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
Web site http://www.ioe.ac.uk/hgoldstn/home.html
Tel: +44 171 612 6682 Fax: +44 171 612 6686

Vol. 7 No. 2 **June, 1995**

Workshops & Courses

Workshop in Norwich: A three-day workshop to be conducted by Dr *Ian Langford* and Professor *Harvey Goldstein* will be held at University of East Anglia (UEA) from 30th August to 1st September 1995. Using *MLn,* the general introductory workshop will cover topics such as basic principles, setting up two and higher level models, repeated measures, logistic models, multivariate analysis and diagnostics. Course fees for academics and non-academics respectively are £350 and £600 inclusive. For further information please contact *Anne-Lise McDonald* at *Health Policy and Practice Unit, UEA, Norwich, NR4 7TJ. Tel 01603 593631, email a.cox@uea.ac.uk.*

Workshop in Glasgow: A workshop on the use of multilevel modelling in Public Health and Health Services Research will be held at the University of Glasgow in 27-29 September 1995. This workshop will give participants the chance to analyse personal data sets using *MLn* as well as following worked examples introducing a variety of models applied in the health field. Course fees for academics and non-academics respectively are £350 and £600 inclusive. Further details are available from *Dr. Alastair Leyland, public Health Research Unit, University of Glasgow, 1 Lilybank Gardens, Glasgow G12 8RZ. Tel: 0141 339 3118 E-mail: a.leyland@udcf.gla.ac.uk.*

Workshop in London: The Multilevel Models Project at the Institute of Education in University of London will run another general workshop using *MLn* in 10-12 October 1995. Two class groups will be formed to suit both experienced participants and beginers. The workshop fee is £300 for academics and £600 for non-academics inclusively. For more details or booking please contact *Min Yang* at the project address.

New World Wide Web site for the Multilevel Models Project has been set up on the Institute of Education Web server. It includes the following:

- An introduction to multilevel models and some application fields
- Some current project activities
- Details of workshops, course and 'clinics'
- Recent issues of the MM Newsletter and example data sets in compressed form for downloading
- *MLn* MACRO files for downloading - currently those for time series and non-linear variance modelling and enhanced loglinear and logistic linear modelling macros
- A description of *MLn* and an order form and other news about releases etc.
- A regularly updated list of known bugs in *MLn*
- Links to other relevant Web sites

The site address for public access is ***http://www.ioe.ac.uk/hgoldstn/home.html.*** Please try this out & give us your comments.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Also In This Issue

aLCD analysis OF LARGE AND COMPLEX DATASETS

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Bookreviews

Applied Multilevel Analysis by *J.J. Hox, TT-Publikaties, Amsterdam, 1994, pp112, NLG 25 or $15, ISBN 90 801073 2 8*

As multilevel modelling becomes more widely used, there is clearly a need for texts which explain why the models are important and how they work, without going into too many technical details. Hox's short book sets out to meet this need for social scientists. The introductory chapter is followed by chapters on multilevel regression models, working with different multilevel packages, applications of multilevel modelling to meta analysis as a way of introducing models for proportions, and finally a chapter on multilevel structural equation modelling. All the exposition is restricted to two level models.

The selection and ordering of the material is a little curious at times. For example, the potentially confusing topic of centering is introduced very early in the first chapter, whereas the important notion of 'shrinkage' of level two residuals is not discussed at all. Also, the chapter on multilevel structural equation modelling is perhaps a little difficult for beginners.

On the whole, the methods are presented accurately. However, it is not true to say (p.14) that the level two variance necessarily rises with the value of the explanatory variable when slopes are random (it depends on the sign and size of the covariance term). And it is generally not advisable (p.17) to judge the importance of random effects by comparing their values with their standard errors (examining differences between the likelihood statistics are more appropriate).

Supplied with the book, there is a disk which contains the example data and some utility programs. Social science researchers new to multilevel modelling could find this book useful, although, if they became regular users, they would need to supplement it with one of the more advanced texts. (*Ian Plewis*)
(Order can be made through authors by email: a716hox%hasara11.bitnet@sara.nl or fax: 31 20 5703500)

∞∞∞

Multilevel Statistical Models by *Harvey Goldstein, London:Edward Arnold, New York: Halsted, 1995, pp178 , price £29.99,ISBN 0 340 59529 9*

Introduction

This book is formally the second edition of Goldstein's *Multilevel Models in Educational and Social Research*, which appeared in 1987. But since 1987 many things have changed. The title, for one thing. The book also *looks* very different, because it is now in the prestigious *Kendall's Library of Statistics*, published by Edward Arnold. The first part of the book is still fairly close to the first edition, but the remainder is more or less completely new. Thus the contents of the book are also very different. We shall treat it, consequently, as a new book, and not as an upgrade of an existing one. The book has a wealth of new material.

The multilevel market place has changed since 1987. Goldstein's first edition was the first book on multilevel models, but in the meantime Bryk and Raudenbush, *Hierarchical Linear Models*, and Longford, *Random Coefficient Models*, have appeared. We shall compare the three works at various places in the review, in order to answer the all-important question for which audience this book is intended.

Basics

We shall first review the book in its role as an expository book on multilevel models. Which audience is it aimed at ? The preface and the first chapter are not completely clear on this, and actually throughout the book there is some ambiguity. Each chapter has appendices with ``technical" material, but unless I am mistaken, the set of people who *can* read and appreciate the main chapters and who *cannot* read and appreciate the appendices is not large.

The introductory chapter is clear, although characteristically short. It discusses, as is usual in books of this type, the ubiquity of hierarchical data, the promise of multilevel models, and the wide variety of existing statistical techniques that can be converted to multilevel versions. There is a section on caveats at the end of the chapter, but this is (also characteristically) even shorter than the other sections.

Chapter 2, which is the core of the expository part of the book, discusses the linear multilevel model. One interpretation of it, is that it tries to condense to 20-30 pages what Bryk and Raudenbush discuss in 200 pages and what Longford discusses in about 150 pages. It fails at this impossible enterprise. A more plausible interpretation is that the chapter defines the notation and terminology, plus some of the basic ideas and problems having to do with algorithms, assumptions, testing, and diagnostics.

A corollary of this analysis of Chapter 2 is that the book is not useful as a textbook on multilevel analysis, even at the graduate or postdoc level. If you want to learn about linear multilevel models, this is not the book for you. Statisticians familiar with variance components analysis, and educational and behavioural statisticians who are already familiar with the multilevel literature, can get quite a bit of mileage out of the chapter, because in a compact form it gives them the necessary background to tackle the remaining chapters.

By the way, the word ``compact" describes much of the book. The amount of information crammed in these pages is astounding. Obviously, a more or less complete treatment of all these topics would require a book of, say, 1000 pages. This is why I think the book is perhaps largely *programmatic*, it gives entries into many additional publications where details are worked out, and it promises a lot of additional research on these topics. Chapter 2, for instance, has a one-page appendix on the EM-algorithm and a one-page appendix on Gibbs sampling. This is just enough to provide one of two references, and a very global idea what these terms refer to, but anybody interested in these matters still has a lot of work to do before they even understand the basics.

Extensions

It is clear that Goldstein thinks of the multilevel idea as a very general one, as indeed it is. If hierarchies are everywhere, then existing statistical models should be adapted to hierarchical data. It must be emphasized that, on the model level, this is a fairly straightforward process. Implementing such a technique is not trivial, however, and making it work in a truly satisfactory way may be quite difficult. The problem with all these extensions and generalizations is that the basic linear multilevel model already has some serious and largely unsolved problems. Most parametrizations tend to be badly conditioned, likelihood functions are flat, and consequently estimation can be problematic. Vendors of software can afford to ignore these problems, but statisticians cannot. If a person is both a vendor of software and a statistician, then this person has a problem. It is not enough to get carried away by all the analyses we can now do with our new software, indeed this is not the statisticians job at all. We have to analyze critically what the properties of the new techniques are, and if they are presented in batches, in a staccato tempo, it becomes very difficult for a statistician to do her job properly.

A first batch of extensions

In the 20 pages of Chapter 3, we encounter complex variance structures, sampling weights, parameter constraints, resampling standard errors, meta analysis, and aggregate level analysis.

Only the first topic gets more than cursory attention. It is based on the observation that first-level residual variances can be modelled in much more detail than is usually done. Indeed, this is a major research topic in regression analysis and generalized linear models. In the spirit of multilevel analysis, we can incorporate first level predictors directly into the first level disturbances. This is easy to do with the MLn software, and the examples are quite interesting, but the full implications of extending the multilevel models in this way are not well understood.

Multivariate Models

In this very short chapter it is pointed out that the multilevel linear model can be extended, in much the same way as the ordinary linear model, to deal with multivariate responses.

Nonlinear Models

Some nonlinear examples of the growth curve type are discussed, by using linearization. The chapter is sketchy, but it appears that experts can indeed handle such nonlinearities in the existing framework.

Repeated Measures Data

The books of Bryk and Raudenbush and of Longford give a lot of attention to repeated measures data, because this is an obvious area in which to apply the linear multilevel model. Goldstein incorporates the possibility of autocorrelated errors, with a quite general autocorrelation function. This gives rise to an enormous number of possible models, of which only a tiny number are illustrated in the example on adolescent growth. It also gives rise to quite a few additional model-choice problems, and these are (characteristically) left not discussed.

Discrete Response Data

We know, from GLM, how to extend linear models so that they can deal with discrete response data. It is most easily done through using a link function. This chapter is a fairly extensive discussion of the various link functions, with applications to counts and multiple responses. Again the treatment is rather sketchy, but less so than in other chapters.

Multilevel Cross Classification

If the data are not purely nested, but the second level is a cross-classification of, say, school and neighbourhood, then of course the multilevel model has to take this cross-classification into account. There is a nice discussion in the book about such designs, and how they translate into variance components. Then cross-classified design are combined with complicated variance structures, and with multivariate data. This is an example of a recurring theme: if you introduce an extension, it can be combined with all previously introduced extensions. This goes without saying, but nevertheless Goldstein says it in various places. This produces a heavy emphasis on the generality, the enormous amount of possible models. It does not emphasize the flip-side of the coin, which is the very serious model choice problems and the possible lack of stability.

Event History Models

Event history models are, or used to be, quite popular in sociology and economics, and they seem to be gaining popularity in education. They are introduced in a couple of pages, and then hierarchies are used to introduce variance components in here as well. This chapter is quite interesting, although it has the usual problem of not telling the reader why particular choices were made, and how a particular analysis was actually done. At least not precisely. Many of Goldstein's examples in the book are not, to use a currently popular term, *reproducible research*. We don't have enough information.

Measurement Errors

Measurement errors can create havoc in ordinary linear model situations, and of course they can do this even more so in multilevel situations. This chapter gives a number of formulas and corrections to deal with measurement error in the covariates at both levels. It is difficult to get the feeling for a general approach to these problems from the chapter, but it seems that there is quite a bit of ongoing research that will clarify the details.

Software and some loose ends

In the last chapter there is a nice unbiased summary of the available multilevel software, although for some largely mysterious reason Goldstein feels that it is necessary to slip in a few more extensions even in this chapter. In general, I want to emphasize that Goldstein's treatment of the multilevel market place is eminently fair. There has been a tendency, especially in the US (of course), to sketch the development of the field as a Darwinian battle between competing software products. Goldstein systematically refuses to enter the fray, and gives major credit to Aitkin, Longford, and others. He barely mentions his computer program, in fact not enough for my taste, because it would be interesting to know how some of the extensions were actually done.

Summary

It is important to emphasize that Goldstein's book does not stand on its own. It is one of the products of the *Multilevel Models Project*, which also produces the computer programs *ML3* and *MLn*, the manuals corresponding to these computer programs, and a stream of both theoretical and applied papers on multilevel analysis. This must be emphasized, because to some extent Goldstein's book reflects the current state and the further research program of the Multilevel Models Project. Thus it can be read both as a progress report, and as a programmatic document. In both these roles it is useful and well-executed. Taken as a whole, of course, the Multilevel Models Project is an impressive effort indeed, and it can serve very well as a model how quantitative educational and behavioural research should be organized.

But, as is perhaps obvious from this review, I have my doubts about the unrelenting expansive approach to multilevel analysis, which looks for generalizations and extensions everywhere. There are quite a few examples in the applied statistical literature of elaborate buildings that have crumbled because the foundations were not solid enough. To use a well-known statistical metaphor, we cannot go on reducing bias by defining more and more elaborate models without seriously jeopardizing the stability of our analyses. It is not clear, from the few examples presented in the book, how useful and how stable these generalizations will be. It is clear, however, that the Multilevel Models Project will provide us with a great deal of additional information about these issues in the future. (*Jan de Leeuw)*

Contributors

We are most grateful to the following people whose contributions have made this issue of the newsletter interest.

Paul C. Lambert & *Keith R. Abrams*
Dept Epidemiology & Public Health
University of Leicester,
Leicester, LE1 6TP, UK
Email: pl4@le.ac.uk

Ian H. Langford & *Toby Plewis*
University of East Anglia
Email: i.langford@uea.ac.uk

Jan de Leeuw
UCLA Statistics Program
University of Califorlia
CA 90024-1555, USA
Email: deleeuw@upf.es

Magdalena Mok
School of Education
Macquarie University
NSW 2109, Australia
Email:ed_mok@hope.ocs.mq.edu.au

Erik Meijer & *Rien van der Leeden*
Dept Psychometrics & Research Methodology
Frank M.T.A. Busing
Department of Data Theory
Leiden University, PO Box 9555
2300 RB Leiden, The Netherlands
Email: vanderleeden@rulfsw.leidenuniv.nl

MLn Clinics in London 1995

Tuesday July 11
Tuesday September 5
Tuesday October 3
Tuesday November 7
Tuesday December 5
AT
Multilevel Models Project
11 Woburn Square, 2nd floor
London WC1A 0SN
contact Min Yang for appointment
Tel: (0)171 612 6682
Email: temsmya@ioe.ac.uk

Theory & Applications

Detecting Outliers in Multilevel Models: an overview

Ian H. Langford and Toby Lewis,
University of East Anglia, UK

Data exploration techniques, including the detection of outlying observations, are a relatively unexplored area of multilevel modelling. For ordinary regression, there is an extensive literature on the detection and treatment of single outliers, and an increasing literature on multiple outliers (Barnett and Lewis 1994). There are a bewildering number of techniques available, the majority of which may be applicable and of use in a particular situation. However, as we have undertaken our research on outliers in multilevel models, we have encountered further complexities which require a somewhat different approach to those used traditionally.

A data exploration approach

The first issue is where to start data exploration in a multilevel model. Rather than looking at individual data points, we have found it most useful to begin at the level of highest aggregation, which may be simply the highest level in the model, or the least complex of a set of cross-classifications. The reasons for this are two-fold. Researchers are often most interested in the highest level of aggregation, and will naturally concentrate their initial efforts here. However, it is also true that if discrepancies can be found in higher level structures, these are more likely to be more indicative of serious problems than a few outlying points in lower level units. After analysing the highest level, then lower levels should be examined in turn, with analysis and initial treatment of outliers at the lowest level of the model. The highest level should then be re-examined after the model has been refitted to the data. The objective is to identify whether an outlying unit at a higher level is consistently outlying, or outlying due to the effects of one or two aberrant lower level units they contain. Similarly, examination of lower level units may uncover the fact that one or two lower level units are aberrant *within* a particular higher level unit which does not appear unusual, and that the higher level unit would be aberrant without these lower level units. Hence, caution must be taken with the analysis not simply to focus on interesting higher level units, but to fully explore lower level units as well.

Detecting structure

The second issue in outlier detection in multilevel modelling concerns emphasis. In ordinary regression, the aim is to detect outliers, usually sequentially as single outliers (though sometimes *en bloc*) and treat them in some fashion, whether by deletion, identification for separate accommodation, model reformulation or whatever. Some recent work has focused on influential subsets of outliers (Peña and Yohai 1995; Lawrance 1995). Sometimes, useful analysis may be undertaken using deletion of single units; but often, at any level, we may be inherently looking for more structure due to the separate estimation of fixed and random effects in a multilevel model. In *MLn*, a single fixed coefficient for an explanatory variable is easily produced, with an estimate of the mean deviation of units around this fixed coefficient at any particular level. In this situation, it may be of primary interest to see how the random effects are determined, whether exclusion of one or more units from the random part of the model removes the necessity for random parameters, or more generally, whether there is any structure to the random effects which is distinct enough to warrant modelling as a fixed effect. This type of analysis is part of the further extension of modelling from fixed effects, to fixed plus random effects, to a general examination of structure. Bernoulli (1777, in Beckman and Cook 1983) first criticised the assumption of identically distributed errors, stating that “every observation should be admitted whatever its quality, as long as the observer is conscious that he has taken care.” The present research on outliers generally follows this principle.

Basic methods for detection

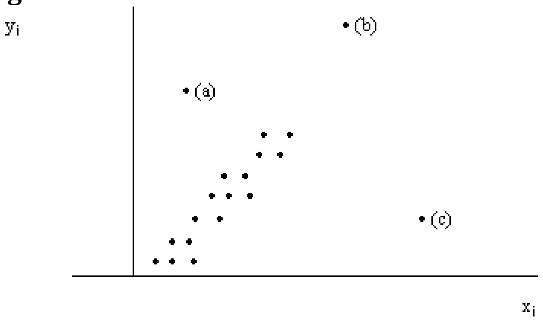
However, it is still true that the basic aims of outlier detection found in ordinary regression apply, and there are three basic situations we can search for, albeit at different levels of a hierarchical model (Rousseeuw and van Zomeren 1990). These are (see Figure 1):

a) vertical outliers, which have low leverage;

b) leverage points which are consistent with the relationships found in the rest of the data (and are not outliers in our sense), and;
c) leverage point outliers, which significantly alter the relationships found in the data by their inclusion, as well as having high leverage.

The difference is that the points displayed in Figure 1 may be individual data points within a level 1 unit, or be the distribution of random coefficients at a higher level unit. In this second case, each point is really a cloud of points whose "centre" has been shrunken towards the fixed parameter estimates in the model.

Figure 1.



Research in progress

Reports on specific examinations of outliers in multilevel modelling applications are beyond the scope of this paper, but will be reported in a later issue. For example, an extensive re-analysis has been undertaken of Aitken and Longford's (1986) data on school effectiveness, modelling exam scores as a function of an intake measure of Verbal Reasoning Quotient in 19 schools, two of which are grammar schools and potential outliers. Data exploration at school level has focused on reductions in deviance from exclusion of each school in turn from the *random part* of the model, and the effects of this on random parameter significance. This was a computationally expensive operation which took over 1300 iterations, and the authors are researching "one-step" alternatives in the fashion of Williams (1987). After unusual schools were identified, attention focused on residuals, leverage and influence measures at pupil level, before re-analysis of school level effects, as suggested above. An examination of a simulated data set, including detection of clustering, is underway, as well as analyses of other educational, social survey and geographical data sets.

References

Barnett V and Lewis T (1994). *Outliers in Statistical Data: 3rd edition.* New York: John Wiley.

Beckman RJ and Cook RD (1983). Outlier..........s. *Technometrics*, 25, 119-149.

Lawrance AJ (1995). Deletion influence and masking in regression. *J R Statist B*, 57, 181-189.

Peña D and Yohai VJ (1995). The detection of influential subsets in linear regression by using an influence matrix. *J R Statist B*, 57, 145-156.

Rousseeuw PJ and van Zomeren BC (1990). Unmasking multivariate outliers and leverage points. *J Am Statist Assoc*, 85, 633-639.

Williams DA (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Appl Stat*, 36, 181-191.

8888888888888888888888888888888

Implementing the Bootstrap for Multilevel Models

Erik Meijer, *Rien van der Leeden* and
*Frank M. T. A. Busing**
Leiden University, The Netherlands

(*Supported in part by SVO project No. 93713)

Introduction

Multilevel models are usually estimated by maximum likelihood methods, be it full information maximum likelihood (**FIML**) or restricted maximum likelihood (**REML**). The maximum likelihood theory is based on several assumptions, some of which are (a) The random errors at all levels are normally distributed, and (b) The sample size is large. More specifically, the properties of the maximum likelihood estimators are derived under the assumption that the sample size goes to infinity.

In practice, these assumptions will at best only be met approximately. The most important effects the violation of these assumptions may have are bias of the estimators and incorrect standard errors.

In many models and situations maximum likelihood estimators are biased in finite samples. For a general class of regression models including multilevel models, however, Magnus (1978) proved that the maximum likelihood estimators of the *fixed* regression coefficients are unbiased. On the other hand, Busing (1993) showed in a Monte Carlo simulation study that the **FIML** estimators of the variance components in multilevel models are biased.

The standard errors of the maximum likelihood estimators that are reported by the various multilevel packages are derived from asymptotic theory. This means that they are based on the idea that as the sample size goes to infinity, the distribution of the estimators will converge to a (multivariate) normal distribution with a certain covariance matrix. The reported standard errors are the square roots of the diagonal elements of this matrix. In finite samples, this approximation may not be very good. The true standard errors may be quite different from the reported ones based on asymptotic theory, and the distributions of the estimators may not be normal. In fact, Busing (1993) showed that the distributions of the variance component estimators can be severely skewed. The focus of this paper is, however, on bias and standard errors, and not on the specific distribution.

A fairly general way to obtain estimates of the bias and correct standard errors is the *bootstrap* (e.g., Efron, 1982). The bootstrap can, however, not straightforwardly be implemented in multilevel models, because the observations are not identically and independently distributed (IID). In this paper, the implementation of three bootstrap methods for multilevel models will be discussed.

The model and its estimation

We will discuss the implementation of the bootstrap procedures for a two-level model. The ideas generalise, however, straightforwardly to models with more levels.

The model is given by

$$y_j = Z_j \beta_j + \varepsilon_j \qquad (1)$$

and $$\beta_j = W_j \gamma + u_j \qquad (2)$$

where j denotes the level-2 unit number, $y_j = (y_{1j}, \ldots, y_{N_j j})'$ is the vector with score on the dependent variable of level-2 unit j, Z_j and W_j are matrices with explanatory variables, β_j and γ are vectors with random and fixed regression coefficients, respectively, and $\varepsilon_j = (\varepsilon_{1j}, \ldots, \varepsilon_{N_j j})'$ and u_j are vectors of level-1 and level-2 random errors, respectively.

Generally, it is assumed that $\varepsilon_j \sim N(0, \sigma^2 I_{N_j})$ and u_j~N(0,Θ), where σ^2, the variance of the level-1 error term, is an unknown (scalar) parameter, and Θ, the covariance matrix of the level-2 error terms, is a (symmetric) matrix of unknown parameters. This model can be elaborated, e.g. by making the level-1 variance a function of other variables, but we shall not consider this here (Goldstein, 1995). From these assumptions and (1) and (2), the likelihood function can be formulated and maximised to obtain maximum likelihood estimators.

The parameters that have to be estimated are the elements of γ (the fixed parameters) and σ^2 and the elements of Θ (the variance components). In order to implement one of the bootstrap methods we need estimators of the level-2 random coefficients β_j and of the random error terms u_j and ε_j (see the next section). Within the multilevel framework, these estimates are usually obtained by *shrinkage* estimation. This yields posterior means and shrunken residuals. An alternative method is using the within-unit **OLS** estimators and corresponding residuals (we call them *raw* residuals).

Bootstrapping multilevel models

Bootstrap methods are concerned with drawing samples from the empirical distribution function. Common bootstrap methods can not be straightforwardly applied to multilevel models, because bootstrap theory requires the observations to be independently distributed. This is not the case with multilevel data, where the observations are subject to intraclass dependency. Therefore, special resampling schemes have to be devised, which take the hierarchical data structure into account. We discuss three methods: (1) the *parametric* *bootstrap*, which uses the normality assumption, and therefore only gives information about sample size effects; (2) the (nonparametric) *error bootstrap,* which assumes that the explanatory variables are fixed; and (3) the (nonparametric) *cases bootstrap*, which assumes that the explanatory variables are random with unknown distribution. For the first two of these methods it is also assumed that the model is true in the population.

The parametric bootstrap uses the parametrically estimated distribution function of the data to generate new samples and compute relevant statistics based on these new (bootstrap) samples. In the two-level model discussed here, it is assumed that the level-1 errors ε_{ij}, *i=1, ..., N_j, j=1,...,J,* are identically and independently distributed N(0,σ^2), and that the level-2 error vectors u_j, *j=1,...,J,* are identically and independently distributed N(0, Θ). Hence, the parametrically estimated distribution functions of ε_{ij} and u_j are the N(0, $\hat{\sigma}^2$) and N(0, $\hat{\Theta}$) distribution functions, respectively.

The parametric bootstrap now draws independent pseudo-random samples from these normal distributions to obtain bootstrap samples $\varepsilon_{ij}^{*(b)}$ and $u_j^{*(b)}$, *i=1,...,Nj, j=1,...,J, b=1,...,B.* Then, for each *b=1,...,B,* a bootstrap sample of the response variable is obtained from

$$\beta_j^* = W_j\hat{\gamma} + u_j^* \qquad (3)$$

and

$$y_j^* = Z_j\beta_j^* + \varepsilon_j^* \qquad (4)$$

where $\varepsilon_j^* = (\varepsilon_{1j}^*, \ldots, \varepsilon_{N_jj}^*)'$ and the superscript (b) is omitted for simplicity. The parametric bootstrap can also be obtained using the **SIMU**lation option of ***MLn*** (Rasbash & Woodhouse, 1995).

For the *nonparametric bootstrap,* several variations can be studied. If the explanatory variables can be considered *fixed* design variables, then, analogously to regression analysis, the *errors* have to be estimated and subsequently resampled (see, e.g., Efron, 1982, pp. 35-36). As mentioned in the previous section, either the shrunken residuals or the raw residuals can be used as estimators of the errors.

Unlike in regression analysis, the estimated residuals in multilevel analysis do not necessarily have a zero mean. Therefore, the residuals must be centered first over the whole data set. Otherwise, the possibly nonzero mean of the errors would necessarily lead to biased estimators of the intercept parameters.

From the centered estimates $\{\hat{u}_j\}$ and $\{\hat{\varepsilon}_{ij}\}$ the (nonparametric) empirical distribution functions of the errors can be obtained. Then, nonparametric bootstrap samples u_j^*, *j=1,...J* and ε_{ij}^*, *j=1,...,J, i=1,...,Nj,* are obtained by drawing samples from these empirical distribution functions. This is equivalent to drawing samples with replacement from the centered residuals. Then, nonparametric bootstrap samples of *y* are obtained from (3) and (4). We call this the *error bootstrap*.

If the *Z* and *W* variables are considered *rando*m, nonparametric bootstrap samples can be drawn by resampling complete cases. The bootstrap samples can be drawn in the following way. First, a sample of size *J* is drawn with replacement from the *level-2 units*. This gives a sample j_k^*, *k=1,...,J* of level-2 unit numbers and accompanying level-2 variables $W_{j_k}^*$. Then for each *k*, a nonparametric bootstrap sample of complete cases from the (original) unit $j = j_k^*$ is drawn, giving (y_{ik}^*, Z_{ik}^*), *k=1,...,J, i=1,...,$N_{j_k^*}$*. We call this the *cases bootstrap* for both levels.

It is also possible to draw cases bootstrap samples from the level-2 units only or from the level-1 units within each level-2 unit only. This can be useful when the level-2 units, or the level-1 units within the level-2 units can not be considered random, for example when countries are compared (fixed level-2 units), or with repeated measures data (fixed level-1 units within each level-2 unit).

Once bootstrap samples are drawn, bias-corrected bootstrap estimators and standard errors are obtained as follows. Consider a typical parameter θ. Its estimator $\hat{\theta}$ is computed from the original sample. For each bootstrap sample *b=1,...,B* (obtained in one of the ways described above), a bootstrap estimator $\theta^{*(b)}$ is obtained in the same way the estimator $\hat{\theta}$ was obtained from the original sample. The mean of these estimates $\theta^{*(b)}$ is called $\theta^{*}_{(.)}$, and its variance is called V*. The bias of $\hat{\theta}$ is now estimated as $Bias(\hat{\theta}) = \theta^{*}_{(.)} - \hat{\theta}$ and a bias-corrected estimator is $\hat{\theta}_B = \hat{\theta} - Bias(\hat{\theta}) = 2\hat{\theta} - \theta^{*}_{(.)}$. The standard error of $\hat{\theta}$ is estimated by $\sqrt{V^*}$, which is also used as a standard error of $\hat{\theta}_B$. See, for example, Efron (1982) for a comprehensive theoretical discussion of bootstrap estimators and standard errors.

Preliminary results and discussion

In order to study the statistical properties of the proposed bootstrap methods, the program **MLA** has been developed. **MLA** provides **FIML** estimation of the two-level model presented in the second section of this paper, as well as the various bootstrap methods discussed in the previous section. Two experimental jackknife methods are implemented too. Other features include simple non-iterative estimators, which, under certain conditions, may be considered useful alternatives to **FIML** estimators (Van der Leeden & Busing, 1994; De Leeuw & Kreft, 1994). Shrinkage and within-unit **OLS** estimators of the random regression coefficients β_j and errors u_j and ε_{ij} can be obtained as well.

The **MLA** program is intended as a vehicle for our own research interests. Therefore, it is not a completely functional program, in the sense that many options of the major programs **MLn**, **VARCL**, and **HLM** are not implemented. For example, the number of levels is restricted to 2. Interested researchers, however, may obtain a copy of the program and the manual (Busing, Meijer, & Van der Leeden, 1994) from us. In the manual the estimation and resampling procedures are discussed in more detail.

Using the MLA program, we are currently performing Monte Carlo simulation studies to give insight in the properties of the bootstrap estimators under various conditions. It appears that γ and σ^2 are usually estimated rather well with **FIML**. Therefore, only bootstrap estimators of Θ, the Level-2 variance components, are considered here. In the simulation, a simple two-level model with one explanatory variable at each level is used. All four elements of γ are set to 1, σ^2 is set to 1, the variance of the Level-1 random intercept and slope terms are set to 0.5 and their covariance is set to 0.25. The explanatory variables are drawn from standard normal distributions. The following settings are varied in the simulation design: number of Level-2 units (10, 25, 65), size of Level-2 units (unbalanced with on average 10, 25, 65 Level-1 units) and distribution of the errors (all normal vs. all skewed). Each generated sample gives **FIML** estimates and standard errors, and raw and shrunken error bootstrap estimates and standard errors. Thus, every parameter is associated with three estimates and three standard errors. The number of B of bootstrap replications is set to 100 and the number of Monto Carlo replications is 500. For each parameter (θ, say) and each estimation method, the relative bias is computed as $(\bar{\theta} - \theta_0) / \theta_0$, where $\bar{\theta}$ is the average of the 500 estimates of θ with given settings and estimation method, and θ_0 is the true value of θ. Additionally, for each parameter, the relative bias of the standard error is estimated in the same way, where the 'true' value of the standard error is estimated by the standard deviation of the parameter estimates in the sample of 500 replications. We are currently analyzing the data obtained from this simulation, and performing simulation with the cases bootstrap and parametric bootstrap.

Table 1 gives some preliminary (and condensed) results for the error bootstrap methods. It indicates that bias correction works fine for the error bootstrap with shrunken residuals, but not for the error bootstrap with raw residuals. As expected, the results also indicate that bootstrapping is only relevant with relatively small samples. Results are, however, preliminary, and more research needs to be done.

Table 1. Relative bias for intercept variance estimator: FIML and raw and shrunken residual bootstrap results

Sample size	Distribution	FIML	Raw	Shrunken
Small	Normal	-0.18	-0.21	0.05
	Skewed	-0.17	-0.20	0.06
Large	Normal	-0.04	-0.05	-0.01
	Skewed	-0.05	-0.06	-0.02

In this paper, we discussed the original bootstrap method, adapted for hierarchical data. Many refinements, extensions, and alternatives to the original bootstrap method have, however, been proposed, especially for regression models (Wu, 1986). It may be relevant to implement these ideas in bootstrap methods for multilevel models as well to improve performance.

References

Busing, F. M. T. A. (1993). Distribution characteristics of variance estimates in two-level models; A Monte Carlo study. Technical report No. *PRM 93-04*. Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

Busing, F. M. T. A., Meijer, E. & Van der Leeden, R. (1994). MLA. Software for multilevel analysis of data with two levels. User's guide for version 1.0b. *PRM 94-01*. Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

De Leeuw, J. & Kreft, I. G. G. (1994). Questioning multilevel models. UCLA Statistics Series #143. Los Angeles: University of California.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. *Philadelphia: SIAM.*

Magnus, J. R. (1978). Maximum likelihood estimation of the {GLS} model with unknown parameters in the disturbance covariance matrix. *Journal of Econometrics*, 7, 281-312.

Rasbash, J. & Woodhouse, G (1995). *MLn* Command Reference, Version 1.0, London: University of London, Institute of Education.

Van der Leeden, R. & Busing, F. M. T. A. (1994). First iteration versus final IGLS/RIGLS estimates in two-level models: A Monte Carlo study with ML3}. *PRM 94-02*. Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis [with discussion]. *Annals of Statistics*, 14, 1261-1350.

88888888888888888888888888888

Sample Size Requirements for 2-level Designs in Educational Research

Magdalena Mok
University of Macquarie, Australia

Introduction

This study is concerned with the effect of sample size on the efficiency of estimates for 2-level survey research designs. One component of sample design is the numbers of level-2 units and level-1 units to be included in the sample to produce accurate information. For example, in the study of school culture, the researcher needs to decide how many schools, and how many students from each school have to be included in the study. Naturally, costs are an important consideration. Costs for a 2 level design would be: (a) those at the school level, comprising, costs of compiling the lists of schools, transportation between center and sites, liaison with gatekeepers, and mailing costs, and (b) the costs at the student level, comprising costs of compiling the lists of students, data collection and analysis. As long as the total number of students remains constant, costs at the student level would be the same, irrespective of the number of schools involved. On the other hand, the design can have serious financial implications for the costs at the school level; the more schools are involved, the more costly it would be. However, cost was not the point of consideration for this paper. Instead, we consider the following problem: suppose 2000 students in total is required for the study, then, in terms of efficiency, should the 2000 students come from 20 schools with 100 students from each school, or should the sampling scheme consist of 100 schools each with 20 students, or some other school- and student-size combinations?

At least for variance component models, the sample design question relates to a 2-level modelling situation, where level-1 units (for example, students) are nested within level-2 units (for example, schools), is analogous to that addressed by Kish (1965:259) in computing the effective sample size in two-stage cluster sampling. Effective sample size of a 2-stage cluster sampling design, n_{eff}, is computed by:

$$n_{eff} = n / [1 + (n_{clus} - 1)\, \rho] \qquad (1)$$

where n is the total number of students in the study, that is, the actual sample size, n_{clus} is the number of students per school, and ρ is the intra-class correlation.

However, the analogy is not straightforward for random slope models, because the intra-class correlation for these models is a function of the independent variable (Goldstein, 1995). For example, for the case of one continuous dependent variable regressing on one continuous explanatory variable, both measured at level-1,

$$y_{ij} = (\alpha_0 + \alpha_1 x_{ij}) + (\varepsilon_{ij} + v_{0j} + v_{1j} x_{ij}) \qquad (2)$$

where $\varepsilon_{ij} \sim N(0, \sigma_e^2)$, and $v_{0j} \sim N(0, \sigma_0^2)$ $v_{1j} \sim N(0, \sigma_1^2)$, $\mathrm{cov}(v_{0j}, v_{1j}) = \sigma_{01}$, The function of intra-class correlation ρ is given by the ratio of level-2 variance to total variance. For the model in equation (2),

$$\begin{aligned} \mathrm{var}(level-2) &= \sigma_0^2 + 2\,\sigma_{01} x_{ij} + \sigma_1^2 x_{ij}^2 \\ \mathrm{var}(level-1) &= \sigma_e^2 \\ \rho &= \frac{\mathrm{var}(level-2)}{\mathrm{var}(level-2) + \mathrm{var}(level-1)} \end{aligned} \qquad (3)$$

It is therefore necessary to re-address the question of sample size requirements for 2 level models when random slope models are involved. The research questions for this study are: (a) What are the effects of level-2 and level-1 sample sizes on sampling accuracy in terms of unbaisedness, efficiency, and consistency of parameter estimates? (b) Does equation (1) give a reasonable estimate of effective sample size for multilevel models when random slope models are involved?

Method

Simulation methods were utilised to answer the research questions. First, a large population was generated from an existing empirical data set, such that the true values of parameters to be estimated were known, and the population had the desired multilevel structure. Next, according to a sample design grid, samples with specific level-1 and level-2 units were drawn randomly from the population. The 11 x 11 sampling design grid was made up of 11 rows of level-1 sizes per level-2 unit: 5, 10, 20, 30, 40, 50, 60, 70, 80, 100, 150, and 11 columns of level-2 sizes: 5, 10, 20, 30, 40, 50, 60, 70, 80, 100, 150 (see Table 1). All designs were balanced. In this way, the total sample size for cell (*i*,*j*) in the grid was the same as the total sample size for cell (*j*,*i*). Cells below, on, and above the diagonal were labelled Type A, B, C designs respectively. The cells for Type A, (respectively, Type B, Type C) designs had smaller (respectively, equal, larger) numerical values of level-1 units than those of level-2 units. So for each actual total sample size, there were at least 2 points of comparison made possible: $i>j$, and $i<j$.

A random sample, satisfying the design sample size specifications, was selected from the population. Each sample point consisted of a pair of observations, one on the independent variable, and the other on the dependent variable. Based on each selected sample, 100 simulations were generated using the *MLn* package (Rasbash and Woodhouse, 1995). For each set of simulated data, a random slope model was fitted and the parameter estimates using the RIGLS method of estimation (Goldstein, 1995), were recorded. The 6 parameters to be estimated comprised the fixed components of the intercept α_0, and of the slope α_1, the level-2 variance of the intercept σ_0^2, and of the slope σ_1^2, the level-2 covariance σ_{01} between the intercept and the slope, and the variance at level-1 σ_e^2. From the parameter estimates over 100 simulations of the conditions of the design, the (signed) bias, the empirical

Table 1. Sampling design grid

	Number of Level-2 units										
Number of level-1 units per level-2 unit	5	10	20	30	40	50	60	70	80	100	150
5	25	50	100	150	200	250	300	350	400	500	750
10	50	100	200	300	400	500	600	700	800	1000	1500
20	100	200	400	600	800	1000	1200	1400	1600	2000	3000
30	150	300	600	900	1200	1500	1800	2100	2400	3000	4500
40	200	400	800	1200	1600	2000	2400	2800	3200	4000	6000
50	250	500	1000	1500	2000	2500	3000	3500	4000	5000	7500
60	300	600	1200	1800	2400	3000	3600	4200	4800	6000	9000
70	350	700	1400	2100	2800	3500	4200	4900	5600	7000	10500
80	400	800	1600	2400	3200	4000	4800	5600	6400	8000	12000
100	500	1000	2000	3000	4000	5000	6000	7000	8000	10000	15000
150	750	1500	3000	4500	6000	7500	9000	10500	12000	15000	22500

Notes:
1. Cell entries denote the total number of level-1 units, that is, the Actual Sample Size.
2. Design A: Cells below the diagonal; Design B: Cells on the diagonal; Design C: Cells above the diagonal.

sampling variance, and the empirical Mean Square Error (MSE) were computed. If β is the true parameter, and b_j is the jth replication of 100, then (signed) bias is given by $(\quad b_j/100-\beta)$, and the (empirical) sampling variance is $\{ \quad [b_j-(\quad b_j/100)]^2/(100-1)\}$. The empirical Mean Squared Error (MSE) is the sum of squared bias and the sampling variance.

The population for this study was simulated from a real data set, which consisted of 4,949 students from 50 New South Wales Catholic schools. The data set was collected by Flynn (1993). The dependent variable was standardised performance at the HSC (Higher School Certificate) examination, and only one explanatory variable, namely the standardised attitude toward achievement, was selected for this study. The population was simulated from a random slope model, as specified in (2) above, fitted to the real data, and consisted of 247,450 simulated students from 440 simulated schools. Students comprised the level-1 units, and schools comprised the level-2 units.

The population parameter values were:

Intercept	α_0	=-.07790	SE=.01359
Slope	α_1	=.2457	SE=.01195
Lev-2 intercept var.	σ_0^2	=.07229	SE=.005227
Lev-2 slope var.	σ_1^2	=.05536	SE=.004039
Lev-2 covariance	σ_{01}	=.01161	SE=.003302
Lev-1 random component	σ_e^2	=.4229	SE=.001731
Intra-class corr.	ρ	=.14598	

Results

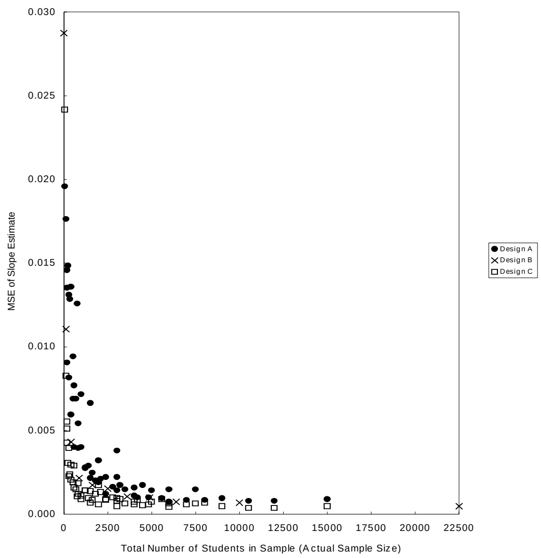
1. Estimates of the fixed components of the intercept and the slope

The *MLn* estimates, a_0 of the intercept α_0, and a_1 of the slope α_1, were obtained for 100 simulations of conditions of the design. The summary statistics of the (signed) bias, sampling variance, and MSE of these estimates are given in Table 2. Results on the intercept and the slope are similar. Four observations were made. First, all designs are consistent: As sample size increases, bias tends to approach zero. Indeed, if the total sample size is more than 800 students, then all estimates of the fixed component of the model lie within 1 standard error of the true value, irrespective of the Type (A, B, or C) of design (Table 2).

Second, for designs with total sample size less than or equal to 800, it appeared that Type A designs show more bias than either Type B or Type C designs: 7 of the 8 designs which had bias more than 1 standard error from the true intercept were of design Type A, and the remainder of design Type C; of the 8 designs with biases more than 1 standard deviation from the true slope, 6 were of design Type A, 1 was of design Type B, and the remaining one was of Type C.

Third, both the intercept sampling variance and the slope sampling variance decrease rapidly as sample size increases from 25 to about 2500, but the curves level off around sample size 2500. This indicated that, for the fixed part of the model, the gain in efficiency by increasing the total sample size beyond 2500 was relatively small. The graphs of the sampling variances and of the MSE of the fixed components are very similar in shape, therefore only the graph of MSE of the slope plotted against sample size is presented here (Figure 1).

Figure 1. MSE of Slope (Fixed Component) Against Actual Sample Size



Four, given a fixed sample size, even when the sample size is considerably large (say, larger than 3000 students), design Type A has relatively larger sampling variance, as well as larger MSE, than either types C or B designs (Figure 1). Based on these observations, for a given total sample size, designs involving more schools and fewer students per school, tend to be less biased and more efficient than designs involving fewer schools and more students per school.

2. Estimates of the level-2 variance of the model: σ_0^2 *and* σ_1^2

The results of the level-2 variance estimates, s_0^2 and s_1^2 are very similar, and three observations are made on them: (a) There is notably more bias, and larger MSEs, in Type A, than in either B or C designs; (b) The majority of Type A designs are biased downwards; (c) Although all designs give consistent estimates, it appears that on increasing sample size, the bias of C designs approaches zero more rapidly than A designs (Table 2).

3. Estimates of the level-2 covariance, σ_{01}

It is not immediately obvious that Type C designs are less biased than A designs in estimating level 2 covariance between slope and intercept, although for Type C designs, it would require a total 400 students in the sample, and for Type A designs. at least 1500 students, for s_{01} to lie within 1 standard error of the true value; all designs are consistent. On the other hand, for a given sample size Type A designs have both smaller sampling variance and smaller MSE than Type A designs.

4. Estimates of the level-1 variance, σ_e^2

All *MLn* estimates of variance at level-1, averaged over the 100 simulations for each sample design lie within 1 standard error of the true level-1 variance if the total sample size is more than 4000. There is no strong indication of which type of design (A, B, or C) is superior to the other, as far as bias is concerned. All designs are consistent. The sampling variance of level-1 variance decreases rapidly as sample size increases from 25 to about 600, beyond which the curve of sampling variance levels off. Type A designs have both larger sampling variance and larger MSE than Type C designs for a given sample size.

5. Adjustment using design effect and effective sample size

It is possible to compute the design effect and the corresponding effective sample size (Kish, 1969) for a 2-level random slope model at selected values of the independent variable, using equations (1) and (3) above. It was decided to compute the design effect and effective sample size at the x-intercept (x=0) in equation (3). Two designs with the same effective sample size are conjectured to have the same bias and the same efficiency.

The results show that, after controlling for effective sample size, Type C designs are more biassed than Type A designs in estimating the intercept (fixed component), level-2 slope-intercept covariance, level-2 slope variance and level-1 variance for a given effective sample size. On the other hand, Type C designs are less bias than Type A designs in estimating level-2 intercept variance. There is no obvious pattern as to which type of design is more or less bias in estimating the slope (fixed component), nevertheless.

In terms of efficiency, if two designs have the same effective sample size, then both their estimates on the fixed components have the same efficiency, irrespective of whether the designs are of Type A, B, or C. However after controlling for effective sample size, Type A designs appear to be more efficient than Type C in estimating the random components at both levels 1 and 2. These findings suggest that perhaps the adjustment using equation (1) computed at the x-intercept over-penalised designs involving more schools with smaller within-school sample size (i.e. Type C) compared to the penalty on those designs involving fewer schools with larger within-school sample size (i.e. Type D).

Conclusions

The object of the study was to investigate sample size requirements for 2-level, random slope, balanced designs using simulation. It was found that, consistent with advice given in the classical literature of cluster sampling designs, if resources are available for a sample size n, comprising J schools with I students from each school, then less bias and more efficiency would be expected from sample designs involving more schools (large J), and fewer students per school (small I) than sample designs involving fewer schools (small J), and more students per school (large I). Converting actual sample size into effective sample size according to equation (3) computed at the x-intercept removed the clustering effect in estimating the fixed components of the slope and intercept, but such conversion failed to remove clustering effects in estimating the random components at either level-2 or level-1. This result is not entirely surprising: By computing intra-class correlation at the x-intercept, one essentially computes the intra-class correlation for a variance component model; given that intra-class correlation varies with the value of the square of the independent variable for random slope models, it might be over-simplistic to expect a single effective sample size to exist for each condition of the designs for such models. Perhaps, in the same spirit as one would center at respective school means for random slope models, and as a second stage of the research, one might compute, for each school within each sample, the intra-class correlation and the associated effective sample size, at the respective school means, and then the effective sample size of the sample would be computed as the sum total of the effective sample sizes of all schools in the sample. To the extent that these data are representative, one might offer as a rule of thumb, in the 2-level random slope balanced case with intra-class correlation of below, say, 0.15, at the x-intercept, that an actual sample size of 3500, and an effective sample size at the x-intercept of 400, to ensure reasonable efficiency and lack of bias.

References

Flynn, M. (1993). *The Culture of Catholic Schools*. NSW: St. Paul's Publication.

Goldstein, H. (1995). *Multilevel Statistical Models*. London: Edward Arnold, New York: Halsted.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Rasbash, J. & Woodhouse, G. (1995). *MLn Command Reference*. London: Institute of Education, University of London.

Acknowledgement

This study was completed while the author was visiting the Multilevel Models Project as part of her OSP in Spring term, 1995; the support of the Project team is gratefully acknowledged. The comments on an earlier draft by Professor Roderick P McDonald, Dr Nick Parr, and Professor George Cooney are gratefully acknowledged.

(***Editor’s note****: A full version of this paper with detailed information on results can be found on the project’s WEB pages in the compressed form for downloading.)*

Meta-Analysis Using Multilevel Models

Paul C. Lambert & Keith R. Abrams
University of Leicester, UK

Introduction

Meta-analysis is the statistical synthesis of results from a number of similar studies (Hedges & Olken, 1985). though originally developed for use in education and social research, over the last ten years it has become an accepted part of medical research (Jones, 1995).

At the simplest level, meta-analysis assumes a fixed effects model in which each study is assumed to estimate an unknown overall population effect (Fleiss, 1993). However, there is sometimes a considerable amount of heterogeneity between the individual studies with respect to their effect sizes, and in such circumstances a random effects model has been advocated (DerSimion & Laird, 1986). In these models each study is assumed to be estimating its own unknown study effect, which are themselves assumed to be distributed about an unknown population effect. More recently mixed effects models have been advocated as a means of analysing meta-analysis in which fixed covariates are used to explain as much of the heterogeneity between studies as possible, and the remaining is modelled using a random component (Breslow & Clayton,1993, Thompson, 1994).

A characteristic of meta-analyses is that usually information is only available at the study level, for example in terms of an odds ratio, together with its standard error. We consider the analysis of such data, using odds ratios adjusted for potential confounding variables, using multilevel models (Goldstein, 1995).

Background - Cholesterol and Mortality

An example of a meta-analysis is Davey-Smith, Song and Sheldon 1993) in which the effects of lowering blood serum cholesterol levels on mortality (Both all-cause and cardiac) were assessed in 35 different randomised controlled trials. Various study-level covariates were collected, amongst the most important were though to be baseline-risk, i.e. the cardiac mortality rate in the control group. Study-level data was available as odds ratios and standard errors. The data used in this example has slightly changed to that published by Davey-Smith, Song and Sheldon (1993), since it has been revised in the light of new information.

Methods

We consider two models, model 1 - a simple random effects model, and model 2 - a random effects model incorporating study specific baseline risk.

Model 1

A simple random effects model (not allowing for study level covariates). Let y_i be the observed odds ratio in the i^{th} study and σ_i its observed standard error. We can specify the model as

$$y_i = \beta_0 + s_i + e_i$$

where β_0 is the estimate of the pooled log odds ratio, s_i is the effect of the i^{th} study and is distributed $s_i \sim N(0, \sigma_s^2)$, e_i is the error associated with the i^{th} study where $E(e_i) = 0$ and $\text{var}(e_i) = \sigma_i^2$.

In order to fit this model in *ML3/MLn* (Rasbash & Woodhouse, 1995), we need the following columns

ID	- The study number (1-34)
CONS	- A constant term
LOR	- The log odds ratios
LOR_SE	- The standard errors of the log odds ratios.

ID is the identifying variable at level 2 with CONS the identifying variable at level 1. LOR is declared as the response variable and CONS as a fixed effect. CONS is also a random effect at level 2. LOR_SE is a random effect at level 1, but the parameter estimate is constrained to equal 1, so that the level 1 variance of the i^{th} study is $1 \times LOR\_SE_i^2 = \sigma_i^2$ as required.

Model based estimates for the log odds ratio in the i^{th} study are given by

$$(\hat{\sigma}_s^2 \hat{y}_i + \hat{\sigma}_i^2 \hat{\beta}_0) / (\hat{\sigma}_s^2 + \hat{\sigma}_i^2)$$

with standard error

$$\sqrt{\frac{\hat{\sigma}_s^2 \hat{\sigma}_i^2}{\hat{\sigma}_s^2 + \hat{\sigma}_i^2} + \left({}_i \ \frac{1}{\hat{\sigma}_i^2 + \hat{\sigma}_s^2} \right)^{-1}}$$

It should be noted that this takes account of the imprecision of $\hat{\beta}_0$ but not $\hat{\sigma}_s^2$. Accommodating uncertainty with regard to $\hat{\sigma}_s^2$. can be performed in a fully Bayesian framework using Monte Carlo simulation methods (Carlin, 1992) or via bootstrapping (Goldstein, 1995).

Model 2

This is similar to Model 1, but where the effect of baseline risk is included as a fixed parameter measured at level 2. Baseline risk was divided into 3 categories depending on the number of deaths from coronary heart disease per 1000 person years on control subjects.

High Risk (>50 deaths)	- HR
Medium Risk (10-50 deaths)	- MR
Low Risk (<10 deaths)	- LR

Thus the model is now

$$y_i = \beta_1 HR_i + \beta_2 MR_i + \beta_3 LR_i + s_i + e_i$$

where s_i and e_i are defined as for the model 1.

Results

Table 1 shows the results from fitting Model 1 using *ML3*. The estimate of the odds ratio is therefore $e^{-0.118}$ (95% CI, 0.79 - 1.00). The between study variance is estimated to be 0.0434.

The analysis was repeated using the method of DerSimion and Laird (1986), in which the between study variance was estimated using a non-iterative method of moments. Similar results were obtained for the overall estimate of the odds ratio, 0.89 (95% CI, 0.78 - 1.02) while the between study variance was estimated to be 0.0679.

Table 1 Results from *ML3 (Model 1)*

	Parameter	Estimate	S.E.
Fixed	β_0	-0.118	0.0609
Level 2	σ_s^2	0.0434	0.0245
Level 1	$\mathrm{var}(e_i)$	1	0

The baseline risk was a priori thought to be important so Model 2 is potentially more appropriate. The Results of this analysis can be seen in Table 2. There appears to be some benefit in lowering cholesterol in the high risk group, but not in the medium or low risk groups.

Table 2 Odds Ratios for Effect of Cholesterol Lowering Stratified by Baseline Risk

	Number of Trials	Number of Subjects	Odds ratio (95% CI)
HR	10	5116	0.75 (0.62-0.91)
MR	15	24090	0.88 (0.76-1.02)
LR	9	27870	1.18 (0.93-1.49)

Figure 1 shows the observed and model based estimates for the odds ratios and 95% confidence intervals for each study together with the pooled estimate. Note that all the model based estimates are shrunk towards their corresponding overall risk group estimate. As already mentioned the model based confidence intervals may be too narrow as no account has been taken of the uncertainty regarding $\hat{\sigma}_s^2$.

Discussion

We have seen that the results of *ML3* compare approximately with those using standard random effects models for meta-analysis. However, the great advantage of *ML3* is that mixed effect models, such as model 2, can be easily accommodated and more complicated scenarios modelled. For example in the meta-analysis reported here some of the trials used drug therapy while others used diet regimens to lower cholesterol, and differences between the types of trials could also be assessed by using multilevel models.

Though not advisable, in some situations there are relatively few studies available for inclusion into a meta-analysis, and therefore estimation of parameters can prove difficult. In such situations methods based on simulation can provide an alternative means of modelling (Gilks, Thomas & Spiegelhalter, 1994).

References

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models, *JASA*, 88:9-25.

Carlin, J.B. (1992) Meta-analysis for 2x2 tables: a Bayesian approach, *Statistics in Medicine,* 11, 141-158.

Davey Smith, G., Song, F. and Sheldon, T.A. (1993) Cholesterol lowering and mortality: the importance of considering initial level of risk, *British Medical Journal* 306: 1367-1373.

DerSimion, R.D. and Laird, N.(1986) Meta-analysis in clinical trials, *Controlled Clinical Trials*, 7: 177-188.

Fleiss, J.L. (1993) The Statistical basis of meta-analysis, *Statistical Methods in Medical Research*, 2: 121-145.

Gilk, W.R., Thomas, A. and Spiegelhalter, D.J. (1994) A language and program for complex Bayesian modelling, *The statistician*, 43(1): 169 - 178.

Goldstein, H. (1995) *Multilevel Statistical Models*, London: Edward Arnold, New York: Halsted Press.

Hedges, J.V. and Olkin, I. (1985) *Statistical methods for meta-analysis*, Orlando: Academic Press.

Jones, D.R. (1995) Meta-analysis: weighting the evidence, *Statistics in Medicine*, 14, 137-149.

Rasbash, J. & Woodhouse, G. (1995) *MLn* Command Reference, Multilevel Models Project, Institute of Education, University of London.

Thompson, S.G. (1994) Why sources of heterogeneity in meta-analysis should be investigated, *British Medical Journal*, 309, 1351-5.

∞∞