LINEAR MULTILEVEL MODELS

JAN DE LEEUW

ABSTRACT. This is an entry for The Encyclopedia of Statistics in Behavioral Science, to be published by Wiley in 2005.

1. HIERARCHICAL DATA

Data are often hierarchical. By this we mean that data contain information about observation units of various levels, where the lower-level units are nested within the higher-level units. Some examples may clarity this. In repeated measurement or growth curve data we have several observations in time on each of a number of different individuals. The time-points are the lowest level and individuals are the higher level. In school effectiveness studies we have observations on students (the lowest or first level), on the schools in which these students are enrolled (the second level), and maybe even on school districts these schools are in (a third level).

Once we have data on various levels, we have to decide at which level to analyze. We can aggregate student variables to the school level or disaggregate school variables to the student level. In the first case we loose potentially

Date: July 19, 2004.

large amounts of useful information, because information about individual students disappears from the analysis. In the second case we artificially create dependencies in the data, because students in the same school by definition get the same score on a disaggregated school variable.

Another alternative is to do a separate analysis for each higher-order unit separately. For example, we do a student-level regression analysis for each school separately. This, however, tends to introduce a very large number of parameters. It also ignores the fact that it makes sense to assume the different analyses will be related, because all schools are functioning within the same education system.

Multilevel models combine information about variables of different levels in a single model, without aggregating or disaggregating. It provides more data reduction than a separate analysis for each higher-order unit, and it models the dependency between lower-order units in a natural way. Multilevel models originated in school effectiveness research, and the main textbooks discussing this class of techniques still have a strong emphasis on educational applications [4, 13]. But hierarchical data occur in many disciplines, so there are now applications in the health sciences, in biology, in sociology, and in econometrics.

2. Linear Multilevel Model

A linear multilevel model, in the two-level case, is a regression model specified in two stages, corresponding with the two levels. We start with separate linear regression models for each higher order unit, specified as

$$\underline{y}_j = X_j \underline{\beta}_j + \underline{\epsilon}_j. \tag{1a}$$

Higher order units (schools) are indexed by j, there are m of them. Unit j contains n_j observations (students), and thus the outcomes $\underline{y}_j$ and the error terms $\underline{\epsilon}_j$ are vectors with n_j elements. The predictors for unit j are collected in an $n_j \times p$ matrix X_j.

In our model specification random variables, and random vectors, are underlined. This shows clearly how our model differs from classical separate linear regression models, in which the regression coefficients β_j are non-random. This means, in the standard frequentist interpretation, that if we were to replicate our experiment then in the classical case all replications have the same regression coefficients, while in our model the random regression coefficients would vary because they would be independent realizations of the same random vector $\underline{\beta}_j$. The difference are even more pronounced, because we also use a second level regression model, which has the first order regression coefficients as outcomes. This uses a second

set of q regressors, at the second level. The sub-model is

$$\underline{\beta}_j = Z_j\gamma + \underline{\delta}_j, \tag{1b}$$

where the Z_j are now $p \times q$ and γ is a fixed set of q regression coefficients that all second order units have in common.

In our regression model we have not underlined the predictors in X_j and Z_j, which means we think of them as fixed values. They are either fixed by design, which is quite uncommon in social and behavioral sciences, or they are fixed by the somewhat artificial device of conditioning on the values of the predictors. In the last case the predictors are really random variables, but we are only interested in what happens if these variables are set to their observed values.

We can combine the specifications in (1a) and (1b) to obtain the linear mixed model

$$\underline{y}_j = X_j Z_j \gamma + X_j \underline{\delta}_j + \underline{\epsilon}_j. \tag{2}$$

This model has both fixed regression coefficients γ and random regression coefficients δ_j. In most applications we suppose that both regressions have an intercept, which means that all X_j and all Z_j have a column with elements equal to one.

For the error terms $\underline{\delta}_j$ and $\underline{\delta}_j$ in the two parts of the regression model we make the usual strong assumptions. Both have expectation zero, they are uncorrelated with each other within the same second-level unit, they are uncorrelated between different second-level units. We also assume that first-level disturbances are homoscedastic, and that both errors have the same variance-covariance matrix in all second-level units. Thus $V(\underline{\epsilon}_j) = \sigma^2 I$ and we write $V(\underline{\delta}_j) = \Omega$. This implies that

$$E(\underline{y}_j) = X_j Z_j \gamma, \tag{3a}$$

$$V(\underline{y}_j) = X_j \Omega X_j' + \sigma^2 I. \tag{3b}$$

Thus we can also understand mixed linear models as heteroscedastic regression models with a specific interactive structure for the expectations and a special factor analysis structure for the covariance matrix of the disturbances. Observations in the same two-level unit are correlated, and thus we have correlations between students in the same school and between observations within the same individual at different time-points. We also see the correlation is related to the similarity in first-order predictor values of units i and k. Students with similar predictor values in X_j will have a higher correlation.

To explain more precisely how the $n_j \times q$ design matrices $U_j = X_j Z_j$ for the fixed effects usually look, we assume that Z_j has the form

$$Z_j = \begin{bmatrix} h_j' & 0 & \cdots & 0 \\ 0 & h_j' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_j' \end{bmatrix},$$

where h_j is a vector with r predictor values for second-level unit j. Thus $q = pr$, and $\underline{\beta}_{js} = h_j'\gamma_s + \underline{\delta}_{js}$. With this choice of Z_j, which is the usual one in multilevel models, the matrix U_j is of the form

$$U_j = \begin{bmatrix} x_{j1}h_j & | & \cdots & | & x_{jp}h_j' \end{bmatrix},$$

i.e. each columns U is the product of a first-level predictor from X and a second-level predictor from H. All $p \times r$ cross-level interactions get their own column in U. If the X_j have their first column equal to one, and the h_j have their first element equal to one, then it follows that the columns of X_j themselves and the (disaggregated) columns of H are among the pq interactions.

In the balanced case of the multilevel model all second-level units j have the same $n \times p$ matrix of predictors X. This happens, for example, in growth curve models, in which X contains the same fixed functions (orthogonal polynomials, for instance) of time. In the balanced case we can collect our

various observations and parameters in matrices, and write $\underline{Y} = \underline{B}X' + \underline{E}$ and $\underline{B} = Z\Gamma + \underline{\Delta}$ or $\underline{Y} = Z\Gamma X' + \underline{\Delta}X' + \underline{E}$. Here the outcomes $\underline{Y}$ are in a matrix of order $m \times n$, individuals by time-points, and the fixed parameters are in a $q \times p$ matrix Γ. This shows, following Strenio et al. [17], how multilevel ideas can be used to generalize the basic growth curve model of Pothoff and Roy [12].

3. Parameter Constraints

If p and q, the number of predictors at both levels, are at all large, then obviously their product pq will be very large. Thus we will have a linear model with a very large number of regression coefficients, and in addition to the usual residual variance parameter σ^2 we will also have to estimate the $\frac{1}{2}p(p+1)$ parameters in Ω. The problem of having too many parameters for fast and stable estimation is compounded by the fact that the interactions in U will generally be highly correlated, and that consequently the regression problem is ill-conditioned. This is illustrated forcefully by the relatively small examples in Kreft and de Leeuw [7].

The common procedure in multilevel analysis to deal with parameter glut is the same as in other forms of regression analysis. Free parameters are set equal to zero, or, equivalently, we use variable selection procedures. Setting regression coefficients (values of γ) equal to zero is straightforward,

because it simply means that cross-level interactions are eliminated from the model. Nevertheless, the usual variable selection problem applies, if we have pq variables to include or exclude, we can make 2^{pq} possible model choices and for large pq there is no optimal way to make such a choice. [7] argue forcefully that either multilevel modelling should be limited to situations with a small numbers of variables or it should only be applied in areas in which there is sufficient scientific theory on the basis of which to choose predictors.

Another aspect of variable selection is that we can set some of the random coefficients in $\underline{\delta}_j$ to zero. Thus he corresponding predictor in X only has a fixed effect, not a random effect. This means that particular row and column of Ω corresponding with that predictor are set to zero. It is frequently useful to use this strategy in a rather extreme way and set the random parts of all regression coefficients, except the intercept, equal to zero. This leads to random intercept models, which have far fewer parameters and are much better conditioned. They are treated in detail in Longford [10].

If we set parts on Ω to zero, we must be careful. In Kreft et al. [6] it is shown that requiring Ω to be diagonal, for instance, destroys the invariance of the results under centering of the variables. Thus, in a model of this form, we need meaningful zero points for the variables, and meaningful zero points

are quite rare in social and behavioral applications. (crossref to centering entry)

4. GENERALIZATIONS

The linear multilevel model can be, and has been, generalized in many different directions. It is based on many highly restrictive assumptions, and by relaxing some or all of these assumptions we get various generalizations.

First, we can relax the interaction structure of $U_j = X_j Z_j$ and look at the multilevel model for general $n_j \times q$ design matrices U_j. Thus we consider more general models in which some of the predictors have fixed coefficients and some of the predictors have random coefficients. We can write such models, in the two-level case, simply as

$$\underline{y}_j = U_j \beta_j + X_j \underline{\delta}_j + \underline{\epsilon}_j.$$

It is possible, in fact, that there is overlap in the two sets of predictors U_j and X_j, which means that regressions coefficients have both a fixed part and a random part. Second, we can relax the homoscedasticity assumptions $V(\underline{\epsilon}_j) = \sigma^2 I$. We can introduce σ_j^2, so that the level of error variance is different for different second-level units. Or we can allow for more general parametric error structures $V(\underline{\epsilon}_j) = \Sigma_j(\theta)$, for example by allowing autocorrelation between errors at different time point with the same individual.

Third, it is comparatively straightforward to generalize the model to more than two levels. The notation can become somewhat tedious, but when all the necessary substitutions have been made we still have a linear mixed model with nested random effects, and the estimation and data analysis proceed in the same way as in the two-level model.

Fourth, the device of modelling parameter vectors as random is strongly reminiscent of the Bayesian approach to statistics. The main difference is that in our approach to multilevel analysis we still have the fixed parameters γ, σ^2 and Ω that must be estimated. In a fully Bayesian approach one would replace these fixed parameters by random variables with some prior distribution and one can then compute the posterior distribution of the parameters vectors, which are now all random effects. The Bayesian approach to multilevel modeling (or hierarchical linear modeling) has been explored in many recent publications, especially since the powerful Markov Chain Monte Carlo tools became available.

Fifth, we can drop the assumption of linearity and consider nonlinear multilevel models or generalized linear multilevel models. Both are discussed in detail in the basic treatises of Raudenbush and Bryk [13] and Goldstein [4], but discussing them here would take us to far astray. The same is true for models with multivariate outcomes, in which the elements of the vector y_j

are themselves vectors, or even matrices. A recent application of multilevel models in this context is analysis of fMRI data [2].

And finally, we can move multilevel analysis from the regression context to the more general framework of latent variable modeling. This leads to multilevel factor analysis and to various multilevel structural equation models. A very complete treatment of current research in that field is in Skrondal and Rabe-Hesketh [16].

5. ESTIMATION

There is a voluminous literature on estimating multilevel models, or, more generally, mixed linear models [15]. Most methods are based on assuming normality of the random effects and then using maximum likelihood estimation. The likelihood function depends on the regression coefficients for the fixed variables and the variances and covariances of the random effects. Itis easily minimized, for instance, by alternating minimization over γ for fixed σ^2 and Ω, and then minimization over σ^2 and Ω for fixed γ, until convergence. This is sometimes known as IGLS, or Iterative Generalized Least Squares [3]. It is also possible to treat the random coefficients as missing data and apply the EM algorithm [13], or to apply Newton's method or Fisher Scoring to optimize the likelihood [8, 9].

There is more than one likelihood function we can use. In the early work the likelihood of the observations was used. Thus is a function of σ^2, Ω and γ. The disadvantage of the FIML estimates obtained by maximizing this full information likelihood function is that variance components tend to be biased, in the same way, and for the same reason, why the maximum likelihood of the sample variance is biased. In the case of the sample variance, we correct for the bias of the estimate by maximizing the likelihood of the deviations of the sample mean. In the same way, we can study the likelihood of a set of linear combinations of the observations, where the coefficients of the linear combinations are chosen orthogonal to the X_j. This means that γ disappears from the residual or reduced likelihood, which is now only a function of the variance and covariance components. The resulting REML estimates, originally due to Patterson and Thomson [11], can be computed with small variations of the more classical maximum likelihood algorithms (IGLS, EM, Scoring), because the two types of likelihood functions are closely related.

Of course REML does not give an estimate of the fixed regression coefficients, because the residual likelihood does not depend on γ. This problem is resolved by estimating γ by generalized least squares, using the REML

estimates of the variance components. Neither REML nor FIML gives estimates of the random regression coefficients or of the random effects. Random variables are not fixed parameters, and consequently they cannot be estimated in the classical sense. What we can estimate is the conditional expectation of the random effects given the data. These conditional expectations can be estimated by plug-in estimates using the REML or FIML estimates of the fixed parameters. They are also known as the *best linear unbiased predictors* or BLUP's [14].

There is a large number of software packages designed specifically for linear multilevel models, although most of them by now also incorporate the generalizations we have discussed in the previous section. The two most popular special purpose packages are HLM, used in Raudenbush and Bryk [13], and MLWin, used in Goldstein [4]. Many of the standard statistical packages, such as SAS, SPSS, Stata, and R now also have multilevel extensions written in their interpreted matrix languages.

6. School Effectiveness Example

We use school examination data previously analyzed with multilevel methods by Goldstein et al. [5]. Data are collected on 4,059 students in 65 schools in Inner London. For each student we have a normalized exam score (`normexam`) as the outcome variable. Student-level predictors are

gender (coded as a dummy `genderM`) and standardized London Reading Test score (`standlrt`). The single school-level predictors we use is school gender (mixed, boys, or girls school, abbreviated as `schgend`). This is a categorical variable, which we code using a boyschool-dummy and a girlschool-dummy.

Our first model is a simple random intercept model, with a single variance component. Only the intercept is random, all other regression coefficients are fixed. The model is

$$\underline{\texttt{normexam}}_{ij} = \underline{\alpha}_j + \texttt{standlrt}_{ij}\beta_1 + \texttt{gender}_{ij}\beta_2 + \underline{\epsilon}_{ij},$$

$$\underline{\alpha_j} = \texttt{schgendboys}_j\gamma_1 + \texttt{schgendgirls}_j\gamma_2 + \underline{\delta}_j$$

We compare this with the model without a random coefficient

$$\underline{\texttt{normexam}}_{ij} = \alpha_j + \texttt{standlrt}_{ij}\beta_1 + \texttt{gender}_{ij}\beta_2 + \underline{\epsilon}_{ij},$$

$$\alpha_j = \texttt{schgendboys}_j\gamma_1 + \texttt{schgendgirls}_j\gamma_2$$

REML estimation of both models gives the followign table of estimates, with standard errors in parentheses. This is a small example, but it illustrates some basic points. The intra-class correlation ρ is only 0.132 in this case, but the fact that it is nonzero has important consequences. We see that if the random coefficient model then the standard errors of the regression coefficient from the fixed model are far too small. In fact, in the fixed

source	Random Model	Fixed Model
intercept	-0.00 (0.056)	-0.03 (0.025)
standlrt	0.56 (0.012)	0.56 (0.017)
genderM	-0.17 (0.034)	-0.17 (0.034)
schgenboys	0.18 (0.113)	0.18 (0.043)
schgengirls	0.16 (0.089)	0.17 (0.033)
ω^2	0.086	–
σ^2	0.563	0.635
ρ	0.132	–

model the schoolvariables schgenboys and schgengirls are highly significant, while they are not even significant on the 5% level in the random model. We also see that the estimate of σ^2 is higher in the fixed model, which is not surprising because the random model allows for an additional parameter to model the variation. Another important point is that the actual values of the regression coefficients in the fixed and random model are very close. Again, this is not that surprising, because after all in REML the fixed coefficients are estimated with least squares methods as well.

7. Growth Curve Example

We illustrate repeated measure examples with a small dataset taken from the classical paper by Pothoff and Roy [12]. Distances between pituitary gland and pterygomaxillary fissure were measured using x-rays in $n = 27$ children (16 males and 11 females) at $m = 4$ time points, at ages 8, 10, 12, and 14. Data can be collected in a $n \times m$ matrix Y. We also use a $m \times p$ matrix X of the first $p = 2$ orthogonal polynomials on the m time-points.

The first class of models we consider is $\underline{Y} = BX' + \underline{E}$ with B a $n \times p$ matrix of regression coefficients, one for each subject, and with $\underline{E}$ the $n \times m$ matrix of disturbances. We suppose the rows of $\underline{E}$ are independent, identically distributed centered normal vectors, with dispersion Σ. Observe that the model here tells us the growth curves are straight lines, not that the deviations from the average growth curves are on a straight line.

Within this class of models we can specify various submodels. The most common one supposes that $\Sigma = \sigma^2 I$. Using the orthogonality of the polynomials in X, we find that in this case the regression coefficients are estimated simply by $\hat{B} = YX$. But many other specifications are possible. We can, on the one hand, require Σ to be a scalar, diagonal, or free matrix. And we can, on the other hand, require the regression coefficients to be all the same, the same for all boys and the same for all girls, or free (all

different). These are all fixed regression models. The minimum deviances (minus two times the maximized likelihood) are shown in the first three rows of Table 1. In some combinations there are too many parameters. As in other linear models this means the likelihood is unbounded above and the maximum likelihood estimate does not exist [1].

	B equal	B gender	B free
Σ scalar	307(3)	280(5)	91(55)
Σ diagonal	305(6)	279(8)	$-\infty$(58)
Σ free	233(12)	221(14)	$-\infty$(64)
random	240(6)	229(8)	$-\infty$(58)

TABLE 1. Mixed Model Fit

We show the results for the simplest case, with the regression coefficients "free" and the dispersion matrix "scalar". The estimated growth curves are in Figure 1. Boys are solid lines, girls are dashed. The estimated σ^2 is 0.85.

We also give the results for the "gender" regression coefficients and the "free" dispersion matrix. The two regression lines are in Figure 2. The regression line for boys is both higher and steeper than the one for girls. There is much less room in this model to incorporate the variation in the data using the regression coefficients, and thus we expect the estimate of the residual variance to be larger. In Table 2 we give the variances and

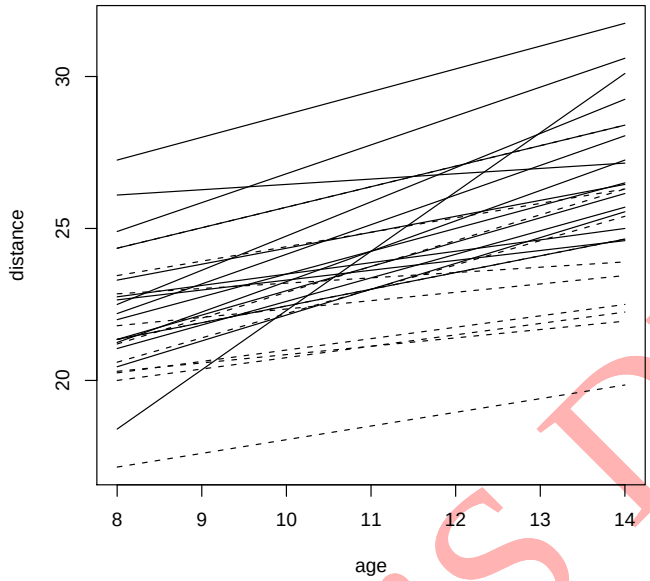

FIGURE 1. Growth Curves for the Free/Scalar Model

correlations from the estimated Σ. The estimated correlations between the errors are clearly substantial.

	8	10	12	14
Correlations	1.00			
	0.54	1.00		
	0.65	0.56	1.00	
	0.52	0.72	0.73	1.00
Variances	5.12	3.93	5.98	4.62

TABLE 2. Σ from Gender/Free Model

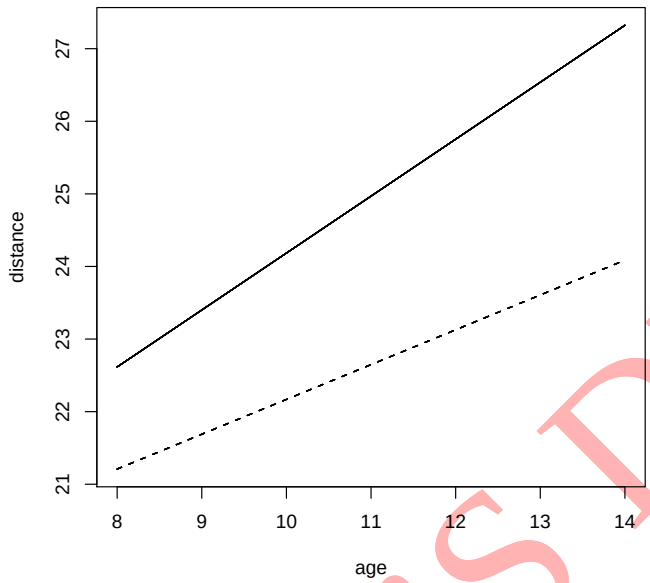

FIGURE 2. Growth Curves for the Gender/Free Model

The general problem with fixed effects models in this context is clear from both the figures and the tables. To make models realistic we need a lot of parameters, but if there are many parameters we cannot expect the estimates to be very good. In fact in some cases we have unbounded likelihoods and the estimates we look for do not even exist. Also, it is difficult to make sense of so many parameters at the same time, as Figure 1 shows.

Next consider random coefficient models of the form $\underline{Y} = \underline{B}X' + \underline{E}$, where the rows of $\underline{B}$ are uncorrelated with each other and with all of $\underline{E}$. By writing $\underline{B} = B + \underline{\Delta}$ with $B = \mathbf{E}(\underline{B})$ we see that we have a mixed linear model of

the form $\underline{Y} = BX' + \underline{\Delta}X' + \underline{E}$. Use Ω for the dispersion of the rows of $\underline{\Delta}$. It seems that we have made our problems actually worse by introducing more parameters. But allowing random variation in the regression coefficients makes the restrictive models for the fixed part more sensible. We fit the "equal" and "gender" versions for the regression coefficients B, together with the "scalar" version of Σ, leaving Ω "free".

Deviances for the random coefficient model are shown in the last row of Table 1. We see a good fit, with a relatively small number of parameters. To get growth curves for the individuals we compute the BLUP, or conditional expectation, $\mathbf{E}(\underline{B}|Y)$, which turns out to be

$$\mathbf{E}(\underline{B}|Y) = \tilde{B}[I - \Omega(\Omega + \sigma^2 I)^{-1}] + \hat{B}\Omega(\Omega + \sigma^2 I)^{-1},$$

where $\tilde{B}$ is the mixed model estimate and $\hat{B} = YX$ is the least squares estimate portrayed in Figure 1. Using the "gender" restriction on the regression coefficients the conditional expectations are plotted in Figure 3.

We see they provide a compromise solution, that shrinks the ordinary least squares estimates in the direction of the "gender" mixed model estimates. We more clearly see the variation of the growth curves for the two genders around the mean gender curve. The estimated σ^2 for this model is 1.72.

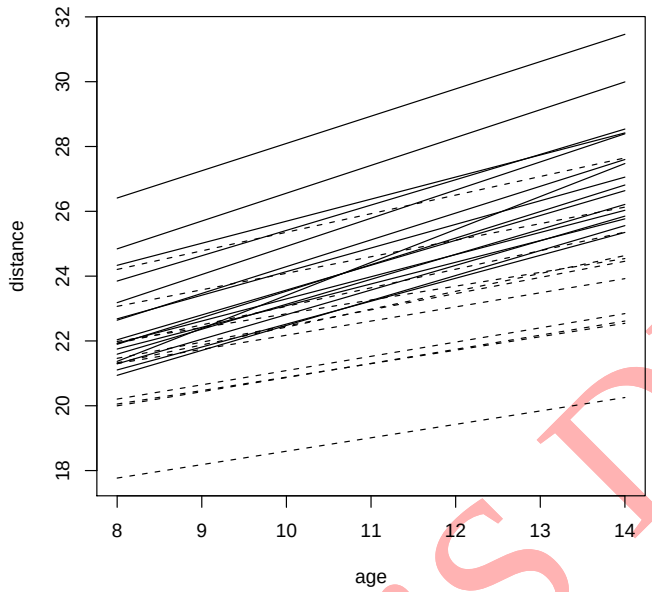

FIGURE 3. Growth Curves for the Mixed Gender Model

REFERENCES

[1] T.W. Anderson. Estimating Linear Statistical Relationships. *The Annals of Statistics*, 12(1):1–45, 1984.

[2] C.F. Beckmann, M. Jenkinson, and S.M. Smith. General Multilevel Linear Modeling for Group Analysis in FMRI. *NeuroImage*, 20:1052–1063, 2003.

[3] H. Goldstein. Multilevel Mixed Linear Model Analysis using Iterative Generalized Least Squares. *Biometrika*, 73:43–56, 1986.

[4] H. Goldstein. *Multilevel Statistical Models*. Arnold, third edition, 2003.

[5] H. Goldstein, J. Rasbash, M. Yang, G. Woodhouse, H. Panand D. Nuttall, and S. Thomas. A Multilevel Analysis of School Examination Results. *Oxford Review of Education*, 19:425–433, 1993.

[6] G. G. Kreft, J. De Leeuw, and L. S. Aiken. The Effects of Different Forms of Centering in Hierarchical Linear Models. *Multivariate Behavioral Research*, 30:1–21, 1995.

[7] G.G. Kreft and J. de Leeuw. *Introduction to Multilevel Modelling*. Sage Publications, Thousand Oaks, CA, 1998.

[8] J. De Leeuw and I.G.G. Kreft. Random Coefficient Models for Multilevel Analysis. *Journal of Educational Statistics*, 11:57–86, 1986.

[9] N.T. Longford. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74:817–827, 1987.

[10] N.T. Longford. *Random Coefficient Models*. Number 11 in Oxford Statistical Science Series. Oxford University Press, Oxford, 1993.

[11] H.D. Patterson and R. Thomson. Recovery of Inter-block Information when Block Sizes are Unequal. *Biometrika*, 58:545–554, 1971.

[12] R.F. Pothoff and S.N. Roy. A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems.

Biometrika, 51:313–326, 1964.

[13] S. W. Raudenbush and A.S. Bryk. *Hierarchical Linear Models. Applications and Data Analysis Methods*. Sage Publications, Newbury Park, CA, second edition, 2002.

[14] G.K. Robinson. That BLUP is a Good Thing: the Estimation of Random Effects (with Discussion). *Statistical Science*, 6:15–51, 1991.

[15] S.R. Searle, G. Casella, and C.E. McCulloch. *Variance Components*. Wiley, New York, NY, 1992.

[16] A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling. Multilevel, Longitudinal, and Structural Equation Models*. Interdisciplinary Statistics. Chapman & Hall, 2004.

[17] J.L.F. Strenio, H.I. Weisberg, and A.S. Bryk. Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates. *Biometrics*, 39:71–86, 1983.

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu