Comparison of Estimation Methods for Complex Survey Data Analysis

Tihomir Asparouhov[1]

Muthen & Muthen

Bengt Muthen[2]

UCLA

[1] Tihomir Asparouhov, Muthen & Muthen, 3463 Stoner Ave. Los Angeles, CA 90066.

[2] Bengt Muthen, UCLA, 2023 Moore Hall, Los Angeles, CA 90095-1521.

This research was supported by SBIR grant R43 AA014564-01 from NIAAA to Muthen & Muthen. The authors are thankful to Stephe Du Toit for assistance with the LISREL software package.

Abstract

Recently structural equation modeling software packages have implemented more accurate statistical methodology for analyzing complex survey data. The computational algorithms however vary across the packages and produce different results even for simple models. In this note we conduct simulation studies to compare the performance of the methods implemented in Mplus and LISREL. The Mplus algorithm produced more accurate results.

Recently several structural equation modeling and multilevel software packages have implemented more accurate statistical methodology for analyzing complex survey data. 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 note we conduct simulation studies to evaluate the performance of the methods implemented in Mplus 4.1 and LISREL 8.8. Mplus is published by Muthen & Muthen and LISREL is published by Scientific Software International. First we conduct a simulation study on a two-level random effect regression model estimated from data that includes within level sampling weights. In a second simulation study we evaluate the performance of the chi-square test statistic for complex survey data using a simple bivariate mean, variance and covariance model.

TWO-LEVEL REGRESSION

For analyzing two-level models with complex survey data Mplus 4.1 implements the multilevel pseudo maximum likelihood (MPML) estimation method, see Asparouhov (2006) and Asparouhov and Muthen (2006). This method allows us to estimate two-level models when the data is obtained from a multistage stratified sampling design and the sampling units at each sampling level are selected with unequal probabilities. Sampling weights at both the within and the between level can be used with this estimator. The sampling weights on the cluster (between) level are obtained by

$$w_j = \frac{1}{p_j} \tag{1}$$

where p_j is the probability that cluster j is included in the sample. The sampling weights on the individual (within) level are obtained by

$$w_{ji} = \frac{1}{p_{i|j}} \tag{2}$$

where $p_{i|j}$ is the probability that individual i in cluster j is selected, given that cluster j is selected. Using the unscaled within level weights in the estimation method can lead to bias in the parameter estimates. A number of different scaling methods have been proposed in Pfeffermann et al. (1998). The choice of scaling method affects the parameter estimates to some extent. The simulation studies in Pfeffermann et alt. (1998) and Asparouhov (2004) conclude that scaling to cluster sample size tends to have the most robust performance. With this scaling method the scaled within level weights w^*_{ji} are obtained by

$$w^*_{ji} = w_{ji} \frac{n_j}{\sum_i w_{ji}} \tag{3}$$

where n_j is the size of cluster j. Note that $\sum_i w^*_{ji} = n_j$. Mplus 4.1 uses the total combined weight variable, which is the product of the within and the between level weight variables

$$w_j w^*_{ji}. \tag{4}$$

LISREL 8.8 implements the PWIGLS method described in Pfeffermann et alt. (1998) using the scaling to cluster sample size for the within level weights as well.

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

$$Y_{ji} = \alpha_j + \beta_j X_{ji} + \gamma Z_{ji} + \varepsilon_{ji} \tag{5}$$

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

$$p_{i|j} = \frac{1}{1 + Exp(-Y_{ij}/2)}. \tag{6}$$

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

$$w_{ji} = \frac{1}{p_{i|j}} = 1 + Exp(-Y_{ij}/2). \tag{7}$$

Consequently we rescale the within level weights using formula (3). To introduce unequal probability sampling on the between level we retain clusters in the sample with probability

$$p_j = \frac{1}{1 + Exp(-\alpha_j)}. \tag{8}$$

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

$$w_{ji} = \frac{1}{p_j} = 1 + Exp(-\alpha_j). \tag{9}$$

We estimate model (5) for each sample using Mplus and LISREL. Within the LISREL software package this kind of models are estimated by the MULTILEV module.

Table 1 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 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 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 non-informative. 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 it is possible to obtain the results given by Mplus within LISREL as well by using the total weight (4) as the within level weight and not using between level weights. Indeed using this approach LISREL will produce results very close to those obtained in Mplus.

CHI-SQUARE ADJUSTMENTS

In structural equation modeling chi-square tests are used to evaluate the overall fit of the model. Typically the chi-square test of fit is simply the likelihood ratio test (LRT) for the structural model against the unrestricted means, variance and covariance model. The test statistic is computed as twice the difference between the log-likelihoods of the two nested models. When we analyze complex survey data with a single or multilevel

Table 1: Bias and MSE of parameter estimates for two-level regression.

Parameter	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

model we use the weighted pseudo log-likelihood. Using the pseudo log-likelihood we can again perform the LRT, however the distribution of this test statistic is no longer a chi-square distribution. This distribution depends on the entire sampling design, including the sampling weights, the stratification and the cluster sampling.

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 which has approximately 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.

Consider 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 adjusted LRT statistic given in Asparouhov and Muthen (2005) is

$$T^* = c \cdot 2(L_1 - L_2), \tag{10}$$

where c is the correction factor

$$c = \frac{d_1 - d_2}{Tr((L_1'')^{-1}Var(L_1')) - Tr((L_2'')^{-1}Var(L_2'))} \tag{11}$$

where L_i' and L_i'' are the first and second derivatives of the pseudo log-likelihoods. The effect of the sampling design on the correction factor is summarized in the score variance terms $Var(L_i')$. 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. The adjusted LRT test performs well also in multilevel analysis of complex survey data, see Asparouhov and Muthen (2006).

When complex survey data is analyzed in Mplus 4.1 the chi-square test of fit are automatically adjusted using formula (10). Note however that LRT testing is of interest not only to conduct chi-square test of fit but also for testing between two nested models. For many advanced latent variable models such as random slope or mixture models there is no naturally defined unconstrained model that can be used to evaluate model fit. In such cases it is important to be able to conduct LRT testing between competing nested models. When we analyze complex survey data it is important to use the adjusted LRT. Mplus 4.1 computes the following log-likelihood correction factors for every likelihood based model estimation

$$c_i = \frac{Tr((L_i'')^{-1}Var(L_i'))}{d_i}. \tag{12}$$

Using the log-likelihood correction factors c_1 and c_2 for two nested models one can compute the LRT correction factor

$$c = \frac{d_1 - d_2}{c_1 d_1 - c_2 d_2}. \tag{13}$$

The LRT adjustment corrects not only for complex survey designs but also for non-normality and other distributional misspecifications. Thus the likelihood correction factors have now enabled us to conduct robust chi-square testing not just for models that have well defined chi-square test of fit but for any latent variable models.

An alternative LRT adjustment has been proposed and implemented in LISREL 8.8. As described in the LISREL documentation (2005) accompanying the software package, the adjustment is given again by equation (10) but the correction factor c is computed by

$$c = \frac{d2}{Tr((L_2'')^{-1}Var(L_2'))}. \tag{14}$$

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

We explore the differences between the adjustments implemented in Mplus and LISREL with a simple simulation study. 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 2 shows the rejection rates for the three LRT statistics, the Mplus LRT adjustment, the LISREL LRT adjustment and the unadjusted LRT. Tables 3 shows the average values of these test statistics. It is clear from these results that the Mplus LRT adjustment performs 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 inflated test statistic values.

In some cases the Mplus approach (11) and the LISREL approach (14) will produce the same results. If the effect of the complex sampling design is similar across all

Table 2: 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 3: 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

variables and parameters in the model then $c_1 \approx c_2$, which in turn implies that formulas (11)and (14) produce the same result.

CONCLUSION

In this note we conducted simple simulation studies to evaluate the estimation methods available in Mplus and LISREL for complex survey data analysis. We found substantial differences between the two software packages. In our simulations studies the results obtained in Mplus 4.1 were more accurate than those obtained in LISREL 8.8.

The simulation studies presented here have limited implications in practice. There are a number of factors that have a substantial impact on the quality of the estimation, see Asparouhov (2006). Four known factors in order of importance are the cluster sample size, the informativeness of the within level weights, the ICC (intra class correlation) and the UWE (unequal weighting effect). This study was not intended to give a complete account on all possible situations but to provide a simple comparison in somewhat artificial settings to evaluate the methodological differences implemented in the two software packages.

In a specific practical situation it is still unclear what the best estimation approach is. In Asparouhov (2006) a six step procedure is recommended as an optimal estimation strategy, however the procedure does not cover all possible practical aspects. For example the situation of large UWE such as the one found in Chantala and Suchindran (2006) is not covered. Additional simulation studies should be conducted to evaluate the quality of the estimation techniques. In addition, simulation study procedure based on the actual data should be developed to incorporate more data specific features.

1 References

Asparouhov, T. (2006). General Multilevel Modeling with Sampling Weights. Communications in Statistics: Theory and Methods, Volume 35, Number 3, pp. 439-460(22).

Asparouhov, T. and Muthen, B. (2005). Multivariate Statistical Modeling with Survey Data. Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference.

Asparouhov, T. and Muthen, B. (2006). Multilevel Modeling of Complex Survey Data. Proceedings of the American Statistical Association, Seattle, WA: American Statistical Association.

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.

LISREL Documetation, Analysis of Structural Equation Models for Continuous Random Variables in the Case of Complex Survey Data (2005).
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.

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

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.

Stapleton, L. (2006) An Assessment of Practical Solutions for Structural Equation Modeling with Complex Sample Data. Structural Equation Modeling, 13, No. 1, 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.