Biometrika (1986), **73**, 1, *pp.* 43–56
Printed in Great Britain

Multilevel mixed linear model analysis using iterative generalized least squares

BY H. GOLDSTEIN

Department of Mathematics, Statistics & Computing, University of London Institute of Education, London WC1H 0AL, U.K.

SUMMARY

Models for the analysis of hierarchically structured data are discussed. An iterative generalized least squares estimation procedure is given and shown to be equivalent to maximum likelihood in the normal case. There is a discussion of applications to complex surveys, longitudinal data, and estimation in multivariate models with missing responses. An example is given using educational data.

Some key words: Errors in variables; Generalized least squares; Hierarchical data; Longitudinal data; Maximum likelihood; Missing data; Mixed model; Multilevel data; Multivariate linear model; Survey.

1. INTRODUCTION

In the social and other sciences, data are often structured hierarchically. Thus, for example, workers are grouped into workplaces, individuals into households, animals into litters, and subjects can be studied repeatedly, so giving rise to measurements grouped within individuals. It has long been recognized that the existence of such 'clustering' presents particular problems of model specification due to lack of independence between measurements, and techniques for dealing with these have been evolved, for example in the analysis of longitudinal data (Goldstein, 1979), and sample surveys (Holt, Smith & Winter, 1980). In addition, however, the groupings themselves are often of interest in their own right, rather than being conveniences as is typical in surveys. Thus, the grouping of children within classrooms and schools implies an interest in the 'contextual' effects of the characteristics of classes and schools on such things as children's achievements.

Despite much interest in the specification of such models, there seems to have been a limited use of them, partly it appears because of the statistical and computational complexities. The present paper sets out a comprehensive multilevel mixed effects model and shows how efficient estimates can be obtained. It is developed using an educational context, and its use in a range of applications is illustrated.

2. BASIC MODEL AND NOTATION

Consider, as an example, a data set with three levels: schools, classrooms within schools and children within classrooms, and assume that simple random sampling takes place at each level. Suppose also that we have measurements on a response variable for the jth child in the ith classroom within the kth school. The full model is

$$Y_{kij} = \alpha^*_{kij} + \beta^*_{ki} + \gamma^*_k. \tag{1}$$

At each level of the 'hierarchy' we set up a linear model relating the terms in (1) to a function of explanatory variables, as follows

$$\gamma_k^* = \gamma_0 + \gamma_1 w_{1,k} + \ldots + v_k = \sum_{l=0}^{q} \gamma_l w_{l,k} + v_k, \tag{2}$$

where v_k is a random variable with $E(v_k) = 0$, $\text{var}(v_k) = \sigma_v^2$, and γ_l is the school level coefficient for the lth explanatory variable $w_{l,k}$ for school k. Also

$$\beta_{ki}^* = \beta_0 + \beta_{1,k} Z_{1,ki} + \ldots + u_{ki} = \sum_{l=0}^{p} \beta_{l,k} Z_{l,ki} + u_{ki}, \tag{3}$$

where u_{ki} is a random variable with $E(u_{ki}) = 0$, $\text{var}(u_{ki}) = \sigma_u^2(k)$, and $\beta_{l,k}$ is the classroom level coefficient for the lth explanatory variable $Z_{l,ki}$ for classroom ki. Lastly, we have

$$\alpha_{kij}^* = \alpha_0 + \alpha_{1,ki} x_{1,kij} + \ldots + e_{kij} = \sum_{l=0}^{r} \alpha_{l,ki} x_{l,kij} + e_{kij}, \tag{4}$$

where e_{kij} is a random variable with $E(e_{kij}) = 0$, $\text{var}(e_{kij}) = \sigma^2(ki)$, and $\alpha_{l,ki}$ is the child level coefficient of the lth explanatory variable $x_{l,kij}$ for child kij.

We may also have interactions of explanatory variables between levels and these can also introduce further random terms. For example, we may have

$$\alpha_{l,ki} = \beta_0' + \beta_1' Z_{ki}' + u_{ki}', \tag{5}$$

which adds an extra random term $u_{ki}' x_{l,kij}$.

Later we shall see how such error terms and more general random coefficient terms can be incorporated, but for the sake of simplicity this is postponed and all coefficients are treated as fixed.

Thus we can write the simpler full model, combining (2), (3) and (4), as

$$Y_{kij} = \alpha_0 + \gamma_0 + \beta_0 + \sum_{l=1}^{r} \alpha_{l,ki} x_{l,kij} + \sum_{l=1}^{p} \beta_{l,k} Z_{l,ki} + \sum_{l=1}^{q} \gamma_l w_{l,k} + (v_k + u_{ki} + e_{kij}), \tag{6}$$

or as $Y = X\beta + E$. The model is assumed to be of full rank.

We have

$$\text{var}(Y_{kij}) = \sigma_v^2 + \sigma_u^2(k) + \sigma^2(ki),$$

assuming that all covariances between the random variables in (6) are zero. Thus the overall variance of Y can be partitioned into components for school, class and child, hence the term 'variance component' models. We also assume further for simplicity that $\sigma_u^2(k) = \sigma_u^2$ and $\sigma^2(ki) = \sigma^2$, but the general formulation can be estimated; see Appendix 2. We obtain

$$\text{var}(Y_{ki.}) = \sigma_v^2 + \sigma_u^2 + \sigma^2/n_{ki}, \quad \text{var}(Y_{k..}) = \sigma_v^2 + \sigma_u^2 t_k + \sigma^2/n_k,$$

where

$$t_k = \sum_{i=1}^{n_k} t_{ki}^2, \quad t_{ki} = n_{ki}/n_k, \quad n_k = \sum_{i=1}^{m_k} n_{ki}, \quad n = \sum_{k=1}^{m} n_k;$$

$$Y_{ki.} = n_{ki}^{-1} \sum_{j=1}^{n_{ki}} Y_{kij}, \quad Y_{k..} = n_k^{-1} \sum_{i=1}^{m_k} n_{ki} Y_{ki.}$$

Here, n_{ki} is the number of children in the ith classroom of the kth school, m_k is the number of classes in the kth school, and m is the number of schools. Note also that

$$\text{cov}(Y_{kij}, Y_{kij'}) = \sigma_v^2 + \sigma_u^2, \quad \text{cov}(Y_{kij}, Y_{ki'j'}) = \sigma_v^2 \quad (i \neq i').$$

Thus for classrooms we can define the 'intra class coefficient'

$$\rho_{ki} = (\sigma_v^2 + \sigma_u^2)(\sigma_v^2 + \sigma_u^2 + \sigma^2)^{-1},$$

and, for schools,

$$\rho_k = \sigma_v^2(\sigma_v^2 + \sigma_u^2 + \sigma^2)^{-1}.$$

Now for each school k we have

$$\mathrm{cov}_k(Y_{kij}) = \sigma^2 I_{(n_k)} + \sigma_u^2 I_{(m_k)} \otimes J_{(n_{ki})} + \sigma_v^2 J_{(n_k)}. \tag{7}$$

The matrices $I_{(p)}$ and $J_{(q)}$ are respectively the $p \times p$ unit matrix and the $q \times q$ matrix of one's. The symbol $\otimes$ denotes the direct product of the matrices. Thus the total covariance matrix is block diagonal, where each school is represented by a block. We have also for the kth school

$$\mathrm{cov}_k(Y_{ki.}) = \mathrm{diag}(\sigma_u^2 + \sigma^2/n_{ki}) + \sigma_v^2 J_{(m_k)}, \quad \mathrm{cov}(Y_{k..}) = \mathrm{diag}(\sigma_v^2 + \sigma_u^2 t_k + \sigma^2/n_k).$$

3. Estimation

Harville (1977) discusses maximum likelihood and restricted maximum likelihood for normal mixed effects models, and gives formulae for special cases, while N. Longford in an unpublished paper gives a computationally efficient method for obtaining maximum likelihood estimates for the general multilevel mixed effects model. For a two-level model, Mason, Wong & Entwistle (1984) obtain restricted maximum likelihood estimates using the EM algorithm, and Aitkin, Anderson & Hinde (1981) use the EM algorithm for a simple version of (5). Fuller & Battese (1973) show how noniterative but consistent moment estimators of the error variances for a simple three-level model can be obtained and used in generalized least squares estimation.

In (6), Y is an $n \times 1$ vector with $\mathrm{cov}(Y \mid X\beta) = \mathrm{cov}(E) = V$, say. If V is known then we have the usual generalized least squares estimators

$$\hat{\beta} = (X^{\mathrm{T}} V^{-1} X)^{-1} X^{\mathrm{T}} V^{-1} Y, \quad \mathrm{cov}(\hat{\beta}) = (X^{\mathrm{T}} V^{-1} X)^{-1}. \tag{8}$$

If β is known but V is unknown, then we can obtain estimators β^* of the parameters of V using generalized least squares as

$$\hat{\beta}^* = (X^{*\mathrm{T}}(V^*)^{-1} X^*)^{-1} X^{*\mathrm{T}}(V^*)^{-1} Y^*, \tag{9}$$

where Y^* is the vector of the upper triangle elements of $(Y - X\beta)(Y - X\beta)^{\mathrm{T}}$, that is the squares and products of the residuals, and V^* is the covariance matrix of Y^*; X^* is the design matrix linking Y^* to V in the regression of Y^* on X^*.

When neither β nor V is known, the iterative generalized least squares estimates are those which simultaneously satisfy both (8) and (9). In Appendix 1 it is shown that the procedure based on (8) and (9) is equivalent to maximum likelihood in the normal case.

The estimation procedure commences from an initial estimate of V, which we use to obtain estimates $\hat{\beta}$, and then obtain an improved estimate of V, and so on, iteratively, until convergence is achieved. If the initial estimate of β is consistent, for example, if we use $V = \sigma^2 I$, which gives ordinary least squares, then so is the estimate of β^* based upon it, and hence the final estimates, by assuming the existence of finite moments up to the fourth moment. Because the $\hat{\beta}, \hat{\beta}^*$'s are each estimated using consistent estimators, they are asymptotically efficient.

Starting with $V = V_1 = \sigma^2 I_{(n)}$, we obtain $\hat{\beta}_1 = (X^T X)^{-1} X^T Y$, which is a consistent estimator of β. Now form residuals $Y - \hat{Y} = Y - X\hat{\beta}_1$, giving for the kth block

$$Y_k^* = \text{vech}\{(Y_k - \hat{Y}_k)(Y_k - \hat{Y}_k)^T\},$$

and for the model (7) the $n_k \times 3$ matrix

$$X_k^* = \{\text{vech}(J_{(n_k)}), \text{vech}(I_{(m_k)} \otimes J_{(n_{kl})}), \text{vech}(I_{(n_k)})\}.$$

Using the results given in Appendix 2, with current estimates $\hat{\beta}_1^*(\hat{\sigma}_v^2 = \hat{\sigma}_u^2 = 0)$, we obtain new estimates $\hat{\beta}_2^* = (\hat{\sigma}_v^2, \hat{\sigma}_u^2, \hat{\sigma}^2)^T$. The vector vech (A) is formed by stacking the columns of the lower triangle of the symmetric matrix A under one another. The vector vec (A) used in the Appendices is formed by stacking the complete columns of A under one another.

Having done this, we obtain a consistent estimator V_2 and so form a new estimate of β,

$$\hat{\beta}_2 = (X^T V_2^{-1} X)^{-1} X^T V_2^{-1} Y.$$

The procedure is repeated until convergence is obtained. The constraints defining the feasible region for the solution are $(\sigma_v^2, \sigma_u^2, \sigma^2) \geqslant 0$. If the solution lies outside the feasible region, then the constrained solution will lie on the boundary. In this case, at any iteration, we can fix each variance in turn to be zero and reestimate the others, choosing the minimum value of the generalized variance $|\text{cov}(\hat{\beta})|$. If further variances become negative, the process can be repeated setting pairs of variances to zero, etc.

We obtain final estimates $\hat{\beta}$ together with cov $(\hat{\beta})$ and also $\hat{\beta}^{*T} = (\hat{\sigma}_v^2, \hat{\sigma}_u^2, \hat{\sigma}^2)$. We can use cov $(\hat{\beta}^*)$ to test hypotheses about the elements of β^*.

4. Random coefficients

Model (6) can be generalized readily to include random coefficients. In single level models, the random coefficients necessarily are defined at that level; see, for example, Fisk (1967). In a multilevel model, however, if a coefficient at any level in (6) is assumed to be a random variable, it can be written in general as the sum of linear functions of explanatory variables at that level and other levels plus error terms which, in principle, could operate at any level.

Consider, as in the example discussed below, the coefficient of the variable 'size of class' in a model with school, class and child nested levels. This coefficient could be a random variable at the level of the school, indicating that the 'effect' of size of class varies randomly between, but not within, schools. It could be a random variable at the class level indicating that it varied between classes but not within classes. It could also be a random variable at the child level indicating that the child level variance was related to size of class. This could, for example, account for a heterogeneous child level variation, and one aim of an analysis might be to determine whether further explanatory variables measuring children's characteristics could account for this random coefficient variation.

Another example would be where the child level variance differed between, say, single sex and mixed schools. Thus, an indicator variable for school gender would be defined at the level of the school and its coefficient assumed to be a random variable at the child level.

The fixed part of such models will generally involve interaction terms at different levels, such as given by (5), and the random part will involve products of explanatory

variables and random error terms. It is assumed that the fixed part of the model has been specified and the estimation of the random part is now considered. Thus, for example, if $\beta_{l,k}$ is random we might have

$$\beta_{l,k} = \beta'_{l,k} + r_{l,ki} + q_{l,k},$$

where $\beta'_{l,k}$ is a linear function of explanatory variables. The new error term in (6) is

$$r_{l,ki} Z_{l,ki} + q_{l,k} Z_{l,ki} + v_k + u_{ki} + e_{kij}.$$

In practice, as discussed earlier, it will often be convenient to make simplifying assumptions of constant variances. The error term in (6), it should be noted, is the special case where the overall constant in the model is the only random coefficient, with error terms at all three levels.

With more than one random coefficient at any level we will in general have covariances between the random errors. Thus for two random coefficients, $\beta_{l,k}, \beta_{p,k}$, the extra error term at the school level is $r_{l,k} Z_{l,ki} + r_{p,k} Z_{p,ki}$, with, say, $\operatorname{cov}(r_{l,k}, r_{p,k}) = \sigma_{lp}$. Thus, in $\operatorname{cov}_k(Y_{kij})$, σ_v^2 is replaced in (7) by

$$\sigma_v^2 + \sigma_l^2 Z_{l,ki} Z_{l,ki'} + \sigma_p^2 Z_{p,ki} Z_{p,ki'} + \sigma_{lp}(Z_{l,ki} Z_{p,ki'} + Z_{l,ki'} Z_{p,ki}),$$

for classrooms i, i' in school k. The contribution to $\operatorname{cov}(Y_{kij})$ from the school level error terms can be written for the kth school as

$$\sigma_v^2(Z_0 Z_0^{\mathrm{T}}) + \sigma_l^2(Z_l Z_l^{\mathrm{T}}) + \sigma_p^2(Z_p Z_p^{\mathrm{T}}) + \sigma_{lp}(Z_l Z_p^{\mathrm{T}} + Z_p Z_l^{\mathrm{T}}),$$

where Z_l is $n_k \times 1$ with kijth element $Z_{l,ki}$ and $Z_0 = J$, also $n_k \times 1$.

In general, there will also be nonzero covariances between v_k and $r_{l,k}, r_{p,k}$, and the above formulae will be modified in obvious ways. For the class level error terms we have analogous formulae, and likewise for random coefficients at the individual and school levels. The covariances are subject also to the constraint that the corresponding correlations have absolute value less than or equal to unity, and a similar procedure can be used as described in §3 for the constraints on the variances.

The use of random coefficient models may be useful in accounting for heterogeneous variances at any level. It is possible in the extreme case, however, to have as many parameters as equations or more, which will necessitate the introduction of constraints.

The specification of random error terms need not be restricted to the coefficients of the explanatory variables in (6). Additional variables can be introduced with random coefficients which have zero expected values, so that they contribute to the random but not the fixed part of the model. Hence, at each step of the iteration these are used in the estimation of β^* but simply omitted from the estimation of β. Thus, for example, we may wish to introduce variance heterogeneity into a model by allowing a variance to be a particular function of time or age without incorporating the same function into the fixed part of the model. A simple example of such separate error specification arises in ordinary least squares in the case of simple regression through the origin, where the constant term does not appear in the fixed part of the model, but is used to specify the error variance.

5. Constraints among parameters

Linear constraints among the β or β^* parameters can be incorporated by suitable reparameterization. In some applications, however, the elements of β^* may be functions

of β and X, say $g(\beta, X)$. Jobson & Fuller (1980) discuss a single level model of this kind. The present procedure provides efficient estimates for these models.

We incorporate new random variables which are proportional to the functions $g(\beta, X)$. Thus, for example, the constant coefficient of variation model, at the class level, specifies

$$\sigma_u^2 = \sigma_{u,1}^2 g(\beta, X) = \sigma_{u,1}^2 \left(\sum_{l=1}^{p} \beta_{l,k} Z_{l,ki} \right)^2.$$

The second column of the matrix X^* will be multiplied by $g(\beta, X)$, where the current values of β are used at each iteration, in model (7).

6. Further applications

6·1. *Sample surveys*

The above model offers an alternative to the traditional approach to the analysis of complex sample surveys, where the survey design involving stratification and clustering is used to calculate the variances and covariances of derived statistics. Computer programs such as SUPERCARP (Hidiroglou, 1981) use this approach to obtain consistent ordinary least squares or weighted least squares estimates of regression coefficients and then obtain consistent estimates of their standard errors using the known sample design. The main problem with this approach is that it is not efficient, and this will be important if intraclass correlations are high.

The present model can incorporate stratification variables by fitting constants for all strata in the model or, more efficiently, attempting to model strata variations, where multiple factors are used for strata definition.

6·2. *Longitudinal data*

Another important area of application is in models for time related longitudinal data, such as the fitting of polynomial growth curves (Goldstein, 1979). Here we have a sample of individuals, possibly classified in terms of measured explanatory variables or nested within a hierarchical structure, each of whom has a set of measurements on a variable at different ages or time points. Thus, there is a within-individual model of the form

$$\alpha_{ij}^* = \sum_{l=0}^{p} \alpha_{l,i} t_{ij}^l + e_{ij},$$

and the simplest between-individual model, if we omit other classifying factors, is $\beta_i = \beta_0 + u_i$. In general, there will be random variables at the level of the individual, which are correlated among themselves and with the u_i:

$$\alpha_{l,i} = \alpha'_{l,i} + u_{l,i},$$

where all individuals are measured at the same set of k time points. This model is considered by Rao (1965), but is more general in allowing different sets of measurements for each individual in its multilevel model formulation.

In fitting growth curves we might also wish to allow for increasing error variance with time. for example $e_{ij} = \sigma t_{ij}$, and the procedures of §4 can be used.

A special case of the growth curve model is when there are $p+1$ time points and the explanatory variables for the within-individual model are $p+1$ dummy variables

specifying at which time point a measurement is made. The coefficients are then estimates of the means of the measurement at each time point. Since the model does not require a measurement at each time point for each individual, this provides a procedure for the efficient estimation of parameters in mixed longitudinal or rotating survey designs, analogous to that given by Jones (1980), but additionally providing an estimate of the covariance matrix rather than assuming it is known.

6·3. *Multivariate multilevel mixed effects models*

In the above special case of the growth curve model, there is in fact no requirement that the measurement at each time point should be the same variable. Thus, we can think simply of measuring $p+1$ variables for each individual and using the model to estimate their means and covariance matrix. We have now, therefore, specified a multivariate linear model in terms of a two-level univariate mixed effects model. In addition, there is no requirement to have each response variable measured on each individual, and the model can include both extra random or fixed explanatory variables for each response variable and further levels of nesting. Thus, we see that the general multivariate multilevel mixed effects model with possibly missing data can be analysed as a special case of the univariate multilevel mixed effects model with an extra within-individual level. Clearly, also, the model can be used to provide efficient estimates for the usual multiple regression model with missing data.

7. Errors of measurement

The usual model is $X = x+u$ and $Y = y+e$, and the linear model is

$$y = x\beta+q, \quad Y = X\beta+v,$$

with the usual independence assumptions. We write

$$V = \operatorname{cov}(q), \quad v = q+e-u\beta, \quad u \sim N(0, \Omega_{uu}).$$

The x, y represent 'true' scores, and the X, Y observed scores, and it is assumed that we wish to make inferences about the relationship between the true scores, namely β and V. The matrix Ω_{uu} is the $p \times p$ covariance matrix of the measurement errors, where p is the number of explanatory variables. This matrix is assumed to be diagonal, with nonzero entries corresponding to those variables where measurement errors exist. The q, e are the residual terms incorporating random errors from each level of the model. Where replications of (X, Y) are available, these can be incorporated directly into the model as a further level of nesting. Otherwise we require external estimates of the relevant quantities and, following a similar procedure to Warren, White & Fuller (1974), who deal with the ordinary least squares case, the following results are obtained for errors of measurement at the child level. A detailed derivation with extensions for errors of measurement at other levels is given in Appendix 3.

At any iteration a consistent estimator of β is given by $\hat{\beta} = \hat{M}_{xx}^{-1} \hat{M}_{xy}$, where

$$\hat{M}_{xx} = X^{\mathrm{T}} \hat{V}^{-1} X - \{n^{-1} \operatorname{tr}(\hat{V}^{-1})\} S_{uu}, \quad \hat{M}_{xy} = X^{\mathrm{T}} \hat{V}^{-1} Y,$$

and $n^{-1} S_{uu}$ is a consistent estimator of Ω_{uu}, estimated independently of the other error terms. Appendix 3 shows how to obtain a consistent estimator of $\hat{V}$ and $\operatorname{cov}(\hat{\beta})$. With

these modifications the iterative procedure proceeds as before until convergence is achieved.

Where reliability estimates are available, the above procedure can be adapted to that case; see Fuller & Hidiroglou (1978).

When independent replication of X, Y is available for individuals, and there are just two levels, the program LISREL (Joreskog & Sorbom, 1979) will give efficient estimates.

8. EXAMPLE

The data consist of 969 cases from a much larger longitudinal study of educational attainment carried out by the Inner London Education Authority (1969) from 1968 to 1973. The variables used here are standardized reading scores on the same children in the first and final years of junior school, that is, with average ages of 8 and 11 years, together with the size of the class at 8 years, in 1968. Very small classes of fewer than 10 children are omitted from the analysis, since these will often be special classes formed of low-attaining children. Table 1 shows the means and variances of these variables. There were 28 schools in the sample, with between 1 and 3 classes per school. Only children who remained in the same school are used in these analyses.

Table 1. *Regression of 8 year reading score on class size. Fitted constants and standard errors*

Explanatory variables	OLS	*A*	*B*
Overall constant	103·4	103·6	103·1
Class size	−0·32 (0·36)	−0·21 (0·48)	−0·26 (0·24)
Mean class size for school	−0·06 (0·39)	−0·06 (0·56)	
Error parameters	**OLS**	***A***	***B***
σ_v^2		18·1 (8·9)	18·1 (8·9)
σ_u^2		8·2 (6·1)	8·1 (6·0)
σ^2	205·7	182·3 (8·5)	182·3 (8·5)
Intraclass correlations		***A***	***B***
Schools		0·09	0·09
Classrooms		0·13	0·13
Number of iterations		5	5

OLS, ordinary least squares analysis;
A, analysis based on all three explanatory variables;
B, analysis based on first two explanatory variables.
The means and standard deviations of 8 yr reading score, 11 yr reading score and class size are respectively: 94·0, 14·5; 93·8, 14·4; 35·3, 12·3.

Table 1 presents results for the regression of 8 year reading score on the child's class size. In all these analyses, the convergence criterion is that the relative change in each random parameter between consecutive iterations is less than 10^{-3}. Analysis *B* fits a simple regression with error terms at each level. Analysis *A* introduces a further explanatory variable which is the mean class size of the classes in the school. The ordinary least squares analysis is given for comparison and it should be noted that the most marked difference is in the smaller standard errors estimated in the usual way for ordinary least squares. Also, the larger standard error for class size in analysis *A* as compared to analysis *B* results from the high correlation of class size with mean class size for school, a variable which adds little information. This effect of underestimating true

standard errors is well known in sample survey theory where there is positive intraclass correlation as here.

From analysis *B* we see that an increase in class size of 10 children is associated with a decrease of 2·6 score units, or about 18% of the standard deviation of reading test score. The standard error of this coefficient is relatively high, however, with a 95% confidence interval which includes zero, and this is consistent with other findings on class size which show only small relationships with attainment. A quadratic term for class size was also studied but had no discernible effect. As would be expected, the child level error term dominates the others, and it is interesting to note the greater homogeneity between classes than between schools.

Table 2. *Regression of* 11 *year reading score on* 8 *year reading score. Fitted constants and standard errors*

Explanatory variables	OLS	*A*	*B*	*C*
Overall constant	28·8	27·8	27·9	23·9
8 yr reading score	0·73 (0·02)	0·73 (0·02)	0·73 (0·02)	0·74 (0·03)
Mean 8 yr reading score for classroom	−0·0039 (0·06)	−0·0072 (0·08)	−0·014 (0·08)	0·012 (0·08)
Within-classroom standard deviation of 8 yr reading score	−0·21 (0·16)	−0·17 (0·21)	−0·14 (0·21)	−0·05 (0·22)
Error parameters	**OLS**	***A***	***B***	***C***
σ_v^2		0·0	0·0	1·85 (1·84)
$\sigma_{u,0}^2$		4·21 (1·83)	0·0	44·19 (49·43)
$\sigma_{u,1}^2 \times 10^3$			0·46 (0·20)	7·0 (6·0)
$\sigma_{u,01}$				−0·56 (0·54)
σ^2	95·4	91·42 (4·25)	91·32 (4·24)	90·74 (4·29)
Intraclass correlations		***A***	***B***	***C***
Schools		0		
Classrooms		0·04		
Number of iterations		5	4	29

Subscripts 0, 1, 01 refer respectively to the simple error, the 8 yr score coefficient error and the covariance of these errors.

In Table 2 the 11 year reading score is regressed on 8 year score, mean 8 year score for the classroom and the within-classroom standard deviation of 8 year score. Strictly speaking, an adjustment for measurement error in 8 year score should be incorporated, but unfortunately no good estimate of this measurement error variance is available. Nevertheless, for this kind of test, values of the reliability coefficient are typically quoted around 0·95, so that any adjustment would not be substantial and none is used in this analysis. Analysis *A* fits only a simple error term at each level and the school level variance now disappears. This suggests that in terms of reading progress between eight and eleven years, there is some variation between classrooms but none between schools, in contrast to the previous analysis of 8 year reading attainment, where there was more variation between schools than between classrooms. The intraclass coefficient for analysis *A* is now rather small so that the OLS analysis is closer to the multilevel model analysis than it is in Table 1.

Analysis *B* introduces a random error term into the regression coefficient of 8 year score with the error being at the classroom level, and the 'constant' error variance

becomes zero. Finally analysis C adds the covariance term at the classroom level. The solution, at the stated level of accuracy, lies on the boundary of the feasible region, in this case where the correlation is equal to $-1{\cdot}0$.

A further important point is that the estimated variances will depend on the origin and scale chosen for the explanatory variables and, where covariance terms are not fitted, this will affect the adequacy of the fit, and the parameter estimates. This is relevant to the implementation of variance component models in existing packages, which do not fit covariance terms.

9. DISCUSSION

Previous work on the mixed effects linear model has dealt largely with estimation procedures for special cases, although the important paper by Harville (1977) discusses general maximum likelihood estimation. The present paper extends this work in a number of ways.

First, it is shown how mixed effects models at each level of a hierarchy simultaneously can be specified and the hierarchical structure utilized in the estimation. Secondly, a general algorithm is given which is straightforward to apply to any design, and which preliminary experience suggests has good convergence properties. Thirdly, the basic model of Harville (1977) assumes that the random error terms associated with the overall constant term, the residual errors in ordinary least squares, are independent of the other random error terms. This is shown to be both unnecessary and undesirable. Fourthly, the generalized least squares approach, while giving maximum likelihood estimates in the case of normal errors, provides a means of obtaining efficient estimates for other distributions. Finally, the present paper shows how parameter constraints and errors of measurement can be incorporated into the model.

The availability of a practical method for fitting multilevel models with many random error terms raises a number of important considerations which are counterparts and extensions to those arising in ordinary least squares models. Thus, for example, decisions are required concerning which error parameters should be included; whether there is a prior order in which they should be introduced; how one interprets the estimates; the use of residuals at different levels and so forth. There is also the general issue of how to deal with coefficients which may be treated either as fixed or random. It is to be hoped that extensive practical use of these models will provide the experience for forming sound judgements on these issues.

Further topics worth investigating include the issue of convergence, failure or convergence to a value which is not a global minimum.

When the boundary of the feasible region is reached during the iterative process, the procedure suggested does allow subsequent estimates to move away from the boundary. This is observed to occur in practice when the solution actually lies within the feasible region, but where the solution is on the boundary a local minimum might be found. In this case the solution could be explored by restarting the iterative process with parameter estimates close to other boundary points which could define local minima and observing the behaviour of the iterations. Further work on this issue would be useful.

With a local minimum we still obtain consistent point and covariance matrix estimates, although if several local minima were found it would be advisable to look carefully at the data. Simulation studies with small samples would be useful, and robust procedures need investigating in the estimation of both β and β^*. A computer program

for a 3-level model has been written with the ability to handle random coefficients, and this is being extended to handle constraints and errors of measurement at any level. A study of the efficiency of the procedure and the sampling properties of the estimators under alternative error distributions also needs study. One important area is the case where the individual level errors are multinomial with variances and covariances which are known functions of β and the methods of § 5 can be used. The error terms at higher levels may be continuously, for example normally, distributed, or have multinomial distributions. In general we can have models with mixtures of error distributions at all levels.

ACKNOWLEDGEMENTS

I am grateful to the Inner London Education Authority for permission to use the data in the example. Comments on drafts from the following were very helpful: Murray Aitkin, Leigh Burstein, David Cox, Russell Ecob, Kate Foot, Nick Longford, Les McLean, Ian Plewis, Ross Traub, Richard Wolfe and Bob Wood. I am especially grateful to Nick Longford whose own work, based on maximum likelihood, has clarified my thinking.

APPENDIX 1

Equivalence of maximum likelihood and iterative generalized least squares estimates

The log likelihood function for the multivariate normal model is, apart from a constant,

$$\log L = -\operatorname{tr}(V^{-1}S) - \log|V|,$$

where $S = (Y - X\beta)(Y - X\beta)^{\mathrm{T}}$. The estimation equations for β^* are

$$\frac{\partial L}{\partial \beta^*} = -\frac{\partial}{\partial \beta^*}\operatorname{tr}(V^{-1}S) + \operatorname{tr}\left(V\frac{\partial V^{-1}}{\partial \beta^*}\right) = 0.$$

Now equation (9) is obtained by minimizing

$$G = \operatorname{vech}(S - V)^{\mathrm{T}}(V^*)^{-1}\operatorname{vech}(S - V).$$

When the error terms are normally distributed, this is equivalent to minimizing

$$G = \operatorname{vec}(S - V)^{\mathrm{T}}(\Omega^{-1} \otimes \Omega^{-1})\operatorname{vec}(S - V) = \operatorname{tr}\{\Omega^{-1}(S - V)\}^2,$$

where Ω is the true unknown covariance matrix with

$$\operatorname{cov}(\hat{\beta}^*) = (X^{**\mathrm{T}}\Omega^{-1} \otimes \Omega^{-1}X^{**})^{-1},$$

where X^{**} corresponds to X^* but based on $\operatorname{vec}(V)$ rather than $\operatorname{vech}(V)$ (Browne, 1974).

Thus estimates of β^* are given by

$$\frac{\partial G}{\partial \beta^*} = 2\operatorname{tr}\left\{\Omega^{-1}\frac{\partial V}{\partial \beta^*}\Omega^{-1}S - \Omega^{-1}V\Omega^{-1}\frac{\partial V}{\partial \beta^*}\right\} = 0.$$

The iterative generalized least squares procedure sets $\Omega = V$, so we have

$$-\operatorname{tr}\left\{\frac{\partial}{\partial \beta^*}(V^{-1}S)\right\} + \operatorname{tr}\left(V\frac{\partial V^{-1}}{\partial \beta^*}\right) = 0,$$

which is the same as the maximum likelihood estimation equations above. The equivalence of the estimation equations for β follows immediately from the fact that both involve the minimization of $(Y-X\beta)^{\mathrm{T}} V^{-1}(Y-X\beta)$. Thus, the iterative generalized least squares procedure is one method for obtaining maximum likelihood estimates. For distributions other than the normal the two sets of estimates will not be identical, although the iterative generalised least squares will still be consistent. For a further discussion in the context of covariance structure models, see Bentler (1983).

APPENDIX 2

Calculation of the inverse of the error covariance matrix and the error variances and covariances

We use the result

$$(A+BCB^{\mathrm{T}})^{-1} = A^{-1} - A^{-1}BC(I+B^{\mathrm{T}}A^{-1}BC)^{-1}B^{\mathrm{T}}A^{-1}. \tag{A1}$$

Consider first the simple two-level model where the block diagonal error matrix is

$$V_2 = \bigoplus_{i=1}^{m} \{\sigma^2 I_{(n_i)} + \sigma_u^2 J_{(n_i)}\} = \bigoplus_{i=1}^{m} \{\sigma^2 I_{(n_i)} + \sigma_u^2 J_{(n_i \times 1)} I_{(1)} J_{(1 \times n_i)}\},$$

where $\oplus$ is the direct sum operator. Using (A1) we obtain

$$V_2^{-1} = \bigoplus_{i=1}^{m} \sigma^{-2}\{I_{(n_i)} - \sigma_u^2(n_i\sigma_u^2+\sigma^2)^{-1} J_{(n_i)}\}.$$

For the simple three-level model (6), the typical block of the covariance matrix can be written as

$$V_{3,k} = V_{2,k} + \sigma_v^2 J_{(n_{ki} \times 1)} I_{(1)} J_{(1 \times n_{ki})},$$

and using (A1) we have

$$V_3^{-1} = \bigoplus_{k} [V_{2,k}^{-1} - V_{2,k}^{-1} J_{(n_k \times 1)}\{\sigma_v^{-2} + J_{(1 \times n_k)} V_{2,k}^{-1} J_{(n_k \times 1)}\}^{-1} J_{(1 \times n_k)} V_{2,k}^{-1}],$$

where the scalar

$$J_{(1 \times n_k)} V_{2,k}^{-1} J_{(n_k \times 1)} = \sigma^{-2}\left\{n_k - \sum_{i=1}^{m} n_i^2 \sigma_u^2 (n_i\sigma_u^2+\sigma^2)^{-1}\right\}.$$

This result is equivalent to that derived, using a direct evaluation of $V_{3,k}^{-1}$, by Searle (1970). The extension to higher level models is obvious.

In the case of random coefficients, again consider first the two-level model. The first-level error terms contribute elements to the covariance matrix for the ith block,

$$V_{1,i} = \bigoplus_{j} \{(X_{ij}, Z_{ij})\,\Omega_1\,(X_{ij}, Z_{ij})^{\mathrm{T}}\}, \tag{A2}$$

where, for the r first-level coefficients which are random variables, including the overall coefficient α_0, X_{ij} is the $1 \times p_x$ vector $(x_{p_1,ij}, \ldots x_{p_x,ij}) = \{x_{p,ij}\}$, and for the second-level coefficients Z_{ij} is the $1 \times p_z$ vector $\{z_{p,ij}\}$, where by definition $z_{p,ij} = z_{p,ij'}$ and $\Omega_1 = \{\sigma_{1,rs}\}$, the covariance matrix of order (p_x+p_z) of the first-level errors.

The set of second-level error terms contribute elements $(X_i, Z_i)\,\Omega_2\,(X_i, Z_i)^{\mathrm{T}}$, where, with the same notation, $X_i = \{X_{ij}\}$, $Z_i = \{Z_{ij}\}$ and Ω_2 is the covariance matrix of the second-level errors.

Thus

$$V_2 = \bigoplus_{i} \{V_{1,i} + (X_i, Z_i)\,\Omega_2\,(X_i, Z_i)^{\mathrm{T}}\}. \tag{A3}$$

Using a similar notation we have, for the three-level model,

$$V_3 = \bigoplus_k \{V_{2,k} + (X_k, Z_k, W_k)\,\Omega_3\,(X_k, Z_k, W_k)^{\mathrm{T}}\}, \tag{A4}$$

where $X_k = \{X_{kij}\}$, and so on, and Ω_3 is the covariance matrix, of order $(p_x + p_z + p_w)$, of the third-level errors. Note, that in general, the error terms for each level will be associated with different sets of explanatory variables. It should also be noted that, while the error terms at any one level are assumed to be independent of those at any other level, within a level, errors from explanatory variables defined at different levels can be dependent. These results apply also to the general error structure given by (7), which can be estimated using a suitable modification to the matrix X^*.

Thus V_3^{-1} is readily evaluated via (A1)–(A4) and the extension to four or more levels is straightforward. The largest matrices requiring inversion by numerical methods are of order equal to the number of random coefficients at each level, and procedures can be programmed for the economical use of core store.

We have, as in Appendix 1, $(V \otimes V)^{-1} = V^{-1} \otimes V^{-1}$, which, for the three-level model, is block diagonal with blocks $V_k^{-1} \otimes V_k^{-1}$.

The estimates $\hat{\beta}, \hat{\beta}^*$ can be calculated at each iteration using the following identity for a block diagonal matrix with blocks indexed by t,

$$A^{\mathrm{T}} V^{-1} B = \Sigma A_t^{\mathrm{T}} V_t^{-1} B_t.$$

Also, we have $X^{**} = \mathrm{vec}\,(\partial V/\partial \beta^*)$, so that the typical component of X_k^{**} can be written as the sum of terms of the form $\mathrm{vec}\,(x_r x_r^{\mathrm{T}})$ or $\mathrm{vec}\,(x_r x_s^{\mathrm{T}}) + \mathrm{vec}\,(x_s x_r^{\mathrm{T}})$ for variance or covariance terms respectively; and we can also write $Y_t^{**} = \mathrm{vec}\,\{(Y_t - \hat{Y}_t)(Y_t - \hat{Y}_t)^{\mathrm{T}}\}$.

Now, for vectors a_i, b_i, c_i,

$$\{\mathrm{vec}\,(a_1 a_2^{\mathrm{T}})\}^{\mathrm{T}}(\Omega_0^{-1} \otimes \Omega_0^{-1}) = \{\mathrm{vec}\,(\Omega_0^{-1} a_1 a_2^{\mathrm{T}} \Omega_0^{-1})\}^{\mathrm{T}} = \{\mathrm{vec}\,(b_1 b_2^{\mathrm{T}})\}^{\mathrm{T}},$$

say, and $\{\mathrm{vec}\,(b_1 b_2^{\mathrm{T}})\}^{\mathrm{T}} \mathrm{vec}\,(c_1 c_2^{\mathrm{T}}) = (c_2^{\mathrm{T}} b_2)(b_1^{\mathrm{T}} c_1)$, so that the estimates $\hat{\beta}^*$ are easily calculated.

Appendix 3

Derivation of estimators when there are errors of measurement in explanatory variables

For the model in §7 we have

$$X^{\mathrm{T}} V^{-1} X = x^{\mathrm{T}} V^{-1} x + u^{\mathrm{T}} V^{-1} x + x^{\mathrm{T}} V^{-1} u + u^{\mathrm{T}} V^{-1} u,$$

$$\mathrm{plim}\,(X^{\mathrm{T}} V^{-1} X) = x^{\mathrm{T}} V^{-1} x + \mathrm{plim}\,(u^{\mathrm{T}} V^{-1} u).$$

For errors of measurement for child level variables, the (j,k)th term of $\mathrm{plim}\,(u^{\mathrm{T}} V^{-1} u)$ is $\Sigma_i \sigma^{ii}\,\mathrm{cov}\,(u_j, u_k)$ with summation over individuals, where σ^{ii} is the ith diagonal term of V^{-1}, and $\mathrm{cov}\,(u_j, u_k)$ is the covariance of the jth and kth measurement errors. The covariance is zero except for $j = k$.

For errors of measurement at the class level, σ^{ii} is replaced by the sum of the elements of V^{-1} corresponding to the ith class, and, for errors of measurement at the school level, it is replaced by the sum of the elements corresponding to the ith school. The definition of T_v below is modified accordingly. For child level errors of measurement we have

$$\mathrm{plim}\,(X^{\mathrm{T}} V^{-1} X) = x^{\mathrm{T}} V^{-1} x + \mathrm{tr}\,(V^{-1})\,\Omega_{uu}.$$

Let $S_{uu} = n\hat{\Omega}_{uu}$, with diagonal elements S_{ui}^2 which are estimated independently of the other error terms.

Writing $\mathrm{tr}\,(V^{-1})/n = T_v$, we obtain a consistent estimator of $M_{xx} = x^{\mathrm{T}} V^{-1} x$,

$$\hat{M}_{xx} = X^{\mathrm{T}} V^{-1} X - T_v S_{uu}.$$

We also have a consistent estimator of M_{xy}, $\hat{M}_{xy} = M_{XY} = X^{\mathrm{T}} V^{-1} Y$. Thus we obtain a consistent estimator of β,

$$\hat{\beta} = \hat{M}_{xx}^{-1} \hat{M}_{xy} = \beta + \hat{M}_{xx}^{-1} \{X^{\mathrm{T}} V^{-1} v + T_v S_{uu} \beta\}. \tag{A5}$$

Using a similar derivation to that given by Warren et al. (1974) we obtain an estimate of cov $(\hat{\beta})$ given by

$$\hat{M}_{xx}^{-1} \{X^{\mathrm{T}} \hat{V}^{-1} X + X^{\mathrm{T}} \hat{V}^{-2} X(\sigma_e^2 + \hat{T}_u) + n^{-1} \hat{T}_v^2 S_{uu} \hat{\beta}\hat{\beta}^{\mathrm{T}} S_{uu} + 2n^{-1} \hat{R}\hat{T}_v^2\} \hat{M}_{xx}^{-1}, \tag{A6}$$

where

$$T_u = n^{-1} \Sigma_j \beta_j^2 S_{uj}^2, \quad \hat{R} = n \operatorname{diag}(d_i^{-1} \hat{\beta}_i^2 S_{ui}^4)$$

and d_i are the degrees of freedom used in calculating S_{ui}^2.

At each iteration we require an estimate of V. We note that

$$\operatorname{cov}(vv^{\mathrm{T}}) = V + I_{(n)}(\sigma_e^2 + T_u)$$

so that the quantity $(\sigma_e^2 + T_u) I_{(n)}$ should be subtracted from the covariance matrix estimate calculated at each iteration, and (A5) is used then to estimate the coefficients, and (A6) their covariance matrix. Typically in practice sample estimates of S_{uj}^2 will be obtained from special studies of measurement errors.

REFERENCES

AITKIN, M., ANDERSON, D. & HINDE, J. (1981). Statistical modelling of data on teaching styles (with discussion). *J. R. Statist. Soc.* A **144**, 419–61.

BENTLER, P. (1983). Some contributions to efficient statistics in structural models: Specification and estimation of moment structures. *Psychometrika* **48**, 493–518.

BROWNE, M. W. (1974), Generalized least squares estimators in the analysis of covariance structures. *S. Afr. Statist. J.* **8**, 1–24.

FISK, P. R. (1967). Models of the second kind in regression analysis. *J. R. Statist. Soc.* B **29**, 266 81.

FULLER, W. A. & BATTESE, G. E. (1973). Transformations for estimation of linear models with nested error structure. *J. Am. Statist. Assoc.* **68**, 626–32.

FULLER, W. A. & HIDIROGLOU, M. A. (1978). Regression estimation after correcting for attenuation. *J. Am. Statist. Assoc.* **73**, 99–104.

GOLDSTEIN, H. (1979). *The Design and Analysis of Longitudinal Studies*. London: Academic Press.

HARVILLE, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Statist. Assoc.* **72**, 320 37.

HIDIROGLOU, M. A. (1981). Computerization of complex survey estimates. In *Proc. Statist. Comp. Sect., Am. Statist. Assoc.*, pp. 1 7.

HOLT, D., SMITH, T. M. F. & WINTER, P. D. (1980). Regression analysis of data from complex surveys. *J. R. Statist. Soc.* A **142**, 474 87.

INNER LONDON EDUCATION AUTHORITY (1969). *Literacy Survey: Summary of Interim Results*. London: I.L.E.A. Research & Statistics Division.

JOBSON, J. D. & FULLER, W. A. (1980). Least squares estimation when the covariance matrix and parameter vector are functionally related. *J. Am. Statist. Assoc.* **75**, 176–81.

JONES, R. G. (1980). Best linear unbiased estimation in repeated surveys. *J. R. Statist. Soc.* B **42**. 221 6.

JORESKOG, K. & SORBOM, D. (1979). *Advances in Factor Analysis and Structural Equation Models*. Cambridge, Massachusetts: Abt Books.

MASON, W. M., WONG, G. Y. & ENTWISTLE, B. (1984). The multilevel linear model: A better way to do contextual analysis. In *Sociological Methodology*. Ed. K. F. Schuessler. pp. 72 103. London: Jossey Bass.

RAO, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika* **52**, 447–58.

SEARLE, S. R. (1970). Large sample variances of maximum likelihood estimators of variance components using unbalanced data. *Biometrics* **26**, 505–24.

WARREN, R. D., WHITE, J. K. & FULLER, W. A. (1974). An errors in variables analysis of managerial role performance. *J. Am. Statist. Assoc.* **69**, 886–93.

[*Received July* 1984. *Revised March* 1985]