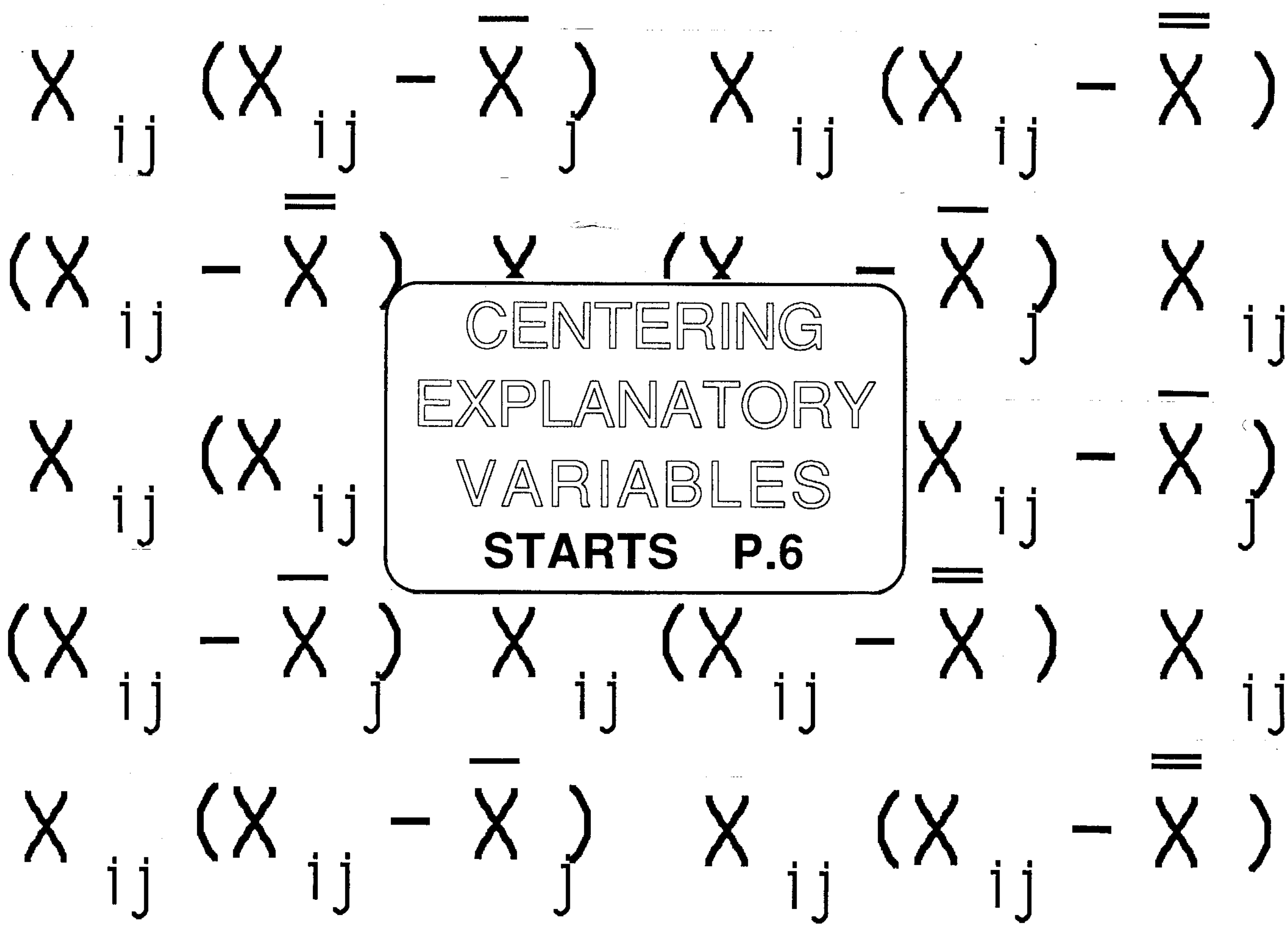
MULTILEVEL MODELLING NEWSLETTER

Produced through the Multilevel Models Project:

c/o Math, Statistics & Computing Department
Institute of Education, University of London,
20 Bedford Way, London WC1H 0AL
e-mail: rprosser%uk.ac.lon.educ.isis@ac.uk
Telephone: 01-636-1500 ext 473

Vol. 1 No. 3 October 1989



ALSO IN THIS ISSUE ...

SOFTWARE DEVELOPMENTS

ML3 V1.0 IS NOW BEING SHIPPED

Version 1.0 of the Multilevel Models Project's software for three-level analysis is now being sent to interested researchers. Its commands and operation are very similar to those of *ML2* (Rasbash, Prosser, & Goldstein, 1989). The program can of course be used to conduct any of the two-level analyses which are possible with *ML2*, so no further development of the latter is planned.

The graphing and utilities commands operate as in the two-level program, but several have expanded capabilities–the SUMMary, MLAVerages, MLSDs, MLCOunts, and RECSplit commands, in particular. Several new general features have been introduced. It is now possible, for example, to work with a set of variables (dummies, for instance) in a block, and to protect worksheet columns against unintended overwriting. You can check MS-DOS directories and view file contents from within *ML3*.

The modelling and output commands have natural extensions to cover the fitting of parameters at a third level. Two new modelling features have been added:

- a command for examining contrasts of random parameters; and
- commands for placing constraints on the estimation of the fixed and random parameters.

The RESIduals and fixed parameter contrasts commands have been changed slightly, and the computing of predicted values and OLS estimates of fixed parameters has been enhanced.

The program is written in C. (FORTRAN-77 was used for *ML2*.) This change will facilitate the addition of improvements planned for 1990. The immediate benefit to users, however, is that the program will now accommodate itself to available RAM (up to 640K); you no longer have to remove all your RAM-resident utilities to do a multilevel analysis. Of course, the more free bytes you have, the larger the worksheet you can use. For users with very large datasets, a VAX/ VMS implementation is being planned.

ML3 is being distributed at no cost to registered users of *ML2*, i.e., individuals who have paid a £50 (US$80) fee or organizations that have paid a negotiated amount. For the rest of 1989, it will be available for the same fee, and those who register now will be given preferential rates when a yet-to-be-finalized pricing policy comes into effect in the new year.

For further information, please contact Bob Prosser at the Project address on the front page.

Reference

Rasbash, J., Prosser, R., & Goldstein, H. (1989). *ML2 Software for Two-level Analysis: Users' Guide.* London: Institute of Education.

REPORT ON COMPARISON OF MULTILEVEL PROGRAMS READY SOON

Ita Kreft

The researchers on the project titled "Development of Computer Programs for Multilevel Analysis in Educational Evaluation" based at UCLA's Center for the Study of Evaluation are nearing completion of a report comparing four computer programs: VARCL, HLM, GENMOD, and *ML2*.

Chapter 1 of a draft report discusses the theory of multilevel modelling. The second chapter provides a comprehensive description of each program:

- the computing algorithm used;
- the setup required for data;
- treatment of missing data;
- special options;
- output files;
- estimation of reliability and related statistics;
- user friendliness;
- unique functions and features of the program;
- program limitations; and
- datasets used for illustration in the program's manual.

In chapter 3, results of timing comparisons are reported. Three simple models and three data sets provided the basis for examining the agreement and convergence properties of the software. A description of desirable extensions to existing software capabilities is provided in the report's final chapter.

Please direct inquiries about the report to Ita Kreft, UCLA, Moore Hall 145D, 405 Hilgard Avenue, Los Angeles, CA 90024-1522. Telephone no.: (213) 206 1528.

BOOK REVIEW

Multilevel Analysis of Educational Data. Edited by D. Bock. (1989). Academic Press. *£30.* Hardcover. 354 pages.

Michael Healy

This book contains the proceedings of a conference held in America in April 1987. It contains twelve papers along with four discussion contributions, plus a late paper by the editor himself.

The title is a little misleading. Some of the papers concentrate on educational applications, but others relate to other fields of application (especially growth problems) or are quite general in their coverage, and it would be a great pity if the wider statistical world were not to be aware of their importance. An area which has been left on one side is econometrics; even two years later, the contacts between multilevel research workers in econometrics and in the social sciences and other fields are far too scanty.

The titles of the papers are:

- Some Applications of Multilevel Models to Educational Data *Donald Rubin*
- Empirical Bayes Methods: A Tool for Exploratory Analysis *Henry Braun*
- A Hierarchical Item-response Model for Educational Testing *Robert Mislevy & Darrell Bock*
- Difficulties with Bayesian Inference for Random Effects *Charles Lewis*
- Multilevel Aspects of Varying Parameters in Structural Models *Bengt Muthén & Albert Satorra*
- Models for Multilevel Response Variables with an Application to Growth Curves *Harvey Goldstein*
- Multilevel Models: Issues and Problems Emerging from their Recent Application in British Studies of School Effectiveness *John Gray*
- Toward a More Appropriate Conceptualization of Research on School Effects: A Three-level Hierarchical Linear Model *Tony Bryk & Steve Raudenbush*
- Quantitative Models for Estimating Teacher and School Effectiveness *Steve Raudenbush & Tony Bryk*
- Multilevel Investigations of Systematically Varying Slopes: Issues, Alternatives and Consequences *Leigh Burstein, Kyung-Sung Kim, & Ginette Delandshere*
- Profile Predictive Likelihood for Random Effects in the Two-level Model *Murray Aitkin*
- Fisher Scoring Algorithm for Variance Component Analysis of Data with Multilevel Structure *Nick Longford*
- Measurement of Human Variation: A Two-stage Model *Darrell Bock*

As to the usefulness of the book, there is primarily the bringing together of so much excellent summarising material on multilevel methodology–I note particularly the papers by Bock, Goldstein and Longford in this regard. Aside from this, I found most interesting the contributions which stressed and tried to investigate some of the often neglected difficulties of multilevel methodology. Many of the arguments in the field concern themselves with different ways of maximising likelihood; fewer concern themselves with the questions of just what should be maximised and of what the rest of the likelihood surface is doing. I am thinking particularly of the paper by Lewis which draws heavily on work by O'Hagan and shows that the problem of estimating individual random effects, even in the simplest case, is liable to be much harder than it looks; and a pithy discussion contribution by McCullagh called *What can go wrong with iteratively re-weighted least squares?* which introduces tensor notation in a way that should prove useful. All that I felt to be missing was a summary of the summaries. It is clear that multilevel modelling is the next great unifying step after the generalised linear modelling of Nelder and Wedderburn, bringing together a whole set of techniques carrying labels like 'empirical Bayes' or 'random coefficient regression'. It would save a lot of unnecessary mental effort if the parallels could be properly systematised.

Nobody active either in the development or in the use of multilevel methods will want to be without this volume. One grouse; in spite of being prepared ready for the typesetter under the editor's supervision (using LaTeX) the book has taken two years to appear. The resulting pages are at least as good as professional typesetting, maybe better (I have only spotted one misprint), and it seems sad that, in this fast-moving area, the publishers could not achieve greater promptness.

APPLICATIONS &

MULTILEVEL MODELS USEFUL FOR EXAMINING INTERVIEWER EFFECTS IN SURVEYS

Dick Wiggins, Colm O'Muircheartaigh, & Nick Longford

The investigation of the effect of interviewers on survey responses has traditionally concentrated on univariate analysis–see for example, Kish (1962) and Fellegi (1964, 1974). The total variance of an estimator of the population mean is typically partitioned into two components, the sampling variance and the interviewer variance, and the impact of interviewers on the precision of the estimator is described using the intra-interviewer correlation coefficient. Such analysis does not address the issue of how the interviewer variance affects the analysis of relationships between variables.

In a forthcoming paper, to be published as part of a methodological series emanating from the London School of Economics/ Social and Community Planning Research centre for survey teaching and research, we present analyses based on variance components models which investigate the extent to which interviewer effects can be incorporated into substantive data analysis. The data come from two epidemiological studies in which the allocation of respondents to interviewers was randomized and in which the primary objective was the construction and evaluation of linear models of relationships between variables. See O'Muircheartaigh and Wiggins (1981) and Wiggins (1985). VARCL (Longford, 1989) was used for these analyses.

In the first application, modelling is demonstrated for a quantitative response variable. The basic analysis produces results analogous to analysis of variance and confirms the presence of a substantial interviewer effect. However, including interviewers as a nesting level in the data structure does not seriously contaminate the interpretation of the fixed part of the model. The analysis is extended to terms in the fixed part of the model to vary. In particular, there is a strong indication that the impact of the interviewer on the responses varies systematically with the age of the respondent. So much so that if the interviewing were confined to a particular subgroup of interviewers, the estimates of some of the coefficients in the fixed part of the model could be seriously biased.

The second application uses a quasi-likelihood adaptation of the estimation procedure for a binary response variable. Again the basic analysis confirms the presence of a substantial interviewer effect. The value of the intra-interviewer correlation coefficient, 0.29, suggests that the potential impact on the analysis of relationships could be overwhelming. However, in this case interviewer variability only appears to "mask" the strength of relationships between variables. An interesting extension illustrates that variation in interviewer experience and response rate can be seen to account for about 50% of the variability introduced into the responses by the interviewers.

On balance, variance component models can be seen to provide a valuable and exciting new approach to the analysis of the impact of variability among interviewers on the results of sample surveys. They supply a framework within which the traditional analysis of interviewer variance can be imbedded. They accommodate extensions of the analysis to include the differential impact of interviewers on respondents with different characteristics, an indication of the sensitivity of the substantive exploratory models to different subsets of interviewers as well as permitting the inclusion of interviewer level characteristics dirctly in the modelling to explain the nature of any interviewer variability. Thus the concerns of the survey methodologist and of the data analyst can be unified in a single approach.

For further information, write to R. Wiggins at The London School of Economics, Houghton St., London, WC2A 2AE or telephone 405 7686, ext. 2639. e-mail: WIGGINSR@LSE.VAX1

References

Fellegi, I. P. (1964). Response variance and its estimation. *JASA, 59,* 1016-1041.

Fellegi, I. P. (1974). An improved method of estimating correlated response variance. *JASA, 69,* 496-501.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *JASA, 57,* 92-115.

Longford, N. T. (1989). Modelling with VARCL. *Multilevel Modelling Newsletter, 1,* 2-3.

Continued on page 11.

ILLUSTRATIONS

CONTEXTUAL EFFECTS AND GROUP MEANS

Nick Longford

In several recent articles dealing with statistical applications in educational research it has been taken for granted that contextual effects in multilevel analysis can be represented by group means. Certainly, contextual effects have to be represented by group-level variables, and aggregates of the individual-level variables have often been found statistically highly significant. However, the group means are only a proxy for the construct (context), the quality of which cannot be properly assessed, and therefore indiscriminant reference to the estimated regression coefficients for these group means as contextual effects may not be theoretically justified.

The following example from an Educational Testing Service study illustrates high collinearity due to excessive application of contextual variables and of cross-level interactions. The dataset consists of test scores from approximately 4200 U.S. graduate students in about 250 graduate school departments that volunteered for the study. The purpose of the study is to report to the graduate schools participating in the study the association of the first-year average grade (F) of its students with the undergraduate grade point average (U), and the Verbal (V), Quantitative (Q) and Analytical (A) scores in the GRE (Graduate Record Examinations) test. Most of the departments have between 8 to 14 students in the data, although the average department-size is about 17. Obviously, it is essential to incorporate the between-department information in the analysis. The general form of the models that have been fitted in the past is that of the two-level model:

$$F_{ij} = p_j + u_j U_{ij} + v_j V_{ij} + q_j Q_{ij} + a_j A_{ij} + \epsilon_{ij}, \quad (1)$$

with linear department-level regressions:

$$\begin{aligned} p_j &= G_{11} + G_{12}\overline{X}_j + \delta_{p,j} \\ u_j &= G_{21} + G_{22}\overline{X}_j + \delta_{u,j} \\ v_j &= G_{31} + G_{32}\overline{X}_j + \delta_{v,j} \\ q_j &= G_{41} + G_{42}\overline{X}_j + \delta_{q,j} \\ a_j &= G_{51} + G_{52}\overline{X}_j + \delta_{a,j} \end{aligned} \quad (2)$$

where G_{kl} are regression parameters. We assume that $(\delta_{p,j}, \delta_{u,j}, \delta_{v,j}, \delta_{q,j}, \delta_{a,j}) \sim N(\mathbf{0}, \Sigma)$. The case of "no context" corresponds to $G_{12} = G_{22} = \ldots = G_{52} = 0$. In the study the department mean for one of the predictors U, V, Q, or A was considered for the variable $\overline{X}_j$ in (2). The unconditional mean for F is equal to

$$\begin{aligned} E(F_{ij}) = G_{11} &+ G_{12}\overline{X}_j \\ &+ (G_{21} + G_{22}\overline{X}_j)U_{ij} \\ &+ (G_{31} + G_{32}\overline{X}_j)V_{ij} \\ &+ (G_{41} + G_{42}\overline{X}_j)Q_{ij} \\ &+ (G_{51} + G_{52}\overline{X}_j)A_{ij} \end{aligned} \quad (3)$$

where G_{kl} are the parameters in the 5×2 matrix $\mathbf{G}$. A simple eigenvalue analysis of the variables $1, \overline{X}_j, U_{ij}, V_{ij}, Q_{ij}, A_{ij}, \overline{X}_j U_{ij}, \overline{X}_j V_{ij}, \overline{X}_j Q_{ij}$, and $\overline{X}_j A_{ij}$ (whichever the choice of X) shows that these 10 regressors represent effectively, a parameter space with only 6 dimensions. If the department-level regressions are extended to include the means of all the four variables, U, V, Q, and A, then the expression for the mean of an observation corresponding to (3) contains 25 regression terms. The eigenvalue analysis for these 25 regressors indicates that they can be effectively replaced by 8 variables. In this context-saturated model, or the model given by (1) and (2), centering around the group-means corresponds to a linear transformation in the regressor space, and so it will not alleviate the confounding among the regression parameters.

The problem with the contextual model given by (1) and (2) has been identified in a roundabout way. The program director of the study was not happy that a large number of departments had at least one negative coefficient in their within department regression prediction formula. Overparameterization, and the resulting collinearity, is the obvious explanation, even though it may be surprising in view of such a small number of parameters and a large number of departments. On closer inspection it transpires that the prediction formulae are very unstable, but the predictions themselves are not! The tremendous flexibility in the multilevel models has thoroughly confused the analyst.

The search for a more parsimonious model is not without hurdles. The variables U, V, Q, and A,

Continued on page 11.

COMMENT ON *"Centering" Predictors in Multilevel Analysis*

Ian Plewis

In the May 1989 issue of the Newsletter, Raudenbush argues that "centering" explanatory variables will often be a sensible way of specifying multilevel models. In other words, rather than using X_{ij} as an explanatory variable for the response Y_{ij}, researchers should use $(X_{ij} - \overline{X}_j)$ where the subscript j refers to the level 2 groups (schools etc.). Although there may be occasions when such an approach is sensible, I shall argue that researchers should be very wary of adopting such a specification routinely.

Contextual Effects and Collinearity

Consider the following model:

$$Y_{ij} = \beta_{0j} + \beta_1 X_{ij} + e_{ij}$$
$$\beta_{0j} = \theta_{00} + \theta_{01}\overline{X}_j + u_{0j} \qquad (1)$$

This is essentially Raudenbush's Equation 1 which he does not like, asserting that estimates of θ_{01} in particular will be imprecise because X_{ij} and $\overline{X}_j$ will often be highly correlated. There may be occasions when $r(X_{ij}, \overline{X}_j)$ is high. However, it approximately equals $\sqrt{\rho_x}$ where ρ_x is the intra-unit correlation, so it will only be high when there is a considerable degree of within-group homogeneity for the explanatory variable. Indeed, in a recent study at the Thomas Coram Research Unit, with a sample of about 800 pupils grouped in 42 classes and tested twice, the correlations for reading and maths attainments were all about 0.4 which is not high collinearity. A greater threat to the precision of θ_{01} is likely to be the size of the sample at level 2 rather than $r(X_{ij}, \overline{X}_j)$.

It is also worth pointing out that contextual effects can be measured in a number of ways. For example, the median group attainment, the within-group standard deviation and the coefficient of variation are all possible candidates. To presuppose that the mean is the best contextual measure seems unwise. This brings me to my second criticism, that of strategy.

Modelling Strategies

A commonsense approach to multilevel modelling would seem to be first to establish whether there is level 2 variation, both for the intercept and for slopes, and if there is, then attempting to model it. However, this strategy will not work with Raudenbush's approach, which prejudges the issues by assuming there are contextual effects (and hence assuming there is level 2 variation). His model is:

$$Y_{ij} = \beta_{0j} + \beta_1(X_{ij} - \overline{X}_j) + e_{ij}$$
$$\beta_{0j} = \theta_{00} + \theta_{01}\overline{X}_j + u_{0j}. \qquad (2)$$

It is, I believe, much more common for X_{ij} to be the best predictor of Y_{ij}, not $X_{ij} - \overline{X}_j$. (The latter will be the best predictor of $Y_{ij} - \overline{Y}_{j\cdot}$.) Consequently, there will often appear to be variation at level 2 for Y_{ij} in model (2) ($\sigma_0^2 > 0$) because the level 1 model is not properly specified. Thus, a researcher will look for explanations of σ_0^2 (and fail to find them) when it would have been sensible to have concluded that there is no level 2 variation. Indeed, if $\sigma_0^2 = 0$, then θ_{01} must equal not zero but β_1 which is confusing.

A General Model

Consider the following more general model:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij}$$
$$\beta_{0j} = \theta_{00} + \theta_{01}Z_{1j} + u_{0j}$$
$$\beta_{1j} = \theta_{10} + \theta_{11}Z_{2j} + u_{1j} \qquad (3)$$

leading to

$$Y_{ij} = \theta_{00} + \theta_{10}X_{ij} + \theta_{01}Z_{1j} + \theta_{11}X_{ij}Z_{2j} + u_{1j}X_{ij} + u_{0j} + e_{ij} \qquad (4)$$

Replacing X_{ij} by $(X_{ij} - \overline{X}_j)$ in (3) gives

$$Y_{ij} = \theta_{00} + \theta_{10}X_{ij} + \theta_{01}Z_{1j} + \theta_{11}X_{ij}Z_{2j} - \theta_{10}\overline{X}_j - \theta_{11}\overline{X}_j Z_{2j} + u_{1j}X_{ij} - u_{1j}\overline{X}_j + u_{0j} + e_{ij} \qquad (5)$$

We can see that models (4) and (5) are not equivalent. Model (4) is more parsimonious than model (5), an advantage which increases as more group-centred explanatory variables are added to the within-group model. Also, if $Z_{2j} = \overline{X}_j$, then (5) includes a quadratic term in $\overline{X}_j$ which is not in (4) and seems difficult to justify.

The problem of centering only arises when there is genuine variation in slopes. Then, σ_0^2 is difficult to interpret, with different values obtained if

Continued on page 11.

TO CENTER OR NOT TO CENTER

Nick Longford

Steve Raudenbush's article (henceforth SR) in the previous issue of this Newsletter discussed the issue of centering of individual-level variables around their group-means in multilevel analysis. My comment contradicts his conclusions about usefulness and relevance of centering in the general setting of linear models.

Let's consider the model formula (1) of SR, in matrix notation,

$$\mathbf{Y} = \mathbf{Z}\beta + \epsilon, \qquad (1)$$

where $\mathbf{Z}$ is the N (students) x 3 design matrix, with the first column containing the l's, the second the students' scores X_{ij} and the third the students' group-means $\overline{X}_j$, and ϵ_{ij} represents the random terms, Var(ϵ) = $\mathbf{V}$. The structure of $\mathbf{V}$ is not relevant until formula (7). The centered version of the formula (1) is the formula (2) of SR, or in matrix notation

$$\mathbf{Y} = \mathbf{A}\beta^* + \epsilon, \qquad (2)$$

where $\mathbf{A} = \mathbf{ZU}$, and

$$\mathbf{U} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \end{pmatrix}.$$

For the model (1) the (estimated) information matrix for β in (1) is

$$\mathbf{M}_1 = \mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z} \qquad (3)$$

where, if unknown, $\mathbf{V}$ is replaced by its estimate. Similarly the (estimated) information matrix for β^* in (2) is

$$\mathbf{M}_2 = \mathbf{U}^T\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{ZU} \qquad (4)$$

The purported advantage of the centered parameterization employed in (2) is the orthogonality of the columns 2 and 3 of $\mathbf{M}_2$. Standard methods for solving the normal equations (in ordinary least squares, Newton-Raphson or Fisher-scoring algorithm, or the M-step of an EM algorithm),

$$\hat{\beta} = (\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Y} \qquad (5)$$

rely on sweeping (diagonalization) of the information matrix (3). If the sweeping starts at the elements (2,3) and (3,2) of $\mathbf{M}_1$ then the first step leads to

$$(\mathbf{U}^T\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{ZU})^{-1}\mathbf{U}^T\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Y} \qquad (6)$$

which corresponds to the information matrix (4). Therefore centering in the sense of the transformation of (1) to (2) avoids one step of sweeping in the solution of the normal equations. Sweeping is preferable because it involves substantially fewer arithmetic operations, and the analyst does not have to keep track of any reparameterizations.

From the computation of the t-test in the Example 1 of SR we can see that the test for $\gamma = 0$ is equivalent to the test of $\beta_w - \beta_b = 0$, and therefore centering is redundant. There is a problem with estimation of γ due to collinearity if and only if there is a problem with estimation of $\beta_w - \beta_b$. In conclusion, centering of a variable in the context of models (1) and (2) corresponds to the numerical procedure of sweeping.

In random slopes models, however, centering may lead to different models. Model (1) would have the random part

$$u_j + v_j X_{ij} + \epsilon_{ij}, \qquad (7)$$

whereas the obvious choice for the random part in the model (2) is

$$u_j + v_j(X_{ij} - \overline{X}_j) + \epsilon_{ij} \qquad (8)$$

(In (7) and (8), v_j is a random variable, possibly correlated with u_j). These two random parts may correspond to substantially different models, as can be seen from their respective variances. In (7) the variance of an observation is a quadratic function of X_{ij} while in (8) it is a function of the deviation from the group mean, ($X_{ij} - \overline{X}_{j}$.) The choice between (7) and (8) should be based on the desired functional form of the fitted variances. The problem, or advantage, of the random part (8) is that the "context" becomes a factor in the variance heterogeneity. If we have doubts about how well the group-mean $\overline{X}_j$ represents the context, this is hardly a desirable feature. If data contain abundant within-group information, we may consider an extension for the random part (8) to the mix of random slopes and variance heterogeneity:

$$u_j + v_j X_{ij} + w_j \overline{X}_j + \epsilon_{ij}, \qquad (9)$$

in which the variance of an observation is a quadratic function of X_{ij} and $\overline{X}_j$.

Continued on page 11.

A RESPONSE TO LONGFORD AND PLEWIS

Steve Raudenbush

The timely thoughtful replies of Nick Longford and Ian Plewis to my piece on "centering" are most welcome. They create an opportunity to clarify issues that have been fuzzy for years–issues which are crucially important in multilevel analyses. However, I find nothing in their replies which contradicts my original assertions.

Plewis warns against routinely centering level one predictors around their group means. And, of course, I agree (see the "Conclusion" of my article). Experience with applications of multilevel methods reveals that rarely are data analytic decisions routine. Multilevel analysts must carefully choose the metric of level-one predictors in the context of each substantive problem if the results are to be interpretable. Centering predictors around the grand mean is sometimes a good choice. But in many applications (including nearly all studies of growth), it is not.

My central argument was and is that in many studies, group-mean centering will provide the metric of choice. Experience in multilevel analysis is rich in examples that reveal the correctness of this choice. I identified three cases in which group-mean centering can be useful: in the study of a) contextual effects b) cross-level interactions; and c) randomly varying slopes. I consider the rebuttals in each context.

Contextual Effects

My article and both replies first consider the simple case of a cross-sectional school effects study in which there is a single student-level variable, X, say, ability, used to predict an achievement outcome, Y. In this context, analysts commonly find that two parameters are needed to represent the relationship between X and Y because the functional relationship between X and Y within schools is different from the relationship between $\overline{X}$ and $\overline{Y}$ between schools (c.f., Burstein, 1980; Plewis, 1989; Zuzovsky and Aitkin, 1989), indicating the presence of a "contextual" or "compositional" effect. My article considered two alternative parameterizations: with X "uncentered" around group means (my original Equation 1) and "group-centered" (my Equation 2).

A test of the existence of the contextual effect in Equation 1 is a test of the null hypothesis $\gamma = 0$. My article pointed out that an equivalent test given Equation 2 is a test of the hypothesis $\beta_b - \beta_w = 0$ and gave a numerical example of how this test could be performed. My article pointed out an advantage of Equation 2: it eliminates the problem of collinearity between X and $\overline{X}$.

Longford demonstrated that the same numerical algorithm could be used for both equations. He reasoned that this demonstration contradicted my conclusions. But it doesn't! My argument was not based on numerical stability or computational efficiency. Rather, it was based on interpretability.

Equation 2 solves the problem noted by Aitkin and Longford (1986) that "the coefficient for $\overline{X}$ will have a large standard error when X is in the model." This quotation applies only to Equation 1. For Equation 2, the opposite is true: including $X_{ij} - \overline{X}_j$ as a predictor actually reduces the standard error for the coefficient of $\overline{X}_j$. That coefficient is perfectly interpretable as the "between-groups" slope. So Equation 2 will often produce what we want: precise estimation of an interpretable parameter.

This is not a minor issue, because Aitkin and Longford's paper uses the collinearity of X and $\overline{X}$ as an argument against study of contextual effects. As the research cited above and much other research shows, contextual effects in studies of school effects cannot be ignored. Equation 2 provides a means of studying them which eliminates the collinearity problem.

Plewis claims, however, that collinearity will rarely be a problem in studies of school effects. He demonstrates an important point: that collinearity becomes a problem only when the intra-school correlation (proportion of variation between schools on X) is large. Because this intra-school correlation for variables like ability and SES is typically small, he reasons, collinearity can be ignored. I would caution readers, however, not to make such an assumption.

During the past six months, I have been involved in three school effects applications in which the intra-group correlation on a crucial X has been high. First, student ratings of instructional quality vary greatly between classrooms in Thailand. And the functional relationship between the mean rating at the classroom level and mean achievement is quite different from the relationship within classrooms. The processes that cause children to rate one teacher higher than another are quite different from the processes that cause differential student ratings of the same teacher, although both are interesting. The same holds for the distribution of textbooks in Thailand, because again, there are two issues: the distribution of texts across schools, and, given the availability of texts at a school, their distribution to different types of children within a school. The third example occurred in the context of a study of growth. For certain time-varying covariates, most of the variation was between-subjects. Variation over time within-subjects was small, but important. In each of these examples, I viewed Equation 2 as superior because collinearity was a serious problem and its elimination meant that I had a "clean" interpretation of a parameter which could be estimated precisely.

Plewis accuses me of "pre-judging the issue by assuming there are contextual effects." This is certainly not true: I showed how to test for contextual effects in each case precisely because I do not prejudge this issue!

Plewis also argues that the two formulations of the model (Equation 1 versus Equation 2) could lead to different assessments of how much level 2 variation exists. This is false: the error terms in both models are identical.

Finally, Plewis suggests that measures of context other than $\overline{X}$ are possible and sometimes desirable. I agree. For example, diversity measures are used profitably in Raudenbush and Bryk (1986) and Lee and Bryk (1989). Nevertheless, contextual effects of $\overline{X}$ are found so frequently that analysts will continue to look for them. Effects of the median X may be similar, though in most cases the median will less stable than the mean; also, the median, unlike $\overline{X}$, is not orthogonal to $X_{ij} - \overline{X}_j$.

Cross-Level Interactions

My article revisited the reanalysis by Cronbach and Webb (1975) of an aptitude-by-treatment interaction study. The crucial failing of the original analysis was its failure to decompose the aptitude-by-treatment interaction into its within-classroom and between-classroom components. The reanalysis performed the decomposition using the group-centering strategy. The result showed clearly that there was no within-class interaction, and, hence, no evidence of a true aptitude-by-treatment interaction.

The model proposed by Plewis (his Equation 4) suffers precisely the deficiency which Cronbach and Webb identified. In that model, the aptitude-by-treatment interaction is represented by the effect of $X_{ij}Z_j$. The problem is that $X_{ij}Z_j$ effect confounds the within-group interaction, ie the effect $(X_{ij} - \overline{X}_j)Z_j$ and the between-group interaction, ie the effect of $\overline{X}_jZ_j$. In Cronbach and Webb's reanalysis, these two interactions differed substantially; it was precisely the original authors' failure to understand this which led to their false inference about aptitude-by-treatment interactions.

Nor would I advise use of Plewis' Equation 4 with $X_{ij} - \overline{X}_j$ replacing X_{ij} (see the equation just below his Equation 4). This method would not adjust for the effect of $\overline{X}_j$ and would not estimate the $\overline{X}_jZ_j$ interaction. I would model Y_{ij} in his Equation 3 as a function of $X_{ij} - \overline{X}_j$; I would model his β_{0j} as a function of $\overline{X}_j$, Z_j, and $\overline{X}_jZ_j$; and I would model the β_{1j} as a function of Z_j.

When this model is written out as a single equation, we have main effects of $X_{ij} - \overline{X}_j$ (aptitude), $\overline{X}_j$ (mean aptitude), and Z_j (treatment); and two-way interactions of $(X_{ij}-\overline{X}_j)Z_j$ (Cronbach and Webb's within-group interaction) and $\overline{X}_jZ_j$ (Cronbach and Webb's between-group interaction). The model can be simplified as the data warrant; and the problem Plewis mentions of the quadratic term in $\overline{X}_j$ disappears.

Two recent studies have adopted the Cronbach and Webb strategy within the framework of the hierarhcical linear model. In both cases, the isolation of a within-group interaction was the critical goal of the study.

First, Raudenbush and Bryk (1986) reanalyzed the US high school data studied by Coleman, Hoffer, and Kilgore (1982). The original analysis tested the effect of $X_{ij}Z_j$ using a structural model similar to Plewis' Equation 4 and found a weaker effect of student SES (X_{ij}) in the Catholic school sector than in the public sector (let's call "sector" Z_j).

The authors concluded that Catholic schools were more egalitarian than were public schools. But this finding was ambiguous: only if the within-school interaction were present would the Coleman et al's interpretation be supported. The reanalysis based on group-mean centering showed that the within-school interaction was, in fact, substantially larger than the between-school interaction, which was non-significant. In this case the original analysis underestimated the all-important within-group interaction.

Second, Lee and Smith (1989) examined evidence regarding sex discrimination in salaries of US high school teachers. US law allows sex differences which reflect different market conditions across school districts. Lee and Smith reasoned that such market effects would be manifest in sex differences between schools. On the other hand, they reasoned if sex differences persist within-schools–after controlling for teacher qualifications, seniority, etc.–evidence of discrimination would be incontrovertible. By school-centering the indicator for sex, they guaranteed that their estimate of the sex effect would be uncontaminated by omitted school-level variables. Similarly, they was able to identify characteristics of schools which predict the sex effect. Even if their model for the mean salary levels of the schools were badly misspecified, their estimate of the sex effect and its correlates would be unbiased as long as the within-school part of the model were adequately specified. This is a major advantage because relatively little is known about determinants of salary differences among schools while a great deal is known about the determinants of salaries within schools.

Randomly Varying Slopes

My article pointed out that when slopes are random and group-mean centering is not employed, intercepts can be uninterpretable and slope variance becomes confounded with intercept variance. Longford agrees that this is can be a problem and offers a more complex model in which variance heterogeneity depends on $\overline{X}_j$ and random slopes are represented by the effect of X_{ij}. This model will at times be quite useful, though the possibility of collinearity between X_{ij} and $\overline{X}_j$ reappears.

So one use of group-mean centering is to clarify interpretation in models with random slopes. But Plewis makes an interesting point: if, for every random slope, one uses both $X_{ij} - \overline{X}_j$ and $\overline{X}_j$ as predictors, is there not danger of lack of parsimony? This would occur if the fixed effects of $X_{ij} - \overline{X}_j$ and $\overline{X}_j$ were equal. In many studies, this lack of parsimony will matter very little because there will be hundreds of classrooms or schools and only a few random coefficients. But when the number of these level 2 units is small, parsimony becomes a concern. In this case, Lindsay Paterson (personal communication) has suggested the model

$$Y_{ij} = b_0 + b_1 X_{ij} + u_{0j} + u_{1j}(X_{ij} - \overline{X}_j) + e_{ij}. \quad (1)$$

In this model, the intercept is interpretable as the group mean adjusted for the average effect of X and the slope variance represents "purely" within-group variance in Y. Moreover, the model is parsimonious, representing the fixed effect of X on Y with one parameter. The model can be estimated using HLM (which allows a random effect predictor even when there is no corresponding fixed effect predictor), and preliminary results using Scottish data produce meaningful results. However, I would suggest use of this model only if the contextual effect of X is found non-significant. When the contextual effect is significant, Equation 1 with the effects of $\overline{X}_j$ added and group-centering of X_{ij} will be superior. Also a "variance heterogenerity" term (random effect of $\overline{X}_j$) as in Longford's Equation 9 could be added to Equation 1 above.

References

Aitkin, M. & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, 149,* 1-43.

Burstein, L. (1980). The analysis of multi-level data in educational research and evaluation. *Review of Research in Education, 8,* 158-233.

Coleman, J.S., Hoffer, T., & Kilgore, S.B. (1982). *High school achievement: Public, Catholic, and other schools compared.* New York: Basic Books.

Cronbach, L.J., & Webb, N. (1975). Between- and within-class effects in a reported aptitude-by-treatment interaction: Reanalysis of a study by G.L. Anderson. *Journal of Educational Psychology, 6,* 717-724.

Continued on page 11.

PLEWIS ... cont'd from p. 6

the relevant explanatory variable is not centered, is centered at the grand mean or at the group means. Interpretation of $\hat{u}_j$ is also problematic. The situation is analogous to treatment-covariate interactions in analysis of covariance-type models when inferences about treatment effects can be difficult to make. However, trying to eliminate the difficulties by using Raudenbush's model seems to create more problems than it solves.

LONGFORD ... cont'd from p. 7

If an unstructured form of the variance matrix for (u_j, v_j, w_j) is assumed, containing three variance and three covariance parameters, replacing X_{ij} with $X_{ij} - \overline{X}_j$, i.e., centering around the sample mean, amounts to a mere reparameterization, because the random part (9), as well as (7) and (8), are invariant with respect to translation. The advantage of this centering is that $\text{Var}(u_j)$ becomes interpretable–it is the group-level variance component for $X = X_j$. In the non-centered model (7) $\text{Var}(u_j)$ is usually very large, and $r(u_j, v_j)$ is very close to either +1 or -1. This is completely analogous to the uninterpretability of the intercept in many ordinary repression, (or multilevel) models.

RAUDENBUSH ... cont'd from p. 10

Lee, V. & Bryk, A.S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education, 62,* 72-92.

Lee, V. & Smith, J.B. (1989, August). *Gender equity in teachers' salary: A multilevel approach.* Paper presented at the International Conference on Multilevel Methods in Educational Research, Edinburgh.

Plewis, I. (1989, August). *Using multilevel models to link educational progress with curriculum coverage.* Paper presented at the International Conference on Multilevel Methods in Educational Research, Edinburgh.

Raudenbush, S.W. & Bryk, A.S. (1986). A hierarhcical model for studying school effects. *Sociology of Education, 59,* 1-17.

Zuzovsky, R. & Aitkin, M. (1989, August). *A Two-level model analysis of achievement of the second IEA science study in Israel.* Paper presented at the International Conference on Multilevel Methods in Educational Research, Edinburgh.

WIGGINS ET AL. ... cont'd from p. 4

O'Muircheartaigh, C. A., & Wiggins, R. D. (1981). The impact of interviewer variability in an epidemiological survey. *Psychological Medicine, 11,* 817-824.

Wiggins, R. D. (1985). A replicated study of the impact of interviewer variability in a community survey of the physically handicapped in an inner London borough. Polytechnic of Central London Research Working Paper, no. 24.

LONGFORD ... cont'd from p. 5

or at least the latter three, should be symmetrically represented in the model; it is difficult to accept a model which contains the department means V_j, but not A_j. But if both variables are included in the model (even without any of their cross-level interactions) the parameter space contains redundancy that cannot be removed by any form of centering.

Confounding among the parameters in the 5×5 variance matrix **E** is a related issue. In most models the estimate of **E** is singular, usually of rank 3. That indicates that the department-level variation could be satisfactorily represented by a 3×3 variance matrix, containing only 6 parameters, instead of the 15 parameters (5 variances and 10 covariances) in the fitted models. The best evidence for acute collinearity among these parameters can be obtained from the estimated information matrix (15×15) for the variances and covariances. It is important to realize that for inference about complex patterns of variation much larger datasets are required than for inference about regression relationships of similar complexity.

This example demonstrates that even a very small number of variables formed as group means and their interactions with individual-level variables may lead to confounding and instability of the regression parameters. An eigenvalue analysis of the crossproduct-matrix of the regressor variables provides a simple diagnostic for such confounding. The estimates of the variances and covariances are also subject to confounding. This is indicated in most algorithms by very slow convergence. Singularity of the variance matrix is usually a sign of overparameterization.

MULTILEVEL WORKGROUP STARTED

Ita Kreft

A new workgroup has been started at UCLA's Center for the Study of Evaluation for applied statisticians and social science researchers interested in multilevel modelling. The current focus of the group is developments in statistical analysis and software, and meetings provide a forum for workers in different fields to share experiences and common problems. Presentations of papers and extensive analyses are most welcome.

At the group's October 6th meeting–one or two are planned each quarter–Bill Mason, director of the Population Studies Center at the University of Michigan, gave a talk.

Interested researchers, within and outside California, are invited to join. Further information is available by writing Ita Kreft, UCLA, Moore Hall 145D, 405 Hilgard Avenue, Los Angeles, CA 90024-1522, or by telephoning (213) 206 1528.

The **Multilevel Modelling Newsletter** is published by the Multilevel Models Project at the Institute of Education (University of London). We welcome your suggestions and ideas, and we request reports on your work, other news, and questions regarding analysis issues.

If you know of interested colleagues who would like to receive future editions, please send a note of their names and addresses to Bob Prosser at the address on the masthead. (Copies of the first two issues are available if you missed them.)

The newsletter is distributed free of charge (in January, May, and October), but reproduction and distribution are costly. We would ask that if you are able, to please send a contribution towards these expenses (say, £3 or US$5). Thank you.

CONTRIBUTORS

Thanks very much to the people who provided articles for this issue.

Michael Healy

Institute of Education
University of London
20 Bedford Way
London WC1H 0AL

Ita Kreft

Center for the Study of Evaluation
Moore Hall, 145D
405 Hilgard Ave.
Los Angeles, CA 90024

Nick Longford

Center for the Study of Evaluation
Moore Hall, 145D
405 Hilgard Ave.
Los Angeles, CA 90024

Colm O'Muircheartaigh

London School of Economics
Houghton St.
London WC2A 2AE

Ian Plewis

Thomas Coram Research Unit
41 Brunswick Square
London WC1

Steve Raudenbush

Department of Education
Michigan State University
East Lansing, Michigan 48824
U.S.A.

Dick Wiggins

London School of Economics
Houghton St.
London WC2A 2AE

IN THE JANUARY ISSUE ...

INTO A NEW DECADE OF MULTILEVEL MODELLING
REVIEW OF *Analysis of Complex Surveys*
NEWS FROM AUSTRALIA
AND MORE ...