Multivariate Statistical Modeling with Survey Data

Tihomir Asparouhov
Muthen & Muthen

Bengt Muthen
UCLA

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

Abstract

We describe an extension of the pseudo maximum likelihood (PML) estimation method developed by Skinner (1989) to multistage stratified cluster sampling designs, including finite population and unequal probability sampling. We conduct simulation studies to evaluate the performance of the proposed estimator. The estimator is also compared to the general estimating equation (GEE) method for linear regression implemented in SUDAAN. We investigate the distribution of the likelihood ratio test (LRT) statistic based on the pseudo log-likelihood value and describe an adjustment that gives correct chi-square distribution. The performance of the adjusted LRT is evaluated with a simulation study based on the Behrens-Fisher problem in a stratified cluster sampling design.

1 Introduction

Estimation of simple statistical models such as linear and logistic regressions with survey data is well established and widely used. These models are however inadequate for analyzing large multivariate data sets that are being made available by governmental agencies and other research institutions. Increasingly analysts are turning to advanced multivariate models to better penetrate these complex data structures. Simultaneous regression equations, structural equation models, time series models, log-linear models, mixture models, mixed models, latent class models, latent variable models and combinations of these are frequently the analysts' choice. Methods for estimating such models however with data obtained by multistage survey designs are not well established. Frequently analysts use methods designed for simple random sampling followed by an ad-hoc adjustment for variance estimation, see Stapleton (2005) for a review of such methods. These methods however are somewhat arbitrary and their theoretical properties are not well known. Until recently many statistical packages implemented such ad hoc methods as well, see Asparouhov (2005a).

Skinner (1989) introduced the pseudo maximum likelihood (PML) method which can be used to estimate any general multivariate parametric model with data from a complex survey design which includes stratification, cluster sampling, and unequal probability sampling with replacement. This method is in fact applicable to a more general sampling design which includes stratified multistage sampling with unequal probability sampling with replacement at the primary sampling stage while allowing for with and without replacement unequal probability sampling on subsequent stages. This sampling design is known as the WR sampling design and is pioneered and implemented in the software package SUDAAN (RTI, 2002). SUDAAN however is based on the general estimating equations (GEE) methodology and is only capable of estimating simple univariate models such as linear and logistic regression. Mplus, Version 3 (Muthen & Muthen, 1998-2004), implements the PML method for the WR design and many multivariate models with normal, discrete and other parametric distributions for observed and latent variables. More information on Mplus modeling capabilities can be obtained at www.statmodel.com. Other multivariate modeling packages, such as LISREL (SSI, 2005) have recently adopted the PML method as well.

The three fundamental sampling designs, WR, WOR and WORUNEQ, pioneered in SUDAAN, are widely used in practice and are being adopted in other software packages. The WOR design is a stratified multistage sampling design with equal probabilities without replacement sampling at the PSU level and equal probabilities with or without replacement sampling at the subsequent stages. The WORUNEQ is a stratified multistage design with unequal probabilities without replacement sampling at the PSU level and with or without replacement equal probabilities sampling at subsequent stages.

The contributions of this paper are as follows. In Section 2, we expand Skinner's (1989) PML method to the WOR and WORUNEQ designs. We also discuss the asymptotic properties of the PML estimator. It is surprising that this flexible estimation method has not been developed yet for these common sampling designs. In Section 3 we conduct a simulation study on a factor analysis model estimated from a two-stage WOR design. In Section 4 we evaluate the performance of four different estimators for samples with small number of PSUs. The four estimators are PML, implemented in Mplus, the GEE method implemented in SUDAAN, the GEE with exchangeable correlation method implemented in SUDAAN and the bias corrected PML method we propose in this article. In Section 5 we investigate the distribution of the likelihood ratio test statistic based on the pseudo log-likelihood value and describe an adjustment that gives a correct chi-square distribution. The effects of various complex sampling features on the distribution of the LRT statistic are illustrated with a simulation study based on the Behrens-Fisher problem in a stratified cluster sampling design. All computations are performed with Mplus, Version 3 (Muthen & Muthen, 1998-2004), unless explicitly noted.

2 Pseudo Maximum Likelihood Estimation in Multistage Sampling

In this section we describe the pseudo maximum likelihood estimation for a general parametric model and the three sampling designs WR, WOR, and WORUNEQ. Suppose that the log-likelihood for individual i is L_i and the model parameters are θ. Let T_i be the vector of first derivative of L_i with respect to θ. Suppose that w_i are the weights produced by the complex sample design, i.e., $w_i = 1/p_i$, where p_i is the probability that individual i is included in the sample. Let n be the size of the sample population and

N be the size of the whole target population. The true model parameters θ_0 are defined as the unique values that maximize the likelihood of the target population

$$L_0 = \sum_{i=1}^{N} L_i.$$

The PML estimates $\hat{\theta}$ are defined as the parameters that maximize the weighted sample log-likelihood

$$L = \sum_{i=1}^{n} w_i L_i.$$

These estimates are obtained by solving the weighted score equations

$$\sum_{i=1}^{n} w_i T_i = 0.$$

For large sample size the weighted sample score equation is an approximation to the total score equation

$$\sum_{i=1}^{n} w_i T_i \approx \sum_{i=1}^{N} T_i = 0. \tag{1}$$

which is solved by the true parameter θ_0. Thus the PML estimate $\hat{\theta}$ is a consistent estimate of θ. The asymptotic variance of $\hat{\theta}$ is given by the asymptotic theory for maximization estimators (see Amemiya (1985), Chapter 4)

$$(L'')^{-1} Var(L')(L'')^{-1}, \tag{2}$$

where $'$ and $''$ denote the first and the second derivatives of the weighted sample log-likelihood. The middle term $Var(L') = Var(\sum_{i=1}^{n} w_i T_i)$ is computed according to the formulas for the variance of the weighted estimate of the total described in Cochran, Chapter 11 (1977) or RTI (2002) taking the appropriate design into account. To describe $Var(L')$ we index the individual observations by membership in each of the sampling stages. That is, individual $i_1, i_2, i_3...$ is individual in strata i_1, PSU i_2, secondary sampling unit i_3, etc. Let $n_{i_1...i_l}$ be the number of sampling subunits in sampling unit $i_1...i_l$, i.e., n_{i_1} is the number of PSUs in strata i_1, $n_{i_1 i_2}$ is the number of secondary sampling units in PSU i_2 in strata i_1, etc. Let $Z_{i_1 \dots i_r} = w_{i_1 \dots i_r} T_{i_1 \dots i_r}$ and let r be the total number of sampling stages.

$$Z_{i_1 \dots i_l} = \sum_{i_{l+1}} Z_{i_1 \dots i_l i_{l+1}} = \sum_{i_{l+1}, \dots, i_r} Z_{i_1 \dots i_r},$$

$$\bar{Z}_{i_1 \dots i_l} = \frac{1}{n_{i_1 \dots i_l}} Z_{i_1 \dots i_l},$$

$$s_{i_1 \dots i_l} = \sum_{i_{l+1}} (Z_{i_1 \dots i_{l+1}} - \bar{Z}_{i_1 \dots i_l})^T (Z_{i_1 \dots i_{l+1}} - \bar{Z}_{i_1 \dots i_l}).$$

Suppose that

$$f^*_{i_1 \dots i_l} = \begin{cases} f_{i_1 \dots i_l} & \text{if the sampling in the } i_1 \dots i_l \text{ unit is WOR} \\ 0 & \text{otherwise} \end{cases}$$

For the WR design, regardless of the number of sampling stages, the variance of the score is given by

$$Var(L') = \sum_{i_1} \frac{n_{i_1}}{n_{i_1} - 1} s_{i_1}.$$

For the WOR design, for compactness, we describe the variance of the score for a stratified 3 stage sampling design

$$Var(L') = V_1 + V_2 + V_3,$$

where

$$V_1 = \sum_{i_1} (1 - f^*_{i_1}) \frac{n_{i_1}}{n_{i_1} - 1} s_{i_1}$$

$$V_2 = \sum_{i_1, i_2} (1 - f^*_{i_1 i_2}) f^*_{i_1} \frac{n_{i_1 i_2}}{n_{i_1 i_2} - 1} s_{i_1 i_2}$$

$$V_3 = \sum_{i_1, i_2, i_3} (1 - f^*_{i_1 i_2 i_3}) f^*_{i_1} f^*_{i_1 i_2} \frac{n_{i_1 i_2 i_3}}{n_{i_1 i_2 i_3} - 1} s_{i_1 i_2 i_3}$$

For the WORUNEQ design we describe the variance of the score again for a stratified 3 stage sampling design. The probability that PSU i_2 in stratum i_1 is selected is denoted by $p_{i_2|i_1}$. The probability that both PSUs i_2 and i_2' in stratum i_1 are selected in the sample is denoted by $p_{i_2 i_2'|i_1}$

$$Var(L') = V_1 + V_2 + V_3,$$

where

$$V_1 = \sum_{i_1}\sum_{i_2}\sum_{i_2'>i_2} \frac{p_{i_2|i_1}p_{i_2'|i_1} - p_{i_2 i_2'|i_1}}{p_{i_2 i_2'|i_1}}(Z_{i_1 i_2} - Z_{i_1 i_2'})^2$$

$$V_2 = \sum_{i_1,i_2}(1 - f^*_{i_1 i_2})p_{i_2|i_1}\frac{n_{i_1 i_2}}{n_{i_1 i_2} - 1}s_{i_1 i_2}$$

$$V_3 = \sum_{i_1,i_2,i_3}(1 - f^*_{i_1 i_2 i_3})p_{i_2|i_1}f^*_{i_1 i_2}\frac{n_{i_1 i_2 i_3}}{n_{i_1 i_2 i_3} - 1}s_{i_1 i_2 i_3}.$$

The above estimation method hinges on the approximation (1) of the total score, which can be achieved if the number of PSU units is large and the residuals of the score estimation within each PSU units satisfy Lindeberg's extension of the central limit theorem (see Feller, 1968). If the number of PSU units is small however the PML parameter estimates can be substantially biased.

3 Factor Analysis Simulation Study

In this section we will evaluate the performance of the PML estimator for a two-stage WOR design for a factor analysis model. The model is described as follows

$$Y_{ij} = \mu_j + \lambda_j \eta_i + \varepsilon_{ij} \tag{3}$$

where $i = 1, ..., n$, n is the sample size, and $j = 1, ..., 5$, i.e., the observed vector for each individual is of dimension 5. Here μ_j is the intercept parameter, λ_j is the loading parameter, η_i is the factor variable, and ε_{ij} is the residual variable. The variables η_i and ε_{ij} are normally distributed zero-mean variables with variances ψ and θ_j respectfully. The parameters we use for this simulation study are as follows

$$\Theta = (\mu_1, ..., \mu_5, \lambda_1, ..., \lambda_5, \theta_1, ..., \theta_5, \psi) = \tag{4}$$

$$(2, 2.7, 3.3, 4.5, 5.5, 1, 0.7, 1.3, 1.5, 0.5, 1, 1, 1, 1, 1, 1.2).$$

First we describe the target and the sample populations. We generate a multivariate target population of 50000 individuals with 5 normally distributed outcomes with mean and variance given by model (3) with parameter values given by (4). We impose the following two-level population structure on the target population. We group the observations into 140 PSUs, the first 120

Table 1: Bias of PML Parameter Estimates for Factor Analysis Model

n	m	μ_3	λ_3	θ_3	ψ
200	20	0.054	0.045	-0.028	-0.129
500	50	0.007	0.009	-0.011	-0.035
1000	100	0.001	-0.003	-0.004	0.003
1400	140	0.003	-0.001	-0.001	0.003

are of size 250 and the remaining 20 are of size 1000. The observations are not placed at random in the PSUs. They are placed according to an ordering based on a function f. That is, the first 250 observations with the highest values of f are placed in PSU 1, the second highest 250 are placed in PSU 2, etc. After all 120 PSUs of size 250 are formed we form the remaining 20 PSUs of size 1000 again according to the order given by the f function. This method of constructing target population was used in Smith and Holmes (1989). The choice of f is to some extent critical to the type of sampling we get. Suppose that f is instead a random function independent of Y. The multi-stage sampling will then be equivalent to simple random sampling (SRS). In a model with dependent variable Y and independent variable X, a function f that depends only on X but not on Y produces non-informative random sampling. The only way to produce informative sampling is to choose f which depends on Y in addition to perhaps other variables. In this target population we choose $f_i = \sum_j Y_{ij}$, which clearly induces informative sampling.

The target population is sampled with a two-stage WOR design. Equal probability sampling is used at each stage. We vary the number m of PSUs included in the sample while the number of units sampled from the $i-$th PSU remains constant, $n_i = 10$. The total sample size is thus $n = 10m$. The ratio between the sampling weights in the large PSUs and the sampling weights in the small PSUs is 4. We use 500 replications, i.e., we sample the target population 500 times and calculate the PML estimates and their 95% confidence intervals.

Table 1 shows the bias of the PML parameter estimates and Table 2 shows the coverage of the PML confidence intervals. We see that the performance of the PML method is very good, bias is almost non-existent and the coverage for the confidence intervals is in line with expectation. The only exception is the estimation of the ψ parameter which has larger bias and consequently

Table 2: Coverage of PML 95% Confidence Intervals for Factor Analysis Model

n	m	μ_3	λ_3	θ_3	ψ
200	20	0.882	0.908	0.912	0.746
500	50	0.940	0.924	0.928	0.850
1000	100	0.950	0.950	0.946	0.926
1400	140	0.954	0.948	0.968	0.952

lower confidence interval coverage for samples with small number of PSUs. In the next section we explore further the bias that arises in model estimation from samples with small number of PSU.

4 Small Number of PSUs

In this section we explore different estimation techniques for dealing with bias that arises in small number of PSUs samples. We conduct a simulation study similar to the simulation study conducted in the previous section. For simplicity we use a two-stage WR design on a smaller target population. Here again equal probability sampling is used at each stage. The target population of size 10000 is generated as in the previous section and 14 PSUs are formed, 12 of size 500 and 2 of size 2000. The sample population again has a varying number m of PSUs while the number of units sampled from the $i-$th PSU remains constant, $n_i = 50$. The total sample size is $n = 50m$. We use 500 replications in this simulation study as well. To be able to compare various estimating techniques we choose a basic regression model for Y_1 and Y_2

$$Y_1 = \alpha + \beta Y_2 + \varepsilon.$$

Using the whole target population we get the true values of the parameters as $\alpha = 0.56$ and $\beta = 0.54$. The variance parameter of the residual ε is not included in this investigation because the methods implemented in SUDAAN do not provide an estimate for this parameter. We compare four different estimation methods. The first method is the PML method implemented in Mplus. The second method is the GEE method implemented in SUDAAN. This method is based on general estimating equations which are identical to the PML score equations and as our simulation study confirms the results produced by the two methods are identical. This observation is valid also for other models such as logistic regression. The third method is the GEE method with exchangeable correlation implemented in SUDAAN. We denote this method by GEE-Ex. The RSTEPS parameter that this method depends on did not affect the results in our simulation significantly and thus we only report the results we obtained with the default RSTEPS=1.

The forth estimation method we examine in this simulation study is a bias corrected PML method (BC) that we describe here. The first step of the BC estimation method is to construct estimates for the mean and the variance/covariance of Y_1 and Y_2, by estimating the bias of the PML mean and variance/covariance estimates. We illustrate this for a single Y variable. The PML estimate for the mean is

$$\hat{\mu}_{PML} = \frac{\sum_i w_i Y_i}{\sum_i w_i}.$$

The BC estimate for the mean is then

$$\hat{\mu}_{BC} = \frac{\sum_i w_i Y_i}{\sum_i w_i} - \hat{C}_0,$$

where $\hat{C}_0$ is an estimate for the bias C_0 of the PML estimate, i.e., if μ is the mean of Y

$$C_0 = E\left(\frac{\sum_i w_i Y_i}{\sum_i w_i}\right) - \mu.$$

The term C_0 is of the form

$$C = \frac{\hat{Z}_1}{\hat{Z}_2} - \frac{E(\hat{Z}_1)}{E(\hat{Z}_2)}.$$

Formula 6.33 in [C] provides a method for estimating such a quantity. An asymptotic estimate for C is

$$\frac{1}{\hat{Z}_2^2}\left(Var(\hat{Z}_2)\frac{\hat{Z}_1}{\hat{Z}_2} - Cov(\hat{Z}_1, \hat{Z}_2)\right).$$

Both Z_1 and Z_2 are estimates of the total quantity for the variables Y and the constant variable 1. Thus the variance/covariance terms above can be estimated just as the variance/covariance of the total score estimates given in Section 2. These estimates take into account the sampling design. The PML estimate for the variance is

$$\hat{v}_{PML} = \frac{\sum_i w_i Y_i^2}{\sum_i w_i} - \left(\frac{\sum_i w_i Y_i}{\sum_i w_i} \right)^2 .$$

The BC variance estimate is

$$\hat{v}_{BC} = \frac{\sum_i w_i Y_i^2}{\sum_i w_i} - \left(\frac{\sum_i w_i Y_i}{\sum_i w_i} \right)^2 - \hat{C}_1,$$

where $\hat{C}_1$ is an estimate of the second order moment bias of the first term $\sum_i w_i Y_i^2 / \sum_i w_i$ constructed just as the bias estimate for the mean. The covariance term is estimated from the multivariate version of the above formula. Once the mean and the variance/covariance estimates for Y_1 and Y_2 are constructed, we estimate the parameters θ=(α,β) by minimizing the quasi ML fit function

$$F(\theta) = tr(\hat{v}_{BC} \;\; v(\theta)^{-1}) - \log |\hat{v}_{BC} v(\theta)^{-1}| + (\hat{\mu}_{BC} - \mu(\theta))^T v(\theta)^{-1} (\hat{\mu}_{BC} - \mu(\theta)),$$

where $\mu(\theta)$ and $v(\theta)$ are the vector mean and variance of the (Y_1, Y_2) vector expressed in terms of the model parameters θ and the following auxiliary parameters: the mean Y_2, the variance of Y_2, and the variance of the residual in the above regression equation.

We study the properties of these four estimators for samples with small number of PSUs. Tables 3 and 4 show the bias and the MSE of the four estimators on samples with 5, 10, 15, and 20 PSUs. The PML method and the GEE method, as expected, produce identical results not only on average but in individual replications as well and are reported in the same column. The bias of the PML/GEE estimator is present for both the intercept and the slope even for $m = 20$ but as expected this bias decreases as the number of PSUs increases. The bias of the BC estimator is almost non-existent except for $m = 5$. The BC method outperforms the PML/GEE estimator in terms of both MSE and bias in this simulation. The BC method, however, may not outperform the PML method in all situations. Examples in Cochran (1977) show that sometimes this method reduces the bias while increasing the MSE of the estimates. The estimator GEE-Ex performs very poorly. This method produces large bias for both parameters and large MSE. It seems also that this bias does not disappear as the number of sampled PSUs

Table 3: Bias and Mean Squared Error for the Intercept in Linear Regression.

		PML/GEE	PML/GEE	GEE-Ex	GEE-Ex	BC	BC
	m	Bias	MSE	Bias	MSE	Bias	MSE
250	5	0.434	0.632	1.576	2.855	0.179	0.442
500	10	0.220	0.281	1.634	2.889	0.016	0.211
750	15	0.132	0.185	1.664	2.902	-0.028	0.157
1000	20	0.103	0.131	1.675	2.917	-0.026	0.111

Table 4: Bias and Mean Squared Error for the Slope in Linear Regression.

		PML/GEE	PML/GEE	GEE-Ex	GEE-Ex	BC	BC
n	m	Bias	MSE	Bias	MSE	Bias	MSE
250	5	-0.237	0.130	-0.672	0.466	-0.133	0.123
500	10	-0.125	0.063	-0.656	0.440	-0.036	0.059
750	15	-0.072	0.040	-0.642	0.419	-0.002	0.038
1000	20	-0.061	0.027	-0.649	0.427	-0.004	0.024

increases. A simulation study based on a logistic regression model produced the same results. The PML/GEE method performed well as the number of PSUs increases for logistic regression as well. In contrast the GEE-Ex method produced large bias regardless of the number of PSUs in the sample.

5 Likelihood Ratio Test in Multistage Sampling

Hypotheses involving several parameters are frequently tested in multivariate modeling. Wald's test can be used for such testing if the asymptotic variance/covariance of the parameter estimates is available. Wald's test however requires additional calculations, which sometimes are quite complex. One such example is the test of a factor analysis model against an unrestricted covariance model. When maximum-likelihood estimation is performed however the likelihood ratio test (LRT) can be obtained without additional computations and this test is frequently used for complex hypothesis testing. In this section we show how the pseudo maximum likelihood can be used to perform LRT for multistage sampling designs. The distribution of the LRT statistic based on the maximized weighted log-likelihood value is not a chi-square distribution. This distribution depends on the sampling design just as the asymptotic covariance of the parameter estimates depends on the sampling design. Here we describe an adjustment of the LRT statistic which takes into account the sampling design and produces a test statistic with a chi-square distribution. This adjustment is constructed similarly to the adjustments of the Yuan-Bentler (2000) and the Satorra-Bentler (1988) robust chi-square tests for mean and variance structures. Similar first and second order adjustments are described also in Rao-Thomas (1989) for contingency tables.

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

$$T^* = c \cdot 2(L_1 - L2), \tag{5}$$

where c is the correction factor

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

The statistic T^* has approximately a chi-square distribution with $d_1 - d_2$ degrees of freedom. The components $Tr((L_i'')^{-1}Var(L_i'))$ are easily available since they are part of the asymptotic covariance for the parameter estimates given in (2). Justification for this approximation is given in the Appendix.

We demonstrate the importance of the LRT adjustment with a simple simulation study which incorporates both cluster and stratified sampling. For simplicity we use a single outcome variable and 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.

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 $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{31} = 3$, $\mu_{12} = 0$, $\mu_{22} = 2$, $\mu_{32} = 4$. Given our choice of parameters the total mean in the two groups is 2. The total variance of y is, however, larger in the second group. We test two hypotheses by LRT. The first hypothesis T_1 is that the means in the two groups are equal and is also known as the Behrens-Fisher problem, see Scheffe (1970). 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. The second test should, however, 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 LRT statistic. These methods are as follows.

- Method A. Adjusted robust LRT which takes both the clustering and the stratification into account.
- Method B. Adjusted robust LRT which takes only the clustering into account and ignores the stratification.
- Method C. Adjusted robust LRT which takes only the stratification into account and ignores the clustering.
- Method D. Adjusted robust LRT which ignores both the clustering and the stratification.

- Method E. Unadjusted LRT.

The results of the simulation study are presented in Table 5. 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 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 LRT statistic and the rejection rates, while including the cluster information decreases the LRT statistic 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 an 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.

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 5 that the sampling features in complex sampling designs can affect dramatically the distribution of the LRT statistics and erroneous conclusions can be reached if the sampling features are not accounted for. The adjusted LRT statistic provides an effective solution for hypothesis testing with complex sampling data. The LRT adjusted statistic is implemented in Mplus, Version 3 (Muthen & Muthen, 1998-2004) for a wide variety of models and complex sampling designs.

6 Conclusion

In this article we demonstrated how the PML estimator can be used with the three basic complex sampling designs WR, WOR and WORUNEQ. The PML estimator can be utilized in advanced multivariate statistical modeling to properly account for various features of complex sampling designs. The

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

PML parameter estimates are affected only by the sampling weights while their standard errors are adjusted to reflect the effects of stratification, cluster sampling, multistage sampling, finite population sampling and unequal probability sampling. Our simulation studies showed that the PML method performs very well as long as the number of PSUs is not small. When the number of PSUs is small alternative estimator such as the bias corrected PML method described here are preferable. Our comparison with the method implemented in SUDAAN showed that the GEE method is equivalent to the PML method for linear and logistic regression. The GEE with exchangeable correlation method performed poorly in our simulation study. The main advantage of the PML method however is its generality. This method can be used to estimate any parametric model.

In this article we described an adjustment to the LRT statistic which takes into account the complex sampling design. The unadjusted LRT can lead to erroneous results when analyzing survey data, while the adjusted LRT performs correctly. Because of its simplicity of use, the adjusted LRT is a valuable alternative to other methods such as Wald's test.

The PML extension described in this article and the LRT adjustment can also be used for multilevel models via the multilevel pseudo maximum-likelihood method described in Asparouhov (2005b).

7 Appendix

We follow the ideas of Yuan-Bentler (2000) to derive a general LRT correction based on the PML method under complex sampling. The arguments below also apply to any consistent estimator obtained by maximizing an objective function l. Such estimators are called extremum estimators; see Amemiya (1985), Chapter 4.

We assume a general hypothesis testing for two nested models M_1 and M_2. Let θ_i be the true parameter values and $\hat{\theta}_i$ the parameters estimates for model M_i that maximize the pseudo log-likelihood function L_i. We are interested in the asymptotic distribution of the test statistic $T = 2(L_2(\hat{\theta}_2) - L_1(\hat{\theta}_1))$ that can be used to test the more restricted model M_1 versus the less restricted model M_2. More specifically we are interested in the asymptotic distribution of T when M_1 is correct. Since M_1 is correct θ_2 is a function of θ_1 and $L_1(\theta_1) = L_2(\theta_2)$. Let $\Delta = \partial\theta_2/\partial\theta_1$. Let $S_i = \partial L_i(\theta)/\partial\theta_i$ and $H_i = -n^{-1}\partial^2 L_i(\theta)/(\partial\theta_i)^2$. Given some basic regularity conditions (see Amemiya, Theorem 4.1.3) we have that

$$\sqrt{n}(\hat{\theta}_i - \theta_i) = O_p(1), \tag{7}$$

where n is the number of observations. Using the Taylor expansion we get that

$$L_i(\hat{\theta}_i) = L_i(\theta_i) + S_i(\theta_i)(\hat{\theta}_i - \theta_i) - \frac{1}{2}n(\hat{\theta}_i - \theta_i)^T H_i(\theta_i)(\hat{\theta}_i - \theta_i) + o_p(1) \tag{8}$$

and

$$0 = S_i(\hat{\theta}_i) = S_i(\theta_i) - nH_i(\theta_i)(\hat{\theta}_i - \theta_i) + o_p(\sqrt{n}) \tag{9}$$

Solving equation (9) for $S_i(\theta_i)$ and substituting that in (8) gives us

$$L_i(\hat{\theta}_i) = L_i(\theta_i) + \frac{1}{2}n(\hat{\theta}_i - \theta_i)^T H_i(\theta_i)(\hat{\theta}_i - \theta_i) + o_p(1) \tag{10}$$

Now

$$T = n(\hat{\theta}_2 - \theta_2)^T H_2(\theta_i)(\hat{\theta}_2 - \theta_2) - n(\hat{\theta}_1 - \theta_1)^T H_1(\theta_i)(\hat{\theta}_1 - \theta_1) + o_p(1) \tag{11}$$

The chain rule for differentiation gives us

$$S_1 = \Delta S_2. \tag{12}$$

Solving (9) for $S_i(\theta_i)$ and substituting in (12) we get that

$$H_1(\theta_1)\sqrt{n}(\hat{\theta}_1 - \theta_1) = \Delta H_2(\theta_2)\sqrt{n}(\hat{\theta}_2 - \theta_2) + o_p(1) \tag{13}$$

Solving now equation (13) for $\sqrt{n}(\hat{\theta}_1 - \theta_1)$ and substituting in (11) we get

$$T = n(\hat{\theta}_2 - \theta_2)^T \Big(H_2(\theta_2) - H_2(\theta_2)\Delta^T H_1^{-1}(\theta_1)\Delta H_2(\theta_2)\Big)(\hat{\theta}_2 - \theta_2) + o_p(1) \tag{14}$$

From equation (9) we also see that the asymptotic distribution of

$$\sqrt{n}(\hat{\theta}_i - \theta_i) \rightarrow N(0, V_i) \tag{15}$$

where

$$V_i = \frac{1}{n} H_i(\theta_i)^{-1} Var(S_i(\theta_i)) H_i(\theta_i)^{-1} \tag{16}$$

Elementary matrix algebra shows that the asymptotic distribution of T is

$$\sum_i \lambda_i \chi^2_{1i} \tag{17}$$

where χ^2_{1i} are independent chi-square distributed random variables and λ_i are the eigenvalues of

$$E = V_2\Big(H_2(\theta_2) - H_2(\theta_2)\Delta^T H_1^{-1}(\theta_1)\Delta H_2(\theta_2)\Big). \tag{18}$$

The p-values of this distribution are easy to compute following a method developed in Imhof (1961). Because θ_1 and θ_2 are not known we use $\hat{\theta}_1$ and $\hat{\theta}_2$ in equation (18).

By equation (9) we get that $S_i(\theta_i) = O_p(\sqrt{n})$. The chain rule for the second derivative gives us

$$H_1(\theta_1) = \Delta^T H_2(\theta_2)\Delta + n^{-1} S_2 \partial^2\theta_2/(\partial\theta_1)^2 = \Delta^T H_2(\theta_2)\Delta + o_p(1) \tag{19}$$

This leads us to the following alternative computation of E

$$E_2 = V_2\Big(H_2(\theta_2) - H_2(\theta_2)\Delta^T(\Delta H_2(\theta_2)\Delta)^{-1}\Delta^T H_2(\theta_2)\Big) = E + o_p(1). \tag{20}$$

While asymptotically equations (18) and (20) are equivalent, they will lead to different results for finite sample size. It is not clear which one of the two should be preferred in specific applications.

Instead of computing the exact p-value of the weighted chi-square distribution (17) we can use the following adjusted test statistic. Let

$$T^* = \frac{d}{\sum_i \lambda_i} T = \frac{d}{Tr(E)} T. \tag{21}$$

where d is the number of parameter restrictions model M_1 imposes, i.e., d is the difference between the number of parameters in the two models. The ratio

$$c = \frac{Tr(E)}{d}$$

is the correction factor. The distribution of T^* is approximated by a chi-square distribution with d degrees of freedom and thus its p-values are readily available. Again we can use E_2 in formula (21) instead of E and get an asymptotically equivalent statistic which in finite sample size may be substantially different.

Now we derive one more formula for computing T^*. Using equations (12) and (16) we get that

$$H_1(\theta_1)V_1H_1(\theta_1) = \Delta H_2(\theta_2)V_2H_2(\theta_2)\Delta^T. \tag{22}$$

Now using formula (18) and (22) we get that

$$Tr(E) = Tr(V_2H_2(\theta_2)) - Tr(V_2H_2(\theta_2)\Delta^T H_1^{-1}(\theta_1)\Delta H_2(\theta_2)) =$$

$$Tr(V_2H_2(\theta_2)) - Tr(\Delta H_2(\theta_2)V_2H_2(\theta_2)\Delta^T H_1^{-1}(\theta_1)) =$$

$$Tr(V_2H_2(\theta_2)) - Tr(V_1H_1(\theta_1)).$$

Again since θ_1 and θ_2 are not know we approximate with $\hat{\theta}_1$ and $\hat{\theta}_2$

$$Tr(E) = Tr(V_2H_2(\hat{\theta}_2)) - Tr(V_1H_1(\hat{\theta}_1)). \tag{23}$$

Formula (23) is the same as formula (5) and is also the formula implemented in Mplus. This formula has several advantages. It is computationally more efficient then formulas (18) and (20) because it does not involve the computation of Δ. It can also be used to easily compute the proper LRT adjustment when two nested hypothesis are involved as follows. Suppose that we have three models M_1, M_2 and M_3 and we have the test statistics T_1^* and T_2^* for testing M_1 versus M_3 and M_2 versus M_3. Suppose that the LRT statistics have been computed according to formulas (21) and (23). Let the correction factors be c_1 and c_2 and the degrees of freedom d_1 and d_2. We want to compute the LRT statistic T^* for testing M_1 versus M_2. Let the degrees of freedom for that test be d and the correction factor be c. We have that $d = d_1 - d_2$ and

$$cd = Tr(V_2 H_2(\hat{\theta}_2)) - Tr(V_1 H_1(\hat{\theta}_1)) = (Tr(V_3 H_3(\hat{\theta}_3)) - Tr(V_1 H_1(\hat{\theta}_1))) -$$

$$(Tr(V_3 H_3(\hat{\theta}_3)) - Tr(V_2 H_2(\hat{\theta}_2))) = c_1 d_1 - c_2 d_2.$$

Thus

$$c = \frac{c_1 d_1 - c_2 d_2}{d}$$

and

$$T^* = \frac{c_1 T_1^* - c_2 T_2^*}{c}.$$

The exact same approach was outlined in Satorra-Bentler (1999) when applied to the Satorra-Bentler (1988) chi-square statistic.

8 References

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

Asparouhov, T. (2005a). Sampling Weights in Latent Variable Modeling. Structural Equation Modeling, 12, 411-434.

Asparouhov, T. (2005b). General Multilevel Modeling with Sampling Weights. Accepted in Communications in Statistics–Theory and Methods.

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

Feller, W. (1968) Introduction to Probability Theory and its Applications, vol. 1. 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. (1998-2005). Mplus User's Guide. Third Edition. Los Angeles, CA: Muthen & Muthen

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.

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

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/

Scheffe, H. (1970). Practical Solutions of the Behrens-Fisher Problem. Journal of the American Statistical Association, 65, 1501-1508.

Scientific Software International (2005). LISREL, Version 8.7.

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.

Stapleton, L. (2005) An Assessment of Practical Solutions for Structural Equation Modeling with Complex Sample Data. Accepted in Structural Equation Modeling.

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.