Adjusting for Unequal Selection Probability in Multilevel Models: Applications to Demographic Survey Data

Kim Chantala, University of North Carolina at Chapel Hill
Chirayath Suchindran, University of North Carolina at Chapel Hill

INTRODUCTION

Demographic surveys commonly use complex sampling plans to survey populations. These sampling plans often involve sampling both clusters and individuals with unequal probability of selection. Special analysis techniques are needed to compute unbiased point estimates and variances when analyzing data collected with these methods. Some analysts decide to adjust for the sampling design by adding covariates to their model that reflect the sampling process rather than using sampling weights in the analysis. Because of the large number of variables that can be involved in the sampling process, adding unwanted complexity to the model can interfere with the scientific reasons for conducting the analysis. Our purpose is to advise those who use multilevel models (MLM) on recent advances in methods to easily correct for the sampling design characteristics by using sampling weights to analyze complex survey data and provide examples of available software packages incorporating these methods.

Extensive research in estimating single-level (population-average or marginal) models from complex survey data has resulted in the availability of several software packages (SUDAAN, svy commands in Stata, and SURVEYFREQ, SURVEYREG, etc. in SAS) that use appropriate analysis techniques for complex survey data. However, research in analysis techniques for estimating MLM from complex survey data is quite recent (Pfefferman (1993), Stapleton (2002)). Not only has this research resulted in new methods for incorporating sampling weights into latent variable models, but has emphasized an important point often overlooked by both analysts and providers of the survey data: the sampling weights used for multilevel analysis need to be constructed differently than the sampling weights used for single-level analysis. The sampling weight used in estimating single-level models is computed as the inverse of the probability that the individual was selected from the population and represents the number of individuals in the population that are likely to answer the survey in a manner similar to the individual interviewed. This type of sampling weight is commonly distributed with data from population surveys. Ideally, estimation of the multilevel models requires scaling weights at each level. Public use data may not provide this information. Special procedures are implemented by several statistical packages to handle this situation.

We first review available software packages for MLM analysis that incorporate sampling weights in analysis. Next, we use data from the National Longitudinal Study of Adolescent Health (Add Health) to show how the sampling weights for MLM need to be constructed for a of these software packages. We conclude by providing examples of estimating a multilevel model with a few of these packages.

SEM AND MLM SOFTWARE FOR COMPLEX SURVEYS

A few structural estimation modeling (SEM) software packages have added the capabilities of analyzing data collected with a complex sampling plan. The most advanced of these packages for analyzing this type of data is MPLUS, but LISREL has recently added this capability to many types of analysis. Gllamm is a user-written Stata program for estimating general latent and linear

mixed models. These SEM packages can be used to estimate MLM. Additional MLM software packages include MLWIN, HLM, PROC MIXED, and PROC NLMIXED from SAS. Not all of these packages produce the same results.

DATA USED IN EXAMPLES

Examples in this paper use data from the National Longitudinal Study of Adolescents (Add Health). Add Health is a longitudinal study of adolescents listed on grade 7-12 enrollment rosters for the 1994 -1995 academic year. A sample of 80 high schools and 52 middle schools were chosen with unequal probability of selection. Incorporating systematic sampling methods and implicit stratification in the study design ensured that these schools were representative of US schools with respect to region of country, location (urban, suburban, rural), school type (private, public, parochial), percentage of students who were white, and school size. Administrators at each school were asked to fill out a special survey that captured characteristics of the school. Add Health has collected four panels of data on adolescents: In-School (1994), the Wave I In-home Survey (1995), the Wave II In-home Survey (1996), and the Wave III In-home Survey (2001). The In-school survey included all students from sampled schools who were in attendance on the day the survey was administered. The Wave I In-home survey selected students from the enrollment rosters of the 132 schools with unequal probability of selection. Several special over-sampled groups were also recruited for the Wave I interview. These include the core sample (roughly equal-sized samples), purposively selected schools (all students selected), non-genetic supplements (Black adolescents whose parents were college graduates, adolescents whose race was Cuban, Puerto Rican, or Chinese.), the disabled sample, and the genetic supplement (biologically related adolescents, non-related adolescents living together). The Wave II and Wave III samples were selected from the Wave I respondents.

For each of these interviews, Add Health provides sampling weights that are designed for estimating population-average models. Sampling weights for the schools selected are also available. Using the final weights from Add Health, we construct a sampling weight appropriate for estimating MLM from different software packages. The weights we use in the construction of the weight for multilevel modeling are the final sampling weight for the schools and the final sampling weight for the Wave I in-home survey.

EXAMPLE

Data for the example used to illustrate the SEM and MLM software packages comes from the School Administrator Survey and the Wave I In-home survey. The analysis used in the example will estimate body mass index of the students in a school from the hours spent watching TV or using computers and availability of a school recreation center. Information on the availability of an on-site school recreation center (variable RC_S) was provided by each school. Each adolescent answered questions used to compute percentile body mass index (BMIPCT) and hours watching TV or playing video or computer games during the past week (HR_WATCH). Our example will fit a MLM with a level for the school and a level for the student. The algebraic formulas describing the model and assumptions appear below.

Student-level model (Within or Level 1):

$$(BMIPCT)_{ij} = \{\beta_{0j} + \beta_{1j}(HR\_WATCHij)\} + e_{ij}$$

where:

$$E(e_{ij}) = 0 \quad \text{and} \quad Var(e_{ij}) = \sigma^2$$

School-level Model (Between or Level 2):

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(RC\_S)_j + \delta_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(RC\_S)_j + \delta_{1j}$$

where:

$$E(\delta_{0j}) = E(\delta_{1j}) = 0, \quad Var(\delta_{0j}) = \sigma^2_{\delta 0}, \quad Var(\delta_{1j}) = \sigma^2_{\delta 1}, \quad Cov(\delta_{0j}, \delta_{1j}) = \sigma_{\delta 01}$$

In this example, we will properly estimate the model taking into account the sampling weights. The sampling weights used by the multilevel modeling software packages need to be constructed in such a way that the software can extract the needed information about sample selection from each level that was sampled.

We use two different methods of scaling the sampling weights for estimating this model. Ideally weights at all levels should be provided. If the weights are available only for the final level of sampling, then these weights are scaled by aggregation. MLWIN provides instruction for the scaling of their weights that follows the different scaling methods discussed in Pfeffermann (1998). We followed Method 2 presented in Pfefferman (1998) to scale the weights for each level in the MLM analysis. These weights were used for the MLWIN and LISREL runs. In MLWIN, it is assumed that weights are assumed to be independent of random effects.

MPLUS uses weights at both levels of sampling to constructed one scaled sampling weight for the two-level analysis. Sampling weights for use with MPLUS two-level model were constructed using instructions given in Asparouhov (Web Note 8, 2004). SAS does not provide information on how the sampling weight should be constructed for PROC MIXED. We chose to use the sampling weight constructed for MPLUS for the PROC MIXED analysis.

The results of the estimation using each package are given in Table 1. The last two columns in table 1 compare the range of parameter estimates computed with sampling weights versus the range computed if sampling weights are ignored. The range for the unweighted estimates is much smaller than the range for the weighted estimates. Although the packages produce very close estimates in the absence of weighting, the estimates becoming much more variable when the multilevel sampling weights are used in the calculation.

PROC MIXED estimates the value of all of the random effects to be more extreme than the value estimated by any of the other packages. Random effects estimated by MLWIN, MPLUS, and LISREL are all within the average of the standard errors for each effect. LISREL and MLWIN consistently estimate the fixed effects to differ by no more than one average of the standard errors, while MPLUS estimates the most extreme values.

Table 1. Results from estimation of Two-level model.

Parameter in 2-Level Model	Mplus Estimate (S.E)	PROC MIXED Estimate (S.E)	LISREL Estimate (S.E.)	MLWIN Estimate (S.E.)	Range of Estimates	
Weights used	MPML Method A[1]	MPML Method A[1]	PWIGLS Method 2[2]	PWIGLS Method 2[2]	Use Weights	Ignore Weights
Fixed Effects						
γ_{00} (Intercept for β_{0j})	60.19 (0.65)	59.09 (0.79)	57.83 (0.72)	58.52 (0.58)	2.36	0.05
γ_{01} (Slope for β_{0j})	-4.49 (0.87)	-2.74 (1.10)	-1.678 (1.06)	-1.41 (0.95)	3.08	0.08
γ_{10} (Intercept for β_{1j})	0.033 (0.016)	0.038 (0.020)	0.045 (0.018)	0.052 (0.013)	0.019	0.001
γ_{11} (Slope for β_{1j})	0.12 (0.021)	0.11 (0.027)	0.099 (0.025)	0.065 (0.022)	0.055	0.003
*Random Effects***						
$\sigma^2_{\delta 0}$ (Var (δ_{0j}))	16.27 (4.04)	24.84 (5.04)	14.13 (3.18)	12.43 (3.05)	12.41	0.53
$\sigma^2_{\delta 1}$ (Var (δ_{1j}))	0.002 (0.002)	0.009 (0.003)	0.002 (0.001)	0.001 (0.001)	0.008	0.0005
σ_{12} (Cov (δ_{0j},δ_{1j}))	-0.065 (0.067)	-0.241 (0.097)	-0.047 (0.047)	-0.007 (0.040)	0.23	0.025
σ^2 (Var (e_{ij}))	794.36 (10.12)	774.08 (8.19)	792.95 (8.72)	793.57 (8.38)	20.28	0.62

[1] MPML Method A from Web note 8, Asparouhov, T. (2004).
[2] PWGLS Method 2 from Pfefferman, (1998)

CONCLUSION

Several software packages have recently incorporated use of sampling weights to adjust for non-response and the design characteristics of complex survey data when estimating structural estimation models and multilevel models. This provides analysts with a simple method for obtaining unbiased estimates from complex survey data. Results from these packages can vary, so users are advised to examine simulation studies that compare these packages. We have provided information on how the weights for multilevel models can be constructed from population average weights. However, the collectors and distributors of complex survey data need to be aware that sampling weights must be provided for every level of sampling for these weights to be constructed.

REFERENCES

Asparouhov, T. (2004). Weighting for Unequal Probability of Selection in Multilevel Modeling, Mplus Web Notes No. 8 available from http://www.statmodel.com/

Asparouhov, T. (2004). Weighting for Unequal Probability of Selection in Latent Variable Models, Mplus Web Notes No. 7 available from /http://www.statmodel.com/

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

Pfeffermann, D., Skinner, C. J., Holmes D. J,and H. Goldstein, Rasbash, J., (1998). Weighting for Unequal Selection Probabilities in Multilevel Models. JRSS, Series B, 60, 123-40.

Stapleton, L. M., (2002) the Incorporation of Sample Weights Into Multilevel Structural Equation Models, Structural Equation Modeling, 9(4), 475-502.