Multilevel Analysis with Informative Weights

Frank Jenkins[1]
[1]Westat, 1650 Research Blvd., Rockville, MD 20850

Abstract
Multilevel modeling has become common in large scale assessments with multistage sampling and unequal probabilities of selection (Raudenbush, 2000, Koretz & McCaffrey, 2001). Pseudo-likelihood estimation with partitioned weights produce asymptotically unbiased estimates of parameters in many applications (Pfeffermann, et al, 1998). However, a case often found in evaluations, longitudinal data of individuals nested within groups, has gotten little scrutiny. In this paper simulation studies will be used to illustrate the effect of informative weights with nonresponse corrections on parameter estimation in the nested longitudinal case with correlated individual-level effects. Continuous and categorical outcomes will be examined.

Key Words: Sampling weights, multilevel, longitudinal

1. Introduction

Multilevel models are a natural choice for analyzing large-scale clustered multistage samples. These models account for clustering of responses in estimating standard errors of regression coefficients. They also allow for modeling different effects for different groups of students (i.e. random group effects) and the prediction of group effects from group-level covariates (i.e. specifying fixed cross-level interactions).

A problem with using multilevel models in large-scale studies is the proper incorporation of sampling weights into the analysis. In a single-level regression, sampling weights are applied to sums of squares and cross products. With a multilevel analysis the standard use of weights leads to inconsistent estimates of fixed and random effects (Koretz & McCaffrey, 2001).

Recently, techniques have been developed to apply sampling weights to the likelihood. The resulting multilevel pseudolikelihood is maximized to yield maximum likelihood estimates of model parameters (see Pfeffermann, et al, 1998). The resulting estimates are asymptotically unbiased over a broad range of likely analysis scenarios.

Simulation studies to gauge parameter recovery of multilevel software that utilizes the pseudo maximum likelihood approach, have usually focused on a limited range of models:

- Models with only two levels of nesting
- With categorical outcomes, usually only the intercept term is modeled at higher levels of nesting.

In response to this limitation, the current study will examine how a popular multilevel analysis program, HLM (Raudenbush, et al, 2006) recovers parameters in a typical model of individual growth. Examples of continuous and categorical outcomes will be examined in subsequent sections. HLM results will be compared to the adaptive quadrature approaches of SAS NLMixed and Stata’s GLLAMM programs.

1.1 The Generating Model

The generating model for the continuous outcome case is depicted in figure 1,

$$\begin{aligned}
&\text{Level 1: Time w/in Person}\\
&y_{ijt} = \pi_{0ij} + \pi_{1ij} * Time_{ijt} + e_{ijt}\\
&\text{Level 2: Person w/in School}\\
&\pi_{0ij} = \beta_{0j} + \zeta^{(2)}_{0ij}\\
&\pi_{1ij} = \beta_{0j} + \beta_{1j} * Treat_{ij} + \zeta^{(2)}_{1ij}\\
&\text{Level 3: School}\\
&\beta_{0j} + \gamma_{000} + \gamma_{001} * SES_j + \zeta^{(3)}_{j}
\end{aligned}$$

Figure 1: Generating model for continuous outcome case

Where y is the outcome for person i in group j at time t,

$\zeta^{(2)}_{0ji}$ and $\zeta^{(2)}_{1ji}$ are the random effects for the intercept and slope at level 2,

$\zeta^{(3)}_{0ji}$ is the random effect of the intercept at level 3 and the level 2 random effects are correlated.

The Variances of the random effects are: $\text{var}(e_{ijt}) = \sigma^2, \text{var}(\zeta^{(2)}_{0ij}) = \psi^{(2)}_0, \text{var}(\zeta^{(2)}_{1ij}) = \psi^{(2)}_1, \text{and var}(\zeta^{(3)}_{ij}) = \psi^{(3)}$. The model is standardized so the variance partitioning is in terms of percent of total variance and the size of fixed effects is in terms of standard deviation units. The factors assumed for the model are based on a couple of large scale, early childhood longitudinal evaluations.

1.2 Multilevel Likelihood

ML estimation requires the evaluation of a multilevel level Likelihood involving complex integration over the random effects. In the normal case an explicit solution is available for the integration (Pinherio & Bates, 2000) so that parameters can be maximized with respect to this likelihood. The multilevel likelihood is calculated for each level of nesting. First there is the likelihood of the repeated t measures of person ij,

$$\text{L}(\underset{\sim}{\text{y}}_{ij} \mid \zeta^{(2)}_{ij}) = \prod_{t=1}^{T} P(y_{ijt} \mid \theta_{ij}, t_{ijt}, \zeta^{(2)}_{ij}) \ ,$$

where θ_{ij} are the individual linear growth parameters for the person and t_{ij} is an indicator of time.

Next, to get the independent contribution of a person to the likelihood, person-level random effects are integrated out.

$$\text{L}(\underset{\sim}{\text{y}}_{ij}) = \int (\text{L}(\underset{\sim}{\text{y}}_{ij} \mid \zeta^{(2)}_{ij})\, \phi(\zeta^{(2)}_{ij})\, d\zeta^{(2)}_{ij} \ ,$$

where $\phi(.)$ is the normal density function.

The conditional likelihood for a school is the product of person likelihoods,

$$\text{L}(school_j \mid \zeta^{(3)}_{m}) = \prod_{i=1}^{n_j} \text{L}(\underset{\sim}{\text{y}}_{ij})$$

And the independent contribution of each school to the likelihood is assessed by integrating out school-level random effects,

$$\text{L}(school_j) = \int (\text{L}(school_j \mid \zeta^{(3)}_{m})\, \phi(\zeta^{(3)}_{j})\, d\zeta^{(3)}_{j},$$

with the final likelihood being the product of school likelihoods:

$$\mathrm{L} = \prod_{j=1}^{J} \mathrm{L}(school_j)$$

Combining the expression for the person and school likelihoods yields,

$$\mathrm{L} = \prod_{j=1}^{J} \{ \int [\prod_{i=1}^{n_j} (\int \mathrm{L}(\underset{\sim}{\mathrm{y}}_{ij} \mid \zeta_{ij}^{(2)}) \phi(\zeta_{ij}^{(2)}) \, d\zeta_{ij}^{(2)}) \;] \phi(\zeta_j^{(3)}) \, d\zeta_j^{(3)} \}$$

1.3 Weighted Analysis

Pfeffermann, et al (1998) devised a scheme for applying sampling weights to multilevel samples by defining a pseudolikelihood. The procedure is the following:

1. Partition the weights by the various levels of the model (e.g. define separate person within group & group weights)
2. Normalized the weights at each level. In one common approach the weights are normalized so that they sum to the number of units. This normalization scheme is the one adopted by several packages (e.g. HLM, LISREL) because it produces relatively unbiased results for fixed and random effects in simulation studies (Pfeffermann, et al, 1998).
3. Apply normalized weights to each level of the multilevel likelihood to create a pseudo likelihood.

The combined the pseudo likelihood is defined by applying the weights as exponents of the person and school likelihoods. These weights are the exponents wij and wj in equation 1,

$$\mathrm{L} = \prod_{j=1}^{J} \{ \int [\prod_{i=1}^{n_j} (\int \mathrm{L}(\underset{\sim}{\mathrm{y}}_{ij} \mid \zeta_{ij}^{(2)}) \phi(\zeta_{ij}^{(2)}) \, d\zeta_{ij}^{(2)})^{w_{i|j}} \;] \phi(\zeta_j^{(3)}) \, d\zeta_j^{(3)} \}^{w_j} \, , \tag{1}$$

where $w_{i|j}$ is the student weight conditional on membership in school j , and wj is the school weight. Pseudo maximum likelihood estimates of parameters are achieved by maximizing the pseudolikelihood, which is accomplished in HLM with an application of the EM algorithm.

2. Defining Informative Weights and Nonresponse Adjustments

2.1 Informative Weights

In the study we simulated informative selection probabilities with a logistic function of the outcome following Asparouhov (2006) and others. If the outcome for an individual or group is x_i, the selection probability is given by the following,

$$\pi_i = \frac{1}{1 + e^{-x_i}} \, ,$$

where x_i is the outcome and π_i is the selection probability. Using a logistic function returns a value between 0 and 1, interpretable as a probability, that is correlated with the outcome, x_i . The sampling weight is defined as the inverse of the selection probability: $w_i = 1/\pi_j$. In a simple random sample the correlation of the outcome with the weight will be about .80. The degree of correlation between the outcome and the weight can be controlled by substituting x_i^* for x_i in the logistic function, where x_i^* is a variate having a pre-defined correlation with x_i . For example if the correlation of x_i and x_i^* is set to .80, the correlation of the original outcome, x_i with the weight will be about .62.

An index of informativeness of the weights can be estimated from the data. In the present study we use an index somewhat different from that suggested in Asparouhov & Muthen (2006). The index we use is

$$I_\theta = |\theta - \hat{\theta}| / \theta ,$$

where θ is the true parameter (e.g. a regression weight) and $\hat{\theta}$ is the weighted estimate of the parameter. This index is only useful in simulations since the true parameter must known.

The present study is conducted by simulating informative student within school and school selection probabilities. Individuals and groups are selected according to these probabilities and sampling weights are defined as the inverse of the selection probabilities. These weights are normalized (within the HLM program) and applied to likelihood as in equation 2.

2.2 Nonresponse Weights

Nonresponse weights for individuals are used in most large scale studies to mitigate the bias of nonrandom nonresponse. In the hierarchical modeling context, the use of nonresponse weights violates the assumption that weights for individuals within a cluster (e.g. school) are conditional only on cluster membership. Nonresponse adjustments are calculated as a function of the response rates of demographic groups within the population. These weights are then applied to individual-within-cluster weights and the resulting weights are no longer strictly conditional on cluster membership. The simulation study will address the question of whether the use of nonresponse weighting factors biases parameter estimates in multilevel models.

To simulate nonresponse weights, the following steps are taken. First individuals are sorted into sample octiles of the outcome, simulating the classification of individuals into demographic nonresponse cells. Individuals are selected in each octile according to an a priori cell nonresponse probability. The nonresponse weight factor is the inverse of this probability, which is multiplied with the individual-within-cluster weight. Note that there is no classification error in the way of simulating nonresponse adjustments, since the focus was on whether multilevel modeling estimates were robust to the use of nonresponse weights

3. Simulations with Informative Sampling Weights and a Continuous Outcome

In this section we present results for a simulation of a normal categorical outcome according to the generating model in figure 1. The parameters for all simulations are listed in table 1. The factors are in table 2, consisting of effect size (small or large), number of time points (3 or 6), individuals per school (8 or 24), level of informativeness of the weights (high or low), and whether nonresponse weights were used, and. This results in 32 cells, simulated to a depth of 500 repetitions.

Table 1: Parameters of simulations

	Level 1 variance	*Variance random effects intercept*	*Variance random effects slope*	*Correlation level 2 random effects*	*Variance of level 3 random effects*	*N of schools*
Parameter	σ^2	$\psi_o^{(2)}$	$\psi_1^{(2)}$	$\rho_{\zeta_{0ij}^{(2)},\zeta_{1ij}^{(2)}}$	$\psi^{(3)}$	J
Values	0.25	0.5	0.15	0.7	0.1	200

Table 2: Factors varied in the simulation

Factor	*Fixed effect level 1* *Time*	*Fixed effect level 2* *Treat*	*Fixed effect level 3* *SES*	*Number of time points* n_t	*N per school* n_j	*Informative wts: hi/low* *I*
Value 1	.15	0.08	.20	3	8	.80
Value 2	.075	0.04	.10	6	24	.40

3.1 Bias in Parameters

Table 3 shows relative bias for four fixed and four variance parameters for the various factors in the simulation. The relative bias is the bias as a proportion of the parameter value (except for the intercept, where the denominator is fixed to 1).

Table 3: Relative bias in parameters for main effects of simulation factors-continuous outcome

		Bias fixed effect parameters				*Bias variance parameters*			
		Intercept	*Time*	*Treat*	*SES*	*Level 1*	*Level 2*	*Level 2*	*Level 3*
Nonresponse Wt						*Sigma*	*Psi0(2)*	*Psi1(2)*	*Psi(3)*
	Yes	0.05	0.11	-0.03	-0.05	0	-0.08	-0.01	0.14
	No	0.03	0.09	-0.02	-0.03	0	-0.05	-0.01	0.12
Effect size	*Small*	0.04	0.13	-0.03	-0.06	0	-0.07	-0.01	0.13
	Large	0.04	0.07	-0.02	-0.02	0	-0.06	-0.01	0.13
Group size	*T3/N8*	0.05	0.13	-0.05	-0.05	0	-0.09	-0.02	0.20
	T3/N24	0.02	0.05	0.00	-0.02	0	-0.04	-0.01	0.07
*# Time points**	*T3/N8*	0.05	0.13	-0.05	-0.05	0	-0.09	-0.02	0.20
	T6/N8	0.04	0.12	-0.03	-0.04	0	-0.07	-0.01	0.11
Informative weights	*High*	0.05	0.11	-0.03	-0.04	0	-0.08	-0.02	0.14
	Low	0.03	0.09	-0.02	-0.03	0	-0.05	-0.01	0.12

* The combination T6/n24 was not included, since T6 and N24 separately yielded relatively unbiased results.

Although the results are not 100 percent consistent some general trends are apparent. All simulation factors reduce bias in the expected directions, i.e. larger sample size, using less informative weights and not using nonresponse weights decrease bias. The effect of specific factors can be summarized below:

1. Informativeness:
   a) When weights are highly informative, using weights dramatically reduces the bias in fixed effects.
   b) When weights are of low informativeness, unweighted estimates have little bias but level 3 variances are biased upward to some degree.
2. N of Cases: Increasing the number of time points is more beneficial for decreasing bias than a similar increase in the number of groups. This is probably because; regarding the effect of most interest, treatment, and the unit of assignment is the individual rather than the group.
3. Nonresponse Weights: Fixed effect estimates are robust to the use of nonresponse weights. Variance estimates show a small increase in bias when these weights are used.

3.2 Power

Table 4 shows the power of the fixed effects associated with the various design factors. Power is simply the percent of the 500 simulations in which the effect was significant. As with bias, the design factors increase power in expected directions, i.e. larger sample size, using less informative weights and not using nonresponse weights increase power. Note that since the intercept is zero, smaller percent significant is better. Also, the time effect was large enough that it was always significant, not an unexpected result for early childhood longitudinal studies. When weights are relatively uninformative, there is a large loss of power when using weights for a small reduction in bias.

Table 4: Power for fixed effects by simulation factors-continuous outcome

		Intercept %	*Time %*	*Treat %*	*SES %*
Nonresponse Wt	*Yes*	30	100	67	44
	No	20	100	69	54
Effect Size	*Small*	23	100	46	24
	Large	26	100	90	74
Group Size	*T3/N8*	32	100	50	45
	T3/N24	16	100	82	54
#Time Points	*T3/N8*	32	100	50	45
	T6/N8	27	100	72	49
Informativeness	*High*	31	100	67	46
	Low	20	100	69	54

4. Simulations with Informative Sampling Weights and a Dichotomous Outcome

Pseudolikelihood estimation in multilevel models when the outcome is dichotomous is problematical because the there is no explicit solution to the multivariate integration of the likelihood. There are three approximations that can be applied in this case:

1. Penalized Quasi-Likelihood (PQL): This involves Linearizing the discrete outcome with a Taylor series expansion (2nd order). Then the outcome is treated as if it was continuous and the standard EM estimation algorithm is applied.
2. Multivariate Laplace Approximation: In this approach, the integrand of the likelihood is approximated with a Taylor expansion. For the HLM program Raudenbush and colleagues derived a 6th order Taylor series expansion of integrand (Raudenbush, et al, 2000). This is relatively fast in computing time and accurate for moderately large sample sizes. PQL is used for starting values.
3. Adaptive Gauss Quadrature (AGQ): The integration is approximated by performing multidimensional discrete quadrature. The great gain in accuracy is paid for in computing time, which can become intractable when there are several nested levels and multivariate random effects.

4.1 Generating Model for Discrete Outcome Simulation

Figure two gives the generating model for the multivariate logistic regression. The outcome is dichotomous and there is a logit link function. The rest of the model is the same as for the continuous case.

$$
\begin{aligned}
&Level\,1:\\
&y_{ijt} \sim Bernoulli(p_{ijt})\\
&Logit(p_{ijt}) = \pi_{0ij} + \pi_{1ij} * Time_{ijt}\\
&Level\,2:\\
&\pi_{0ij} = \beta_{00j} + \zeta^{(2)}_{0ij}\\
&\pi_{1ij} = \beta_{10} + \beta_{11} * Treat_{ij} + \zeta^{(2)}_{1ij}\\
&Level\,3:\\
&\beta_{00j} = \gamma_{000} + \gamma_{001} * SES_j + \zeta^{(3)}_j
\end{aligned}
$$

Figure 2: Generating model for the dichotomous outcome simulation

Problems were encountered in using the Laplace approximation in HLM for a 3-level model. The only 3-level model estimable was one without a random slope effect, ζ_{1ij} , at level 2. As a result we explored the behavior of HLM estimates with a 2-level model that had correlated random effects at level 2. The 3-level model was explored the Stata GLLAMM software, which uses an adaptive quadrature.

Table 5 gives the result of the 2-level simulation run without weights compared with the adjusted Gauss approximation approach using the SAS Proc NLMixed program.

Table 5: Bias in unweighted estimates for 2-level simulation- HLM compared to NLMixed
(2-Level results: No weights: Proportional bias (bias/theta))
(Intercept=0, Time=.15, Treat=.08, V(Intercept)=.50, V(Slope)=.15, R=.70. NSchools= 200)

Model		*HLM Laplace*	*HLM Laplace*	*HLM Laplace*	*NLMixed AGQ*
# Tm Pts		3	6	12	3
Bias Fixed	*Intercept*	- .01	.01	.00	.00
	Time	- .11	-.12	.01	-.01
	Treatment	- .11	-.15	.00	-.02
Bias Var.	$\psi_0^{(2)}$	-.49	-.31	-.13	-.02
	$\psi_1^{(2)}$	-.96	-.42	-.08	.08

HLM Laplace approximation only performs satisfactorily with 12-time points. While the adaptive Gauss-Hermite quadrature performs well with only 3-time points.

The 3-level weighted model was explored with GLLAMM (Rabe-Hesketh & Skrondal, 2008). Table 6 gives the bias of estimates with 3-time points. However, the computing time for this model was at least 100 times that of HLM. A modest simulation as a proof of concept was conducted with just 100 repetitions and 125 groups (vs. 200 in the previous simulation).

Results are mixed. First consider a simulation with three time points, labeled as (T3) in the table. Using weights yields considerable bias in the intercept and the level 3 variance. But the parameters of most interest, the fixed effects and the level 2 variances, are estimated well considering relatively small sample size (1000 cases in 125 groups). The last two columns list the results when 6 time points per person were simulated. In this case, there was about 1/3 less bias in the intercept and level 3 variance estimates. Note that unequal group size was imposed by the nonresponse selection. However GLLAMM doesn't normalize the weights as is recommended by Pfeffermann, et al (1998). An analysis with pre-normalized weights would probably yield less biased results.

Table 6: Gauss quadrature estimates of 3-level logistic regression via GLLAMM

Effect	*Name*	*True Value*	*Estimate (T3)*	*Bias (T3)*	*Estimate (T6)*	*Bias (T6)*
Fixed	*Intercept*	.00	.09	.09	.07	.07
	Time	.15	.17	.02	.16	.01
	Treat	.08	.08	.00	.07	.01
	SES	.20	.19	-.01	.19	-.01
Random	$\psi_0^{(2)}$	.50	.42	-.08	.44	-.06
	$\psi_1^{(2)}$	.15	.16	.01	.15	.00
	$\psi^{(3)}$	.10	.15	.05	.13	.03

4. Conclusions

For continuous outcomes with three time points, HLM yields relatively unbiased estimates of fixed effects but demonstrates negative bias for level 2 variances and positive bias for Level 3 variance. Power for fixed effects at levels 2 and 3 are poor except when group sizes are 24 or there are 6 or more time points. Finally, when weights are not highly informative, unweighted analysis results in little bias and more power than weighted analysis. Surprisingly, estimates were fairly robust with respect to the use of nonresponse weights.

For dichotomous outcomes, analysis using adaptive quadrature outperforms Laplace approximation in unweighted two-level analyses. For weighted 3-level analyses with the standard longitudinal model explored in this paper, adaptive quadrature via GLLAMM performs reasonably well even with relatively small sample sizes.

We conclude that unless the number of time points or individuals is large, adaptive quadrature is the recommended analysis approach for dichotomous outcomes, even though there is a large computing overhead.

It is difficult to assess how generalizable these results are. For longitudinal data and two levels of nesting, the number of possible models is unlimited, with different numbers and magnitudes of fixed and random effects. It is recommended that for large scale evaluations, issues of power and bias should be explored through simulation studies similar to those presented in this paper. Plausible ranges of effect sizes and variances of random effects can be posited. Also estimates of the informativeness of weights as well as the magnitude of nonresponse can be built into simulations to give plausible ranges of power and bias for a particular study. Efficient easy to use utility programs need to be devised that simulate data for a wide range of longitudinal designs.

For dichotomous outcomes the assessment of power is made problematical by the large computing time associated with adaptive quadrature. However, the relative efficiency of different designs can be explored using HLM and Laplace approximations, and final model can be assessed with adaptive quadrature analysis.

References

Asparouhov, T. (2006). General multilevel modeling with sampling weights, *Communication in Statistics- Theory and Method*, 35(3), 439-460.

Asparouhov, T. & Muthen, B. (2006). Multilevel modeling of complex survey data, Proceedings of the American Statistical Association, Seattle, WA: American Statistical Association.

Koretz, D. & McCaffrey, D. (2001). Using TIMSS to analyze correlates of performance variation in mathematics, US department of education, OERI, Working Paper series, No. 2001-05.

Pinheiro, J.C. & Bates, D. M. (2000). Mixed-Effects Models in S and S-Plus, Springer.

Pfeffermann, D.; Skinner, C., Goldstein, H. & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models, *Journal of the Royal Statistical Society*, Series B, 60, part 1, pp.23-40.

Rabe-Hesketh, S. and Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata.* (Second Edition). College Station, TX: Stata Press.

Raudenbush, S.W. (2000). Synthesizing Results for NAEP Trial State Assessment. In Grissmer, D.W. and Ross, Michael (Ed.), *Analytic Issues in the Assessment of Student Achievement*, Washington, DC: National center for Educational Statistics.

Raudenbush, S.; Bryk, A.; Cheong, Y.; & Congdon, R.T. (2006). *HLM 6: Hierarchical linear and nonlinear modeling.* Chicago: Scientific Software International.

Raudenbush, S.; Yang, M. & Yosef, M.(2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation, *Journal of computational and Graphical Statistics, 9(1), 141-157.*