Bootstrapping For Variance Estimation in Multi-Level Models Fitted To Survey Data

Milorad S. Kovacevic, Huang Rong, Yong You
Statistics Canada, Ottawa, Canada

Abstract

For estimation of parameters of a multi-level model fitted to hierarchical survey data, the standard assumptions are that the survey weights are available at all sampling levels and that the hierarchy of sampling levels coincides with the hierarchy used in modeling. Under these two assumptions, we propose two bootstrap methods for the variance estimation of the estimated parameters in the multi-level model. These methods are essentially modifications of the well-known survey bootstrap methods of Rao and Wu (1988). In a simulation study designed according to the Canadian Workplace and Employee Survey (CWES), we study and compare the properties of these methods, and in turn we compare them to the more prevalent Taylor linearization method.

Keywords: Multi-stage design, Informative sampling, Survey weights, Variance components, Taylor linearization

1. Introduction

There has been an increased interest in fitting the multi-level models to survey data in recent years. Problems related to the use of survey weights in estimation of model parameters were addressed by several authors: Graubard and Korn (1996), Pfeffermann et al. (1998), Korn and Graubard (2003), Kovacevic and Rai (1999, 2003), Huang and Hidiroglou (2003), You, Rao and Kovacevic (2003), Asparouhov (2004, 2006), Grilli and Pratesi (2004), to mention only few. Some of these papers also include suggestions for standard error estimation. Pfeffermann et al. (1998) consider a robust "sandwich" estimator, a variant of the Taylor linearization method. Similarly Asparouhov (2004, 2006) advocates the use of the "sandwich" estimator. The multi-level software that can currently handle survey data (Mplus and HLM) also favour the linearization method. Korn and Graubard (2003) suggest variance estimation based on resampling the PSUs, in particular the delete one PSU jackknife method. Grilli and Pratesi (2004) describe a possible two-stage bootstrap, but use only the PSU bootstrap for variance estimation when fitting multilevel ordinal and binary models. None of these papers, with the exception of Grilli and Pratesi (2004), however, studies the properties of any of the resampling methods.

Bootstrapping is a variance estimation technique well researched and often applied for bias reduction and variance estimation in multi-level models by the model-based researchers. The variants of the model-based bootstrap method used in multi-level inference can be categorized as: parametric, residual, and cases bootstrap. The parametric bootstrap simulates level-1 and level-2 residuals from an estimated model distribution. For example, for linear models it is a normal distribution with a zero mean and an estimated variance (Goldstein, 1995). The residual bootstrap resamples the estimated residuals at both levels, while keeping the explanatory variable fixed (Carpenter et al., 2003). The cases bootstrap resamples entire cases of response variables together with their explanatory variables. Resampling may occur at different levels, separately (only at one level) or jointly (resampling at all levels). For a recent review of bootstrapping in multilevel models see van der Leeden, Meijer and Busing (2005).

In this paper, we focus on design-based variance estimation by using the modified rescaled bootstrap method of Rao, Wu and Yue (1992). As an alternative, we investigate a two-stage rescaled bootstrap method (Rao and Wu, 1988, Davison and Hinkley, 1997). Our research is motivated by the practical needs of analysts of survey data who usually rely on the weights prepared for them by survey statisticians. The methods studied allow computation of bootstrap weights and make estimation of the variance and the bias correction of the estimates of model parameters straightforward.

In a limited simulation study motivated by the Canadian Workplace and Employees Survey (CWES), and designed according to a simulation study presented in Pfeffermann et al. (1998), we examine the properties of the proposed bootstrap methods with an emphasis on their performance for inference about the variance components. We also compare these bootstrap methods to the linearized (sandwich) estimator.

The outline of the paper is as follows. In Section 2 we recognize that most of data used for multi-level analysis are obtained by sample surveys based on multi-stage sample designs. We discuss some sample design related issues, such as weighting, clustering, intraclass correlation and informativeness. Section 3 presents the multi-level linear model as a linear mixed model, and in

particular, discusses a random intercept model. We briefly review some of the methods available for estimation of the multi-level model parameters and discuss their merits and problems. We describe the two design-based bootstrap methods for resampling of the clustered survey data in Section 4. In Section 5 we describe in detail the simulation study conducted to compare the performances of the two bootstrap methods as well as the performance of the linearization method. Our findings are summarized and some limited conclusions are derived in the last section of the paper.

2. Multi-Stage Sample Designs

A typical hierarchical data set analysed by fitting multi-level models is obtained by a survey based on a multi-stage sample design. Without loss of generality, we will consider a two-stage design. The first stage sample, s_1, consists of m clusters drawn from a population of M clusters with, most likely, unequal probabilities of selection: $\pi_c = Prob\{c \in s_1\}$. Note that the clusters are sometimes denoted as primary sampling units (PSU), sometimes as groups or level-2 units. These clusters are possibly stratified. The probability weights at the first stage are simply the inverses of the selection probabilities: $d_c = 1/\pi_c$. It may also happen, although rarely, that the weights at this level are calibrated, i.e., d_c becomes w_c, so that the calibrated weights add up to the total number of clusters in the population: $\sum_{c=1}^{m} w_c = M$. From a sampled cluster c, containing N_c individuals (elements, level-1 units), a sample of n_c individuals is drawn usually with equal probabilities of selection, $\pi_{i|c} = n_c / N_c$. The total (unconditional) probability of inclusion into the sample of the ith element from the cth cluster is $\pi_{ci} = \pi_{i|c}\pi_c$. The corresponding probability weights are $d_{i|c} = 1/\pi_{i|c}$, and $d_{ci} = 1/\pi_{ci}$. These probability weights are usually additionally modified to adjust for nonresponse, post-stratification, calibration, outliers, etc., so that $d_{i|c}$ transforms into $w_{i|c}$, and d_{ci} into w_{ci}. In this paper we assume that these adjustments do not affect the cluster weights w_c.

Ignoring the weights when fitting a multi-level model to survey data leads to biased estimation of the model parameters when the sampling is informative, i.e., when the distribution of the sampled units, based on the sample design, is different from the distribution that would be obtained by sampling directly from the model. Also, the presence of clustering implies a positive intraclass correlation between the elements in the same cluster, and has to be accounted for in estimation of standard errors and in test procedures. One of the efficient ways of accounting for the sample design when analysing survey data is to use the design-based approach.

In this paper we assume that the sample design hierarchy coincides with the model hierarchy and that the probability weights are available at all levels of the model hierarchy.

3. Multi-Level Linear Model

A multi-level linear model can be represented as a linear mixed model

$$y_c = \mathbf{X}_c\beta + \mathbf{Z}_c v_c + e_c, \; c = 1,...,M$$

where y_c is an $(N_c \times 1)$ vector of outcome variable, $\mathbf{X}_c$ and $\mathbf{Z}_c$ are known $(N_c \times p)$ and $(N_c \times q)$ covariate matrices, β is a $(p \times 1)$ vector of fixed effects, v_c is a $(q \times 1)$ vector of random effects: $v_c \underset{ind}{\sim} N(0, \mathbf{G}(\theta_v))$, where $\mathbf{G}(\theta_v)$ is a covariance matrix dependant on up to $q(q+1)/2$ unknown dispersion parameters θ_v, and e_c is an $(N_c \times 1)$ vector of within cluster errors $e_c \sim N(0, \sigma_e^2 \mathbf{I}_c)$. Here $\mathbf{I}_c$ denotes the identity matrix. Together, $\theta = \{\theta_v, \sigma_e^2\}$ are the dispersion parameters.

The random part of the model is:

$$\mathbf{Z}_c v_c + e_c \sim N(0, \mathbf{\Sigma}_c(\theta)),$$

where $\mathbf{\Sigma}_c(\theta) = \mathbf{Z}_c \mathbf{G}(\theta_v)\mathbf{Z}_c' + \sigma_e^2 \mathbf{I}_c$.

In the model-based context, completely ignoring the sample design information, the fixed effects β are estimated by ML (or GLS) as

$$\hat{\beta} = (\sum_c \mathbf{X}_c' \mathbf{\Sigma}_c^{-1} \mathbf{X}_c)^{-1} \sum_c \mathbf{X}_c' \mathbf{\Sigma}_c^{-1} y_c,$$

assuming that $\mathbf{\Sigma}_c = \mathbf{\Sigma}_c(\theta)$ is known. Some known large sample properties of $\hat{\beta}$ are: i) If θ is consistently estimated by $\hat{\theta}$, then using $\mathbf{\Sigma}(\hat{\theta})$ in the ML (GLS) estimator $\hat{\beta}$, results in an asymptotically efficient estimate; ii) $\hat{\beta}$ is model-unbiased for any well-defined working $\mathbf{\Sigma}_0$, but possibly inefficient.

For estimation of dispersion parameters θ, there are several methods available in the literature. Some of them,

such as ML, REML, IGLS, RIGLS, are based on alternating between estimation of $\beta(\theta)$ and $\theta(\beta)$. Others include the method of moments, the method of fitting the constants, etc.

In this paper we focus on the random intercept model, a relatively simple two-level model that is prevalent in practice,:

$$y_c = \mathbf{X}_c\beta + j_c u_c + e_c, \quad c = 1,...,M \qquad (1)$$

where $j_c = (1,...,1)'$, $u_c \sim N(0,\sigma_u^2)$, $e_c \sim (0,\sigma_e^2\mathbf{I}_c)$. The variance of the random part of the model is

$$\mathbf{\Sigma}_c = \sigma_u^2 j_c j_c' + \sigma_e^2\mathbf{I}_c .$$

Its inverse has a convenient form

$$\mathbf{\Sigma}_c^{-1} = \{\mathbf{I}_c - \tau_c\, j_c j_c' / N_c\}/\sigma_e^2 ,$$

where $\tau_c = \sigma_u^2 / \left[\sigma_e^2 / N_c + \sigma_u^2\right]$. The finite population parameters β_N can be defined, for example, as the GLS estimate of the model parameters β, i.e., as a solution of the normal equation:

$$\sum_{c=1}^{M} \mathbf{X}_c'\left[y_c - \mathbf{X}_c\beta_N - \tau_c(\bar{y}_c - \bar{\mathbf{X}}_c\beta_N)\right] = 0 , \quad (2)$$

where $\bar{\mathbf{X}}_c$ is the matrix of the means of the $\boldsymbol{X}$ variables for the cth cluster repeated N_c times (i.e., $\bar{\mathbf{X}}_c = (\bar{x}_c,...,\bar{x}_c)'$), and $\bar{y}_c$ is the vector with the mean of the response variable repeated N_c times (i.e., $\bar{y}_c = (\bar{y}_c,...,\bar{y}_c)'$).

For a model fitted to survey data, the estimate of β, as well as of β_N, $\hat{\beta}$ is obtained from the weighted normal equation (2), and is equal to

$$\hat{\beta} = \left[\sum_c \mathbf{X}_c'\mathbf{W}_c\hat{\mathbf{D}}_c\right]^{-1} \sum_c \hat{\mathbf{D}}_c'\mathbf{W}_c y_c ,$$

where $\hat{\mathbf{D}}_c = \mathbf{X}_c - \hat{\tau}_c\hat{\bar{\mathbf{X}}}_c$, $\mathbf{W}_c = \mathbf{diag}(w_{c1},...,w_{cn_c})$, $\hat{\tau}_c = \hat{\sigma}_u^2 / \left(\hat{\sigma}_e^2 / \hat{N}_c + \hat{\sigma}_u^2\right)$, $\hat{N}_c = \sum_i w_{i|c}$, and

$\hat{\bar{\mathbf{X}}}_c = (\hat{\bar{x}}_c,...,\hat{\bar{x}}_c)'$ is the $n_c \times p$ matrix of estimated cluster means of $\boldsymbol{X}$ variables repeated n_c times:

$$\hat{\bar{x}}_c = \sum_{i\in c} w_{i|c} X_{ci} / \sum_{i\in c} w_{i|c} = \sum_{i\in c} w_{ci} X_{ci} / \sum_{i\in c} w_{ci} .$$

Similarly, $\hat{\bar{y}}_c$, needed for the weighted normal equation, is the $n_c \times 1$ vector containing the cluster mean for the response variable,

$$\hat{\bar{y}}_c = \sum_{i\in c} w_{i|c} y_{ci} / \sum_{i\in c} w_{i|c} = \sum_{i\in c} w_{ci} y_{ci} / \sum_{i\in c} w_{ci}$$

Note that even if σ_u^2 and σ_e^2 are assumed known, the estimation of β still requires knowledge of the within cluster weight, $w_{i|c}$, since the calculation of $\hat{N}_c$ requires $w_{i|c}$. All other weighting is based on the joint weights w_{ci}. An analyst usually has two pieces of sample-design information at the element level: the (joint) final weight, w_{ci}, and a cluster indicator $I_{ci} = 1$ if $i \in c$, otherwise $I_{ci} = 0$. Having only these two pieces of information one can still estimate consistently the fixed parameters assuming a working covariance structure with $\tau_c = 0$. However, any estimation of the dispersion parameters σ_u^2 and σ_e^2, using only these two pieces of information, can be seriously biased.

For estimation of dispersion parameters θ there are several methods proposed in the recent literature. One of them, the Probability-Weighted IGLS (Pfeffermann, et al., 1998) is the IGLS adapted to complex sample designs by the proper weighting. This method iterates between estimation of $\hat{\beta}$ and $\hat{\theta}$ in two weighted GLS regressions. From the first regression, the "raw" residuals $\tilde{y}_{ci} = y_{ci} - x_{ci}'\hat{\beta}$ are computed. Then, a cross-product matrix of $\tilde{y}\,\tilde{y}'$, written in vector form $\{\tilde{y}\,\tilde{y}'\}$, is considered as a response vector in the second GLS equation, where θ is a vector of unknown regression coefficients. The procedure alternates between the two regressions until convergence. Estimates $\hat{\theta}$ are consistent when sample and population sizes mutually increase. A reduction of the small-sample bias in estimation of the variance components by the PWIGLS method, particularly for σ_e^2, was addressed in the paper by proposing the scaling of element weights so that the scaled weights sum up to the within-cluster sample size or to the within-cluster effective sample size.

Other proposed methods for variance components estimation include the method of moments (MM) (Korn and Graubard, 2003), the Pseudo-BLUP (Huang and Hidiroglou, 2003) with the application of a weighted version of Henderson's method III for estimation of variance components, the Pseudo-EBLUP (You and Rao, 2002) which uses the unweighted Henderson's method, the Iterative Weighted Estimating Equations (You, Rao, Kovacevic, 2003) which uses iterative updating of the weighted Henderson's method III, the Multilevel Pseudo Maximum Likelihood (MPML) (Asparouhov, 2006) which uses the weighted maximum likelihood equations.

Based on a chosen estimation method (PWIGLS, PML, MM, etc.) the estimates $\hat{\beta}$ and $\hat{\theta}$ are ultimately obtained as functions (although not in a closed form) of the observed y, x and z variables (y_s, X_s, Z_s), and the weights at different levels $\{w_{ci}\}, \{w_c\}, \{w_{i|c}\}$, i.e.,

$\hat{\beta} = \hat{\beta}\left(y_s, X_s, Z_s, \{w_{ci}\}, \{w_c\}, \{w_{i|c}\}\right)$, and

$\hat{\theta} = \hat{\theta}\left(y_s, X_s, Z_s, \{w_{ci}\}, \{w_c\}, \{w_{i|c}\}\right)$.

4. Variance Estimation

4.1 Bootstrap Procedures

Without loss of generality, we present two bootstrap procedures for a simplified situation where there is no stratification of the clusters. An extension to a stratified population is straightforward. The original sample has m clusters drawn according to a sample design $p_1(s)$. From each sampled cluster c, a subsample s_{1c} of n_c elements is drawn by a sample design $p_2(s)$. We assume that the probability weights are known at different levels: cluster weights d_c, the within-cluster weights $d_{i|c}$, and the joint weights $d_{ci} = d_c d_{i|c}$. Further, we assume that the weight adjustments done to the final weights are not affected the cluster weights.

Bootstrap Procedure 1 (BS1): The Rescaled Bootstrap Method (Rao, Wu, Yue, 1992) modified to suit hierarchical data:

A bootstrap replicate b is obtained as follows:

i) A SRSWR of m-1 clusters is drawn. A counter $t_c^{(b)}$ counts the number of times that the c-th cluster is included in the bootstrap replicate b.

ii) The cluster weights d_c are rescaled to obtain the cluster bootstrap weights:

$$d_c^{(b)} = d_c \frac{m}{m-1} t_c^{(b)}.$$

If any adjustment was done to d_c it should be applied to $d_c^{(b)}$ too, in which case $d_c^{(b)}$ becomes $w_c^{(b)}$.

iii) The unadjusted joint bootstrap weights are calculated as

$$d_{ci}^{(b)} = w_c{}^{(b)} d_{i|c}.$$

iv) The adjusted joint bootstrap weights, $w_{ci}^{(b)}$, are obtained from $d_{ci}^{(b)}$ by applying all the adjustments made in the process of calculating the full sample joint survey weights.

v) The within-cluster bootstrap weight, $w_{i|c}^{(b)}$, is then calculated as:

$$w_{i|c}^{(b)} = w_{ci}^{(b)} / w_c^{(b)}$$

Steps i) - v) are repeated many times to obtain many bootstrap replicates.

This bootstrap procedure preserves the exact hierarchical structure of the original sample and is relatively simple computationally. If there is no adjustment done to the joint bootstrap weights then the within-cluster bootstrap weights remain equal to the original full sample within-cluster weight, $w_{i|c}^{(b)} = w_{i|c} = d_{i|c}$. That is why this procedure may not be very accurate in capturing the variability of estimates that are "within-cluster dependent", such as $\hat{\sigma}_e^2$, even when the within-cluster sample size is large. Also, if the bootstrap procedure is to be used for the bias correction, it may not be very efficient in removing the bias for these estimates.

Note that this is the same procedure as the original Rao, Wu, Yue (1992) with an additional calculation of the weights at different levels.

Bootstrap Procedure 2 (BS2): A Two-stage Bootstrap for Hierarchical Data. The procedure, originally proposed by Rao and Wu (1988), is modified to suit hierarchical data. Note that a similar procedure is given by Davison and Hinkley (1997, page 100) in a model-based context (thus without weighting).

The first two steps are identical to the steps i) and ii) in the Bootstrap Procedure 1.

iii) A SRSWR of $n_c - 1$ elements is drawn from a cluster selected in Step i). Independent selections are made from the same cluster chosen more than once. A counter $t_{i|c}^{(b)}$ counts the total number of times that the i-th element is resampled.

iv) The within-cluster probability weights are rescaled first to obtain the unadjusted within-cluster bootstrap weights as

$$d_{i|c}^{(b)} = d_{i|c}\frac{n_c}{n_c-1}t_{i|c}^{(b)} / t_c^{(b)}$$

v) The unadjusted joint bootstrap weights are

$$d_{ci}^{(b)} = w_c{}^{(b)} d_{i|c}^{(b)} .$$

vi) The adjusted joint bootstrap weights, $w_{ci}^{(b)}$, are obtained from $d_{ci}^{(b)}$ after applying all the same adjustments done in the process of calculating the original full sample joint survey weights.

vii) The within-cluster bootstrap weights are then calculated as:

$$w_{i|c}^{(b)} = w_{ci}^{(b)} / w_c^{(b)}$$

Steps i) -vii) are repeated many times.

This procedure also preserves the hierarchical structure of the original sample. It should give approximately the same results as the first procedure for the estimates whose variability is mainly due to the sampling of clusters. However, it should be more accurate in estimating the variability of "within-cluster dependent" estimates and more efficient in removing their biases. It is important to emphasize that it may encounter computational problems when within-cluster sample sizes are small, and that it is computationally more intense than the first procedure.

From either of the above procedures, an analytic file which contains *B* sets of bootstrap weights at each level and the joint bootstrap weights along with the original full sample weights and the survey variables is produced:

$$Data = \left\{ \begin{array}{l} (y_{ci}, x_{ci}', z_{ci}', \underbrace{w_{i|c}, w_c, w_{ci}}_{full\ sample}, \\ \underbrace{w_{i|c}^{(1)}, w_c^{(1)}, w_{ci}^{(1)}}_{replicate\ 1}, ..., \underbrace{w_{i|c}^{(B)}, w_c^{(B)}, w_{ci}^{(B)}}_{replicate\ B}) \end{array} \right\}_{ci \in s}$$

Estimates of $\xi = (\beta', \theta')'$ are obtained using the full sample weights as well as using the bootstrap weights, yielding $\hat{\xi} = (\hat{\beta}', \hat{\theta}')'$ and $\hat{\xi}^{(b)} = \left(\hat{\beta}^{(b)'}, \hat{\theta}^{(b)'} \right)'$, $b = 1, ..., B$. The bootstrap estimate of the variance matrix of estimates of fixed effects $\hat{\beta}$ and of dispersion parameters $\hat{\theta}$ is

$$\hat{V}_{BS}(\hat{\xi}) = \sum_{b=1}^{B} (\hat{\xi}^{(b)} - \hat{\xi})(\hat{\xi}^{(b)} - \hat{\xi})' / B$$

4.2 Taylor Linearization

For standard error estimation of both the regression and the dispersion parameters, Pfeffermann et al. (1998) consider a robust sandwich estimator, a variant of the Taylor linearization method. Similarly, Asparouhov (2004, 2006) advocates the use of the robust sandwich estimator. This estimator considers only the variation among the clusters (level-2) units. Assuming a small sample fraction of clusters m/M, but with *m* still large (more than 50), and when the weights are the probability weights, the design-based Taylor linearization variance estimator for $\hat{\beta}$ proved to be a good estimator (Pfeffermann et al., 1998). In the case of the random intercept model, the sandwich estimator takes the following form:

$$\hat{V}(\hat{\beta}) = \hat{\mathbf{J}}^{-1} \hat{V}_p \left\{ \sum_{c=1}^{m} \hat{\mathbf{D}}_c' \mathbf{W}_c (y_c - X_c' \hat{\beta}) \right\} \hat{\mathbf{J}}^{-1} \qquad (3)$$

where $\hat{\mathbf{J}} = \sum_{c=1}^{m} \mathbf{X}_c' \mathbf{W}_c \hat{\mathbf{D}}_c$, $\hat{\mathbf{D}}_c = \mathbf{X}_c - \hat{\tau}_c \hat{\bar{\mathbf{X}}}_c$, and $\hat{V}_p\{.\}$ is the design-based covariance estimator for the vector of totals of the form:

$$\left\{ \sum_{c=1}^{m} \sum_{i=1}^{n_c} w_{ci} (X_{ci,k} - \hat{\tau}_c \hat{\bar{X}}_{c,k})(y_{ci} - \sum_{j=1}^{p} X_{ci,j} \hat{\beta}_j) \right\}_{k=1,...p}$$

Note that in the above expression both $\hat{\beta}$ and $\hat{\tau}$ are the estimates from the last iteration.

Similarly, the robust sandwich variance estimator for $\hat{\theta}$, following Pfeffermann et al. (1998), is obtained in a closed form.

Typically, the "sandwich" estimator (3) underestimates the design-based standard errors of $\hat{\beta}$ and $\hat{\theta}$ due to the plug-in principle for the "nuisance" parameters, and also due to a possible ignoring of the stochastic adjustments to the weights, such as the unit non-response, poststratification, calibration, etc.

5. Simulation Study

In order to make an empirical assessment of the properties of the two proposed bootstrap procedures and to compare them to the Taylor linearization method, we carried out a simulation study. We simulated a model of the relationship between the hourly wage and several factors associated with employees and employers, motivated by a model fitted to data from the 1999 CWES as presented in Drolet (2002).

The simulated model expresses the logarithm of hourly wage (***hwg***) as a linear function of the employee's years of experience (***yexp***, continuous variable), the highest education (***hedu***, 3 categories), and the occupation group (***ocgrp***, 3 categories). The model also uses a binary variable (***nprft***) at the level of employer (company) indicating whether or not the company is non-profit. The model has the form of a linear random intercept model:

$$\log(hwg_{ci}) = \beta_0 + \mu_{ci} + u_c + e_{ci},$$

where

$$\begin{aligned}\mu_{ci} &= \beta_1 \cdot nprft_c + \beta_2 \cdot yexp_{ci} \\ &+ \beta_{31} \cdot I\{hedu_{ci} = 1\} + \beta_{32} \cdot I\{hedu_{ci} = 2\} \\ &+ \beta_{41} \cdot I\{ocpgr_{ci} = 1\} + \beta_{42} \cdot I\{ocpgr_{ci} = 2\}\end{aligned} \qquad (4)$$

$$u_c \sim N(0, \sigma_u^2 = 0.05),\ e_{ci} \sim N(0, \sigma_e^2 = 0.1),$$
$$c = 1,...,M,\ i = 1,...,N_c.$$

The variables used in the model were generated using the joint distributions and parameters estimated from the CWES.

5.1 Simulation of a Two-level Finite Population

First, we generated a finite population of *M*=1000 clusters (level-2 units). The cluster sizes were obtained as a function of cluster random effects u_c, (see Pfeffermann et al., 1998), so that

$$N_c = Integer[100\exp(2.5u_c)],$$

where $u_c \sim N(0, \sigma_u^2 = 0.05)$. In this way we allowed for the possibility of informative sampling since the same random effects, u_c, will be used for generation of the response variable ***hwg*** using model (4). The resulting cluster sizes took values in the interval (30, 400) with the mean size about 140, determining the total size of the population of individuals (level-1 units) of approximately $\sum_{c=1}^{1000} N_c \approx 140{,}000$. Then, *M* values of the cluster-level variable ***nprft*** were simulated from its distribution estimated from the CWES. Thus the population of clusters is created as $\{u_c, N_c, nprft_c\}_{c=1,....,M}$.

Next, we generated a population of individuals in the following way: (i) $N = \sum_{c=1}^{1000} N_c$ values of *X* covariates were first simulated from their joint distributions estimated from the CWES. Also, we simulated *N* values of *e* from $e \sim N(0, \sigma_e^2 = 0.1)$. (ii) For each *c*, *c*=1,…,1000, we generated N_c values of the response variable ***hwg*** according to model (4) using the same u_c that was used to generate the cluster size N_c. The coefficients $\beta = (\beta_0, \beta_1, \beta_2, \beta_{31}, \beta_{32}, \beta_{41}, \beta_{42})$ were adapted from the estimates obtained by fitting the same model to the CWES data, thus we used $\beta = (2.5, 0.2, 0.01, -0.15, -0.05, 0.4, 0.1)$. (iii) In addition, we stratified the individuals in each cluster into two strata according to the sign of the corresponding e_{ci} value. This stratification will later allow us to oversample from one stratum to reinforce the informativeness of the sampling design (see Pfefferman et al., 1998).

5.2 Sample Design

We considered a two-stage sample design. Clusters were selected with a probability proportional to size (N_c) without replacement using Sampford's method as implemented in SAS procedure Surveyselect. We drew clusters into samples of two different sizes, *m*=50 and *m*=100. The cluster probability weights were adjusted so that they added up to the total number of clusters in the population. The individuals were selected by simple random sampling without replacement from two strata within each of the sampled clusters. We used two sample sizes: (i) fixed size of $n_c = 8$, and (ii) variable size $n_c = \max\{6, 0.1N_c\}$. The sample of individuals was allocated (25%, 75%) in two strata. The resulting total sample sizes were 400 and 800 individuals for the fixed within-cluster sample size, and about 700 and 1400 in the case of variable within-cluster sample size.

We also imposed a poststratification of individuals into four poststrata according to two binary variables that were generated along with the variables used in the model. These two variables were ***sex*** and ***union*** (membership in trade-union). The poststratification was applied to the joint weights.

The sampling weights were calculated according to the sample design and the poststratification. They were produced at both levels. In addition to the original weights, we calculated the scaled within-cluster weights using the scaling method 2 of Pfefferman et al. (1998):

$$\tilde{w}_{i|c} = w_{i|c} \frac{n_c}{\sum_i w_{i|c}}. \qquad (5)$$

For each sample we produced *B*=200 bootstrap replicates using the Bootstrap Method 1 and *B*=200 bootstrap replicates using the Bootstrap Method 2. The resulting bootstrap weights were obtained as described in Section 4.1.

5.3 Monte Carlo Setup

Considering the two sizes of cluster sample (50 and 100) and two types of within-cluster sample sizes (fixed and variable) we had four different sample settings. We selected 300 samples from the finite population using each of these four settings. From each sample we estimated fixed and dispersion model parameters by using the PWIGLS method of Pfeffermann et al., 1998. For estimation we used the original (unscaled) weights, and the scaled weights (5). The standard errors were estimated using the two bootstrap methods and the Taylor linearization method. All variance estimation methods were applied with the unscaled and scaled weights. All programming was done using SAS IML.

5.4 Performance Measures for Variance Estimation Methods

The methods were compared with respect to accuracy, stability and the coverage properties of the resulting interval estimates. The measures used for the comparisons are given below. Here, ξ is used to represent any one of the seven β regression coefficients or the two dispersion parameters, $\theta = (\sigma_e^2, \sigma_u^2)$. Naturally, $\hat{\xi}$ is used to denote an estimate of ξ, in general, while $\hat{\xi}_k$ denotes the estimate of ξ_k based on the *k*th sample. The letter *M* is used to denote any one of the variance estimation methods being compared {BS1, BS2, Taylor, BS1_s, BS2_s, Taylor_s}. Here '_s' denotes the use of the scaled weights. Finally, *K* denotes the number of samples being used (*K*= 300).

The accuracy of method M is assessed by the relative bias of the resulting bootstrap variance estimates:

$$rel.bias(\hat{V}_M(\hat{\xi})) = \frac{1}{K}\sum_k^K \frac{\hat{V}_{M,k}(\hat{\xi}_k) - EMSE(\hat{\xi})}{EMSE(\hat{\xi})}.$$

The empirical mean square error (EMSE) is calculated over the K independent samples as $EMSE(\hat{\xi}) = \frac{1}{K}\sum_k^K (\hat{\xi}_k - \hat{\xi})^2$. It is considered as a 'true' MSE for the comparison of the methods. The smaller the relative bias the more accurate is the method for variance estimation. A negative bias would mean that the method underestimates the variance leading to overstated precision and significance.

The stability method M is quantified by the Relative Root Mean Square Error of the estimated variance:

$$rel.rmse(\hat{\xi}) = \sqrt{\frac{1}{K}\sum_k^K \left(\frac{\hat{V}_M(\hat{\xi}_k)}{EMSE(\hat{\xi})} - 1\right)^2}.$$

Certainly, the more stable method will have the smaller *rel.rmse*.

The empirical coverage rates for method *M* are computed in order to assess the coverage properties of normal-theory confidence intervals:

$$e.c.r_{1-\alpha}(\hat{V}_M(\hat{\xi})) = \frac{1}{K}\sum_k^K I\left\{\frac{\left|\hat{\xi}_k - \xi_k\right|}{\sqrt{\hat{V}_M(\hat{\xi}_k)}} \le z_{\alpha/2}\right\}.$$

We report the coverage rates for nominal confidence of 100(1-α)%=90 percent. In the above expression $I\{a\}$=1 if a is true, and 0 otherwise, and $z_{\alpha/2}$ is the upper α/2th standard normal percentile. Upper and lower tail error rates were also computed but are not reported here.

5.5 Results of Simulation Study

The graphical presentations of the results of the comparisons of the three methods for variance estimation, using the unscaled and scaled weights, are given in the Appendix. They are given in three charts referring to the accuracy (Chart 1), stability (Chart 2) and coverage properties (Chart 3) of the resulting variance estimates. Each chart has four graphs representing the four different arrangements of sample sizes. Each variance method is represented by a different color, and one solid and one dashed line, and a different symbol. The solid line corresponds to the use of original (unscaled) weights while the dashed line indicates the use of scaled (5) weights. In each graph, the four regression coefficients and two dispersion parameters are given along the horizontal axis and the value of the measure is on the vertical axis. For visual clarity, we omit the intercept and the two categories of the ***ocpgr*** variable from the graphical presentation. We denoted the variables as *NP* (***nprft***), *ED1 and ED2* (two categories of the ***hedu***, the reference category is high education), EXP (***yexp***), s2u (σ_u^2) and s2e (σ_e^2).

6. Discussion and Conclusion

6.1 Accuracy of Variance Estimation (Chart 1)

We found little difference between BS1 and BS2 with respect to the accuracy of the variance estimation of the fixed effects. BS1 and BS2 overestimate the variances of the fixed effects at level-1, but underestimate the variance of the fixed effects at level-2 (variable ***nprft***). The Taylor method underestimates the variances of all estimates and for all settings.

Scaling of weights affects very little the relative bias of variance estimates for fixed effects. Scaling of weights slightly reduces the bias of all variance estimates for $\hat{\sigma}_e^2$.

The number of clusters in the sample has higher impact than the within-cluster sample sizes on the accuracy of variance estimates.

6.2 Stability of Variance Estimation (Chart 2)

The Taylor method provides the uniformly most stable variance estimates. There is virtually no difference between results obtained using unscaled and scaled weights.

There is a negligible difference in stability among the bootstrap methods for variance estimation of the fixed effects. For all methods the least stable is the variance of $\hat{\sigma}_e^2$.

For stability of the variance estimates, the number of clusters in the sample is more important than the within-cluster sample sizes. The exception is estimation of the variance of $\hat{\sigma}_e^2$ which shows sensitivity on the within-cluster sample sizes as well.

6.3 Coverage Properties (Chart 3)

Regardless of sample sizes, all methods yielded similar coverage for the fixed effect associated with the cluster-level variable NP (nprft). For other fixed effects, the Taylor method always understated the nominal level and the bootstrap methods overstated it. Two bootstrap methods performed similarly for the dispersion parameters, both significantly overstated the coverage for $\hat{\sigma}_e^2$.

Scaling of weights improved the coverage properties of all methods, especially for estimation of $\hat{\sigma}_e^2$.

The number of clusters affects more the coverage rates than the within-cluster sample sizes.

6.4 Conclusion

From our simulation study, it follows that the survey bootstrap method BS1 can be applied for variance estimation in linear multi-level models under very general conditions. Method BS2 with the resampling within clusters seems to be a complex undertaking with a relatively small gain in accuracy of variance estimation for the dispersion parameters. The performance of the bootstrap methods depends on both sample sizes, with BS2 being slightly more sensitive to the within-cluster sample size. Overall there is a gain in both accuracy and stability of variance estimation when using the scaled weights. The robust "sandwich" variance estimator systematically underestimates the design-based variance.

References:

Asparouhov, T. (2004), "Weighting for Unequal probability of selection in multilevel modeling." *Mplus Web Notes, No. 8*.: http://www.statmodel.com

Asparouhov, T. (2006), "General Multi-Level Modelling with Sampling Weights." *Communications in Statistics – Theory and methods*, 35, 439-460

Carpenter, J.R., Goldstein, H., and Rasbah, J. (2003), "A novel bootstrap procedure for assessing the relationship between class size and achievement." *Appl. Statist.* , Vol. 52, pp 431-443.

Davison, A.C. and Hinkley, D.V. (1997), "*Bootstrap Methods and Their Application.*" Cambridge University Press

Drolet, M. (2002), "The "Who, What, When and Where" of Gender Pay Differentials," *Research Paper Series, Catalogue No.:71-584-MIE2002004, Statistics Canada*

Goldstein, H. (1995), "*Multi level statistical models*," 2nd edition. London, Arnold.

Graubard, B., and Korn, E. (1996), "Modeling the Sampling Design in the Analysis of Health Surveys." Statitics. *Statist. Meth. Med. Res*., 5, 263-281

Grilli, L. and Pratesi, M. (2004), "Weighted estimation in multi-level ordinal and binary models in the presence of informative sampling designs." *Survey Methodology*, 30, 93-103.

Huang, R. and Hidiroglou, M. (2003), "Design Consistent Estimators for a Mixed Linear Model on Survey Data." *ASA Proceedings of Survey Research Methods Section*, pp. 1897-1904.

Korn, E. and Graubard, B. (2003), "Estimating the variance components by using survey data.." *Journal of the Royal Statistical Society, Series B,* 65, 175-190

Kovacevic, M. S. and Rai, S. N. (1999), "Multi-level modelling of survey data." *ASA Proceedings of Survey Research Methods Section,* pp. 605-610.

Kovacevic, M. S. and Rai, S. N. (2003), "A pseudo maximum likelihood approach to multi-level modeling of survey data." *Communications in Statistics – Theory and methods, 32,* 103-121.

van der Leeden, R., Meijer, E., & Busing, F. M. T. A. (2005), "Resampling multilevel models." *In Handbook of Multilevel Analysis*. Ed. J. de Leeuw , New York: Springer.

Pfefferman, D., Skinner, C.J., Holmes, D.J., Goldstein, H., and Rasbash, J. (1998), "Weighting for unequal

selection probabilities in multi-level models." *J. Roy. Statist. Soc., Ser. B, (with discussion),* 60:23-76.

Rao, J.N.K. and Wu, C.F.J. (1988), "Resampling Inference with Complex Survey Data." *J. Amer. Statist. Assoc.,* Vol.83, No. 401, 203-241.

Rao, J.N.K, Wu , C.F.E., and Yue, K (1992), "Some Recent Work on Resampling Methods for Complex Surveys." *Survey Methodology*, Vol. 18, No. 2, pp. 209-217

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, pp. 431-439

You, Y., Rao, J.N.K., and Kovacevic, M.S. (2003), "Estimating Fixed Effects and variance componrnts in a random intercept model using survey data." *In Proceeding of Statistics Canada Symposium* 2003, Ottawa

Appendix

Chart 1. Relative Bias of the Variance Estimates

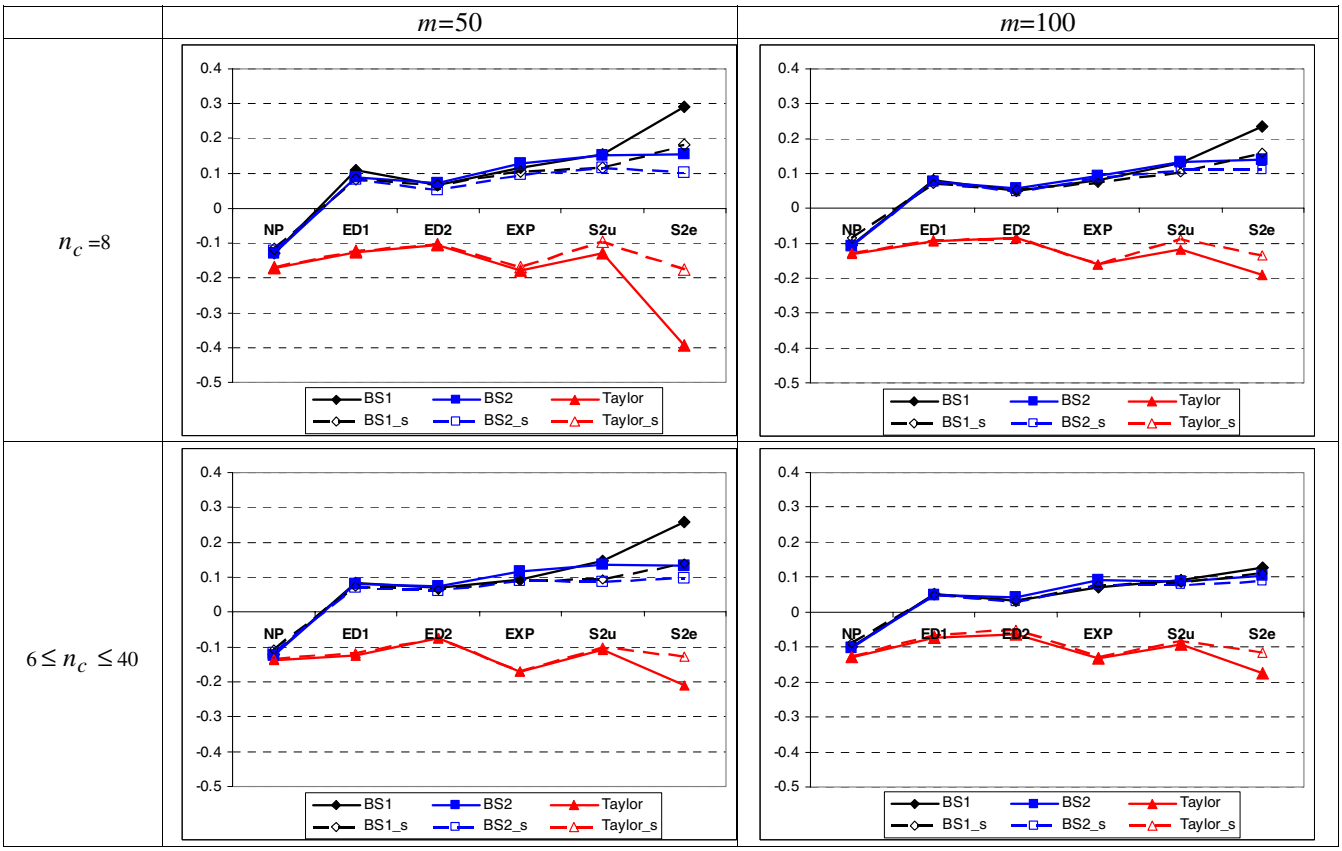


Chart 2: Relative Root MSE of the Variance Estimates

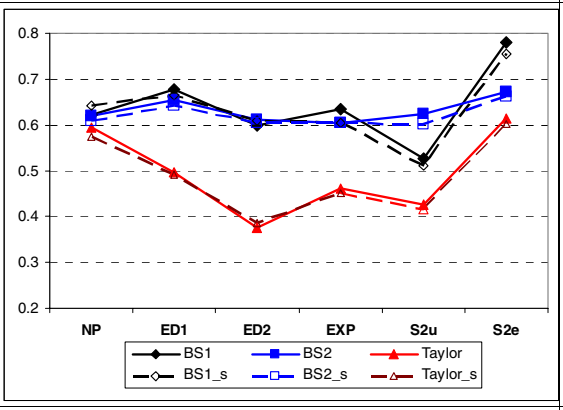
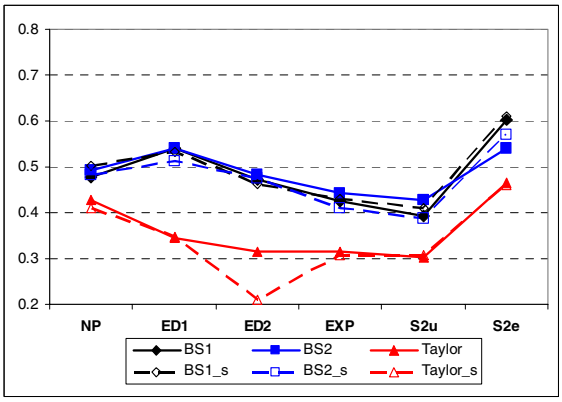
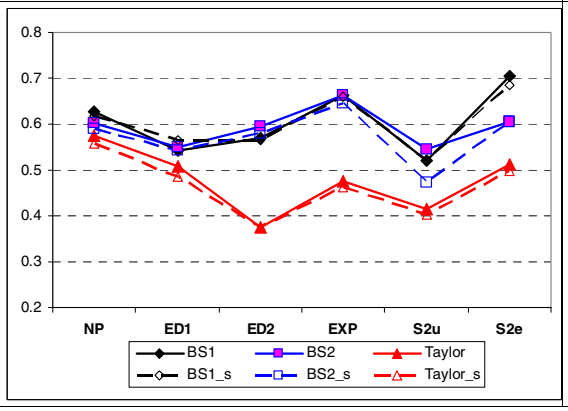
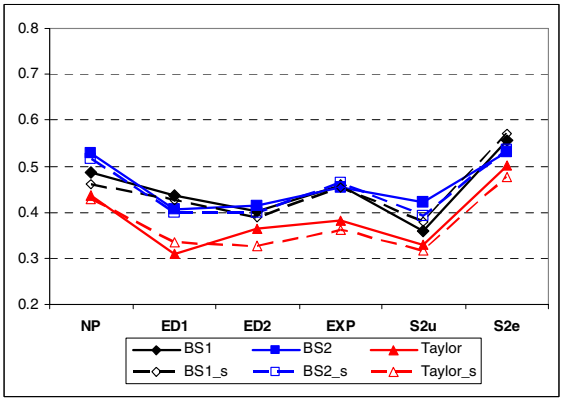
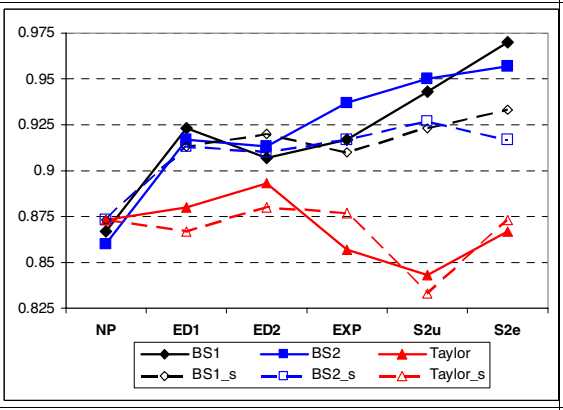
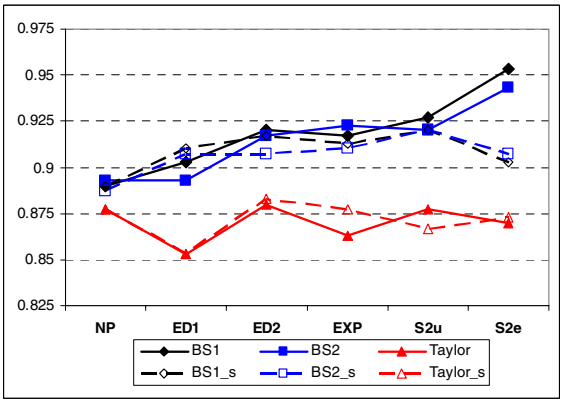
	m=50	m=100
n_c =8		
$6 \le n_c \le 40$		

Chart 3: Coverage Rate for the Nominal 90% (Normal) Confidence Interval

	m=50	m=100
n_c =8		
$6 \le n_c \le 40$