SAMPLE SIZES FOR SURVEY DATA ANALYZED WITH HIERARCHICAL LINEAR MODELS

Michael P. Cohen, National Center for Education Statistics*
555 New Jersey Avenue NW, Washington, DC 20208-5654

Key Words: Multilevel, random effects, covariance components

Abstract: Behavioral and social data commonly have a nested structure (for example, students nested within schools). Recently techniques and computer programs have become available for dealing with such data, permitting the formulation of explicit hierarchical linear models with hypotheses about effects occurring at each level and across levels. An example of such a model is given. If data users are planning to analyze survey data using hierarchical linear models rather than concentrating on means, totals, and proportions, this needs to be accounted for in the survey design. The implications for determining sample sizes (for example, the number of schools in the sample and the number of students sampled within each school) are explored.

1. Introduction and Example

There has been an upsurge in interest among behavioral and social scientists and education researchers in analyzing data in a way that accounts for the naturally occurring nested structure, for instance, in analyzing students nested within schools. Linear models appropriate for such data are called *hierarchical.* In part, the increased interest has been sparked by the availability of new software that properly handles the nested structure and facilitates the analyses. There has also been a realization that one can take advantage of the nested structure to explore relationships not amenable to other approaches.

Bryk and Raudenbush (1992), Goldstein (1987), and Longford (1993) are recommended for book-length discussions related to hierarchical linear models.

To illustrate these models, an example of Bryk and Raudenbush (1992, Chapter 4) will be summarized. This example is based on data from a subsample of the 1982 High School and Beyond Survey, a survey of high school students by the National Center for Education Statistics. The socioeconomic status (SES) of the student is a variable computed from the income, education, and occupation of the student's parents. The MEAN SES is the average over the students in the school of the SES values for the students. The following questions, quoted from Bryk and Raudenbush (1992, p. 61), were being explored:

> 1. *How much do U.S. high schools vary in their mean mathematics achievement?*
> 2. *Do schools with high MEAN SES also have high math achievement?*
> 3. *Is the strength of association between student SES and math achievement similar across schools? Or is SES a more important predictor of achievement in some schools than others?*
> 4. *How do public and Catholic schools compare in terms of mean math achievement and in terms of the strength of the SES-math achievement relationship, after we control for MEAN SES?*

These are the kinds of questions that hierarchical linear models (HLMs) can handle.

The student-level model for this example is

$$Y_{ij} = \beta_{0j} + r_{ij}$$

and the school level model is

$$\beta_{0j} = \gamma_{00} + u_{0j}.$$

The r_{ij} are mean zero, independent, normally distributed random variables, each with variance σ^2, for the $i = 1, \ldots, n_j$ students in school j. The u_{0j} are independent of each other and of the r_{ij}. They are normally distributed, each with mean zero and variance τ^2. The σ^2 are called the *student-level variances*, and the τ^2 are called the *school-level variances.*

2. Simple Two-Stage Design with a Simple Cost Function

In order to gain insight into the problem, we restrict our attention to a simple two-stage sampling design with a simple cost function. We select m schools, and from each of the m schools, we select n students (a balanced sample design). It costs C_s to include a school in the sample and an additional C_k for each student ("kid") sampled at the school. We wish to hold total sampling costs to our budgeted amount C where

$$C = C_s m + C_k mn.$$

We refer to the *first stage units* as *schools* and the *second stage units* as *students* throughout this paper in order to avoid cumbersome terminology. Of course, the results apply much more broadly (for example, to beds within hospitals or to books within libraries).

In reality we would almost certainly select the schools by a stratified design. Additional levels (e.g., school districts, classrooms) are possible. Unequal probability sampling might be used at any level. Our assumption of a balanced sample design (same number of students from each school) would almost certainly not hold exactly, but we do not expect that our results are very sensitive to this assumption, provided that the design is not too unbalanced.

3. Traditional Sample Size Determination

Hansen, Hurwitz, and Madow (1953, pp. 172-73) have developed the formula for the optimal size n for the number of students to sample from each school. It applies to estimating means, totals, and ratios. A simple approximate version of the formula is as follows:

$$n_{\text{opt}} \doteq \sqrt{\frac{C_s}{C_k} \times \frac{1-\rho}{\rho}}, \qquad (3.1)$$

where ρ is the measure of homogeneity, also called the intraclass (*intra-school* in our example) correlation coefficient. The number of schools sampled is then

$$m_{\text{opt}} = \frac{C}{C_s + C_k n_{\text{opt}}}.$$

Under the HLM model, we have

$$\rho = \frac{\tau^2}{\sigma^2 + \tau^2},$$

where σ^2 is the student level variance and τ^2 is the school level variance. It will also be convenient to work with the *variance ratio* ω defined by $\omega = \tau^2/\sigma^2$. In terms of the variance ratio, (3.1) becomes

$$n_{\text{opt}} \doteq \sqrt{\frac{C_s}{C_k} \times \frac{1}{\omega}}, \qquad (3.2)$$

so that the optimal number of students to sample from each school in the traditional setting varies inversely with the square root of the variance ratio ω.

It is perhaps worth mentioning that we are interested in finding the optimal values of n and m, not with the notion that they should be adhered to exactly, but rather with the idea that they can serve as a guide in survey planning.

4. Sample Size Determination for Hierarchical Linear Modelling

In analyzing HLM models, it is important to be able to estimate not only the regression coefficients but also the school-level and student-level variances (τ^2 and σ^2) because these quantities are of substantive interest. In this section, we first explore the sample size implications of needing to estimate τ^2 and σ^2. We then study, for a simple special case, the corresponding problem for the regression coefficients.

4.1 The Student-Level and School-Level Variances

Longford (1993, p. 58) shows that the maximum likelihood estimates of τ^2 and σ^2 have asymptotic variances

$$\text{var}(\widehat{\sigma}^2) = \frac{2\sigma^4}{mn - m} \qquad (4.1)$$

and

$$\text{var}(\widehat{\tau}^2) = \frac{2\sigma^4}{mn}\left(\frac{1}{n-1} + 2\omega + n\omega^2\right) \qquad (4.2)$$

as the number of *schools* m grows large. As before, $\omega = \tau^2/\sigma^2$ denotes the variance ratio. We aim to minimize these variances subject to the cost constraint of Section 2.: $C = C_s m + C_k mn$ where C is the total allowable cost, C_s is the cost of sampling each school, and C_k is the additional cost of sampling each student. But then $m = C/(C_s + C_k n)$ so that m can be eliminated from the equations (4.1) and (4.2)

For fixed values of C, C_s, C_k, σ^2, and ω (the latter two would have to be estimated from previous data), it is relatively easy to find the values of n and m that minimize $\text{var}(\widehat{\sigma}^2)$ or $\text{var}(\widehat{\tau}^2)$ with $m = C/(C_s+C_k n)$. We merely evaluate the variance equations for all reasonable values of n. This can be done very quickly on a computer. But the result does not convey an understanding of how the sample should be apportioned as the various parameters vary. We therefore seek analytical solutions.

Let us consider $\text{var}(\widehat{\sigma}^2)$ first. Although (4.1) is minimized subject to the cost constraint by taking n (students per school) as large as possible, in fact, $\text{var}(\widehat{\sigma}^2)$ is relatively flat even for moderate n. It is (4.2), again subject to the cost constraint, that is the critical one to minimize.

The expression for minimizing $\text{var}(\widehat{\tau}^2)$ with $m = C/(C_s+C_k n)$ reduces to solving a fourth degree polynomial in n. We have obtained the solution, but the expression is too cumbersome to be of any practical use. We can, however, study the closely related expression

$$\mathrm{var}(\hat{\tau}^2) \doteq \frac{2\sigma^4}{mn}\left(\frac{1}{n} + 2\omega + n\omega^2\right) \qquad (4.3)$$

where we have replaced $n-1$ by n in the denominator of the first term. We have made informal numerical comparisons of (4.2) and (4.3) and found, in our experience, that the best values of n are usually the same and, if not, almost always within one for the two equations. See Figure 1 for an illustration. It turns out, moreover, that (4.3) is the correct asymptotic expression for $\mathrm{var}(\hat{\tau}^2)$ when σ^2 is *known* (cf. Longford, 1993, p. 59). Although σ^2 would seldom be known in practical situations, the expressions should be close asymptotically (that is, when the number of schools m is large).

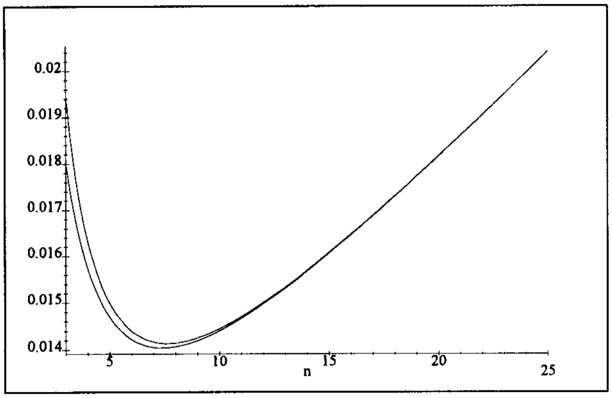

Figure 1: Plot of Two Expressions for $\mathrm{var}(\hat{\tau}^2)$

The solution to (4.3), subject to $C = C_s m + C_k mn$, is

$$\begin{aligned} n_{\text{opt}} &= \frac{\sqrt{C_k(C_k + 8C_s\omega)} + C_k}{2C_k\omega} \\ &= \frac{1}{2\omega} + \sqrt{2\frac{C_s}{C_k} \times \frac{1}{\omega} + \frac{1}{4\omega^2}}. \end{aligned}$$

In particular, for small values of the variance ratio ω, n_{opt} will be inversely proportional to ω. This contrasts with the traditional case of (3.2) where n_{opt} is inversely proportional to the *square root* of ω. For small ω, estimation of τ^2 requires a larger sample of students within each school (and hence fewer schools) for a fixed cost than does estimation of traditional quantities (means, totals, ratios).

4.2 The Regression Coefficients

It is also, of course, important to be able to estimate the regression coefficients themselves. We denote, as usual, the vector of regression coefficients by $\boldsymbol{\beta}$, the design matrix by $\mathbf{X}$, and the vector of outcomes by $\mathbf{y}$. Then the maximum likelihood estimator of $\boldsymbol{\beta}$ and its covariance matrix are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{y} \quad \text{and}$$

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{X})^{-1},$$

(Longford, 1993, p. 54), where $\mathbf{V}$ is a matrix of the form

$$\mathbf{V} = \begin{pmatrix} \tau^2\mathbf{J}_n & & & & \\ & \tau^2\mathbf{J}_n & & 0 & \\ & & \ddots & & \\ & 0 & & \tau^2\mathbf{J}_n & \\ & & & & \tau^2\mathbf{J}_n \end{pmatrix} + \sigma^2\mathbf{I}_{mn}.$$

We are using $\mathbf{I}_d$ to denote the $d \times d$ identity matrix and $\mathbf{J}_d$ to denote the $d \times d$ matrix of all 1's. So $\mathbf{V}$ is a block diagonal matrix with entries of $\tau^2 + \sigma^2$ on the main diagonal, entries of τ^2 in the blocks but off the main diagonal, and 0's elsewhere. Note, in particular, that for $\tau^2 = 0$, $\mathbf{V}$ reduces to $\sigma^2\mathbf{I}_{mn}$, and the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ reduces to the familiar ordinary least squares estimator $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$.

Investigating the properties of the estimators of the regression coefficients is made difficult by the dependence on the design matrix $\mathbf{X}$. We will only consider here a very simple design for a very balanced situation. We will let the first column of $\mathbf{X}$ be all 1's; this corresponds to estimating an intercept term in $\boldsymbol{\beta}$. The second column of $\mathbf{X}$ will be a student-level indicator ("dummy") variable, and the third column will be a school-level indicator variable. We assume the student-level indicator variable is balanced within a school and that the school-level indicator is balanced overall. This design is illustrated in (4.4) for the case of $n = 6$ students sampled per school and $m = 2$ schools sampled (but we are really interested in large m).

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix} . \quad \begin{matrix} \left.\begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix}\right\} \text{school 1} \\ \left.\begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix}\right\} \text{school 2} \end{matrix} \qquad (4.4)$$

When m and n are both even, an explicit expression can be derived for the matrix $\mathrm{cov}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}^{\mathsf{T}}\mathbf{V}\mathbf{X}\right)^{-1}$ in terms of σ^2, ω, m, and n:

$$\begin{pmatrix} \frac{\sigma^2(2n\omega+3)}{mn} & -\frac{2\sigma^2}{mn} & -\frac{2\sigma^2(n\omega+1)}{mn} \\ -\frac{2\sigma^2}{mn} & \frac{4\sigma^2}{mn} & 0 \\ -\frac{2\sigma^2(n\omega+1)}{mn} & 0 & \frac{4\sigma^2(n\omega+1)}{mn} \end{pmatrix}.$$

Let us minimize $\text{var}(\beta_0) = \frac{\sigma^2(2n\omega+3)}{mn}$, $\text{var}(\beta_1) = \frac{4\sigma^2}{mn}$, and $\text{var}(\beta_2) = \frac{4\sigma^2(n\omega+1)}{mn}$ subject to the simple cost constraint $C = C_s m + C_k mn$. The results are

$$\begin{aligned} n_{\text{opt},0} &= \sqrt{\tfrac{3}{2}\tfrac{C_s}{C_k} \times \tfrac{1}{\omega}}, \\ n_{\text{opt},1} &= \tfrac{C-C_s}{C_k}, \qquad \text{and} \\ n_{\text{opt},2} &= \sqrt{\tfrac{C_s}{C_k} \times \tfrac{1}{\omega}}, \qquad \text{respectively.} \end{aligned}$$

The $n_{\text{opt},2}$ value is the same and the $n_{\text{opt},0}$ value is similar to that obtained in the traditional case (3.2). The $n_{\text{opt},1}$ value is equivalent to $m_{\text{opt},1} = 1$; we should only sample one school (were this practical) if we *only* want to estimate β_1. The variance of β_1, though, will be small for any reasonable design (no n in the numerator of the variance expression) so other considerations are more important.

The author has informally explored some more complicated and less balanced cases, and the results were qualitatively like those given above. The variance of β_1 may depend on ω (hence τ^2) but, in the cases looked at, does so in a bounded way.

It seems that traditional sample designs may do very well in enabling us to estimate the regression coefficients. Estimating the variance components, τ^2 in particular, could present additional difficulties.

5. Final Comment

As hierarchical models become more widely used by researchers analyzing survey data, the need grows for survey design statisticians to understand the implications of such use for good survey design. This paper is the beginning of an effort to develop such an understanding. But we have scarcely scratched the surface. Opportunities abound for further research on this topic.

REFERENCES

Bryk, A.S., and Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, California: Sage.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research.* London: Griffin.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory,* Volume II (Theory). New York: Wiley.

Longford, N.T. (1993). *Random Coefficient Models.* Oxford: Clarendon.

*This paper is intended to promote the exchange of ideas among researchers and policy makers. The views are those of the author, and no official support by the U.S. Department of Education is intended or should be inferred.