Journal of Educational and Behavioral Statistics
Summer 1995, Vol. 20, No. 2, pp. 171–189

Questioning Multilevel Models

Jan de Leeuw
University of California, Los Angeles
Ita G. G. Kreft
California State University, Los Angeles

In this article, practical problems with multilevel techniques are discussed. These problems, brought to our attention by the National Center for Education Statistics (NCES), have to do with terminology, computer programs employing different algorithms, and interpretations of the coefficients in one or two steps. We discuss the usefulness of the hierarchical linear model (HM) in the most common situation in education—that of a large number of relatively small groups. We also point to situations where the more complicated HMs can be replaced with simpler models, with statistical properties that are easier to study. We conclude that more studies need to be done to establish the claimed superiority of restricted versus unrestricted maximum likelihood, to study the effects of shrinkage on the estimators, and to explore the merits of simpler methods such as weighted least squares. Finally, distinctions must be made between choice of model, choice of technique, choice of algorithm, and choice of computer program. While HMs are an elegant conceptualization, they are not always necessary. Traditional techniques perform as well, or better, if there are large groups and small intraclass correlations, and if the researcher is interested only in the fixed-level regression coefficients.

In this article, we discuss some of the practical problems in using multilevel techniques, by looking into the choices users of these techniques have to make. It is difficult, of course, to define "user." Different users have different degrees of statistical background, computer literacy, experience, and so on. We adopt a particular operational definition of a "user" in this article, which certainly does not cover all users. Our "user" is defined by the set of questions asked by the Statistical Standards and Methodology Division of the National Center for Education Statistics (NCES). These questions were asked in the context of a grant, which has as one of its specific purposes to evaluate the practical usefulness of multilevel modeling in educational statistics. We cannot discuss, let alone answer, all the questions from NCES in this article. Even the ones we discuss will usually require additional statistical and computational research, but they illustrate some of the practical methodological problems in using hierarchical linear models.[1]

The statistics and mathematics will be kept as simple as possible. We shall concentrate on the situation in which we have a relatively large number of relatively small groups. The situation in which we have only two or three groups does not really interest us here, and the situation in which we have

a large number of very small groups (twins, couples) also requires a slightly different emphasis. There should be at least 20, but maybe as many as 1,000, groups of size at least 5, but maybe as large as 50. This seems to cover most studies in which the individuals are students and the groups are schools or classes.

The NCES questions will be discussed in terms of a number of *choices* the user has to make. Here is a brief list. The user has to choose (a) a selection and coding of her variables, (b) a model from the class of regression models, (c) a loss function to measure goodness-of-fit, (d) an algorithm to minimize the loss function, and (e) a computer program to implement the algorithm. All these choices are nontrivial, but our discussion will mainly emphasize the choice of the model, the loss function, and the technique—and, of course, the consequences of these choices.

Generalities on Linear Regression

The first, rather general question posed by NCES is

> **Question 1:** Is some form of hierarchical linear model always preferable when conducting analysis with independent variables from two levels of a hierarchical data set?

We shall try to answer this question in a very roundabout way. First, some terms need to be defined. *Hierarchical data* occur if the objects we study are classified into groups. Students within classes is one classical example; individual in census tracts or political districts is another one; and time points within individuals is a third one. We want to describe our hierarchical data by using a linear model, or, more precisely, a linear model which takes the hierarchical structure of the data into account. This is rather vague, but we shall make it more precise as we go along.

In the usual (nonhierarchical) linear model there are n individuals and p predictors. The outcomes are collected in an n-element vector $\mathbf{y} = \{y_i\}$, the values on the predictors in an $n \times p$ matrix $\mathbf{X} = \{x_{is}\}$. We suppose

$$\underline{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \underline{\boldsymbol{\epsilon}}. \tag{1}$$

Random variables are distinguished from fixed quantities by underlining them (Hemelrijk, 1966). In discussing this class of regression models, the distinction between what is fixed and what is random is quite important, and the underlining helps to emphasize the difference between the two.

Throughout the article, frequentist terminology is used. Thus (1), for instance, is a model that describes a hypothetical sequence of replications of the experiment that generated the data. This statistical model does not describe the outcome of a single experiment, or of an actual sequence of replications, but it models a hypothetical sequence of replications. In this hypothetical sequence $\mathbf{X}$ remains fixed (i.e., it is exactly the same in each replication).

The coefficients $\boldsymbol{\beta}$ are also fixed over replications, but we do not know what their values are. They are *parameters* that have to be *estimated.* The *disturbances* $\underline{\boldsymbol{\epsilon}}$ are different for each of the hypothetical replications, and they vary according to specific probability distributions. In particular, it is assumed, for their expected value and dispersion matrix, that

$$E(\underline{\boldsymbol{\epsilon}}) = \mathbf{0}, \tag{2}$$

and

$$V(\underline{\boldsymbol{\epsilon}}) = \sigma^2\mathbf{I}. \tag{3}$$

The model (1)–(3) says, in essence, that the disturbances are independent and identically distributed; that is, they are not systematically related to $\mathbf{X}$—or to anything else, for that matter.

Model (1)–(3) is really meant for situations in which the predictors in $\mathbf{X}$ are under experimental control, and can be assumed to be measured without error. That is, it is meant for designed experiments. The x_{is} are fixed quantities; that is, they remain the same over the hypothetical replications, which means that in order to use the model we must have a way of physically keeping them the same. This does not happen very often in educational statistics. If school success is regressed on IQ, we are usually not interested in replications in which the individual has the same IQ all the time, only different school success. Both variables covary; that is, it looks as if we should use a model with a random predictor. Fortunately, this problem can be solved quite easily, at least from a formal point of view. If it is assumed that (1) models the conditional distribution of $\underline{y}$ given $\underline{\mathbf{x}}_i = \mathbf{x}_i$, then the marginal distribution of $\underline{\mathbf{x}}_i$ can be modeled separately to get a model for the joint distribution of $(\underline{y}_i, \underline{\mathbf{x}}_i)$.

A second problem of (1)–(3) is that to assume linearity and homoscedasticity of the regression in the joint distribution is to make a very strong assumption which is unlikely to be even approximately true. It forces us to take a more modest approach, in which models are used as tools for compact *description* and/or as tools for *prediction.* There is no need to worry about the model being true (it obviously is not); the question is only if it does its job of summarizing the information in the data and extrapolating into the future well enough. It is still the general consensus that the linear regression model (1)–(3) does quite well, especially considering how strong and unrealistic it is. It is still the workhourse of applied statistics, and in fact it sometimes seems as if applied statistics *is* linear regression analysis.

This leads to a third general point, which is of considerable importance, and which is not often discussed. Statistical models are languages that users in a particular field have to learn, and thay they use to talk to each other efficiently. Regression analysis, path analysis, factor analysis, and survival

analysis are all examples of this. There is a tendency to narrow down the language even more, so that in the 1970s, for example, LISREL became the language of choice for a large group of scientists in various disciplines. In educational statistics, the multilevel framework provides a language that encompasses and supersedes the older language of contextual analysis, and there seems to be a tendency to narrow it down even more to the language of the HLM program. But this means that in the field it becomes difficult to talk about hierarchical data structures without adoping the terminology (and constraints) of the HLM program.

On Random Coefficients

Another critical assumption in (1) is that the regression coefficients β are the same for all individuals. Starting with Wald in 1947, economists have (sometimes) been critical about this assumption. In his text book, Klein (1953) says,

> Individuals differ greatly in behavior, and it may not be possible to obtain observations on a sufficiently large number of variables so that each unit may be considered to behave according to the same structural equation. We are then faced with the problem of interpreting a single estimated equation as representative in some sense of a large number of underlying equations. (p. 216)

This quotation is interesting in that it states explicitly that we need more than one regression equation because we do not have enough predictors. If we had all relevant predictors in our study, we could use a single equation for all individuals, but because this is impossible, or at least impractical, the equations will vary around some average equation.

This can be formalized by using the notion of random coefficients. The model is

$$\underline{y}_i = \mathbf{x}_i'\beta_i + \underline{\epsilon}_i, \tag{4}$$

$$\underline{\boldsymbol{\beta}}_i = \boldsymbol{\beta} + \underline{\boldsymbol{\delta}}_i, \tag{5}$$

where $\underline{\boldsymbol{\delta}}_i$ are independent and identically distributed with zero expected value and dispersion $\boldsymbol{\Omega}$. Moreover, they are independent of the $\underline{\epsilon}_i$. Thus $\underline{\mathbf{y}}$ has expectation $\mathbf{X}\boldsymbol{\beta}$, as in the fixed coefficient model, but now there is heteroscedasticity because

$$V(\underline{y}_i) = \mathbf{x}_i'\,\boldsymbol{\Omega}\mathbf{x}_i + \sigma^2. \tag{6}$$

Once again we emphasize that the distinction between fixed and random coefficients is important, because it changes the definition of the population over which we want to generalize. If we repeat our experiment, then we do

not expect individuals i to have the same regression coefficients in each replication. The regression coefficients vary, both within individuals and between individuals, around a population mean.

Regression in Multiple Populations

The more general situation in which there are m groups, indexed by j, is analyzed next. This moves towards the situation in which we have hierarchical data. A straightforward generalization of (1) is

$$\underline{\mathbf{y}}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \underline{\boldsymbol{\epsilon}}_j. \tag{7}$$

Now the $\underline{\mathbf{y}}_j$ and the $\underline{\boldsymbol{\epsilon}}_j$ are vectors of length n_j, the number of individuals in group j. Matrix $\mathbf{X}_j$ is $n_j \times m$. It makes sense also to assume that

$$E(\underline{\boldsymbol{\epsilon}}_j) = \mathbf{0}, \tag{8}$$

and

$$V(\underline{\boldsymbol{\epsilon}}_j) = \sigma_j^2\mathbf{I}. \tag{9}$$

Finally, we assume the different $\underline{\boldsymbol{\epsilon}}_j$ are independent of each other.

There is nothing wrong with model (7)–(9). It takes the hierarchical structure of the data into account, although it merely says that the same regressors apply to each of the groups. The model can be fitted to each of the m groups separately, because none of the parameters are common to the groups. This is not very attractive, especially if there is a large number of relatively small groups—for instance, students from many school classes, where each class has somewhere around 10–20 students. It ignores the fact that schools are all part of the same system, and that consequently the regressions are likely to have something in common. One way to incorporate this commonality into the model is to require that some of the parameters are equal in all groups. There are two obvious choices:

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_m, \tag{10}$$

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2. \tag{11}$$

But this is a clear case of throwing the baby away with the bathwater. Although all schools are related, and must have something in common, we do not want to assume they are identical. There a number of ways out of this dilemma. We discuss two of them in this section, in order to be able to compare them at a later stage. The first approach uses *linear restrictions*. The

second uses a *random coefficient model* that takes the hierarchical structure of students in schools into account.

Linear Restrictions on Parameters

The analysis of covariance is an example of the first approach. It assumes (11), and that all slopes, but not all intercepts, are equal. More generally, the assumption is that $\boldsymbol{\beta}_j = \mathbf{Z}_j\boldsymbol{\gamma}$, where the $\mathbf{Z}_j$ are chosen in such a way that some elements of the $\boldsymbol{\beta}_j$ are equal. In ANCOVA, for instance, we have mp parameters β_{sj}, and we replace them by $p - 1$ slopes and m intercepts. If the slopes are in a vector $\boldsymbol{\beta}$ and the intercepts in a vector $\boldsymbol{\alpha}$, then we can set

$$\boldsymbol{\beta}_j = \mathbf{Z}_j\boldsymbol{\gamma} = \begin{pmatrix} \mathbf{e}_j' & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}, \tag{12}$$

with $\mathbf{e}_j$ the jth unit vector. Thus, the partial identity approach leads to linear constraints on the $\boldsymbol{\beta}_j$. Combining these constraints with (7) simply gives the fixed-effects linear model

$$\underline{\mathbf{y}}_j = \mathbf{X}_j\mathbf{Z}_j\boldsymbol{\gamma} + \underline{\boldsymbol{\epsilon}}_j, \tag{13}$$

which can be fitted with ordinary least squares methods, because it is still the case that $V(\underline{\mathbf{y}}_j) = \sigma^2\mathbf{I}$.

Random Coefficients Revisited

In random coefficient models, a slightly different route is traveled. The model is in between "separate models for all schools" and "complete equality." It is given by

$$\underline{\mathbf{y}}_j = \mathbf{X}_j\underline{\boldsymbol{\beta}}_j + \underline{\boldsymbol{\epsilon}}_j, \tag{14}$$

with

$$\underline{\boldsymbol{\beta}}_j = \boldsymbol{\beta} + \underline{\boldsymbol{\delta}}_j. \tag{15}$$

Compare this with (4)–(5). In the earlier model it is assumed that each *individual* has her own regression coefficients, and these coefficients are independent over individuals. In (14) the assumption is that each *group* has its own regression coefficients, which are independent over groups. But the coefficients are identical for different individuals in the same group. The coefficients are modeled as random, which means that the definition of a population (our hypothetical sequence of replications) is modified. The slopes and intercepts are no longer fixed numbers, which are constant within schools

and maybe even between schools, but they also vary over replications. In order to complete the specification we also assume that the *second-level disturbances* $\underline{\boldsymbol{\delta}}_j$ are independent of each other, are independent of the first-level disturbances $\underline{\boldsymbol{\epsilon}}_j$, have zero expectation, and have dispersion matrix $\boldsymbol{\Omega}$. If the second-level disturbances are identically equal to zero, then we are back in the situation (10)–(11).

By combining (14) and (15) we see that

$$\underline{\mathbf{y}}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{X}_j\underline{\boldsymbol{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j, \tag{16}$$

which implies

$$V(\underline{\mathbf{y}}_j) = \mathbf{X}_j\boldsymbol{\Omega}\mathbf{X}_j' + \sigma_j^2\mathbf{I}. \tag{17}$$

Individuals in the same school have correlated disturbances, and the correlation will be larger if their predictor profiles are more similar, in the metric $\boldsymbol{\Omega}$. This is an interesting consequence of the specification (14)–(15), but understanding (14)–(15) itself is clearly more basic. It will be difficult, even for sophisticated users, to interpret the variance and covariance components in (17) directly.

Random coefficient models are a convenient compromise between separate fixed coefficient models for each group, and models with all coefficients equal for each group. They are "convenient" because we expect them to give more stable estimates than separate models and more interesting parameters than equal coefficients. They are also more plausible, by the Klein argument, because it cannot be assumed that we have included all relevant variables.

Multilevel Models

Our regression situation becomes more complicated, but also more interesting, if we have variables describing individuals (students) as well as variables describing groups (schools). Combining them in a single analysis is called *multilevel analysis.* In multilevel analysis we combine the two approaches discussed earlier in this section. Linear restrictions of the form $\boldsymbol{\beta}_j = \mathbf{Z}_j\boldsymbol{\gamma}$ are used to reduce the number of free regression parameters, and the idea of random coefficients is used to model the idea that schools are sampled, and that we cannot expect to explain all relevant variation with only a few regressors. The combined model, which replaces (14)–(15), is

$$\underline{\mathbf{y}}_j = \mathbf{X}_j\underline{\boldsymbol{\beta}}_j + \underline{\boldsymbol{\epsilon}}_j, \tag{18}$$

$$\underline{\boldsymbol{\beta}}_j = \mathbf{Z}_j\boldsymbol{\gamma} + \underline{\boldsymbol{\delta}}_j. \tag{19}$$

There are clearly two different regression models on two different levels.

The first-level model (18) is complemented by the second-level model (19). If we substitute (19) into (18), we have

$$\underline{\mathbf{y}}_j = \mathbf{X}_j\mathbf{Z}_j\boldsymbol{\gamma} + \mathbf{X}_j\underline{\boldsymbol{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j. \tag{20}$$

Compare this with (13).

Are Hierarchical Models Necessary?

If we summarize our results so far, trying to answer the first NCES question, we see that the answer to this question is difficult to give, because of some possible conceptual confusion. Hierarchical models are any models that take the hierarchical nature of data into account. The multilevel model with random coefficients, (18)–(19), sometimes also known as the *slopes-as-outcomes* model (Burstein, Linn, & Capell, 1978), is only one single specific hierarchical model. If the question is interpreted as "Should we take the hierarchical nature of the data into account in our models?," then the answer is yes. We should because it is important prior information that can be used to increase the power and precision of our techniques, and also because it often reflects the way the sample is collected.

But even if the hierarchical nature of the data is taken into account, and even if we have multilevel data, we still have a scale of modeling possibilities. Regression coefficients in all groups can be restricted to be equal, or they can be completely unrestricted, and vary freely over groups. The first possibility may be too restrictive, and the second one may be too unrestrictive in typical school-effectiveness situations. Two natural intermediate classes of models can be formed by using linear restrictions on the parameters, or by using random coefficients. Neither of these is inherently superior to the other. In many school-effectiveness studies, however, the second-level units (i.e., the schools) are sampled from a population of schools. In those cases, the notion of random variation on both levels is very appealing.

To be sure, even in cases in which the multilevel model with random coefficients is the natural choice, it still does not follow that statistical techniques based on maximum likelihood or empirical Bayes (ML/EB) methodology should be used. This is a separate question which requires separate study. And finally, even if we decide to use ML/EB methods, this again does not imply the choice of a specific computer program. If we interpret the NCES question as "Is ML/EB necessary?" or "Is HLM necessary?," then we have at this point not enough information to give a reasonable answer.

Separate or Single Equations

The next NCES question is

> **Question 2:** Some analysts are more comfortable presenting HLM results in terms of a combined model, i.e., a single regression equation containing

interaction terms. Others prefer to discuss the coefficients without recourse to a single regression equation. Are the two approaches equally valid?

There are two aspects of this question that we discuss separately. The first one is *modeling* in one or two steps; the second one is *estimating* in one or two steps. Again, there is some confusion in the literature about these two aspects of the question.

One-Step or Two-Step Models?

Let us translate this part of the question into formulas, because it is at least partly a question about formulas. If we look at the fixed part of (20) we see that

$$E(\underline{\mathbf{y}}_j) = \mathbf{X}_j\mathbf{Z}_j\boldsymbol{\gamma}. \tag{21}$$

In (21) the cross level interactions are formed as products of the first-level regressors $\mathbf{x}_s$ and the second-level regressors $\mathbf{z}_r$. In a sense, there is not much to choose. The single-equation and two-equation formulations describe the same model.

From the interpretational point of view, however, the two formulations are quite different. We feel it is very difficult, perhaps impossible, to interpret (20) without going back to (18)–(19). It is, of course, possible to interpret the fixed effects in (20), because there is a lot of experience with interpreting interactions in fixed-effect situations. Compare the useful reviews by Aiken and West (1991) and Cox (1984). It is, however, quite impossible to come up with a convincing interpretation of the structure of the disturbance term in (20) without referring to (18)–(19). The disturbance term in question is $\mathbf{X}_j\underline{\boldsymbol{\delta}}_j + \underline{\boldsymbol{\epsilon}}_j$, and its dispersion matrix is $\mathbf{X}_j\boldsymbol{\Omega}\mathbf{X}_j' + \sigma_j^2\mathbf{I}$. We have seen, in the previous section, that it is difficult to make direct sense of these, and especially of the covariance components.

One-Step or Two-Step Estimates?

The one-step (20) and the two-step (18)–(19) specifications of the multilevel model, discussed in the previous section, suggest two different ordinary least squares (OLS) methods for fitting the model. This has already been discussed in detail by Boyd and Iverson (1979). We follow the treatment of de Leeuw and Kreft (1986).

The two-step method first estimates the $\boldsymbol{\beta}_j$ by

$$\hat{\boldsymbol{\beta}}_j = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\underline{\mathbf{y}}_j, \tag{22}$$

and then $\boldsymbol{\gamma}$ by

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{j=1}^{m} \mathbf{Z}_j'\mathbf{Z}_j\right)^{-1} \sum_{j=1}^{m} \mathbf{Z}_j'\hat{\boldsymbol{\beta}}_j. \tag{23}$$

The one-step method estimates $\boldsymbol{\gamma}$ directly from (20) as

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{j=1}^{m} \mathbf{Z}_j'\mathbf{X}_j'\mathbf{X}_j\mathbf{Z}_j\right)^{-1} \sum_{j=1}^{m} \mathbf{Z}_j'\mathbf{X}_j'\underline{\mathbf{y}}_j. \tag{24}$$

By using (22) we see immediately, however, that the one-step method can also be written as

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{j=1}^{m} \mathbf{Z}_j'\mathbf{X}_j'\mathbf{X}_j\mathbf{Z}_j\right)^{-1} \sum_{j=1}^{m} \mathbf{Z}_j'\mathbf{X}_j'\mathbf{X}_j\hat{\boldsymbol{\beta}}_j. \tag{25}$$

Thus the "one-step estimate" can be computed in two steps, as well; in fact, this is often the best way to compute it because the matrices in (25) are much smaller than those in (24).

Both methods provide unbiased estimates of $\boldsymbol{\gamma}$, are noniterative, and are easy to implement; and because they are linear in the observations, it is trivial to give an expression for their dispersion matrices. Nevertheless they have fallen into disgrace, because they are neither best linear unbiased estimates (BLUEs) nor best linear unbiased predictors (BLUPs). On the basis of the computational experience we have so far (which is quite minimal), we feel that they still deserve a fighting chance.

The next candidate that comes to mind is based on the BLUE. If we knew σ_j^2 and $\boldsymbol{\Omega}$, then we could compute the BLUE by

$$\begin{aligned}\hat{\boldsymbol{\gamma}} = &\left\{\sum_{j=1}^{m} \mathbf{Z}_j'\mathbf{X}_j'(\mathbf{X}_j\boldsymbol{\Omega}\mathbf{X}_j' + \sigma_j^2\mathbf{I})^{-1}\mathbf{X}_j\mathbf{Z}_j\right\}^{-1} \\ &\times \sum_{j=1}^{m} \mathbf{Z}_j'\mathbf{X}_j'(\mathbf{X}_j\boldsymbol{\Omega}\mathbf{X}_j' + \sigma_j^2\mathbf{I})^{-1}\underline{\mathbf{y}}_j.\end{aligned} \tag{26}$$

This looks horrible, but it can be simplified to

$$\hat{\boldsymbol{\gamma}} = \left\{\sum_{j=1}^{m} \mathbf{Z}_j'\mathbf{W}_j^{-1}\mathbf{Z}_j\right\}^{-1} \sum_{j=1}^{m} \mathbf{Z}_j'\mathbf{W}_j^{-1}\hat{\boldsymbol{\beta}}_j, \tag{27}$$

where

$$\mathbf{W}_j = \boldsymbol{\Omega} + \sigma_j^2(\mathbf{X}_j'\mathbf{X}_j)^{-1}. \tag{28}$$

Observe that $\mathbf{W}_j$ is the dispersion of the OLS estimate $\hat{\boldsymbol{\beta}}_j$.

The formal similarity of (23), (25), and (27) is clear. They can all be thought of as two-step methods, which first compute the $\hat{\boldsymbol{\beta}}_j$, and then do a weighted regression of the $\hat{\boldsymbol{\beta}}_j$ on the $\mathbf{Z}_j$. Of course, (27) is useless by itself, because we do not know what σ_j^2 and $\boldsymbol{\Omega}$ are, but a method to compute consistent estimates of these variance parameters from the OLS residuals is discussed in de Leeuw and Kreft (1986). This adapts a method proposed by Swamy (1971) to the multilevel model. The resulting method is fully efficient, noniterative, and uses unbiased estimates of the variance components. Of course, unbiasedness in this context is not necessarily good, because it inevitably leads to negative variance estimates.

Again, we think a more detailed comparison of these simpler methods with the complicated iterative ML/EB methods such as HLM (Bryk, Raudenbush, Seltzer, & Congdon, 1988), or VARCL (Longford, 1990), or ML/3 (Rasbash, Prosser, & Goldstein, 1989) would be useful. The least squares methods are computationally simpler, and easier to understand and explain. Moreover, it is generally simpler to study their statistical properties. In the case in which the variance components have to be estimated first, the statistics are still quite complicated (Johansen, 1982). Some interesting Monte Carlo results on weighted least squares versus ML/EB estimation have been published by Kim (1990) and van der Leeden and Busing (1994). Also, see Kreft and Yoon (1994) for an overview of Monte Carlo results so far.

Loss Functions and Global Fit Measures

This also brings us to the next question asked by NCES:

> **Question 3:** Most discussion of HLM results centers on the individual coefficients: the betas and gammas. There is, of course, some interest in the overall measures, such as the proportion of variance explained. What is the best way to obtain and present overall measures when using HLM?

Each of the two-step methods discussed above gives one way to compute the "proportion of variance explained." We have residual sums of squares in each of the two steps.

The Analysis of Deviance

We get a somewhat more integrated picture by using the *analysis of deviance,* which is based on the multinormal likelihood function. Fixed and random coefficient models are combined in

$$\underline{\mathbf{y}}_j = \mathbf{X}_j\underline{\boldsymbol{\beta}}_j + \underline{\boldsymbol{\epsilon}}_j, \tag{29}$$

$$\underline{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j + \underline{\boldsymbol{\delta}}_j. \tag{30}$$

The two additional specifications we can either impose, or test, or both, within this model are

$$\boldsymbol{\beta}_j = \mathbf{Z}_j\boldsymbol{\gamma}, \tag{31}$$

and

$$\boldsymbol{\Omega} = \mathbf{0}. \tag{32}$$

A special case of (31) is equality of the $\boldsymbol{\beta}_j$; another special case is (random effects) ANCOVA. Of course, (32) is the hypothesis that the regression coefficients have no random variation.

The multinormal deviance for model (29)–(32) is, ignoring the usual constants,

$$\begin{aligned}\Delta = &\sum_{j=1}^{m} \log|\mathbf{X}_j\boldsymbol{\Omega}\mathbf{X}_j' + \sigma_j^2\mathbf{I}| \\ &+ \sum_{j=1}^{m} (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)'[\mathbf{X}_j\boldsymbol{\Omega}\mathbf{X}_j' + \sigma_j^2\mathbf{I}]^{-1}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j).\end{aligned} \tag{33}$$

This can be simplified by writing $\mathbf{y}_j = \mathbf{X}_j\hat{\boldsymbol{\beta}}_j + \mathbf{r}_j$, where $\hat{\boldsymbol{\beta}}_j$ is any OLS estimate. We find that, except again for some constants,

$$\begin{aligned}\Delta = \sum_{j=1}^{m} \Bigg\{ &\log|\mathbf{W}_j| + (n_j - p)\left\{\log\sigma_j^2 + \frac{\hat{\sigma}_j^2}{\sigma_j^2}\right\} \\ &+ (\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)'\mathbf{W}_j^{-1}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\Bigg\}.\end{aligned} \tag{34}$$

Here $\hat{\sigma}_j^2$ is the OLS estimate of the residual variance, that is,

$$\hat{\sigma}_j^2 = \frac{\mathbf{r}_j'\mathbf{r}_j}{n_j - p}. \tag{35}$$

The derivation of (34) from (33) is, for example, in de Leeuw and Kreft (1986).

It seems that all the "overall measures" that are useful are components of (34). The deviance itself is an overall measure of fit. We also see the residual individual level variance σ_j^2 in each group, while the two components of $\mathbf{W}_j$

are the *parameter variance* $\mathbf{\Omega}$ and the *estimation variance* $\sigma_j^2(\mathbf{X}_j'\mathbf{X}_j)^{-1}$. This is discussed extensively in Bryk and Raudenbush (1991).

If we want to establish how much variance of the $\hat{\boldsymbol{\beta}}_j$ is "explained" by the $\mathbf{Z}_j$, we merely have to compute the matrix

$$\sum_{j=1}^{m} \hat{\mathbf{W}}_j^{-1}(\hat{\boldsymbol{\beta}}_j - \mathbf{Z}_j\hat{\boldsymbol{\gamma}})(\hat{\boldsymbol{\beta}}_j - \mathbf{Z}_j\hat{\boldsymbol{\gamma}})' \tag{36}$$

and look at its diagonal or trace. Here $\hat{\mathbf{W}}_j$ and $\hat{\boldsymbol{\gamma}}$ are the maximum likelihood estimates, computed by minimizing the deviation (34) over the free parameters. An alternative discussion of measures of fit, based on the notion of "modeled variance," is given in Snijders and Bosker (1994).

REML Versus ML Deviance

In HLM, and in some of the other multilevel programs as well, the deviance that is actually minimized is defined slightly differently. Instead of minimizing the deviance of the data, they minimize the deviance of the least squares residuals. This leads to restricted maximum likelihood (REML) estimates (Harville, 1977). In the multilevel context, the relevant algebra is in the appendix of the book by Bryk and Raudenbush (1991), or in the paper by de Leeuw and Liu (1993). REML estimates are generally considered to be superior to the maximum likelihood estimates based on the deviance of the data, but the evidence of their superiority in complicated cases, and in multilevel analysis in particular, is not too convincing. The precise asymptotics for both ML and REML have been worked out (Cressie & Lahiri, 1993; Miller, 1977), but, as usual, the results are not very helpful. Careful Monte Carlo studies in simpler cases (Swallow & Monahan, 1984) do not lead to unambiguous recommendations. Clearly, a great deal more research, of the theoretical and the Monte Carlo varieties, is needed here.

Shrinkage Estimates

Another question which is of some interest from the practical point of view is how the $\underline{\boldsymbol{\beta}}_j$ are estimated. Obviously, the unbiased and consistent estimates $\hat{\boldsymbol{\beta}}_j$ or $\underline{\boldsymbol{\beta}}_j = \mathbf{Z}_j\boldsymbol{\gamma}$ can be used, just as in the fixed coefficient case. This is not what is normally done, however. One of the key selling points of multilevel approaches is the *shrinkage estimator,* which is used to borrow strength from the other contexts (groups, schools). In this approach we estimate $\boldsymbol{\beta}_j$ by using the conditional expectation (or the linear regression, in the nonnormal case) of $\underline{\boldsymbol{\beta}}_j$, given $\underline{\mathbf{y}}$. The shrinkage estimate has the simple expression

$$\tilde{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\Theta}}_j\hat{\boldsymbol{\beta}}_j + (\mathbf{I} - \hat{\boldsymbol{\Theta}}_j)\mathbf{Z}_j\hat{\boldsymbol{\gamma}}, \tag{37}$$

with

$$\hat{\boldsymbol{\Theta}}_j = \hat{\boldsymbol{\Omega}}\hat{\mathbf{W}}_j^{-1}. \tag{38}$$

Thus, the shrinkage estimator $\tilde{\boldsymbol{\beta}}_j$ is in the class of matrix weighted averages, and the algebra and geometry derived in Chamberlain and Leamer (1976) apply. Using the weighted average interpretation can help in understanding the regression coefficients. It can also help in understanding the frustration of the principal of an excellent school who sees the predictions of success of her students shrunken towards the mean.

The fact that we actually have three different estimates of $\boldsymbol{\beta}_j$ offers many opportunities for diagnostics which have not really been explored so far. In fact, the emphasis in the literature has been on the appropriateness and the plausibility of the model, and not on the ways in which the model can be violated. This is perhaps a useful attitude in the initial stages of development, but the time has come to become more critical. One possibility is to relax the assumptions and to fit more general models. As we know, going this route means going further into the minefield of Plausibility, declaring war on Parsimony and its faithful ally Stability. The other possibility is to use diagnostics, either graphical or computational. There have been a few attempts to develop such tools for the mixed linear model (Beckman, Nachtsheim, & Cook, 1987; Christensen, Pearson, & Johnson, 1992; Lange & Ryan, 1989), but their usefulness for multilevel analysis is just beginning to be explored by Hilden-Minton (1994, 1995).

Algorithms and Computer Programs

Some people think, perhaps, that it is irrelevant for the ordinary user which algorithm is used to compute, say, maximum likelihood estimates. Moreover, some may think, it is equally irrelevant which computer program is used to compute the estimates. But this is true in the same sense that it is irrelevant which means of transportation you use to get to your work. Eventually you will get there all right, no matter what means of transportation you use, but walking takes hours, the bus is unpleasant, and an old car breaks down all the time. The review by Kreft, de Leeuw, and Kim (1990) (see also Kreft, de Leeuw, & van der Leeden, 1994) shows that algorithms do matter, and that, consequently, the NCES question about software makes perfect sense. Related comparisons are in van der Leeden, Vrijburg, and de Leeuw (1991). On the basis of this comparison, the answer to

> **Question 4:** Are there alternatives to the HLM software that NCES should consider using?

is a resounding yes.

In the first place, this is a "yes" in the general sense. The two-step ordinary and weighted least squares methods deserve some additional study. The

nonparametric and semiparametric methods, and the path analysis and latent variable versions of the multilevel models, should also be studied in detail. And, perhaps most importantly, software should be developed that studies the deviations from the multilevel model, preferably in a graphical and interactive way.

Secondly, it is a "yes" in the narrow sense. As far as algorithms for maximum likelihood estimation are concerned, the alternatives are clear. We can choose between the scoring method in Longford's VARCL, the iterative generalized least squares (IGLS) methods in Goldstein's ML/2 and ML/3, and the EM algorithm in HLM and Mason's GENMOD. It is also obvious that there is no uniformly best method, and that none of the three may provide the final answer. Hilden-Minton's (1994) TERRACES package combines EM and scoring in a single algorithm.

If we compare advantages and disadvantages, then EM has global convergence from any starting point to a solution which is always feasible (no negative variances). This advantage, however, is also its undoing in other situations. Global convergence means small steps, and thus slow convergence. If there is convergence to a boundary point, EM slows down to a crawl, and it will not get there in our lifetime. Technically, EM becomes sublinear in such circumstances. The user will have stopped long before this, at a point which looks stationary because nothing is really changing anymore. Because EM typically does not give information about the quadratic component of the likelihood function in the region in which it meanders, there is very little information available that can be used to diagnose this situation. Scoring is often said to have locally quadratic convergence, but this is true only if the model is true, which it is not, and if convergence is not to a boundary point or a point where the information matrix is singular. In examples that are ill-conditioned, VARCL also slows down and becomes linear or worse. Both VARCL and IGLS, however, give better indications that something is wrong. Variances become negative, inverses explode, and so on.

From the results of Kreft et al. (1990) and Kreft et al. (1994), we conclude that VARCL is more difficult to use than HLM, but gets one to the same solution faster if the model is well-conditioned. If the model is way off, then VARCL has better ways of showing this. More or less the same thing is true for ML/3, but ML/3 is really an interactive software package with a much more general range than HLM. With ML/3 we can study residuals, compute summary statistics, make plots, and so on. The learning curve is much steeper, but this is unavoidable. Even steeper learning curves result if the user decides to write multilevel software in Xlisp-Stat or S-Plus, interactive statistical environments that are rapidly becoming more popular. These give the maximum amount of user control, but also require the maximum amount of prior knowledge.

To put it somewhat differently, the HLM program assumes from the start that the basic model is correct, and the number of variations and tests within

the basic model that can be tried out is consequently quite limited. Clearly, the developers of HLM have a different class of users in mind. Because NCES has to deal also with more sophisticated users, who want to explore their data, experiment with models, and investigate the residuals, we think ML3 should be available as well. We see the simple order, in terms of *precookedness,*

$$\text{HLM} \geq \text{VARCL} \geq \text{ML/3} \geq \text{XLISP}.$$

This ordering implies that the programs cater to different groups of users.

Discussion

We have seen that the NCES questions can only partially be answered. This is partly because of the confusion between choice of model, choice of technique, choice of algorithm, and choice of computer program. Each of these choices requires some care.

Multilevel models with random coefficients are an elegant conceptualization. In some cases they are not really necessary—for instance, with very large groups, with very small intraclass correlation, and for researchers who are interested only in the regression coefficients γ.

If we decide to use these models, then it is unclear so far what the best estimation method is. Results of Busing (1993), van der Leeden and Busing (1994), and Kim (1990) show that γ can be reliably estimated with any weighted or unweighted least squares method. This implies, by the way, that we cannot expect large differences between OLS and ML/EB as far as scientific conclusions based on γ are concerned. The reason why this is especially important is that most researchers seem to be interested in the fixed-level regression coefficients, not in the shrinkage estimates for each school, and not in the variance and covariance components (Kreft & Yoon, 1994).

If we decide to use these models, and to use ML/EB, then we still have a choice of algorithm. So far, it seems that a safeguarded version of scoring and an accelerated version of EM are about equally fast and equally reliable. Finally, the choice of computer program is becoming more and more interesting. TERRACES (Hilden-Minton, 1994) is interactive, works on the Mac, with MS Windows, and with X11, and is free. It also has diagnostics, and is embedded in Xlisp-Stat, which means that additional statistical computations can very easily be done on-line. Don Hedeker has published public domain versions of his MIXOR and MIXREG programs, which can deal with autoregressive residuals and categorical responses. These programs, which are not covered by Kreft et al. (1994), will present a serious challenge to the older packages.

Note

[1]They also illustrate the dominant position of the terminology and notation of Bryk and Raudenbush (1991), and of the computer program HLM (Bryk, Raudenbush, Seltzer, & Congdon, 1988) in the field of "official" educational statistics. In many cases it seems as if "fitting a multilevel model" and "using HLM" are seen as identical activities. They are not, of course. To avoid confusion, we shall not use the term "hierarchical linear models," and if we say HLM we mean the computer program of that name.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interaction.* Newbury Park, CA: Sage.

Beckman, R. J., Nachtsheim, C. J., & Cook, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics, 29,* 413–426.

Boyd, L. H., & Iversen, G. R. (1979). *Contextual analysis: Concepts and statistical techniques.* Belmont, CA: Wadsworth.

Bryk, A. S., & Raudenbush, S. (1991). *Hierarchical linear models for social and behavioral research: Applications and data analysis methods.* Newbury Park, CA: Sage.

Bryk, A., Raudenbush, S. W., Seltzer, M., & Congdon, R. T. (1988). *An introduction to HLM: Computer program and user's guide.* Chicago: University of Chicago.

Burstein, L., Linn, R. L., & Capell, F. J. (1978). Analyzing multilevel data in the presence of heterogeneous within-class regressions. *Journal of Educational Statistics, 3,* 347–383.

Busing, F. M. T. A. (1993). *Distribution characteristics of variance estimates in two-level models* (Tech. Rep. No. PRM 93-04). Leiden, The Netherlands: University of Leiden, Department of Psychometrics.

Chamberlain, G., & Leamer, E. E. (1976). Matrix weighted averages and posterior bounds. *Journal of the Royal Statistical Society, B38,* 73–84.

Christensen, R., Pearson, L. M., & Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics, 34,* 38–45.

Cox, D. R. (1984). Interaction. *International Statistical Review, 52,* 1–31.

Cressie, N., & Lahiri, S. N. (1993). The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis, 45,* 217–233.

de Leeuw, J., & Kreft, I. G. G. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 11,* 57–86.

de Leeuw, J., & Liu, G. (1993). *Augmentation algorithms for mixed model analysis.* Los Angeles: University of California, Los Angeles, Department of Statistics.

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association, 57,* 369–375.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and related problems. *Journal of the American Statistical Association, 72,* 320–340.

Hemelrijk, J. (1966). Underlining random variables. *Statistica Neerlandica, 20,* 1–7.

Hilden-Minton, J. (1994). *TERRACES: An XLISP-STAT package for multilevel modeling with diagnostics* (Tech. Rep.). Los Angeles: University of California, Los Angeles, Department of Statistics.

Hilden-Minton, J. (1995). *Multilevel diagnostics for mixed and hierarchical linear models.* Unpublished doctoral dissertation, University of California, Los Angeles.

Johansen, S. (1982). Aymptotic inference in random coefficient regression models. *Scandinavian Journal of Statistics, 9,* 201–207.

Kim, K.-S. (1990). *Multilevel data analysis: A comparison of analytical alternatives.* Unpublished doctoral dissertation, University of California, Los Angeles.

Klein, L. R. (1953). *A textbook of econometrics.* Evanston, IL: Row, Peterson and Co.

Kreft, I. G. G., de Leeuw, J., & Kim, K.-S. (1990). *Comparing four different statistical packages for hierarchical linear regression: GENMOD, HLM, ML3, and VARCL* (CSE Tech. Rep. 311). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Kreft, I. G. G., de Leeuw, J., & van der Leeden, R. (1994). Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, VARCL. *The American Statistician, 48,* 324–335.

Kreft, I. G., & Yoon, B. (1994). *Are multilevel techniques necessary? An attempt at demystification.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. TM 021737)

Lange, N., & Ryan, L. (1989). Assessing normality in random effects models. *The Annals of Statistics, 17,* 624–642.

Longford, N. T. (1990). VARCL software for variance component analysis of data with nested random effects (maximum likelihood) [Computer software]. Princeton, NJ: Educational Testing Service.

Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annals of Statistics, 5,* 746–762.

Rasbash, J., Prosser, R., & Goldstein, H. (1989). *M12 software for two-level analysis: User's guide.* London: University of London, Institute of Education.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science, 6,* 15–51.

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods and Research, 22*(3), 342–363.

Swallow, W. H., & Monahan, J. F. (1984). Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics, 26,* 47–57.

Swamy, P. A. V. B. (1971). *Statistical inference in a random coefficient model.* New York: Springer.

van der Leeden, R., & Busing, F. M. T. A. (1994). *First iteration versus IGLS/RIGLS estimates in two-level models: A Monte Carlo study with ML3* (Tech. Rep. No. PRM 94-02). Leiden, The Netherlands: University of Leiden, Department of Psychometrics.

van der Leeden, R., Vrijburg, K., & de Leeuw, J. (1991). *A review of two different approaches for the analysis of growth data using longitudinal mixed linear models: Comparing hierarchical linear regression (ML/3, HLM) and repeated measures design with structured covariance matrices (BMDP-5V)* (preprint). Los Angeles: University of California, Los Angeles, Department of Statistics.

Authors

JAN DE LEEUW is Professor, Departments of Psychology and Mathematics, UCLA, 405 Hilgard Ave., Los Angeles, CA 90024-1555; deleeuw@stat.ucla.edu. He specializes in multivariate analysis and computational statistics.

ITA G. G. KREFT is Associate Professor, School of Education, California State University, 5151 State University Drive, Los Angeles, CA 90032-8143; kreft@stat.ucla.edu. She specializes in methods of research and data analysis.