MULTILEVEL MODELLING NEWSLETTER

The Multilevel Models Project
Mathematical Sciences
Institute of Education, University of London
20 Bedford Way, London WC1H 0AL, ENGLAND
Web site: http://www.ioe.ac.uk/multilevel/
Enquiries about newsletter to Ian Plewis
E-mail: i.plewis@ioe.ac.uk
Tel: +44 (0)20 7612 6688 Fax: +44 (0)20 7612 6686

Vol. 11 No. 1 **December, 1999**

IF YOU WISH TO CONTINUE TO RECEIVE THE NEWSLETTER, PLEASE SEND YOUR E-MAIL ADDRESS TO a.burch@ioe.ac.uk AS SOON AS POSSIBLE, AND READ THE EDITORIAL IN THIS ISSUE.

Editorial

I have taken over editing the newsletter from Min Yang, and I am sure you would all want to join me in thanking Min for all her hard work over the last eight years. The change in editorship coincides with a change in publishing policy. This is the last newsletter to be sent out as hard copy. From the first issue in 2000, we will send it out as an e-mail attachment in PDF format so you will need the latest version of Acrobat (freely available from http://www.adobe.com/products/acrobat/readstep.html) to be able to read it. In addition, we will put all the articles on our web site. We have had to make this change because our latest research grant from the ESRC (see page 2) does not cover printing and mailing costs for the newsletter. We hope you will not find this change too inconvenient. At present our mailing list is approximately 2100.

We plan to send out two issues a year and you might find it helpful to know what our editorial policy is. We aim for a mixture of technical and less technical material, look for material generated by the electronic discussion list (to subscribe, mail multilevel@mailbase.ac.uk with the message 'join multilevel') and we encourage articles reporting work in progress. In addition, we will always publish relevant news items. The newsletter is not a refereed journal. We do not, however, accept everything submitted to us, and we do ask two people attached to the Multilevel Models Project to give an opinion on articles submitted.

I hope you continue to find the newsletter useful, and I look forward to receiving plenty of material for future issues. It would be helpful if you could submit articles as Word documents.

IAN PLEWIS

Also In This Issue

aLCD analysis OF LARGE AND COMPLEX DATASETS

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

New ESRC Funding

The multilevel models project has recently started a new 3-year ESRC funded project, under the direction of Harvey Goldstein and Jon Rasbash, which is aimed at providing materials, including software, to help users gain understandings of complex data structures. Full time research officers are Jon Rasbash, Min Yang and Bill Browne, with support from Geoff Woodhouse and Ian Plewis. Amy Burch has been appointed as the half time administrator for the project. A wide network of collaborators has also been established.

More forthcoming Workshops

5-7 April 2000, a three-day introductory workshop to multilevel modelling using *MLwiN* will take place in the Institute of Education, University of London.
Enquiries to Anne-Lise McDonald at *Health Policy and Practice Unit, UEA, Norwich, NR4 7TJ*. Tel: +44 (0)1603 514867, Fax: +44 (0)1603 593739, e-mail: **a.cox@uea.ac.uk**.

A Non-parametric bootstrap for multilevel models

James Carpenter[1], *Harvey Goldstein*[2], *Jon Rasbash*[2]
[1] **London School of Hygiene and Tropical Medicine, London**
[2] **Institute of Education, London**
e-mail: j.carpenter@lshtm.ac.uk

1. Introduction

Bootstrapping is now a well established procedure for assessing the bias and standard error of parameters in statistical models (Davison and Hinckley, 1997). Given a fitted model and parameter estimates, the idea is to generate synthetic (termed bootstrap) data from the fitted model, and then refit the model to the synthetic data, thus obtaining a set of synthetic (termed bootstrap) parameter estimates. These synthetic parameter estimates stand in approximately the same relationship to the model parameter estimates as the model parameter estimates stand in relationship to the population parameters. Thus, we can estimate quantities of interest relating the population parameters and the estimated parameters (such as bias, confidence intervals) by looking at the relationship between the estimated parameters and the synthetic, or bootstrap, parameters.

Broadly speaking the synthetic data can be generated in one of two ways, termed the *parametric* and *non-parametric* bootstrap. The parametric bootstrap, already implemented in *MLwiN*, generates the bootstrap data from the full parametric model. For example, consider the 2-level model fitted to the tutorial data example in the *MLwiN* user's guide,

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + u_{1j} x_{1ij} + e_{ij}$$
$$(i=1,2,\ldots,I_j\ ;\ j=1,\ldots,J) \qquad (1)$$

where the response is the normalised exam score, the explanatory variable is the standardised LRT score and there are I_j pupils within school j. Suppose we have fitted the model and obtained estimates of all the parameters. Then the parametric bootstrap simulates

1) $e_{ij}^{*} \sim N(0, \hat{\sigma}_e^2)$, $(i=1,2,\ldots,I_j\ ;\ j=1,\ldots,J)$,

where $\hat{\sigma}_e^2$ is the estimate of $\sigma_e^2 = \mathrm{Var}(e_{ij})$ obtained from the data

2) $\begin{bmatrix} u^*_{0j} \\ u^*_{1j} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2_{u0} & \\ \hat{\sigma}_{u01} & \hat{\sigma}^2_{u1} \end{pmatrix}\right]$ (j=1,…,*J*),

where the $\hat{\sigma}^2_u$'s are elements of the variance-covariance matrix of the u's estimated from the data.

The bootstrap data set is then (y^*_{ij}, x_{1ij}), (i=1,2,…, I_j ; j=1,…,*J*), where

$$y^*_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{1ij} + u^*_{0j} + u^*_{1j} x_{1ij} + e^*_{ij}$$

A large number, B, typically 1000, such bootstrap data sets are generated, and the model fitted to each one. We thus obtain B bootstrap estimates of each parameter in the model, which we can then use to estimate bias, standard error and confidence intervals, as described in the *MLwiN* user's guide. We can also obtain bootstrap estimates of other quantities, such as the level 2 residuals.

Here we outline a non-parametric alternative to the parametric bootstrap, and show that it can yield a substantial reduction in the coverage error of parametric bootstrap confidence intervals when the data are not truly normally distributed.

2. A Non-parametric bootstrap for multilevel models

Non-parametric bootstrapping can take two forms. In the first kind, case re-sampling, we build a bootstrap data set from the original data by sampling with replacement from the (y_{ij}, x_{1ij}) pairs that make up the data. However, in a multilevel context doing this crudely would break the structure of the dataset; if, as an alternative, we resample 'blocks' of data, it is not at all obvious which 'level' the blocks should correspond to. Furthermore, work in the standard regression context suggests that while this approach might be useful for deciding between models, it does not give accurate inference for parameters within such models, which is our principal goal.

We therefore propose to generalise the residual non-parametric bootstrap for regression models to the multilevel case. A crude residual bootstrap for model (1) would be the following:

1) Fit the model (1) to the data, and calculate the set of residuals $\{e_{ij}\}_{i=1,\ldots I_j;\, j=1,\ldots J}$ and $\{(u_{0j}, u_{1j})\}_{j=1,\ldots J}$
2) Sample with replacement from these two sets, obtaining two new sets $\{e^*_{ij}\}_{i=1,\ldots I_j;\, j=1,\ldots J}$ and $\{(u^*_{0j}, u^*_{1j})\}_{j=1,\ldots J}$
3) The bootstrap data set is then (y^*_{ij}, x_{1ij}), where

$$y^*_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{1ij} + u^*_{0j} + u^*_{1j} x_{1ij} + e^*_{ij}$$

The drawback of this simple approach is that we will underestimate variances in particular because the crude residuals are 'shrunk' towards zero. We therefore need to 'reflate' the residuals before passing them back through the fitted model as in step (3) above. We now outline a procedure for doing this. For convenience we shall illustrate the procedure using the level 2 residuals, but analogous operations can be carried out at all levels. Rewrite model (1) as

$$y_{ij} = (X\beta)_{ij} + (ZU)_j + e_{ij} \qquad (2)$$
$$U^T = \{U_0, U_1 \ldots.\}$$

Having fitted the model we calculate the residuals:

$$\hat{U} = \{\hat{u}_0, \hat{u}_1 \ldots.\}$$

Write the empirical covariance matrix of the estimated residuals at level 2 in model (2) as

$$S = \frac{\hat{U}^T \hat{U}}{J}$$

and the corresponding model estimated covariance matrix of the random coefficients at level 2 as *R*. The empirical covariance matrix is estimated using the number of level 2 units, *J*, as divisor rather than *J-1*. We assume that the estimated residuals have

been centered, although centering will only affect the overall intercept value.

We now seek a transformation of the residuals of the form

$\hat{U}^* = \hat{U}A$

where A is an upper triangular matrix of order equal to the number of random coefficients at level 2, and such that

$$\hat{U}^{*T}\hat{U}^* / J = A^T\hat{U}^T\hat{U}A = A^T SA = R \qquad (3)$$

The new set of transformed residuals $\hat{U}^*$ now have covariance matrix equal to that estimated from the model, and we sample sets of residuals with replacement from $\hat{U}^*$, as described in the residual bootstrap algorithm above.

To complete the residual bootstrap, we repeat this process at every level of the model, with sampling being independent across levels. Details of how to form A are given in the appendix below.

3. Example

Consider the following 2-level model fitted to the tutorial data example in the *MLwiN* User's Guide, using RIGLS estimates. The model is

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + u_{1j}x_{1ij} + e_{ij}\,. \qquad (4)$$

We simulate data from this model using the parameter estimates given in the second column of Table 1, with residuals at level 2 simulated from

$$N\left[\begin{pmatrix}1.0\\0.5\end{pmatrix},\begin{pmatrix}0.2 & \\ 0.05 & 0.2\end{pmatrix}\right]$$

and at level 1 we simulate from a chi-squared distribution with 1 degree of freedom**.**

Five hundred data sets were generated from this model, containing 4059 level 1 and 65 level 2 units. For each of these data sets the bootstrap parameter estimates and confidence intervals were constructed using 500 parametric and 500 non-parametric bootstrap data sets.

Table 1 gives the parameter estimates and estimated coverage probability for a nominal 95% interval computed directly from the ranked bootstrap replications for each bootstrap set, for the parametric bootstrap and Table 2 for the non-parametric bootstrap.

Both bootstrap procedures produce unbiased estimates for all the parameters. The coverage proportions are satisfactory except for the level 1 variance in the parametric bootstrap where it is only 0.55 compared to the nominal value of 0.95.

Table 1. Parametric bootstrap estimates

Parameter	Expected value*	Bootstrap mean	Coverage proportion
β_0	2.00	2.00	0.95
β_1	0.50	0.500	0.93
σ^2_{u0}	0.20	0.200	0.94
σ_{u01}	0.05	0.049	0.96
σ^2_{u1}	0.20	0.202	0.95
σ^2_e	2.00	2.00	0.55

***The expected value for a chi squared distribution with 1 degree of freedom (=1) is added to the intercept. The variance of a chi squared distribution with 1 degree of freedom is 2.**

Table 2. Non-parametric bootstrap estimates			
Parameter	Expected value*	Bootstrap mean	Coverage proportion
β_0	2.00	2.00	0.95
β_1	0.50	0.500	0.95
σ^2_{u0}	0.20	0.198	0.93
σ_{u01}	0.05	0.050	0.95
σ^2_{u1}	0.20	0.202	0.94
σ^2_e	2.00	1.99	0.93
***The expected value for a chi squared distribution with 1 degree of freedom (=1) is added to the intercept. The variance of a chi squared distribution with 1 degree of freedom is 2.**			

4. Conclusions

We have briefly described a residuals non-parametric bootstrap for multilevel models. This residuals bootstrap provides a robust alternative to a fully parametric bootstrap, and can be used, for example where standardised residual plots indicate departures from normality. The bootstrap can also be used to estimate other functions. For example we can estimate residuals for each bootstrap replicate and use the resulting chains for inference about the residual estimates themselves.

This non-parametric bootstrap procedure is implemented in *MLwiN* release 1.1 (Autumn 1999).

5. References

Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and their Application*, Cambridge University Press, Cambridge, 1997.

6. Appendix

To form *A* we note the following.
Write the Cholesky decomposition of *S*, in terms of a lower triangular matrix as

$$S = L_S L_S^T$$

and the Cholesky decomposition of *R* as

$$R = L_R L_R^T$$

We have

$$L_R L_S^{-1} \hat{U}^T \hat{U} (L_R L_S^{-1})^T / J =$$
$$L_R L_S^{-1} S (L_S^{-1})^T (L_R)^T = L_R (L_R)^T = R$$

Thus, the required matrix is

$$A = (L_R L_S^{-1})^T .$$

Random effects meta-analysis of trials with binary outcomes using multilevel models in MLwiN

Rebecca M Turner[1], *Rumana Z Omar*[1], *Min Yang*[2], *Harvey Goldstein*[2], *Simon G Thompson*[1]
[1] **Imperial College School of Medicine, London**
[2] **Institute of Education, London**
e-mail: rebecca.turner@ic.ac.uk

Introduction

In meta-analysis we aim to combine the results from a series of similar trials and draw a unified conclusion from an overall estimate of treatment effect. Multilevel models provide a general framework for meta-analysis; meta-analysis corresponds to a multilevel model (patients within trials) whether summary data at trial level or individual patient data are available. A multilevel modelling approach also allows meta-analysis of a combination of summary and individual patient data (Goldstein *et al.*, 1999). The aim of this article is to describe implementation of random effects meta-analysis methods, and related confidence interval construction, in MLwiN (Turner *et al.*, 1999).

Summary data methods

The traditional random effects model for meta-analysis assumes that the true treatment effects vary randomly between trials. Here we consider the log odds ratio to be the measure of treatment effect. Assuming normality of the log odds ratio, the model can be written for trials i=1…n:

$$y_i \sim \mathrm{N}\left(\theta + v_i, \sigma_i^2\right)$$
$$v_i \sim \mathrm{N}\left(0, \tau^2\right)$$

where y_i is the observed log odds ratio from the ith trial, σ_i^2 is its (within trial) variance, usually assumed known and τ^2 is the between trial variance of the true log odds ratios. Maximum likelihood (ML) or restricted maximum likelihood (REML) estimates of the parameters can be obtained within MLn/MLwiN using IGLS or RIGLS respectively; implementation of this model is described by Lambert and Abrams (1995). The conventional method of estimation of this model is non-iterative, employing a method of moments estimator for the between trial variance τ^2 (DerSimonian & Laird, 1986).

It is possible within MLn to construct likelihood based and parametric bootstrap confidence intervals for both θ and τ^2, as well as Wald intervals derived from asymptotic standard errors. For θ, likelihood and bootstrap intervals are preferable to Wald intervals since they allow for the imprecision in estimation of τ^2; bootstrap intervals also relax the assumption of normality required when interpreting likelihood intervals as confidence intervals. Both Wald and likelihood methods are problematic for τ^2 since the normality assumption is likely to be invalid; therefore bootstrap intervals are preferred. When using REML estimation, likelihood intervals should not be constructed for θ since REML likelihood ratio tests may not be strictly valid for fixed effects (Welham & Thompson, 1997). REML likelihood based intervals may be computed for τ^2 (Morrell, 1998). However REML log likelihood is not available in MLn or MLwiN.

The procedure for obtaining likelihood based intervals is described below and may be used directly in MLn or within the Command interface window in MLwiN. To construct bootstrap intervals we use the parametric bootstrap (Efron & Tibshirani, 1993), in which a series of data sets are simulated under the distributional assumptions of the initial model. This is directly available in the

MLwiN software but can alternatively be implemented as an MLn macro. A bootstrap set of parameter estimates are generated, and confidence intervals may be based on the smoothed percentiles of the bootstrap distributions.

Construction of likelihood based intervals in MLn/MLwiN

First we fit the simple random effects model as described by Lambert and Abrams (1995): we use trial identifiers at level 2, CONS (vector of 1s) as identifying variable at level 1, LOR (log odds ratios) as response, CONS as a fixed effect and as a random effect at level 2, LOR_SE (standard errors of log odds ratios) as a random effect at level 1 with the corresponding parameter estimate constrained to equal 1. We use IGLS to provide maximum likelihood estimation and note the resulting value of –2*log(likelihood), call this *L*. To find a 95% likelihood based interval for θ, we calculate the maximised log likelihood for different possible values $\tilde{\theta}$ until we find values (θ^L, θ^U) for which –2*log (likelihood) is equal to $L \pm 3.84$, where 3.84 is the 95% point of the χ_1^2 distribution. A similar procedure is used for τ^2.

To examine the likelihood corresponding to a value $\tilde{\theta}$, we use the command FCON to constrain the fixed parameter to equal $\tilde{\theta}$. To try the value –0.5 for example, assuming that the model still has only one fixed effect CONS, we store the vector $(1,–0.5)^T$ in an empty column (c20 say) and then type 'FCON c20'. Unfortunately we cannot use the corresponding command RCON to constrain the level 2 random parameter to possible values $\tilde{\tau}^2$, because the level 1 random parameter is already constrained and models with all random parameters constrained may not be estimated in MLn. Instead we use the value $\tilde{\tau}^2$ as an offset at level 2. First remove CONS from the level 2 random part, create a new column ZERO (vector of zeros) of the same length and declare this to be random at level 2. Then, for example, to try the value $\tilde{\tau}^2$=0.8, create a column (c21 say) of the same length as CONS in which every entry is 0.8 and type 'OFFS 2 c21'.

Example

To demonstrate the methods we use a data set consisting of 22 trials for prevention of respiratory tract infection. Patients in intensive care units were randomised to receive treatment by a combination of non-absorbable antibiotics or to receive no treatment (Smith *et al.*, 1995). The table below presents results from employing either conventional non-iterative moment estimation or maximum likelihood estimation in MLn. The bootstrap intervals are based on 1000 replications.

Differences between the widths of the two Wald confidence intervals for θ merely reflect differences in the estimates of τ^2. The bootstrap and Wald intervals for τ^2 are of similar width, but both bootstrap and likelihood intervals cover ranges further from zero than those covered by the Wald intervals. The discrepancy between likelihood and bootstrap intervals is greater for τ^2 than for θ because the normality assumption required for interpretation of likelihood intervals is less sound for τ^2.

Estimates and 95% confidence intervals from summary data meta-analysis example

		Log OR (θ)	(95% CI)	Between trial variance (τ^2)	(95% CI)
Conventional moment estimation	(Wald CI)	-1.27	(-1.61, -0.92)	0.36	
ML estimation	(Wald CI)	-1.29	(-1.65, -0.92)	0.42	(0.01, 0.83)
	(likelihood based CI)		(-1.73, -0.92)		(0.12, 1.19)
	(bootstrap CI)		(-1.63, -0.92)		(0.06, 0.86)

Individual data methods

When a binary outcome meta-analysis data set is set up with one record per patient, the above model may be fitted as a random effects logistic regression, involving terms v_i representing the deviation of each trial's true log odds ratio from the average. We need to allow the log odds to vary across trials as well as allowing the log odds ratio to vary; these trial effects can be assumed either fixed or drawn from a random distribution. If the former, we write y_{ij}=0/1 to indicate the outcome for the *j*th individual in the *i*th trial where π_{ij} = Pr(y_{ij}=1), and include as explanatory variables TREAT (treatment group indicator x_{ij}=0/1) together with *n* dummy variables w_{ijk}=0/1, *k*=1…*n* representing trial effects, we then also declare TREAT as random at level 2:

$$\text{logit}(\pi_{ij}) = (\theta + v_i)x_{ij} + \sum_{k=1}^{n} \beta_k w_{ijk}$$

$$v_i \sim \text{N}(0, \tau^2)$$

A similar procedure can be used for grouped binary data, where *j* represents the group (treatment or control) rather than individual. Since even PQL with second-order approximations may give downwardly biased estimates of the between trial variance τ^2 when the number of trials is small or when the probabilities of events are extreme, we used the bias corrected parametric bootstrap procedure now directly available in MLwiN (Goldstein, 1996). Bias corrected estimates may be obtained and confidence intervals for θ and τ^2 constructed from the quantiles of the bootstrap distributions using a method suggested by Kuk (1995). Likelihood based intervals may not be used since reliable log likelihood values are not available in MLn/MLwiN for multilevel models with binary outcomes (Goldstein, 1995).

Rather than using fixed trial effects to allow the log odds to vary across trials, we can instead assume these to be random by removing the *n* trial indicators and adding an explanatory variable CONS (vector of 1s) as both a fixed effect and a random effect at level 2:

$$\text{logit}(\pi_{ij}) = \alpha + u_i + (\theta + v_i)x_{ij}$$

$$u_i \sim \text{N}(0, \sigma^2), v_i \sim \text{N}(0, \tau^2), \text{Cov}(u_i, v_i) = \rho\sigma\tau$$

It is important to include $\text{Cov}(u_i, v_i)$ rather than assume this to be zero, because of implications for the between trial variance-covariance matrix of the bivariate log odds parameter (i.e. log odds in control groups, log odds in intervention groups). In data sets with few trials it may not be possible to achieve convergence when including $\text{Cov}(u_i, v_i)$; we experienced this in a data set involving nine trials (Thompson & Pocock, 1991). If in such cases $\text{Cov}(u_i, v_i)$ must be

excluded from the model, it may be preferable to use ±½ rather than 0/1 coding for the treatment covariate TREAT, thereby placing more realistic assumptions on the variance-covariance matrix.

Summary

This article describes how to perform the basic methods required for meta-analysis of binary outcome data within the multilevel models framework. The principal advantage of multilevel modelling for meta-analysis is its flexibility which enables a number of desirable extensions. Individual-level and trial-level covariates are easily included, three-level models are a possibility for meta-analysis of multi-centre or cluster randomised trials, and methods are available for more complex outcomes, such as ordinal, survival or mixed multivariate data.

References

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

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

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

Goldstein, H. (1996). Consistent estimators for multilevel generalised linear models using an iterated bootstrap, *Multilevel Modelling Newsletter*, 8, 3-6.

Goldstein, H., Yang, M., Omar, R. Z., Turner, R. M., and Thompson, S. G. (1999). Meta-analysis using multilevel models with an application to the study of class size effects, *Journal of the Royal Statistical Society, Series C*, in press.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalised linear models with random effects, *Journal of the Royal Statistical Society, Series B*, 57, 395-407.

Lambert, P. C. and Abrams, K. R. (1995). Meta-analysis using multilevel models, *Multilevel Modelling Newsletter*, 7, 17-19.

Morrell, C. H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood, *Biometrics*, 54, 1560-1568.

Smith, T. C., Spiegelhalter, D. J., and Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study, *Statistics in Medicine*, 14, 2685-2699.

Thompson, S. G. and Pocock, S. J. (1991). Can meta-analyses be trusted?, *Lancet*, 338, 1127-1130.

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H., and Thompson, S. G. A multilevel model framework for meta-analysis of clinical trials with binary outcomes, *submitted for publication*, 1999.

Welham, S. J. and Thompson, R. (1997). Likelihood ratio tests for fixed model terms using residual maximum likelihood, *Journal of the Royal Statistical Society, Series B*, 59, 701-714.

Standard errors in multilevel analysis

Nick Longford
De Montfort University, Leicester
ntl@dmu.ac.uk

The subject of this note is estimation of standard errors in multilevel analysis, and its essence is that the traditional approach has a profound flaw which is ubiquitous in multilevel analysis as we use it at present. I illustrate this on simple examples.

By standard error of a (general) estimator $\hat{\beta}$ I mean the square root of its sampling variance, $\sqrt{\text{var}(\hat{\beta})}$. When we attach interpretation to the 'standard errors' produced by a (*any*) software package, most of us are not aware that the standard errors are themselves *estimated*. I agree that assessing the precision of the standard errors may appear to be taking things too far, but what if the standard errors are estimated so poorly as to render them totally useless? The standard errors are not only *estimated* but also *asymptotic*; however, the errors due to the asymptotics not holding for the realised sample (size) are often minute in comparison to the errors due to estimation.

For a simple illustration, consider the 'empty' two-level model

$\mathbf{y}_j = \mu + \delta_j + \varepsilon_j$;

$\mathbf{y}_j$ are n_j x 1 vectors of outcomes $(j = 1,...,N_2)$, and $\{\delta_j\}$ and the $N = \Sigma_j n_j$ elements of $\{\varepsilon_j\}$ are two mutually independent random samples from $N(0,\sigma_2^2)$ and $N(0,\sigma^2)$ respectively. Instead of σ_2^2 it is practical to estimate the variance ratio $\tau = \sigma_2^2/\sigma^2$ because it is unaffected by a linear transformation of the outcomes.

The asymptotic standard error of the ML estimator of τ is

$$SE(\hat{\tau};\tau) = \left\{\frac{1}{2}\sum_{j=1}^{n}\frac{n_j^2}{(1+\tau n_j)^2}\right\}^{-\frac{1}{2}} \qquad (1)$$

(Longford, 1993, Section 2.3). As a matter of routine, it is estimated by $SE(\hat{\tau};\hat{\tau})$, given by default in most software implementations. In the balanced design, when $n_1 = n_2 = ... = n_m (= n)$, equation (1) simplifies to

$$SE(\hat{\tau};\tau) = \sqrt{\frac{2}{m}\left(\frac{1}{n}+\tau\right)} \qquad (2)$$

Thus, the estimator $\hat{\tau}$ is perfectly correlated with the estimated standard error $SE(\hat{\tau};\hat{\tau})$; the error committed in estimating τ is present also in estimating its standard error. For unbalanced designs, equation (2) does not apply, but a good approximation is obtained by replacing n with the arithmetic average of the sample sizes n_j, even when the imbalance is severe.

In Figure 1, the (exact) values of $SE(\hat{\tau};\tau)$ given by (1) are compared with the expression (2) in which n is replaced by the arithmetic, geometric, and harmonic averages of the sample sizes. The cluster-level sample sizes used,

$n = (5,6,6,6,7,8,10,10,10,10,10,11,11,11,11,11,11,14,14,14)$

$(m = 20, N = 196)$

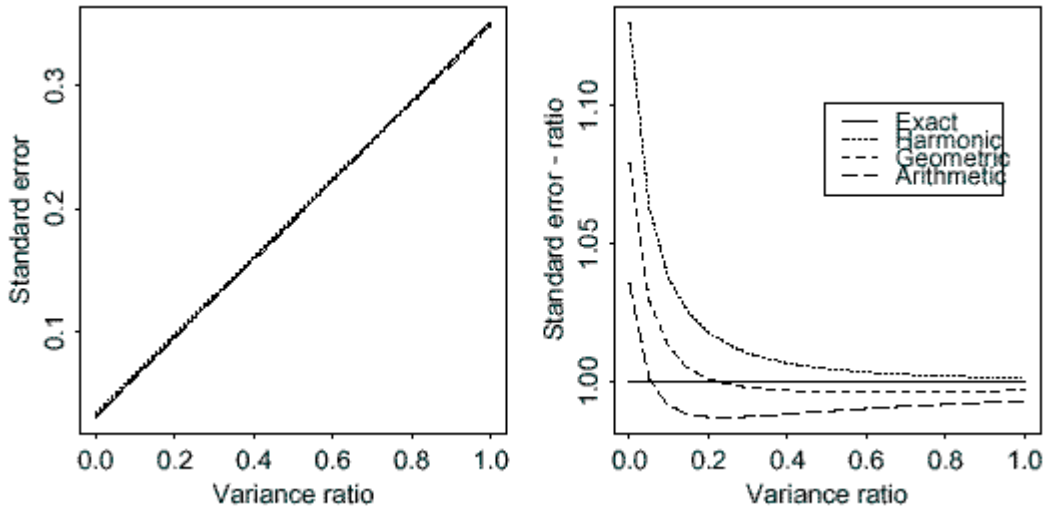
were generated by a random process.

In the left-hand panel, the standard errors are plotted against the variance ratio for $0 \le \tau \le 1$. (In general, $0 \le \tau + \infty$.) The range of the exact (and approximate) standard errors is so wide $(0.031 - 0.350)$ that the four linear functions are indistinguishable. Better insight is gained by plotting the ratios of the approximations to the exact values, as done in the right-hand panel. The

approximations dominate the exact values for τ close to zero, by not more than 4% for the arithmetic average. For $\tau > 0.2$, the approximations are very good; the error is smaller than 2% for either average. These errors are minute in comparison with the errors $SE(\hat{\tau};\hat{\tau}) - SE(\hat{\tau};\tau)$.

Suppose that in fact $\tau = 0.15$. Then $SE(\hat{\tau};\tau) = 0.081$, so the value of $\hat{\tau} = 0$ would be judged as quite feasible. If we do obtain an estimate $\tau \cong 0$, then the standard error is estimated by $SE(\hat{\tau},\hat{\tau}) = 0.031$; it suggests that $\tau = 0.15$ is not feasible. This paradox can be resolved by evaluating (1), or its approximation based on (2), for a range of values of τ, and then ascertaining for which τ the estimate obtained (say, $\tau \cong 0$) is feasible. For instance, $SE(\hat{\tau},0.2) = 0.097$, the z-ratio is $0.2/0.097 \cong 2.0$, so $\tau = 0.2$ is on the borderline of feasibility.

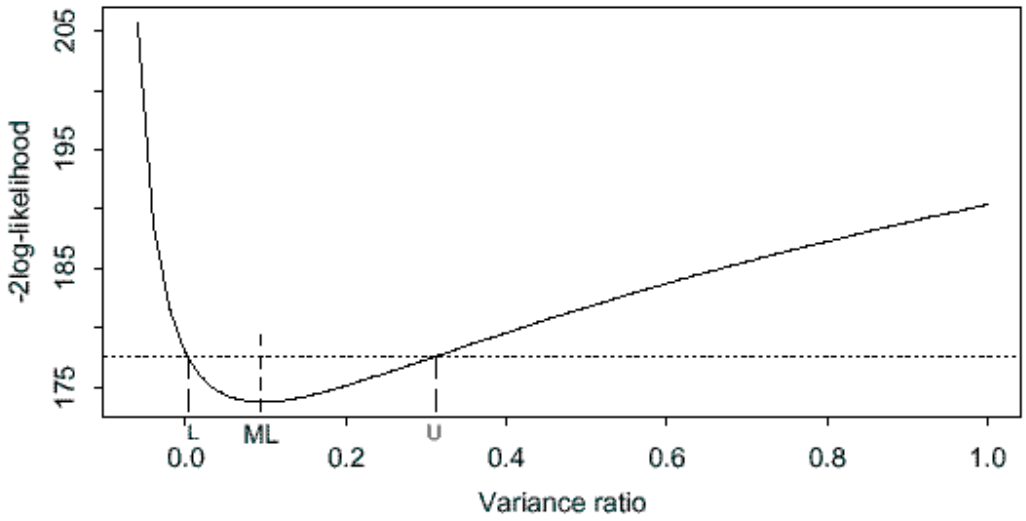
Figure 1: Exact asymptotic standard errors for the variance ratio $\hat{\tau}$, and its approximations.



The dependence of $\text{var}(\hat{\tau})$ on τ can be reduced (though not removed completely) by estimating $\gamma = \log(\bar{n}^{-1} + \tau)$, where $\bar{n} = (n_1 + ... + n_m)/m$. Then, asymptotically, $\text{var}(\hat{\gamma}) = 2/m$. But this is a rather unnatural a scale.

Of course, the problem with misleading $SE(\hat{\tau},\hat{\tau})$ would not arise had we explored the profile likelihood for τ. This is done in Figure 2 for a simulated example with the 20 within-cluster sample sizes listed above. The ML estimator is $\hat{\tau} = 0.094$ and the conventional estimate of the standard error is $SE(\hat{\tau},\hat{\tau}) = 0.062$. The symmetric 95% confidence interval based on these estimates is $(-0.029, 0.216)$. The confidence interval based on the profile likelihood, $(0.003, 0.311)$, is much more appropriate; see Lindsey (1999) for a discussion.

Figure 2: The profile likelihood for the variance ratio τ.
The horizontal line is drawn 3.84 above the minimum deviance attained at the ML estimator $0.093('ML')$. The lower and upper limits of the 95% confidence interval are $0.003('L')$ and $0.311('U')$.



The discrepancy between the two ways of generating confidence intervals persists when REML is applied, or when the variance $\sigma_2^2 = \sigma^2\tau$ or the proportion of the aggregate-level variance $\rho = \tau/(1+\tau)$ or their square roots are estimated. Equation (1) holds even when the model is supplemented by regressors:

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \delta_j + \boldsymbol{\varepsilon}_j, \qquad (3)$$

so all conclusions about $SE(\hat{\tau},\hat{\tau})$ apply also to these models. Between-cluster variation is often a nuisance, so estimation of τ is often of limited concern. However, the estimate of the regression parameter vector $\boldsymbol{\beta}$ and of its sampling variance matrix depend on the (unreliable) estimator $\hat{\tau}$.

The ML and REML estimators of $\boldsymbol{\beta}$ in a general two-level model are given by the expression

$$\hat{\boldsymbol{\beta}} = \left(\sum_j \mathbf{X}_j^T\hat{\mathbf{V}}_j^{-1}\mathbf{X}_j\right)^{-1}\sum_j \mathbf{X}_j^T\hat{\mathbf{V}}_j^{-1}\mathbf{y}_j,$$

where $\hat{\mathbf{V}}_j$ is the ML or REML estimator of $\mathrm{var}(\mathbf{y}_j)$; $\mathrm{var}(\hat{\boldsymbol{\beta}})$ is estimated by the inverse in the above expression. For the model (3) with a single covariate x,

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2\begin{pmatrix} \Sigma_j w_j & \Sigma_j w_j\bar{x}_j \\ \Sigma_j w_j\bar{x}_j & \mathbf{x}^T\mathbf{x} - \tau\Sigma_j n_j w_j\bar{x}_j^2 \end{pmatrix}^{-1} \qquad (4)$$

where $\mathbf{x}$ is the vector of values of the covariate, $\bar{x}_j$ its mean in cluster j, and $w_j = n_j/(1+n_j\tau)$. In the balanced case, when $n_1 = n_2 = \ldots = n_m$ *and* the within-cluster totals $\mathbf{x}_j^T\mathbf{1}_{n_j}$ are constant, (4) reduces to

$$\sigma^2\begin{pmatrix} mw & 0 \\ 0 & \mathbf{x}^T\mathbf{x} - N\bar{x}^2 \end{pmatrix}^{-1},$$

where w is the common value of the w_j. Similarly, in the balanced case,

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = \sigma^{-2}\begin{pmatrix} Nw\bar{y} \\ \mathbf{x}^T\mathbf{y} - N\bar{x}\bar{y} \end{pmatrix}$$

($\bar{y}$ is the sample mean of y). Hence, in the balanced case, the estimate of the slope β_2 does not depend on τ, and coincides with the ordinary least squares estimator. When x is close to balance, $\hat{\tau}$ has little impact on estimating the slope β_2.

Without balance, an adequate approximation to $\text{var}(\beta_2)$ is

$$\frac{\sigma^2}{N\sigma^2_{W,x} + (m-1)\sigma^2_{B,x,\tau}},$$

where $\sigma^2_{W,x}$ is the within-cluster variance of the covariate and $\sigma^2_{B,x,\tau}$ the between-cluster variance, calculated with weights $1/(1+n_j\tau)$.

For a cluster-level covariate x, $\hat{\boldsymbol{\beta}}$ is profoundly affected by $\hat{\tau}$ only in some esoteric situations. In the balanced case, $n_j \equiv n$, $\hat{\boldsymbol{\beta}}$ does not involve $\hat{\tau}$. However, the sampling variance of $\hat{\beta}_2$, $(\bar{n}^{-1} + \tau)/(\mathbf{x}^T\mathbf{x} - N\bar{x}^2)$ does depend on τ. When the cluster-level sample size m is moderate or small substituting a subjectively chosen value of τ may yield a more reliable estimated standard error than substituting the unreliable estimate $\hat{\tau}$. Or, we can consider the entire range of feasible values of τ.

Similar conclusions are arrived at for estimating the regression parameter for a covariate that is associated with variation. For the model

$$\mathbf{y}_j = \beta_1 + \beta_2\mathbf{x}_j + \delta_{j,1} + \delta_{j,2}\mathbf{x}_j + \varepsilon_j,$$

with the usual assumptions, $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\Sigma_j \mathbf{G}_j^{-1}\mathbf{X}_j^T\mathbf{X}_j)^{-1}$, where $\mathbf{G}_j = \mathbf{I}_2 + \boldsymbol{\Omega}\mathbf{X}_j^T\mathbf{X}_j$ (Longford, 1993, Ch. 4). In the balanced case, when $\mathbf{X}_j^T\mathbf{X}_j$ is constant,

$$\text{var}(\hat{\beta}_2) = \frac{(N-1)\sigma^2}{\hat{\text{var}}(x)} + \frac{\sigma^2}{m}\Omega_{2,2},$$

where $\Omega_{2,2}$ is the $(2,2)$ element of $\boldsymbol{\Omega}$ and $\hat{\text{var}}(x)$ is the sample variance of $\mathbf{x}$. So, the standard error of $\hat{\beta}_2$ depends on an (often poorly) estimated $\Omega_{2,2}$.

In the balanced case, $\hat{\beta}_2$ does not depend on $\hat{\boldsymbol{\Omega}}$. Heuristically, as we depart from balance $\hat{\beta}_2$ depends more and more strongly on $\hat{\boldsymbol{\Omega}}$.

In approximations to the standard error of $\hat{\beta}_2$ we may consider an average matrix $\overline{\mathbf{G}}$, so long as the magnitudes of the within-cluster cross-product matrices $\mathbf{X}_j^T\mathbf{X}_j$ do not vary a great deal.

Of course, these problems are avoided if inferences are based on the profile likelihood. The results and their discussion suggest that in approximately balanced designs $\text{var}(\hat{\beta}_2)$ depends most strongly on $\boldsymbol{\Omega}_{2,2}$, so exploring the (one-dimensional) profile likelihood on it is useful. As we depart from balance, the two other unique elements of $\boldsymbol{\Omega}$ exert their influence, not only on the estimated standard error of $\hat{\beta}_2$, but on the estimate itself.

In conclusion, the estimated standard errors for variance parameters have a strong element of 'self-fulfilling prophecy' - smaller estimate is associated with smaller *estimated* standard error. The error in estimating between-cluster variation parameters is present also in estimating the sampling variance of the regression parameters. Variation balance (constant cross-products $\mathbf{Z}_j^T\mathbf{Z}_j$) obviate some (but not all) of these problems. Likelihood ratio and exploration of the profile likelihood are much more appropriate tools for model comparisons.

References

Lindsey, J. K. (1999) Some statistical heresies. *The Statistician* 48, 1-40.

Longford, N. T. (1993). *Random Coefficient Models*. Clarendon Press, Oxford.

Review of 'Multilevel Analysis'
T A B Snijders & R B Bosker. Pp ix & 266. London: Sage, 1999.

M J R Healy
Multilevel Models Project, Institute of Education, London

It is a sign of the maturity of multilevel modelling techniques that introductory textbooks are starting to appear. Hox's *Applied Multilevel Analysis* dates back to 1994, Kreft and de Leeuw's *Introducing Multilevel Modelling* appeared in 1998 and now Tom Snijders and his colleague Roel Bosker have produced this *Introduction to Basic and Advanced Multilevel Modelling*. As an indication of its expected audience, the reader should have 'a good working knowledge of statistics', including 'the basics of hypothesis testing and multiple regression analysis' and should be able to 'understand formulae of the kind that occur' in this context.

Chapters 1 to 3 are an introduction to the main concepts of multilevel data and associated models, pointing out the importance of allowing for the variance structure in the analysis. Chapter 4 introduces the random intercept model and chapter 5 extends this to random slopes. Chapter 6 describes various significance tests and includes a useful discussion of the issues underlying model building. These chapters constitute the more elementary part of the book.

Chapter 7 is something of a digression into assessment of explained variance. Chapter 8 deals with complex variances and chapter 9 gives a deeper treatment of model checking. Chapter 10 contains a useful but complex discussion of study design and sample size determination. Crossed models are the subject of chapter 11, longitudinal data are treated in chapter 12, chapter 13 describes multivariate models and chapter 14 is devoted to discrete data. The final chapter describes several of the generally available software packages.

Many examples are given, some of them available along with MLwiN and HLM programs on a website. Most of these, apart from some which use artificial data, are based on social science material, largely from educational studies. The datasets are usually large and the models often complex. Interpretation of the results of a multilevel analysis is often tricky and a fuller treatment of some of the examples would have been useful.

It will be seen that most of the topics encountered in the handling of multilevel data are covered in the book (a notable exception is meta-analysis which receives no more than a passing mention). The level of sophistication assumed is fairly high; the reader must not be put off by quite lengthy algebraic equations, nor by the elaborate notation of suffices, superfices, bars and hats which seem inevitable in the multilevel context. Variance being defined as what any two statisticians are at, I would quarrel mildly with some of the material included, such as Fisher's combination of probabilities test which has the unfortunate property of ignoring the (possibly inconsistent) signs of the effects to be combined. But putting together a book with this coverage is a horribly difficult exercise, and the authors have tackled it very successfully. I would have no hesitation in recommending it to beginners in multilevel analysis – at least, to those with a statistical or mathematical background – and I shall keep my copy handy for reference in the future.

Modelling Structured Categorical Data

Following the successful one-day conference on 'Applications of Random Effects/Multilevel Models to Categorical Data in Social Science and Medicine' in October 1998, the most recent issue (Volume 162, Part 3) of 'Statistics in Society' (Journal of the Royal Statistical Society, Series A) is devoted to a selection of papers presented then. The issue is available from Blackwell Publishers Journals (e-mail: jnlinfo@blackwellpublishers.co.uk). (*Ian Plewis, Institute of Education, University of London*)

Some Recent Publications Using Multilevel Models

Cohen, M. P. (1998). Determining sample sizes for surveys with data analyzed by hierarchical linear models. *Journal of Official Statistics*, **14** (3), 263-275.

Fieldhouse, E. & Gould, M. I. (1998). Ethnic minority unemployment and local labour market conditions in Great Britain. *Environment and Planning, A,* **30**: 833-853.

Fielding, A. (1998). Methodological innovations in the use of teaching groups for evaluating cost-effectiveness. *International Congress on School Effectiveness and Improvement, 1998 Proceedings*, CD ROM ISBN 0-9-2252-62-3, School of Education, University of Manchester.

Fielding, A. (1998). Perspectives on performance indicators: GCE Advanced Level and differences between institution types in cost effectiveness. *School Effectiveness and School Improvement*, **9** (2): 218-231.

Fielding, A., Belfield, C. R. & Thomas, H. R. (1998). The consequences of drop-outs on the cost-effectiveness of 16-19 colleges. *Oxford Review of Education*, **24** (4): 487-511.

Gould, M. I. & Fieldhouse, E. (1997). Using the 1991 Census SAR in a multilevel analysis of male unemployment. *Environment and Planning, A*, **29**: 611-628.

Gould, M. I. & Jones, K. (1996). Analysing perceived limiting long-term illness using UK Census Microdata. *Social Science and Medicine*, **42**: 857-869.

Gould, M. I., Jones, K. & Moon, G. (1997). Guest editorial: The scope of multilevel models. *Environment and Planning, A,* **29**: 581-584.

Hardy, M. A. & Hazelrigg, L. E. (1999). A Multilevel model of early retirement decisions among autoworkers in plants with different futures. *Research on Aging*, **21**: 275-303.

Hedeker, D. & Mermelstein, R. (1998). A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research*, **33** (4), 427-455.

Huang, G. G., Weng, S., Zhang, F. & Cohen, M. P. (1997). Outmigration among high school graduates: the effect of academic and vocational programs. *Education Evaluation and Policy Analysis*, **19** (4), 360-372.

Jardí, A. C. I. & Longford, N. T. (1998). Monitoring the university admission process in Spain. *Higher Education in Europe*, **23** (3): 385-395.

Johnston, R. J., Jones, K. & Gould, M. (1995). Department size and research in English Universities: inter-university variations. *Quality in Higher Education*, **1**: 41-47.

Jones, K., Gould M. I. & Watt, R. (1998). Multiple contexts as cross-classified models: the Labour vote in the British General Election of 1992. *Geographical Analysis*, **30**: 65-93.

Spencer, A. & Fielding, A. (1998). Comparison of modelling strategies for value added analyses of educational data. *Statistical Modelling: Proceeding of the 13th International Workshop on Statistical Modelling*, B. Marx & H. Friedl (eds.), Baton Rouge, Louisiana State University.

Van den Oord, E. J. C. G., & Rowe, D. C. (1998). An examination of genotype-environment interactions for academic achievement in an U.S. national longitudinal survey. *Intelligence*, **25** (3): 205-228.

Williams, D. R. and Goldstein, M. (1999) Graphical Diagnostics for the Bayes Linear Analysis of Hierarchical Linear Models with Applications to Educational Data, *Bayesian Statistics 6*, J. M. Bernardo et al (eds.). Oxford University Press, 859-867.

Wright, D. B. (1998). Modelling clustered data in autobiographical memory research: the multilevel approach. *Applied Cognitive Psychology*, **12**: 339-357.

Yang, M., Goldstein, H., Rath, T. and Hill, N. (1999). The Use of Assessment Data for School Improvement Purposes. Oxford Review of Education, **25**: 469-483.

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

MLwiN/MLn Clinics in London 1999/2000

Tuesday 11 January 2000
Tuesday 8 February 2000
Tuesday 7 March 2000
Tuesday 4 April 2000

at

Multilevel Models Project
11 Woburn Square, London WC1A 0SN

Contact Min Yang for appointment
Tel: (0)20 7612 6682 / 6657, Email: **m.yang@ioe.ac.uk**

IF YOU WISH TO CONTINUE TO RECEIVE THE NEWSLETTER, PLEASE SEND YOUR E-MAIL ADDRESS TO a.burch@ioe.ac.uk AS SOON AS POSSIBLE, AND READ THE EDITORIAL IN THIS ISSUE.