2 Multilevel Analysis for Education Research

See also the Summer 1995 Special Issue, *Hierarchical Linear Models: Problems and Prospects*, of the *Journal of Educational and Behavioral Statistics*, *20*, 109-240, as well as Kreft, de Leeuw, and Aiken (1995) and de Leeuw and Kreft (1995).

2.1 Aspects of the Analysis of Large Educational Databases *(Jan de Leeuw)*

Large educational databases (LEDs) such as the National Longitudinal Survey of 1972, High School and Beyond, the National Educational Longitudinal Survey of 1988, and the National Assessment of Educational Progress (NAEP), have an incredible richness of data. They can be used to answer a large number of questions about relationships between variables, and even about mechanisms in the school-attainment system (comparisons of schools, of states, of race and gender, about discrimination, tracking, and so on).

In fact, the number of questions that can be asked is very, very large, and some questions are framed in such general terms (or in such explicit causal terminology) that they are impossible to answer. We are only interested here in questions which are framed in terms of relationships between variables, either about the existence of such relationships, or about their strength. Answers to such questions can be used by educators and policy makers, they can be translated into causal terminology, and they can be related to choices that have to be made in either small-scale or large-scale educational policy making.

Making such translations and interpretations is not the primary task of the statistician. The statistician's job is to describe the relationship in a clear and convincing way, and to give an indication about the stability of the description.

We shall concentrate on some of the LED projects which have been and which are still being carried out at the Division of Statistics, University of California, Los Angeles, as part of the NISS education research project.

2.1.1 Paradigm

The paradigm for analysis of LEDs is still linear regression analysis, the workhorse of applied statistics, and this is as it should be, because many of the questions LEDs suggest are formulated in terms of the relationship between an outcome variable and a number of predictors. Nevertheless, the linearity, homoscedasticity, and independence assumptions could conceal more than they reveal, and with increasing sample size and increasing computer power, we can perhaps become a little bit more daring.

2.1.2 Complications

Hierarchical structure. Most LEDs contain information about students, teachers, and schools, i.e., there are variables describing units at various levels. For some questions, it is necessary to combine units from various levels in a single analysis. This creates a multilevel problem, in which more complicated error structures are needed in the regression analysis to take care of the correlation within teachers, classes, or schools.

Various design and implementation questions related to multilevel models have been studied in considerable detail in the last few years, but the answers are still rather tentative.

Sampling design. Typically, LEDs are not simple random samples (certainly not with replacement). They are stratified and/or clustered, with oversampling of certain groups of students or certain geographical regions. It is unclear, so far, what the consequences of the sampling design are for the standard errors typically computed in LEDs. But the problem is certainly important, and one that we are working on.

Temporal dependencies. Although LEDs are sometimes cross-sections, they often have longitudinal aspects as well. This is especially true for cohort studies in which a number of students are followed for a comparatively long period. The longitudinal aspects of LEDs have not been explored much, and this is rather unfortunate. A lot of structure can be derived from the temporal aspect of a study, and process variables can be used to enter the black box and get a better idea of mechanisms (think of tracking or choice of classes). The few attempts to analyze longitudinally have been mostly along LISREL lines. We should explore explicitly the state-space and event-history approaches (the latter, especially, seem quite promising).

Spatial dependencies. LEDs include spatial information. This could be in the crude form of state, but also could involve the actual location of the school in some smaller scale studies (such as the California Learning Assessment System). Spatial-statistics techniques have not been used much, until now, to map school achievement and related variables. An increasing number of techniques and software are available now to explore these spatial dependencies (variograms, kriging, ALEX).

Mixed measurement level. LEDs have many, many variables, and inevitably some of these variables will be numerical, some will be ordinal, some will be nominal. In the same way that the mixing of levels from the hierarchy can be problematic, mixing levels of measurement can be problematic, too. Standard techniques of multivariate analysis tend to treat all variables as nominal (log-linear), or all variables as numerical (normal). Mixed-level analysis is comparatively rare, and often is only available for a small number of variables in the analysis.

Nonlinearities. There is no reason to suppose that the relationships between variables in a LED (regressions, for instance) will necessarily be adequately approximated by linear structures. By now, various computerized forms of nonlinear and nonparametric regression are available, using kernels or local linear fits, and they should at least be tried out on LED regression problems (comparatively small ones, preferably).

Censoring of school careers. Among the missing-data problems that are common in LEDs, censoring is an important one. In data from Delaware, we have information about all school children in the state between 1981 and 1993 (tests, school achievement, family background, schools, courses, discipline, etc.). This means that we have one full cohort of students from kindergarten to twelfth grade; the other cohorts in the data are either censored on the left or censored on the right. It makes sense in the analyses (planned for the next couple of years) to take this censoring into account. In the same way, it seems interesting to analyze data on dropouts, using survival analysis techniques.

Selection. There are various missing-data problems in LEDs. We have already mentioned dropouts and censoring, but attrition in cohorts is another example. There may be other less systematic forms of missing data (schools in the Delaware cohort do not take all tests in all years, some NAEP information may be collected in some states and not in others, and there may be various forms of "undercount"). In general, missing information may not be missing and random, and we have to think of ways of describing the process generating the missing data.

2.2 The Effects of Centering in Multilevel Analysis: Is the Public School the Loser or the Winner? A New Analysis of an Old Question *(Ita G. G. Kreft)*

Multilevel models raise new problems that need solutions. One of them is centering, with two centering choices, centering predictors within context, or grand mean centering. Both are statistically sound ways to improve the estimation of the parameters in the model. Users of the software package, HLM (Bryk, Raudenbush, Seltzer, & Congdon, 1989), commonly center within context. The literature shows that this type of centering is applied in two ways, with or without adding the mean back to the model. It is not discussed in this literature that centering — especially when means are not reintroduced in the model — produces different results from raw-score models, especially for second-level estimates.

The effects of centering are examined using a large national data set, NELS:88. The research question is an old one, regarding the success of the private school over the public school,