MULTILEVEL MODELLING NEWSLETTER

Centre for Multilevel Modelling

Bedford Group for Lifecourse and Statistical Studies
Institute of Education, University of London
20 Bedford Way, London WC1H 0AL, ENGLAND
Web site: http://www.mlwin.com/
Enquiries about newsletter to Ian Plewis
E-mail: i.plewis@ioe.ac.uk
Tel: +44 (0) 20 7612 6238
Fax: +44 (0) 20 7612 6572

Vol. 17 No. 1 **September, 2005**

Editorial

This is the last issue of the newsletter to be edited by me. I have done the job for the last six years but, to coincide with the move of the Centre for Multilevel Modelling from the Institute of Education in London to Bristol (see Jon Rasbash's article in this issue), I am handing over the reins to Harvey Goldstein. I would like to thank all of you who have contributed articles, news and reviews and also to encourage everyone to continue to think about the newsletter as an outlet for material on multilevel modelling whose importance continues to grow, not only as a set of statistical techniques but also as a valuable way of thinking about many aspects of the world.

Ian Plewis

Fifth International Amsterdam Conference

The Fifth International Amsterdam Conference on Multilevel Analysis was held in Amsterdam on 21-22 March 2005. The following papers were presented:

Carlos A. Q. Coimbra and Tom A. B. Snijders
Estimation of Non-Linear Models by Stochastic Approximation

G. W. Jacobusse, S. Van Buuren and C. G. M. Groothuis-Oudshoorn
Multiple Imputation of Missing Data in a Multilevel Setting

Tom A.B. Snijders
MLwiN Macros for Nonlinear Transformations of Independent Variables

S. Teerenstra, R. J. F. Melis, P. G. M. Peer and G.F. Born
Pseudo Cluster Randomization

Also in this issue
Learning Environment for Multilevel Methodology and Applications (LEMMA)
Selection bias in random intercept models
Review of 'Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models'

Mirjam Moerbeek
On the Design of Cluster Randomized Trials for the Comparison of Group versus Individual Therapies

Michael Schweinberger and Tom A.B. Snijders
Random Effects Models for Digraph Panel Data

Bonne J.H. Zijlstra and Marijtje A.J. Van Duijn
A Multilevel P2 Model

Harvey Goldstein
Multilevel Multidimensional Structural Equation Models with an Application to the Analysis of Comparative Data on Educational Performance

Neil H. Spencer
Defining Structures for Multilevel SEMS of Pupil Progress

Johannes Hartig and Nina Jude
Effects of Different Estimators for Student Proficiencies on Multidimensional Multilevel Structures

Leigh Anne Shafer, Janet Rice, Leann Myers, John Lefante and Jim Todd
Comparison of Methods in Regression Analysis with Longitudinal Data: A Simulation Study

Susanne Eschmann, Daniel Zimprich, Christa Winkler Metzke and Hans-Christoph Steinhausen
Internalizing Problems in Adolescence: Risk-Groups Based on Developmental Trajectories

Olga Mitina
Using Multilevel Modeling for Analysis Data Got from Semantic Differential Method

Michela Battauz, Ruggero Bellio and Enrico Gori
A Multilevel Measurement Error Model for Value-Added Assessment in Education

Markus Hadler
Individual Attitudes Towards Different Political Regimes

Fernando De Maio
Health Inequities in Argentina: Multilevel Analysis of the Contextual Effect of Provincial Income Inequality

Ian Plewis
Comparing Parents' Responses over Time: A Multilevel Perspective

Bradley A. Corbett
Risk and Protective Factors Influencing the Prevalence of Smoking among Canadian Adolescents: A Two Level Logistic Regression Growth Model using Data from Canada's National Longitudinal Study of Children and Youth (NLSCY)

Marc Callens and Christophe Croux
Poverty Dynamics in Europe. A Multilevel Discrete-Time Recurrent Hazard Analysis

Lars-Erik Malmberg
Father- and Mother-Child Interaction During Feed and Play

Carla Rampichini and Leonardo Grilli
Sample Selection in Multilevel Models

Jean-Paul Fox
Linear Mixed Models for Randomized Responses

John F. Bell and Eva Malacova
Outliers and Multilevel Models

William Browne
An Illustration of the Use of Reparameterisation Methods for Improving MCMC Efficiency in Crossed Random Effect Models

Jeroen Vermunt
Random-Effects Regression Modeling Using Latent Class Methods

Peter C. Austin
A Logistic-Mixture of Normal Distributions Multilevel Model for Hospital Mortality

Jay Magidson and Jeroen K. Vermunt
Analysis of Repeated and Multilevel Discrete Choice, Ranking and Rating Data

Enrico Gori and Luca Grassetti
Linear Mixed ModelsiIn Efficiency Identification

P. Van Dommelen, S. Van Buuren, G.R.J. Zandwijken and P.H. Verkerk
A Nonlinear Mixed Model for Detecting Girls With Turner Syndrome

Laura Green
Use of Statistical Models to Understand Footrot in Sheep, an Infectious Disease

Joop J. Hox and Cora J.M. Maas
Approximating Cross-Classified Models by Confounding Classifications

Alastair H. Leyland and Øyvind Næss
Using Correlated Cross-Classified Multilevel Models to Estimate Area Influences on Health Throughout the Lifecourse

Omar Paccagnella
The Accuracy of Estimates in Discrete Responses Multilevel Models. New Simulation Results

Learning Environment for Multilevel Methodology and Applications (LEMMA)

Jon Rasbash
University of Bristol
j.rasbash@bristol.ac.uk

The Centre for Multilevel Modelling is moving from the Institute of Education in London to the University of Bristol. The centre staff, with other Bristol academics, were successful in an application for a new research project called LEMMA. The LEMMA project is one of a set of six nodes funded under the ESRC’s National Centre for Research Methods (NCRM). The brief of the NCRM (www.ncrm.ac.uk) is to provide a step change in the quality of social science research in the UK. This article describes the LEMMA project.

The project has three inter-related elements: (i) statistical methodology; (ii) flagship substantive research projects; (iii) sets of materials and systems for training and capacity building.

Methodology

The usual distributional assumption of Normality for higher level random effects can be overly restrictive. To give two examples:

(1) Nagin (1999) describes a formulation for growth curves where a discrete set of latent groups is posited and each individual has a membership distribution across the groups. Muthén (2004) has also implemented models in this area.

(2) In binary response and event history models, many higher level units have response patterns of all zero or all one and this leads to the class of mover/stayer models, which also utilize latent categories. Such models have been implemented in SABRE (http://www.cas.lancs.ac.uk/).

The existing work in the area of latent categorical effects in multilevel models has focused on fitting latent categorical distributions to hierarchical models. (Rabe-Hesketh et al., 2004, Chapter 5; Vermunt and Magidson, 2005). The LEMMA project will build on existing work so that latent categorical distributions can be fitted to any level in multilevel models which can contain mixtures of nested, crossed and multiple membership classifications.

Flagship substantive research projects

A number of projects are planned that will demonstrate how multilevel models can be applied to substantive social science problems.

Geography of school effects

This project addresses the relationship between school effectiveness, school choice and parental relocation, thereby addressing current debates about the effects of quasi-markets in education. This will be tackled by using Pupil Level Annual School Census (PLASC) data and also possibly the exceptionally detailed Avon Longitudinal Study of Parents and Children (ALSPAC) data. These data have a highly complex structure including multiple membership and crossed classifications; repeated measures on individuals within areas; movement of individuals between areas and schools; repeated measures on individuals within primary year cohort within primary school; and repeated measures within secondary year cohort within schools. Spatial models will also be used to model ‘competition’ between the higher level units, such as schools with overlapping catchments, differentiated by school type.

Modelling the duration of episodes in hospital

The effects of consultants, hospitals and geographical areas on the length of time a patient stays in hospital will be explored using data from the Hospital Episodes Statistics dataset. These data have a highly complex non-hierarchical structure. A patient might have more than one stay in hospital, leading to repeated episodes nested within individuals. Episodes and consultants have a multiple-membership structure since a patient might see more than one consultant during an episode and different consultants between episodes. In addition, consultants are crossed by hospitals, as consultants can work at more than one hospital, and patients are nested within a cross-classification of hospitals and geographical areas.

Voting choice

The substantive issue focuses on the individual, household and neighbourhood determinants of voting abstention and party choice. Using the British Household Panel Study (BHPS) we have repeated binary measures on voting intention for individuals within households within areas at a variety of scales. Normally distributed individual level random effects are unrealistic as part of the mover/stayer problem as are Normal household effects. We will compare the autocorrelation approach of Goldstein and Barbosa (2000) and the discrete latent-effects model, in particular a doubly-nested model, with latent classifications at the individual and household level.

Mental health and psychosocial development

This uses measures on mental well-being from two data sets. The first is the BHPS and will compare the use of Normal random effects and discrete latent effects to describe between individual variations in patterns of change over time with a multiple-membership model to take account of changes in household composition. The second dataset is the Avon Brothers and Sisters Study with repeated measures on psychosocial adjustment, for multiple children within families. Again we will explore the use of continuous and discrete latent effects to describe between individual patterns of variation in psycho-social development.

Modelling group diversity

Traditional statistical models have concentrated on modelling mean effects as functions of predictor variables. Multilevel models allow us to model the variation for any classification as a function of further variables. For example Goldstein and Noden (2003) model the between school and between LEA variation in the percentage of pupils eligible for free school meals as functions of LEA characteristics. Such models, applied for example to measures of poverty or service delivery, are highly relevant to current debates about diversity since they avoid certain arbitrary features of traditional index measures, and provide efficient and objective estimates of between-unit variation. A further development is to construct models that use such estimates of diversity, for example estimated for each unit in a classification, as predictors in a further model where outcomes are a function of area level measures of diversity.

Training and capacity building

The resources we plan to develop will be useful for solo and group learning.

The resources are aimed at social scientists with a wide spectrum of statistical expertise.

Our planned training materials will be designed to give users the necessary skill to carry out quantitative research on data with complex structure. They will be

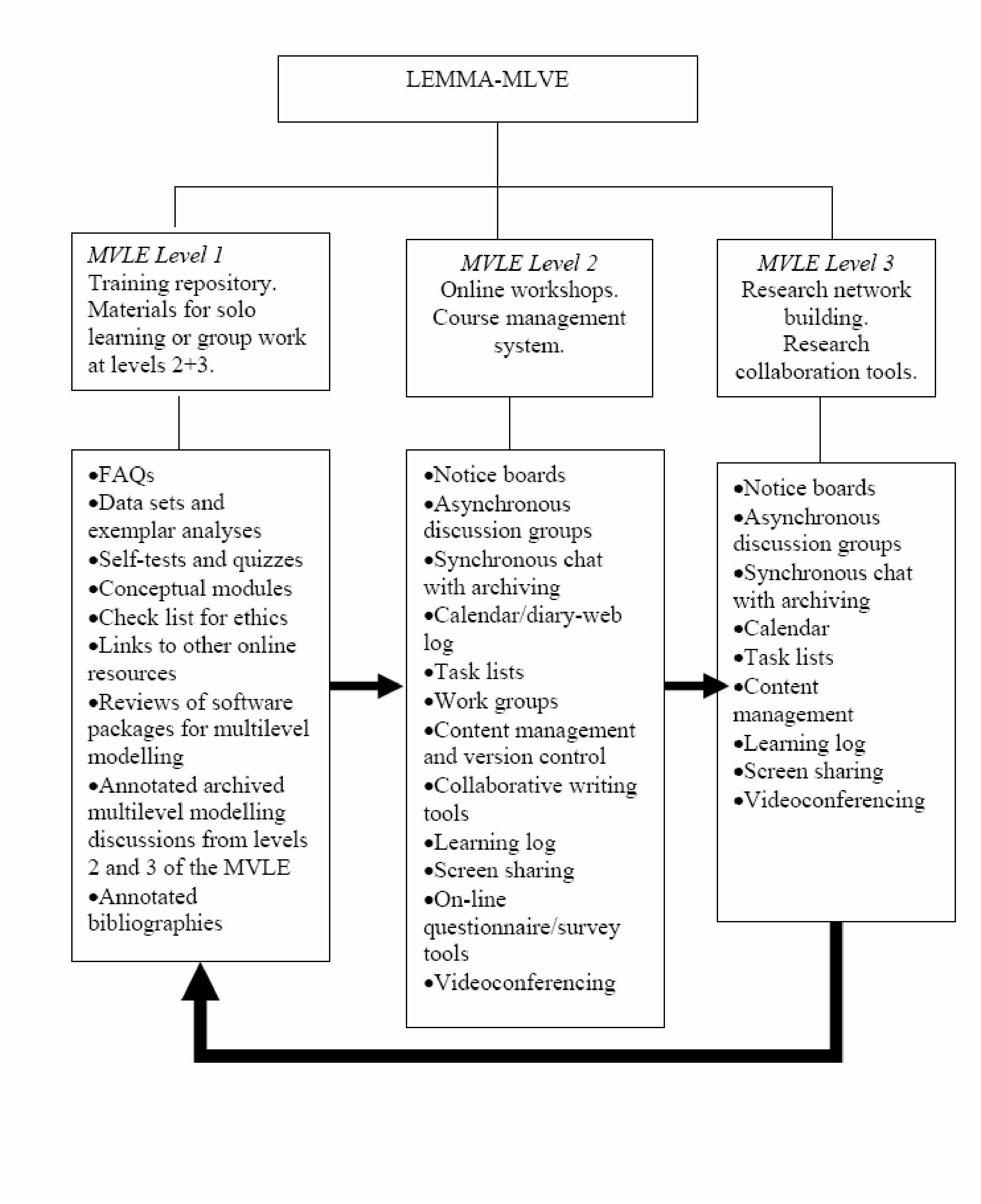
- Carefully graded
- Model-based
- Realistic
- Authentic
- Contextualised

The project will be running a mixture of conventional face-to-face training workshops and clinics. These modes of training have limitations in that workshops are over-subscribed and there are insufficient resources to provide follow-up support. This lack of follow up support often prevents researchers from converting the methodological insights and practical skills gained at a workshop into routine use of these skills in their day to day research work. The LEMMA project aims to address these training shortfalls with a web based learning environment which will support solo and group learning and provide support to try and foster self-supporting sustainable groups of researchers.

Web-based training

A substantial and ambitious new venture that builds on our web experience will be the provision of a range on online resources, including a repository of training materials, a series of collaborative and moderated online workshops, as well as a number of research networks which together form a multilevel modelling virtual learning environment (MVLE) designed to initiate, develop, and support dispersed researchers.

The anticipated architecture of the Information and Communications infrastructure that will be adopted as the MVLE is sketched out in the diagram below:


LEMMA-MLVE
MVLE Level 1
Training repository.
Materials for solo learning or group work at levels 2+3.
MVLE Level 2
Online workshops.
Course management system.
MVLE Level 3
Research network building.
Research collaboration tools.
•FAQs
•Data sets and exemplar analyses
•Self-tests and quizzes
•Conceptual modules
•Check list for ethics
•Links to other online resources
•Reviews of software packages for multilevel modelling
•Annotated archived multilevel modelling discussions from levels 2 and 3 of the MVLE
•Annotated bibliographies
•Notice boards
•Asynchronous discussion groups
•Synchronous chat with archiving
•Calendar/diary-web log
•Task lists
•Work groups
•Content management and version control
•Collaborative writing tools
•Learning log
•Screen sharing
•On-line questionnaire/survey tools
•Videoconferencing
•Notice boards
•Asynchronous discussion groups
•Synchronous chat with archiving
•Calendar
•Task lists
•Content management
•Learning log
•Screen sharing
•Videoconferencing

The proposed ICT architecture as shown maps on to the pedagogical design for the whole system. This anticipates three levels of activity, which feed into each other:

Level 1

Repository of Training Materials

This is essentially a database of teaching and learning materials which can be used by solo learners or tutors for delivering their own training.

Level 2
Online Workshops
Moderators use materials provided in level one to facilitate group learning in online courses. These are designed to promote online group formation which will be carried through into level three after the end of formal teaching.

Level 3
Online Research Communities
This forms the core of collaborative knowledge building in intensely focused groups. Computer-supported, collaborative work-tools will be used to help form and sustain online networks. It is anticipated that some of the outputs of these networked activities will result in learning objects being deposited in MVLE level one such as new exemplars and annotated archived discussions.

We regard the feedback of new knowledge from level three back into levels one and two of the MVLE as our mechanism for the spread of knowledge, concepts and new practices into the wider research community. The above learning architecture is no more than a representation of what the applicants envisage as the final pedagogical structure as every level of the MVLE will go through an iterative and participatory process of design, delivery, evaluation and re-design.

The first release of the LEMMA MVLE will be in 2006. Following the link to LEMMA on the National Centre for Research Methods site (www.ncrm.ac.uk) will provide progress reports on the LEMMA node's activities.

References

Barbosa, M. F. & Goldstein H. (2000). Discrete response multilevel models for repeated measures: An application to voting intentions data. *Quality & Quantity*, **34**: 323-330.

Jones, K (2005). An introduction to statistical modeling. In Somekh, B. and Lewin, B. (eds). *Research Methods in the Social Sciences*, Sage.

Goldstein, H., and Noden, P. (2003) Modelling social segregation. *Oxford Review of Education*, **29**: 225-237.

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (ed.), *Handbook of Quantitative Methodology for the Social Sciences*.

Nagin, D. (1999) Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods*, **4**: 139-177.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). GLLAMM Manual (http://www.bepress.com/ucbbiostat/paper160/). U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 160.

Vermunt, J.K, and Magidson, J. (2005). Hierarchical mixture models for nested data structures. In C. Weihs and W. Gaul (eds), *Classification: The Ubiquitous Challenge*. Heidelberg: Springer.

Selection bias in random intercept models

Leonardo Grilli and Carla Rampichini
Department of Statistics, University of Florence
carla@ds.unifi.it

Introduction

Sample selection is a common problem in observational studies, as it arises when the response variable of principal interest Y^P is observed conditionally on the value of another variable, say $Y^S > 0$. For example, the desired number of labour hours supplied is observed only for people actually working. In regression analysis, sample selection leads to biases if the selection mechanism depends on unobserved variables correlated with the model errors.

Starting from the work of Heckman (1979), the problem of selection bias has been thoroughly studied in the context of standard single level models and linear models for panel data. See Vella (1998) for a general review.

Applications dealing with sample selection in multilevel settings are rare (Borgoni and Billari, 2002; Bellio and Gori, 2003; Grilli and Rampichini, 2004) and there is no systematic study on selection bias in multilevel models.

The phenomenon of selection in a multilevel model is more complex than in a single level model for the following reasons: (a) the selection process can act at different levels, giving rise to a wide variety of patterns; (b) the variance-covariance structure is often of primary interest, so the effect of selection on the variance-covariance structure must be carefully assessed; (c) the selection process modifies the hierarchical structure of the data, in terms of number of clusters and cluster sizes, a feature that is relevant in the estimation phase, as it influences the behaviour of the estimation algorithms, the accuracy of the asymptotic approximations and the power of the tests.

The aim of this contribution is to outline the consequences of sample selection in the relatively simple case of a random intercept model, focusing on the linear case and only mentioning the extension to the binary case.

The model

Let us consider a two-level hierarchy, where $j = 1, 2, \ldots, J$ is the index of the level two units (clusters) and $i = 1, 2, \ldots, n_j$ is the index of the level one units (elementary units). Then let us denote with Y_{ij}^S and Y_{ij}^P two continuous response variables, where S stands for Selection and P for Principal. A bivariate linear two level random intercept model can be written as

$$\begin{aligned} Y_{ij}^S &= \mathbf{z}_{ij}^S \boldsymbol{\theta}^S + u_j^S + e_{ij}^S \\ Y_{ij}^P &= \mathbf{z}_{ij}^P \boldsymbol{\theta}^P + u_j^P + e_{ij}^P \end{aligned} \qquad (1)$$

For any equation, $\mathbf{z}_{ij}$ are covariates at the elementary or cluster levels (a given covariate may enter one or both equations), $\boldsymbol{\theta}$ are regression coefficients, u_j are cluster level errors (random effects), e_{ij} are elementary level errors. The errors are assumed to be independent at different levels with distributions:

$$\begin{bmatrix} e_{ij}^S \\ e_{ij}^P \end{bmatrix} \overset{iid}{\sim} N\left(\mathbf{0}, \begin{bmatrix} \sigma_S^2 & \\ \sigma_{SP} & \sigma_P^2 \end{bmatrix}\right)$$

$$\begin{bmatrix} u_j^S \\ u_j^P \end{bmatrix} \overset{iid}{\sim} N\left(\mathbf{0}, \begin{bmatrix} \tau_S^2 & \\ \tau_{SP} & \tau_P^2 \end{bmatrix}\right)$$

The distributional assumption of Normality is not essential for the general discussion of selection bias, but it is used to derive the analytical results later shown in Table 1.

When both Y^S and Y^P are observed for every unit, the model for Y^P can be fitted separately without any bias (though with a possible loss of efficiency). Now let us consider the consequences stemming from the following selection mechanism: Y^P is observed if and only if Y^S>0.

Such a selection mechanism operates at the elementary level, as it causes the missingness of single elementary units (even when the elementary-level covariance σ_{SP} is null, as in many models for panel or longitudinal data) and so it modifies the hierarchical structure of the data in terms of cluster sizes and possibly also in terms of number of clusters. Note that the assumed selection mechanism is general as, within a given cluster, the pattern of missingness can be of any kind.

The selection mechanism is ignorable when *both* covariance parameters σ_{SP} and τ_{SP} are null. In this case the model for the *Principal* equation can be fitted separately, without any bias or loss of efficiency. When the selection mechanism is not ignorable, it is of interest to determine the bias which arises when fitting the *Principal* equation alone.

Selection in the linear model: general

Let us define $w_{ij}^S = u_j^S + e_{ij}^S$ as the composite error of the *Selection* equation, so that Y_{ij}^P is observed if and only if $w_{ij}^S > -\mathbf{z}_{ij}^S \boldsymbol{\theta}^S$. Moreover, the set of truncation events of the whole cluster is

$$A_j = \left\{ \bigcap_{i:Y_{ij}^S > 0} \left\{ w_{ij}^S > -\mathbf{z}_{ij}^S \boldsymbol{\theta}^S \right\} \right\} \bigcap \left\{ \bigcap_{i:Y_{ij}^S \le 0} \left\{ w_{ij}^S \le -\mathbf{z}_{ij}^S \boldsymbol{\theta}^S \right\} \right\}$$

Truncation is below for the elementary units which are observed and above for the others.

Now let us consider the first observed elementary unit $(i=1)$ of cluster j, with the corresponding model

$$Y_{1j}^P = \mathbf{z}_{1j}^P \boldsymbol{\theta}^P + u_j^P + e_{1j}^P . \qquad (2)$$

To evaluate the consequences of selection on model (2), it is necessary to condition on A_j, while the observations pertaining to other clusters are irrelevant, as independence is assumed among clusters. The key quantities are the conditional mean

$$E\left(Y_{1j}^P \mid u_j^P, A_j\right) = \mathbf{z}_{1j}^P \boldsymbol{\theta}^P + u_j^P + E\left(e_{1j}^P \mid u_j^P, A_j\right) \quad (3)$$

the marginal mean

$$E\left(Y_{1j}^P \mid A_j\right) = \mathbf{z}_{1j}^P \boldsymbol{\theta}^P + E\left(u_j^P \mid A_j\right) + E\left(e_{1j}^P \mid A_j\right) \quad (4)$$

and the marginal variance

$$V\left(Y_{1j}^P \mid A_j\right) = V(u_j^P \mid A_j) + V(e_{1j}^P \mid A_j) + 2cov(u_j^P, e_{1j}^P \mid A_j) \quad (5)$$

The conditioning on $\mathbf{z}_{1j}^P$ is implicit. The key point is that, due to the conditioning on A_j, the means and variances after selection depend on some features of the cluster: the cluster size, the missingness pattern and ($\mathbf{z}_{1j}^S, \ldots, \mathbf{z}_{n_j j}^S$), i.e. *all* the covariates of the *Selection* equation for *all* the elementary units of the cluster.

If selection is not ignorable, fitting the *Principal* equation alone creates potential problems for both the regression coefficients (slopes) and the variances.

The slopes are biased for those covariates that enter both the *Selection* and the *Principal* equations. In fact, if z_{k1j} enters both equations, from (3) it follows that the conditional slope of z_{k1j} (i.e. given u_j^P) is:

$$\frac{\partial}{\partial z_{k1j}} E\left(Y_{1j}^P \mid u_j^P, A_j\right) = \theta_k^P + \frac{\partial}{\partial z_{k1j}} E\left(e_{1j}^P \mid u_j^P, A_j\right)$$

For a given elementary unit, the selection bias on the slope depends on A_j and therefore on the features of the cluster it belongs to. Moreover, comparing (3) and (4) it is clear that, in general, the well-known equivalence between conditional and marginal slopes in multilevel linear models is corrupted.

As in single level models affected by selection, the slope differs among elementary units, so the resulting estimated slope is a sort of *average* of the elementary-unit values. However, in a multilevel setting the *true* slope may be random: in such a case, the variability of the *true* slope is confused with the variability induced by selection. Also note that if the *true* slope is *not* random, the researcher ignoring the selection bias might incorrectly believe that the model should include a random slope.

As for the variances (see equation (5)), fitting the *Principal* equation alone when selection is not ignorable leads to the following potential problems: (a) the errors are no longer homoscedastic, nor independent, undermining the efficiency of the estimators; (b) the ICC is biased, leading to false conclusions about the hierarchy.

Selection in the linear model: special cases

It is interesting to locate the configurations of the model parameters for which some of the moments of the errors are not affected by selection (even if selection is not ignorable), so that some of the potential biases just described do not operate.

The relevant random variables are the two errors in Y_{ij}^P, namely u_j^P and e_{1j}^P, plus all the composite errors determining selection in the cluster under consideration, i.e. $(w_{1j}^S, w_{2j}^S, \ldots, w_{n_j j}^S)$. The distribution of the relevant errors before truncation is assumed multivariate Normal. However, the joint distribution of $(e_{1j}^P, u_j^P \mid A_j)$, i.e. after truncation on (w_{1j}^S, w_{2j}^S), is no longer Normal (Arellano-Valle and Azzalini, 2005).

When *both* equations are multilevel ($\sigma_S^2 > 0$, $\sigma_P^2 > 0$, $\tau_S^2 > 0$, $\tau_P^2 > 0$) the model errors have a full rank distribution governed by the two covariance parameters (τ_{SP} at the cluster level and σ_{SP} at the elementary level). The study of the (conditional) independencies of the errors reveals that some simplifications occur when one of the covariance parameters vanishes.

Table 1 reports in each cell the moments corresponding to the errors entering, respectively, the conditional mean (3), the marginal mean (4) and the marginal variance (5) of model (2) after truncation. Each cell of Table 1 is defined by a given combination of the covariance parameters, distinguishing between null and non-null values. The cell with $\sigma_{SP}\neq 0$ and $\tau_{SP} \neq 0$ corresponds to the general case where all the potential biases are in effect, while in the cell with $\sigma_{SP} = 0$ and $\tau_{SP}=0$ selection is ignorable. If $\sigma_{SP}\neq 0$ and $\tau_{SP}=0$, i.e. only the elementary level errors are correlated, the conditional slopes are biased and equal to the marginal slopes; moreover the ICC is overestimated, due to underestimation of the elementary level variance. Alternatively, if $\sigma_{SP}=0$ and $\tau_{SP}\neq 0$, i.e. only the cluster level errors are correlated, the conditional slopes are unbiased, but different from the marginal slopes, that are biased; moreover the ICC is underestimated, due to underestimation of the cluster level variance.

If one of the equations is *not* multilevel, then τ_{SP} is necessarily zero: therefore, for $\sigma_{SP} \neq 0$ the moments have the form reported in the upper right cell of Table 1. However, if the equation which is not multilevel is the *Selection* equation (i.e. $\tau_S^2 = 0$), then the form of the moments further simplifies. In this case, indeed, when conditioning on A_j, the elementary units other than the one under consideration (i.e. unit 1) can be ignored, so the relevant conditioning set reduces to $A_{1j}=\left\{w_{1j}^S > -\mathbf{z}_{1j}^S \boldsymbol{\theta}^S\right\}$. As a consequence, the moments of interest have simple expressions (which are well-known in the literature on selection bias: e.g. Heckman, 1979; Copas and Lee, 1997).

Table 1. Some moments of errors of model (2) after truncation (both equations multilevel)

Elementary-level errors covariance	*Cluster-level errors covariance*	
	$\tau_{SP} \neq 0$	$\tau_{SP} = 0$
$\sigma_{SP} \neq 0$	$E(e_{1j}^P \mid u_j^P, A_j)$ $E(u_j^P \mid A_j) + E(e_{1j}^P \mid A_j)$ $Var(u_j^P + e_{1j}^P \mid A_j)$	$E(e_{1j}^P \mid A_j)$ $E(e_{1j}^P \mid A_j)$ $\tau_P^2 + Var(e_{1j}^P \mid A_j)$
$\sigma_{SP} = 0$	0 $E(u_j^P \mid A_j)$ $Var(u_j^P \mid A_j) + \sigma_P^2$	0 0 $\tau_P^2 + \sigma_P^2$

The formulae for the moments are simple only when the *Selection* equation is not multilevel. Otherwise the expressions are very complex (though a reasonably simple form can be derived when $n_j = 2$, e.g. panel data with two waves). Therefore, even with the linear two level random intercept model, simulation studies are needed to assess the bias caused by a multilevel selection mechanism.

Consequences of selection in the binary case

When the response of the *Principal* equation is binary, the corresponding model is a random intercept GLM. In such a case the situation is more complex but by exploiting the threshold representation of the binary model, some of the theoretical results on the linear model can be used. Denoting with $\breve{Y}^P$ the observable binary response of the *Principal* equation, a random intercept GLM is induced by assuming that $\breve{Y}^P$ is generated by a latent continuous response Y^P following model (1) through a threshold rule: $\breve{Y}^P$=1 if and only if Y^P>0. It follows that (Grilli and Rampichini, 2002):

$$P(\breve{Y}_{ij}^P = 1 \mid u_j^P) = \Phi\left(\mathbf{z}_{ij}^P \boldsymbol{\theta}^P \frac{1}{\sigma_P} - \frac{u_j^P}{\sigma_P} \right).$$

The estimable slopes are in σ_P units, so the slope bias has a *direct* component through $\boldsymbol{\theta}^P$ and an *indirect* component through σ_P. Depending on the selection mechanism, these two components might balance each other.

Note that similar arguments also hold when the response variable of the *Principal* equation is ordinal, since a threshold representation is possible as well.

Remedies to selection bias

In principle, the problems induced by selection can be circumvented by explicitly modelling the selection mechanism, thereby fitting a bivariate model such as (1), where the *Selection* equation is binary.

Many of the statistical packages for multilevel analysis can fit a bivariate model with at least one binary response, e.g.: ML estimates can be obtained using the `gllamm` command of Stata, the NLMIXED procedure of SAS, M-*plus* and aML; MCMC solutions are possible with *MLwiN* and Winbugs. The bivariate and multilevel nature of the model can sometimes lead to computational problems.

Moreover, the full modelling approach has several drawbacks such as weak identification (unless one can rely on instrumental variables); strong dependence of the estimates on distributional assumptions; low power of the tests for detecting non-ignorable selection.

A preliminary simulation study on the model with a binary response variable shows that ML estimation of a bivariate probit model effectively corrects for the bias, also when the errors have an asymmetric distribution. However, the LR test used to detect selection has very low power.

The problems connected with the full modelling approach have stimulated other approaches, such as semiparametric estimation (see Vella, 1998) and sensitivity analysis (Copas and Li, 1997), whose potentialities in the multilevel setting have still to be investigated.

Final remarks

Sample selection in multilevel models involves additional and somewhat unexpected problems: the hierarchical structure of the data is modified; as in single-level models, the slopes are biased, but the bias depends on many factors (cluster size, missingness pattern, covariates of other units); the conditional and marginal slopes are different even in the linear case; the variance-covariance structure changes, so the error terms are no longer homoscedastic, nor independent, and the ICC is biased.

However, there are important cases where some of the potential problems are irrelevant: (1) the slopes are unbiased when selection depends on unobserved factors only at the cluster level; (2) the conditional and marginal slopes in the linear model are equal (though biased) when selection depends on unobserved factors only at the elementary level.

Further theoretical work and simulation studies are needed: (a) to fully appreciate the consequences of selection bias in multilevel models, especially in models with a complex variance structure; (b) to develop and evaluate the performance of methods to correct for selection bias.

References

Arellano-Valle R.B. and Azzalini A. (2005). On the unification of families of skew-normal distributions, *mimeo*, submitted.

Bellio R. and Gori E. (2003). Impact evaluation of job training programmes: Selection bias in multilevel models, *Journal of Applied Statistics*, **30**, 893-907.

Borgoni R. and Billari F.C. (2002). A multilevel sample selection probit model with an application to contraceptive use, in: *Proceedings of the XLI meeting of the Italian Statistical Society*. Padova: CLEUP.

Copas B.J. and Li H.G. (1997). Inference for non-random samples (with discussion), *Journal of the Royal Statistical Society B*, **59**, 55-95.

De Fraine, B., Van Landeghem, G., Van Damme, J., & Onghena, P. (2005). An analysis of well-being in secondary school with multilevel growth curve models and multilevel multivariate models. *Quality & Quantity*, **39**, 297-316.

Grilli L. and Rampichini C. (2002). Specification issues in stratified variance component ordinal response models, *Statistical Modelling*, **2**, 251-264.

Grilli L. and Rampichini C. (2004). A polytomous response multilevel model with a non ignorable selection mechanism, in: *Proceedings of the 19th International Workshop on Statistical Modelling*. Firenze: FUP.

Heckman J. (1979). Sample selection bias as a specification error, *Econometrica*, **47**, 153-161.

Vella F. (1998). Estimating models with sample selection bias: A survey, *Journal of Human Resources*, **33**, 127–169.

Review of 'Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models'

Skrondal, A., and Rabe-Hesketh, S. (2004)

Boca Raton, Florida: Chapman and Hall CRC

ISBN: 1584880007 £54.99, pp +508.

David Bartholomew

Statisticians are at last waking up to the fact that latent variable models are not some esoteric (and suspect) preserve of certain social scientists but all of a piece with the traditional statistical approach. A statistical model is a statement about the joint distribution of a set of random variables. A latent variable model is one in which some of those random variables are unobserved (or unobservable). Alternatively, it may be thought of as an ordinary statistical model in which some of the variable values are missing. For these reasons latent variable models can arise in almost any statistical context. This makes them relevant, as the authors say to 'multi-level or generalized linear mixture models, longitudinal or panel models, item response or factor models, latent class or finite mixture models, and structural equation models'. It is a great virtue of this book that the authors emphasize that latent variables have a natural role in statistical modelling.

In their introduction, the authors put their finger on one of the key problems in opening up this field to statisticians. It is so well put that it is worth quoting in full. 'we strongly believe that progress is hampered by the use of 'local' jargon leading to compartmentalization. For instance, econometricians and biostatisticians are rarely seen browsing each other's journals. Even more surprising is tribalism within disciplines, as reflected by a lack of cross-referencing between item response theory and factor modeling in Psychometrics (even within the same journal!)'

.

Formally, then, there is nothing special about latent variable models. There are, of course, subtle questions which arise over what latent variables represent and what kind of inference one wishes to make about them. These questions need careful thought. It falls to statisticians, I think, to unify this vast field by presenting it within a common framework. The authors are not the first to do this but they have probably done it in the most comprehensive and practical way yet. It is to be fervently hoped that their pioneering work will be noticed and followed up.

A particular virtue of the book is the discussion of identifiability and equivalence in Chapter 5. The danger of adopting a comprehensive model is that it is tempting to multiply the number of variables and parameters well beyond what the data will bear. Quite simple examples have been around for a long time which show that it is actually very difficult to learn much about the distributions of latent variables without vast quantities of data. The authors at least recognize the dangers and direct readers in the direction of caution.

In spite of the book's many merits, it has certain limitations. It is a pity that the brief discussions at the ends of the chapters are not more complete. The authors have obviously tried to make sure that nothing was omitted but have not always found the space to give a considered view. The impressive list of references occupies 42 pages, but could still have been improved. There are comments on a wide variety of computer software programs but no mention is made, for example, of Moustaki's GENLAT program which has been available on the Chapman and Hall website for some years and is referenced in several of her publications. At the conceptual level, it might have been helpful to mention that the Response model, given here in Chapter 4, bears a marked similarity to the General Linear Latent Variable Model (GLLVM), which is the centrepiece of Bartholomew and Knott (1999). However, Skrondal and Rabe-Hesketh take matters farther by showing, for example, that multilevel models and survival modelling are included within the general framework. They do not note that the linear model is, in fact, a special case of a much more general class in which the linearity assumption is relaxed.

The very broad coverage, which is strength of this book, sometimes leads to an uncritical presentation. A good example is provided by Chapter 7, ‘Assigning values to latent variables’. The tendency is simply to summarise the extensive literature and offer all theory and methods without caveat. The dilemma the authors face is well illustrated by the remark at the beginning of section 6.10 (p.200) where they say ‘We depart from this interpretation (i.e. of latent variables as ***random variables***) and instead consider latent variables as unknown fixed parameters’. Does it really matter, one may ask, whether the latent variables are treated as variables or parameters? The reader would get no inkling that the issues raised here have been the subject of long and acrimonious debate among psychologists and that it may really matter how these issues are resolved. In short, the underlying ideas which, ultimately, really count are not always exposed to critical examination.

Seeing things within an abstract or more general framework does not necessarily make things more difficult - often the reverse. Sweeping away the peculiar customs and practices of the self-contained latent variable tribes will remove much useless lumber. In particular cases, however, the efficiency with which the special case can be handled outweighs the advantages conferred by generality and it remains to see where the balance of advantage will lie in this field. Another value of a unified approach is that it offers the possibility of a single software package with which the methodology can be applied. There is such a package, in which the authors are major players, known as GLLAMM. It can be downloaded free and experience will show whether its conceptual advantages are sufficient to displace the many special purpose programs now in use.

There is a wealth of material here, much of which will be new to statisticians and it is to be hoped that this book marks a further step towards making latent variable models a standard part of statistical practice and education. It is neither the first nor the last word on the subject but those who come after should certainly not ignore it.

Some Recent Publications using Multilevel Models

Gray, B. R., Haro, R. J., Rogala, J. T., and Sauer, J. S. (2005). Modeling fingernail clam (Family: Sphaeriidae) abundance-habitat associations at two spatial scales using hierarchical count models. *Freshwater Biology,* **50** (4): 715-729.

Leyland A. H., and Davies, C. A. (2005). Empirical Bayes methods for disease mapping. *Statistical Methods in Medical Research,* **14**: 17-34.

Leyland, A. H. (2005). Assessing the impact of mobility on health: implications for life course epidemiology. *Journal of Epidemiology and Community Health,* **59**: 90-91.

Manca, A., Rice, N., Sculpher, M. J. , and Briggs, A. H. (2005). Assessing generalisability by location in trial-based cost-effectiveness analysis: the use of multilevel models. *Health Economics*, **14** (5): 471-485.

Mok, M. M. C., Ma, H. S., Liu, Y. F., and So, Y. P. (2005). Multilevel analysis of primary students' perception and deployment of self-learning strategies. *Educational Psychology*, **25** (1): 129 - 148.

Priestley, G., Watson, W. E., Rashidian, A., Mozley, C., Russell, D., Wilson, J., Cope, J., Hart, D., Kay, D., Cowley, K., and Pateraki, J. (2004). Introducing critical care outreach: a ward-randomised trial of phased introduction in a general hospital. *Intensive Care Medicine*, **30** (7): 1398-1404.

Rabe-Hesketh, S., and Skrondal, A (2005). *Multilevel and Longitudinal Modeling using Stata*. College Station, TX: Stata Press.

Zaslavsky, A. M., Zaborski, L. S., and Cleary, P. D. (2004). Plan, geographical, and temporal variation of consumer assessments of ambulatory health care. *Health Services Research*, **39** (5): 1467-1486.

Please send us your new publications in multilevel modelling for inclusion in this section in future issues.