SAMPLE SIZE CONSIDERATIONS FOR MULTILEVEL SURVEYS

Michael P. Cohen
Bureau of Transportation Statistics, U.S. Department of Transportation

Abstract

Social and economic data commonly have a nested structure (for example, households nested within neighborhoods). Recently techniques and computer programs have become available for dealing with such data, permitting the formulation of explicit multilevel models with hypotheses about effects occurring at each level and across levels. 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 neighborhoods in the sample and the number of households sampled within each neighborhood) are explored.

Keywords: hierarchical model, cost function, regression coefficients, intraclass correlation

1 Introduction

There has been an upsurge in interest in analyzing data in a way that accounts for the naturally occurring nested structure, for instance, in analyzing households nested within neighborhoods. Linear models appropriate for such data are called "hierarchical" or "multilevel." 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.

If researchers are planning to analyze data from federal surveys using hierarchical linear models rather than concentrating on means, totals, and proportions, it is best to account for this in the survey design. One important aspect of the design is the sample size at each level (for example, the number of neighborhoods in the sample and the number of households sampled within each neighborhood). Cost models can be developed to determine the most efficient allocation of the sample.

To date, there has been only a limited amount of research on this topic. A key paper is Snijders and Bosker (1993). Afshartous (1995) and Mok (1995) did empirical studies for particular datasets. The most up-to-date account is given in Chapter 10 of Snijders and Bosker (1999). Except for the author's article (Cohen, 1998), the emphasis has been on small single-purpose surveys rather than on large federal surveys.

This article will begin with a brief description of multilevel models. No prior knowledge of these models will be assumed. The roles of cost functions in survey design will then be discussed. There will be a review of traditional sample size determination as in, for

example, Hansen, Hurwitz, and Madow (1953). Extensions of previous work on sample size determination for surveys to be analyzed by multilevel analysis will be made to estimating the fixed coefficients in the multilevel model and to estimating the intraclass correlation coefficient.

2 Multilevel Models

Goldstein (1987, 1995), Bryk and Raudenbush (1992), Longford (1993), Hox (1995), Kreft and de Leeuw (1998), and Snijders and Bosker (1999) are recommended for book-length discussions related to multilevel models.

Consider a simple example. Suppose the household-level model is

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

and the neighborhood-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 $i = 1; : : : ; n_j$ persons in neighborhood 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 the household-level variances, and the τ^2 are the neighborhood-level variances. This is a two-stage model. We could, of course, have further levels below the household (persons, person trips); we could have more levels above the neighborhood (city, state or province).

3 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 neighborhoods, and from each of the m neighborhoods, we select n households (a balanced sample design). It costs C_2 to include a neighborhood in the sample and an additional C_1 for each household sampled at the neighborhood. We wish to hold total sampling costs to our budgeted amount C where

$$C = C_2 m + C_1 mn:$$

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

In reality we would almost certainly select the neighborhoods by a stratified design. Additional levels (e.g., cities, persons) are possible. Unequal probability sampling might be used at any level. Our assumption of a balanced sample design (same number of households from each neighborhood) 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.

4 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 households to sample from each neighborhood. It applies to estimating means, totals, and ratios. A simple approximate version of the formula is as follows:

$$n_{opt} \doteq \sqrt{\frac{C_2}{C_1} \pounds \frac{1 ¡ ½}{½}}; \tag{1}$$

where ½ is the measure of homogeneity, also called the intraclass correlation coe¢ cient. The number of neighborhoods sampled is then

$$m_{opt} = \frac{C}{C_2 + C_1 n_{opt}}:$$

In the two-level setting, we have

$$½ = \frac{¿^2}{¾^2 + ¿^2};$$

where $¾^2$ is the household level variance and $¿^2$ is the neighborhood level variance. It will also be convenient to work with the variance ratio ! de…ned by ! = $¿^2/¾^2$. In terms of the variance ratio, (1) becomes

$$n_{opt} \doteq \sqrt{\frac{C_2}{C_1} \pounds \frac{1}{!}}; \tag{2}$$

so that the optimal number of households to sample from each neighborhood in the traditional setting varies inversely with the square root of the variance ratio ! .

It is perhaps worth mentioning that we are interested in …nding 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.

5 Sample Size Determination for Regression Coe¢ cients

For household i in neighborhood j, let us consider the simple the multilevel model

$$Y_{ij} = ¯_{0j} + r_{ij}$$

where

$$¯_{0j} = °_{00} + °_{01} z_{1j} + ¢¢¢ + °_{0q} z_{qj} + u_{0j}$$

and the $\{r_{ij}; u_{0j}\}$ are mutually independent random variables with $E(r_{ij}) = E(u_{0j}) = 0$, $var(r_{ij}) = ¾^2$, and $var(u_{0j}) = ¿_0^2$. Notice that this simple model has no explanatory variables at the household level.

Suppose we want to estimate $a^0°$ where $° = (°_{00}; : : : ; °_{0q})^0$ and a is a vector of constants $(a_0; : : : ; a_q)^0$. This includes the case in which we are mainly interested in estimating a single coordinate of ° . Let $\hat{°}$ be an (asymptotically e¢ cient) estimator of ° . Let m denote the

number of neighborhoods in the sample; let n denote the number of households in each neighborhood in the sample (assumed constant); and let $E(z_j) = \gamma$ and $\mathrm{var}(z_j) = \Sigma_z$. As in Snijders and Bosker (1993, pp. 248–249),

$$\mathrm{var}(\hat{a}) \approx \frac{1}{m}\left(\sigma_0^2 + \frac{\sigma^2}{n}\right)\left(a'(1\ z)^{-1}\begin{pmatrix}\tau_0^2 & \\ & \Sigma_z\end{pmatrix}^{-1}a\right).$$

They show that for the cost model $C = C_2 m + C_1 mn$,

$$n_{\mathrm{opt}} = \sqrt{\frac{C_2\sigma^2}{C_1\tau_0^2}}.$$

For this choice of n,

$$\mathrm{var}(\hat{a}) \approx \frac{1}{C}\left(\sqrt{C_1\sigma^2} + \sqrt{C_2\tau_0^2}\right)^2 \left(a'(1\ z)^{-1}\begin{pmatrix}\tau_0^2 & \\ & \Sigma_z\end{pmatrix}^{-1}a\right).$$

Clearly, if we want to know the total cost C needed to achieve a specified value of $\mathrm{var}(\hat{a})$, this will be

$$C \approx \frac{1}{\mathrm{var}(\hat{a})}\left(\sqrt{C_1\sigma^2} + \sqrt{C_2\tau_0^2}\right)^2 \left(a'(1\ z)^{-1}\begin{pmatrix}\tau_0^2 & \\ & \Sigma_z\end{pmatrix}^{-1}a\right).$$

6 Sample Size Determination for the Intraclass Correlation Coefficient

The variance for estimating the intraclass correlation coefficient ρ is

$$\mathrm{var}(\hat{\rho}) = \frac{2(1-\rho)^2(1+(n-1)\rho)^2}{n(n-1)(m-1)}$$

(Snijders and Bosker, 1999, p. 21). We would like to find the value of n that minimizes this expression subject to the cost constraint $C = C_2 m + C_1 mn$ (so that $m = (C_2 + C_1 n)/C$). This can be done, but the expression is cumbersome and not of any use. We will instead optimize the nearly equal expression

$$\frac{2(1-\rho)^2(1+(n-1)\rho)^2}{(n-1)^2 m}. \tag{3}$$

This gives

$$n_{\mathrm{opt}} = \frac{\sqrt{8\rho(1+C_2/C_1)+1}}{2\rho} + \frac{1}{2\rho} + 1.$$

The variance expression (3) at n_{opt} can be easily solved for C, giving the cost needed to achieve a given (approximate) variance for $\hat{\rho}$.

7 Final Remark

The importance of multilevel models among today's data analysts poses a challenge to designers of surveys. The surveys should be well designed for multilevel analysis. Research into the design of such surveys is an exciting and relatively new area.

References

[1] Afshartous, D. (1995). Determination of sample size for multilevel model design. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

[2] Bryk, A.S., and Raudenbush, S.W. (1992). Hierarchical Linear Models, Applications and Data Analysis Methods. Sage.

[3] Cohen, M.P. (1998). Determining sample sizes for surveys with data analyzed by hierarchical linear models. J. O¢ cial Statist., 14, 3, 267-275.

[4] Goldstein, H. (1987). Multilevel Models in Educational and Social Research. Gri¢ n.

[5] Goldstein, H. (1995). Multilevel Statistical Models. Edward Arnold.

[6] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Volume II (Theory). Wiley.

[7] Hox, J.J. (1995). Applied Multilevel Analysis. TT-Publicaties.

[8] Kreft, I., and de Leeuw, J. (1998). Introducing Multilevel Modeling. Sage.

[9] Longford, N.T. (1993). Random Coe¢ cient Models. Clarendon.

[10] Mok, M. (1995). Sample Size Requirements for 2-level designs in educational research. Multilevel Modelling Newsletter, 7, 2, 11-15.

[11] Snijders, T.A.B., and Bosker, R.J. (1993). Standard errors and sample sizes for two-level research. J. Educational Statist., 18, 237-259.

[12] Snijders, T.A.B., and Bosker, R.J. (1999). Multilevel Analysis. Sage.