Multilevel Modeling of Complex Survey Data

Tihomir Asparouhov[1], Bengt Muthen[2]
Muthen & Muthen[1]
University of California, Los Angeles[2]

Abstract

We describe a multivariate, multilevel, pseudo maximum likelihood estimation method for multistage stratified cluster sampling designs, including finite population and unequal probability sampling. Multilevel models can be estimated with this method while incorporating the sampling design in the standard error computation. Design based adjustment of the likelihood ratio test (LRT) statistic is proposed. We also discuss multiple group and subpopulation analysis in this context. Simulation studies are conducted to evaluate the performance of the proposed estimator and test statistic. We also compare the estimators and the LRT adjustments implemented in Mplus and LISREL in simulation studies.

Keywords: Multilevel Models, Multistage Sampling.

1 Introduction

Multilevel models are frequently used to analyze data from multistage sampling designs. Such sampling designs use unequal probability of selection at each sampling stage, stratified sampling, cluster sampling and finite population sampling. Multilevel models are used to study the effect of cluster level variables on the individual outcomes, however in multistage samples there are various levels of clustering. For example consider a survey of school aged children. School districts can be the primary sampling units (PSU), classrooms can be the secondary sampling units (SSU) and the students within the classroom can be the third level sampling unit (TSU). In addition the school district sampling can be stratified by more homogeneous regions to improve the quality of the estimation. For example urban schools districts may form one stratum, suburban school district can form another stratum and rural school districts can form another stratum.

Typically two-level models will be used to study the effects of the lowest level of clustering, e.g. classrooms, on individual outcomes. This is because the lowest level of clustering usually has the greatest impact and because it is of greater substantive interest. In addition, to model the cluster level effects a relatively large number of units are needed. In our example the teacher's effect on the students performance can be included in the model, while school district effects and strata effects will typically not be included in a two-level model. Since the SSUs are nested within the PSUs they will not be independent. Thus inference assuming independence of the SSU, which is the basic assumption of the two-level model, will not be accurate. In addition if we ignore the stratification in the sampling design the precision gains obtained by this design feature will be unaccounted for.

To adjust the estimation for the unequal probability of selection, sampling weights are assigned at one or both levels in the two-level model. Let p_j be the probability of selection for SSU j and let $p_{i|j}$ be the probability that individual i in SSU j is selected, given that SSU j is selected. The sampling weights on the cluster (between) level are then obtained by

$$w_j = 1/p_j.$$

The sampling weights on the individual (within) level are then obtained by

$$w_{i|j} = 1/p_{i|j}.$$

If the sampling weights are ignored at either level the parameter estimates can be substantially biased. There are a number of articles that propose two-level estimation methods that utilize the sampling weights to reduce or eliminate the bias. None of these proposed methods have achieved this objective completely. In general, the unequal probability of selection for the within level units remains problematic especially when the within level selection is highly informative and the cluster sample sizes are small. For large cluster sample sizes however it is possible to obtain consistent parameter estimates. For a detailed discussion on two-level estimation with sampling weights see Asparouhov (2006).

Unlike unequal probability of selection it is easy to incorporate stratification, cluster sampling and finite population sampling in the estimation of two-level models. We build upon the pseudo-maximum likelihood estimation method developed by Skinner (1989), following ideas of Binder (1983). The pseudo-maximum likelihood (PML) is generally defined for a single level models, however it has been adopted for two-level models as well, see Grilli and Pratesi (2004), Asparouhov (2004), Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006). We call the two-level version of the PML estimator the multilevel pseudo maximum likelihood estimator (MPML). The main advantages of the MPML estimator over other two-level weighted estimators is that it applies to all two-level models, including all multilevel models implemented in Mplus such as two-level latent variable models, multilevel Probit and logistic regressions, multilevel multinomial logistic regression, multilevel mixture models and multilevel continuous and discrete time survival models. Another advantage of the MPML estimator is that it can easily incorporate missing data under the standard assumption of missingness at random (MAR).

Incorporating stratification, cluster sampling and finite population sampling into the MPML estimator amounts to adjusting the variance computation of the score vector. Variance computation for stratified cluster samples with and without replacement are well established techniques, see for example Cochran (1977) or SUDAAN User Manual (2002). In Asparouhov and Muthen (2005) this approach was used to derive explicit estimation formulas for the PML estimator for single level models and for the three most common multistage survey designs: WR (with replacement sampling), WOR (without replacement sampling) and WORUNEQ (without replacement unequal probability sampling). This terminology was pioneered in the software program SUDAAN and it has been adopted widely in practice. The WOR design is a stratified multistage sampling design with equal probabilities without replacement sampling at the PSU level and equal probabilities with or without replacement sampling at the subsequent stages. The WORUNEQ is a stratified multistage design with unequal probabilities without replacement sampling at the PSU level and with or without replacement equal probabilities sampling at subsequent stages. In Section 2 we review the MPML estimator and provide explicit formulas for the variance estimation for the three cluster designs using the same approach as in Asparouhov and Muthen (2005). In Section 4 we describe a design based adjustment to the likelihood ratio test (LRT) that can be used with the MPML estimator. In Section 4 we discuss Multiple Group and Subpopulation Analysis for multilevel models. In Section 5 we conduct a simple simulation study to evaluate the performance of the MPML estimator and the adjusted LRT test for a stratified three-stage cluster sample design. In Section 6 we compare the adjusted LRT implemented in Mplus with the adjusted LRT implemented in LISREL in a simulation study using a cluster sampling design. In Section 7 we conduct a simulation study to compare the parameter estimates and their standard errors obtained by Mplus and LISREL for a two level regression model with sampling weights at both levels. All computations are performed with Mplus 4.2 (Muthen & Muthen, 1998-2006) and LISREL 8.8 (SSI, 2006).

2 Multilevel Pseudo Maximum Likelihood Estimation in Multistage Sampling

In this section we describe the MPML estimator for a general parametric model and the three sampling designs WR, WOR, and WORUNEQ. We describe the MPML estimator for a 3-stage stratified cluster sampling design. Suppose that the population is divided into S strata. In stratum s we sample n_s PSUs. From the k-th PSU in stratum s we sample n_{sk} SSUs (clusters) and finally from the j-th SSU in the k-th PSU in stratum s we sample n_{skj} TSUs (individuals). Denote the observed individual variables y_{skji} for individual i in cluster j in PSU k in stratum s. Denote the cluster random effect by η_{skj}, the individual level covariates by x_{skji} and the cluster level covariates by x_{skj}. Denote the density function of y_{skji} by $f(y_{skji}|x_{skji}, \eta_{skj}, \theta_1)$ and the density function of η_{skj} by $\phi(\eta_{skj}|x_{skj}, \theta_2)$, where θ_1 and θ_2 are the parameters to be estimated on the within and the between level respectively. Let $w_{skj} = 1/p_{skj}$ and $w_{skji} = 1/p_{skji}$ be the sampling weights for the cluster and the individual level. The within level weights w_{skji} are consequently scaled to improve the estimation method. A number of different scaling methods have been considered in the literature, see for example Pfeffermann et al. (1998), Stapleton (2002) and Asparouhov (2006). Several scaling methods are implemented in the software package Mplus, see Muthen & Muthen (1998-2006). In this article we consider only the most common scaling method where the weights are standardized to add up to the sample size of the corresponding cluster

$$w'_{skji} = n_{skj} \frac{w_{skji}}{\sum_i w_{skji}}. \qquad (1)$$

Let l_{skj} be the weighted pseudo likelihood of the observed data in the $j-$th cluster

$$l_{skj} = \int \Big(\prod_i f(y_{skji}|x_{skji}, \eta_{skj}, \theta_1)^{w'_{skji}} \Big) \phi(\eta_{skj}|x_{skj}, \theta_2) d\eta_{skj}. \qquad (2)$$

The MPML estimates $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ are obtained by maximizing the total weighted pseudo likelihood

$$l = \prod_{s,k,j} l_{skj}^{w_{skj}}. \qquad (3)$$

Denote by $L = log(l)$ the weighted pseudo log-likelihood and by $L_{skj} = log(l_{skj})$ the weighted log-likelihood of the j-th cluster. The asymptotic covariance matrix of the parameters θ is then given by

$$(L'')^{-1} Var(L')(L'')^{-1}, \qquad (4)$$

where $'$ and $''$ refer to the first and the second derivative of the log-likelihoods with respect to the parameters θ. The middle term

$$Var(L') = Var(\sum_{s,k,j} w_{skj} L_{skj})$$

is computed according to the formulas for the variance of the weighted estimate of the total described in Cochran, Chapter 11 (1977) taking the appropriate design into account.

We now describe $Var(L')$ for the three sampling designs WR, WOR and WORUNEQ. Let $Z_{skj} = w_{skj} L_{skj}$. Also define $Z_{sk} = \sum_j Z_{skj}$, $Z_s = \sum_k Z_{sk}$ and

$$s_{sk} = \sum_j (Z_{skj} - \bar{Z}_{sk.})^T (Z_{skj} - \bar{Z}_{sk.})$$

$$s_s = \sum_k (Z_{sk} - \bar{Z}_{s.})^T (Z_{sk} - \bar{Z}_{s.})$$

Let f_{sk} and f_s be the sampling fractions at each of the corresponding sampling stages if the sampling at that stage is without replacement. If the sampling is with replacement these quantities are assumed to be 0.

For the WR design the variance of the score is given by

$$Var(L') = \sum_s \frac{n_s}{n_s - 1} s_s.$$

For the WOR design the variance of the score is given by

$$Var(L') = V_1 + V_2,$$

where

$$V_1 = \sum_s (1 - f_s) \frac{n_s}{n_s - 1} s_s$$

$$V_2 = \sum_{sk} (1 - f_{sk}) f_s \frac{n_{sk}}{n_{sk} - 1} s_{sk}$$

For the WORUNEQ design we need the joint probability of selection. The probability that PSU k in stratum s is selected is denoted by $p_{k|s}$. The probability that both PSUs k_1 and k_2 in stratum s are selected in the sample is denoted by $p_{k_1 k_2|s}$. The variance of the score is given by

$$Var(L') = V_1 + V_2,$$

where

$$V_1 = \sum_s \sum_{k_1} \sum_{k_2 > k_1} \frac{p_{k_1|s} p_{k_2|s} - p_{k_1 k_2|s}}{p_{k_1 k_2|s}} (Z_{sk_1} - Z_{sk_2})^2$$

$$V_2 = \sum_{sk} (1 - f_{sk}) p_{k|s} \frac{n_{sk}}{n_{sk} - 1} s_{sk}$$

3 Likelihood Ratio Test

Hypotheses involving several parameters are frequently tested in multivariate modeling. Here we will explore the possibility to use the MPML pseudo maximum-likelihood value to perform likelihood ratio test (LRT). The distribution of the LRT statistic based on the maximized weighted log-likelihood value is not a chi-square distribution. This distribution depends on the sampling design just as the asymptotic covariance of the parameter estimates depends on the sampling design. In Asparouhov and Muthen (2005) we describe an adjustment of the single level LRT statistic which takes into account the sampling design and produces a test statistic with a chi-square distribution. This adjustment is constructed similarly to the adjustments of the Yuan-Bentler (2000) and the Satorra-Bentler (1988) robust chi-square tests for mean and variance structures. Similar first and second order adjustments are described also in Rao-Thomas (1989) for contingency tables. In the multilevel context we know that when the cluster sample sizes are large the MPML log-likelihood can be approximated by PML log-likelihood, see Asparouhov(2006), thus for designs with large cluster the adjustment of the PML-LRT can be used to adjust the MPML-LRT. In this section we will explore the quality of this approximation through a simulation study.

First let's describe the LRT adjustment and see how it applies in the multilevel context. We assume a general hypothesis testing for two nested models M_1 and M_2. Let θ_i be the true parameter values and $\hat{\theta}_i$ the parameters estimates for model M_i that maximize the pseudo log-likelihood function L_i. Let d_i be the number of parameters in model M_i. The corrected LRT statistic is

$$T^* = c \cdot 2(L_1 - L_2), \qquad (5)$$

where c is the correction factor

$$c = \frac{d1 - d2}{Tr((L_1'')^{-1} Var(L_1')) - Tr((L_2'')^{-1} Var(L_2'))}. \qquad (6)$$

The above description of the adjusted LRT can be used with the MPML estimator as well. For single level models and multilevel models with large cluster size the statistic T^* has approximately a chi-square distribution with $d_1 - d_2$ degrees of freedom. The components $Tr((L_i'')^{-1} Var(L_i'))$ are easily available since they are part of the asymptotic covariance for the parameter estimates given in (4).

4 Multiple Group and Subpopulation Analysis

A multiple group model is a model which estimates several submodels for several subpopulations or groups. The different submodels can have some parameters that are held equal across the different subpopulations and some parameters that are different across the subpopulations. The subpopulations can be defined by a grouping variable such as gender or race. The grouping variable is essentially a categorical predictor variable. In linear regression for example when a categorical variable is included as a predictor, a regression model is essentially equivalent to a multiple group model with different means across the subpopulations. In multivariate modeling however multiple group models are more general than the regression models in that they allow not only means to vary across subpopulations but also slopes and variance covariance parameters as well.

Subpopulation analysis is conducted when only a particular subpopulation D is of interests. If the complement of the subpopulation is D_1 the subpopulation analysis estimates a model for group D while it does not estimate a model for the complement D_1. Subpopulation analysis is equivalent to a multiple group analysis based on the two groups D and D_1 where the models for D and D_1 have no parameters in common. In that respect subpopulation analysis is a special case of multiple group analysis.

One way to conduct a subpopulation analysis with complex survey data is to estimate a two-group model. However for single level models a more popular approach is to estimate one model for the entire population by using zero sampling weights for all observations that are not in the domain of interest D, see for example Korn and Graubard (1999) Section 5.4. Let's call this approach the zero-weight approach. For single level models the zero-weight approach and the two-group approach are equivalent. For multilevel models however the two-group approach is better than the zero-weight approach because the zero-weight approach could underestimate the standard errors.

To demonstrate this, we first focus on the definition of a multiple group multilevel model. One popular approach is to assume that there is one set of cluster level random effects that applies to all groups. The multiple group version of the pseudo loglikelihood (2) is then given by

$$l_{skj} = \int \Big(\prod_i f_g(y_{skji}|x_{skji}, \eta_{skj}, \theta_1, g_{skji} = g)^{w'_{skji}} \Big) \phi(\eta_{skj}|x_{skj}, \theta_2) d\eta_{skj} \quad (7)$$

where g is the grouping variable taking values $1, ..., G$ where G is the total number of groups. This however is not the most general formulation. A more general formulation is to assume that there are different cluster level random effects η_{skjg} for each group, which may or may not be correlated. In this case the pseudo loglikelihood is

$$l_{skj} = \int \Big(\prod_i f_g(y_{skji}|x_{skji}, \eta_{skjg}, \theta_1, g_{skji} = g)^{w'_{skji}} \Big) \phi(\eta_{skj}|x_{skj}, \theta_2) d\eta_{skj} \quad (8)$$

where η_{skj} is the vector of all random effects $(\eta_{skj1}, ..., \eta_{skjG})$. The model with likelihood (7) is a special case of the model with likelihood (8). The model with likelihood (7) essentially assumes that the random effects η_{skjg} have correlation 1 and equal variance across groups.

If we now estimate subpopulation analysis by the zero-weight approach the pseudo loglikelihood is

$$l_{skj} = \prod_g \int \Big(\prod_{i, g_{skji}=g} f_g(y_{skji}|x_{skji}, \eta_{skjg}, \theta_1)^{w'_{skji}} \Big) \phi(\eta_{skjg}|x_{skj}, \theta_2) d\eta_{skjg}. \quad (9)$$

The model with likelihood (9) is also a special case of the model with likelihood (8). This model assumes that the random effects η_{skjg} are independent. This assumption could be incorrect in practical applications because these random effects usually are highly correlated. Such incorrect independence assumptions could result in underestimation of standard errors of the parameter estimates and even in biased parameter estimates.

In conclusion, for correct multilevel subpopulation analysis it is best to use the multiple group approach rather than the zero-weight approach. Note however that when the subpopulations are nested above the cluster level the two approaches are equivalent and thus the zero-weight approach is still valid. Also if the correlation between the random effects is small the two approaches will produce the similar result.

Another obstacle to the zero-weight approach is the fact that some software packages automatically rescale the weights. Thus adding observations with zero weights to the sample will actually inflate the weights for the observations in the subpopulation of interest. Such weight inflation produces unscaled weights, which results in biased estimates, see Asparouhov (2006). To avoid this problem the zero-weight approach should be used with software packages that allow model estimation without automatic weights scaling or software packages that have a special subpopulation implementation. Both approaches are possible in Mplus.

5 Simulation Study

In this section we conduct a basic simulation study to evaluate the performance of the variance computation and the LRT adjustment described in Sections 2 and 3. First we construct a target population of size 50000, which consists of 5000 clusters of size 10. Each observation consist of 10 dependent observations Y_1,..., Y_{10} which are generated from a two-level factor model with one factor on the between level and two factors on the within level. For a general discussion on multilevel structural equation models and their applications for example see Muthen (1994).

We denote the $r-$th observation for individual i in cluster j by Y_{jir}. The model used to generate the target population is described by the following equation

$$Y_{jir} = Y_{jrb} + Y_{jirw}$$

where Y_{jrb} is the between, cluster specific, part of Y_{jir} and Y_{jirw} is the within, individual specific, part of Y_{jir}, i.e., Y_{jrb} is the random intercept in a two-level hierarchical settings and Y_{jirw} is the residual variable. The unobserved random variables Y_{jrb} and Y_{jirw} are assumed to be independent of each other and are normally distributed. The following equations describe the distribution of Y_{jrb} and Y_{jirw} used in the data generation process

$$Y_{jirw} = \lambda_{rw}\eta_{ji1} + \varepsilon_{jir}, r = 1, ..., 5$$

$$Y_{jirw} = \lambda_{rw}\eta_{ji2} + \varepsilon_{jir}, r = 6, ..., 10$$

$$Y_{jrb} = \mu_i + \lambda_{rb}\eta_j + \varepsilon_{jr}, r = 1, ..., 10$$

where $\lambda_{rw} = 1$, for $i = 1, ..., 5$, $\lambda_{rw} = 0.7$, for $i = 6, ..., 10$, $\lambda_{rb} = 0.8$ are the within and the between level loading parameters that are to be estimated. The variables η_{ji1} and η_{ji2} are zero-mean within level factors which are normally distributed random variables with variance 1 and correlation parameter to be estimated $\rho = 0.5$. The variable η_j is the normally distributed between level factor with zero mean and variance 1. The variables ε_{jir} and ε_{jr} are zero-mean normal residuals with variance $\theta_{rw} = 1$ and $\theta_{rb} = 0.5$. The intercept parameters are $\mu_r = 0$. We call the parameter values used in the generation routine the true parameters. First we estimate the above model using the entire target population. We call the resulting

Table 1: Bias of MPML Parameter Estimates and Population Value Coverage

Parameter	True Value	Population Value	$L = 10$	$L = 20$	$L = 30$
λ_{1w}	1.0	0.996	0.007(89%)	-0.002(93%)	-0.005(96%)
λ_{6w}	0.7	0.698	-0.004(94%)	-0.007(91%)	-0.007(92%)
λ_{1b}	0.8	0.789	-0.006(100%)	-0.008(98%)	0.002(97%)
ρ	0.5	0.492	0.008(92%)	0.002(93%)	-0.002(94%)
μ_1	0.0	-0.023	-0.002(96%)	0.011(94%)	-0.001(97%)
θ_{1w}	1.0	1.003	-0.004(91%)	-0.004(93%)	-0.001(97%)
θ_{1b}	0.5	0.480	-0.016(94%)	0.000(96%)	-0.012(91%)

parameter estimates the population values. As expected, the population parameters are very close to the true parameters because the sample size is very large.

To construct strata and PSUs within our target population we first compute the following two cluster level variables

$$M_j = \sum_{i,r} Y_{jir}$$

$$S_j = \sum_i \sum_r (Y_{jir} - \overline{Y}_{ji.})^2.$$

M_j is the sum of all observations in a cluster while S_j is a multiple of the sum of the sample variances of all observations in the cluster. We now reorder the entire target population so that the cluster variable M_j increases monotonically as j increases. We use the first 1000 clusters to form stratum 1, the next 3000 clusters to form stratum 2, and the last 1000 clusters to form stratum 3. Thus the clusters with large Y_{jir} will appear in the first stratum, the clusters with small Y_{jir} will appear in the last stratum and the clusters with medium Y_{jir} will appear in the second stratum. Within each stratum we order the clusters in ascending S_j order and we combine every 10 consecutive clusters to form a PSU. Thus strata 1 and 3 have 100 PSUs and stratum 2 has 300 PSUs. This choice of constructing the strata and the PSUs guarantees that the multistage sampling is informative and that it is not equivalent to simple random sampling.

We then construct the sampling scheme as a stratified 3 stage random sampling as follows. From each stratum we select L PSUs at random with replacement. From each PSU we select 5 clusters at random with replacement. Finally from each cluster we select 5 observations at random without replacement. It is very important that at the last sampling stage we use without replacement sampling. We will discuss this point later in this section. The total sample size is $75L$ and the total number of PSUs in the sample is $3L$. We conduct simulation study with $L = 10$, 20 and 30. For each value of L we select 100 samples and estimate the correct two-level factor model. We compute the parameter estimates and their standard errors, as well as the LRT statistic and the adjusted LRT statistic.

Table 2: Rejection LRT rates

Test	$L = 10$	$L = 20$	$L = 30$
Unadjusted LRT	88%	92%	89%
Adjusted LRT	20%	13%	7%

The results of the simulation study are presented in Tables 1 and 2. Table 1 contains the bias of the parameter estimates, computed as the difference between the average parameter estimates and the population values. Table 1 also contains in brackets the coverage values, which represent the percentage of replications for which the population values are covered by the estimated 95% confidence intervals. We include the results only for a representative selection of parameters. The results in Table 1 show that in all cases the bias of the parameter estimates is very small and the coverage value is very close to the nominal 95% value. Table 2 contains the rejection rates of the adjusted and the unadjusted LRT tests for the true factor model against the saturated two-level mean and variance/covariance model. The rejection rate is the percentage of replications with p-value smaller than 5%. Since the model is correct we expect the rejection rates to be close to the nominal value of 5%. The estimated factor model has 51 parameters while the saturated model has 120, and therefore there are 69 degrees of freedom. Table 2 shows that the unadjusted LRT has very poor performance. The unadjusted LRT overestimates the test statistic, underestimates the p-value and inflates the rejection rate. On the other hand the unadjusted LRT appears to perform well especially when the number of PSUs increases.

Let's now evaluate the individual effect of the stratification and the clustering on the variance estimation. In Table 3 we compute the ratio between the average of the standard errors and the standard deviation of the parameter estimates. When the standard error computation is correct this ratio should be close to 1. Any deviation from 1 would indicate underestimation or overestimation of the standard errors. In Table 3 we report this ratio for the full design variance computation, for the full design

Table 3: Ratio of Average Standard Errors to Standard Deviation of Parameter Estimates

Parameter	Full Design	Excluding Stratification	Excluding Clustering
λ_{1w}	1.000	0.991	0.911
λ_{6w}	0.947	0.939	0.886
λ_{1b}	1.062	1.491	0.931
ρ	0.985	0.977	0.882
μ_1	1.017	1.496	0.900
θ_{1w}	1.048	1.032	0.959
θ_{1b}	1.014	1.007	0.869

but excluding the stratification in the sampling and for the full design but excluding the clustering in the sampling. The full design excluding the clustering is a stratified multistage sampling design where we ignore the PSU sampling and we assume that the clusters are sampled at random and not as they really are.

Table 3 contains the results for $L = 20$. The results show that excluding the stratification results in overestimation of the standard errors for some of the parameters, namely the intercepts and the between level loadings. Excluding the clustering, results in a small underestimation of the standard errors for all parameters. For the full design the results suggest that the estimated standard errors are very close to the correct values.

Finally, let's focus on the sampling method at the last sampling stage in the multistage sampling. For single level analysis when the sampling at the primary stage is WR, the method of sampling at the consecutive stages does not affect the variance estimation. This however is not the case for multilevel models. The sampling method for the lowest level, the level that is included in the model has to be WOR. If the WR method is used instead, any repetitions of observations can inflate the correlation between the observations in the cluster which will result in bias estimates. To illustrate this we compare the results for sample designs with WR sampling at the last stage and with WOR sampling at the last stage. Again we present the results for $L = 20$. Table 4 contains the bias of the parameter estimates and the coverage of the population values by the estimated confidence intervals. The results indicate that the parameter estimates based on the design with WR sampling at the last sampling stage are biased and the coverage for some of the parameters is low. Thus we conclude that the estimator described in Section 2 is inappropriate for multilevel modeling based on sampling designs with WR sampling at the last sampling stage.

6 Comparing Mplus and LISREL LRT Adjustments

An alternative LRT adjustment has been proposed and implemented in the LISREL software package. In this section we explore the differences between the LISREL adjustment and the adjustment described in this article and implemented in Mplus. As described in the LISREL documentation (2005) accompanying the software package, the adjustment is given by equation (5) where the correction factor c is computed by

$$c = \frac{d2}{Tr((L_2'')^{-1}Var(L_2'))}. \tag{10}$$

This formula can also be found in Stapleton (2006). The adjustment is available in LISREL for single level models.

Table 4: Bias of MPML Parameter Estimates and Population Value Coverage for WOR and WR Sampling at the Last Stage

Last Stage Sampling	WOR	WR
λ_{1w}	-0.002(93%)	-0.049(76%)
λ_{6w}	-0.007(91%)	-0.042(87%)
λ_{1b}	-0.008(98%)	0.040(93%)
ρ	0.002(93%)	-0.010(95%)
μ_1	0.011(94%)	-0.012(97%)
θ_{1w}	-0.004(93%)	-0.094(67%)
θ_{1b}	0.000(96%)	0.106(84%)

We illustrate the differences between the two adjustments with a simple simulation study. For simplicity we use a single level model but this discussion applies to multilevel models as well. We generate a target population of size 5000 with two observed variables Y_1 and Y_2 from a bivariate normal distribution with means $\mu_1 = \mu_2 = 0$, variances $\psi_1 = \psi_2 = 1$ and covariance $\rho = 0$. We reorder the target population so that the values of Y_1 are in ascending order. Clusters of size 10 are then constructed as follows. The first 10 observations are placed in cluster 1, the next 10 observations are placed in cluster 2, etc. The target population then contains 500 clusters. We select 100 samples from the target population by cluster sampling, i.e., for each sample we select at random L clusters and use all observations from that cluster. Thus the sample size is $10L$. Using the entire target population we estimate the population values $\mu_1 = -0.018$, $\mu_2 = 0.014$, $\psi_1 = 1.011$, $\psi_2 = 1.041$ and $\rho = 0.025$. The LRT is used to test between the following two models, the saturated model where all 5 parameters are estimated and a restricted model where the parameters μ_1, ψ_1 and ρ are fixed to their population values. Since the model restrictions are correct the LRT test should have a rejection rate of approximately 5%. The test between the two models has 3 degrees of freedom and thus the mean value of the LRT statistic should be approximately 3. Table 5 shows the rejection rates for the three LRT statistics, the Mplus LRT adjustment, the LISREL LRT adjustment and the Unadjusted LRT. Tables 6 shows the average values of these test statistics. It is clear from these results that the Mplus LRT adjustment performs

Table 5: LRT Rejection Rates

Test	L=50	L=100	L=200
Mplus LRT adjustment	10%	5%	6%
LISREL LRT adjustment	66%	67%	65%
Unadjusted LRT	68%	69%	67%

Table 6: LRT Average Values

Test	L=50	L=100	L=200
Mplus LRT adjustment	3.2	2.7	2.8
LISREL LRT adjustment	19.7	16.3	16.7
Unadjusted LRT	21.0	18.3	18.6

very well in all cases, the rejection rates are close to the nominal 5% value and the average test statistic values is close to 3. In contrast the LISREL LRT adjustment and the unadjusted LRT produced incorrectly large rejection rates and average inflated test statistic values.

7 Comparing Mplus and LISREL Estimation of Hierarchical Regressions with Sampling Weights

Recently several structural equation modeling and multilevel software packages have implemented more accurate statistical methodology for analyzing complex survey data, see Asparouhov (2005). Despite these improvements large differences in the results obtained from different packages are being reported in practical applications, see Chantala and Suchindran (2006) for example. In this section we conduct a simulation study to evaluate the performance of the estimation methods implemented in Mplus and LISREL for estimating a two-level random effect regression model with informative sampling weights on both levels. The Mplus estimation is based on the multilevel pseudo maximum likelihood estimation method described in this article while LISREL implements the PWIGLS method described in Pfeffermann et al. (1998). In both software packages we use the scaling to cluster sample size for the within level sampling weights given in equation (1).

We conduct a simulation study on a two-level regression model with a normally distributed dependent variable Y and two normally distributed independent variables X and Z. The covariate Z has a fixed effect on Y while the covariate X has a random effect on Y. This two-level regression model is described as follows

$$Y_{ji} = \alpha_j + \beta_j X_{ji} + \gamma Z_{ji} + \varepsilon_{ji} \tag{11}$$

where α_j and β_j are normally distributed cluster level random effects with means $\alpha = 0.5$ and $\beta = 0.1$ and variances $\psi_\alpha = 1$ and $\psi_\beta = 0.2$ and covariance $\rho = 0.3$. The residual effect ε_{ij} is a mean zero independent normal random variable with variance $\theta = 1$. The covariates X_{ji} is generated from a normal distribution with mean 3 and variance 2 while Z_{ji} is generated from a standard normal distribution. The fixed effect γ is set at 0.5. The model has a total of seven parameters. We generate 100 samples of size 25000. Each sample has 1000 clusters of size 25. To introduce unequal probability sampling on the within level we retain each observation in the sample with probability

$$p_{i|j} = \frac{1}{1 + Exp(-Y_{ij}/2)}. \tag{12}$$

For all observations in the sample we compute the weight variable as

$$w_{ji} = \frac{1}{p_{i|j}} = 1 + Exp(-Y_{ij}/2). \tag{13}$$

Consequently we rescale the within level weights using formula (1). To introduce unequal probability sampling on the between level we retain clusters in the sample with probability

$$p_j = \frac{1}{1 + Exp(-\alpha_j)}. \tag{14}$$

For all clusters in the sample we compute the between level weight as

$$w_j = \frac{1}{p_j} = 1 + Exp(-\alpha_j). \tag{15}$$

We estimate model (11) for each sample using Mplus and LISREL. Within the LISREL software package this kind of models are estimated by the MULTILEV module.

Table 7 contains the bias, the mean squared errors (MSE) and the confidence interval coverage for both software packages. The Mplus bias for all parameters is very close to 0, however the LISREL bias is relatively large for the α and ψ_α parameters. When conducting the simulation study with informative selection on the within level only or on the between level only the parameter estimates and standard errors between Mplus and LISREL are identical. The differences reported in Table 1 occur only when we use sampling weights at both levels. The LISREL bias is also directly affected by the informativeness of the selection on the between level. The stronger the association between α_j and the probability of selection the bigger the bias is. This fact also explains why only the mean and the variance parameters α_j have this bias. If the selection on the between level was associated with β_j we would see this bias for the mean and variance of β_j. The LISREL bias also resulted in larger MSE when compared to Mplus MSE. The coverage probabilities were overall better in Mplus although both packages were far from the nominal 95% probability. The ratio between the standard deviation of the parameter estimates and the standard errors were close to 1 in both programs. This means that the drop in the coverage is caused primarily by the bias in the parameter estimates, which tends to

Table 7: Bias and MSE of parameter estimates for two-level regression estimation in Mplus and LISREL.

Para- meter	True Value	Mplus Bias	LISREL Bias	Mplus MSE	LISREL MSE	Mplus Coverage	LISREL Coverage
α	0.5	0.03	0.10	0.004	0.013	0.94	0.51
β	0.1	0.02	0.01	0.001	0.001	0.88	0.94
γ	0.05	-0.01	-0.01	0.000	0.000	0.88	0.90
ψ_α	1.0	0.03	-0.12	0.011	0.019	0.98	0.61
ψ_β	0.2	-0.01	-0.02	0.000	0.001	0.81	0.64
ρ	0.3	-0.03	-0.03	0.002	0.002	0.78	0.67
θ	1.0	-0.03	-0.02	0.001	0.001	0.61	0.79

disappear as the number of clusters in the sample and the cluster sample sizes increase.

Even though in our simulation study the results obtained with Mplus were somewhat more accurate than those obtained with LISREL, there is no guarantee that this will be the case for other simulation studies or in specific practical applications. When the data is obtained via simple random sampling the maximum likelihood estimator (MLE) is known to be the most accurate estimator at least when the sample size is sufficiently large. Consequently most software packages are based on the MLE and the applied researchers are accustomed to obtaining the same results from different statistical packages. When the data is obtained from a complex survey design however, there is no one estimator that is always more accurate than all other estimators. Such a most accurate estimator does not exist even for the most basic estimation problems with sampling weights. Consider for example the case when the sampling weights are noninformative. An estimator that completely ignores the weights will be more accurate than an estimator that facilitates the weights. However this will not be the case if the weights are informative. Because there is no one estimator that is the most accurate in all cases, the applied researchers should not expect to obtain identical results from different software packages since the packages could be based on different estimators. In cases when the software packages show critical differences, the applied researcher should conduct a simulation study similar to the one described in this note to evaluate the accuracy of the different packages. Note however that even if all software packages show identical results, these results may still not be very accurate. One example is the case of uninformative sampling weights. Thus the applied researcher should always include sampling weights analysis as an essential part of their overall data analysis.

Stephen Du Toit communicated to the authors that the problems with the LISREL estimation are due to the LISREL implementation rather than the Pfeffermann et al. (1998) estimation method. A future release of the LISREL program implementing correctly the Pfeffermann et al. (1998) method yields results that are very close to the Mplus results.

8 Conclusion

In this article we described the MPML method which can used to estimate multilevel models with survey data. The estimator incorporates survey sampling features such as stratification, multistage sampling, cluster sampling, finite population sampling and unequal probability sampling at every sampling stage. The method is likelihood based and thus applies to multivariate outcomes from any parametric family of distributions, including for example the generalized linear models. In the simulation study described in this article the MPML estimator performed very well.

We also conducted simple simulation studies to compare the estimation methods available in Mplus and LISREL. We found substantial differences between the two software packages. In our simulations studies the results obtained in Mplus were more accurate than those obtained in LISREL.

The biggest challenge in estimating a two-level model with survey data remains the presence of within level sampling weights. While our study did not address this topic, detailed information can be found in Asparouhov (2006). There are a number of factors that can have a substantial impact on the quality of the estimation when within level sampling weights are present in the data. In order of importance these factors are the cluster sample size, the informativeness of the within level weights, the ICC (intra class correlation) and the UWE (unequal weighting effect).

In the past researchers frequently had to make a choice between estimating a multilevel model or estimating a single level model but incorporating the sampling design in the standard error estimation. The MPML estimator implemented in Mplus allows the researcher to combine these two techniques and thus conduct more accurate and informative analyses.

References

Asparouhov, T. (2004), “Weighting for unequal probability of selection in multilevel modeling,” *Mplus Web Note # 8.* http://statmodel.com/download/webnotes/MplusNote81.pdf

Asparouhov, T. (2005), "Sampling weights in latent variable modeling," *Structural Equation Modeling*, **12**, 411–434.

Asparouhov, T. (2006), "General Multilevel Modeling with Sampling Weights," *Communications in Statistics: Theory and Methods*, **35**, 439–460.

Asparouhov, T. and Muthen, B. (2005), " Multivariate Statistical Modeling with Survey Data," *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference.* http://www.fcsm.gov/events/papers05.html

Binder, D. A. (1983), "On the variance of asymptotically normal estimators from complex surveys," *Int. Statist. Rev.*, **51**, 279–292.

Chantala, K. Suchindran, C. (2006), "Adjusting for Unequal Selection Probability in Multilevel Models: A Comparison of Software Packages," *Proceedings of the American Statistical Association*, Seattle, WA: American Statistical Association.

Cochran, W. G.(1977), *Sampling Techniques*, John Wiley & Sons.

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

Korn, E. L., and Graubard B. I. (1999), *Analysis of Health Surveys*, Wiley.

Muthen, B. (1994), "Multilevel covariance structure analysis," *Sociological Methods & Research*, **22**, 376–398.

Muthen, L.K. and Muthen, B.O. (1998-2006), *Mplus User's Guide*, Third Edition. Los Angeles, CA: Muthen & Muthen.

LISREL Documetation (2005), "Analysis of Structural Equation Models for Continuous Random Variables in the Case of Complex Survey Data." http://www.ssicentral.com/lisrel/techdocs/compsem.pdf

Pfeffermann, D.; Skinner, C.J.; Holmes, D.J.; Goldstein, H.; Rasbash, J. (1998), "Weighting for unequal selection probabilities in multilevel models," Journal of the Royal Statistical Society, Series B, **60**, 23–56.

Rabe-Hesketh, S. and Skrondal, A. (2006), "Multilevel modeling of complex survey data," *Journal of the Royal Statistical Society, Series A*, **169**, 805–827.

Rao, J. N. K., & Thomas, D. R. (1989), "Chi-Square Tests for Contingency Table," *Analysis of Complex Surveys (eds. C.J.Skinner, D.Holt and T.M.F. Smith)*, 89–114, Wiley.

Research Triangle Institute (2002), *SUDAAN User Manual Release 8.0.*

Satorra, A., & Bentler, P.M. (1988), "Scaling Corrections for Chi-Square Statistics in Covariance Structure Analysis," *Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, 308–313.

Skinner, C. J. (1989), "Domain Means, Regression and Multivariate Analysis," *Analysis of Complex Surveys* (eds. C.J.Skinner, D.Holt and T.M.F. Smith), 59–87, Wiley.

Stapleton, L. (2002), "The Incorporation of Sample Weights Into Multilevel Structural Equation Models," *Structural Equation Modeling*, **9**, 475–502.

Stapleton, L. (2006), "An Assessment of Practical Solutions for Structural Equation Modeling with Complex Sample Data," *Structural Equation Modeling*, **13**, 28–58.

Yuan, K., & Bentler, P. M. (2000), "Three Likelihood-Based Methods for Mean and Covariance Structure Analysis With Nonnormal Missing Data," *Sociological Methodology*, **30**, 167–202.