Random cross-classifications of units can arise at any level of a data hierarchy. For example, school students may be classified both by the schools they attend and their neighborhoods of residence. This article explores the issues of efficiently modeling such data and gives an example from a study of parental choice of schools.

Multilevel Cross-Classified Models

HARVEY GOLDSTEIN
Institute of Education, London

The standard multilevel model for analyzing hierarchically structured data is now well understood and being applied in widely different areas. Although most of the applications have, in the past, been to educational data, the power of these models is being applied increasingly in new fields, such as epidemiology, area studies, growth, genetics, and in the modeling of complex survey data.

In many of these areas, data also arise that do not have a purely hierarchical structure. For example, school students might be classified both by the schools they attend and their neighborhoods of residence. Or they might be classified by their elementary/primary school and their high school/secondary school. If we are studying educational achievement, this will generally be influenced by both such classification units. Thus the total variation in secondary school achievement can be modeled as the sum of contributions from both the elementary and the secondary schools. A model that incorporates such a partitioning of the variation might be more informative than one that only models the variation between the secondary schools; the latter also might be seriously misspecified if there is substantial between-elementary school variation.

Random cross-classifications of units can arise at any level and, in the following sections, I shall introduce a simple formal description

AUTHOR'S NOTE: *My thanks are due to Barry Bastow and Jon Rasbash for their helpful comments. This work was supported by a research grant from the Economic and Social Research Council (U.K.).*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 22, No. 3, February 1994 364-375

of such models and then discuss the analysis of a real data set concerning parents' choice of secondary schools in England.

MODELS FOR RANDOM CROSS-CLASSIFICATIONS

LEVEL 2 CROSS-CLASSIFICATION

Consider first the simple example given above of a 2-level model where the *i*th student is classified by the *j*th school and the *k*th neighborhood ($j = 1, \ldots, m_1$; $k = 1, \ldots, m_2$). The response (Y) is, for example, an achievement score, and a pretest score (x) measuring intake achievement is available. We write

$$y_{i(jk)} = \beta_0 + \beta_1 x_{i(jk)} + u_{1j} + u_{2k} + e_{i(jk)} \tag{1}$$

where the level 2 variation results from the sum of random variables from school and neighborhood, with $\mathrm{var}(u_{1j}) = \sigma^2_{u1}$, $\mathrm{var}(u_{2k}) = \sigma^2_{u2}$, $\mathrm{var}(e_{i(jk)}) = \sigma^2_e$. If one of these, such as U_2, has zero variance, then (1) becomes the usual simple 2-level variance components model for students and schools. The model can be further elaborated by introducing additional explanatory variables and random parameters, including higher levels of variation. The model implies the following covariance structure.

For two students in the same school and neighborhood, the covariance between their achievement scores is $\sigma^2_{u1} + \sigma^2_{u2} + \sigma^2_e$. For two students in the same school but different neighborhoods, it is σ^2_{u1}. For two students in different schools but the same neighborhood, it is σ^2_{u2}. For two students in different schools and neighborhoods, it is 0.

We can contrast this structure with that of the level 2 unstructured model, given by

$$y_{i(jk)} = \beta_0 + \beta_1 x_{i(jk)} + u_{(jk)} + e_{i(jk)}, \tag{2}$$

which is a standard 2-level model with m_1m_2 level 2 units. In this model, the covariance between two students is nonzero only when they belong to the same school and neighborhood. We can regard (2) as a fully interactive model and a likelihood ratio test for (2) against the marginal structured model (1) is available. Note that (2) implies no consistent effect of belonging to a particular school or neighborhood, and this would seem generally an unreasonable assumption to make. Nevertheless, it can be useful to fit (2) as a preliminary model to explore patterns in the data before going on to fit the more structured model (1).

LEVEL 1 CROSS-CLASSIFICATION

Suppose each student in a school responds to a practical science task, which is rated by all of the teachers in that school and a score assigned, and this is repeated for a sample of schools. We wish to model the score as a function of explanatory variables, school effects, and teacher effects. For randomly sampled units, we can write a general model as follows

$$\begin{gathered} y_{(ij)k} = \sum_m \beta_m \chi_{m,(ij)k} + u_k + e_{1i} + e_{2j} \\ \text{with } \mathrm{var}(u_k) = \sigma_u^2,\ \mathrm{var}(e_{1i}) = \sigma_{e1}^2,\ \mathrm{var}(e_{2j}) = \sigma_{e2}^2. \end{gathered} \tag{3}$$

As in the level 2 cross-classified model, we can use the corresponding unstructured model to carry out a preliminary exploration of the data. We note that, for estimability, we do not require a complete design, so that not all of the teachers in a school need to rate each student. In the extreme case, however, where only one teacher rates each student, the between-teacher variation is confounded with the between-student variation and we have a standard two-level hierarchical model.

A more complex model arises where, instead of different teachers rating the tasks, a common sample of raters is used, the same for each school. In this case, there will exist covariances between students in different schools due to the common raters. This complicates the analysis, increasing the computational burden, but otherwise yields similar interpretations. A common example of such a situation is where the same random selection of questions or items is administered to a sample of students in different schools. So called generalizability models for test scores are of this kind, although typically the traditional estimation procedures do not recognize the complex structure of the data and so are not fully efficient and generally will also be biased.

In these more complex models, it should be noted that raters or questions are thought of as level 1 units because the principal level 1 units, namely the students, are crossed with this classification and not nested within it as is the case with neighborhoods.

ESTIMATION

The iterative generalized least squares (IGLS) estimation procedure for (1) and (2) is used in this article and is described in Goldstein (1986, 1987a, 1987b). When the random terms have a multivariate normal distribution, these yield maximum likelihood or restricted maximum likelihood estimates.

Unlike the purely hierarchical case, the computations for model (1) involve the inversion of large matrices; of order ($m_1 + m_2$). Where there is a random cross-classification at level 1, the computational problems become more severe and we shall not discuss that case.

The calculations in this article have been carried out using ML3 (Prosser, Rasbash, and Goldstein 1991), a software package for multilevel analysis. A set of ML3 macros has been written to carry out the present analyses, and future releases of ML3 will incorporate built-in procedures for cross-classified models.

DATA

The data are taken from a survey of parental choice of secondary school (Bastow 1991). Parents of 10- and 11-year-old children in their final two years of primary education were asked for their preferred choice of secondary school. Fifteen secondary schools were involved, situated in an area north of London. At the same time, information was obtained about the social background of the parents, their visits to schools, and so forth.

The present analysis uses a subsample of 509 respondents for whom information is available on preferred school and geographically closest school, together with the number of visits made to secondary schools to compare and contrast the schools and make a choice for their child.

As a result of the 1988 Education Reform Act, parental choice of school, other than the geographically closest, is encouraged and one aim of the analysis is to study the way in which parents from different social backgrounds exercise choice. The model uses an index of social status as the response variable, with the number of visits made to schools as the explanatory variable. At level 2 there is a cross-classification of preferred school by closest school. The model is thus the one given by (1), with school and neighborhood replaced by closest and preferred school. The results of the analysis are not to be interpreted in terms of causal effects on social status, but simply in terms of how social status varies across schools.

ANALYSIS

UNSTRUCTURED MODEL

We first run the unstructured model (2),

$$y_{i(jk)} = \beta_0 + \beta_1 x_{i(jk)} + u_{(jk)} + e_{i(jk)},$$

and the results are given in Table 1. A normal plot of the level 1 standardized residuals is given in Figure 1, and a normal plot of the level 2 residuals is given in Figure 2. The level 1 plot is reasonably close to linear, but on the level 2 plot, we see two extreme points. These are two groups of parents with highest average social status with a particular combination of closest and preferred school. The closest school in both cases is School 7, and the preferred schools are Schools 12 and 13, respectively a boys' and a girls' school, which had previously been grammar schools, where children are selected by their ability to do well in specially constructed tests. These schools still tend to be regarded as such by parents in the area. As expected, the number of visits is positively related to social status, and there is only a moderate intraschool correlation of 0.11.

The closest school identifies the locality in which a parent lives, and there is geographical stratification by social class. The mean social status by closest and preferred school is given in Table 2 and we see that School 7 has the second highest average value for closest school

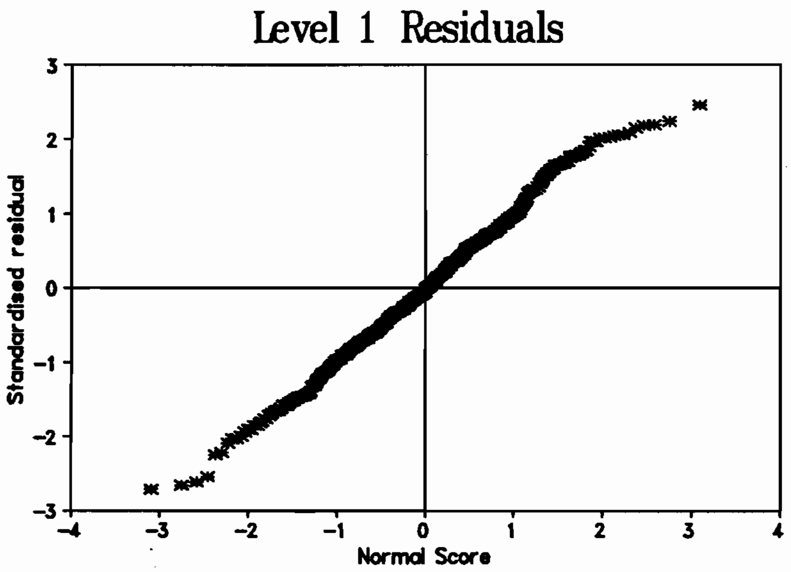

Figure 1: Level 1 Results

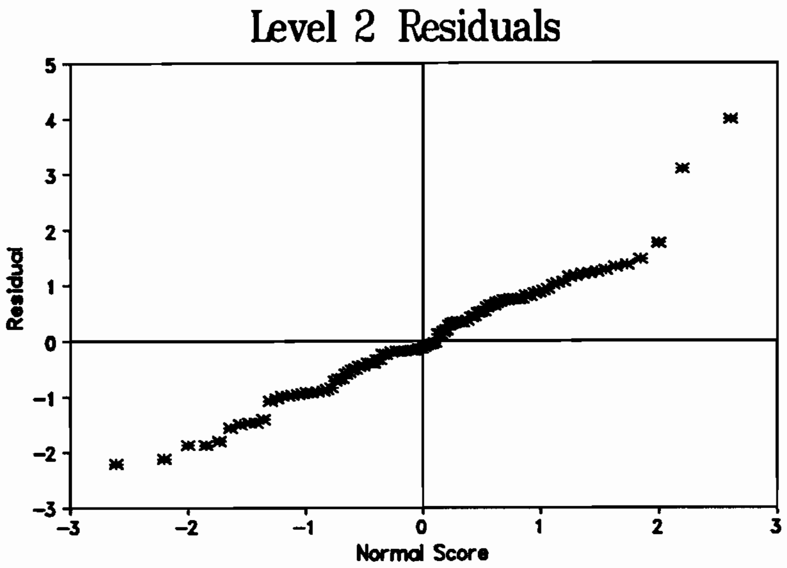

Figure 2: Level 2 Results

TABLE 1: Social Status Related to Number of School Visits (unstructured level 2 variation)

Explanatory Variable	*Estimate*	*Standard Error*
Fixed		
Intercept	21.8	
Number of visits	0.77	0.17
Random		
Level 2		
σ_u^2	3.55	1.48
Level 1		
σ_e^2	29.65	2.01

NOTE: Number of level 2 units = 108; Number of level 1 units = 509.

TABLE 2: Mean Social Status Index by School

School Number	*Closest School*	*Preferred School*
1	24.2	22.0
2	25.4	24.1
3	24.5	25.2
4	21.6	24.6
5	23.2	21.6
6	24.0	25.8
7	26.1	24.2
8	23.5	23.6
9	25.0	26.1
10	24.6	23.3
11	22.7	22.5
12	26.1	27.5
13	25.2	25.8
14	19.1	19.0
15	29.4	22.3

and 12 and 13 have the first and third highest values for preferred school. School 12 also has the second highest value for closest school and School 13 the sixth highest value. It appears, therefore, that School 12 is in a locality with a high social status index and is attractive to high social status parents generally, and the other attractive school, 13, is geographically close and also tends to have a high social status locality. It also tends to be preferred by high social status parents, although not as much as School 12, perhaps because it is a girls' grammar school rather than a boys' grammar school. School 7, on the other hand, is ranked seventh as the preferred school, so that there is a tendency for the local high social status parents to prefer other than this local school. In fact, 89 parents, or 18% of the total, live closest to School 7, and of these, just under half prefer School 7 and a further third prefer Schools 1 or 13.

THE CROSS-CLASSIFIED MODEL

We now look at the results from fitting model (1),

$$y_{i(jk)} = \beta_0 + \beta_1 x_{i(jk)} + u_{1j} + u_{2k} + e_{i(jk)},$$

and these results are given in Table 3. In overall terms, the results are very similar. An approximate comparison between the models can be obtained by computing the likelihood ratio test statistic (with one degree of freedom), which has a value of 9.16 and is highly significant.

The overall conclusions from Table 3 differ little from those of Table 1. The number of visits is positively associated with social status: the larger the number of visits the higher the social status of the parents tends to be. The total level 2 variance estimate is 3.13, which is a little less than the estimate of 3.55 in Table 1. The relatively large standard errors indicate that these separate variances are not estimated very precisely.

Table 4 gives the residual estimates for each school and we see that School 7 is the closest school with the highest residual estimate and Schools 12 and 13 are again the first and third highest residuals. Thus, after fitting the model, that is after allowing for the number of visits and the level 2 structure, our general conclusions remain the same.

SINGLE-CLASSIFICATION MODELS

It is instructive to look at the simplest models, where only the closest school or the preferred school is used as the level 2 classification. Table 5 presents the results of the analysis where closest school is chosen as the level 2 unit.

The estimate of the between-closest school variance is some 50% higher than for the cross-classified model, with an increase in the level 1 variance also. Table 6 shows the corresponding results when preferred school is chosen as the level 2 unit. As with the closest school,

TABLE 3: Social Status Related to Number of School Visits (level 2 cross-classification model)

Explanatory Variable	*Estimate*	*Standard Error*
Fixed		
Intercept	21.6	
Number of visits	0.73	0.17
Random		
Level 2		
σ^2_{u1}	1.28	0.94
σ^2_{u2}	1.85	1.13
Level 1		
σ^2_e	29.94	1.93

NOTE: The subscript 1 refers to closest school and 2 to preferred school.

TABLE 4: Residual Estimates for Each School From Cross-Classification Model

School Number	*Closest School*	*Preferred School*
1	0.2	–1.1
2	0.9	0.1
3	0.0	0.4
4	–1.4	0.3
5	–0.2	–1.3
6	–0.1	0.9
7	1.11	–0.1
8	–0.15	0.1
9	0.7	1.5
10	0.1	–0.1
11	–0.3	–0.8
12	0.9	2.5
13	0.6	1.0
14	–1.5	–1.4
15	1.0	–0.5

the between-preferred school variance estimate is increased, although this time by about 15%, reflecting the greater contribution to the level 2 variation made by this factor.

In neither of these analyses are the inferences concerning the fixed part of the model altered appreciably. The important distinction is made in the random part, where the choice of unit at level 2 can substantially alter conclusions about the relative amount of between-

TABLE 5: Closest School as Level 2 Unit

Explanatory Variable	*Estimate*	*Standard Error*
Fixed		
Intercept	21.7	
Number of visits	0.79	0.17
Random		
Level 2		
σ^2_{u1}	1.95	1.14
Level 1		
σ^2_e	31.36	1.99

TABLE 6: Preferred School as Level 2 Unit

Explanatory Variable	*Estimate*	*Standard Error*
Fixed		
Intercept	21.7	
Number of visits	0.74	0.17
Random		
Level 2		
σ^2_{u2}	2.11	1.18
Level 1		
σ^2_e	30.81	1.96

TABLE 7: Residual Estimates From Single-Classification Models

School Number	*Closest School*	*Preferred School*
1	0.0	–1.3
2	0.8	0.0
3	0.0	0.4
4	–2.0	0.3
5	–0.5	–1.5
6	–0.2	0.8
7	1.4	–0.1
8	–0.5	–0.2
9	0.9	1.5
10	0.1	–0.1
11	–0.7	–1.1
12	1.2	2.7
13	0.6	1.1
14	–2.3	–1.8
15	1.4	–0.7

unit variation. This is analogous to the usual situation in multiple regression, where two correlated explanatory variables are fitted both separately and together, and where similar interpretational issues occur.

The residual estimates for preferred school are little changed from those in Table 4; those for closest school are somewhat more altered, and in some cases, the rank order changes, although these changes are well within the standard error estimates for the residuals.

DISCUSSION

The purpose of the analyses reported here is to describe the features of randomly cross-classified designs. Current computational restrictions (March 1991) do not allow the analysis of very extensive data sets, but this limitation should be removed in the near future. Thus the relatively small size of this data set results in rather large standard error estimates, so that clear-cut conclusions are not possible. Nevertheless, several features are clear.

Where units can be classified at the same level in more than one way, it is important to establish whether the resulting structure affects the covariance structure of the data. Where it does, then failure to take account of that structure can lead to misleading inferences and cause us to overlook important features of the between-unit variation. In the example mentioned in the introduction, where students are naturally grouped in terms of the successive schools they attend, there is a considerable educational interest in determining the contributions from each school. This will be especially important in systems where there is a great deal of mobility between schools from year to year, and where the current school attended might not be the most influential. Analyses that ignore this structure of the data will tend to underrate the importance of school differences.

In survey analysis, there are similar concerns. Individuals simultaneously belong to communities that can be defined in a number of ways; for example, in terms of geographical proximity and transportation connections or working relationships. Being able to model these classifications simultaneously could provide important new insights.

REFERENCES

Bastow, B. 1991. "A Study of the Factors Affecting Parental Choice of Schools." Ph.D. dissertation, University of London, Institute of Education.

Goldstein, H. 1986. "Multilevel Mixed Model Analysis Using Iterative Generalised Least Squares." *Biometrika* 73:43-56.

———. 1987a. "Multilevel Covariance Component Models." *Biometrika* 74:430-31.

———. 1987b. *Multilevel Models in Educational and Social Research*. London: Griffin/New York: Oxford University Press.

Prosser, R., J. Rasbash, and H. Goldstein. 1991. *ML3: Software for Three Level Analysis*. London: Institute of Education.

Harvey Goldstein has been a professor of statistical methods at the London Institute of Education since 1977. His two principal research interests are in the methodology of multilevel statistical models and the theory and practice of educational assessment. Since 1986 he has directed funded research projects aimed at developing the methodology of multilevel modeling and disseminating the techniques among social scientists. A comprehensive software package (ML3) has been produced by his coworker Jon Rasbash and this is now in use around the world. He has made numerous contributions to the literature on mental test construction and analysis, and has provided a wide ranging critique of much current activity in this area. He has also written about the social and political aspects of assessment and is currently codirector of the International Centre for Research on Assessment at the Institute of Education.