MULTILEVEL MODELLING NEWSLETTER

Centre for Multilevel Modelling

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

Vol. 16 No. 2 **December, 2004**

Forthcoming Workshops

Multilevel Discrete-time Event History Analysis workshop

10-11 February 2005. A free two-day Multilevel Discrete-time Event History Analysis workshop will take place at Institute of Education. The workshop is now fully booked, with priority given to researchers who intend to use event history analysis in their own research.
Materials will be available to download after the workshop from:
http://multilevel.ioe.ac.uk/team/mmmpceh.html#workshop

Multilevel Modelling Workshop

6-8 April 2005. A three-day introductory workshop in multilevel modelling using *MLwiN* will take place at University of Bristol.
Enquiries to: Theresa Andrews, School of Geographical Sciences, University of Bristol, University Road, Bristol BS8 1SS, United Kingdom
Tel: +44 (0) 117 954 5977
Fax: +44 (0) 117 928 7878
Email: theresa.andrews@bristol.ac.uk

If you plan to run any workshops using *MLwiN*, please notify Amy Burch a.burch@ioe.ac.uk and she will advertise these workshops on the multilevel web site.

Publications reminder

Don't forget to send us any details of publications in multilevel modelling for the next newsletter.

Also in this issue

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

Fifth International Amsterdam conference on "Multilevel Analysis" and a Workshop on multilevel latent class models

The 5th International Amsterdam Multilevel Conference is going to be held on 21-23 March 2005; the first two days with lectures, the third day with a workshop that will be taught by Jeroen Vermunt on multilevel latent class models. There are at present two invited presentations: Harvey Goldstein on "Multilevel Structural Equation Modelling via MCMC" and Jeroen Vermunt on "Multilevel Latent Class and Mixture Models".

The conference will be about all aspects of statistical multilevel analysis: theory, software, methodology, and innovative applications. It is organized by Joop J. Hox and Cora Maas (University of Utrecht) and Tom A.B. Snijders (University of Groningen).

The conference and course will be held in the Amsterdam ARENA conference centre, which is close to the centre of the city and easy to reach by public transport.

Conference: your contributions

You are invited to submit abstracts for contributed presentations.

Abstracts of 10 to 50 lines of text can be submitted until February 15, 2005. Those who submit an abstract will be notified before February 21 about the acceptance for presentation at the conference.

Abstracts should be sent as an ASCII file (not in a different word processor format!!) by email to multilevelconference@mail.fss.uu.nl.

Abstracts can also be submitted by postal mail to:
Multilevel Conference
c/o R. Holdinga
Dept. of Methodology & Statistics
FSW, UU
P.O.B. 80140
NL-3508 TC Utrecht
The Netherlands
or by fax to:
+ 31 30 2535797

Information

Further information about the conference is available on the conference web site.
http://www.uu.nl/uupublish/defaculteit/organisatie/capaciteitsgroep/methodenleerenst/methodenleerstat/multilevelconfer/34527main.html

A registration form is also available at this web site. Further information can be obtained from j.hox@fss.uu.nl or c.maas@fss.uu.nl

Prices

The conference fee is € 250 and the price for the course is € 100. However, people who pay before 1 February 2005 will only pay € 200 and € 75, respectively. For participants from Eastern Europe or Third World countries, a further € 50 reduction in fees is possible. Coffee and tea are included in these prices.

The Role of the Hausman Test and whether Higher Level Effects should be treated as Random or Fixed

Antony Fielding
University of Birmingham
a.fielding@bham.ac.uk

Introduction

For expository purposes, and using standard notation, consider a basic two level linear model of the form:

$$y_{ij} = X_{ij}\beta + u_j + e_{ij}$$

Initially the vector of coefficients β is regarded as fixed. The terms u_j represent the effect of the j'th level-two unit (panel units in many econometric applications) in shifting the intercept. The central question is whether these terms should be considered fixed or random? In the fixed case, they are handled using fixed coefficients of dummy variables for level-two units. Under classical regression assumptions, a computationally efficient variant of OLS is an appropriate estimation procedure. In the random case the u_j are treated as i.i.d. drawings from some population of effects, often assumed $N(0,\sigma_u^2)$. For this model, diverse estimation procedures are available. However, the main focus here is feasible generalised least squares and in particular the iterative form (IGLS). One of the crucial assumptions for consistent estimation of β is that u_j are uncorrelated with the covariates X of the model. It is as a test of this assumption that the Hausman procedure has been used. The validity of the assumption is of no consequence for the consistency of OLS in the fixed case, since the level-two effects are partialled out in this process.

Research contexts

Design based inference criteria in traditional experimental frameworks through analysis of variance, and also for complex multi-stage sample survey data are well established. Here fixed u_j are usual if the level-two units are treated as pre-determined groupings and restricted inferences for those units only are of interest. On the other hand, if generalisation is required beyond the units in the data to the broader population from which the units have been (randomly) drawn then it may seem appropriate to use random u_j.

Model based inference can lead to somewhat different perspectives, which have been the basis for developments of multilevel modelling methodology. It is not unusual for instance, to use random effects models where the set of level-two units are fixed by design or arise naturally from routine databases including where a full finite population of level-two units is present in a database. This is quite common, for example, in educational progress studies where a full set of schools or universities at a point in time is under

study. The desire to generalise and to acknowledge uncertainty not explicitly modelled, underpinned by an appeal to the superpopulation ideas of model based inference, seem compelling here.

It seems then that the criteria on which to base the central decision are not quite so clear-cut as an initial focus on design considerations might imply. Considerations seem to extend beyond these. However, there is a criterion which links design based and model based inference and seems to synthesise them: de Finetti's notion of *exchangeability*. This, according to Hausman (1978), is necessary and sufficient for the random effects specification to be justified. Briefly, the notion says that, given a set of sample units, we can consider exchanging any pair of the u_j without changing the subjective distributional characteristics of the model. Hausman refers to these two kinds of inference as logical as opposed to statistical questions, which focus on properties of various types of estimator leading naturally from whether to use fixed or random effects models. An appeal to exchangeability might resolve some issues, which focus on the model itself and the context of its application. However, there may be circumstances where a random effects specification seems appropriate on these logical grounds but estimation might proceed by conditioning on the particular sample of units and their u_j. This is equivalent to adopting a fixed effects model. It is these sorts of matters that lie behind the Hausman procedure, which will shortly be discussed.

There are other practical issues, which subtly touch on the decision though these are not often made explicit in the literature. In many econometric applications the focus is on models where the u_j are treated as nuisance factors, the X are all level-one variables, and the main interest is good estimation of the β. If this is the case then a decision may be made between the conditional estimation using a fixed effects specification and a random effects specification on other pragmatic grounds. Greene (2003) discusses many of these. Fixed effects for u_j might be avoided, for example, if there were a large number of them since each requires a separate parameter to estimate. This is implicit even if deviations from level two means were used, as for example in panel studies, since degrees of freedom on which standard error estimates are based are then effectively small. If there were few level-one observations per level-two unit this might lead to poor precision. On the other hand, with very few level-two units a fixed effects specification might be preferred since the inferences about level-two variation might be weak. An extremely important consideration here though is the restriction of a fixed effects model to situations where no level-two variables are present in X. If they were then they are confounded with fixed u_j and the parameters of the model are not identifiable. It is only a random effects specification that can handle level-two covariates and the extent to which level-two covariates can explain level-two variation. It is clear that fixed

effects specifications for u_j are unsuitable for many of the complex questions to which multilevel methodology has been addressed.

Hausman test

The Hausman (1978) test procedure is ubiquitous in econometric software and applications, particularly to panel data, to the extent that it is often interpreted as an automatic resolution of the specification decision we have been examining. It is, however, limited to situations where logic and context can leave the decision open. Thus, for example, models that entertain level-two covariates cannot fall within its remit.

Hausman suggests a very general test that can be applied to a wide variety of possible model mis-specifications. Its application to the current question is but one particularisation. Holly (1982) also notes that it is rather curious since there are no stated parametric restriction hypotheses underlying the test. The arguments are also conducted in asymptotic terms. The idea is to compare two possible estimators of β. $\hat{\beta}_0$ will be asymptotically consistent with $p\lim \hat{\beta}_0 = \beta$ and efficient under a certain set of specifying assumptions. This will be the case for a GLS in the fully specified random effects model discussed above providing the u_j are uncorrelated with all of the variables defining X_{ij}. There may also be another estimator $\hat{\beta}_1$ such that $p\lim \hat{\beta}_1 = \beta$ regardless of whether the above specifying restriction is valid or not. This is the case in the circumstances under consideration for OLS estimates $\hat{\beta}_1$ of a fixed effects model, which equivalently yield estimates of β conditioned on sample u_j. We might, however, expect $\hat{\beta}_1$ to be less efficient than $\hat{\beta}_0$. The technical details of the procedure and derivation of the test statistic are given fully in the original paper and in standard econometric texts. Broadly however, the argument centres on whether the sample evidence does or does not support $p\lim \hat{\beta}_0 = p\lim \hat{\beta}_1$, as it will if the specifying restriction is valid. With $q = \{\hat{\beta}_1 - \hat{\beta}_0\}$ and consistent estimators of the variances of $\hat{\beta}_1$ and $\hat{\beta}_0$, the Hausman test statistic $q'(\hat{V}ar\hat{\beta}_1 - \hat{V}ar\hat{\beta}_0)^{-1}q$ is shown to have asymptotically a null χ^2_k distribution, where k is the number of elements in β. If it leads to the conclusion that the u_j are not uncorrelated with covariates then a fixed effects specification and OLS estimation is used. It may be noted that such a decision does make a real difference in the practical estimation of β. Many contrasting applications in the econometrics literature, including an original one by Hausman, point to sharp differences in estimates from the two approaches even in situations where they both may be expected to be consistent.

It is clear that the Hausman test is simply a diagnostic of one particular assumption behind the estimation procedure usually associated with the random effects model. This may be fine in the restricted situations in which such a diagnostic is relevant. However, it is equally clear that it does not address the decision framework for a wider class of problems where the logical status of the effects may be relevant. Indeed this was recognised by Hausman (1978). That it has become an almost routine procedure purporting to resolve this decision has, however, led to many reservations. Detailed work on these broader reservations is, however, limited. The aim of the rest of this article is to summarise what has been done, and to highlight areas where further work would be fruitful. Ongoing work by the present writer and colleagues will be reported at a later date.

Limitations of the Hausman test

The reservations uncovered in the literature fall into three broad groups. Firstly, even if attention is restricted to the use of Hausman in its diagnostic sense, is it a good diagnostic for these purposes? Secondly, what is lost in terms of model building and development to answer important substantive questions if its use leads to the adoption of a fixed effects framework? Thirdly, even if it was an acceptable diagnostic of the inconsistency of GLS, why are available consistent procedures not used more frequently?

Skrondal and Rabe-Hesketh (2004) is one of the few general texts on modelling which address the first perspective. They point out that the test appears sensitive to a variety of mis-specifications other than the one under consideration. Thus a significant value for the test statistic may be misleading as a diagnostic for this mis-specification. Some limited simulations by the present writer indicate that dropping a key covariate in the model can have this effect but there is scope for deeper work. They also point out that the null distribution might not be well approximated by the asymptotic chi-square in finite samples. This can lead to an overstated size of test. Long and Trivedi (1993) have also commented on its poor power as a diagnostic in many typical multilevel structures. Some aspects of this have been confirmed more recently by Ejrnaes and Holm (2004). Concern has also been expressed at the possible narrowness of focus on consistency of estimators. There is some thought that, due to its efficiency, a random effects GLS estimator may have advantages even if it is slightly biased and inconsistent. Work is obviously required on all these questions but there is a certain sufficiency about the doubts they raise.

Turning to the second group of reservations, there are examples in the literature where the problem has been finessed by switching from a fixed effects to a random effects model midway through an application. Fielding (2004) discusses one such application. There is also a sense in which Hausman is used as if the model

under consideration is finished in that the covariates are pre-determined and all that remains is the question of how to treat the u_j. However, in model exploration we might suppose that Hausman on a base model has led to adoption of a fixed u_j framework. We know that fixed effect model estimates of coefficients have different precisions than for the random scenario. In a stepwise approach to model building with strict criteria governing the addition of covariates this might lead to a fixed effects model excluding some which might have been included in a random effects model. Differential effects of covariates between level-two units is one type of model development that it is frequently desired to explore. This can be accommodated in a random effects framework with the addition of one or more variance parameters. In a fixed effects framework degrees of freedom might be stretched if large numbers of fixed interaction effects between dummies and covariates are included. Implicitly in Hausman applications there is no corresponding test of the fixed effects model, which becomes a default option, which may in itself not be the correct specification. Model building strategies, which, at each stage of refinement, include an alternative decision making framework, might be more suitable. One possibility might be a DIC criterion (Spiegelhalter et al, 2002) where the inclusion of the effective number of parameters might favour random effects.

The last group of reservations really centres on why we should drop the flexibility of the random effects model if it is an appropriate specification. The diagnostic test may question the key specifying assumption necessary for desirable properties of its usual estimator. However, there are other estimators, which accommodate this. For instance, with correlation between u_j and specific level-one covariates, one could extend the random effects specification by including level-two means of these covariates. Snijders and Berkhof (2004) have demonstrated that GLS will then yield consistent estimators of β in the original model. This extension is equivalent to the auxiliary regressions and testing the additional coefficients suggested by Hausman as an easy way of implementing his test. There are however, possible objections: the addition of context effects changes the characterisation of level-two variation when for substantive reasons such adjustments might not be desired in the study of level-two effects themselves. However, the consistent conditioned iterative generalised least squares procedure (CIGLS) developed by Rice et al. (1998) exploits these ideas without departing in this way from the specification of the original model. The auxiliary regressions form an estimation step in an adaptation of an IGLS procedure, which can retain the original characterisation of level-two heterogeneity.

Consistent instrumental variable (IV) estimation of various forms have also been proposed. The main difficulty here has been finding suitable instruments and the recognition of low precision of IV in many circumstances.

Hausman and Taylor (1981), however, suggest that suitable instruments may be formed from transformations of the covariates themselves. Arellano and Bover (1995) discuss a general framework for IV estimators. Spencer and Fielding (2002) provide some evidence that IV estimators can be relatively precise if instruments are well chosen. Monte Carlo Markov Chain estimation also offers further possibilities. Congdon (2004) comments that it can accommodate u_j correlated with X quite flexibly.

In conclusion, the general lessons appear to be that the widely held belief that a significant Hausman test implies that a random effects framework be abandoned, is somewhat premature. This paper presents some of the issues as a stimulus to further debate and work on what is really quite a complex issue.

References

Arellano, M and Bover, O. (1995). Another look at instrumental variable estimation of error-component models. *Journal of Econometrics,* **68***,* 29-51.

Congdon, P. (2003). *Applied Bayesian Modelling*. Chichester: Wiley.

Ejrnaes, M., and Holm, A. (2004). Comparing fixed effects and covariance structure estimators. *Proceedings of RC33, 6th International Conference on Social Science Methodology, University of Amsterdam*.

Fielding, A. (2004). Invited comments on the papers by Draper and Gittoes and Bratti et al., *Journal of the Royal Statistical Society, Series A,* **167**, 498.

Greene, W. H. (2003). *Econometric Analysis (5th Ed.)*. Upper Saddle River: Prentice-Hall.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, **46**,1251-1271.

Hausman, J. A., and Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica*, **49**,1377-1398.

Holly, A. (1982). A remark on Hausman's specification test. *Econometrica*, **50**, 749-759.

Long, J. S. and Trivedi, P. K. (1993). Some specification tests for the linear regression model. In Bollen, K. A., and Long, J. S. (eds) *Testing Structural Equation Models,* Newbury Park, CA: Sage.

Rice, N., Jones, A., and Goldstein, H. (1998). *Multilevel models where random effects are correlated with fixed predictors*. York: University of York Centre for Health Economics.

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalised Latent Variable Modelling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton: Chapman Hall.

Snijders, T., and Berkhof, J. (2004). Diagnostic checks for multilevel models. In de Leeuw, J., and Kreft, I. (Eds), *Handbook of Quantitative*

Multilevel Analysis, Sage (forthcoming).

Spencer, N., and Fielding, A. (2002). A comparison of modelling strategies for value added analyses of educational data, *Computational Statistics*, **17**,103-116.

Spiegelhalter, D., Best, N.G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B,* **64**, 583-539.

Multiple Imputation using *MLwiN*

James R Carpenter
London School of Hygiene & Tropical Medicine
Harvey Goldstein
Institute of Education, University of London
james.carpenter@lshtm.ac.uk

Introduction

Missing data are ubiquitous in social science research. They introduce ambiguity into the analysis beyond the familiar sampling imprecision. However, in analysing partially observed datasets, the aim is still to produce a valid analysis. In this context, valid means that point estimates home in on the true values as the sample size increases, confidence intervals achieve their nominal coverage and inferences are correct. However, when data are missing, routine analyses are only valid if certain assumptions are made.

The key to understanding when analyses are going to be valid lies in the relationship between the process by which data become missing and the observed and unobserved data. As described by Rubin (1976), if the process by which data become missing is unrelated to any observed or unobserved data, missing data are said to be 'missing completely at random (MCAR)'. If the process depends on the data, but conditional on the observed data does not depend on the unobserved data, missing data are said to be 'missing at random (MAR)'. If the process depends on the unobserved data, even after accounting for the information in the observed data, missing data are said to be 'not missing at random (NMAR)'. The two important points to note are (i) on the basis of the observed data alone, it is not possible definitively to say which of these three processes cause the data to be missing, and (ii) if the data are MCAR or MAR then likelihood based methods (such as used in multilevel models) can proceed without explicitly modelling the dropout mechanism.

When data are missing, it is generally inappropriate to restrict analysis to units

with complete observations. This is only valid under the restrictive MCAR assumption, and even then can waste information. An intuitively attractive approach is to use the relationship amongst the observed variables to impute a plausible value for the missing variables. Although this can reduce bias, such a single imputation underestimates the variance of estimators. This led Rubin (1987) to propose multiple imputation, whereby a number of 'complete' datasets are formed by drawing the missing data from the estimated distribution of the missing given the observed data. Each of the resulting 'complete' datasets can then be analysed in the usual way. The attraction of the approach is that the results of these analyses can be combined using simple rules to obtain valid point estimates, variances and hence inferences. Furthermore, the rules for combining the imputations are the same in almost every setting.

Below, we briefly review the details of multiple imputation. We then describe *MLwiN* macros that implement multiple imputation, and give an illustration of their use, before drawing some conclusions.

Brief review of multiple imputation

Suppose that the data Y can be split into observed and missing parts denoted by $Y=(Y_O, Y_M)$. Note that different units will typically have different combinations of missing and observed variables. Let the quantity we are interested in calculating from the data be $Q=Q(Y_O, Y_M)$, for example a regression coefficient or a variance term. Ideally, we want to estimate the distribution of

$$f(Q \mid Y_O) = \int Q(Y_O, Y_M) f(Y_M \mid Y_O) dY_M.$$

Multiple imputation assumes this distribution is approximately Normal, so that it can be described by its mean and variance. These are estimated as follows. First we note that under MCAR or MAR, regression models, *in which only the responses are missing*, give valid parameter estimates. We therefore set up a (typically multivariate response) regression model in which all the partially observed variables are on the left-hand side. Fitting this then gives a valid estimate of the distribution of $Y_M \mid Y_O$. This model can be fitted by maximum likelihood or in a Bayesian framework, with uninformative priors. Having done this, we can then impute the missing data for partially observed units by drawing them from the distribution estimated by this model. This can be done a number of times, giving rise to a series of imputed datasets, each of which can be analysed as the original complete dataset was intended to be.

Note that having only the partially observed variables on the left-hand side is not necessary but it makes computation quicker. Under multivariate normality, the joint distribution of the missing and observed we estimate will be the same if we have all the variables on the left-hand side and only the constant on the right-hand side. Thus, this method should work if a substantial majority of the variables have some missing data.

Suppose the analysis of each of K imputed datasets gives rise to estimates $Q_1, Q_2, \ldots Q_K$, with standard errors σ_1, $\sigma_2, \ldots \sigma_K$. Then the mean of the distribution of Q, denoted Q_{MI} is approximated by the average of the estimates from the imputed datasets:

$$Q_{MI} = E_{Y_M|Y_O}\{Q(Y_M, Y_O)\} \approx \frac{1}{K}\sum_{k=1}^{K} Q_k.$$

The formula for the variance of the distribution of Q is more complicated; to overcome the drawbacks of single imputation it takes into account both between and within imputation components of variance. Using conditional expectations, we can estimate this in terms of the Q_k's and σ_k's:

$$\begin{aligned} V(Q_{MI}) &= E_{Y_M|Y_O} V\{Q(Y_O, Y_M)\} + \\ &V_{Y_M|Y_O} E\{Q(Y_O, Y_M)\} \\ &\approx \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2 + \left(1+\frac{1}{K}\right)\left(\frac{1}{K-1}\right)\sum_{k=1}^{K}(Q_k - Q_{MI})^2, \\ &= \sigma_W^2 + \sigma_B^2, \text{say,} \end{aligned}$$

where the term (1 + 1/K) in the second expression in the second line compensates for the finite number of imputations.

Then, valid tests of hypotheses $Q=Q_0$ are obtained by referring

$$\left(\frac{Q_{MI} - Q_0}{\sqrt{V(Q_{MI})}}\right)$$

to a t distribution with v degrees of freedom, where

$$v = (K-1)\left[1 + \frac{\sigma_W^2}{(1+1/K)\sigma_B^2}\right]^2.$$

For multiple imputation to be valid, a number of conditions must hold. First, the missing data must be MAR. While multiple imputation can be applied if data are NMAR (where a joint model of the missingness mechanism and the data are required), this is not entirely straightforward, although it is possible to implement in *MLwiN*. Second, the imputation model must include all the structure that we wish to investigate in the model of interest (i.e. in the model we intended to fit to the full data). So, for example, covariates and interactions that are to be investigated in the full dataset must be included in the imputation model, otherwise the imputed observations will not have this structure. Thirdly, and most importantly in this context, the imputation models must have the right variance structure. If the data are multilevel, the imputation model must be too.

Macros for multiple imputation in *MLwiN*

We have already commented that if a dataset is multilevel, then the imputation model should be multilevel too. Thus *MLwiN* (Rasbash et al., 2004) is a natural tool. This is especially so as *MLwiN* can fit a range of Bayesian models using Markov Chain Monte Carlo, and draw samples from the posterior for missing response variables (Browne, 2004, chapter 17).

We have written a macro that uses these MCMC routines in *MLwiN* and enables multilevel multiple imputation to be performed semi-automatically, as part of the analysis process. A user can set up and fit a model in the *equations* window, and then invoke the macro from the command interface window. The macro does the following:

1. Records the model of interest
2. Sets up a multilevel multivariate imputation model with the partially observed variables as responses, and fits this model in a Bayesian framework with uninformative priors using Markov Chain Monte Carlo methods
3. Imputes a number of completed datasets
4. Fits the model of interest to each of these datasets
5. Combines the results, as set out in the previous section
6. Presents the results in the equations window.

The user has the option to specify the number of imputations, the number of updates of the sampler between imputations and whether the imputations are single or multilevel. In addition, the user can suspend the macros before the imputation model is fitted, and add variables so that the imputation model is more general than the model of interest. The multilevel imputation model has an unstructured covariance matrix at each level, to capture the multilevel structure of the data and draw appropriate imputations.

For a more detailed introduction to multiple imputation, see the 'getting started' section of our website (www.missingdata.org.uk) or Schafer (1997, 1999).

Example

Consider the following data, extracted from Blatchford et al. (2002). Note this dataset is not representative of the project dataset. We have data on 4873 pupils from 172 schools, and are interested in the relationship between literacy score at the end of reception year (variable `nlitpost`) on literacy score at baseline (variable `nlitbase`), eligibility for free school meals (variable `fsmn`: 1=eligible), term of entry (variable `tentry`: 1=Spring or Summer, 0=Autumn) and gender (variable `gend`: 1=male). Both baseline and end of reception literacy score variables are standardised versions of the test results which are approximately Normal.

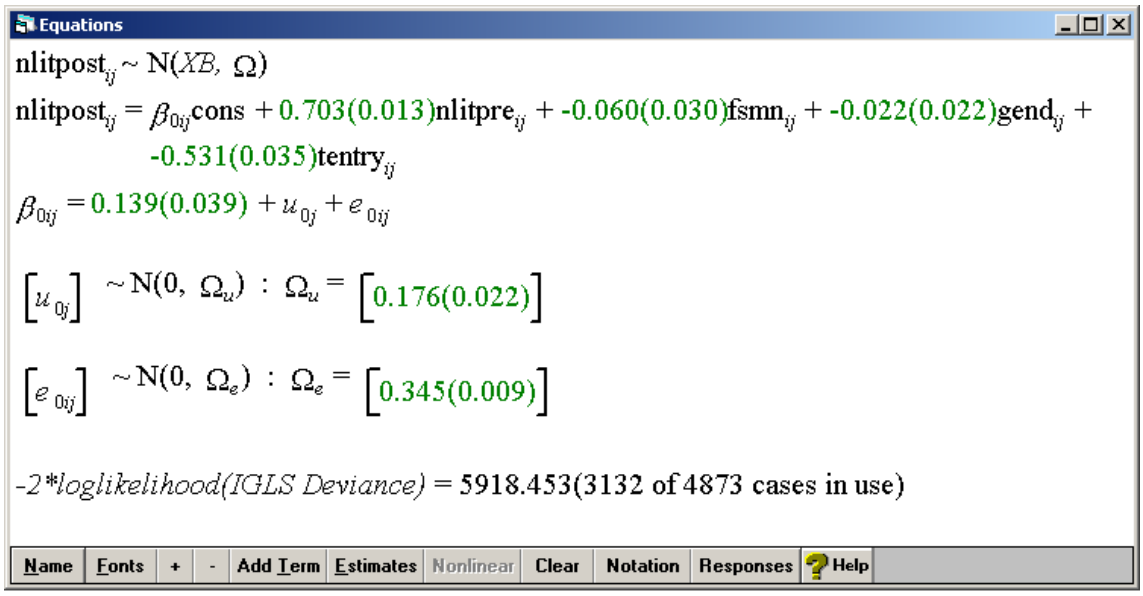
Clearly, the dataset is multilevel, with pupils nested within schools. For brevity, we focus on the coefficient for gender. The results from the full data are shown in the first column of Table 1. After adjustment for the other variables, gender is significant, with boys doing worse than girls. Note that ignoring the school level (which is highly significant) in the modelling leads to a marked inflation in the gender effect (second line of Table 1).

We now make some of the pre-reception literacy observations missing, according to the following rule:

$$\Pr(\text{Observe } nlitpre) = 1/(1+\exp\{0.5\,nlitpost - 1[\text{pupil is male}] - 1[\text{pupil eligible for free school meals}]\})$$

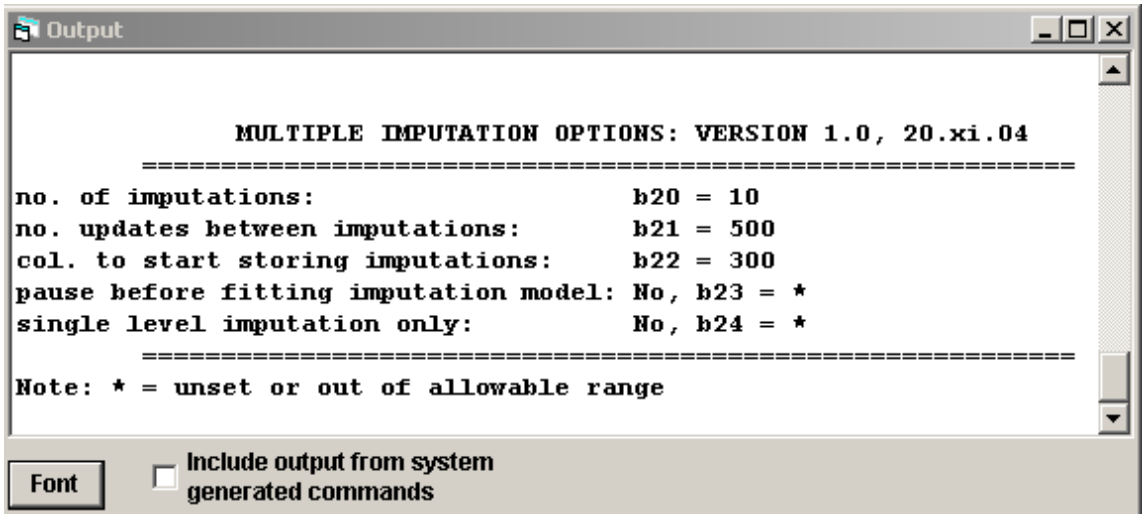
In other words, for each pupil, we calculate this quantity, say *p*. We then generate a uniform random variable, *u*, on [0,1]. If *u*>*p* we set the pupil's *nlitpre* score to missing.

The resulting dataset has only 3132 complete observations (65% of the total) from 171 schools. Fitting the model to these data gives the *equations* window shown below. The coefficient for gender is reproduced in the third row of Table 1; it is biased down by ~58%, and no longer significant.

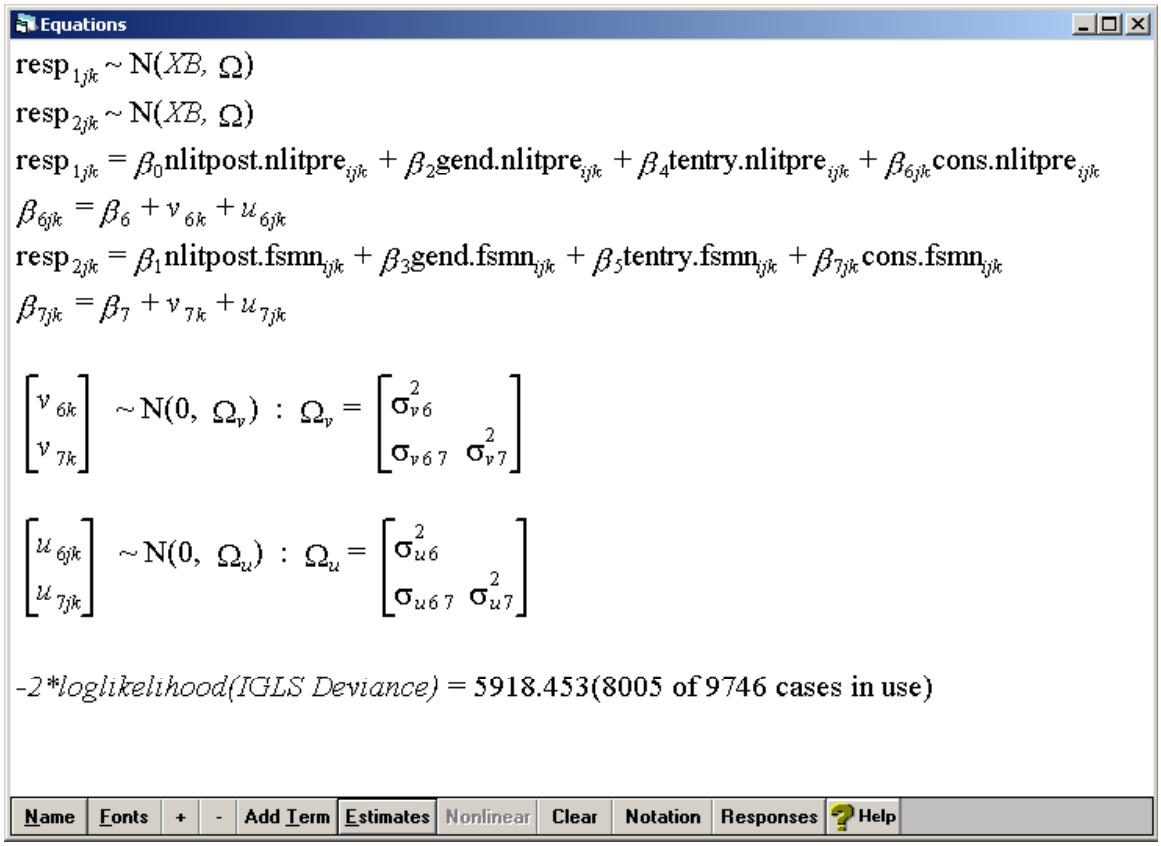


We now use the multiple imputation macro to impute the missing data under MAR. In the command interface, the imputation options can be viewed by typing *obey mi_options.*

We will carry out 10 imputations, as this is usually enough to make the fraction of the variance of the imputation parameter estimates due to simulation acceptably small (see Schafer, 1999). Further, we update the sampler 500 times between each imputation. As we have mean centred all our variables, this should be more than sufficient to make the imputations independent. We will carry out multilevel imputations. Setting these options gives:

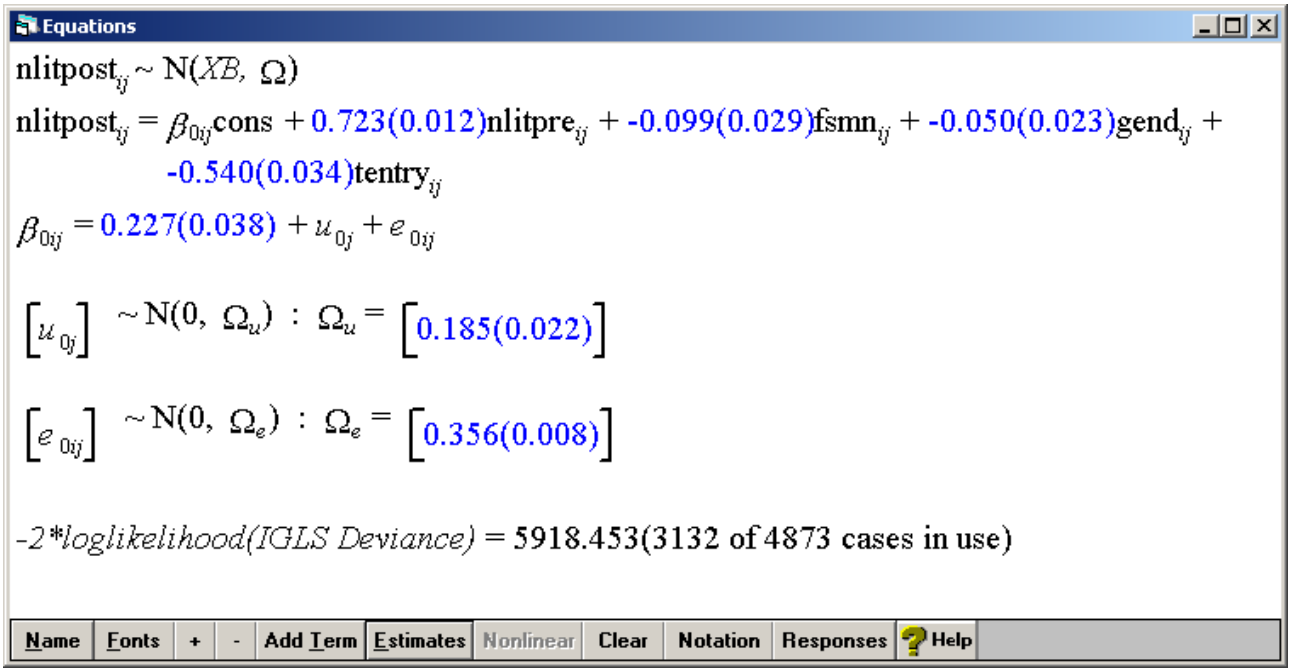


The multiple imputation macro is then started by typing *obey mi* in the command interface. The macro takes 35 seconds on a 2.5Ghz PC. The imputation model, which is set up automatically by the macro, is as follows:



Although only one variable has missing observations, the model is multivariate as *MLwiN* needs a fully observed variable on the left-hand side; the program chooses this to be free school meal eligibility. For both responses, the model has components of variance for school and pupil, together with covariances between them.

The imputation model can be displayed by selecting the appropriate option before starting the imputation process. Otherwise, it is only displayed fleetingly, and the macro proceeds automatically, displaying the results in the *equations* window, which is reproduced below. The coefficient for gender is also shown in Table 1.



The between and within imputation variances are stacked in columns named *fp_withvar* and *fp_betvar.* From these we can calculate the degrees of freedom for the t-distribution for testing the hypothesis about gender; this comes to 64. The two-sided p-value for gender is thus 0.03. Using the quantiles from the t-distribution we can calculate the 95% confidence interval shown in Table 1. Multilevel multiple imputation has moved the estimate for gender very close to the estimate for the fully observed data, but reduced the significance as a consequence of the lost information.

Finally, for comparison, we present the results for single level imputation. These are obtained by setting *b24=1* and re-running the macro. The results are shown in the bottom row of Table 1. We see that the coefficient for gender is now much smaller; further the p-value is 0.005, much more significant. Note that the coefficient for gender is now much smaller than for the multilevel analysis of the fully observed data, but much closer to that for the single level

analysis for the fully observed data. The same goes for the confidence interval for gender. We conclude that when data are multilevel, using a multilevel imputation model is important to avoid misleading conclusions.

Table 1. Estimates of the (adjusted) effect of gender on literacy at the end of reception year, for multilevel and single level analyses of the full, partially observed and imputed datasets.

Model	Estimated effect of gender	Standard error	90% confidence interval
Full data (n=4873)	-0.053	0.017	(-0.086, -0.020)
Full data (n=4873), single level model	-0.086	-0.021	(-0.127, -0.045)
'Observed' data (n=3132)	-0.022	0.022	(-0.065, 0.021)
Multiple imputation	-0.050	0.023	(-0.096, -0.004)
Single level multiple imputation	-0.073	0.025	(-0.123, -0.022)

Finally, we note that this analysis has also been carried out in WinBUGS, which gave very similar answers, but required a lot more programming and data manipulation.

Discussion

The *mi* macro provides users of *MLwiN* with a general framework for analysing datasets with missing observations under the assumption of missing at random. For models where all the variables needed for imputation are included in the model of interest, the macro can be used semi-automatically, as described above. While users do not require a deep knowledge of multiple imputation to do this, this is obviously no substitute for carefully looking at the data before modelling. In particular, the user will often want to include covariates, or interactions, in the imputation model, which are not in the model of interest. The macros allow this to be done; there is an option to suspend the macro once the model of interest has been set up. The imputation model can then be generalised by the addition of extra terms. Additional variables which themselves have missing values should be added to the response (left-hand side). Additional fully observed variables should be added to the explanatory variables (right-hand side). The macro can then be resumed to automatically draw the imputations and analyse the imputed datasets. Full details are given in the instructions available at www.missingdata.org.uk

The example shows the importance of using a multilevel imputation model when the data are multilevel. Of the other software currently available for multiple imputation, SAS PROC MI, and Schafer's NORM routine (See

multiplep-imputation.com) perform single level imputation only; they are equivalent to specifying a single level imputation in our macro. Schafer's PAN package is a stand alone Windows program suitable for repeated measures data, that uses a similar hierarchical structure to the *mi* macro; it is not however clear from the website whether it copes with more than two levels. By contrast, the macro described here will handle up to four levels, and both imputation and analysis are combined in one package. Further, the *mi* macro works by calling sub-macros, which can be exploited by experienced users for analyses where automatic use of the macro is inappropriate. Automatic use of the macro will be inappropriate when, for example, the model of interest has linear and quadratic terms in, say, X, and X has missing observations. This is because the macro would put X and X^2 on the left hand side of the imputation model, which would then fail to fit. What is needed is to impute X, then calculate X^2 for each imputed dataset and fit the model of interest to this.

Note that currently the *mi* macro only works for Normally distributed data. Schafer (1997, Ch. 6) recommends transforming non-Normal quantitative variables to approximate Normality before imputation, and imputing binary or ordinal data under the Normal model before rounding off to the nearest category. His simulations suggest that under certain circumstances multiple imputation is fairly robust to model misspecification. Handling non-Normal data in this way is an option in the current framework but the macro components cannot be used automatically in this case. Currently, we are developing more appropriate methods for handling discrete data.

As discussed in the introduction, the possibility that the data are not missing at random (NMAR) can never be ruled out. Thus, it is often appropriate to look at the sensitivity of the analysis to NMAR. We are currently developing an additional macro, which would use a weighting approach (Carpenter and Kenward, 2005) to test the robustness of conclusions to certain forms of NMAR.

The macros described will be available from our website (missingdata.org.uk) from the beginning of January. They require *MLwiN* release 2.0, and two additional patches, which can also be downloaded from the website.

Acknowledgements

We are grateful to Peter Blatchford and colleagues at the Institute of Education for permission to use their data. James Carpenter is supported by ESRC Research Methods Programme grant H333250047, titled 'Missing data in multi-level models'.

References

Blatchford, P., Goldstein, H., Martin, C. and Browne, W. (2002). A study of class size effects in English school reception year classes. *British Educational Research Journal*, **28**, 169-185.

Browne, W. (2004). *MCMC estimation in MLwiN (version 2.0)*. London:

Institute of Education. Available from http://multilevel.ioe.ac.uk

Carpenter, J., and Kenward, M. (2005). Contribution to discussion of Greenland, S., 'Multiple-bias modelling for the analysis of observational data'. *Journal of the Royal Statistical Society, Series A, to appear*.

Rasbash, J., Steele, F., Browne, W. and Prosser, B. (2004). *A User's guide to MLwiN (version 2.0).* London: Institute of Education. Available from http://multilevel.ioe.ac.uk

Rubin, D. B., (1976). Inference and missing data. *Biometrika*, **63**, 581-592.

Rubin, D. B., (1987). *Multiple Imputation for Non-response in Surveys*. New York: Wiley.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Schafer, J. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research,* **8,** 3-15.

A Review of Multilevel Software Packages

Harvey Goldstein

Institute of Education, University of London

In 2002 the Centre for Multilevel Modelling decided to undertake a review of existing statistical packages that had facilities for multilevel modelling. The review was co-ordinated by Min Yang and it commissioned a number of people with extensive experience of multilevel modelling to undertake these reviews and the Centre is extremely grateful to those who gave their time to do this. The aim was to place each review on the Centre's website together with a set of comparative summary tables. A number of datasets were chosen to cover the main types of user models and each model was run with each package to evaluate numerical accuracy and timings.

The full results of the review are available from:
http://multilevel.ioe.ac.uk/softrev/index.html
although there are still a few gaps waiting to be filled.

The datasets covered the following areas:

- Two and three level Normal response models
- Random coefficient models and heterogeneous variance models
- Repeated measures models
- Two level binary response models
- Two level ordered category response models
- Cross classified models

Not all the packages could handle all the dataset types and where additional data structures could be modelled, e.g.

structural equation models, these are mentioned. In addition, authors of reviews were invited to respond if they wished with their responses placed alongside the reviews. In the event, none has to date taken up this offer.

In terms of timings, because the reviews were carried out on different machines we have had to apply conversion factors so that there will be some uncertainty over these. Nevertheless, while most packages performed similarly for most datasets there were certain notable differences. All of the packages, except WINBUGS, use a likelihood based algorithm so that timings for these are comparable. For a simple three level variance components model (on a 433 MhZ Pentium II PC) AML took over 10 hours compared to just over an hour for WINBUGS, a few minutes for SAS and SPSS and a few seconds for the remainder! For the binary response data, different procedures were used: those using PQL methods took between one second (*MLwiN* and GENSTAT) and one minute (SAS). Those using maximum likelihood varied between about three seconds (EGRET) and a minute (MIXOR), and WINBUGS took just under 45 minutes. For the cross-classified variance components model, most packages took less than a minute with about 40 minutes for WINBUGS. For the ordered category model, MIXOR was the quickest of the packages using maximum likelihood (under 10 seconds) and SAS the slowest (just under one minute). There was good numerical agreement among packages.

In terms of ease of use, to some extent this is a subjective perception and potential users should read the detailed reviews which take the reader through the analyses of the datasets. User support varies among packages. For example, most packages have adequate manuals and either email support or an active user group. Flexibility in terms of the number of different models that can be fitted varies between *MLwiN*, WINBUGS, GENSTAT, STATA, aML and SAS that can fit a very wide range, to EGRET that is effectively restricted to handling discrete response data only, and SPSS that can handle a limited range of Normal response models.

The web site is intended to respond to new software and new versions of existing software and will be updated periodically. Software writers and anyone who would like to review any package are welcome and should contact Amy Burch (a.burch@ioe.ac.uk).

Review of 'Multilevel Modeling. Methodological Advances, Issues and Applications'. Reise, S. P. and Duan, N. (Eds.) (2003). Mahwah, NJ: Lawrence Erlbaum Associates Inc. ISBN: 0-8058-5170-4 $37.50, pp. 314.

Ian Plewis
Institute of Education, University of London

As the title implies, this is a varied collection of 13 chapters, ranging from demanding theoretical chapters on modelling to a number of interesting applications, and with just one introductory chapter (Ch. 13 by the editors) on design. The book is nicely produced; its strength is that anyone interested in multilevel modelling is likely to find at least one chapter of interest but this strength might also be a weakness when it comes to thinking about buying the book.

At one end of the methodological spectrum, there are introductions to the two level model by Bachmann and Hornung (Ch. 8) and Rowe (Ch. 12) and to repeated measures data by Baumler et al. (Ch. 7). At the other end, there is a chapter on mean and covariance structures (Bentler and Liang, Ch. 3) which is hard going and includes an unhelpful example, and one by Cudeck and Du Toit (Ch. 1) on nonlinear models for repeated measures with a focus on estimation issues.

The remaining chapters cover a range of applications but, perhaps surprisingly, only one (Fielding, Ch. 9) deals with categorical response variables. Fielding, in a thorough and wide-ranging chapter, presents illustrations of both hierarchical and cross-classified models for ordered categorical outcomes. Two chapters deal with sensitivity issues: Seltzer and Choi (Ch. 2) use different t-distributions to represent level-one and level-two variance in a fully Bayesian analysis of growth curve data with very few units at level two whereas Ecob and Der develop an iterative procedure for dealing with level-one outliers with similar kinds of data. Growth curve modelling is, in fact, a recurrent theme, further illustrated by two chapters (4 and 6) from the Muthén stable, one on relating two growth curves, contrasting multilevel modeling and one form of trajectory analysis, and the other using the models to estimate so-called complier effects in intervention studies.

The collection is completed by a review of multilevel meta-analysis by Hox and de Leeuw (Ch. 5) and a chapter by Hutchison (Ch. 10) that tackles the important but often-ignored topic of the effects on estimates of measurement errors in explanatory variables and extends it, using bootstrap methods, to the multilevel context.

Some Recent Publications using Multilevel Models

Bressoux, P., and Bianco, M. (2004). Long-term teacher effects on pupils' learning gains. *Oxford Review of Education*, **30** (3): 327-345.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika*, **69** (2): 167-190.

Skrondal, A., and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/ CRC Press.

Zimprich, D., Hofer, S. M., and Aartsen, M. J. (2004). Short-term versus long-term longitudinal changes in processing speed. *Gerontology*, **50** (1): 17-21.

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

MLwiN Clinics in London

Wednesday 2 March 2005
Wednesday 6 April 2005

at

Centre for Multilevel Modelling
11 Woburn Square, London WC1H 0NS

Contact *MLwiN* Technical Support for appointments
Tel: +44 (0) 20 7612 6688
mlwin.support@ioe.ac.uk

Future clinic dates will be announced at:
http://multilevel.ioe.ac.uk/support/clinics.html