Title
Determination of Sample Size for Multilevel Model Design

Author
David Afshartous

Publication Date
2011-10-24

available, then the value of the variable SEX assigned on the school roster was used. If SEX was still missing, it was imputed from the respondent's name. On any records for which this could not be done unambiguously, this variable had a value of 1 or 2 randomly assigned. The values for SEX are:

1 = Male
2 = Female

Label	Code	Frequency	Percent Raw	Percent WGTD
Male.....	1	12241	49.8%	50.1%
Female.....	2	12358	50.2%	49.9%
TOTALS:		24599	100.0%	100.0%

****

Variable: RACE COMPOSITE RACE
Module: 1S2 Position: 384-384

RACE was constructed from BYS31A. See NELS:88 First Follow-Up: Student Component Data Users' Manual Vol. 1 for more details on how this composite was constructed. The values for RACE are:

1 = Asian or Pacific Islander
2 = Hispanic, regardless of race
3 = Black, not of Hispanic origin
4 = White, not of Hispanic origin
5 = American Indian or Alaskan Native
8 = Missing, BYS31A was not answered or more than one race category was chosen

Label	Code	Frequency	Percent Raw	Percent WGTD
Asian or Pacific Islander.....	1	1527	6.2%	3.5%
Hispanic, regardless of race..	2	3171	12.9%	10.4%
Black, not of Hispanic origin.	3	3009	12.2%	13.2%
White, not of Hispanic origin.	4	16317	66.3%	71.6%
American Indian or Alaskan Native.....	5	299	1.2%	1.3%
MISSING.....	8	276	1.1%	(MISS)
TOTALS:		24599	100.0%	100.0%

```
*********************************************************************
Variable:  BYSES      SOCIO-ECONOMIC STATUS COMPOSITE
Module: 1S2         Position: 416-420 .3

     BYSES was constructed using the following parent questionnaire
 data:  father's education level, mother's education level, father's
 occupation, mother's occupation, and family income (data coming from
 BYP30, BYP31, BYP34B, BYP37B, and BYP80).

     For cases where all parent data components were missing (8.1
 percent of the participants), student data were used to compute the
 BYSES. The first four components from the student data are the same
 as the components used from parent data (i.e., educational-level
 data, BYS34A and BYS34B, similarly recoded; occupational data, BYS4B
 and BYS7B of student questionnaire part one, also recoded). The fifth
 component for BYSES from the student data consisted of summing the
 non-missing household items listed at BYS3A-P (after recoding "Not
 Have Item" from "2" to "0"), calculating a simple mean of these
 items, and then standardizing this mean.  The actual range for BYSES
 is -2.97 through 2.56, with 99.998 indicating - Missing.  See
 NELS:88 First Follow-Up: Student Component Data File User's Manual
 for more details.

                                                        Percent
  Label                              Code  Frequency  Raw  WGTD
-----------------------------------------------------------------
-2.97 thru 2.56.............    1.000    24588  100.0% 100.0%
  MISSING.....................   99.998       11     .0% (MISS)
                                           ----   ------ ------
 TOTALS:                                  24599  100.0% 100.0%

> (def ses (remove nil ses))
SES
> (length ses)
24588
> (mean ses)
-0.06753810802017091
> (standard-deviation ses)
0.7994172446174057
> (min ses)
-2.97
> (max ses)
2.56


*********************************************************************
Variable:  SEX        COMPOSITE SEX
Module: 1S2         Position: 383-383

     SEX was taken first from the "Your Background" (BYS12) section
 of the student questionnaire. If this source was missing or not
```

(Metropolitan Statistical Area)
3 = Rural -- outside MSA

Label	Code	Frequency	Percent Raw	Percent WGTD
Urban	1	7484	30.9%	25.0%
Suburban	2	10068	41.5%	43.5%
Rural	3	6694	27.6%	31.5%
TOTALS:		24246	100.0	100.0

URBAN IS A 0-1 INDICATOR FORMED FROM ABOVE
0 = NON-URBAN
1 = URBAN

****

Variable: G8LUNCH PERCENT FREE LUNCH IN SCHOOL
Module: 2C2 Position: 262-262

G8LUNCH categorizes the percentage of free or reduced price lunch at the school calculated from the school questionnaire. It was constructed by dividing BYSC16A by BYSC2, multiplying by 100, rounding to the nearest whole number and coding the result. If the school questionnaire was missing or if BYSC16A was missing, G8LUNCH was set to missing. The value for G8LUNCH are:

0 = None
1 = 1-5%
2 = 6-10%
3 = 11-20%
4 = 21-30%
5 = 31-50%
6 = 51-75%
7 = 76-100%
8 = Missing

NOTE: This variable was recoded by NCES in accordance with the confidentiality provisions of PL100-297 (1988).

Label	Code	Frequency	Percent Raw	Percent WGTD
None	0	4323	17.8%	11.6%
1-5%	1	3125	12.9%	14.2%
6-10%	2	2406	9.9%	10.5%
11-20%	3	3823	15.8%	17.4%
21-30%	4	3228	13.3%	14.9%
31-50%	5	3807	15.7%	16.5%
51-75%	6	2274	9.4%	10.5%
76-100%	7	1175	4.8%	4.5%
MISSING	8	85	.4%	(MISS)

Morris, Carl (1983). "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, v78, pp. 47-55.

Robinson, G.K. (1991). "That BLUP is a good thing: The estimation of random effects," *Statistical Science*, v6, pp. 15-32.

Van der Leiden, R. and Busing F. (1994). "First Iteration versus IGLS/RIGLS Estimates in Two-Level Models: A Monte Carlo Study with ML3," Technical Report PRM 94-03, Department of Psychometrics and Research Methodology, University of Leiden, Leiden, Netherlands.

Appendix A: Variable Descriptions

```
**********************************************************************
Variable:  BYTXMSTD   MATHEMATICS STANDARDIZED SCORE
Module: 1S2         Position: 483-488 .3

 Mathematics Standardized Score

                                                      Percent
  Label                              Code  Frequency  Raw  WGTD
-----------------------------------------------------------------
26.747 thru 71.222.............    1.000    23629   96.1% 100.0%
  MISSING.....................  999.998      970    3.9% (MISS)
                                            ----   ------ ------
 TOTALS:                                   24599  100.0% 100.0%

> (length math)
23629
> (mean math)
50.642
> (standard-deviation math)
10.218
> (min math)
26.747
> (max math)
71.221
**********************************************************************
Variable:  G8URBAN    URBANICITY COMPOSITE
Module: 2C2         Position: 259-259

      G8URBAN classifies the urbanicity of the student's school. It
 was created directly from QED (Quality Education Data) data (position
 199-199). The classifications are the Federal Information Processing
 Standards as used by the U.S. Census.  Classifications reflect the
 sample school's metropolitan status at the time of the 1980 decennial
 census.  The values for G8URBAN are:

           1 =  Urban -- central city
           2 =  Suburban -- area surrounding a central city
                within a county constituting the MSA
```

the scope of this paper. The sub-sampling method employed in this analysis represents a simple but effective method with which to examine multilevel models. There exists several extensions for future research:

1. Sub-sampling can be viewed from a variety of perspectives, e.g., boostrap, data compression, or sampling design. I have taken the perspective sampling design perspective in this paper. Research needs to be directed towards the other perspectives.

2. My analysis is specific to a given data set. Future research should investigate these issues with other real data sets.

3. My analysis of the covariance components relies upon a composite statistics, the correlation, that is necessarily bounded. Further research that examines the distribution of the covariance components, not a function of them as I've done here, may prove to be more illustrative.

4. Sub-sampling of both level-1 and level-2 units simultaneously would may provide additional insight.

References

Bassiri, D. (1988). "Large and Small Sample Properties of Maximum Likelihood Estimates for the Hierarchical Linear Model," Ph.D. dissertation, Michigan State University.

Bryk, A. S., and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and data analysis methods.* Newbury Park, CA: Sage Press.

Busing, F. (1993). "Distribution Characteristics of Variance Estimates in Two-level Models," Technical Report PRM 93-04, Department of Psychometrics and Research Methodology, University of Leiden, Leiden, Netherlands.

Cochran, W. (1976). *Sampling Techniques*, 3rd Edition, John Wiley & Sons, New York.

De Leeuw, Jan & Kreft, Ita (1995). "Questioning Multilevel Models," *Journal of Educational and Behavioral Statistics.* Forthcoming.

Hartigan, J.A. (1969). "Using Subsample Values as Typical Values," *Journal of the American Statistical Association*, v64, pp. 1303-1317.

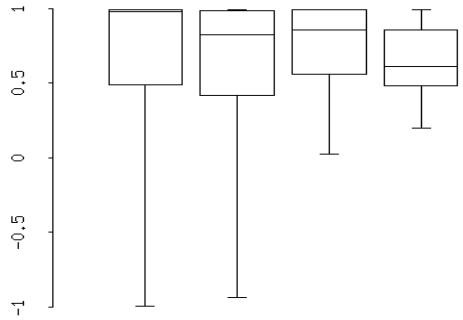
Harville, David A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation," *Journal of the American Statistical Association*, v72, pp. 320-338.

Hilden-Minton, James (1994). *TERRACE-TWO: A New Xlisp-Stat Package for Multilevel Modeling with Diagnostics*, UCLA Statistics Series.

Hilden-Minton, James (1995). *Multilevel Diagnostics for Mixed and Hierarchical Linear Models*, Ph.D. dissertation, UCLA.

Kim, K.S. (1990). "Multilevel Data Analysis: A Comparison of Analytical Alternatives," Ph.D. dissertation, UCLA.

increased to 320 schools, but this is eight times as large as the corresponding size that was necessary to produce unbiasedness for the fixed effects. With regard to the spread of our estimates, the situation is also worse than that for the fixed effects. The initial doubling of the sub-sample design has little effect; Indeed, the interquartile range actually *increases*. Moreover, given that the correlation statistic lies in the $[-1, 1]$ interval, the repeated sub-samples of 40 and 80 schools do not provide much guidance in narrowing down the original parameter space.




```
           Tau (correlation)

Intercept       1.0000   0.6002
BYSES           0.6002   1.0000
```

5 Summary

The results of this paper provide some guidelines with regard to sample size consideration. For instance, the fixed effects and variance components behave quite differently under small sample size situations. Thus, if one's research interests are mainly concerned with obtaining accurate and reliable estimates of variance components, a relatively large number of level-2 units are necessary. On the other hand, if one is solely interested in the estimates of fixed effects, the number of necessary level-2 units that are necessary decreases substantially. In either case, additional level-2 units improves the accuracy and reliability of the estimates. Moreover, the reliability of the fixed effects estimates may be related to the type of fixed effect, e.g., intercept or slope, being studied.

Although my preliminary results generally agree with the results of the Monte Carlo studies mentioned previously, a full discussion of the similarities and differences is beyond

corresponding reduction for slope fixed effects is relatively constant each time the sub-sample is doubled.

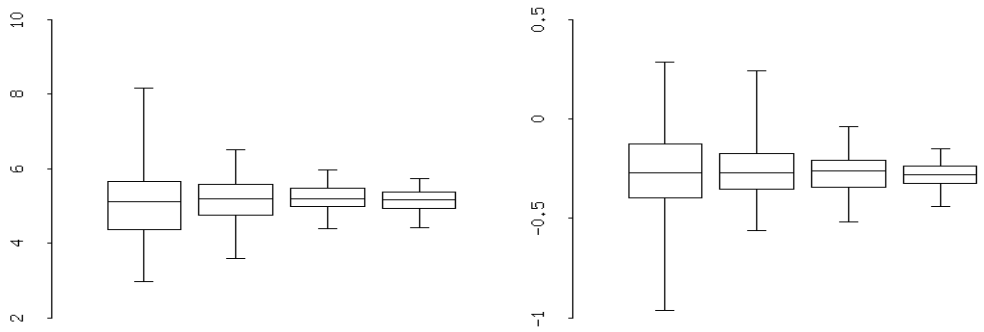


Figure 4: γ_{30} (intercept) and γ_{31} (G8LUNCH), respectively.

```
BYSES
  By Intercept        5.1726   (  0.1568)    32.9888
  By G8LUNCH         -0.2746   (  0.0411)    -6.6758
```

4.2 Variance Components

With regard to the variance components, I concentrate on the T matrix of level-2 variance components. Recall that this is a 2×2 matrix for our given model, containing elements for the estimated variance of level-1 intercept, level-1 SES effect, and covariance between them. With these three estimates, one may estimate the correlation between level-1 intercept and level-1 SES effect.[11] Thus, for each design condition, we obtain 100 values of the estimated correlation between level-1 intercept and level-1 SES slope. The estimate based on the entire data is .6, which suggests that schools with high average mathematics score are likely to exhibit a high SES effect, i.e., the impact of student SES on student mathematics score is likely to be more pronounced in such schools. The boxplot below shows the distribution of this statistic over the various sub-sample design conditions. Unlike the results for the fixed effects, a relatively unbiased estimate is unlikely to be obtained from a small samples of schools. Indeed, even for samples as large as 160 schools, the boxplot clearly demonstrates that an unbiased estimate is unlikely. Matters improve greatly once the sub-sample size is

[11] Recall that the correlation between two random variables is simply the ratio of their covariance to the square root of the product of their respective variances.

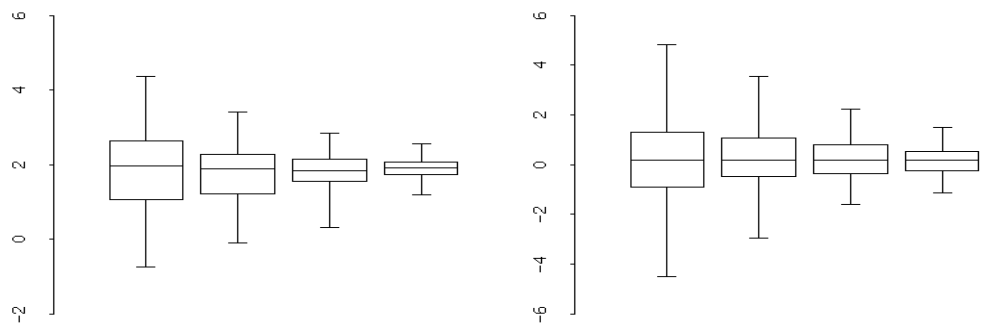


Figure 3: γ_{20} (intercept) and γ_{21} (Urban), respectively.

```
     Fixed Effects for race: 100 repeated samples of 40, 80,
     160, and 320 schools.  Complete-data estimates below:

Whites
  By Intercept       1.8862   (  0.1770)    10.6569
  By Urban           0.1684   (  0.3018)     0.5581
```

Now, let us examine the distribution of the fixed effects for the student variable SES. Recall that there exists two fixed effects with regard to the level-1 SES variable: an overall fixed effect and a fixed effect conditional upon the extent to which school lunches are subsidized. The boxplots below graphically lay out the distribution of these parameters over the different sub-sample designs. Compared to the estimates obtained from the entire data, the boxplots demonstrate that a relatively unbiased estimate of each fixed effect may be obtained from repeated sub-samples of size as small as 40 schools. With regard to variability, a substantial reduction in spread is again evident. In addition, there exists differential reduction for the two fixed effects. The reduction in variability for the intercept fixed effect is again somewhat quadratic, while that for the Lunch fixed effect behaves erratically. Thus, the reduction in spread for the intercept fixed effect is relatively more pronounced the first time the sub-sample is doubled, while there the second design condition (80 schools) impedes a simple statement about the relationship for the Lunch fixed effect.

In summary, unbiased estimates of fixed effects are readily obtainable from sub-samples of relatively small size, e.g., 40 schools. With regard to the variability of these estimates, there exists substantial improvement each time the sub-sample size is doubled. Furthermore, there exists preliminary evidence that the rate of this improvement is dependent upon the type of fixed effect being considered. Specifically, intercept fixed effects evince a proportionally greater reduction in spread the first time the sub-sample size is doubled, while the

this parameter estimate over the different sub-sample designs. Compared to the estimate obtained from the entire data, the boxplot demonstrates that a relatively unbiased of the sex fixed effect may be obtained from repeated sub-samples of size as small as 40 schools. With regard to variability, there is once again a substantial decrease in the spread of these estimates each time we double the number of schools selected. Connecting the "whiskers" of the boxplots does not form a relatively straight line; Rather, the relationship appears to be somewhat quadratic, indicating that the reduction in spread is relatively more pronounced the first time the sub-sample is doubled.

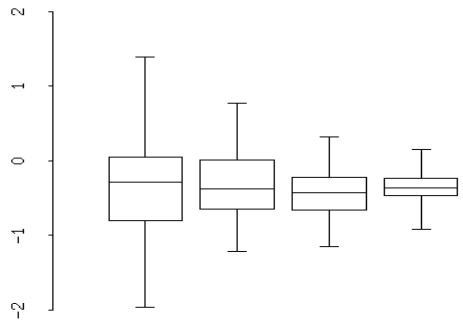


Figure 2: γ_{10} (intercept)

```
sex
  By Intercept      -0.3253   (  0.1123)    -2.8974
```

Now, let us examine the distribution of the fixed effects for the student level racial variable. Recall that there exists two fixed effects with regard to the level-1 racial variable: an overall fixed effect and a fixed effect conditional upon the urbanicity of the school. The boxplots below graphically lay out the distribution of these parameter estimates over the different sub-sample designs. Compared to the estimates obtained from the entire data, the boxplots demonstrate that a relatively unbiased estimate of each fixed effect may be obtained from repeated sub-samples of size as small as 40 schools. With regard to variability, although a substantial reduction in spread is once again evident, the two fixed effects behave somewhat differently. The reduction in variability for the intercept fixed effect is somewhat quadratic, while that for the Urban fixed effect is more linear. Thus, the reduction in spread for the intercept fixed effect is relatively more pronounced the first time the sub-sample is doubled, while there is more of a constant relationship for the urban fixed effect.

4.1 Fixed Effects

Let us first examine the fixed effects, i.e., the γ estimates. Recall that there exists three fixed effects with regard to the level-1 intercept: an overall fixed effect, a fixed effect conditional upon the urbanicity of the school, and a fixed effect conditional upon the extent to which school lunches are subsidized at the school. The boxplots below graphically lay out the distribution of these parameter estimates over the different sub-sample designs. For instance, the four boxplots in each figure correspond to a sample design condition, e.g., the boxplot on the far left of each figure displays the distribution for repeated samples 40 schools, while the boxplot on the far right of each figure displays the distribution for repeated samples of 320 schools. Estimates obtained from the entire data are given below the boxplots. Compared to the estimates obtained from the entire data, the boxplots demonstrate that a relatively unbiased estimate of each fixed effect may be obtained from repeated sub-samples of size as small as 40 schools. With regard to variability, there is a substantial decrease in the spread of these estimates each time we double the number of schools selected. Connecting the "whiskers" of the boxplots forms a relatively straight line, indicating that the relative reduction in spread is constant each time the sub-sample size is doubled.

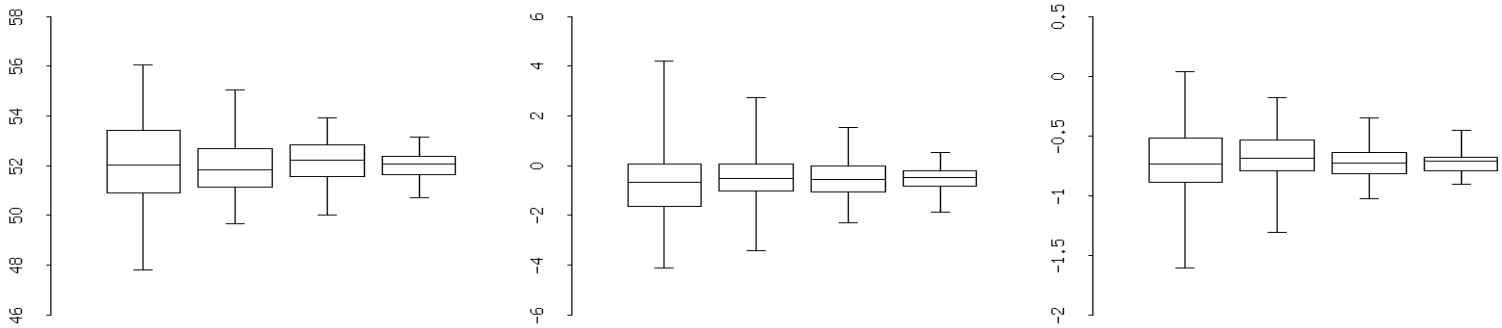


Figure 1: γ_{00} (intercept), γ_{01} (Urban), and γ_{02} (G8LUNCH), respectively.

```
Intercept
  By Intercept      52.0761   (  0.2952)   176.4347
  By Urban          -0.4413   (  0.2854)    -1.5462
  By G8LUNCH        -0.7164   (  0.0525)   -13.6528
```

Now, let us examine the distribution of the fixed effects for the student level variable sex. Recall that there exists only one fixed effect (intercept) with regard to the level-1 sex effect. Moreover, since level-1 sex was *not* modeled as random, the estimate of this fixed effect and level-1 sex effect are identical.[10] The boxplot below graphically lays out the distribution of

[10]For random level-1 variables, the corresponding estimate is obtained from the mean of a posterior distribution. However, as stated previously, these estimates are not of interest in this paper.

```
Intercept
  By Intercept      52.0761   (  0.2952)   176.4347
  By Urban          -0.4413   (  0.2854)    -1.5462
  By G8LUNCH        -0.7164   (  0.0525)   -13.6528

sex
  By Intercept      -0.3253   (  0.1123)    -2.8974

Whites
  By Intercept       1.8862   (  0.1770)    10.6569
  By Urban           0.1684   (  0.3018)     0.5581

BYSES
  By Intercept       5.1726   (  0.1568)    32.9888
  By G8LUNCH        -0.2746   (  0.0411)    -6.6758


Sigma^2:   69.0143

          Tau (covariance)

Intercept       8.1002   1.3878
BYSES           1.3878   0.6600


          Tau (correlation)

Intercept       1.0000   0.6002
BYSES           0.6002   1.0000
```

The results above are based upon analysis of over 1,000 schools. Specifically, we have $J = 1034$[9], and a given n_j for each school, ranging from 1 to 70. The sub-sampling routines were carried out under the following design conditions. Random samples of size 40, 80, 160 and 320 schools were drawn from the sample population of schools, and the above model was fit to each of these sub-samples. This procedure was repeated 100 times, thereby providing data with which to assess sampling variability of estimates both within and across the given design conditions. Thus, for each design condition, e.g., a sample of 40 schools, we have 100 values for each parameter in our given model.

[9] Eighteen schools were dropped by Terrace-Two due to missing data.

Within-school model:

$$\mathbf{Math}_{ij} = \beta_{0j} + \beta_{1j} * \mathbf{Sex}_{ij} + \beta_{2j} * \mathbf{White}_{ij} + \beta_{3j} * \mathbf{SES}_{ij} + r_{ij}$$

Between-school model:

$$\begin{aligned}
\beta_{0j} = &\quad \gamma_{00} + \gamma_{01} * \mathrm{Urban}_j + \gamma_{02} * \mathrm{GLUNCH}_j + u_{0j} \\
\beta_{1j} = &\quad \gamma_{10} \\
\beta_{2j} = &\quad \gamma_{20} + \gamma_{21} * \mathrm{Urban}_j \\
\beta_{3j} = &\quad \gamma_{30} + \gamma_{31} * \mathrm{GLUNCH}_j + u_{3j}
\end{aligned}$$

Although the substantive implications of this model, along with the diagnostic methods used to test the fit of the model, are both important and interesting, they are both beyond the scope of this paper.[7] I focus on the distribution of sub-sampled estimates. This model is estimated with respect to the entire NELS:88 data.[8] The output is as follows:

```
> Maximizing Likelihood...

                        Deviance    Method

      Iteration   1: 167354.2684    EM, init.
      Iteration   2: 166247.2691    Fisher
      Iteration   3: 165657.9647    Fisher
      Iteration   4: 165642.0690    Fisher
      Iteration   5: 165641.0928    Fisher
      Iteration   6: 165641.0408    Fisher
      Iteration   7: 165641.0380    Fisher
      Iteration   8: 165641.0378    Fisher
Final Iteration   9: 165641.0378    Fisher


TERRACE-TWO:  Full Maximum Likelihood Estimates

Parameters        Estimates     (S.E.)        T
```

[7] See Hidden-Milton (1995) for discussion of diagnostics.

[8] An XLISP-STAT program written by James Hilden-Minton, which incorporates both the EM algorithm and Fisher scoring for parameter estimation. See "Terrace-Two User's Guide: An XLISP-STAT Package for Estimating Multi-Level Models" by Afshartous & Hilden-Minton for a full description of Terrace-Two. Software and manuals accessible via World Wide Web site http://www.stat.ucla.edu. XLISP-STAT was developed by Luke Tierney and is written in the Xlisp dialect of Lisp, which was developed by David Betz.

4 Sub-Sampling Analysis

I investigate the effects of sample size on multilevel model estimates from a different perspective. Moreover, my method fits nicely into the "sample reuse" or "resampling" methods that are currently popular in various statistical literatures. I investigate an actual hierarchical data set as follows. First, care is taken to specify a reasonable two-level model with respect to the entire data. Next, repeated sub-samples of various sizes are taken from the population of level-two units (schools).[4] Given the already small number of level-1 units (students) within each level-2 unit (school), only level-2 units are sub-sampled. Under this resampling scheme, both fixed effects and covariance component estimates are observed. Finally, the results are presented in a graphical manner such that the researcher may obtain useful information, given his/her specific needs. This sub-sampling scheme may be viewed from a variety of perspectives:

- Model Bootstrap: Given an estimated model, the distribution of the sub-sampled estimates is an empirical measure of the stability of the original estimates. In addition, the location of the original estimate within this distribution provides an indicator of the "extremeness" of the original estimate.

- Data Compression: Using multilevel models to analyze large data sets is often tediously slow. It is often useful to perform exploratory analysis with a smaller portion of the data, formulating a number of hypotheses that may be examined with respect to the entire data? How large should this sub-sample be in relation to the entire data?

- Sampling Design: If one desires to collect multilevel data, how should resources be allocated to collecting data at various levels of the design, e.g., amongst schools and students.[5]

The data consists of the base-year sample from the National Educational Longitudinal Study of 1988 (NELS:88). This data set consists of 24,599 eighth grade students, distributed amongst 1052 school nationwide. (See Appendix A for more detailed description of NELS:88 data.) Given the plethora of student and school level variables available from the NELS:88 data (over 6,000), an endless number of multilevel models may be proposed and estimated. The multilevel model used in this paper represents a reasonable model that, in the interests of parsimony, relies on a small number of variables. Student mathematics score is modeled as a function of the sex, race, and socio-economic status (SES) of the student, while schools are differentiated according to urbanicity and the extent to which school lunches are subsidized.[6] Formally, the model is as follows:

[4] For a general discussion of sub-sampling methods, see Hartigan (1969).

[5] If one's sample consists of the entire population, there exists little difference between the issues involved in the previous item.

[6] The extent to which school lunches are subsidized may be considered as an indicator of the poverty level of the students within a given school.

viewed as a constrained optimization problem. Although the sampling literature (Cochran, 1976) provides some simple results for cluster sampling, there exists little discussion of this issue within the multilevel model literature, where most papers deal with estimation issues.

Depending upon one's research interests, multilevel models may be utilized for a variety of purposes. Similar to Empirical Bayes estimation (Morris, 1983), multilevel modeling is a way of "borrowing strength" in order to obtain improved estimates of individual effects (β's). On the other hand, one might be interested in the impact of a particular group-level characteristic on a specific individual effect (β_{jp}); The multilevel modeling framework provides a conceptual framework with which to model such hypotheses.[3] In addition, one might be interested in the variation and covariation of the individual effects (β's) from group to group. If so, one pays close attention to the T dispersion matrix.

Given these various aspects of multilevel modeling, the "determination of sample size" is a somewhat amorphous task. It is necessary to first specify one's priorities, e.g., from a statistical perspective, one must rank the types of inferences one wishes to make. For the purposes of this paper, two characteristics of multilevel models will be analyzed. First, the stability of the fixed effects under a variety of sample size situations will be investigated. The stability or lack thereof is directly relevant to inference associated with cross-level interaction. Second, the stability of the covariance components in the standard multilevel model will be analyzed. The stability or lack thereof is directly relevant to inference associated with level-1 parameter variance and covariance.

Although the multilevel model literature is quite extensive and growing rapidly, most of the research deals with estimation issues, e.g., producing improved algorithms to generate parameter estimates. However, there exists several articles dealing with sample size considerations. For instance, Busing (1993) and Van der Leeden & Busing (1994) examined the small sample behavior of multilevel model variance components. Bassiri (1988) examined the behavior of fixed parameter estimates, while Kim (1990) reviewed several existing estimation methods for multilevel models, paying attention to sample size considerations. Each of these studies is a Monte Carlo study, i.e., data from a known distributional form is simulated in order to learn about the sampling variability of parameter estimation.

Although the results from simulation studies are often instructive, additional research is useful. To be sure, one should be wary of situations where true models are "created." In actual data analysis, the model is never true. In the aforementioned studies, a "correct" model is fit to data from a known distribution. Given the heightened specification difficulties of multilevel models, above and beyond those of general linear models, the thought of fitting the "true" model in practice is highly optimistic. Consider the following alteration: one researcher generates the data and another unknowing researcher, given a long list of explanatory variables, fits a multilevel model to the data. Surely there will exist substantial variation in model specification, even amongst experienced users of multilevel models. To be sure, I am not advocating that Monte Carlo studies are not useful; They should be combined with analyses based on more realistic situations.

[3] In our notation, γ measures the magnitude of this impact.

models by considering the β_j as random.[1] Given that β_j is random, a second regression model may be specified as follows:

$$\beta_j = Z_j\gamma + u_j, \tag{2}$$

where Z_j is a matrix of group level explanatory variables, γ is a vector of fixed coefficients, u_j is a vector of error terms, and $u_j \sim (0, T)$. Thus, we may consider β_j as normally distributed with mean $Z_j\gamma$ and dispersion matrix T, independent of j. The diagonal elements of T represent the variances of each element of β_j, while the off-diagonal elements represent covariance between different elements of β_j. Along with σ^2, the elements of T comprise the covariance components of the multilevel model.[2] For an extensive review of covariance component estimation, see Harville (1977).

Combining (1) and (2), one obtains

$$Y_j = X_jW_j\gamma + X_ju_j + r_j. \tag{3}$$

Thus, given the aforementioned distributional assumptions, Y_j is normally distributed with mean $X_jW_j\gamma$ and dispersion matrix $V_j = X_jTX_j{}' + \sigma^2 I_{n_j}$. The combined model clearly demonstrates that a multilevel model is a special mixed model, since Y_j is modeled according to both fixed (γ) and random (u_j, r_j) effects.

3 Sample Sizes

The purpose of this paper is to examine the small sample properties of multilevel model estimates. Or, since the total sample size is merely the sum of the level-1 units, this problem is similar to the problem of examining the behavior of parameter estimates under various specifications of level-1 (n_j) and level-2 (J) sample size. For example, if one were planning to gather educational data on a national scale, one would need to determine (amongst other things) two things:

1. How many schools to sample.

2. How many students to sample from each school.

The differential monetary costs of these two processes makes sample size determination an important issue. For instance, although it may be relatively inexpensive to obtain information from an additional student within an already sampled school, the sampling of an additional school may be prohibitively costly. Thus, sample size determination may be

[1] Some or all of the elements of β_j may be considered random; For the sake of clarity, I will focus on the "full" model where all of the level-1 coefficients are random. To be sure, a fixed constant may be viewed as a random variable as well.

[2] A generalization of this model would consist of separate ${\sigma_j}^2$ for each j. However, this generalization is often difficult to implement, since small n_j makes the estimation of ${\sigma_j}^2$ difficult.

1 Introduction

Statisticians and social scientists must often analyze data that comes in hierarchical form. A classic example is educational data, where students are nested within schools. There exists a growing literature concerning the statistical techniques that should be employed to analyze such data. To be sure, one technique is to ignore the hierarchical structure in the data and merely employ conventional techniques, by either aggregating or disaggregating the data. However, the problems associated with such methods are well documented, e.g., ecological fallacies, underestimated standard errors, etc. (See De Leeuw & Kreft, 1995, and Bryk & Raudenbush 1992, for a review). Multilevel modeling is an increasingly popular technique for analyzing hierarchical data. The major purpose of this paper is to investigate the small sample properties of multilevel model estimates, thereby providing information with which to guide sample size considerations. If one desires to gather multilevel data on a large scale, the cost savings incurred by having a firm understanding of sample size determination could be quite significant.

Although relevant software has surfaced only in the past ten to fifteen years, relevant multilevel model theory and applications have been around for quite some time. To be sure, when one recognizes that a multilevel model is special kind of mixed linear models, e.g., a model containing both fixed and random effects, the extensive mixed model literature becomes relevant. Mixed models are extensively employed in agriculture experiments, where Best Linear Unbiased Predictors (BLUP) of fixed and random effects have been derived by various methods. See Robinson (1991) for an excellent review. Nevertheless, since the mixed linear model does not deal with hierarchical data *per se*, the statistical issues that arise from multilevel modeling are distinct as well as similar. Given the many potential applications of multilevel models, a thorough understanding of their performance characteristics will surely aid their implementation. Before going any further, it is instructive at this point to introduce some notation.

2 Description of a multilevel model

Suppose we have N subjects naturally grouped into J units, where there are n_j subjects in the jth unit and $\sum_{j=1}^{J} n_j = N$. In addition, suppose that for the J units we want to regress the response variable Y_j on a matrix of P predictor variables X_j. Thus, for the jth unit we model

$$Y_j = X_j\beta_j + r_j, \tag{1}$$

where each X_j has dimensions $n_j \times P$, and

$$r_j \sim n(0, \sigma^2 I_{n_j}).$$

These models will be referred to as level-1 models. So far, model (1) is no different than the conventional regression model. Multilevel models diverge from conventional regression

Determination of Sample Size for Multilevel Model Design *

David Afshartous
afsharto@math.ucla.edu
Department of Mathematics
University of California, Los Angeles
Los Angeles, CA 90024

KEY WORDS: Hierarchical Linear Model, Sub-sampling,
National Educational Longitudinal Study

ABSTRACT:

Multilevel modeling is an increasingly popular technique for analyzing hierarchical data. Suppose a data set consists of J level-2 units with n_j level-1 units within each level-2 unit, e.g., J schools with n_j students per school. If there are no covariates examined at either level, the scenario is identical to simple cluster sampling. Given that one wants to model clustered data, the determination of optimal values for J and n_j, for which a closed form solution does not exist, is of interest. The small sample properties of multilevel model parameter estimates provides insight to this problem. I investigate these small sample properties as follows: A fixed data set exists, from which I repeatedly sub-sample according to various specifications of J. After a reasonable model is estimated for the entire data, the same model is estimated for each of these sub-samples. The data used is the National Educational Longitudinal Study (NELS), a large multilevel data set from the U.S. Department of Education.

*Paper presented at the AERA meeting in San Francisco, April 1995