*Journal of Educational and Behavioral Statistics*
*Summer 1995, Vol. 20, No. 2, pp. 201–204*

# Hierarchical Data Modeling in the Social Sciences

**Harvey Goldstein**
*Institute of Education, University of London*

The last 10 years of active research in the area of hierarchical, multilevel data modeling has brought problems as well as benefits. The three conference papers reflect well both the potentialities of the new procedures and some of the dangers we need to guard against. As in all statistical modeling of the real world, our inferences are no better than the data upon which they are based and the adequacy of the assumptions we are prepared to make.

The paper by de Leeuw and Kreft sounds some useful warnings, and I will discuss that one first. The paper by Rogosa and Saner focuses in detail on a repeated measures application and one software package, and asks questions about the usefulness of the available analysis procedures. I shall have some general remarks about ways of handling repeated measures data, but leave comments about the HLM software to Professor Raudenbush to respond to. The paper by Draper is concerned with causal inference and ways in which this can be strengthened by using multilevel models. He also places these models in their historical context, and his discussion of competing estimation procedures raises some interesting topics for future research.

## de Leeuw and Kreft

This paper considers the relatively simple linear two-level model with a continuous response variable. It provides a useful introduction by taking the reader from a series of separate equation regressions to a random coefficients model. The authors are right to emphasize the need to provide interpretational guidance for users, but, in my view, tend to exaggerate some of the difficulties. For example, the Level 2 covariance matrix of random coefficients can be used to provide estimates of the between-school variance as a function of the predictor variables, and this can be plotted to give insights into how, say, the school level variation changes with social background or gender. In addition, by calculating posterior means of the coefficients for each school, the individual (estimated) school relationships can be plotted—remembering, of course, that these are "shrunken" estimates.

While the use of a relatively simple model has advantages, it ignores some interesting extensions. It is a pity that the authors, having got as far as considering a two-level random coefficient model, do not discuss, for example, the modeling of the Level 1 variance. In many educational data sets we find heteroscedasticity at Level 1. Thus, boys tend to have higher variances for test scores than girls, and in a longitudinal study one often finds that those students with low pretest scores have smaller variance on a posttest

score than those with high pretest scores. Indeed, in some cases, fitting complex variation at Level 1 considerably improves the overall explanatory power of the model and the stability of other parameters. It is also the case, of course, that there is now considerable interest in nonlinear multilevel models, especially generalized linear models for proportions and count data, but I shall return to that below.

The distinction drawn between simple noniterative estimation procedures and iterative maximum likelihood (ML) or restricted maximum likelihood (REML) is now, I think, rather artificial. The standard advantages of ML or REML in terms of efficiency are important and the computational penalty is not usually very severe. The simpler methods are, however, sometimes more robust—a property shared with the iterative generalized estimating equation (GEE) approach (Liang & Zeger, 1986). This property may be useful, for example, when we suspect that multivariate Normality does not hold, but for most social scientists it is the structure of the model which requires explication rather than the details of the estimation procedure.

I am, of course, delighted that the authors, in their final section on software, speak well of the flexibility of the ML3 software. This flexibility was designed from the outset because we wished to have an open general system that could easily incorporate new developments. This has allowed us to add facilities, such as the ability to handle random cross-classifications, measurement errors, and nonlinear, especially generalized linear, models, as the relevant estimation theory has been developed. This is currently coming to fruition in the form of the next, many-level version, MLn.

There is a danger, and this paper reminds us of it, that multilevel modeling will become so fashionable that its use will be a requirement of journal editors, or even worse, that the mere fact of having fitted a multilevel model will become a certificate of statistical probity. That would be a great pity. These models are as good as the data they fit: they are powerful tools, not universal panaceas.

## Rogosa and Saner

Repeated measures data constitutes a very good example of a situation in which a two-level model is really essential, because most of the variation typically is at the higher level. The literature on fitting repeated measures data, especially from growth studies, has a long history (see, for example, Goldstein, 1979) and its formulation as a two-level model immediately solves a great number of outstanding problems.

One of these is that previous models, based upon a multivariate formulation, were able to handle only measurements made at discrete times, possibly with some missing responses. In the two-level formulation, this requirement is completely unnecessary and we can have any pattern and number of repeated measurements per individual, including individuals who contribute only one measurement, and obtain fully efficient (ML or REML) estimates using any

of the existing multilevel software packages. It is a pity, therefore, that the authors stick with discrete time data sets, because their conclusions about comparisons among estimation procedures rely heavily on the fact that their example data sets are highly balanced.

The authors make a useful point about data description and presentation. Nevertheless, after fitting a two-level model we can estimate residuals (posterior means) and plot their standardized values in a number of ways, which generally will be more reliable than the simple OLS plots when the number of measurements per individual is small. The authors are also right to point to the little work that has been done on study design.

Finally, it is worth pointing out that the basic two-level repeated measures model can be extended in a number of useful directions. At the Institute of Education, we have recently completed work on fitting models where the Level 1 residuals have an additional time series structure, which often occurs in growth data with measurements taken close together in time (Goldstein, Healy, & Rasbash, 1994). The models can also be extended to multivariate responses and can be used to provide efficient methods for growth prediction (Goldstein, 1995).

## Draper

David Draper's discussion of justifiable inference is clear and a further useful reminder that we should pay as much attention to the source of our data as to the methods of their analysis. The discussion of Huttenlocher's analysis, however, raises a further issue which is not discussed.

When researchers use convenience samples, they sometimes do so because they have evidence (or a view based on their professional experience) which leads them to believe that there is a close correspondence between their convenience population and the real population of interest. The problem is that this correspondence is uncertain and difficult to quantify and is often not made explicit. Yet it does sometimes happen that inferences based upon formally inadequate samples give accurate inferences or predictions—voting intention surveys are a case in point and this may be more than just luck. Of Draper's examples, some fall into this category. Among them is one on fitting growth curves to London children which I used in my book (Goldstein, 1987). This is an interesting case because I was clearly guilty of improperly contextualizing the study which produced the data. In fact, that study was one of a series of collaborative studies across Europe of which one of the intentions was to see whether growth patterns could be replicated. It turns out that in the area of child growth there is indeed a considerable uniformity of pattern across different population groups (Tanner, 1962) so that there is good reason to feel confident about the generalizability of the results. From a scientific point of view, it is the replicability of findings in very different contexts that is usually more convincing than the evidence from a single representative sample. The moral would seem to be that investigators should

be more explicit about all their sources of evidence when they attempt to produce generalizable statistical inferences. This could be added to Draper's list of desiderata in his section "The Value of Explicitness in Inferential Conclusions."

From my point of view, the main reason for producing the *Guardian* value-added survey was to counter the misuse by the British government of raw school examination results to produce league tables. The intention was to demonstrate that both adjusting for intake achievement and presenting uncertainty intervals were necessary, although not sufficient, conditions for valid comparisons. We were not primarily interested in causal inferences, although I believe that the data are adequate enough for that, and we are currently pursuing it.

The emergence of Markov-chain Monte Carlo (MCMC) methods such as Gibbs sampling is clearly very important for a wide range of estimation problems, especially where there are small numbers of units. It is not at all surprising, of course, that in Rubin's example with eight Level 2 units, the likelihood estimate of the variance is zero and that the inclusion of prior information gives a positive estimate. In a likelihood framework this emphasizes the importance of procedures such as bootstrapping, which, like MCMC methods, allows accurate assessment of parameter uncertainty.

It is interesting that Draper quotes Rodríguez's findings on the bias in estimation for multilevel models with binary responses. This has led to collaborative methodological work resulting in a considerable improvement (Goldstein, 1995) and is an example of the kind of critical evaluation of techniques which Draper emphasizes.

## References

Goldstein, H. (1979). *The design and analysis of longitudinal studies.* London: Academic Press.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Griffin.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold; New York: Halstead Press.

Goldstein, H., Healy, M. J. R., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine, 13,* 1643–55.

Liang, K., & Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika, 73,* 45–51.

Tanner, J. M. (1962). *Growth at adolescence.* Oxford: Blackwell.

## Author

HARVEY GOLDSTEIN is Professor, Institute of Education, 20 Bedford Way, London, WC1HOAL, England; hgoldstn@ioe.ac.uk. He specializes in the modeling of hierarchical data structures.

## References

**Longitudinal Data Analysis Using Generalized Linear Models**

Kung-Yee Liang; Scott L. Zeger

*Biometrika*, Vol. 73, No. 1. (Apr., 1986), pp. 13-22.

Stable URL:

http://links.jstor.org/sici?sici=0006-3444%28198604%2973%3A1%3C13%3ALDAUGL%3E2.0.CO%3B2-D