MULTILEVEL MODELLING NEWSLETTER

The Multilevel Models Project:

Dept. of Mathematics, Statistics & Computing
Institute of Education, University of London

20 Bedford Way, London WC1H 0AL, ENGLAND
e-mail: temsmya@ioe.ac.uk

Telephone: 071-612-6682 Fax: 071-612-6686

Vol. 5 No. 1 March, 1993

Workshops on Three-level Modelling: The free course of using the ML3 package, lectures, seminars, data analysis and group discussions on **18-21 May 1993** is still open for booking. The hands-on lecture will cover two level models, three level models, complex variance models, models for repeated measures data, multilevel binary response logistic models and multivariate models. The further workshop is scheduled for 2-5 November 1993. If you wish to book a place on one of them or to be on the waiting list for further workshops, please contact *Min Yang* at the project address.

Introductory booklets for ML3 macros: Three booklets of introduction to ML3 macros are available now from the Multilevel Models Project. They will handle multilevel binary response logistic models, multilevel time series models and multilevel multinomial response logistic models. With one or two worked examples in each booklet, they are designed to lead you through data manipulation, model specification and result interpretion. Please ask *Min Yang* for them.

Successful Grant Application: The Economic & Social Research Council (U.K.) has awarded a project grant to the Institute of Education to develop and apply multilevel modelling techniques to the handling of large and complex data. Further details will be given in the next newsletter.

ML3 Clinics in London 1993

Free for users of *ML3/ML3-E*

Tuesday May 4
Tuesday June 8
Tuesday July 13
Tuesday September 14
Tuesday October 12
Tuesday November 9
Tuesday December 7

10.30 am - 5.30 pm,
Multilevel Models Project
11 Woburn Square, Second flore
London WC1H 0SN
Call *Min Yang* for an appointment
Tel: 071 612 6682

Also In This Issue

Book Review: Hierarchical Linear Models by Anthony S. Bryk and Stephen W. Raudenbush.

Book Review: Combining information - statistical issues and opportunities for research.

Multilevel methods for estimation in surveys with complex sampling design.

Comparison between fixed effects and mixed effects logistic models in multicentre studies.

Multilevel models for comparing the effects of clinical treatments over time.

References to Multilevel Modelling: some recent articles.

The Multilevel Models Project is supported by a grant from the Economic and Social Research Council, U.K.

Book Reviews

COMBINING INFORMATION - Statistical Issues and Opportunities for Research. Washington DC: National Academy Press, Pp217. 1992. ISBM 0-309-04730-7

If a camel is a horse designed by a committee, this book is a dromedary - it is the product of the Panel on Statistical Issues and Opportunities for Research in the Combination of Information, of the Committee on Applied and Theoretical Statistics, of the Board on Mathematical Sciences, of the Commission on Physical Sciences, Mathematics and Applications, of the National Research Council, USA! Its topic field is extremely broad. As the report points out, forming a simple average is a form of combination of information (CI) - but the main theme is the expected one of forming a combined estimate of a single quantity from separate estimates obtained under different conditions.

A brief Chapter 1 gives three examples of topics which the report as a whole will deal with. The first is a simple meta-analysis of six clinical trials testing the effect of aspirin in preventing death following a heart attack. The second refers to the provision of agreed values of physical constants based upon numbers of experiments, and the third is the military problem of combining information from different detectors on potentially hostile targets. Methods suggested for further discussion include the use of fixed and random models and the combination of P-values using Fisher' s method.

Chapter 2 is titled 'What, why and when to combine'. The most interesting question is the third of these, and the report repeatedly stresses the role of judgment in deciding upon the comparability of different studies, whether or not this is expressed in formal Bayesian fashion. The uncertainty inherent in this type of judgment is seldom quantified and properly incorporated into the final conclusions of the combining study. Complementary to this is the emphasis laid upon the use of sensitivity analyses and validation exercises.

Chapter 3 provides examples of CI from no fewer than eleven different fields. These include the examples from Chapter 1 (treated in more detail) and others including education (a simple meta-analysis of studies on the effect of coaching on Standardised Attainment Scores), astronomy, oil exploration and image processing. Chapter 4 is devoted to statistical methodology. Topics discussed include homogeneity judgments, selection bias, robustness and record linkage as well as fixed and hierarchical modelling. Multiple comparison methods are treated at some length as tools for investigating homogeneity. Forecasting is also included though the link with CI is tenuous; the main issue is that of tackling uncertainty due to model choice. The combination of P-values is also discussed and its use as a CI technique is discouraged. Chapter 5 is an overview and summary.

The book contains much interesting material but its standard is uneven, with much space given to secondary topics and more important matters treated superficially. The 20-page bibliography is predominantly American and includes an infuriating number of unpublished reports and theses. Considering the importance of hierarchical models in CI, it is astounding to find that the term *multilevel* does not appear in the report, and that names such as Goldstein, Bock, Raudenbush, Bryk and Longford are absent from the bibliography. Linear modelling, let alone multilevel modelling, is barely discussed at all. One of the panel's conclusions is that 'a general purpose statistical computing package allowing investigators to routinely perform interactive Bayesian analyses in hierarchical models would gain immediate and widespread acceptance'. The existence of ML3, HLM and VARCL seems to have gone unnoticed.

M J R Healy

Hierarchical Linear Models: Applications and Data Analysis Methods, by A.S.Bryk and S.W.Raudenbush. Newbury Park: Sage, pp xvi + 265. ISBN 0-8039-4627-9

The pioneering work of Bryk and Raudenbush is well-known to anyone who uses multi-level models. Their new book is a clear introduction to the subject, or at least to its theory and principles, and I would be happy to recommend it to statisticians who want a thorough mathematical account of what multilevel modelling is about. Its failings are partly a product of this strength - it being far more useful for the mathematics than for applications - and partly an inevitable consequence of the rapid developments in the theory since their manuscript was completed.

What it does do well, it does excellently. It provides a carefully argued exposition of the mathematics, starting with the familiar heuristic explanation of multi-level models as formalisations of "slopes as outcomes" regressions, and proceeding to the full generality of the relevant matrix algebra. The general algebra is always kept close to particular instances, so that we can develop an intuitive feel for what the symbols mean. That is a rare accomplishment in a textbook of mathematical statistics.

Bryk and Raudenbush employ an explicitly Bayesian framework for inference, which is probably more cogent for explanation than the alternatives. It allows us to think clearly about the crucial assumptions - for example, that a school is an instance (or a case) of a population of schools, from which we can "borrow strength" to make the inferences about that particular school more reliable. On the whole, the algebra in this Bayesian framework is identical to that which would be used in a frequentist approach, although a reader who is familiar with, say, Goldstein's work should note some different conventions in the terminology - for example, the slightly different meanings of the terms "fixed" and "random".

There are, unavoidably, gaps in the mathematics, because the subject has been changing rapidly. For example, it is now possible to do multi-level loglinear modelling, but this would, at best, have been only a theoretical possibility when the manuscript was completed. However, points like this are not serious indictments: a textbook in a changing subject is bound to seem out of date when it is published, but that does not detract from its main purpose, which is exposition of general principles. Exactly the same point can be made about Goldstein's 1987 book (Multilevel Models): it too has been superseded, but remains valuable. I should add, moreover, that Bryk and Raudenbush are generally more comprehensive in their treatment of the mathematics than Goldstein was.

Thus, as a mathematical text, the book is unrivalled (although afflicted by the unaccountable absence of an index). Its main limitations have to do with the practice. It does present some thoroughly developed examples, drawn from research in education. And the authors are skilled at linking the arithmetical details to the mathematical theory. So, for teaching on a fairly theoretical course, the book would be useful. But it does fall into a familiar problem with statistical textbooks - of tending to treat data analysis as an "application" of theory. The rather more chaotic informal rules of actual educational (or psychological or geographical or medical) science are then ignored.

This is probably not a view I would have taken a few years ago. But one of the effects of moving from even a very applied statistics department to a social-science department is to find that the ways in which statistical methods are used are not those which we, as mathematical theorists, would prescribe. Sometimes, of course, these uses are downright wrong, and books of this type are invaluable for correcting that. For example, in what is their most innovative chapter, Bryk and Raudenbush discuss various ways of judging whether a multi-level model adequately fits the data, and what to do if we conclude that it does not. This kind of theory is the point where the elegance of the general models begins to engage with reality.

But the problems have more to do with the conventions which grow up within disciplines as to what counts as an explanation. Take, for example, the crucial assumption mentioned above - that individual schools are "cases" of a general population. That this is controversial is becoming acutely obvious in the current debate about performance indicators, which Bryk and Raudenbush suggest can be estimated by empirical-Bayes residuals at the school level. The problem is to know whether a particular outlying residual is a consequence of miss-specification of the model, or of a genuinely highly effective (or ineffective) school. Even if the residuals perfectly fit a Normal distribution, the 5% of "outliers" might be, not outliers at all, but schoools with unique impacts on their pupils. There is, quite simply, no mathematical way of resolving this dilemma: the only way of doing that is by including detailed data on the school's practices, but in routine monitoring that might not be available. To collect that data in a sensible way requires explicit educational theories about how schools have their effects, and so the statistical consideratons have to be embedded in educational ones.

This is where Goldstein"s book was, and remains, the best available. But neither it nor the one by Bryk and Raudenbush can keep up with new developments in the use of multi-level models. Significant examples are now available from geography (especially on labour markets), medicine, and political science, although some of these are not yet in the public domain. In a few years' time, therefore, it will be possible to write a quite different book contrasting and comparing the emerging conventions surrounding the use of multi-level models in different disciplines. An analogy can be drawn with log-linear modelling. There are now several books that do take applications as their main focus - for example, Nigel Gilbert's Modelling Society. Such books do not supersede the key mathematical texts by, for example, McCullagh and Nelder or Bishop, Fienberg and Holland. But they do base their exposition on social science, not mathematical theory.

So this book by Bryk and Raudenbush will remain valuable, especially for statisticians wanting a clear exposition of the mathematical theory. For synoptic accounts of the practice we must, inevitably, wait a little longer.

Lindsay Paterson

Some Recent References to Multilevel Modelling

Albandar, J.M. & Goldstein, H. (1992) Multi-Level Statistical Models in Studies of Periodontal Diseases. J Periodontal, 63, 690-695.

Garner, C.L., & Raudenbush, S.W. (1991) Neighborhood effects on educational attainment: A multilevel analysis. *Sociology of Education*, 64, 4, 251-262.

Hoesksma, J.B. & Koomen, H.M.Y. (1992) Multilevel Models in Developmental Psychological Research: Rationales and Applications. In Early Development and Parenting, Vol. 1(3) 157-167, New York: Wiley.

Raudenbush, S.W. (in press). Hierarchical linear models as generalizitions of certain common experimental design models. In Edwards, L.(Ed). *Applied analysis of Variance in Behavioral Science.* New York: Marcell Decker.

Raudenbush, S.W. (in press). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. To appear in the *Journal of Educational Statistics.*

Raudenbush, S.W. and Chan, W.S. (1992) Growth curve analysis in accelerated longitudinal designs with application to the National Youth Survey. *Journal of Research on Crime and Delinquency*, 29 4 387-411.

Raudenbush, S.W. and Chan, W.S. (in press). Application of a hierarchical linear model to the study of adolescent deviance in an overlapping cohort design. to appear in the *Journal of Clinical and Consulting psychology.*

Raudenbush, S.W., rowan, B., and Kang, S.J. (1991). A multilevel, multivariate model for studying school climate in secondary schools with estimation via the EM algorithm. *Journal of Educational Statistics*, 16 4 295-330.

Rowan, R., Raudenbush, S.W. & Kang, S.J. (1991). Organizational design in high schools: A multilevel analysis. *American Journal of Education*, 99 2 238-266.

****

PLEASE SEND US ANY MULTILEVEL MODELLING PUBLICATIONS FOR INCLUSION IN THIS SECTION IN FUTURE ISSUES.

****

APPLICATIONS

Multilevel methods for estimation in surveys with complex sampling design

N.T.Longford
Educational Testing Service, Princeton, NJ

Many large scale surveys employ stratified clustered designs with unequal sampling weights. Stratification is an important device for improved estimation of population characteristics, effective especially in well-researched populations. Clustering promotes cost efficiency and organizational manageability. Human subjects and educational (and other) institutions cannot be coerced to respond in a survey. The resulting non-response is likely to be informative; certain kinds of institutions and/or individuals are more (less) likely to abstain from responding. The bias due to such informative missingness can be reduced by *adjustment* of weights. In such a procedure the original sampling weights (proportional to the reciprocal of the probability of inclusion in the sample assuming no non-response), set by the design, are adjusted, usually by a set of multiplicative factors, so that the resulting weights are closer to being proportional to the reciprocals of the 'true' probabilities of inclusion (that take non-response into account).

Assessment of the quality of weight adjustment is often difficult. The weight adjustment depends on the sample drawn, and therefore the adjusted weights are *random* variables. Standard methods for estimation of population means in surveys ignore the stochastic nature of the sampling weights, and use a more or less arbitrary normalization of the weights. For instance, the weights are scaled so that their total is equal to the sample size.

Longford (1992) applied multilevel models in the context of the National Assessment of Educational Progress (NAEP), an on-going programme of surveys of the U.S. primary and secondary education. The main outcome of the analysis of each survey, as mandated by the contractor, the U.S. Department of Education, are tables of estimated population and subpopulation means of proficiencies in a number of academic subjects, and the associated estimated standard errors. For example, the subpopulations of interest are defined by ethnicity, region, gender, school-type, responses to attitudinal and experiential questionnaire items, and the like. Mathematics, English, History, and Geography are some of the academic subjects. I abstain from discussion of utility of such tables but consider the situation in which the means of a variable are estimated for a large number of subpopulations.

In brief, an effective analysis of the survey has to accomodate the following features:

- stratification
- clustering
- unequal sampling weights
- 'estimated' sampling weights (weight adjustment).

The principal outcome variable, the proficiency score, is itself estimated from the responses to cognitive items. The proficiency score is represented by five draws (called feasible values) from the posterior distribution of the proficiency. Thus each estimate is averaged over the five feasible values, and the estimated standard errors are adjusted similarly.

This article, based on my report (Longford, 1993), describes three generic approaches to estimation in surveys with stratification, clustering, and unequal sampling weights, and presents a framework for assessing importance of the stochastic nature of the weight adjustment.

Let Y_i, $i = 1,2,...,N$, be the values of the variable of interest for the target population. The population mean is defined as $\overline{Y} = \Sigma_{i=1}^{N} Y_i/N$. The values of the variable for the subjects included in the survey are denoted by y_{ijk} for student $i = 1,...,n_{jk}$ in school $j = 1,..., m_k$ in the stratum $k = 1,...,K$ (stratification is applied to schools). The corresponding original and adjusted weights are denoted by $w_{0,ijk}$ and $w_{a,ijk}$. The ratio estimator

$$\hat{\mu} = \sum_{ijk} w_{a,ijk} y_{ijk} / \sum_{ijk} w_{a,ijk} \qquad (1)$$

is commonly used for estimating the population mean. Three methods for estimating the sampling variance of this estimator are discussed: the jackknife, a variance component (VC) method, and an

ANOVA-like method. A stratified single-stage clustered sampling design with unequal weights is assumed throughout.

Jackknife

A typical survey involves a large number of strata, e.g., $K = 56$, with up to three schools (clusters) in each stratum. For each stratum k a *pseudo-sample* is defined by copying the values of y and w in all strata except k, and:

- if stratum k contains not more than one school, the values for the stratum are copied also
- if the stratum contains two schools, the second school is replaced by the first
- if the stratum contains three schools, the third school is replaced by the first two, with halved weights.

Each pseudo-sample is again subjected to weight adjustment.

Let $\hat{\mu}^{(k)}$ be the estimator of $\overline{Y}$, as in (1), based on the pseudo-sample k. The jackknife estimator is defined as the mean of the *pseudo-estimators* $\hat{\mu}^{(k)}$, $\hat{\mu}^{(J)} = \Sigma_k\hat{\mu}^{(k)}/K$, and its sampling variance is estimated by

$$\hat{\sigma}_J^2 = \sum_k (\hat{\mu}^{(k)} - \hat{\mu})^2.$$

See Wolter (1985) and Johnson and Rust (1992) for details.

Variance components

We assume the model

$$y_{ijk} = \mu_k + \delta_{jk} + \varepsilon_{ijk} \tag{2}$$

where $\delta_{jk} \sim N(0,\sigma_B^2)$ and $\varepsilon_{ijk} \sim N(0,\sigma_{W,jk}^2)$, and all these random variables are mutually independent. Submodels, e.g.,assuming some or all the variances $\sigma_{,jk}^2$ to be equal, can be considered. The stratum means μ_k are unknown constants.

If the sampling weights were equal the variance component model in (2) (with the standard assumptions of normality) could be fitted by EM, IGLS, or Fisher-scoring methods. In straightforward adaptations of these methods the various crossproducts are replaced by their weighted versions; this is proposed in Longford (1992). The model in (2) involves a large number of parameters. The simplifying assumption that the stratum means μ_k be regarded for a random sample is not appropriate, as demonstrated by Longford (1992). Also, once the estimated means $\hat{\mu}_k$ are obtained, it is not immediately clear how to combine them to form the estimate of the population mean $\hat{\mu}$.

Longford (1992) compared the jackknife and the VC estimators of the population and subpopulation means, concluding that there is little to choose between them, but the VC estimator of the sampling variance is much more efficient than its jackknife counterpart. The VC method is iterative -- this may be perceived as a distinct disadvantage when a large number of subpopulations is considered.

An ANOVA-like method

We adopt the estimator (1), and evaluate its sampling variance assuming
the model (2). Elementary algebra leads to the identity

$$var(\hat{\mu}) = \frac{\sigma_B^2}{n_B^2} + \frac{1}{n_B}\frac{\Sigma_{jk}\sigma_{W,jk}^2\Sigma_i w_{a,ijk}^2}{\Sigma_{jk}(\Sigma_i w_{a,ijk})^2} \tag{3}$$

where $n_B = (\Sigma_{ijk}w_{a,ijk})^2/\Sigma_{jk}(\Sigma_i w_{a,ijk})^2$. An estimate of this sampling variance is obtained by substituting estimates of the variances σ_B^2 and $\sigma_{W,jk}^2$ in (3). These variances can be obtained by moment matching.

Let $w_{A,ijk} = \frac{\Sigma_i w_{a,ijk}}{\Sigma_i w_{a,ijk}^2} w_{a,ijk}$ and $n_{A,jk} = \Sigma_i w_{A,ijk}$. Note that also $\Sigma_i w_{A,ijk}^2 = n_{A,jk}$. Potthoff, Woodbury, and Manton (1992), whose approach is adopted here, refer to $w_{A,ijk}$ as the *normalized weights* and to $n_{A,jk}$ as the *effective sample size.*

The within-school variances $\sigma_{W,jk}^2$ can be estimated as the corrected weighted within-school sums of squares

$$v_{A,jk} = \frac{1}{n_{A,jk} - 1}\sum_i w_{A,ijk}(y_{ijk} - \overline{y}_{jk})^2 ,$$

where $\overline{y}_{jk}$ is the weighted (sample) mean of y in school jk. It is an unbiased estimator of $\sigma_{W,jk}^2$ and its distribution is approximately χ^2 with $n_{A,jk} - 1$ degrees of freedom. This is the rationale for definition of $w_{A,ijk}$ and $n_{A,jk}$.

Estimation of the between-school variance is based on the statistic

$$v_B = \sum_{jk} u_{jk}(\bar{y}_{jk} - \bar{y}_k)^2, \qquad (4)$$

where u_{jk} are suitable constants (e.g., functions of the weights $w_{A,ijk}$, such as $u_{jk} = \Sigma_i w_{ijk}$). Moment matching yields the estimator

$$\hat{\sigma}_B^2 = \frac{v_B - \Sigma_{jk} u_{jk} U_{jk} \hat{\sigma}^2_{W,jk}/n_{A,jk}}{\Sigma_{jk} u_{jk} U_{jk}}, \qquad (5)$$

where

$$U_{jk} = 1 - 2\frac{\Sigma_i w_{a,ijk}}{\Sigma_{ij} w_{a,ijk}} + \frac{\Sigma_j(\Sigma_i w_{ijk})^2}{(\Sigma_{ij} w_{a,ijk})^2}.$$

Properties of the three types of estimators described above were explored by a simulation study considering a variety of nesting designs. In general, the jackknife is much less efficient than the VC methods, and it does not deliver on the promise of unbiasedness. Some care is required in choosing the constants u_{jk} in (4). The REML estimator, using the normalized weights has the least bias, but is less efficient than the ANOVA-like method with $u_{jk} = \Sigma_i w_{a,ijk}$.

Weight adjustment

Several exploratory analyses document that the weights 'matter'; analyses using equal weights lead to conclusions substantially different from those using adjusted weights. Since the weights are important every effort should be made to adjust the weights as appropriately as possible. We consider the following model for the weight adjustment:

$$\log(w_{a,ijk}) = \log(w_{ijk}) + \delta_{w,jk} + \varepsilon_{w,ijk,} \qquad (6)$$

where w_{ijk} are the 'true' weights (exactly proportional to the reciprocal probabilities of inclusion), and $\delta_{w,jk}$ and $\varepsilon_{w,ijk}$ are two mutually independent random samples, with zero means and variances σ_δ^2 and σ_ε^2. These weight-variances are not known because the weights $w_{A,ijk}$ are realized only once. However, an idea of their size can be gained by considering the log-adjustment, $\log(w_{a,ijk}/w_{o,ijk})$ as having two components: adjustment from the original design weights to the true sampling weights, and the noise ('error' in adjustment). Presumably the error variance is smaller than the variance of the logs of the adjustment factors, otherwise the adjustment is very ineffective. This consideration can be applied separately to the school- and student-level variances, σ_δ^2 and σ_ε^2. These variances can be used in a Monte Carlo study (perturbing the adjusted weights) to assess the impact of uncertainty about the weight adjustment. The most likely outcome is that the stochastic nature of weight adjustment can be ignored, but the weight adjustment itself is not negligible.

Extensions

The ANOVA approach for two stages of clustering is described in Longford (1993). The approach readily extends to estimation of regression and variance parameters. First, the population quantity, such as a regression parameter, is defined. It is a function of a small number of population summaries, e.g., totals of crossproducts. Then these crossproducts are estimated by their sample estimators. Standard errors can be derived by the delta method which can also be used for bias correction. An important advantage of this approach is that distinction is made between the uncertainty due to representation of the population by the sample, and the imperfect fit of an underlying model in the population.

References

Johnson, E.G., and Rust, K. (1992) Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics* **17**, 175--190.

Longford, N.T. (1990) Standard errors for the means of proficiencies in NAEP: Jackknife versus variance components. *UCLA Statistics Series* **47**, Los Angeles, CA.

Longford, N.T. (1992) Comparison of efficiency of jackknife and variance component estimators of standard errors. *ETS Technical Report* RR-92-24. Princeton, NJ.

Longford, N.T. (1993) Model-based methods for analysis of data from NAEP Trial State Assessment. Manuscript in preparation.

Potthoff, R.F., Woodbury, M.A., and Manton. K.G. (1992) "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association* **87**, 383--396.

Wolter, K. (1985) *Introduction to Variance Estimation.* John Wiley and Sons, New York, NY.

APPLICATIONS

A comparison between fixed effects and mixed effects logistic models in multicentre studies

H.Poirel[1], *S.Richardson*[1], *M.Chavance*[2], *T.Moreau*[2], *M.Busson*[3]

1.I.N.S.E.R.M. U170 16, avenue Paul Vaillant-Couturier 94807 Villejuif cedex, France.
2.I.N.S.E.R.M. U169 16, avenue Paul Vaillant-Couturier 94807 Villejuif cedex, France.
3.I.N.S.E.R.M. U93 16, Hôpital Saint-Louis-Centre Hayem 2, Place du docteur Fournier 75010 Paris, France.

Multicentre studies are being increasingly used in clinical medicine and epidemiology in order to gain more power. It is then necessary to include a centre effect in the analysis, either to measure it or to avoid a possible bias when estimating the other effects. Two types of models may be considered. In fixed effects models, the centres are viewed as a population of centres, and so, the number of parameters increases with the number of centres. In large cooperative studies, this inflates substantially the number of parameters to be estimated. In mixed effects models, the studied centres are viewed as a sample of all interesting (or potential) centres, as suggested by Gilks (1987). The random effects which characterize each centre are assumed to be i.i.d., generally $N(0, \sigma^2)$. Thus only one parameter needs to be specified, the variance σ^2.

Population

One-year graft survival was registered for 3,530 patients who received a kidney transplant between 1/1/1985 and 3/1/1990 in 17 of the 46 french transplantation centres. Numbers of patients by centre ranged from 46 to 591. Eight prognosis factors, listed in table 1, were available for each patient.

Models

Let Y_{ij} be the binary outcome observed for patient i in centre j, and X_{ij}, the p-dimensional vector of the fixed effects prognosis factors for this patient. We assume Y_{ij} to be Bernoulli (Binomial) (π_{ij}), with $logit(\pi_{ij})$ a linear function of these factors.

1) Fixed effects logistic model:

centre effects are considered as fixed, and we set

$logit(\pi_{ij}) = \beta_0 + \beta^T X_{ij} + \gamma_j$

where β and $\{\gamma_j\}$ are vectors of unknown coefficients. We choose the largest centre as a reference with corresponding γ_j equal to 0. The maximum likelihood estimates were obtained using the $LR^{\alpha R}$ program from the BMDP statistical software.

2) Mixed effects logistic model:

Centre effects are considered as random and we set

$logit(\pi_{ij}) = \beta_0 + \beta^T X_{ij} + u_j$

with the u_j i.i.d. $N(0, \sigma^2)$. The parameters β_0, β, σ^2 are estimated by the Iterative generalized Least Square (IGLS) method, as proposed by Goldstein (1988, 1991), using the statistical software ML3 (Prosser, Rasbash and Goldstein 1991). After a linearisation of the fixed part of the model, this algorithm estimates alternatively: a) the fixed effects parameters conditionally on the last estimate of σ^2 (taken as 0 for the first step), and b) the random effect variance conditionally on the last estimate of the fixed effects. Once all the parameters of the model are known, posterior estimates of the random effects u_j can be obtained from the observed residuals and the covariance structure of the data postulated by the model. The random effect for centre j can be interpreted as the averaged residuals for centre u_j, weighted appropriately.

Results

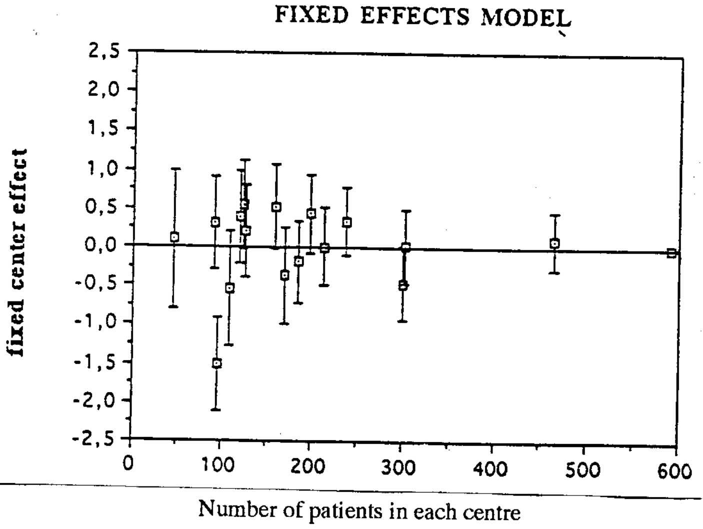
As can be seen in Table 1, estimates of the odds-ratios corresponding to the prognosis factors and their confidence intervals were similar for both models.

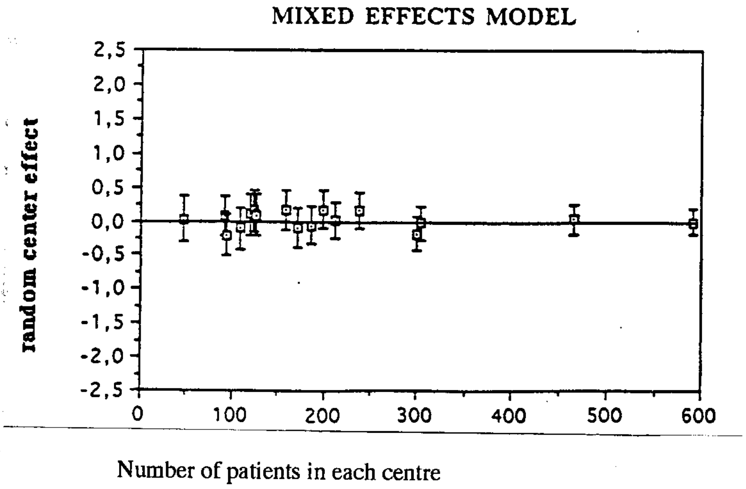
According to the mixed effects model, the inter-centre variance was 0.032 with standard-deviation of 0.027. The posterior estimates of the random centre effects ranged from 0.79 to 1.17 while the estimates of the fixed centre effects range from 0.22 to 1.77. This illustrates the usual feature of 'shrinkage' for such moderate estimates. In the Figure these estimates are represented according to the number of patients in each centre. In the fixed effects model, estimates for small centres are more widely spread than large ones. In the mixed effects model, estimates for small centres are attracted

Table 1. Comparison of estimated relative risks between fixed effects and mixed effects model

Variables	Fixed effects model			Mixed effects model		
	e^{β}	*95%CI*		e^{β}	*95%CI*	
Fixed effects for covariates						
constant	0.07	(0.04,0.13)	***	0.08	(0.04,0.16)	***
male recipient with female donor	2.07	(1.31,3.25)	**	2.05	(1.30,3.25)	***
antibodies>25% panel cells	1.50	(1.15,1.96)	**	1.52	(1.17,1.97)	***
donor age ≥ 55(years)	1.75	(1.13,2.72)	**	1.73	(1.12,2.69)	**
donor age <15 (years)	1.09	(0.72,1.65)		1.09	(0.72,1.67)	
year of graft linear effect	1.32	(1.21,1.44)	***	1.31	(1.21,1.42)	***
year of graft quadratic effect	1.13	(1.08,1.19)	***	1.13	(1.08,1.17)	***
number of HLA identities	0.91	(0.83,1.01)		0.92	(0.83,1.00)	
cold ischemia ≥24h	1.20	(0.94,1.54)		1.08	(0.85,1.38)	
dialysis ≤5 years	1.28	(0.92,1.80)		1.28	(0.92,1.80)	
pregnancies ≥3	1.19	(0.82,1.71)		1.17	(0.82,1.68)	
Random effect						
inter-centre variance: σ^2				0.032	(0.027)	

* p<0.05   ** p<0.01   *** p<0.0001





towards 0, while for large centres, estimates for both models are similar.

We also considered random coefficient models where the effect of a prognosis factor could fluctuate randomly from one centre to another. Estimates were imprecise, even when such an interaction was introduced for only one factor, probably because of the relatively small number of centres.

Discussion

In this paper, we have compared two different ways of taking into account the centre effects in multicentre studies, considering them either as fixed effects or as a random effect. We have been able to use both methods only because number of centres was moderate. When the number of fixed parameters become too large, the fixed effects model is impractical. Nevertheless, with only 17 centres, the estimate of the variance of random effects cannot be very precise. Further, we note that there are other variables which would be necessary to study before attempting to interpret the centre effect in these data.

References

Gilks W.R. (1987) Some application in hierarchical models in kidney transplantation. *The Statistician*, 36: 127.

Goldstein H. (1986) Multilevel mixed linear model analysis using iterative generalized least squares estimation. *Biometrika*, 76: 622.

Goldstein H. (1991) Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, 78: 45.

Prosser R., Rasbash J. & Goldstein H.(1991) ML3 Software for Three-level analysis: User's Guide for V.2. Institute of Education, University of London.

An application of multilevel models to compare the effects of clinical treatments over time

Xu Yong-yong Department of Health Statistics, The Fourth Military Medical University, Xi'an, P.R.China

It is very common in clinical practice to record symptoms or other examination results longitudinally in order to assess a therapeutic effect over time. This generates repeated measurement data with time or occasion being nested within each individual patient. In particular, patients are often divided into groups, each group receiving a different treatment. It is important to compare treatment effects among groups by taking the time trend into account. Multilevel models for repeated measures reflect the data structure most adequately. In this article, both the fixed occasion and basic multilevel models are used to compare the treatment effects among three patient groups of drug addicts in a special hospital in north China.

Data

The data consist of 179 drug addicts allocated randomly into three groups given different medicines, artificial hibernation (AH), Diazepam plus Clonidine (DPC) and Diazepam only (DON). The AH serves as a control. A score reflecting the relief of symptoms was recorded at the day before treatment and at days 1, 2 and 3 after treatment. The score ranges from 0 to 10, the larger it is, the less relief of symptoms. Table 1 gives the means and SDs of scores for each group by measurement occasion.

Table 1 The mean scores (SD)

GRP	N	Before tr.	1 day aft.	2 days aft.	3 days aft.
AH	59	9.9 (0.5)	9.1 (1.5)	7.5 (1.6)	6.0 (1.8)
DPC	60	10.0(0.0)	7.8 (1.1)	5.2 (1.2)	2.3 (1.2)
DON	60	10.0(0.0)	7.6 (1.3)	5.3 (1.5)	4.5 (1.6)

It can be seen from the table that there are few differences between groups before treatment, and that scores consistently decrease over time after treatment for all groups. The questions to be answered on the treatment effects are: 1) whether the three treatments have same effect in general, and 2) whether their effects significantly differ at each measurement occasion, and 3) which treatment relieves symptoms more quickly in the first three days.

Models

First, the two-level fixed occasion model is considered,

$$y_{ij} = \beta_{ij} t_{ij} \qquad (1)$$

where $i = 0, 1, 2, 3$ indicates time occasion at level 1, $j = 1, 2, \ldots, 179$ indicates patients at level 2, t_{ij} is a dummy variable at occasion i for patient j. β_{ij} are occasion effects, and can be expressed as an average effect of occasion plus some level 2 covariates together with random variation as follows:

$$\beta_{0j} = \beta_0 + u_{0j}$$
$$\beta_{1j} = \beta_1 + d_{11}x_{1j} + d_{12}x_{2j} + u_{1j}$$
$$\beta_{2j} = \beta_2 + d_{21}x_{1j} + d_{22}x_{2j} + u_{2j}$$
$$\beta_{3j} = \beta_3 + d_{31}x_{1j} + d_{32}x_{2j} + u_{3j}$$

where x_{1j}, x_{2j} are dummy variables for DPC and DON respectively. They are level 2 covariates. β_0 is the grand mean that is the effect before treatment, β_1, β_2, β_3 are mean effects of occasions 1, 2 and 3 after treatment for group AH. The d_{i1} are effects of DPC and d_{i2} are those of DON at the *ith* occasion compared to the AH group. The u_{ij} are random residuals across patients at occasions. As level 1 occasion affects are estimated in the fixed part, there are no level 1 random terms in the model.

In the second model, we treat t_{ij} in (1) as continuous, and the following model is used to fit the data,

$$y_{ij} = \beta_0 + \beta_{1j} t_{ij} + e_{ij} \qquad (2)$$
$$\beta_{1j} = \beta_1 + \gamma_1 x_{1j} + \gamma_2 x_2 + u_{1j}$$

where β_0 is a before treatment mean, β_1 is the mean slope on time, γ_1, γ_2 covariate differences of slopes between groups DPC and AH, DON and AH respectively. The u_{1j} are residuals of slopes across patients and the e_{ij} are level 1 residuals.

Thus 4 parameters in the fixed part will be estimated. Two random parameters $\sigma_e^2, \sigma_{u_1}^2$ are estimated. All the slopes are assumed to pass through the origin β_0.

Results and discussion

Table 2 gives estimates of the fixed part of model (1). The answer to question (1) is obtained straight away by comparing estimates of d_{ij} to their own SEs. They are all significant, which indicates that the effects of the three treatments are not the same, both treatments DPC and DON have better effects than treatment AH at days 1, 2 and 3.

Table 2 Estimates in fixed part of model (1)

Parameter	Estimate	S.E.
β_0	9.978	0.022
β_1	9.132	0.165
β_2	7.520	0.188
β_3	6.034	0.199
d_{11}	-1.338	0.232
d_{12}	-1.571	0.232
d_{21}	-2.351	0.264
d_{22}	-2.217	0.264
d_{31}	-3.701	0.280
d_{32}	-1.568	0.280

To find whether DPC and DON have the same effect, hypotheses $d_{11} = d_{12}, d_{21} = d_{22}$ and $d_{31} = d_{32}$ are tested. This gives us three χ^2s with 1 d.f. each as 1.02, 0.26 and 58.72 respectively, indicating that DPC and DON have the same effect during the first two days after treatment, but DPC relieves symptoms by 1.5 score points more than DON after the third day's treatment.

Table 3 Random parameter estimates: covariance (correlation) matrix

	u_{0j}	u_{1j}	u_{2j}	u_{3j}
u_{0j}	.089(1.0)			
u_{1j}	.025(.07)	1.61(1.0)		
u_{2j}	-.011(-.02)	.442(.24)	2.08(1.0)	
u_{3j}	.001(.00)	-.191(-.10)	.677(.31)	2.32(1.0)

From the random parameter estimates at the patient level in Table 3, correlation coefficients of score between days 1 and 2, 2 and 3 were calculated as 0.24 and 0.31 respectively.

Fitting model (2), with estimates in Table 4, it is seen that three treatments reduce the symptom scores by 1.3 a day in average, and treatment DPC relieves symptoms quicker than both AH and DON by 1.2 and 0.48 score units per day respectively. Treatment DON does better than AH by 0.76 units. This answers question (3).

Table 4 Fixed and random parameter estimates of model (2)

	Parameters	Estimates	S.E.
Fixed	β_0	9.992	0.081
	β_1	-1.271	0.056
	γ_1	-1.225	0.066
	γ_2	-0.741	0.066
Random	σ_u^2	0.137	0.059
	σ_e^2	1.490	0.091

Analyses based on both models showed that treatments DPC and DON have better effects in relieving the symptoms than AH at each time occasion, and DPC does better than DON in terms of the change over time.

To fit model (1), our interest was in the main effect estimates at each time occasion and two covariates were introduced at level 2 to study group differences at each time occasion. The correlations between occasions are provided by the random parameter estimates at level 2. This model served our aims satisfactorily because the data have only 4 time occasions. It would be difficult to use when the number of time occasions is large, say 13, since the full number to be estimated at level 2 would be 78. In this case, model (2) would be a better choice. With this model more explanatory variables, such as age of patient and sex, can be easily introduced into the fixed part.

With fewer parameters to be estimated, model (2) is easier to explain in practice and describes the relationship between the treatment effects and time more clearly. In particular, when scores before treatment are different among three groups of treatment, covariates can be introduced in the model to make the adjustment. If the decrease of scores over time is not linear, a polynomial or other growth curve can be used. Again we can use other covariates such as age and sex in the model to make more precise comparisons between treatment groups.

& & & & & & & & & & & & & & & & & & &

CONTRIBUTORS

We are grateful to the people who provided articles for this issue.

Healy, M.J.R.	Institute of Education, University of London, London, U.K.
Longford, N.T.	Educational Testing Service, Princeton, NJ, USA
Paterson, L.	Centre for Educational Sociology, Edinburgh University, U.K.
Poirel, H.	INSERM U170 16, Avenue Paul Vaillant-Couturier 94807 Villejuif cedex, France
Raudenbush, S.W.	College of Education, Michigan State University, East Lansing, Michigan 48824-1034, USA
Xu, Y.Y.	Department of Health statistics, The Fourth Minitary Medical University, Xi'an, China

& && && && && && && && && && && && &