GENERAL MULTILELEVEL MODELING WITH SAMPLING WEIGHTS

Tihomir Asparouhov
Muthen & Muthen
Los Angeles, California, USA
tihomir@statmodel.com

Key Words: multilevel pseudo maximum likelihood; sampling weights; multilevel models; multilevel mixture models; weights scaling; informative selection;

ABSTRACT

In this article we study the approximately unbiased multilevel pseudo maximum likelihood (MPML) estimation method for general multilevel modeling with sampling weights. We conduct a simulation study to determine the effect various factors have on the estimation method. The factors we included in this study are scaling method, size of clusters, invariance of selection, informativeness of selection, intraclass correlation and variability of standardized weights. The scaling method is an indicator of how the weights are normalized on each level. The invariance of the selection is an indicator of whether or not the same selection mechanism is applied across clusters. The informativeness of the selection is an indicator of how biased the selection is. We summarize our findings and recommend a multistage procedure based on the MPML method that can be used in practical applications.

1. INTRODUCTION

Multilevel models are frequently used to analyze data from cluster sampling designs. Such sampling designs however often use unequal probability of selection at the cluster level and at the individual level. Sampling weights are assigned at one or both levels to reflect these probabilities. If the sampling weights are ignored at either level the parameter estimates can be substantially biased.

Weighting for unequal probability of selection in single level models is a relatively well established procedure. The pseudo maximum likelihood (PML) method developed by Skinner

(1), following ideas of Binder (2), can estimate any single level model when the data is obtained by unequal probability sampling. The PML parameter estimates and their asymptotic variance/covariance estimates are consistent. For multilevel models however the situation is completely different. There is currently no well established general multilevel consistent estimation method. Several methods have been proposed recently in the literature for example in Graubard and Korn (3), Grilli and Pratesi (4), Korn and Graubard (5), Kovacevic and Rai (6), Pfeffermann et al. (7), Pfeffermann et al. (8) and Stapleton (9). However, the asymptotic properties of these estimators, when increasing the number of clusters to infinity but bounding the cluster sample size, are unknown. These estimators can produce biased parameter estimates and the size of the bias can be evaluated only through a limited simulation study. General comparison between the methods is not available. Many of the limited comparisons that are available are inconclusive to some degree and depend on the particular model and sampling mechanism. Simulation studies have clearly indicated that as the number of clusters and the cluster sample sizes increase the parameter bias can generally be eliminated. Estimation methods that possess this property are called approximately unbiased. Often however this property has been verified only through simulation studies and the exact conditions for it to hold are unclear. In addition the previously proposed methods apply only to certain multilevel models and parameters, and cannot be extended beyond the framework they are defined in.

In this article we introduce the multilevel pseudo maximum likelihood (MPML) estimation method for a general multilevel model which can be seen as a natural extension of the PML method as defined in Skinner (1) for the single level models. This method has been previously studied also in Grilli and Pratesi (4) in the context of multilevel probit regression. In this article we find exact conditions which guarantee that the parameter estimates are approximately unbiased and find the asymptotic covariance of the parameter estimates. For certain special models we obtain a closed form expression for the parameter estimates, which enable us to compare this method with other previously proposed methods. The main advantage of this method is that it can be applied to any general multilevel model just as

the PML method can be applied to single level models. The method is also flexible and it can be modified for different estimation problems. In Section 2 we provide more background information on unequal probability sampling in multilevel settings. In Section 3 we define the MPML method and describe the variance estimation for the parameter estimates. In Section 4 we clarify the concept of weights scaling, which has been the primary bias reduction tool for example in Pfeffermann et al. (7) and Stapleton (9), and we show that different scaling methods should be used depending on whether or not the sampling mechanism is of two different types: invariant and non-invariant. In Section 5 we develop a measure for the level of informativeness of the selection mechanism which can be used to evaluate the effect of different selection mechanisms on the estimation method and to compare different simulation studies and simulation studies with practical applications. The measure of informativeness can also be used in evaluating the need for incorporating the weights in the analysis. In Sections 6 and 7 we conduct simulation studies that evaluate the effect of several components on the MPML estimation method. These components are informative index, sample cluster size, invariance of sampling mechanism and scaling method. In Section 8 we evaluate the effect of intraclass correlation on the MPML estimates. In Section 9 we show that the MPML method produces biased estimates even for non-informative sampling and study the effect of the variability of the standardized weights on that bias. In Section 10 and 11 we conduct a simulation study on a multilevel logistic regression and a multilevel mixture model, which currently can be estimated only by the MPML method. In Section 12 we summarize our findings and outline a 6 step procedure that should be followed for proper usage of weights in multilevel models. In Appendix A we provide theoretical justification for the MPML method and underline the conditions needed for the approximate unbiasedness of the parameter estimates. In Appendix B we obtain a closed form expression for the balanced random intercept model and compare the MPML method to other previously proposed methods.

2. BACKGROUND

In a general multilevel model the observed variable in cluster $j = 1, ..., M$ of individual

$i = 1, ..., n_j$ is the vector y_{ij} and the level 2 random effect in cluster j is the vector η_j. The predictors on the individual level are denoted by x_{ij} and the predictors on the cluster level are denoted by x_j. A general multilevel model is specified by the density function of y_{ij} which we denote by $f(y_{ij}|x_{ij}, \eta_j, \theta_1)$ and the density function of η_j which we denote by $\phi(\eta_j|x_j, \theta_2)$. The parameters are denoted by θ_1 and θ_2. These density functions can belong to any parametric family of density functions or any parametric family of probability functions for the case when y_{ij} or η_j are discrete variables. The general multilevel modeling framework includes but is not limited to multilevel linear regression, multilevel logistic regression, multilevel probit regression, multilevel factor analysis and multilevel growth mixture models. A detailed description of a simple multilevel linear regression example is given in Appendix B.

Suppose that the data is sampled with unequal probability of selection on both levels. Suppose that the probability of selection for cluster j is p_j and $w_j = 1/p_j$. Suppose that the probability of selection for individual i in cluster j, given that cluster j is selected, is p_{ij} and the sampling weight is $w_{ij} = 1/p_{ij}$. Our goal is to estimate the parameters θ_1 and θ_2 by incorporating the sampling weights w_j and w_{ij} in the estimation method and thus eliminate the selection bias.

The intricacies of multilevel weighted analysis begin with the choice of the sampling weights. Frequently data sets are made available with weights prepared for computing means. Unfortunately these weights are not appropriate for multilevel models and can produce erroneous results if used with multilevel models.

The usual description of weighted two-level analysis includes weighting for unequal probability of selection at level 2 - the cluster level, weighting for unequal probability of selection for level 1 - the individual level, or both. However, because of the PML method developed by Skinner (1), the only case that requires new theoretical considerations is the weighting for unequal probability of selection for level 1 units. When weights are present at level 2 only, that is to say that independent units, namely clusters, have been sampled with unequal probability, we identify this framework as the framework of single level weighted modeling and methods available for single level weighted analysis can be applied. A two level model

with weights on level 2 can be presented as a multivariate single level model with weights, which can be estimated by the single level PML method. The model estimates will be consistent regardless of the size of the clusters. As an example, consider the linear growth model $Y_{it} = \alpha_i + \beta_i t + \varepsilon_{it}$ where Y_{it} is a normally distributed observed variable at time t for individual i. Here α_i and β_i are normally distributed random effects and ε_{it} is a zero-mean residual. This model can be viewed as a two level model. Suppose that for each individual there are 5 observations, given at times $X = (0, ..., 4)$. This model is equivalent to a single level, mean and variance model, with observed vector $Y_i = (Y_{i0}, ..., Y_{i4})$. This model is given by $Y_i = \mu + \varepsilon_i$, where the parameter μ are estimated under the constraint equation $\mu = E(\alpha_i) + E(\beta_i)X'$. Similar parameter constraint is needed for the variance of ε_i. The level 2 weight variable has the role of a single level weight variable and the estimation can be done by the single level PML technique. Simulation studies based on this approach are conducted in Asparouhov (10).

The situation is completely different when weights are present at level 1 because the unequal probability of selection is applied to dependent units and thus the assumptions of the single level methodology are violated. The MPML method defined in Section 3 applies to the general case of weighting on both levels. Our primary interest in multilevel weighted analysis however is in the case when the level 1 units are weighted for unequal probability of selection.

If the selection mechanism is not informative we should exclude the weights from the analysis. The estimates will remain consistent and in fact will be more precise. Including non-informative weights in the analysis can result in a substantial loss of efficiency. The PML parameter estimates, however, will remain consistent. This is not quite the case however for the MPML estimator. In Section 9 we conduct a simulation study which demonstrates that a relatively small large sample bias can arise in the estimation of multilevel models when the cluster sample sizes are small even if the weights are non-informative on level 1. Non-informative weights on level 2 are not a source of bias for the MPML estimates in general.

In practical applications it may not be possible to easily determine whether the selection

is informative or not. A general method for testing the informativeness of the weights is described in Pfeffermann (11) and the test can be used for multilevel models as well as single level models. A simpler but incomplete method is proposed in Section 5 based on the informative index. If the weights are determined to be non-informative they should not be used in the analysis. If any informativeness test is inconclusive, including the weights in the analysis is necessary. In this article we are generally concerned with informative selection mechanisms at level 1.

One of the key issues in the multilevel weighted estimation literature has been the fact that the parameter estimation are usually only approximately unbiased, i.e., they are unbiased for sufficiently large cluster sample size, but can be severely biased when the cluster sample size is small. Different scaling of the weights has been one of the focal points of the bias reducing techniques for example in Pfeffermann et al. (7) and Stapleton (9). There have been no theoretical results to support one scaling method over another. Scaling of the weights comprises of multiplying the weights by a scaling constant so that the sum of the weights is equal to some kind of characteristic of the sample, for example, the total sample size. In multilevel models the scaling modification methods scale the weights differently across clusters so that the total weight of the cluster is equal to some cluster characteristic. In single level modeling the scaling of the weights does not affect the PML estimator at all. In multilevel models that is not the case however and the ratio between the true cluster sample size and the total weight within the cluster is an important quantity since it affects for example the distribution of the level 2 random effects conditional on all observed data. The different scaling methods however may have different effect on different estimation techniques. If a scaling method performs well with the MPML approach proposed in Section 3, it will not necessarily perform well with other estimation techniques.

Another perspective that exposes the need for scaling the weights is the following. Let the weight for individual i in cluster j be w_{ij}. Typically w_{ij} is computed as $w_{ij} = 1/p_{ij}$ where p_{ij} is the probability that individual i in cluster j is included in the sample. Often however these probabilities are very small. Thus the resulting weights are very large, which is not very

practical and the weights are consequently rescaled. Alternatively the size of the population that is being sampled may not be known and consequently the exact probabilities of selection may not be known. Nevertheless unequal probability of selection could be implemented, for example a certain ethnicity could be oversampled at a given rate even when the size of the population is not known. In these cases the scale of the weights is undetermined and only the relative value of the weights has practical meaning. Methods that rely on the equations $w = 1/p$ are not applicable directly, while scale invariant methods are.

3. MULTILEVEL PSEUDO MAXIMUM LIKELIHOOD

Let the observed variable in cluster $j = 1, ..., M$ of individual $i = 1, ..., n_j$ be y_{ij} and the level 2 random effect in cluster j be η_j. Let the individual level covariates be x_{ij} and the cluster level covariates be x_j. Let the density function of y_{ij} be $f(y_{ij}|x_{ij}, \eta_j, \theta_1)$ and the density function of η_j be $\phi(\eta_j|x_j, \theta_2)$, where θ_1 and θ_2 are the parameters to be estimated. Let $w_j = 1/p_j$ and $w_{ij} = 1/p_{ij}$ be the sampling weights for the cluster and the individual level. We define the MPML estimates $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ as the parameters that maximize the weighted pseudo likelihood

$$l(\theta_1, \theta_2) = \prod_j \left(\int \Big(\prod_i f(y_{ij}|x_{ij}, \eta_j, \theta_1)^{w_{ij} s_{1j}} \Big) \phi(\eta_j|x_j, \theta_2) d\eta_j \right)^{w_j s_{2j}}, \tag{1}$$

where s_{1j} and s_{2j} are level 1 and level 2 scaling constants. Let $\hat{n}_j = \sum_i w_{ij} s_{1j}$ be sum of the scaled weights within cluster j. The effect of this value on the estimation is very similar to the effect of the cluster sample size on the maximum likelihood estimation without sampling weights. Within the EM maximization algorithm for example $\hat{n}_j$ is an indicator of the extent to which η_j is determined by the observed data and is inversely proportional to the variance of η_j conditional on all observed data. Thus the scale factor s_{1j} can be used to balance this effect. The second level scaling s_{2j} can be used to counter balance the first level scaling s_{1j}. We will be interested primarily in the following choices $s_{2j} = 1$ or $s_{2j} = 1/s_{1j}$. The MPML estimates are approximately unbiased if

1. n_j and $\hat{n}_j$ are sufficiently large,
2. s_{2j} and η_j are conditionally independent given all model covariates,

3. $\hat{n}_j/n_j$ and η_j are conditionally independent given all model covariates.

The proof of the approximate unbiasedness can be found in the Appendix A. A closed form solution for the MPML estimates for the balanced random intercept model is derived in Appendix B. There is a considerable flexibility in the definition of MPML because the scaling constants are relatively unrestricted. A natural choice for s_{1j} is $s_{1j} = n_j/\sum_i w_{ij}$. In that case $\hat{n}_j = n_j$ and the third condition above is automatically satisfied even when η_j and n_j are not independent.

Denote by $L = log(l)$ the weighted log-likelihood and by $L_j = log(l_j)$ the weighted log-likelihood of the j-th cluster, where l_j is

$$l_j = \int \Big(\prod_i f(y_{ij}|x_{ij}, \eta_j, \theta_1)^{w_{ij}s_{1j}} \Big) \phi(\eta_j|x_j, \theta_2) d\eta_j. \tag{2}$$

The asymptotic covariance matrix given by standard asymptotic theory is

$$(L'')^{-1} \Big(\sum_j (s_{2j}w_j)^2 L'_j L'^T_j \Big) (L'')^{-1}, \tag{3}$$

where $'$ and $''$ refer to the first and the second derivative of the log-likelihoods.

The MPML just as the PML is defined as a general estimator not connected to any optimization algorithm. Virtually any ML optimization algorithm can be adapted to include the weights and maximize the MPML objective function, for example the EM-algorithm, accelerated EM-algorithm, the Quasi-Newton algorithm or the Fisher scoring algorithm. In the extensive simulation study described in this article, all of which were done with Mplus 3 (www.statmodel.com; Muthen & Muthen (13)) using the accelerated EM algorithm, we encountered almost no convergence problems at all.

4. SCALING

In this section we discuss in detail the concept of scaling of the weights. We assume that the probability of selection has only a relative meaning. If $p_{i_1 j} : p_{i_2 j} = 2$ we interpret this as indication that individuals similar to individual i_1 are oversampled at a rate of 2:1 in comparison to the individuals similar to individual i_2. Therefore the weights w_{ij} should be standardized by the scale factors s_{1j} to some meaningful values. In choosing the proper

standardization several considerations should be made. First we need to understand whether or not the ratio between the weights of individuals from different clusters is a meaningful quantity. That is, we have to know whether $w_{i_1j_1} : w_{i_2j_2}$ can be interpreted in the usual sense of oversampling. This would be so if the same sampling mechanism is used across the clusters, i.e., no cluster random effect has an influence on the sampling. If different sampling mechanisms have been used, depending on cluster specific random effects, then the weights within each cluster could be on different scales and direct comparison between the weights can be detrimental. If the same sampling mechanism has been used the ratio between the weights across clusters is a meaningful quantity we need to choose scaling that takes advantage of that quantity and the relative information is not lost.

We say that the sampling mechanism is invariant across clusters if the sampling weight on the individual level w_{ij} and the cluster random effects η_j are conditionally independent given all model covariates x_{ij} and x_j. In practice a good understanding of the selection mechanism would be sufficient to determine if invariance holds. For example, oversampling adolescents to adults in rates 3:1 is invariant across clusters. Oversampling the subpopulation with the lowest y values within each cluster is invariant. However, oversampling all individuals with y values below a certain threshold value is not invariant since the sampling weights depend on the cluster average.

Note that scaling concerns only level 1 weights. The MPML estimates are independent of the scale of the level 2 weights, just as in a single level model the PML estimates are independent on the scale of the weights. In fact the scale constants s_{2j} are not needed to standardize the level 2 weights but to possibly counter the standardization on level 1 and recover any information that may be lost after the level 1 standardization. Of course s_{2j} could not be considered as a scale factor for the level two weights also because they depend on j. The scaling constants s_{2j} also do not need any standardization, that is the MPML estimates depend only on the relative values of s_{2j}. In choosing scaling we also need to make sure that conditions 1, 2 and 3 in Section 3 are satisfied so that the approximate unbiasedness is guaranteed at least for large cluster sizes.

We consider 6 different weighting methods. Since $\hat{n}_j/s_{1j} = \sum_i w_{ij}$ the definition of the scale constants s_{1j} can be expressed as a definition of $\hat{n}_j$. Also define the effective sample size n_{0j} as in Potthoff et al. (12), $n_{0j} = (\sum_j w_{ij})^2/\sum_j w_{ij}^2$. The weighting methods we are interested in are

Method A. $\hat{n}_j = n_j$ and $s_{2j} = 1$

Method AI. $\hat{n}_j = n_j$ and $s_{2j} = 1/s_{1j}$

Method B. $\hat{n}_j = n_{0j}$ and $s_{2j} = 1$

Method BI. $\hat{n}_j = n_{0j}$ and $s_{2j} = 1/s_{1j}$

Method C. $s_{1j} = \sum_j n_j/\sum_{ij} w_{ij}$ and $s_{2j} = 1$

Method D. Unweighted analysis

Method E. Unscaled weighted analysis, $s_{1j} = 1$ and $s_{2j} = 1$

The scaling methods A and B have been proposed in the literature already in Grilli and Pratesi (4), Potthoff et al. (12), Pfeffermann et al. (7) and Stapleton (9). For scaling method A conditions 2 and 3 in Section 3 are always satisfied and thus only condition 1 is needed for approximate unbiasedness. The scaling methods AI and BI are based on scaling methods A and B but also include a level two offset scaling. The index letter I in the name of the scaling methods AI and BI is to indicate that the methods are appropriate only for invariant selection mechanisms. Indeed if the selection is not invariant η_j would influence s_{1j} and in turn s_{2j} which would be a violation of condition 2 in Section 3. For invariant selection mechanisms these methods are expected to perform better since they would use all the information including the ratio between weights across clusters. That is because the product of the level 1 and level 2 weights is unchanged by these scaling methods. We also demonstrate these facts in the simulation studies in Sections 6 and 7 below. Scaling method C is the scaling method that is traditionally used with mean estimation. This method has constant scaling across clusters and $\sum_j \hat{n}_j = \sum_j n_j$.

If simple random sampling is used at both sampling stages, then $w_j = w_k$ and $w_{ij} = w_{lj}$. As a result, scaling methods A and B are identical and the MPML estimates are equal to the ML estimates. Therefore these methods are consistent with SRS even when the cluster

sample sizes are bounded. This is a very important property from practical point of view because frequently sampling mechanisms only mildly deviate from SRS. Scaling methods AI and BI are also equivalent under SRS however they are different from the ML estimates for unbalanced designs. Only condition 2 in Section 3 is needed for the consistency of the MPML estimates with AI and BI scaling under SRS, however the estimates will be less efficient than the ML estimates. When the design is balanced, i.e., cluster sizes and cluster sample sizes are constant, all 5 scaled methods A, AI, B, BI and C become equivalent to the ML method (method D) and the parameter estimates are consistent and asymptotically efficient even when the cluster sample sizes are bounded. This is not however the case for method E which remains biased unless the cluster sample sizes n_j are large. We illustrate this fact in Appendix B with the multilevel random intercept model and show exactly what that bias is. The fact that method E is biased even in the most simple case of balanced design with simple random sampling shows that this method is of little general interest and we will not include this method in our simulation studies.

5. INFORMATIVE INDEX

The quality of the MPML estimates is driven primarily by two factors, the sample size of the clusters and the degree of informativeness of the selection mechanism. While the sample size is an explicit variable, the informativeness is not directly measurable. Such measurement is needed however to study the dependence of the quality of the MPML estimates on the informativeness. Pfeffermann (11) constructs a test statistic that allows us to determine whether the selection mechanism is informative. If $\hat{\theta}_w$ and $\hat{\theta}_0$ are the parameter estimates of the weighted and unweighted analysis respectively, and $\hat{V}(\hat{\theta}_w)$ and $\hat{V}(\hat{\theta}_0)$ are their variance estimates,

$$I = (\hat{\theta}_w - \hat{\theta}_0)^T [\hat{V}(\hat{\theta}_w) - \hat{V}(\hat{\theta}_0)]^{-1} (\hat{\theta}_w - \hat{\theta}_0) \sim \chi_{(p)}, \tag{4}$$

has approximately a chi-square distribution with $p = \dim(\theta)$ degrees of freedom. The value of I is a clear measurement of the informativeness of the selection. We now define a similar test statistic for a specific variable Y in the model. Suppose that $\hat{\mu}_w$ is the weighted estimated mean of Y and $\hat{\mu}_0$ the unweighted estimated mean of Y. Let $\hat{\sigma}_w^2$ and $\hat{\sigma}_0^2$ be the estimated

variance for $\hat{\mu}_w$ and $\hat{\mu}_0$ respectively. The informativeness of the selection mechanism for variable Y can then be measured by the T statistic

$$I_1(Y) = \frac{\hat{\mu}_w - \hat{\mu}_0}{\sqrt{\hat{\sigma}_w^2 - \hat{\sigma}_0^2}}. \tag{5}$$

Large absolute values would indicate that the selection is informative for Y. The values of $I_1(Y)$ are easy to compute once the weighted and unweighted analysis have been completed. As the sample size increases, the variance of the mean estimates decrease, and all selection mechanisms produce high I_1 values. Thus it is not possible to form recommendations based on the value of I_1 that apply to any sample size. Another problem with I_1 is that it is not defined when $\hat{\sigma}_w < \hat{\sigma}_0$. This happens quite often when the sample size is not large or when the weights are not sufficiently different across individuals. An informativeness measure that is independent of the sample size is

$$I_2(Y) = \frac{|\hat{\mu}_w - \hat{\mu}_0|}{\sqrt{v_0}}, \tag{6}$$

where v_0 is the unweighted estimate of the variance of Y. This informativeness measure depends however on the scaling and the cluster sample size since the MPML method depends on the scaling when the cluster size is small. This leads us to define

$$I_3(Y) = \frac{|\mu - \hat{\mu}_0|}{\sqrt{v_0}}, \tag{7}$$

where μ is the true mean of Y. This informativeness measure would be relatively independent of the cluster size and the sample size, however in practice we can only estimate $I_2(Y)$, and approximate $I_3(Y)$ by $I_2(Y)$. This approximation will be sufficient as long as there is no substantial bias in the mean parameter estimate. In a simulation study of course I_3 can be easily computed since the true parameter value μ is known.

The mean parameters are the parameters that usually are the most sensitive to selection bias. Thus evaluating the informative index for all dependent variables will generally be sufficient to determine whether the sampling and the resulting weights are informative. If the sampling weights appear to be non-informative or very slightly informative it is very

likely that the reduction in the bias due to the weighting of the analysis will be overwhelmed by the increase in the variance of the parameter estimates. Therefore including the weights in the analysis will in effect only increase the mean squared error. It is recommended that the weights are not used at all in such cases. The informative index is a very easy to compute and practical tool, however it is not a universal tool for detecting informative selections. For example if the weights are only informative for variance and covariance parameters the informative index would not detect that. In practice, however, such situations are probably very rare.

6. INVARIANT SELECTION

We conduct a simulation study on the following model

$$y_{ij} = \mu + \eta_j + \varepsilon_{ij}, \tag{8}$$

where $\mu = 0.5$, and η_j and ε_{ij} are zero mean normal random variables with variance $\psi = 0.5$ and $\theta = 2$ respectively. The selection model is defined by $P(I_{ij} = 1) = 1/(1 + e^{-\varepsilon_{ij}/\alpha})$, i.e., the probability of inclusion is dependent only on the level 1 residual and is invariant across clusters. We use 3 different values for α to achieve 3 different levels of informativeness of the selection. For $\alpha = 1, 2, 3$ the approximate values for $I_3(Y)$ are 0.5, 0.3 and 0.2 respectively. Within each simulation the cluster size is constant. We use three different values for the cluster size $n_j = 5, 20$, and 100. Each of the analyses is replicated 500 times. In all cases we used 100 cluster units. The results of the simulation are presented in Table 1. The table contains the absolute bias and the percentage of times the true parameter was covered by the 95% confidence intervals in the parenthesis. All estimates for the θ parameter were negatively biased. All estimates for the ψ parameter were positively biased. All estimates for the μ parameter were positively biased except for method BI which had negatively biased μ estimates. We make the following conclusions based on the results in Table 1.

- The unweighted parameter estimates are substantially biased. The bias increase as the informative index I_3 increases but is unaffected by the cluster size n_j.

- The cluster sample size and the informative index I_3 affect the quality of the results for all 5 weighting methods. The most difficult case is when the cluster size n_j is small and the selection mechanism is strongly informative. Rows 1, 2 and 4 are of this type. None of the scaled weighted methods produced satisfactory results for all parameters when $n_j = 5$ and $I_3 = 0.3$ or 0.5 or when $n_j = 20$ and $I_3 = 0.5$, while different parameters were estimated well by different scaling methods. In the remaining 6 cases, i.e., the cases when either the cluster size is sufficiently large or the informative index I_3 is small, the asymptotics appear to become valid to a varying extent and all 5 scaled weighted methods appear to perform reasonably well.

- Choosing a scaling method is not an easy task. The best method for estimating μ is AI, the best method for θ is BI and the best method for ψ is D in our simulation study. Nevertheless some conclusions are quite clear. Methods AI and BI, which are designed to take advantage of the fact that the selection is invariant outperform their non-invariant counterparts A and B, except in the first row in the ψ subtable, which however can not outweigh all other cases. As expected also methods A and AI perform somewhat better than B and BI for the mean parameters and worse for variance covariance parameters.

- It is hard to recommend one scaling method for all situations. Nevertheless we single out method AI. This method performs well when the informativeness is not very strong or the cluster sizes are not small. If the cluster sizes are small and the informative index is large however we would recommend that a detailed analysis is undertaken, using several different scaling methods and perhaps a single level model is used as a final model.

- It is hard to determine a specific threshold level for I_3 and n_j that would guarantee unbiased results in general. From what we see in this simulation study, we conclude that if $I_3 < 0.2$, the AI method will work sufficiently well with any sample size. If

Table 1: Absolute Parameter Bias (Coverage). Invariant Selection.

μ parameter

n_j	I_3	A	AI	B	BI	C	D
5	0.5	0.27(27)	0.00(95)	0.31(13)	0.13(93)	0.21(50)	0.73(0)
5	0.3	0.12(79)	0.00(94)	0.13(75)	0.02(95)	0.08(89)	0.45(1)
5	0.2	0.07(89)	0.00(96)	0.08(88)	0.00(95)	0.04(92)	0.32(10)
20	0.5	0.10(78)	0.00(93)	0.12(67)	0.15(86)	0.09(82)	0.73(0)
20	0.3	0.03(95)	0.00(96)	0.10(80)	0.02(96)	0.02(96)	0.45(0)
20	0.2	0.02(95)	0.00(96)	0.02(95)	0.00(96)	0.02(95)	0.32(1)
100	0.5	0.03(94)	0.00(94)	0.03(93)	0.09(87)	0.02(93)	0.73(0)
100	0.3	0.00(93)	0.00(92)	0.01(93)	0.01(92)	0.00(93)	0.45(0)
100	0.2	0.01(97)	0.00(96)	0.01(97)	0.00(96)	0.01(97)	0.32(0)

θ parameter

5	0.5	0.62(0)	0.47(3)	0.65(0)	0.30(82)	0.49(0)	0.51(1)
5	0.3	0.22(66)	0.14(78)	0.26(59)	0.09(91)	0.14(73)	0.19(66)
5	0.2	0.09(87)	0.05(91)	0.11(87)	0.03(95)	0.05(90)	0.08(88)
20	0.5	0.30(3)	0.21(36)	0.42(1)	0.15(94)	0.21(22)	0.53(0)
20	0.3	0.07(83)	0.04(90)	0.10(80)	0.03(97)	0.04(90)	0.20(8)
20	0.2	0.03(92)	0.02(95)	0.04(93)	0.01(97)	0.02(94)	0.10(60)
100	0.5	0.09(58)	0.06(78)	0.19(36)	0.04(99)	0.06(74)	0.53(0)
100	0.3	0.01(92)	0.01(93)	0.02(93)	0.00(97)	0.01(93)	0.20(0)
100	0.2	0.00(95)	0.00(95)	0.00(96)	0.00(96)	0.00(95)	0.10(0)

ψ parameter

5	0.5	0.29(58)	0.45(55)	0.17(80)	0.34(67)	0.30(72)	0.02(92)
5	0.3	0.11(92)	0.12(92)	0.08(93)	0.08(93)	0.11(93)	0.02(92)
5	0.2	0.05(94)	0.04(93)	0.03(94)	0.02(93)	0.04(94)	0.02(92)
20	0.5	0.15(77)	0.21(75)	0.09(88)	0.23(78)	0.15(79)	0.00(93)
20	0.3	0.03(95)	0.03(96)	0.02(95)	0.02(95)	0.03(95)	0.01(93)
20	0.2	0.01(95)	0.01(96)	0.01(95)	0.01(95)	0.01(95)	0.01(94)
100	0.5	0.04(94)	0.05(93)	0.02(93)	0.07(91)	0.04(93)	0.01(92)
100	0.3	0.01(94)	0.01(94)	0.00(93)	0.00(93)	0.01(94)	0.01(91)
100	0.2	0.00(93)	0.00(92)	0.00(93)	0.00(92)	0.00(93)	0.00(92)

$0.2 < I_3 < 0.3$ a cluster sample size of at least 10 is needed. If $0.3 < I_3$ we would recommend using the AI method only with cluster size 35-40 and above.

- This simulation study shows that the intraclass correlation (ICC=$\psi/(\psi+\theta)$) is overestimated in general by all methods to a varying degree when some selection bias remains.

- Method AI seems to produce unbiased estimates for the mean parameter with any cluster size and informative index.

7. NON-INVARIANT SELECTION

In this section we use the same setup as in the previous section. We only change the selection model to $P(I_{ij} = 1) = 1/(1 + e^{-y_{ij}/\alpha})$, for $\alpha = 1, 2, 3$. The informative index I_3 is slightly lower now but approximately unchanged. The selection is clearly non-invariant, because η_j influences the selection. Thus methods AI and BI are not expected to perform well even asymptotically. Table 2 contains the results of this simulation study. All estimates for the θ parameter were negatively biased. All estimates for the ψ parameter were positively biased except for method D which had negatively biased ψ estimates. All estimates for the μ parameter were positively biased except for method AI and BI which had negatively biased μ estimates. The following conclusions can be made from these results.

- Methods AI and BI are not suitable for this situation, the fact that they violate condition 2 from Section 3 results in biased estimates for the μ parameter even when the cluster size is large. For methods AI and BI the selection bias for the μ parameter fails to decrease as n_j increases.

- Methods A, B and C perform well when the cluster size is large or the informative index is small. The case of small cluster size and large informative index is again a difficult one. There is no large difference between the three methods. However, method C slightly outperforms method A, which slightly outperforms method B.

Table 2: Absolute Parameter Bias (Coverage). Non-Invariant Selection.

μ parameter

n_j	I_3	A	AI	B	BI	C	D
5	0.5	0.23(35)	0.15(87)	0.28(15)	0.31(68)	0.13(74)	0.61(0)
5	0.3	0.10(83)	0.11(83)	0.11(78)	0.13(79)	0.02(94)	0.40(0)
5	0.2	0.07(89)	0.07(89)	0.07(89)	0.08(88)	0.01(94)	0.29(11)
20	0.5	0.08(83)	0.16(70)	0.11(70)	0.39(29)	0.05(89)	0.61(0)
20	0.3	0.03(91)	0.10(77)	0.04(90)	0.13(67)	0.01(92)	0.40(0)
20	0.2	0.01(93)	0.08(83)	0.01(93)	0.09(78)	0.00(93)	0.29(4)
100	0.5	0.02(95)	0.16(52)	0.03(94)	0.39(9)	0.01(95)	0.61(0)
100	0.3	0.01(92)	0.10(72)	0.01(92)	0.13(61)	0.00(92)	0.40(0)
100	0.2	0.01(95)	0.07(85)	0.01(95)	0.08(83)	0.01(95)	0.30(1)

θ parameter

5	0.5	0.52(0)	0.42(8)	0.54(3)	0.27(80)	0.45(2)	0.47(3)
5	0.3	0.19(64)	0.13(81)	0.23(64)	0.08(92)	0.14(75)	0.19(66)
5	0.2	0.10(84)	0.07(90)	0.12(84)	0.04(92)	0.07(88)	0.10(83)
20	0.5	0.24(13)	0.18(47)	0.32(9)	0.12(96)	0.18(35)	0.48(0)
20	0.3	0.06(84)	0.04(89)	0.09(82)	0.03(95)	0.04(88)	0.19(9)
20	0.2	0.03(90)	0.02(91)	0.04(89)	0.01(93)	0.02(91)	0.10(65)
100	0.5	0.07(66)	0.05(78)	0.15(50)	0.04(98)	0.05(74)	0.48(0)
100	0.3	0.01(92)	0.01(92)	0.02(93)	0.00(97)	0.01(91)	0.19(0)
100	0.2	0.00(95)	0.00(95)	0.01(96)	0.00(97)	0.00(95)	0.10(5)

ψ parameter

5	0.5	0.15(91)	0.42(75)	0.01(93)	0.33(68)	0.20(93)	0.21(32)
5	0.3	0.08(93)	0.13(92)	0.04(95)	0.09(93)	0.09(95)	0.09(78)
5	0.2	0.04(94)	0.05(94)	0.02(94)	0.04(94)	0.04(96)	0.06(87)
20	0.5	0.08(93)	0.20(86)	0.00(91)	0.26(73)	0.08(94)	0.22(5)
20	0.3	0.02(94)	0.04(94)	0.01(94)	0.04(93)	0.02(95)	0.10(63)
20	0.2	0.01(94)	0.01(94)	0.00(94)	0.01(94)	0.01(94)	0.05(83)
100	0.5	0.02(93)	0.07(92)	0.00(90)	0.14(84)	0.02(93)	0.21(2)
100	0.3	0.00(92)	0.01(92)	0.00(92)	0.02(93)	0.00(92)	0.10(60)
100	0.2	0.00(93)	0.01(94)	0.00(93)	0.01(93)	0.00(93)	0.05(82)

8. INTRACLASS CORRELATION

Kovacevic and Rai (6) discovered that the intraclass correlation (ICC=$\psi/(\psi+\theta)$) affects the performance of multilevel weighted estimation. For a different estimator they report that the bias of the estimates increases as ICC decreases. This is a very important finding because in practical applications ICC is typically small and that could be a source of substantial bias in multilevel weighted analysis. In this section we evaluate through a simulation study the effect of ICC on the MPML parameter estimates under different levels of informative selections and cluster sample sizes. We use the model described in the previous section and the same selection mechanism $P(I_{ij}=1)=1/(1+e^{-y_{ij}/\alpha})$. To achieve different ICC we use different values for ψ and θ while keeping constant the total variance of y, $\psi+\theta=2.5$. We simulate 5 different ICC levels 0.50, 0.20, 0.10, 0.05 and 0.01. These values are obtained by setting ψ to be 1.25, 0.5, 0.25, 0.125 and 0.025 respectively, while the value of θ is $2.5-\psi$. Since the selection mechanism is non-invariant we use scaling method A only. Note also that the simulation study in the previous section has ICC value of 0.20. The results of the simulation are presented in Table 3.

- We clearly see that for all parameters the bias increases as the ICC decreases which confirms Kovacevic and Rai (6) finding. The intuitive explanation for why this occurs is that as the ICC decreases the estimation on the individual level becomes more influential, but that is exactly where the weakness of the weighted estimation is. Alternative explanation is that when ICC converges to 1 then level 1 variation converges to zero and the model can be approximated by a level 2 model only, i.e., a single level model. The MPML estimator can then be approximated with the PML estimator which is consistent and thus the bias decreases as ICC increases.

- We also see that the effect of ICC is substantial in rows 1, 2 and 4. This means that the effect of ICC is more pronounced when the cluster sizes are small and selection is very informative.

- The informativeness index I_3 in this simulation study depends on both ICC and the

Table 3: Absolute Parameter Bias (Coverage) For Various ICC Levels, Scaling Method A.

μ parameter

n_j	α	icc 0.50	icc 0.20	icc 0.10	icc 0.05	icc 0.01
5	1	0.14(75)	0.23(35)	0.25(19)	0.27(14)	0.28(12)
5	2	0.07(94)	0.10(83)	0.11(76)	0.12(69)	0.13(62)
5	3	0.03(92)	0.07(89)	0.07(85)	0.08(84)	0.08(79)
20	1	0.04(94)	0.08(83)	0.10(71)	0.10(67)	0.10(51)
20	2	0.01(94)	0.03(91)	0.03(90)	0.03(89)	0.04(84)
20	3	0.03(93)	0.01(93)	0.02(92)	0.02(92)	0.02(90)
100	1	0.01(94)	0.02(95)	0.02(93)	0.03(89)	0.03(83)
100	2	0.00(93)	0.01(92)	0.01(96)	0.01(95)	0.01(92)
100	3	0.01(95)	0.01(95)	0.00(93)	0.01(95)	0.01(95)

θ parameter

5	1	0.24(12)	0.52(0)	0.62(1)	0.68(0)	0.73(0)
5	2	0.08(80)	0.19(64)	0.24(60)	0.27(53)	0.29(51)
5	3	0.03(91)	0.10(84)	0.12(83)	0.13(84)	0.14(80)
20	1	0.09(46)	0.24(13)	0.31(7)	0.34(4)	0.36(4)
20	2	0.02(88)	0.06(84)	0.08(81)	0.08(80)	0.09(81)
20	3	0.01(94)	0.03(90)	0.03(92)	0.04(92)	0.04(92)
100	1	0.03(82)	0.07(66)	0.10(58)	0.11(54)	0.13(49)
100	2	0.00(94)	0.01(92)	0.02(91)	0.02(91)	0.02(91)
100	3	0.00(95)	0.00(95)	0.01(93)	0.00(94)	0.01(91)

ψ parameter

5	1	0.03(94)	0.15(91)	0.22(57)	0.28(30)	0.32(10)
5	2	0.01(92)	0.08(93)	0.11(88)	0.14(79)	0.16(59)
5	3	0.00(93)	0.04(94)	0.06(93)	0.07(93)	0.10(87)
20	1	0.00(90)	0.08(93)	0.12(68)	0.16(29)	0.17(4)
20	2	0.01(93)	0.02(94)	0.04(92)	0.04(88)	0.05(56)
20	3	0.01(94)	0.01(94)	0.02(93)	0.02(93)	0.02(90)
100	1	0.01(92)	0.02(93)	0.04(92)	0.05(69)	0.06(7)
100	2	0.02(93)	0.00(92)	0.01(95)	0.01(96)	0.01(78)
100	3	0.01(92)	0.00(93)	0.00(96)	0.00(95)	0.00(93)

Table 4: Absolute Bias (Coverage) with Non-Informative Selection, Scaling Method AI

	n_j	ICC	$n_0/n = 0.95$	$n_0/n = 0.91$	$n_0/n = 0.78$	$n_0/n = 0.64$
θ	5	0.20	0.02(93)	0.04(92)	0.12(84)	0.25(60)
θ	10	0.20	0.01(94)	0.02(93)	0.06(89)	0.12(77)
θ	20	0.20	0.00(95)	0.01(95)	0.03(93)	0.05(88)
θ	5	0.05	0.03(92)	0.06(92)	0.14(84)	0.30(60)
θ	10	0.05	0.01(94)	0.02(93)	0.06(89)	0.14(77)
θ	20	0.05	0.00(95)	0.01(95)	0.03(93)	0.05(88)
ψ	5	0.20	0.01(92)	0.03(92)	0.10(88)	0.23(73)
ψ	10	0.20	0.00(92)	0.01(93)	0.04(94)	0.10(90)
ψ	20	0.20	0.01(92)	0.00(92)	0.02(92)	0.04(92)
ψ	5	0.05	0.03(97)	0.05(95)	0.13(84)	0.28(36)
ψ	10	0.05	0.01(93)	0.02(92)	0.06(87)	0.13(53)
ψ	20	0.05	0.00(92)	0.01(92)	0.03(92)	0.06(74)

selection model parameter α. For example when $n_j = 5$, the selection index I_3 is 0.31, 0.45, 0.50, 0.52 and 0.53 for the 5 different ICC values in Table 3 given in the same order. Small ICC values are thus generally associated with more informative selection as well. One could presume here that the bias increases simply because the selection becomes more informative as ICC decreases, however that is not entirely the case. In the next section we demonstrate that the bias increases when ICC decreases even when I_3 is held constant.

9. NON-INFORMATIVE SELECTION

In this section we examine the behavior of the MPML estimates under non-informative sampling. In the Sections 6 and 7 we observed that as the informativeness of the selection decreases the bias of the estimates decreases. It is important to know whether this bias decreases to 0 while preserving the distribution of w. Denote by $v_w = Var(w_{ij})/(E(w_{ij}))^2$ the relative variance of the weights. This quantity can be interpreted as the variance of the weights which are standardized as to have an average of 1 (see scaling method C). It can

be interpreted also as $v_w = (n - n_0)/n_0$ where n is the sample size and n_0 is the effective sample size $n_0 = (\sum_{ij} w_{ij})^2 / \sum_{ij} w_{ij}^2$. If the sampling is simple random sampling $v_w = 0$ and as v_w increases the weights are more disproportionate and the weighted estimation in general becomes less efficient. Here we conduct a simulation which shows also that as v_w increases not only the mean squared error of the estimator increases but also the bias of the MPML estimates increases even when the weights are non-informative. This bias decreases as the cluster sample size increases.

The model we use for this simulation is the same as the model used in Section 6 and the selection model is $P(I_{ij} = 1) = 1/(1 + e^{-\xi_{ij}})$ where ξ_{ij} is a zero mean normal random variable independent of y_{ij}. This selection is non-informative. Let $\alpha = Var(\xi_{ij})$. By varying the values of α we obtain different values of v_w. The values $\alpha = 2/9, 1/2$, and 2 give the same distribution of the weights as the once used in Section 6. We also simulate data with $\alpha = 10$. For $\alpha = 2/9, 1/2, 2, 10$, the corresponding values for v_w and $n_0 : n$ are $v_w = 0.05, 0.10, 0.28, 0.56$ and $n_0 : n = 0.95, 0.91, 0.78, 0.64$. We use two different ICC values 0.20 and 0.05 generated as in Section 8. Since the sampling is invariant we use AI scaling. Table 4 shows the bias of the MPML estimate for θ and ψ. The estimates for μ are unbiased and are not reported. We make the following conclusions from these simulation results.

- Even if the weights are not informative the MPML estimates can be biased, although this bias disappears as the cluster sample sizes increase or the relative variance v_w of the weights decreases. This bias can be substantial particularly when the cluster sample size is as small as 5 and the effective sample size (based on the level one weights) is more than 20% smaller than the actual sample size.

- By comparing rows 1-3 and 7-9 in Table 4 with the second column of Table 1 we see that around 1/4 of the small cluster sample size bias of the MPML estimates is due to the variation in the weights and does not originate in the informativeness of the weights.

- Decrease in ICC leads to an increase in the parameter bias. Since $I_3 = 0$, the increase

Table 5: Bias (Coverage) in Multilevel Logistic Regression, Scaling Method A.

n_j	s	I_3	μ_α	μ_β	σ_α	σ_β	ρ
5	1/3	0.57	−0.06(89)	0.04(95)	0.39(91)	0.32(96)	0.03(95)
20	1/3	0.57	−0.04(78)	0.01(93)	0.06(96)	0.07(95)	−0.01(95)
5	1/2	0.35	−0.04(92)	0.00(94)	0.09(95)	0.10(92)	−0.03(95)
20	1/2	0.35	−0.03(92)	0.00(94)	0.00(92)	0.02(92)	0.00(94)

is due purely to the decrease in ICC and not to the level of informativeness.

10. MULTILEVEL LOGISTIC REGRESSION

In this section we illustrate the MPML estimator with a simple multilevel logistic regression. Let u_{ij} be a binary outcome variable and x_{ij} be a predictor variable with standard normal distribution. We consider the multilevel logistic regression defined by

$$P(u_{ij} = 1) = \frac{1}{1 + e^{-(\alpha_j + \beta_j x_{ij})}} \tag{9}$$

where α_j and β_j are normally distributed random effects with means $\mu_\alpha = 0.5$, $\mu_\beta = 0.5$ variances $\sigma_\alpha = 0.5$, $\sigma_\beta = 0.5$ and covariance $\rho = 0.25$. The selection mechanism is defined by $P(I_{ij} = 1|u_{ij} = 1) = s < 1$ and $P(I_{ij} = 1|u_{ij} = 0) = 1$. This selection mechanism oversamples zero outcomes at a rate of $1 : s$. When $s = 1/2$ and $s = 1/3$ the oversampling rates are 2:1 and 3:1 respectively. This selection mechanism is not invariant because the weight variable and the random effects are not independent. We use scaling method A. The computations are performed with Mplus 3 (www.statmodel.com; Muthen & Muthen (13)) and are based on adaptive numerical integration. The results of the simulation study are presented in Table 5 and confirm our previous findings. The bias increases as the sample size decreases and the informativeness increases. The performance of the MPML method is satisfactory as long as either the sample size is not small or the informativeness is not strong.

11. MULTILEVEL MIXTURES

Table 6: Selection Bias (Coverage) in Multilevel Mixtures, Scaling Method AI.

n_j	μ	β	θ	ψ
5	−0.01(94)	0.01(90)	−0.18(85)	0.15(91)
10	0.00(92)	0.02(92)	−0.11(92)	0.07(92)
20	0.00(94)	0.01(92)	−0.05(96)	0.05(93)

In this section we conduct a simulation study based on the following model

$$y_{ij} = \mu + \beta x_{ij} + \eta_j + \varepsilon_{ij}, \tag{10}$$

where x_{ij} is a binary covariate taking values 0 and 1 equally likely and where x_{ij} has missing values resulting in a finite mixture model. As in the previous sections we assume that η_j and ε_{ij} are normally distributed random variables with variance θ and ψ. The parameter values we use in the simulation are as follows $\beta = 1$, $\theta = 2$, $\psi = 0.5$ and $\mu = 0$. The probability that x_{ij} is missing is $1/(1 + e^{(y_{ij}-0.75)/4})$. Unequal probability of selection is induced by the following inclusion model $P(I_{ij} = 1) = 1/(1 + e^{-\varepsilon_{ij}/2})$ which produces an informative index $I_3(y) = 0.27$. We use three different values for the cluster size $n_j = 5, 10$, and 20. Each of the analysis is replicated 500 times. In all cases we used 100 cluster units.

This model is only a slight modification of the model considered in Sections 6 and 7 and usually its estimation would not be very different. That is not the case however if the covariate x_{ij} has missing values. When x_{ij} is missing the conditional distribution of $[y_{ij}|\eta_j]$ is not normal but it is a bimodal distribution which is obtained as a mixture of two normal distributions with equal variance and different means. This is why the estimation of this model is much more complicated than the usual multilevel model. We used the multilevel mixture track in Mplus 3 (www.statmodel.com; Muthen & Muthen (13)) to estimate the model with the MPML method. Because the probability that x_{ij} is missing depends on y_{ij} the missing data type is not MCAR but it is MAR and therefore listwise deletion method would not only reduce the sample size but would also produce biased estimates. Thus bias in the estimates for this model can arise from not including the incomplete cases or

from not including the weights in the analysis. The MPML method however resolves both problems. The maximum likelihood estimates are consistent when the missing data is MAR. This generalizes trivially to multilevel maximum likelihood estimates and to the pseudo maximum likelihood for single and multilevel models as well. The selection is invariant across clusters and we therefore can use the AI scaling method. Table 6 contains the bias and coverage of the MPML estimates. The bias of the estimates is quite small except for the two variance estimates when $n_j = 5$. This is in line with the observations and conclusions we made in the previous section. We therefore extend these conclusions to the multilevel mixture models and other multilevel models. The MPML estimate would be approximately unbiased as long as the cluster sizes are not small when the informative index is large.

12. CONCLUSION

In this article we discussed some of the intricacies of weighting for unequal probability of selection in multilevel models. We introduced the multilevel pseudo maximum likelihood estimation method with general scaling options and provided conditions guaranteeing that these estimates are approximately unbiased. Through our simulation studies we demonstrated how sample size, the informative index, the nature of the selection mechanism, the type of scaling, the variance of the standardized weights and the intraclass correlation affect the quality of the MPML estimates.

We summarize our findings in the following 6 steps procedure that can be used to steer away from the pitfalls of weighting in multilevel modeling.

Step 1. Verify that weights are designed for a multilevel analysis. If the weights are designed for a single level analysis, and multilevel weights can not be obtained, instead of attempting multilevel modeling we recommend single level modeling that is designed for stratified and cluster sampling designs. Examples of this modeling approach can be found in Asparouhov (10).

Step 2. If weights are available only at level 2 there is no need to proceed with the rest of the steps in this procedure. Although the model is multilevel, the nature of the weighting is not. There are no complications in this case. The MPML estimation method simplifies to the

usual single level PML estimates, where the likelihood of a cluster takes the role of individual likelihood. Thus the MPML estimates, obtained by maximizing the weighted log-likelihood, are simply PML estimates and as such they are consistent regardless of how informative the selection is and how large the cluster sample sizes are. However the informativeness of the weights should still be examined as in step 6 below and if the weights are not informative they should not be used at all.

Step 3. Determine whether or not the selection mechanism of the level 1 units is invariant across cluster. This information could be extracted from a short description of the sampling design. If such information is not available assume that the selection is non-invariant.

Step 4. If the selection is invariant use scaling method AI. If the selection is non-invariant use scaling method A. Compute the weights taking the scaling method into account.

Step 5. Perform unweighted ML and weighted MPML analysis. Compute the informative index for all dependent variable in the model. Compute the ICC for all dependent variables in the model using the MPML parameter estimates. Compute the ratio between the effective sample size based on the individual level weights and the actual sample size.

Step 6. If all informative indices are below 0.02 a complete test of informativeness using Pfeffermann (11) test should be conducted. If the weights are non-informative an unweighted analysis, ignoring the weights, is recommended. If the informative index is less than 0.3 for all variables or the average sample size is larger than 10 the MPML estimates are expected to be trustworthy. If any of the informative indices is above 0.3 and the average cluster size is smaller than 10 we recommend a single level analysis as described in step 1. Borderline cases would require additional examination.

Additional steps could involve for example comparing the results from different scaling methods and also conducting simulations studies such as the ones described in this article based on a specific application.

The main advantage of the MPML method is its generality. While for some specific estimation problems perhaps a more accurate non-ML based estimator could be constructed, the generality of MPML is unparalleled. Cautiously using the MPML method can be a very

effective tool in dealing with selection bias in multilevel modeling.

13. AKNOWLEDGEMENT

The author is thankful to Bengt Muthen for his guidance, to Linda Muthen for her support and commitment and to Thuy Nguyen for computational assistance. The author is thankful to Rod Little, Laura Stapleton and the reviewer for helpful comments on the earlier draft. This research was supported by SBIR grant R43 AA014564-01 from NIAAA to Muthen & Muthen.

14. APPENDIX A

Here we show that the MPML estimates are approximately unbiased when the conditions 1-3 given in Section 3 are satisfied. For brevity denote by $F_{0j} = \phi(\eta_j|x_j, \theta_2)$ and by $F_{ij} = f(y_{ij}|x_{ij}, \eta_j, \theta_1)$. Let $\theta = (\theta_1, \theta_2)$ be the total parameter vector. Using the Laplace approximation (De Bruijn (16), Lindley (17))

$$l_j \approx e^{g_j(\hat{\theta}_{0j})}\sqrt{2\pi}\hat{\sigma}_{0j}\hat{n}_j^{-0.5}, \tag{11}$$

where

$$g_j(\theta) = \log(F_{0j}) + s_{1j}\sum_i w_{ij} log(F_{ij}), \tag{12}$$

$\hat{\theta}_{0j}$ is the mode of $g_j(\theta)$, and $\hat{\sigma}_{0j}^2 = -\hat{n}_j/g_j''(\hat{\theta}_{0,j})$. Since $g_j/\hat{n}_j \to E_j(\log(F_{ij}))$ as $n_j \to \infty$, where E_j is the expectation with respect to y_{ij} conditional on individual i being in cluster j, we get that $\hat{\theta}_{0j} \to \hat{\theta}_j-$ the mode of $E_j(\log(F_{ij}))$. Similarly $\hat{\sigma}_{0j} \to \hat{\sigma}_j$ as $n_j \to \infty$, where $\hat{\sigma}_j^2 = -1/E_j(\log(F_{ij}))''$. Thus for sufficiently large n_j

$$l_j \approx e^{g_j(\hat{\theta}_j)}\sqrt{2\pi}\hat{\sigma}_j\hat{n}_j^{-0.5}. \tag{13}$$

Maximizing the approximate likelihood amounts to solving the following approximate score equations obtained by omitting the lower level terms and terms that do not depend on the θ parameters

$$\frac{\partial}{\partial\theta_1}\sum_j\sum_i w_j w_{ij} s_{2j} s_{1j} \log(F_{ij}(\hat{\theta}_j)) = 0 \tag{14}$$

$$\frac{\partial}{\partial\theta_2}\sum_j w_j s_{2j} log(F_{0j}(\hat{\theta}_j)) = 0. \tag{15}$$

These two equations can be approximated by

$$\frac{\partial}{\partial\theta_1}E(s_{2j}\hat{n}_j E_j(log(F_{ij}(\hat{\theta}_j)))) = 0 \tag{16}$$

$$\frac{\partial}{\partial\theta_2}E(s_{2j}log(F_{0j}(\hat{\theta}_j))) = 0. \tag{17}$$

Under the assumptions 2 and 3 in Section 3, these score equations are equivalent to

$$\frac{\partial}{\partial\theta_1}E(n_j E_j(log(F_{ij}(\hat{\theta}_j)))) = 0 \tag{18}$$

$$\frac{\partial}{\partial\theta_2}E(log(F_{0j}(\hat{\theta}_j))) = 0. \tag{19}$$

At this point we see that the score equations are independent of the weights and the sampling scheme and thus sampling weights of 1, i.e., using simple random sampling, would yield the same approximate score equations. We conclude that the MPML estimates under unequal probability sampling and the ML estimates under simple random sampling are asymptotically equivalent and thus the MPML estimates are approximately unbiased for large enough n_j and $\hat{n}_j$.

15. APPENDIX B

Here we illustrate the MPML estimator by deriving a closed form expression for the parameter estimates of a random intercept model. Suppose that

$$y_{ij} = \mu + \eta_j + \varepsilon_{ij}, \tag{20}$$

where η_j and ε_{ij} are zero mean normally distributed variables with variances ψ and θ respectively. The weighted likelihood of the j cluster is

$$l_j = \int (2\pi\theta)^{-\hat{n}_j/2}(2\pi\psi)^{-1/2}Exp\bigg(-\frac{s_{1j}}{2\theta}\sum_i w_{ij}(y_{ij}-\mu-\eta_j)^2 - \frac{\eta_j^2}{2\psi}\bigg)d\eta_j =$$

$$(2\pi\theta)^{-\hat{n}_j/2}(2\pi\psi)^{-1/2}Exp\Big(-\frac{s_{1j}}{2\theta}\sum_i w_{ij}(y_{ij}-\bar{y}_j)^2\Big)\int Exp\bigg(-\frac{\hat{n}_j}{2\theta}(\bar{y}_j-\mu-\eta_j)^2 - \frac{\eta_j^2}{2\psi}\bigg)d\eta_j \tag{21}$$

where $\bar{y}_j$ is the weighted mean $\bar{y}_j = \sum w_{ij}y_{ij}/\sum w_{ij}$. Using the Laplace approximation (De Bruijn (16), Lindley (17)) formula which is exact when the function in the exponent is

quadratic we get that

$$l_j = (2\pi\theta)^{-(\hat{n}_j-1)/2}(2\pi(\theta+\hat{n}_j\psi))^{-1/2}Exp\bigg(-\frac{s_{1j}}{2\theta}\sum_i w_{ij}(y_{ij}-\bar{y_j})^2 - \frac{\hat{n}_j(\bar{y_j}-\mu)^2}{2(\theta+\hat{n}_j\psi)}\bigg). \quad (22)$$

Explicit maximization of the pseudo log-likelihood is possible only when $\hat{n}_j$ is constant across all clusters. In that case the parameter estimates are

$$\hat{\mu} = \frac{\sum_j s_{2j}w_j\bar{y_j}}{\sum_j s_{2j}w_j} \quad (23)$$

$$\hat{\theta} = \frac{\sum_j s_{2j}s_{1j}w_j\sum_i w_{ij}(y_{ij}-\bar{y_j})^2}{\sum_j s_{2j}w_j(\hat{n}_j-1)} \quad (24)$$

$$\hat{\psi} = \frac{\sum_j s_{2j}w_j(\bar{y_j}-\hat{\mu})^2}{\sum_j s_{2j}w_j} - \frac{\hat{\theta}}{\hat{n}_j}. \quad (25)$$

Weighting methods A and AI would allow explicit maximization when all cluster sample sizes n_j are equal. In that case when implementing scaling method A we get

$$\hat{\mu}_A = \frac{\sum_j w_j\bar{y_j}}{\sum_j w_j} \quad (26)$$

$$\hat{\theta}_A = \frac{1}{\sum_j w_j(n_j-1)}\sum_j n_j w_j\frac{\sum_i w_{ij}(y_{ij}-\bar{y_j})^2}{\sum_i w_{ij}} \quad (27)$$

$$\hat{\psi}_A = \frac{\sum_j w_j(\bar{y_j}-\hat{\mu})^2}{\sum_j w_j} - \frac{\hat{\theta}}{n_j} \quad (28)$$

and implementing scaling method AI we get

$$\hat{\mu}_{AI} = \frac{\sum_{ij} w_j w_{ij}y_{ij}}{\sum_{ij} w_j w_{ij}} \quad (29)$$

$$\hat{\theta}_{AI} = \frac{\sum_{ij} w_j w_{ij}(y_{ij}-\bar{y_j})^2}{\sum_{ij} w_j w_{ij}(1-1/n_j)} \quad (30)$$

$$\hat{\psi}_{AI} = \frac{\sum_{ij} w_j w_{ij}(\bar{y_j}-\hat{\mu})^2}{\sum_{ij} w_j w_{ij}} - \frac{\hat{\theta}}{n_j}. \quad (31)$$

The parameter estimate with scaling method A are asymptotically equivalent to Method 2 in Stapleton (9), note however that the asymptotic covariance of the parameter estimates is not the same. The asymptotic covariance estimates for Method 2 in Stapleton (9) are negatively biased while the asymptotic covariance estimates of the MPML method are generally consistent, as it is seen in the simulation studies presented in Sections 6 and 7.

Explicit maximization with scaling method C is possible if $\sum_i w_{ij}$ is constant across clusters. In that case we get that

$$\hat{\mu}_C = \frac{\sum_j w_j \bar{y_j}}{\sum_j w_j} \tag{32}$$

$$\hat{\theta}_C = \frac{\sum_{ij} w_j w_{ij} (y_{ij} - \bar{y_j})^2}{\sum_j w_j (\sum_i w_{ij} - \bar{w})} \tag{33}$$

$$\hat{\psi}_C = \frac{\sum_j w_j (\bar{y_j} - \hat{\mu})^2}{\sum_j w_j} - \frac{\hat{\theta}}{\hat{n}_j} \tag{34}$$

where $\bar{w} = \sum_{ij} w_{ij} / \sum_j n_j$ and $\hat{n}_j = \sum_i w_{ij} / \bar{w}$. This method is somewhat similar to method C described in Graubard and Korn (3) and Korn and Graubard (5), the only difference for example in the θ estimation is that $\bar{w}$ is replaced by 1. However, that estimator becomes biased for small cluster sample sizes even with SRS as noted in Korn and Graubard (5), where as the MPML estimator is consistent even for small cluster sample sizes. In fact $\hat{\theta}_C$ could be used even in the unbalanced cases as a moment based estimator that avoids the pitfalls of estimator C of Graubard and Korn (3) and Korn and Graubard (5).

Scaling methods B and BI have a closed form solution when the effective sample size $(\sum_i w_{ij})^2 / \sum_i w_{ij}$ is constant across clusters. The exact formulas are derived similarly. In that case scaling method B produces the same parameter estimates for θ as Method 3 in Stapleton (9), while the parameter estimate for ψ is approximately the same especially for large cluster sample sizes, because in Stapleton (9), $\hat{n}_j$ is replaced by n_j in the computation of the average cluster size.

Under SRS and balanced design all conditions needed for the closed form expressions are satisfied and methods A, AI, B, BI, C and D produce the same estimates as the ML estimates

$$\hat{\mu}_{ML} = \bar{y} \tag{35}$$

$$\hat{\theta}_{ML} = \frac{\sum_{ij} (y_{ij} - \bar{y})^2}{n - m} \tag{36}$$

$$\hat{\psi}_{ML} = \frac{\sum_j (\bar{y_j} - \hat{\mu}_{ML})^2}{m} - \frac{\hat{\theta}_{ML}}{n_j} \tag{37}$$

When using the unscaled weights method E, as the sampling fraction approaches zero, $\hat{n}_j$ approaches infinity. Thus the estimates for μ and θ are the same while the estimate for ψ

$$\hat{\psi}_E \approx \frac{\sum_j (\bar{y}_j - \hat{\mu}_{ML})^2}{m} \tag{38}$$

The bias of ψ_E is approximately θ/n_j. In the example considered in Sections 6 and 7 for $n_j = 5$ the bias of ψ_E would be approximately 0.4 under simple random sampling.

BIBLIOGRAPHY

(1) Skinner, C. J. Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys* (eds. Skinner, C.J.; Holt, D.; Smith, T.M.F.), Wiley, **1989**, 59-87.

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

(3) Graubard, B.; Korn, E. Modeling the Sampling Design in the Analysis of Health Surveys. Statistical Methods in Medical Research **1996**, *5*, 263-281.

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

(5) Korn, E.; Graubard, B. Estimating the variance components by using survey data. Journal of the Royal Statistical Society, Series B, **2003**, *65, Part 1*, 175-190.

(6) Kovacevic, M. S.; Rai, S. N. A pseudo maximum likelihood approach to multilevel modeling of survey data. Communications in Statistics, Theory and Methods, **2003**, *32*, 103-121.

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

(8) Pfeffermann, D.; Moura, F.; Silva, P. Multi-Level Modeling Under Informative Sampling. Submitted, **2004**.

(9) Stapleton, L. The Incorporation of Sample Weights Into Multilevel Structural Equation

Models. Structural Equation Modeling, **2002**, *9(4)*, 475-502.

(10) Asparouhov, T. Sampling Weights in Latent Variable Modeling. Forthcoming in Structural Equation Modeling, **2005**. See also Mplus Web Note #7 www.statmodel.com.

(11) Pfeffermann, D. The Role of Sampling Weights when Modeling Survey Data. International Statistical Review, **1993**, *61, 2*, 317-337.

(12) Potthoff, R.F.; Woodbury, M.A.; Manton, K.G. "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. Journal of the American Statistical Association, **1992** *87*, 383-396.

(13) Muthen, L.K.; Muthen, B.O. *Mplus User's Guide*, Third Edition. Los Angeles, CA: Muthen & Muthen, **1998-2005**.

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

(15) Asparouhov, T. Stratification in Multivariate Modeling, Submitted. See also Mplus Web Note #9 www.statmodel.com.

(16) De Bruijn, N. G. *Asymptotic Methods in Analysis*, Amsterdam: North-Holland, **1961**.

(17) Lindley, D.V. Approximate Bayesian Methods. In *Bayesian Statistics* (eds. Bernardo, J.M.; Degroot, M.H.; Lindley, D.V.; Smith, A.N.F.). Valencia, Spain: University Press, **1980**.