Random coefficient models

Jan de Leeuw
Departments of Psychology and Mathematics UCLA
Ita G.G. Kreft
Institute for Social Science Research UCLA

Introduction

Hierarchical linear models, also known as *multilevel models*, have been studied quite intensively in educational statistics in the last five years. We refer to the basic papers by Aitkin and Longford (1986), Raudenbush and Bryk (1986), and De Leeuw and Kreft (1986), and to the books by Goldstein (1987), Bock (1989), and Raudenbush and Bryk (1991). Most of these papers and books concentrate on two-level models, with measurements (variables) on both levels of the hierarchy.

In this review paper we develop notation, terminology, and algorithms for fitting general hierarchical models. The usual notation for multilevel analysis rapidly becomes very complex if the number of levels increases (compare Kreft, De Leeuw, and Kim, 1990). Therefor we shall use the somewhat simpler random coefficient notation.

Random coefficient models

Notation

First we define a *hierarchical index structure* (of level s). Suppose $\mathcal{I} = \{1, \ldots, n\}$ is an index set, and $\mathcal{P}_r$ with $r = 0, \ldots, s+1$ is a hierarchical sequence of *partitionings* of $\mathcal{I}$. By this we mean that $\mathcal{P}_{r+1}$ is a *refinement* of $\mathcal{P}_r$: $\mathcal{P}_{r+1}$ is the union of partitionings of the sets in $\mathcal{P}_r$. For our real example we take the GALO data, used previously in De Leeuw and Kreft (1986). There are 37 schools in the city of Groningen, The Netherlands, with 1290 students. School 1 has 12 students, school 2 has 26 students, ... , school 37 has 30 students. Thus the partitionings are

$$\{\{1,2,3,\cdots,1290\}\}$$
$$\{\{1,2,3,\cdots,12\}\{13,14,15,\cdots,38\}\cdots\{1261,1266,1267,\cdots,1290\}\}$$
$$\{\{1\}\{2\}\{3\}\cdots\{1290\}\}$$

This defines a hierarchical structure with three levels. The first and the last partition in this example are *trivial*, the middle one is *nontrivial*. We use index s for the highest (i.e. finest) nontrivial level. For the GALO example $s = 1$. Level $s+1$ is always the individuals, level 0 is always the whole index set.

There is some additional information about the data set that we shall need. For each of the students we know gender, intelligence quotient (measured with the GIT, a general intelligence test developed in The Netherlands in the fifties), father's profession (classified in six SES-type categories), and teacher's advice. This last variable is the opinion of the sixth grade teacher about the most appropriate form of secondary education for the student. This is classified in seven categories, which are (or were) the seven main types of secondary education in The Netherlands. For the time being we shall treat all four variables as quantitative. This means that we have three quantititative predictors, and a quantitative dependent variable. Clearly this is not entirely appropriate. Both father's profession and teacher's advice are really categorical variables, even the "correct" order

of the categories is somewhat in doubt. IQ is numerical, and SEX is binary, so it can be treated as a numerical variable without any problems. If we decide to treat SES as categorical, our number of predictors will increase from three to eight. If we decide to treat teacher's advice as categorical, we need to look for extensions of the basic linear model which allow for categorical dependent variables.

We need some additional notation in order to get started. Remember the definition of the Kronecker symbol, which is

$$\delta^{ik} = \begin{cases} 1, & \text{if } i = k; \\ 0, & \text{if } i \neq k. \end{cases}$$

We generalize this by defining

$$\delta_r^{ik} = \begin{cases} 1, & \text{if } i \text{ and } k \text{ are in the same set on level } r; \\ 0, & \text{otherwise.} \end{cases}$$

The hierarchical nature of the partitionings implies that if $\delta_r^{ik} = 0$ then $\delta_t^{ik} = 0$ for all $t \geq r$. If students are not in the same school district, then they are not in the same school, and certainly not in the same class.

Also remember the definition of the *direct sum* of a number of matrices. If $A_1, \cdots, A_m$ are m matrices, where A_j has r_j rows and c_j columns, then the direct sum $A = A_1 \oplus \ldots \oplus A_m$ is a block-diagonal matrix with the A_j along the diagonal. Thus the direct sum A has $\sum_{j=1}^m r_j$ rows and $\sum_{j=1}^m c_j$ columns. It looks like

$$A = \begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_m \end{pmatrix},$$

Using this definition makes it possible to write quite a number of expressions in a more compact way.

Basic Model

Now let us define the random coefficient regression model we are interested in. In order to be perfectly explicit, we underline random variables or random vectors . Suppose for each $i \in \mathcal{I}$ we have an observed random outcome $\underline{y}_i$ and observations on, say, m predictors variables, collected in a vector x_i, which is supposed to be nonrandom. We also have a hierarchical index structure with s nontrivial partitionings. The model for the outcome of individual i is

$$\underline{y}_i = x_i'\beta + x_i'(\underline{\vartheta}_{i1} + \ldots + \underline{\vartheta}_{is}) + \underline{\epsilon}_i$$

For the vector disturbances $\underline{\vartheta}_{ir}$ and the scalar disturbances $\underline{\epsilon}_i$ we assume that they have expectation zero. Moreover

$$\mathbf{E}(\underline{\epsilon}_i, \underline{\epsilon}_k) = \delta^{ik}\sigma^2,$$
$$\mathbf{E}(\underline{\vartheta}_{ijr}, \underline{\vartheta}_{klt}) = \delta^{rt}\delta_r^{ik}\omega_{jlr}.$$

It follows that each of the $\underline{\vartheta}_{ir}$ can be identified with one of the levels of the partitioning. Disturbances of different levels are uncorrelated. Disturbances of the same level are correlated if the individuals are in the same subset on that level, and the correlation is the same for all pairs of individuals. Thus the parameters in the problem are the m elements of β, the single parameter σ^2, and for each of the s levels an $m \times m$ matrix Ω_r. Of course the Ω_r are symmetric and positive semidefinite. We shall usually make the model more specific, by requiring that some of the elements of β are zero, or that some of the elements of the Ω_r are zero. If a β_j is zero, we say that the variable x_j has no *fixed effect*. If row j and column j of Ω_r are zero, we say that variable j has no *random effect* on level r.

It follows from the specifications so far that

$$\mathbf{E}(\underline{y}_i) = x_i'\beta,$$

$$\mathbf{E}((\underline{y}_i - \mathbf{E}(\underline{y}_i))(\underline{y}_k - \mathbf{E}(\underline{y}_k))) = x_i'\{\sum_{r=1}^{s(i,k)} \delta_r^{ik}\Omega_r\}x_k + \sigma^2\delta^{ik},$$

where $s(i,k)$ is the highest level for which i and k are still in the same group. Thus if the levels are district-school-class-student, then students in different classes in the same school have $s(i,k) = 2$ and students in different districts have $s(i,k) = 0$.

Matrix notation

For computational purposes it is convenient to rewrite our model in matrix notation. For the expected values this is trivially $\mathbf{E}\,\underline{y} = X\beta$.

The covariances are a bit more complicated. Remember that level r has k_r groups. We define X_r as the direct sum of k_r matrices X_{vr}, where the X_{vr} stacked on top of each other are X. Level 0 has only one group, and thus $X_0 = X$. All X_r contain the same observed numbers, but these numbers are organized differently in the direcxt sum matrices, reflecting the group structure on level r. Thus X_{11} in GALO has the values of the independent variables for the 12 students in school 1, and so on. Because we have three predictors plus an intercept, the matrix X_1 has 1290 rows and 37 x 4 = 148 columns.

Moreover we also define $\breve{\Omega}_r$ by using the direct sum. Thus

$$\breve{\Omega}_r = \overbrace{\begin{pmatrix} \Omega_r & 0 & \cdots & 0 \\ 0 & \Omega_r & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_r \end{pmatrix}}^{k_r \text{ times}}.$$

Using this notation we can now write

$$\mathbf{E}((\underline{y} - X\beta)(\underline{y} - X\beta)') = \sum_{r=1}^{s} X_r\breve{\Omega}_r X_r' + \sigma^2 I_n.$$

There is an interesting property of the X_r that will be important in further derivations. If $span(X_r)$ is the space spanned by the columns of X_r, then $span(X_0) \subset span(X_1) \subset$

$\cdots \subset span(X_s) \subset span(X_{s+1})$. If $null(X_r)$ is the space of all z such that $z'X_r = 0$, then $null(X_0) \supset null(X_1) \supset \cdots \supset null(X_s) \supset null(X_{s+1})$. If all rows of X are nonzero, then $null(X_{s+1}) = \emptyset$ and $span(X_{s+1}) = \mathcal{R}^n$. In fact, it is true for all r that $span(X_r)$ and $null(X_r)$ are two orthogonal complementary subspaces of $\mathcal{R}^n$.

Relation to multilevel models

As explained in the introduction, random coefficient models often occur in the social and behavioural sciences as *multilevel models*. In this subsecetion we explain the relationship between the two types of mdoels. Multilevel models are usually specified in two or more steps. We first assume a model of the form

$$\underline{y}_i = x_i'\underline{b}_i + \underline{\epsilon}_i.$$

Observe that the regression coefficients are now random variables. The first-level disturbances $\underline{\epsilon}_i$ have the usual properties. The next step is to assume that there are second-level predictors z_i which predict the variation in the first-level regression coefficients $\underline{b}_i$. Thus

$$\underline{b}_{ij} = z_i'\underline{\gamma}_{ij} + \underline{\delta}_{ij}.$$

The disturbances $\underline{\delta}_{ij}$ have expectation zero, as usual, but their covariance is

$$\mathbf{E}(\underline{\delta}_{ij}, \underline{\delta}_{kl}) = \delta^{ik}_{s-1}\omega_{jl}.$$

Thus disturbances for individuals which are not in the same subset on the highest non-trivial level (in the same class in our small example, or in the same school in GALO) are uncorrelated. Very often, although not always, we assume that the predictors z_i also have level $s-1$, i.e. $z_i = z_k$ if i and k are in the same set on level $s-1$.

It is now easy to see how we can proceed with a third step, specifying a model with disturbances (and perhaps predictors) of level $s-2$. In our artificial example this means incorporating school-level errors. The only question which must be resolved is how to stop. In the GALO example the school level is the highest (and only) non-trivial level. In that case we stop after two specification steps, and assume that the $\underline{\gamma}_{ij}$ are actually fixed, and equal to γ_j. Thus the model now becomes, switching to element-wise notation,

$$\underline{y}_i = \sum_{j=1}^{m} x_{ij}\underline{b}_{ij} + \underline{\epsilon}_i,$$

$$\underline{b}_{ij} = \sum_{u=1}^{v} z_{iu}\gamma_{ju} + \underline{\delta}_{ij}.$$

If we substitute the second equation into the first one we find

$$\underline{y}_i = \sum_{j=1}^{m}\sum_{u=1}^{v} x_{ij}z_{iu}\gamma_{ju} + \sum_{j=1}^{m} x_{ij}\underline{\delta}_{ij} + \underline{\epsilon}_i.$$

We see that the multilevel model is a special random coefficient model in which the fixed effects γ_{ju} correspond with interactive variables $x_{ij}z_{iu}$. These interactive variables do not have random effects. The only random effects come from the predictors in the first step specification. It is possible, however, that some of the x_{ij} or z_{iu} are constants (they are the same for all i). If $z_{iu} = 1$ for all i, then obviously $x_{ij}z_{iu} = x_{ij}$ and thus x_{ij} can have both fixed and random effects. This treatment generalizes easily to more than two levels of specification.

The Loglikelihood function

We now compute the multinormal likelihood of the outcomes. As explained in De Leeuw and Kreft (1986) this does not necessarily mean that we actually *assume* multivariate normality, it merely means that we choose one particularly simple and appealing way to measure loss. We measure, simultaneously, the fit of the expected values and the fit of the residual dispersions. Expected values are fitted well if the residuals are small, residuals are fitted well if they have the appropriate covariance structure. The multinormal log likelihood combines the (weighted) sums of squares metric for the size of the residuals and the log-determinant covariance metric for the fit of the structural part into one convenient loss function.

The negative log likelihood function is simply (ignoring the usual irrelevant constants)

$$\mathcal{L} = \ln \ det[\Sigma(\theta)] + (\underline{y} - X\beta)'[\Sigma(\theta)]^{-1}(\underline{y} - X\beta)$$

Here θ contains all the variance parameters, i.e. the Ω_r and σ^2, and

$$\Sigma(\theta) = \sum_{r=1}^{s} X_r \breve{\Omega}_r X_r' + \sigma^2 I_n.$$

A simplification at no cost

In order to simplify the calculations we first transform the outcomes by an orthonormal transformation. This will not change the likelihood.

Start with X_s, with s the highest nontrivial level. We know that X_s is the direct sum of k_s matrices X_{vs}, one for each group v on level s. Matrix X_{vs} has, say, n_v rows and m columns, and rank ρ_v. This means that we can write $X_{vs} = K_v T_v$, with K_v of dimensions $n_v \times \rho_v$ and orthonormal, and with T_v of dimensions $\rho_v \times m$ and of rank ρ_v. Of course ρ is the sum of the ρ_v. Moreover K can be chosen as $K = K_1 \oplus \ldots \oplus K_{k_s}$.

Suppose K is an orthonormal basis for the column space of X_s, with s the highest nontrivial level, and suppose $K_\perp$ is an orthonormal basis for the complementary subspace (the null space of X_s). Suppose the dimension of K (the rank of X_s) is ρ. Define

$$\begin{pmatrix} \underline{u} \\ \underline{v} \end{pmatrix} = \begin{pmatrix} K'\underline{y} \\ K_\perp'\underline{y} \end{pmatrix}$$

Because of the hierarchical nature of the data we have $K_\perp' X_r = 0$ for all $r = 1, \ldots, s$, and also $K_\perp' X = 0$. This implies that $\underline{u}$ and $\underline{v}$ are uncorrelated. Moreover

$$\mathbf{E}(\underline{v}) = 0,$$
$$\mathbf{E}(\underline{v}\,\underline{v}') = \sigma^2 I_{n-\rho}.$$

Now let us look at $\underline{u}$. We know that X_s is the direct sum of k_s matrices X_{vs}, one for each group v on level s. Matrix X_{vs} has, say, n_v rows and m columns, and rank ρ_v. This means that we can write $X_{vs} = K_v T_v$, with K_v of dimensions $n_v \times \rho_v$ and orthonormal, and with T_v of dimensions $\rho_v \times m$ and of rank ρ_v. Of course ρ is the sum of the ρ_v. Moreover K can be chosen as $K = K_1 \oplus \ldots \oplus K_{k_s}$.

Define T, of order $\rho \times m$, by stacking the T_v on top of each other. Moreover we write T_r for $K' X_r$, and we find that T_r has exactly the same blockdiagonal structure as X_r, but with each of the X_{vr} replaced by smaller matrices T_{vr}. We illustrate this with our small example which has four groups on the highest nontrival level (four classes), and two groups on the level before that (schools).

$$\begin{pmatrix} K_1' & 0 & 0 & 0 \\ 0 & K_2' & 0 & 0 \\ 0 & 0 & K_3' & 0 \\ 0 & 0 & 0 & K_4' \end{pmatrix} \times \begin{pmatrix} X_1 & 0 \\ X_2 & 0 \\ 0 & X_3 \\ 0 & X_4 \end{pmatrix} = \begin{pmatrix} T_1 & 0 \\ T_2 & 0 \\ 0 & T_3 \\ 0 & T_4 \end{pmatrix}.$$

It follows that

$$\mathbf{E}(\underline{u}) = T\beta,$$
$$\mathbf{E}((\underline{u} - T\beta)(\underline{u} - T\beta)') = \sum_{r=1}^{s} T_r \breve{\Omega}_r T_r' + \sigma^2 I_\rho.$$

Let us see what the effects of this simplication are on the GALO data. We have three predictors (SEX, SES, IQ) and 37 schools. This means that T has (at most) $\rho = 111$ rows. If some of the X_{1v} are singular, ρ is even smaller.

Partitioning the loss even further

We now make an additional transformation, by defining

$$\hat{\beta}_1 = T_1^+ \underline{u}$$

and

$$r_1 = \underline{u} - T_1 \hat{\beta}_1.$$