A Simple Multilevel Technique

Jan de Leeuw

Departments of Psychology and Mathematics

Ita Kreft

Center for the Study of Evaluation

The problem

Multilevel models, also known as hierarchical linear models, random coefficient models or Bayesian linear models, are becoming quite popular data analysis tools in the social and behavioural sciences. The ~~important~~ defining characteristic of the ~~techniques~~ is that data are collected on more than one *level*. In this paper we shall discuss the case of two levels, for definiteness using *students* to describe the units on the first level and *classes* for the units on the second level. Of course many other examples can be given, in many cases with more than two levels. A fixed regressor linear model, with different parameters, is fitted on the student-level in each of the classes. It is assumed, however, that the linear model parameters satisfy a second regression model, which relates them to a number of fixed class-level variables. By combining these two regression models, one for each level, into a single comprehensive model we have a parsimoneous and easily interpretable model of our hierarchical data structure.

We shall not go into details here, but merely refer the interested reader to the voluminous literature. Some of the more important review papers in the area are .... Computer programs have been written by Longford (..), Goldstein et al. (..), Raudenbusch (), Mason (). The techniques all compute maximum likelihood estimates. Although they use various different algorithms, they are all iterative and computationally quite demanding. This is not such a large disadvantage any more as it used to be, but still ~~it do~~es mean that special purpose software is needed, and that the statistical properties of the resulting estimates are ~~very~~ complicated. There is, except in ~~very~~ special cases, no simple relationship any more between the estimates computed by the programs and the more familiar regression parameters. This, much more than the actual cost of computation, is a disadvantage, although perhaps not a very serious one.

one can disagree on the question how serious it is

Outline of model

Consider the following model, ~~called model~~ I. In the formulas below index i stands for students (i=1,...,n), j for classes (j=1,...,m). ~~Class~~ j has n_j students. Random variables are underlined.

School

1: $\underline{y}_{ij} = \underline{a}_j + \underline{\varepsilon}_{ij}$,

2: $\underline{a}_j = \alpha + \beta z_j + \underline{\delta}_j$,

3: $\underline{\varepsilon}_{ij} \sim \mathcal{N}(0,\sigma^2)$,

4: $\underline{\delta}_j \sim \mathcal{N}(0, \omega^2/n_j)$,

5: the $\underline{\varepsilon}_{ij}$ and $\underline{\delta}_j$ are mutually independent.

It follows from these assumptions that the $\underline{y}_{ij}$ are normal, with mean $\alpha + \beta z_j$, with variance $\sigma^2 + \omega^2/n_j$, with covariance ω^2/n_j if the students are from the same class, and with covariance zero if they are from different classes. For obvious reasons we call this a *random intercept model*, there are no first level regressors other than the intercept.

In this model the deviance, i.e. twice the negative logarithm of the likelihood function, has a very simple form. It is, except for irrelevant constants, $\Delta = \Delta_1 + \Delta_2$, where

$$\Delta_1 = m \ln \lambda^2 + \lambda^{-2}\Sigma_{j=1}^{m} n_j(y_j - \alpha - \beta z_j)^2,$$

$$\Delta_2 = (n - m) \ln \sigma^2 + \sigma^{-2} \Sigma_{j=1}^{m} \Sigma_{i \in I(j)} (y_{ij} - y_j)^2.$$

In these expressions we use y_j for the class mean, we use λ^2 for $\sigma^2 + \omega^2$, and we use I(j) for the index set of all students in class j.

The proof of this basic theorem , in fact of a much more general version, is given in the next section. It follows that the maximum likelihood estimates of α and β are computed by performing a weighted regression of the school means y_j on the z_j. If $y_{\heartsuit}$ and $z_{\heartsuit}$ are the overall means, then

$$\text{est}(\beta) = \Sigma_{j=1}^{m} n_j(y_j - y_{\heartsuit})(z_j - z_{\heartsuit}) / \Sigma_{j=1}^{m} n_j(z_j - z_{\heartsuit})^2$$

$$\text{est}(\alpha) = (y_{\heartsuit} - \text{est}(\beta) z_{\heartsuit})$$

The maximum likelihood estimate of λ^2 is the residual sums of squares of this regression. Thus

$$\text{est}(\lambda^2) = \tfrac{1}{m} \Sigma_{j=1}^{m} n_j(y_j - \text{est}(\alpha) - \text{est}(\beta) z_j)^2.$$

The maximum likelihood estimate of σ^2 is the pooled within class variance, i.e.

$$\text{est}(\sigma^2) = \tfrac{1}{n-m} \Sigma_{j=1}^{m} \Sigma_{i \varepsilon I(j)} (y_{ij} - y_j)^2.$$

This is only true if $\text{est}(\lambda^2) > \text{est}(\sigma^2)$. Otherwise

$$\text{est}(\lambda^2) = \text{est}(\sigma^2) = \tfrac{1}{n} \{\Sigma_{j=1}^{m} n_j(y_j - \text{est}(\alpha) - \text{est}(\beta) z_j)^2 + \Sigma_{j=1}^{m} \Sigma_{i \varepsilon I(j)} (y_{ij} - y_j)^2\},$$

which is simply the residual sum of squares of the regression of $\underline{y}_{ij}$ on z_j.

The fact that we can do our computations in two separate simple steps is, of course, very convenient. It depends critically on Assumption 4, which says that the dispersions of the $\underline{\delta}_j$ are inversely proportional to the n_j. This implies, among other things, that the correlation ρ between errors of students in the same class is equal to $\omega^2/(\omega^2 + n_j\sigma^2)$, which shows that ρ is relatively small for large classes and relatively large for small classes. Although the assumption is made for mathematical convenience, it seems at least as plausible as the assumption that the dispersions of the $\underline{\delta}_j$ are the same for all classes, which is the usual assumption in random intercept models like this. Of course there will be very little difference between the two assumptions if all classes have roughly the same size.

Extension of the model

An obvious extension of the model is obtained if we allow for first level fixed regressors, for more than one second level fixed regressors, and if we drop the assumption of homogeneity of the error variances in the different classes. With p fixed student level regressors, collected in the n_j x p matrix X_j, the model becomes

1: $\underline{y}_j = X_j\underline{b}_j + \underline{\varepsilon}_j$,
2: $\underline{b}_j = Z_j\gamma + \underline{\delta}_j$,
3: $\underline{\varepsilon}_j \sim \mathcal{N}(0, \sigma_j^2 I)$,
4: $\underline{\delta}_j \sim \mathcal{N}(0, \omega_j^2(X_j'X_j)^{-1})$,
5: the $\underline{\varepsilon}_j$ and $\underline{\delta}_j$ are mutually independent.

Here $\underline{b}_j$ is a random p-element vector for each j. Assumption 2 says that element s of $\underline{b}_j$ is equal to a linear combination of class characteristics, plus a random disturbance term. The characteristics of class j predicting coefficient s are collected in a vector z_{js} of, say, q_s elements, corresponding with q_s second level variables. It is not necessary, therefor, that the same variables are used to predict different coefficients. In (2) $Z_j = z_{j1}' \oplus \ldots \oplus z_{jp}'$, i.e. Z_j is the direct sum of the z_{js}, the $p \times \Sigma_{s=1}^{p} q_s$ matrix with the row vectors z_{js} in the $1 \times q_s$ diagonal submatrices, and zeroes everywhere else. Our previous model is the special case in which p = 1 and $x_{ij1} = 1$ for all i and j, in which q = 2 and $z_{j1} = 1$ for all j, and in which σ_j^2 and ω_j^2 are the same for all j.

Observe that Assumption 4 generalizes the assumption that the second level dispersions are inversely proportional to the class sizes. This multivariate version of

Assumption 4 seems more far-fetched than the univariate version we used earlier, but we shall show that it retains the advantages of simple estimation and interpretation. In order to interpret it, let's look at the ordinary least squares estimate $\mathbb{b}_j = (X_j'X_j)^{-1}X_j'y_j$. By making the usual substitutions we see that $\underline{y}_j$, the scores for the students in class j, satisfy

$$\underline{y}_j = U_j\gamma + X_j\underline{\delta}_j + \underline{\varepsilon}_j,$$

where γ has pq elements, and $U_j = X_jZ_j$. It follows that $E(\mathbb{b}_j) = (X_j'X_j)^{-1}X_j'E(\underline{y}_j) = E(\underline{b}_j) = Z_j\gamma$, and $VAR(\mathbb{b}_j) = (X_j'X_j)^{-1}X_j'VAR(\underline{y}_j)\, X_j(X_j'X_j)^{-1}$. Now $\underline{y}_j$ is normal with dispersion $\omega_j^2X_j(X_j'X_j)^{-1}X_j' + \sigma_j^2I = \lambda_j^2P_j + \sigma_j^2Q_j$. Here $P_j = X_j(X_j'X_j)^{-1}X_j'$ and $Q_j = I - P_j$, while $\lambda_j^2 = \omega_j^2 + \sigma_j^2$, as usual. Thus $VAR(\mathbb{b}_j) = \lambda_j^2(X_j'X_j)^{-1}$, which is consequently very much like the variance of the ordinary least squares estimate in the usual case.

The deviance is, except for irrelevant constants,

$$\Delta = \Sigma_{j=1}^{m} \{ \ln\det(\lambda_j^2P_j + \sigma_j^2Q_j) + (y_j - U_j\gamma)'(\lambda_j^2P_j + \sigma_j^2Q_j)^{-1}(y_j - U_j\gamma)\}.$$

Now

$$\ln\det(\lambda_j^2P_j + \sigma_j^2Q_j) = p\ln\lambda_j^2 + (n_j - p)\ln\sigma_j^2,$$

$$(\lambda_j^2P_j + \sigma_j^2Q_j)^{-1} = \lambda_j^{-2}P_j + \sigma_j^{-2}Q_j.$$

It follows that

$$(y_j - U_j\gamma)'(\lambda_j^2P_j + \sigma_j^2Q_j)^{-1}(y_j - U_j\gamma) =$$

$$\lambda_j^{-2}(y_j - U_j\gamma)'P_j(y_j - U_j\gamma) + \sigma_j^{-2}(y_j - U_j\gamma)'Q_j(y_j - U_j\gamma).$$

Now $P_jU_j = U_{js}$, which means that $Q_jU_j = 0$. Also $P_jy_j = X_j\mathbb{b}_j$, with $\mathbb{b}_j$ the ordinary least squares estimate, i.e. $\mathbb{b}_j = (X_j'X_j)^{-1}X_j'y_j$. This means that

$$(y_j - U_j\gamma)'(\lambda_j^2P_j + \sigma_j^2Q_j)^{-1}(y_j - U_j\gamma) =$$

$$\lambda_j^{-2}(\mathbb{b}_j - Z_j\gamma)'X_j'X_j(\mathbb{b}_j - Z_j\gamma) + \sigma_j^{-2}y_j'Q_jy_j.$$

It follows that

$$\text{est}(\gamma) = (\Sigma_{j=1}^m \text{est}(\lambda_j)^{-2}Z_jX_j'X_jZ_j)^{-1}\Sigma_{j=1}^m \text{est}(\lambda_j)^{-2}Z_jX_j'X_j\mathbb{b}_j,$$

$$\text{est}(\lambda_j)^2 = \tfrac{1}{p}(\mathbb{b}_j - Z_j\text{est}(\gamma))'X_j'X_j(\mathbb{b}_j - Z_j\text{est}(\gamma)),$$

$$\text{est}(\sigma_j^2) = \tfrac{1}{n_j-p}\, y_j'Q_jy_j.$$

Thus no iteration is required to find the $\text{est}(\sigma_j^2)$. If all dispersions are supposed to be equal, i.e. $\lambda_j^2 = \lambda^2$ and $\sigma_j^2 = \sigma^2$, then no iteration is needed at all.

In this case our estimates are (exactly) unbiased, est(γ) is normal, with mean γ, and dispersion

$$\lambda_j^2 (\Sigma Z_j'X_j'X_j Z_j)^{-1}$$

$$\mathbb{b}_j - Z_j[\gamma + \overset{\delta}{\lambda_j^2}\ ]$$

$$\mathbb{b}_j = b_j$$