Stratification in Multivariate Modeling

Tihomir Asparouhov
Muthen & Muthen

Mplus Web Notes: No. 9

Version 2, December 16, 2004

[1]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. This research was supported by SBIR grant R43 AA014564-01 from NIAAA to Muthen & Muthen.

Abstract

In this note we illustrate stratified complex sampling with several simulation studies implemented in Mplus 3.1 and discuss the effect of stratification on parameter and variance estimation and on log-likelihood chi-square testing. We compare the results obtained by Mplus with those obtained by SUDAAN on linear and logistic regression models.

1 Introduction

The purpose of this note is to illustrate the concept of stratification through several simulation studies conducted with Mplus Version 3.1 (Muthen and Muthen (1998-2004); www.statmodel.com). The pseudo maximum likelihood (PML) developed by Skinner (1989), which is implemented in Mplus, can be used to estimate any model with stratified sample. The PML method for non-stratified samples was discussed in detail in Mplus Web Note 7, Asparouhov (2004).

Stratified sampling in practice arises naturally for example when the strata are different geographical areas, school districts etc. Stratification is also used as a tool to improve the precision of the parameter estimates. Unlike cluster sampling, stratified sampling actually reduces the variance of all parameter estimates. Stratification can also arise as post-stratification, balancing a non-stratified sample through weights, see Cochran (1977) 5A.9.

The PML estimates are obtained by maximizing the weighted pseudo log-likelihood

$$l = \sum_{i,h} w_{hi} l_{hi},$$

where w_{hi} is the sampling weight and l_{hi} is the log-likelihood of subject i in stratum h. The variance of the parameter estimates is given by

$$V = H^{-1} Var\Big(\sum_h \sum_i w_{hi} s_{hi} \Big) H^{-1},$$

where $H = L''$ is the information matrix and $s_{hi} = l'_{hi}$ is the score. Thus the variance is approximated by

$$V = H^{-1} \Bigg(\sum_h \frac{n_h}{n_h - 1} \sum_i (w_{hi} s_{hi} - \overline{s}_h)(w_{hi} s_{hi} - \overline{s}_h) \Bigg) H^{-1}, \tag{1}$$

where

$$\overline{s}_h = \frac{\sum_i w_{hi} s_{hi}}{\sum_i w_{hi}}$$

is the weighted average score and n_h is the number of samples in stratum h. If we fail to account for stratification and use the usual sandwich variance estimate

$$H^{-1} \Bigg(\sum_{hi} w_{hi}^2 s_{hi} s_{hi}^T \Bigg) H^{-1}, \tag{2}$$

we would overestimate the variance by approximately

$$H^{-1}\left(\sum_h n_h \overline{s}_h^2\right)H^{-1}.$$

2 Factor Analysis

We first consider a factor analysis model with 3 indicators Y_1, Y_2 and Y_3 and a single factor η

$$Y_j = \mu_j + \lambda_j \eta + \varepsilon_j,$$

where ε_j is a zero mean uncorrelated to η normal residual with variance θ_j. The factor η is a standard normal random variable with mean zero and variance 1. For identification purposes we assume in our model that these parameters are fixed.

The population is constructed in the following way. Within each stratum the variance of Y_j is 1 and the covariance between them is 0.5. The means of Y_j vary across strata. In this example we use 3 strata with sample sizes 250, 750 and 150 and the means in these strata are 1, 3, and 6 respectively and are the same for the three variables. For simplicity we assume that all sampling weights are 1, which corresponds to the situation where the sample size within each stratum is proportional to the size of the stratum. All files used to conduct this simulation are available on Mplus web site www.statmodel.com. The data are generated with the StrataEx1.inp file. We consider 5 different estimation methods.

- Method A. This method properly accounts for the stratified sampling using formula (1).
- Method B. This method ignores the stratified sampling, it uses robust standard errors that avoid normality assumptions using formula (2).
- Method C. This method uses the usual maximum likelihood information matrix based estimation.
- Method D. This method incorrectly specifies stratified sampling as cluster sampling. The difference between the two is that in cluster sampling a random selection occurs on the higher level while in stratified sampling an exhaustive selection occurs on the higher level. Using this method, our sample consists of three clusters.

Table 1: Ratio Between Average Standard Errors (SE) and Standard Deviation of Parameter Estimates (SD) in Factor Analysis

Parameter	SD A-D	SE/SD A	SE/SD B	SE/SD C	SE/SD D	SD E	SE/SD E
μ_1	0.030	0.993	1.731	1.731	24.037	0.053	0.970
λ_1	0.030	0.980	1.382	1.316	21.174	0.045	0.939
θ_1	0.034	0.959	0.959	0.962	0.740	0.034	0.959

- Method E. This method is provided here just for an illustration and is not directly comparable with the above methods. Here we replace the stratified sampling with simple random sampling to show how much precision is gained by the stratification.

All of the above methods use maximum likelihood parameter estimation. The difference in the methods is only in the estimation of the standard errors of these parameter estimates. Method A is implemented in input file StrataEx1A.inp, method B in StrataEx1B.inp, etc. Note that Method E sampling is implemented via Mixture modeling, that is, a stratum variable corresponds to the different classes, and each sample is selected by first selecting stratum with probability proportional to size and then selecting a random sample from that stratum. This sampling is equivalent to simple random sampling. All simulation results are obtained by replicating the analysis 500 times. Since all parameter estimates are obtained via maximum likelihood they are generally consistent and we do not report them here. We only report the standard deviation of these parameter estimates and the average standard errors. If the analysis is correct the ratio between the two should at least asymptotically approach 1. Any systematic deviation from 1 indicates overestimation or underestimation of the standard errors. Table 1 shows the results for parameters μ_1, λ_1 and θ_1, the remaining parameters by symmetry are equivalent to these.

Table 1 shows that the only acceptable results for stratified sampling are given by method A. Methods B and C produce results that match the standard deviation of parameter estimates under simple random sampling, implemented in Method E. Method D shows the worst results, that is because cluster sampling tends to decrease the precision of the parameter estimates

while stratified sampling tends to increase it and confusing the two can produce particularly bad results.

An alternative approach to analyzing this data is to estimate a multiple group model or a multilevel model. These however may not be of interest in some situations. For example it may be desirable to estimate the model for the entire population rather than have separate models for different segments of the population. In every specific population segment the correlation between the observed variables would usually be smaller than the correlation for the entire population. Multilevel analysis would also not accommodate the stratification structure, that is, the multilevel model would not use the fact that the strata represented in the sample are indeed all strata. Another drawback of multiple group analysis, when the number of strata is large, is that it may actually lead to a model with too many parameters which are estimated with greater imprecision than models with fewer parameters. Finally, both multilevel and multiple group analysis tend to model the dependence within group with linear equations, while stratified single level analysis would not engage in such modeling and thereby be less risky in terms of model misspecification.

3 Linear Regression

In this example we consider a simple regression of a dependent variable Y on another variable X

$$Y = \mu + \beta X + \varepsilon,$$

where ε is a zero mean residual with variance θ. We generate a finite target population of size 10000 using a normal distribution for X with mean 4 and variance 1 and the following parameter values $\mu = 1.5$, $\beta = 0.75$ and $\theta = 1$. If we estimate these parameters using the entire population we get $\mu = 1.452$, $\beta = 0.758$ and $\theta = 0.992$ which as expected are very close to the original values. We rank the observations according to the value of the function f defined by $f(Y, X) = Y/X$. We then split the target population into 14 strata the first two of size 2000 and the remaining of size 500, so that the first 2000 ranked observations are placed in the first stratum, the second 2000 are placed in the second stratum; the next 500 are placed in the third stratum etc. The formation of the strata can be done with any other ranking function. The effects of the stratification however depend directly on this function. For example if the ranking function is independent of Y

Table 2: Bias and SE/SD Ratio in Linear Regression Analysis

Parameter	Bias A	SE/SD A	Bias B	SE/SD B	Bias C	SE/SD C
μ	0.005	1.035	0.005	1.035	0.005	1.189
β	−0.001	1.043	−0.001	1.041	−0.001	1.130
θ	−0.001	1.017	−	−	−0.001	1.250

and X the stratified sampling will be equivalent to simple random sampling as far as this model is concerned and thus the stratification will have no effect at all on the analysis. In practice we can argue that strata are formed this same way with a ranking function that is not known and most likely is very complicated. Whatever the true ranking function is however, as long as we have the correct sampling weights we can produce consistent parameter estimates and correct standard errors as we illustrate this below.

The data generation procedure that we used above is completely different from the one used in the previous section. We use this procedure to avoid the problem of combining strata generated with different models and analyzed with one model, which may not hold for the entire population. Similar data generation technique has been used for example by Smith and Holmes (1989).

In our simulation study we sample 50 observations with replacement from each stratum, and assign sampling weight of 4 for the observations from the first two strata and 1 for the rest. We repeat this sampling procedure 500 times. We analyze these 500 samples using the following three methods

- Method A. Mplus 3.1 implementation of PML for stratified sample.
- Method B. SUDAAN 8.0.2 (2003, Research Triangle Institute) implementation of GEE for stratified sample.
- Method C. Mplus 3.1 implementation of PML ignoring the stratification sampling and assuming unequal probability sampling to reflect the sampling weights.

The results of this simulation are presented in Table 2. In general the estimating equations of the GEE approach coincide with the score equations of the PML approach and thus we expect methods A and B to produce exactly the same results and indeed this was confirmed by our simulation

Table 3: Bias and SE/SD Ratio in Logistic Regression Analysis

Parameter	Bias A	SE/SD A	Bias B	SE/SD B	Bias C	SE/SD C
α	-0.011	1.004	-0.011	1.004	-0.011	1.137
β	0.004	0.980	0.004	0.981	0.004	1.105

study. Methods A and B not only agree on average but also in individual replications both in parameter estimates and in standard errors. All three methods produce almost no bias at all. As expected method C overestimated the standard errors for all parameters.

4 Logistic Regression

In this example we have a binary variable Y with observed values 0 and 1 and a predictor variable X. The logistic regression we consider is

$$P(Y = 1) = \frac{1}{1 + e^{-\alpha - \beta X}}.$$

We generate data using a normal distribution for X with mean 2 and variance 1 and parameters $\alpha = -0.5$ and $\beta = 0.5$. We generate a target population of 10000. If we estimate these parameters using the entire target population we get $\alpha = -0.436$, $\beta = 0.468$. We construct the strata and the sample as in the previous example using the function $f(X, Y) = (Y + U)/X$ to rank the observations where U is an independent uniformly distributed random variable on the interval [0,1]. We use the same estimating methods as in the previous example.

The results are presented in Table 3. Again for the logistic model the estimating equations of the GEE approach coincide with the score equations of the PML approach and thus methods A and B produce exactly the same results not only on average but also in individual replications. Both parameter estimates and standard errors are exactly the same. All three methods produce almost no bias at all, and the parameter estimates are identical. Method C again shows small overestimation of the standard errors because it ignores the stratification.

5 Multilevel Probit Regression

In this section we demonstrate the effect of stratification on the weighted least square (WLS) estimation method of Muthen (1984) and Muthen et al. (1997) for a multivariate probit regression. In Asparouhov (2004) we show how the PML approach of Skinner (1989) to complex sample modeling is extended to the WLS estimators. This approach is implemented in Mplus Version 3.1. The WLS estimator can be used to estimate for example a factor analysis models with binary variables or growth models with binary variables. In this section we consider a growth model with binary variables which can be interpreted also as a multivariate probit regression. Suppose that Y_{it} is a binary variable for individual i at time t. This variable can be either 0 or 1 and

$$P(Y_{it} = 0) = \Phi(\alpha_i + \beta_i t) \tag{3}$$

where α_i and β_i are random effect variables with means μ_1 and μ_2, variances ψ_1 and ψ_2 and covariance ρ. The function Φ is the standard normal distribution function. In this example we assume that we have 6 observations for each individual at times t=0,1,2,...,5, however we could have individually varying number of observations at individually varying times of observations. The probit binary growth model estimated by the WLS estimator has a computational advantage over the maximum-likelihood logistic growth model because it does not involve numerical integration, which tends to be slower and less accurate. The WLS estimator tends to have convergence problems with small sample size and thus we use its modified version, the WLSMV estimator, which is less efficient in general but less prone to computational problems. The true parameter values for this simulation study are as follows $\mu_1 = -1$, $\mu_2 = 0.15$, $\psi_1 = 0.6$, $\psi_2 = 0.05$ and $\rho = 0$. Stratified sampling is implemented as follows. We generate a target population according to model (3) with 10000 individuals with 6 observations each. The target population is ordered with the following ranking function $f_1 = -3Y_1 - 2Y_2 - Y_3 + Y_4 + 2Y_5 + 3Y_6$. We separate the total population into 2 equal size groups. The first 5000 individuals with the highest f_1 values are placed in the first group and the remaining in the second group. Within each of the two groups the observations are ordered in a lexicographical (alphabetical) order. This ordering is equivalent to ordering based on the following function $f_2 = 32Y_1 + 16Y_2 + 8Y_3 + 4Y_4 + 2Y_5 + Y_6$. Then each group is split into two equal parts so that the higher lexicographical order individuals are

Table 4: Bias and SE/SD Ratio in Multilevel Probit Regression

Parameter	Bias A	SE/SD A	Bias C	SE/SD C
μ_1	-0.001	0.905	-0.004	1.727
μ_2	-0.002	0.949	0.002	1.622
ψ_1	0.062	0.961	0.064	1.033
ψ_2	0.005	0.973	0.005	0.968
ρ	-0.013	0.941	-0.013	1.094

separated from the lower lexicographical order individuals. In total we get 4 equal size strata. We have basically stratified our sample by two separate factors. The lexicographical order f_2 affects the random intercept while the stratification based of f_1 affects the random slope. From each stratum we sample 100 observations at random and with replacement. The total sample size is thus 400. We draw 500 samples and analyze them with method A and method C defined as in Section 3

- Method A. Mplus 3.1 implementation of WLSMV for stratified sample.
- Method C. Mplus 3.1 implementation of WLSMV ignoring the stratification sampling.

It is not possible to analyze this model with SUDAAN because SUDAAN does not support the probit regressions models and multilevel models. All of the strata are of equal size and are sampled equally. Therefore the weights for all individuals are 1.

The results are presented in Table 4. It is clear from this simulation that the WLSMV method provides the correct stratified sampling modifications for the standard errors. Method C overestimated the standard errors of μ_1 and μ_2 by 73% and 62% respectively. The bias produced by both methods is negligible and presumably will decrease to zero as the sample size increases. Note that unlike the PML method, the stratification information with the WLSMV method affects the parameter estimates because the weight matrix is affected by the stratification. The difference between the parameter estimates of method A and C are very small however. In our simulation the variance parameters ψ_1, ψ_2 and ρ are virtually unaffected by the stratification

and both methods produced correct standard errors for these parameters. This can be explained with the fact that the particular stratification used in the simulation study did not improve the precision of these parameter estimates. If different stratification had been used or additional stratification had been done based on information related to these parameters the results would have been similar to the results for the mean parameters. The values of SE/SD ratio for method A differ slightly from 1 however that difference decreases as the sample size increases.

6 Log-Likelihood Chi-Square

The log-likelihood chi-square difference testing, just as the standard errors of the parameter estimates, can be affected by stratification. The adjusted chi-square test provided by Mplus for complex sampling estimation can take into account the stratification and the clustering features of the complex sampling. The chi-square adjustment is constructed similarly to the adjustments of the Yuan-Bentler (2000) and the Satorra-Bentler (1988) robust chi-square tests. We demonstrate the importance of such adjustments with a simple simulation study which incorporates both cluster and stratified sampling. For simplicity we will use a single outcome variable and will compare the mean and the variance of this outcome across two groups. Each of the two groups contains three strata. Within each stratum we sample at random entire clusters. For example the two groups can be private and public schools, the strata can be different regions in the country, the clusters can be the classrooms and the students can be the individual observations. While in this example the groups actually contain entire strata and clusters, this doesn't necessarily have to be the case. For example the grouping variable could be gender which is not nested above the strata and the cluster variables. This kind of grouping variable can be used with the MLR estimator in Mplus 3.11 for complex data but not with WLS based estimators, which require that the grouping variable be nested above the strata variable.

All six strata in our simulation study have equal size and we sample 200 observations from each by cluster sampling. Within each stratum the clusters are of equal size. We denote the size of the clusters in stratum s in group g by n_{sg}. The cluster sizes in the six strata are as follows $n_{11} = 5$, $n_{21} = 10$, $n_{31} = 20$, $n_{12} = 10$, $n_{22} = 20$, $n_{32} = 40$. The distribution of observations i in

cluster j in stratum s in group g is described by

$$Y_{ijsg} = \mu_{sg} + \eta_{jsg} + \varepsilon_{ijsg}$$

where η_{jsg} and ε_{ijsg} are zero mean normally distributed variables with variance 1, and the parameters μ_{sg} are as follows $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{31} = 3$, $\mu_{12} = 0$, $\mu_{22} = 2$, $\mu_{32} = 3$. Given our choice of parameters the total mean in the two groups is 2 however the total variance of y is larger in the second group. We test two hypotheses by the log-likelihood chi-square difference test. The first hypothesis T_1 is that the means in the two groups are equal. The second hypothesis T_2 is that both the means and the variance parameters are equal in the two groups. The first test should not reject the hypothesis because the means are indeed equal however the second test should reject the hypothesis because the variances are not equal. In addition the test statistic T_1 should have a chi-square distribution with 1 degree of freedom because it tests just one constraint. Test statistic T_2 has two degrees of freedom because it tests two constraints. The null hypothesis for the second test is not correct however and therefore the T_2 test statistic is not expected to have a chi-square distribution with 2 degrees of freedom. This test statistic is expected to be sufficiently large so that the test is rejected.

To evaluate the effect of stratification and clustering on the test we compare five different methods for computing the chi-square statistic. These methods are as follows.

- Method A. Adjusted robust chi-square test which takes both the clustering and the stratification into account.
- Method B. Adjusted robust chi-square test which takes only the clustering into account and ignores the stratification.
- Method C. Adjusted robust chi-square test which takes only the stratification into account and ignores the clustering.
- Method D. Adjusted robust chi-square test which ignores both the clustering and the stratification.
- Method E. Unadjusted log-likelihood difference chi-square test.

The results of the simulation study are presented in Table 4. We report the average values of the T_1 and T_2 test statistics over 500 replications and the rejection rates for the two tests based on the 5% rejection level. As expected method A performs correctly producing a test statistic T_1 with an average

Table 5: Effect of Stratification and Clustering on the Chi-Square Test

Method	A	B	C	D	E
T_1 Average	1.042	0.349	9.141	5.052	4.984
T_1 Rejection	0.054	0.002	0.524	0.380	0.380
T_2 Average	12.827	8.057	75.884	61.236	53.856
T_2 Rejection	0.760	0.500	0.990	0.982	0.980

value of approximately 1 and rejection rate of approximately 5%, while all the other methods produced erroneous results. From the table we clearly see that including the stratification information results in an increase of the chi-square statistic and the rejection rates, while including the cluster information decreases the chi-square and the rejection rates. The result of not including the stratification information in the first test is that there are virtually no rejections, while the result of not including the cluster information is that the test rejects the null hypothesis incorrectly additional 47% of the time above the nominal 5% level. Methods D and E both produce rejection rates that are too high and in our simulation the results of the two methods are quite close. This can be explained by the fact that there is no severe non-normality in the data. The mixture of the normal distributions from the three strata produces a distribution that is somewhat close to normal.

The most important effect of stratification is actually seen in the second test. Methods C, D and E all have inflated power largely because the clustering information is ignored. Method A rejects 76% of the time for this sample size. As the sample size increases this rejection rate converges to 100%. Not including the stratification information in method B results in a decrease of power. As a result of that, method B does not reject the second hypothesis as it should an additional 26% of the time.

It is clear from Table 4 that the sampling features in complex sampling designs can affect dramatically the chi-square statistics and erroneous conclusions can be reached if the sampling features are not accounted for. The adjusted chi-square test provided in Mplus deals effectively with complex sampling features such as stratified and cluster sampling.

7 Conclusion

In this note we demonstrated the effects of stratification on parameter estimation, standard error estimation and chi-square testing. We showed that we can improve the precision of the parameter estimates if we can provide a good quality stratification that groups the samples into more homogeneous subsets than the entire populations is. If the sample is poorly stratified however, i.e., if the stratification is irrelevant to the estimated model, we would not gain any efficiency and could actually decrease the precision of the estimates through an increase in the weights variation. We also showed that when stratified sampling is used we have to account for it while estimating the standard errors. The regular robust standard errors are generally biased upwards. This overestimation can be of varying magnitude depending on how strong the stratification is, i.e., the magnitude depends on the level of homogeneity improvement we achieve with the stratification. If we do not use robust standard errors and use regular information matrix estimation the effect is unpredictable, that is the information matrix can lead to overestimation or underestimation when we fail to account for stratification. Failing to include the stratification information in the analysis can also result in underestimation of the adjusted log-likelihood chi-square test and as a consequence the power of the test will be decreased. We also demonstrated through simulation studies that the current implementation of the variance estimation for stratified complex sample in Mplus 3.1 leads to exactly the same results as the GEE implementation in SUDAAN 8.0.2 for both linear and logistic regression. While the framework of the GEE method is limited to generalized linear models the PML method can be used to estimate any model, and this appears to be its main advantage.

8 References

Amemiya, T. (1985). Advanced Econometrics. Harvard University Press.

Asparouhov, T. (2004). Sampling weights in latent variable modeling. Mplus Web Note #7. Forthcoming in Structural Equation Modeling.

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

Imhof, J.P. (1961) Computing the Distribution of Quadratic Forms in Normal Variables. Biometrika 48, 419-429.

Muthen, L.K. and Muthen, B.O. (1998-2004). Mplus User's Guide. Third Edition. Los Angeles, CA: Muthen & Muthen

Muthen, B. (1984). A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators. Psychometrika, 49, 115-132.

Muthen, B., du Toit, S.H.C. & Spisic, D. (1997). Robust Inference Using Weighted Least Squares and Quadratic Estimating Equations in Latent Variable Modeling with Categorical and Continuous Outcomes. Accepted for publication in Psychometrika.

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.

Satorra, A., & Bentler, P.M. (1999). A Scaled Difference Chi-square Test Statistic for Moment Structure Analysis. UCLA Statistics Series # 260. http://www.stat.ucla.edu/papers/preprints/260/

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

Smith, T., and Holmes, D. (1989) Multivariate Analysis. In Analysis of Complex Surveys (eds. C.J.Skinner, D.Holt and T.M.F. Smith) 165-190, Wiley.

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

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.