of individuals sampled from the selected design groups, n_g. In addition, we assume that a number of the model groups realized in the sample, say k, is known as well as the number of individuals that fall into a model group. The retrospective sampling that we are proposing, 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 inclusion probabilities of the design groups.

We say that a model group is "retrospectively" selected using the Bernoulli sampling with the probability 1 if there is at least one level-1 design unit with the known positive weight (or the inclusion probability) found in that model group, i.e.:

$$Prob\{M_j \mid D_{gi}\} = \begin{cases} 1, & \text{if unit } (gi) \in j, \\ 0, & \text{otherwise.} \end{cases}$$

for j=1,..., k, i=1,..., n_g, and g= 1,..., m. Here M_j denotes the event "jth model group is selected" and D_{gi} stands for the event "gi-th level-1 design unit is selected". Then the "retrospective" probability of selecting a jth model group, $Prob\{M_j\}$, is obtained from the Bayes formula in the following way:

$$1 = Prob\{M_j \mid D_{gi}\} = \frac{Prob\{M_j\}\ Prob\{D_{gi} \mid M_j\}}{Prob\{D_{gi}\}}$$

for $(gi) \in j$ and j=1,..., k, implying that

$$Prob\{D_{gi}\} = Prob\{M_j\}\ Prob\{D_{gi} \mid M_j\}$$

for $(gi) \in j$ and j=1,..., k. Taking reciprocals on both sides and summing up over all $(gi) \in j$, we arrive at

$$\sum_{(gi)\in j} \frac{1}{Prob\{D_{gi}\}} = \frac{1}{Prob\{M_j\}} \sum_{(gi)\in j} \frac{1}{Prob\{D_{gi} \mid M_j\}} \qquad (30)$$

which is equivalent to

$$\sum_{(gi)\in j} \frac{1}{\pi_{gi}} = \frac{1}{\pi_j^*} \sum_{(gi)\in j} \frac{1}{\pi^*_{gi|j}} \tag{31}$$

for j=1,...,k, where π_{gi} denotes the inclusion probability of the (gi)-th design level-1 unit into the sample s, π_j^* is the "retrospective" inclusion probabilities of the jth model unit, and $\pi^*_{gi|j}$ is the conditional inclusion probability for the (gi)-th design level-1 unit given that the jth model unit occured in the sample. In terms of the sampling weights, equation (31) is equivalent to

$$\sum_{(gi)\in j} w_{gi} = w_j^* \sum_{(gi)\in j} w^*_{gi|j}. \tag{32}$$

The sum on the right hand side, $\sum_{(gi)\in j} w^*_{gi|j}$, is equal to the estimated size of the jth model unit, say $\hat{N}_j$. Therefore,

$$w_j^* = \sum_{(gi)\in j} w_{gi} / \hat{N}_j, \text{ and}$$

$$w^*_{gi|j} = \hat{N}_j w_{gi} / \sum_{(gi)\in j} w_{gi}$$

for $(gi)\in j$ and j=1,...,k.

The value of $\hat{N}_j$ can be replaced by the value of the parameter N_j if it is known (which is often the case.) For example, the school principle knows the total number of students in school, or the total number of families in an enumeration area (considered as a neighbourhood) is known from the Census.

7. DISCUSSION

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 heterogeneity. 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. As one of the referee points out, we are investigating the effects of changing the ratio between the first level and second level variance factors. Since the first level variance was fixed, changes in this ratio may be confounded with changes in the overall variance of the observations. To evaluate our procedure further, one may consider additional simulation study keeping the total variance fixed and varying both the first level and second level variances. Also a problem of obtaining appropriate weights for the different levels of the hierarchy is pointed out. Two different approaches were suggested depending on if the design and the model hierarchies are the same or different,

Acknowledgment: The authors thank Professor J.N.K. Rao and Dr. Harold Mantel for their useful comments.

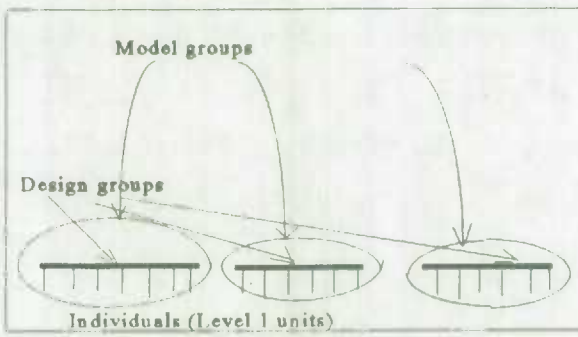
APPENDIX:

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

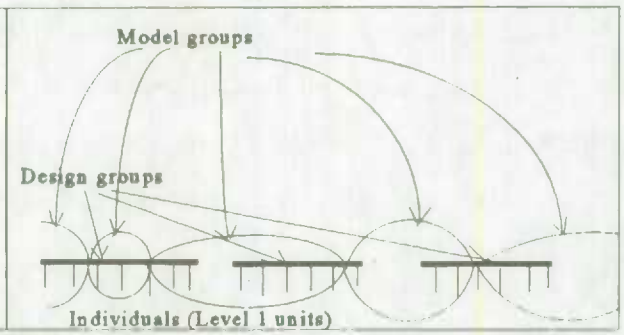
		m=5				m=50				m=200			
n_g	σ_1^2/σ_2^2	$b(\hat{\eta})$	$se(\hat{\eta})$	$rb(\hat{\sigma}^2)$	$re(\hat{\sigma}^2)$	$b(\hat{\eta})$	$se(\hat{\eta})$	$rb(\hat{\sigma}^2)$	$re(\hat{\sigma}^2)$	$b(\hat{\eta})$	$se(\hat{\eta})$	$rb(\hat{\sigma}^2)$	$re(\hat{\sigma}^2)$
All large [50-100]	0.1	3.0	4.5	0.9	2.0	0.3	1.4	0.7	0.6	-0.6	0.7	0.1	0.3
	0.2	1.7	3.2	0.4	1.9	0.8	1.0	-0.1	0.5	0.2	0.5	0.3	0.3
	0.5	-1.8	2.0	0.6	1.5	0.2	0.6	0.5	0.4	0.3	0.3	0.5	0.2
	1	2.8	1.5	1.9	1.1	-0.1	0.4	1.2	0.7	-0.2	0.2	1.1	0.1
	2	2.2	1.0	0.2	0.8	0.3	0.3	0.8	0.2	0.3	0.2	0.9	0.1
Some small [5-100]	0.1	8.0	5.0	-1.2	2.0	0.3	1.6	0.5	0.6	-0.7	0.8	0.3	0.3
	0.2	1.0	4.0	-0.2	2.0	1.1	1.1	1.4	0.5	-0.6	0.6	0.5	0.3
	0.5	2.0	2.0	-0.7	1.7	0.1	0.7	1.1	0.4	0.7	0.4	1.5	0.2
	1	0.0	2.0	1.0	1.0	0.0	0.5	1.1	0.3	0.2	0.3	1.2	0.1
	2	0.0	1.0	2.7	0.7	0.1	0.4	1.5	0.2	0.1	0.2	1.9	0.1
All small [5-10]	0.1	2.9	4.6	3.4	2.1	2.4	1.5	1.6	0.6	-1.6	0.7	1.3	0.3
	0.2	3.1	3.2	3.3	2.0	0.1	1.0	3.1	0.5	0.0	0.5	2.0	0.3
	0.5	-3.5	2.0	2.4	1.5	0.3	0.7	3.6	0.5	-0.5	0.3	4.6	0.2
	1	0.1	1.5	5.2	1.3	0.2	0.5	6.6	0.4	-0.1	0.2	6.4	0.2
	2	0.8	1.1	7.5	1.1	-0.1	0.4	8.0	0.3	0.0	0.2	9.0	0.1

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

a) Hierarchies are the same



b) Hierarchies are different



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 Resources Development 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.

Kovacevic, M. S. and Rai, S.N. (1999). A pseudo maximum likelihood approach to inference on hierarchically structured data. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* (to appear).

Krieger, A.M. and Pffefermann, D. (1992). Maximum Likelihood Estimation from Complex Sample Surveys. *Survey Methodology*, 18, 225-240.

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.