

REALCOM: methodology for realistically complex multilevel modelling

Training materials manual

by

Harvey Goldstein
Fiona Steele
Jon Rasbash
Christopher Charlton

Centre for Multilevel Modelling
Graduate School of Education
University of Bristol
BS8 1JA.
United Kingdom

<http://www.cmm.bristol.ac.uk>

June 2008

Acknowledgements

This work was supported by the Economic and Social Research Council (ESRC) grant number RES-000-23-0140. We are also grateful to the following for their advice and comments: Anthony Robinson, William Browne, James Carpenter.

Contents

INTRODUCTION	1
1. <i>A brief introduction to MCMC estimation.....</i>	<i>1</i>
2. <i>Using the MATLAB routines.....</i>	<i>2</i>
CHAPTER 1. MODELLING MEASUREMENT ERRORS IN MULTILEVEL MODELS.....	3
1. <i>Introduction.....</i>	<i>3</i>
2. <i>Defining measurement errors.....</i>	<i>3</i>
3. <i>The effect of adjusting for measurement errors.....</i>	<i>5</i>
4. <i>An example data set.....</i>	<i>6</i>
5. <i>Adjusting for measurement errors.....</i>	<i>8</i>
6. <i>Adjusting for missclassification probabilities.....</i>	<i>12</i>
7. <i>Discussion.....</i>	<i>14</i>
APPENDIX A1: MODELLING MEASUREMENT ERROR – MCMC ESTIMATION,.....	15
1. <i>Introduction.....</i>	<i>15</i>
2. <i>Correlated measurement errors.....</i>	<i>15</i>
3. <i>Binary and ordered category explanatory variables.....</i>	<i>16</i>
4. <i>Missclassification probabilities.....</i>	<i>17</i>
5. <i>Extensions.....</i>	<i>18</i>
APPENDIX B1: AGGREGATING LEVEL 1 VARIABLES WITH MEASUREMENT ERRORS	20
1. <i>Introduction.....</i>	<i>20</i>
2. <i>Sampling level 1 units.....</i>	<i>20</i>
3. <i>Further considerations.....</i>	<i>21</i>
CHAPTER 2. MODELLING MULTILEVEL LATENT VARIABLE STRUCTURES.....	23
1. <i>Introduction.....</i>	<i>23</i>
2. <i>One level factor models.....</i>	<i>23</i>
3. <i>Adding further explanatory variables and structural models.....</i>	<i>25</i>
4. <i>Two level factor and structural equation models.....</i>	<i>26</i>
5. <i>Models for binary and ordered responses.....</i>	<i>26</i>
6. <i>An educational example.....</i>	<i>27</i>
7. <i>Fitting single level factor models.....</i>	<i>28</i>
8. <i>Tables.....</i>	<i>32</i>
CHAPTER 3. MULTILEVEL MULTIVARIATE MODELS WITH MIXED RESPONSE TYPES AT 2 LEVELS.....	34
1. <i>Introduction.....</i>	<i>34</i>
2. <i>Models for mixed multivariate responses at 2 levels.....</i>	<i>34</i>
3. <i>Growth data example.....</i>	<i>35</i>
4. <i>Class size data example.....</i>	<i>40</i>
5. <i>Multiple imputation.....</i>	<i>40</i>
6. <i>Class size data example.....</i>	<i>41</i>
7. <i>Conclusions.....</i>	<i>45</i>
APPENDIX A3. AN MCMC ALGORITHM FOR ESTIMATING MULTIVARIATE MIXED RESPONSE TYPES AT 2 LEVELS	46
1. <i>The model.....</i>	<i>46</i>
2. <i>Multicategory (unordered) responses:.....</i>	<i>46</i>
3. <i>Ordered responses.....</i>	<i>47</i>
4. <i>Sampling the fixed coefficients.....</i>	<i>49</i>
5. <i>Sampling the random effects.....</i>	<i>49</i>
6. <i>Sampling the level 1 (multivariate) covariance matrix.....</i>	<i>49</i>
7. <i>Sampling the level 2 covariance matrix.....</i>	<i>50</i>

8.	<i>Responses at both level 1 and level 2</i>	50
9.	<i>Imputing categories</i>	51
REFERENCES		52

Introduction

REALCOM is an ESRC supported research project at the University of Bristol, Centre for Multilevel Modelling. The team consists of Professor Harvey Goldstein, Professor Jon Rasbash, Dr Fiona Steele and Dr Christopher Charlton. Dr Edmond Ng was a research officer on the project until April 2006. The project's aims are to develop methodology and associated training materials in the following areas of multilevel modelling: structural equation models, measurement errors and multivariate responses at more than one level of the data hierarchy. A description of the research grant application can be found at <http://www.cmm.bristol.ac.uk>.

The methodology builds upon that already implemented in MLwiN version 2.02 which is described in the MLwiN manuals (<http://www.cmm.bristol.ac.uk>). The software is written in MATLAB (<http://www.mathworks.co.uk>.) and is available as a set of free-standing programs. They have their own graphical user interfaces for setting up models and displaying results. Development work on these programs is continuing and updates will be provided from time to time on the Centre website.

This volume contains a set of training materials that can be used as introductions to the methodology and as guides to using the software. Each of the chapters also has an appendix describing the estimation algorithm.

The project team are continuing to develop the methodology and training materials and feedback to the authors is very welcome.

1. A brief introduction to MCMC estimation

For a detailed introduction you should look at chapter 2 of the MCMC MLwiN manual, downloadable from the MLwiN web site (<http://www.cmm.bristol.ac.uk/>) . This manual also has introductions to structural equation models and measurement errors.

Briefly, MCMC estimation is a Bayesian estimation method that generates random draws from the 'posterior' distribution of the model parameters (including higher order residuals), The term MCMC stands for Markov Chain Monte Carlo, the Markov Chain consists of a set of 'iterations' of the algorithm: each iteration is a set of random (Monte Carlo) parameter draws where each parameter is drawn from its conditional (on the other parameters and the data) distribution. The idea is to produce a large (e.g. 5000) set of these that are then used for inference – e.g. the mean for a parameter is the simple average of the set, the standard deviation of the set is equivalent to the 'standard error' in the classical sense and quantiles are read off from the empirical (or smoothed) distribution. The chain needs to be 'stationary' and so the burn in period is used to get to this position and then the burn in draws are discarded. To speed up the analyses it is suggested you start with a small burn-in – say 100 – and say 500 iterations. Inspection of these chains provides certain diagnostics for our models and they will be discussed when we study the individual data sets. To compare one model with another we use a statistic called the Deviance Information Criterion (DIC) which is essentially a measure of model complexity and allows us to carry out an overall comparison between models. For more discussion see Chapter 3 of the MLwiN MCMC manual.

2. Using the MATLAB routines

For each of the models described in Chapters 1-3, there is an executable program file together with example datasets and files containing starting values and other parameters required. It is recommended that the data should be placed in a subdirectory of the directory containing the program files. In each chapter examples are given of how to set up the models via the graphical user interface. These interfaces will generate the algebraic representation of the model as the model components are defined. If you wish to copy any of the windows as displayed you will need to use a screen capture utility or the 'print screen' key. The Matlab command window will appear and record commands, and you may wish to minimise this. Results in tabular form and MCMC chain plots can be produced and are also displayed in separate windows. In some cases output files, for example of imputed values, can be requested and these are stored as tab delimited files suitable for import to other programs.

The executable files are as follows, together with the data sets used in the examples:

measurement-error.exe	classsize
mixed-responses.exe	growthdata.txt classsize_impute
structural-equation.exe	pisadata

These should be copied to separate directories on your computer.

Chapter 1. Modelling measurement errors in multilevel models.

1. Introduction

The implementation of measurement error modelling in MLwiN 2.02 will handle the case of measurement errors in Normally distributed variables where it is assumed that there is no correlation between the measurement errors for different variables. The REALCOM software extends the model to allow correlations between measurement errors and also has the capability of modelling misclassification probabilities in binary variables.

In many of the variables used in the social and medical sciences measurement errors are found. These can arise from unreliable measuring instruments, problems with variable definitions or simply reflect temporal fluctuations, for example within individual units. The errors we are concerned with are essentially considered as random and distinct from systematic errors which can lead to biases.

There is a large statistical literature on the modelling of such errors, mostly dealing with the case of continuously distributed variables in single level linear and non-linear models. Fuller (2006) provides a comprehensive treatment. In our research we have developed existing work based upon MCMC estimation for multilevel models (Browne et al., 2001) and incorporated in the MLwiN software (Browne, 2004). We deal with the 2-level case in detail with extensions to three levels being relatively straightforward. Extensions to handle cross classified and multiple membership models (Goldstein, 2003, Chapters 11 &12) also involve just the addition of appropriate sampling steps within the MCMC algorithm. The consequences of ignoring measurement errors are well known and typically lead to underestimation of coefficients and biased standard errors. In multilevel models we will additionally obtain biased estimates of covariance matrices.

The multilevel model of interest is assumed to be the Normal 2-level model including random coefficients, given by

$$y_{ij} = X_{ij}\beta + Z_j u + e_{ij} \quad (1)$$
$$u \sim MVN(0, \Omega_u), \quad e \sim N(0, \sigma_e^2)$$

Where $X_{ij}\beta$ is the fixed part of the model involving regression coefficients β and $Z_j U$ describes the random effects contribution at level 2, with a simple level 1 residual term e_{ij} . Details of the estimation of the parameters of this model, using maximum likelihood or Bayesian MCMC procedures can be found, for example, in Goldstein (2003, Chapter 2). We next introduce some basic definitions and assumptions. For completeness we review the salient features of measurement error models for the single level case.

2. Defining measurement errors

We consider measurement errors of two types. The first occurs with continuously distributed variables where the observed value can be written in the form, omitting subscripts, as

$$x^0 = x + m \quad (2)$$

where x^0 is the observed value, x the true value and m the measurement error. We shall assume that the model of interest, that is (1), is that which uses the true variable values rather than those observed with error. In some cases we may wish to use the variables as observed with error, for example if we were interested solely in prediction based on these. More generally, however, for purposes of explanation we would like to model the true values. The procedures described here will allow us to do this, after making certain assumptions.

Thus, to enable us to identify model parameters we must make the following assumptions (or equivalent ones). First, the true values and the measurement errors are assumed to be uncorrelated, and the mean value of m is zero. Secondly, we need to specify a distribution for m , typically Normal so that we have

$$m \sim N(0, \sigma_m^2)$$

Finally we need to consider properties for x . Suppose that we were able to obtain independent replications of x^0 , say x_1^0, \dots, x_k^0 . We could then write a simple model

$$x_i^0 = x + m_i, \quad i = 1, \dots, k \quad (3)$$

This will provide estimates for x , σ_m^2 . In a more complex model involving x the existence of replications will likewise generally allow us implicitly to incorporate the estimation of x , σ_m^2 into the model. In many applications replication can in fact be considered as a lower level of a data hierarchy and thus handled by standard multilevel modelling procedures.

In most practical social science applications, however, we do not have the possibility of independent replications. For example, in administering an educational test, a residual memory effect will preclude the possibility of independent replications. Hence the following exposition does not assume the existence of such replications. Instead we assume:

1. An independent value of σ_m^2 is available, recognising that this is typically a sample estimate, so that we may wish to incorporate uncertainty about σ_m^2 into our analysis, either by supplying a prior distribution (not considered here) or, generally more usefully, by carrying out a sensitivity analysis over the likely range of values for σ_m^2 .
2. A distribution for x . This is required because we cannot condition on x in (1) (as we can do in the replicated situation) and we only directly observe the distribution for x^0 . We shall assume that $x \sim N(\mu_x, \sigma_x^2)$ where μ_x , the mean true value, is typically estimated by the observed mean \bar{x}^0 . We can extend (3) to the multivariate case in a straightforward way by replacing the variance by a covariance matrix.

In the 'classical' measurement error model we define the reliability of x^0 by

$$R = R(x^0) = \sigma_x^2 / \sigma_{x^0}^2, \quad \sigma_{x^0}^2 = \sigma_x^2 + \sigma_m^2 \quad (4)$$

and we shall make use of this term. Thus, given a sample of values $\{x_i^0\}$ we can estimate $\sigma_{x^0}^2$, and hence σ_x^2 since σ_m^2 is assumed known. This step effectively becomes incorporated into the MCMC algorithm described in Appendix A.

There is, of course, the problem of obtaining a suitable estimate of σ_m^2 and also recognising that this may vary across subgroups of the population. We shall not get involved in any debate about this, but see Ecob and Goldstein (1983).

The second type of error is a *misclassification error* where the observed category of a discrete response variable is not necessarily the true category.

Suppose we have a binary (0,1) variable, for example whether or not a school pupil is eligible for free school meals (yes=1). We assume that the allocation to a category is not perfect and we denote the probability of observing a zero (no eligibility), given that the true value is zero, by $P_{obs}(0|0)$ and the probability of observing a one given that the true value is zero by $P_{obs}(1|0)$. Similarly we have $P_{obs}(0|1)$ and $P_{obs}(1|1)$. In Appendix A we show that knowledge of these misclassification probabilities allows us to compute the true probabilities of a zero and a one and how these are used in the estimation.

We shall only consider, for simplicity, misclassification error for a binary variable; the extension to multicategory variables raises no fundamentally new issues. In all cases we assume that our interest is in measuring the relationship with the true rather than observed explanatory variables.

3. The effect of adjusting for measurement errors

Consider the simple single level linear model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

with measurement error in the single explanatory 'true' variable, x . As above we have the adjusted variances and covariances for the 'true' model

$$\sigma_x^2 = \text{var}(x) = R \text{var}(x^o), \quad \text{cov}(xy) = \text{cov}(x^o y) = c_{xy}$$

Thus the estimate of the regression coefficient is given by

$$\frac{c_{xy}}{R \text{var}(x^o)} = b_{obs} / R$$

where b_{obs} is the coefficient for the regression based on the observed values and since the reliability is always less than or equal to 1.0, the regression coefficient is greater in absolute value. The estimate of the residual variance is given by

$$\text{var}(y) - \frac{c_{xy}^2}{R \text{var}(x^o)}$$

compared with

$$\text{var}(y) - \frac{c_{xy}^2}{\text{var}(x^o)}$$

for the regression using the observed values, and hence smaller than the latter.

Before we go on to an analysis of a data set we will note some restrictions that our models impose.

Consider the case of two explanatory variables with measurement error, and suppose for simplicity that they have the same observed variance equal to 1 and the same reliability, R . Let us also suppose that their measurement errors have a correlation of ρ_m and that the correlation between the observed variables is ρ_o .

For example, we require that the correlation between the true values lies between -1 and 1 and this implies

$$\frac{\rho_o + R}{1 - R} > \rho_m > \frac{\rho_o - R}{1 - R} \quad (5)$$

Thus, say, if $R=0.7$ and $\rho_o = 0.8$ then we require $\rho_m > 0.33$. A corresponding condition can be derived for categorical variables. In an example we shall explore correlated measurement errors further, but note that these can easily arise in practice

when a set of variables such as obtained from ratings, educational tests etc., are carried out under the same conditions or at the same time and where random variation over conditions or times is present.

In the case of categorical predictors adjusting for misclassifications will often have little effect on the size of the coefficient but may be expected to increase its standard error. Thus, for a binary predictor the coefficient of the dummy (0,1) variable estimates the adjusted difference between the two categories. If there is a weak relationship with the other variables in the model then the process of (randomly) reassigning values from one category to the other will have little effect on the estimated difference but will add random variation to the chain estimates resulting in a larger value for the variability estimate.

Appendix A sets out the steps involved in the MCMC estimation. The general model allows for the possibility that the measurement error covariance matrix can differ from individual (level 1 unit) to individual thus allowing for different groups, for example males and females to have different measurement error distributions. In particular we can allow different measurement error covariance matrices for individuals according to the category observed for a categorical variable where this is assumed to have misclassification errors. This therefore allows for an association between measurement errors and misclassifications. In addition to binary predictors it allows for general categorical variables with misclassification.

4. An example data set

The data we shall use come from a study of the relationship between class size and pupil progress (Blatchford et al., 2002). A cohort of pupils was followed from entry to reception class until the end of the school year, with assessments at the start and end. The response variable is a normalised maths score (end of reception year) *postmaths*. The 5 explanatory variables are: *constant* (=1), *regcls-30* (regular class size centered at 30), normalised pretest maths *pre-maths*, normalised pre test literacy *prelit*, free school meals eligibility *fsmn*.

In the original analysis (Blatchford et al., 2002) a ‘regression spline’ smoothed relationship with class size was fitted rather than the linear relationship examined here¹. Figure 1 shows the resulting relationship.

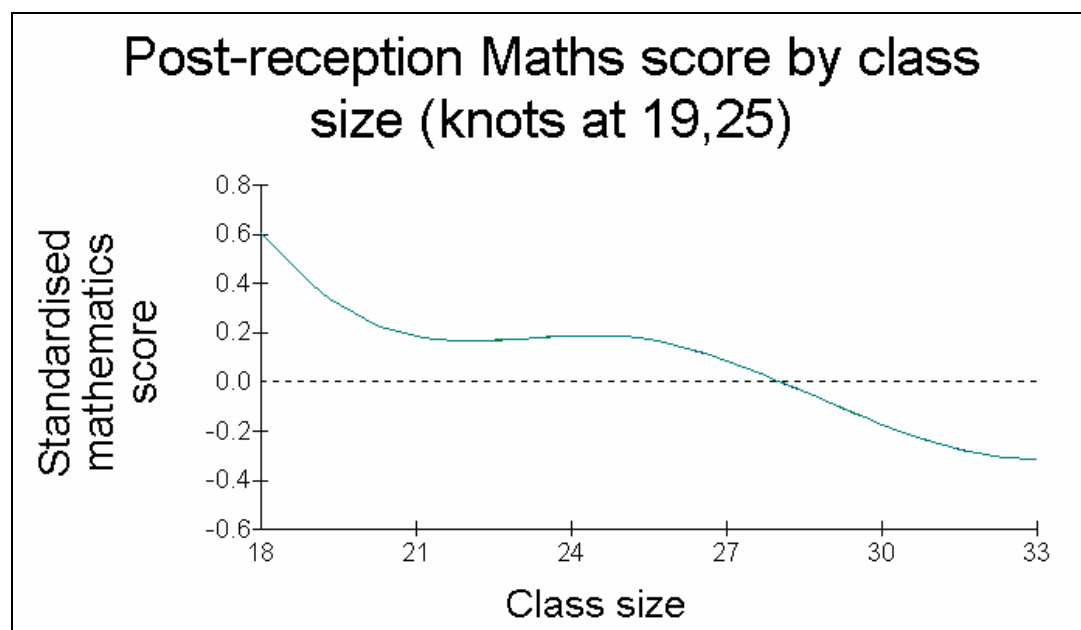
¹ a single level cubic regression with a spline term is defined as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 z_i^3 + e_i$$

$$z_i = \begin{cases} 0 & \text{if } x_i < k \\ x_i - k & \text{if } x_i \geq k \end{cases}$$

This provides a smooth join at the value k , the knot, and allows us better to calibrate the curve for high values of x .

Figure 1.



We now show the MCMC estimates assuming no errors of measurement. There are 4625 pupils in 248 classes (a subset of the original data) and no missing data. All the test scores are transformed to have a standard Normal distribution.

Table 1. Post-test Mathematics related to prior achievements with no measurement errors. Class size measured around value of 30. MCMC estimates; burn in = 500, iterations= 5000.

Coefficient	Estimate	Standard error
Intercept β_0	-0.242	
Class size β_1	-0.068	0.0074
Pre-test Maths score β_2	0.358	0.016
Pre-test literacy score β_3	0.379	0.016
Free school meals (Yes=1) β_4	-0.065	0.028
Between-class variance σ_u^2	0.260	0.027
Between pupil variance σ_e^2	0.381	0.008

It is clear that there is a significant effect of being eligible for free school meals equivalent to a decrease in the adjusted maths score of 0.12 of the pupil level residual standard deviation standard. Likewise, the greater the class size the smaller the posttest mathematics score.

Kounali et al (2007) have analysed the stability of free school meals data at Key Stage 2 and their data suggest that approximately 2% of those not eligible for free school meals at any one time may be classified as eligible. Likewise they suggest that as many as 60% of those eligible may be classified as not eligible. We shall use the illustrative values 2% and 60% respectively in our example. The pretest scores are

based upon teacher assessments and can be expected to have relatively low reliability: we can assume a range of values from 0.6 to 0.9 for these reliabilities. In the following analyses, for illustration, we shall assume a range of values for these reliabilities. It is also reasonable to assume that misclassification errors in FSM are independent of measurement errors in the test scores since the former are ascertained from the school records.

All the following analyses use a burn in of 500 with a sample of 5000 iterations. For training purposes a burn in of 250 and sample of 750 is recommended.

5. Adjusting for measurement errors.

We begin by studying the effect of allowing for measurement errors in the prior test scores, Mathematics and Literacy and we shall assume that both of these have the same reliability. In Table 6 we have summarised the results from all the separate models fitted. We start with Table 2 that shows the parameter estimates where the reliability is assumed to be 0.9 and the measurement errors independent. Table 3 assumes the lowest value of 0.6 for the measurement error. We cannot now, however, assume a zero correlation between the measurement errors, as pointed out above, since the correlation between the observed values is greater than the reliability, being 0.75 . We have assumed a moderate correlation between the measurement errors of 0.5.

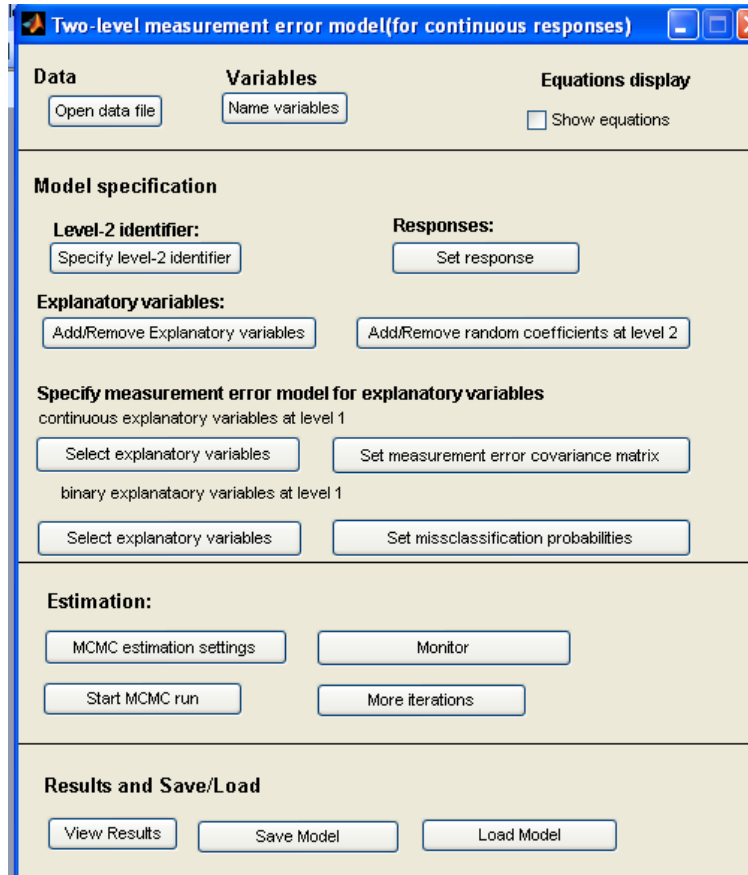
Coefficient	Estimate	Standard error
Intercept	-0.242	
Class size	-0.068	0.007
Pre-test Maths score	0.376	0.026
Pre-test literacy score	0.407	0.025
Free school meals (Yes=1)	-0.066	0.028
Between-class variance	0.260	0.027
Between pupil variance	0.352	0.008

Coefficient	Estimate	Standard error
Intercept	-0.249	
Class size	-0.068	0.008
Pre-test Maths score	0.618	0.094
Pre-test literacy score	0.525	0.090
Free school meals (Yes=1)	-0.063	0.028
Between-class variance	0.260	0.028
Between pupil variance	0.127	0.012

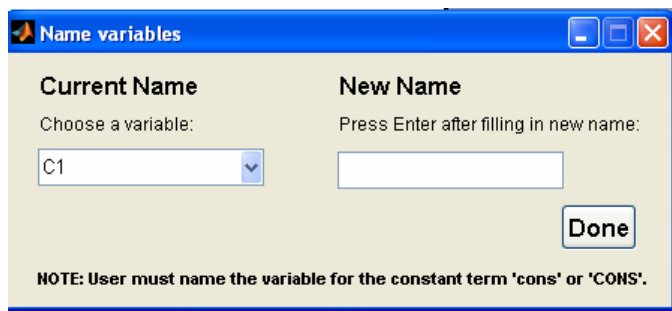
Note that the pretest coefficients are greatly increased with the lowest reliability with also a very large increase in standard error, and also the level 1 variance is reduced as expected.

To fit these models using the software first of all open up the settings window by clicking on the file 'measurement-error.exe' in the directory this was placed in.

A window will appear as follows:

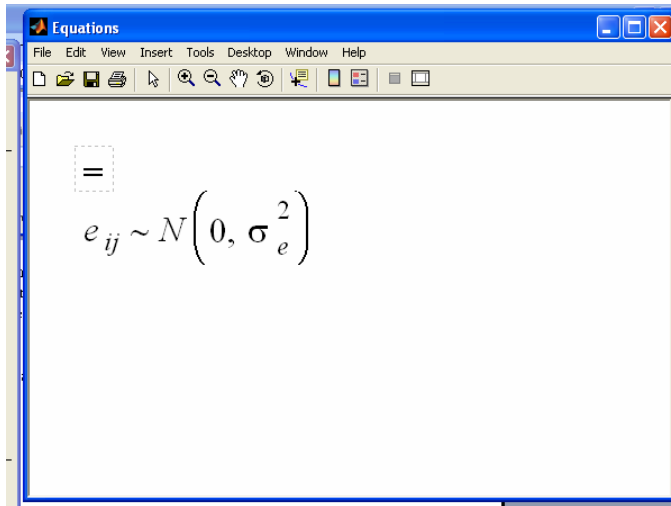


To load a data file click on **open data file** and select the file 'classsize'. Click on **Name variables** and the following window will appear:



The first column is the response, Normalised Maths score so type in a name, say **Normexam**. Note that the constant term is in column 2 and should be called **CONS**. Fill in the remaining names in the order of the fixed effects in Table 2. Note that the final variable is the school identifier. When finished click **Done**.

Now tick the box to display equations and you will see the following screen:



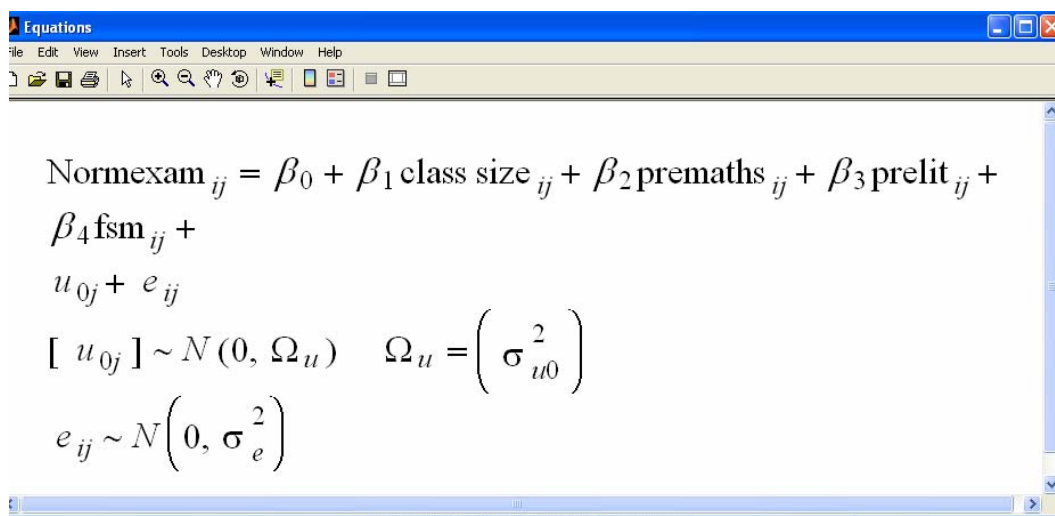
This has just the basic default statement for the level 1 residual distribution. As we define further elements of the model you will be able to see how this changes.

Click on **Specify level 2 identifier**, select **school** (if this is what you named it) and click **Done**.

Now click on **set response** and select the normalised maths score. You will see a box that allows you to set a measurement error variance for the response – for now we shall just leave this as zero, but you may wish to experiment with changing this later.

Now click on **Add/remove explanatory variables** and select all the variables listed in Table 2 – you can hold down the CTRL key to make a multiple selection and then click on the + sign and the **Done**.

Now click on **Add/remove random coefficients at level 2** add **CONS** click the + sign and then **Done**. By this stage the equations screen should appear as follows:



You may need to drag the sides to display the full model.

We have now specified the basic model and it remains to specify any measurement error covariance matrices. You will see that you can do this for the continuous or binary predictors or both. We shall just do this for now for the continuous predictor variables, so click on **select explanatory variables**, and select the prior maths and literacy variables (click + then **Done**). Now click **Set measurement error covariance matrix**, select the two variances ‘maths/math’s’ and ‘literacy/literacy’ and enter 0.111 as the measurement error variance for each – this corresponds to a reliability of 0.9 if we assume the variances of the observed variables are 1.0. Click **Done**. If we wished to set a non-zero covariance we would click the button again and set a value for this term. The equations window now contains the model specification including the measurement error component:

$$\text{Normexam}_{ij} = \beta_0 + \beta_1 \text{class size}_{ij} + \beta_2 \text{premaths}_{ij}^{(o)} + \beta_3 \text{prelit}_{ij}^{(o)} + \beta_4 \text{fsm}_{ij} + u_{0j} + e_{ij}$$

$$[u_{0j}] \sim N(0, \Omega_u) \quad \Omega_u = \begin{pmatrix} \sigma_{u0}^2 & \\ & \end{pmatrix}$$

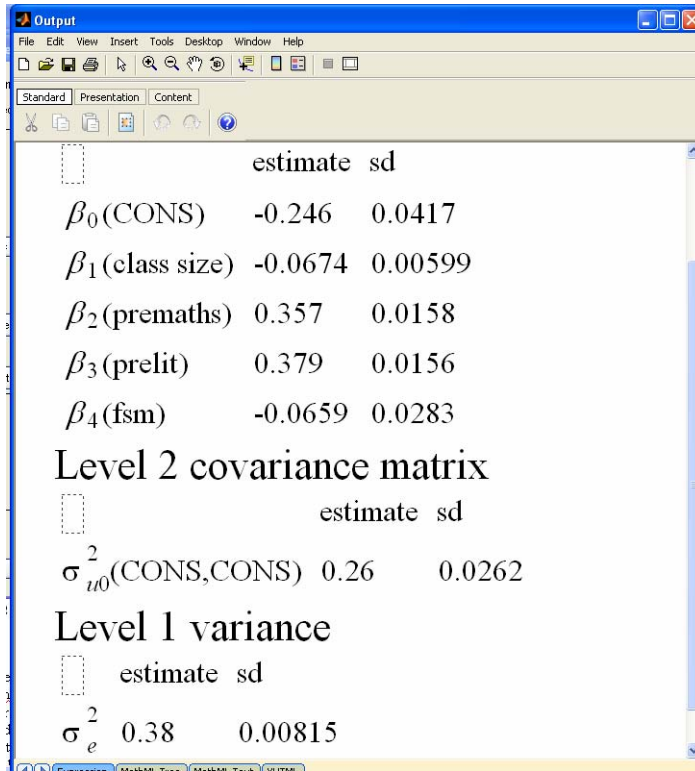
$$e_{ij} \sim N(0, \sigma_e^2)$$

$$\text{premaths}_{ij}^{(o)} = \text{premaths}_{ij} + m_{2ij}$$

$$\text{prelit}_{ij}^{(o)} = \text{prelit}_{ij} + m_{3ij}$$

$$\begin{bmatrix} m_{2ij} \\ m_{3ij} \end{bmatrix} \sim N(0, \Omega_m) \quad \Omega_m = \begin{pmatrix} 0.111 & \\ & 0.111 \end{pmatrix}$$

We now specify the estimation control values. Click on the **MCMC estimation settings** and set burnin to 250, **Number of iterations to** 750 and **Screenrefresh rate** to 10. The screen refresh rate specifies how often the displayed chain is updated. Now click on **Monitor** and select the parameters you wish to see chains displayed for. Note that variance parameters are denoted by a covariance (cov) term where both variables are the same. You will see the chains updating. When iterations have finished you can view the results by clicking on **view results**. You will see the resulting estimates – you will need to drag the screen to see all the results – as follows.



The results are comparable with those in Table 2 which are obtained from a longer chain.

Finally, you can save the model by clicking on **save model** and specifying a distinctive name, for example 'classsizemodel1'. The '.dat' extension will be added by default. If you now wish to modify the model you can click on **load model** and change the specification.

6. Adjusting for missclassification probabilities

We now introduce misclassification probabilities for free school meals. We shall not go through the detailed set up process but just give the results that you can compare with your own analysis.

Table 4. Post-test Mathematics related to prior achievements with measurement errors. Reliability=0.9; Measurement error correlation=0. Missclassification probabilities P(0 1)= 0.60; P(1 0)=0.02 . Class size measured around value of 30.		
Coefficient	Estimate	Standard error
Intercept	-0.224	
Class size	-0.066	0.007
Pre-test Maths score	0.378	0.026
Pre-test literacy score	0.408	0.025
Free school meals (Yes=1)	-0.061	0.035
Between-class variance	0.258	0.027
Between pupil variance	0.351	0.008

In Table 4 we note that the only real change from Table 2 is that the free school meal coefficient standard error has increased.

Finally we allow for measurement error in the response variable, post test mathematics, and the results are given in Table 5.

Table 5. Post-test Mathematics related to prior achievements with measurement error reliability of response $R_y = 0.9$. Reliability of pretest measures=0.9; Measurement error correlation=0. Missclassification probabilities $P(0 1)= 0.60$; $P(1 0)=0.02$. Class size measured around value of 30.		
Coefficient	Estimate	Standard error
Intercept	-0.233	
Class size	-0.067	0.008
Pre-test Maths score	0.374	0.025
Pre-test literacy score	0.403	0.025
Free school meals (Yes=1)	-0.046	0.031
Between-class variance	0.258	0.027
Between pupil variance	0.320	0.008

Now, in addition to a rather smaller increase in standard error of the free school meal coefficient the coefficient itself has decreased in absolute value as expected. Also, as expected, the level 1 variance is reduced.

For comparison purposes the above estimates are set out side by side in Table 6.

Table 6. Estimates from Tables 1-5.					
Coefficient	R=1.0 $\rho = 0.0$ P(0 1)=0 P(1 0)=0 $R_y = 1$	R=0.9 $\rho = 0.0$ P(0 1)=0 P(1 0)=0 $R_y = 1$	R=0.6 $\rho = 0.5$ P(0 1)=0 P(1 0)=0 $R_y = 1$	R=0.9 $\rho = 0.0$ P(0 1)=0.60 P(1 0)=0.02 $R_y = 1$	R=0.9 $\rho = 0.0$ P(0 1)=0.60 P(1 0)=0.02 $R_y = 0.9$
Intercept	-0.242	-0.242	-0.249	-0.229	-0.232
Class size	-0.068	-0.068	-0.068	-0.068	-0.067
Pre-test Maths	0.358	0.376	0.618	0.378	0.374
Pre-test literacy	0.379	0.407	0.525	0.408	0.402
FSM	-0.065	-0.066	-0.063	-0.058	-0.045
Level 2 var.	0.260	0.260	0.260	0.260	0.257
Level 1 var.	0.381	0.352	0.127	0.351	0.320

7. Discussion

We have seen how inferences about both fixed and random effects are changed when we allow for measurement error and misclassification probabilities. An important issue remains that of obtaining suitable estimates for the measurement error variance and misclassification probabilities. In general a range of values should be used in the spirit of a sensitivity analysis since typically these estimates, and especially of measurement error correlations, will at best be approximate.

We also note a further limitation of the current models which assume that measurement errors are limited to variables defined at level 1. However, as suggested in Appendix B, at least for level 2 variables that are aggregates of a level 1 variable, we can often ignore such level 2 measurement errors.

It is suggested that you try assigning different values of measurement error covariance matrices and misclassification probabilities to explore the effect on the estimates.

Appendix A1: Modelling measurement error – MCMC estimation,

1. Introduction

Browne et al., (2001) considered a model with measurement errors in explanatory variables and Browne (2004) implemented this in MlwiN. The features of this model are:

Measurement errors are independent across explanatory variables

The measurement error variance is assumed known

The unknown true values are assumed to have Normal distributions

In addition they considered the case of a polynomial in x and where there was a level 2 variable that was the mean of a level 1 variable. In the latter case they found that ignoring the measurement error of the mean seemed acceptable, since it was relatively small (see Appendix B). However, allowing polynomials complicates the estimation, MH is needed, and it is difficult to implement a general procedure. We therefore do not deal with this here.

The MATLAB routines have implemented the following extensions:

Allowing correlated measurement errors

Allowing for binary, ordered and multcategory explanatory variables.

Thus, apart from the polynomial case, this will provide a quite general procedure for handling measurement errors.

The following is a description of the extended algorithm.

2. Correlated measurement errors

Suppose we have p explanatory variables containing measurement error and q that do not. The model is:

$$y_{ij} = [X_{1ij}(\beta_1 + Z_{1ij} \cdot U_{1j})] + [X_{2ij}(\beta_2 + Z_{2ij} \cdot U_{2j})] + e_{ij} \quad (\text{A.1})$$

$$\beta^T = \{\beta_1^T, \beta_2^T\}, \quad Z^T = \{Z_1^T, Z_2^T\}, \quad U = \{U_1, U_2\}$$

where the explanatory variable matrix of true values for those with measurement error is X_1 ($N \times p$) and that for those without error is X_2 ($N \times q$). For the random part explanatory variables Z_1, Z_2 are indicator vectors of dimensions $(p \times 1)$ and $(q \times 1)$, with ones or zeros, so that the dot (Hadamard) product with the level 2 residuals selects the explanatory variables for the random part of the model – assuming that these are a subset of the fixed part explanatory variables. Using the notation of Browne et al., we have

$$X_1^O \sim MVN(X_1, \Omega_m), \quad X_1 \sim MVN(\theta, \Omega_\phi) \quad (\text{A.2})$$

where X_1^O is the matrix of observed values and Ω_m is the covariance matrix of measurement errors, initially assumed to be common to all level 1 units, θ is the mean vector and Ω_ϕ is the, assumed known, covariance matrix of the true values of X_1 .

We may sample as follows.

$$\begin{aligned} p(\theta | X_1, \Omega_\phi) &\sim MVN(\hat{\theta}, \hat{V}_\theta), \quad \hat{\theta} = \bar{X}_1, \quad \hat{V}_\theta = \Omega_\phi / N \\ p(\Omega_\phi^{-1} | X_1, \theta) &\sim Wishart(N-3, [(X_1 - \hat{\theta})^T (X_1 - \hat{\theta})]^{-1}) \end{aligned} \quad (\text{A.3})$$

where N is the number of level 1 units. Since θ is a row vector of means we assume a uniform prior for θ . For Ω_m we assume this known and in practice we may wish to try different values in a sensitivity analysis. In principle we might also be able to elicit a prior distribution from users that could be used directly in the analysis. For Ω_ϕ we could assume an inverse Wishart prior, but it is not clear what parameters we should use, so we have assumed a uniform prior here.

The sampling for the fixed parameters, β , the residuals, level 2 covariance matrix and level 1 variance, conditional on the X_1, X_2 and given priors, is as in the standard case.

For sampling the X_1 we write

$$p(X_1 | y, X_1^O; \beta, U, \sigma_e^2, \Omega_\phi, \Omega_m) = p(y | X_1; \beta, U, \sigma_e^2) p(X_1^O | X_1, \Omega_m) p(X_1 | \Omega_\phi) \quad (\text{A.4})$$

which leads to the following sampling for each row of X_1 .

$$\begin{aligned} X_{1ij} &\sim MVN(\hat{X}_{1ij}, \hat{V}_{ij}) \\ \hat{V}_{ij} &= \left[\frac{(\beta_1 + Z_1 \cdot U_{1j})(\beta_1 + Z_1 \cdot U_{1j})^T}{\sigma_e^2} + \Omega_m^{-1} + \Omega_\phi^{-1} \right]^{-1} \\ \hat{X}_{1ij} &= \hat{V}_{ij} \left[\frac{(\beta_1 + Z_1 \cdot U_{1j})(y_{ij} - X_{2ij}(\beta_2 + Z_2 \cdot U_{2j}))}{\sigma_e^2} + X_{1ij}^O \Omega_m^{-1} + \theta \Omega_\phi^{-1} \right] \end{aligned} \quad (\text{A.5})$$

where $Z \cdot U$ denotes the Hadamard vector product. The level 1 residuals are obtained by subtraction.

In some applications the measurement error covariance matrix may vary across level 1 (or level 2) units, for example as a known function of predictor variables. In this case we simply replace Ω_m^{-1} by Ω_{mij}^{-1} in (5).

If we have measurement error in the response

$$y^O = y + e_y, \quad e_y \sim N(0, \sigma_{e_y}^2) \quad (\text{A.6})$$

in order to ensure identification we must know variance $\sigma_{e_y}^2$. We apply this to the residuals using the adjusted value $\sigma_{e_y}^{*2} = \sigma_e^2 \sigma_{e_y}^2 / \sigma_y^2$ and we insert the extra step to sample y_{ij} from

$$N[(\sigma_e^2 - \sigma_{e_y}^{*2})\sigma_e^{-2}\tilde{y}_{ij} + \hat{y}_{ij}, (\sigma_e^2 - \sigma_{e_y}^{*2})\sigma_e^{-2}\sigma_{e_y}^{*2}] \quad (\text{A.7})$$

where \hat{y}_{ij} is the predicted value and $\tilde{y}_{ij} = y_{ij}^O - \hat{y}_{ij}$.

3. Binary and ordered category explanatory variables

We consider the binary variable case.

One possibility is to assume a threshold model as follows:

$$P(\text{observed} = 1) = \int_{\bar{x}^o}^{\infty} \phi(t) dt, \quad x^o = x + m \quad (\text{A.8})$$

where \bar{x}^o is the mean of the observed underlying variable, $\phi(t)$ is the standard Normal distribution, x^o is the underlying continuously distributed variable including measurement error, m , and P is the probability of observing a one. We assume, for now, that for the binary variables the measurement errors are mutually independent and independent of the measurement errors for the continuous variables. Otherwise, for each binary variable, when we make a draw from the underlying continuous variable, we will need to condition also on the observed values of the other variables.

We insert an extra step into the MCMC algorithm as follows:

We first draw from the underlying continuous distribution choosing a random draw from the standard Normal distribution in the range (\bar{x}^o, ∞) if the response is a one and from $(-\infty, \bar{x}^o)$ if the response is a zero. We then draw from the distribution of the true value given the observed value, i.e. from:

$$N[(1 - \sigma_m^2)(x^o - \bar{x}) + \bar{x}, (1 - \sigma_m^2)\sigma_m^2]$$

We can also allow a prior distribution for σ_m^2 .

Now $(1 - \sigma_m^2)$ is the reliability, but typically we will have no direct estimate of this. We note, however, that the probability of observing a one, given that the true value is zero is given by

$$\int_{-\infty}^{\bar{x}} \left[\int_{\bar{x}-x}^{\infty} f(m) dm \right] \phi(t) dt = \int_{-\infty}^{\bar{x}} \left[\int_{(\bar{x}-x)/\sigma_m}^{\infty} \phi(t) dt \right] \phi(t) dt \cong \int_{-\infty}^{\bar{x}^o} \left[\int_{(\bar{x}^o-x)/\sigma_m}^{\infty} \phi(t) dt \right] \phi(t) dt \quad (\text{A.9})$$

which can be evaluated to provide an estimate for the parameter σ_m numerically. A corresponding expression holds for the probability of observing a zero when the true value is one. Thus with estimates of the misclassification probabilities we can obtain two estimates of the measurement error variance that can be combined. In fact, we could also in principle obtain a further estimate of the true mean value \bar{x} from the above that could be used, but this may not provide much extra information. In practice it is common to assume that the two misclassification probabilities are equal, in which case we only need (9).

The other possibility, that we have implemented, is to work directly with the observed values and misclassification probabilities.

4. Missclassification probabilities

Suppose we write the probability of observing a zero given that the true value is zero as $P_{obs}(0|0)$ and the probability of observing a one given that the true value is a zero as $P_{obs}(1|0)$, etc. Then the probability of observing a zero is $P_{obs}(0) = P_{true}(0)P_{obs}(0|0) + P_{true}(1)P_{obs}(0|1)$ and the probability of observing a one is $P_{obs}(1) = P_{true}(1)(1 - P_{obs}(0|1)) + P_{true}(0)(1 - P_{obs}(0|0))$ where $P_{true}(0)$, $P_{true}(1)$ are the true probabilities of a zero and one.

This gives the following values for the true (prior) probabilities

$$P_{true}(0) = \frac{P_{obs}(1|1) - P_{obs}(1)}{P_{obs}(1|1) + P_{obs}(0|0) - 1}, \quad P_{true}(1) = 1 - P_{true}(0)$$

Consider a Normal response model. The probability for an observation that has true value zero where we *observe* a zero for the binary variable x_1 with coefficient β_1 which is assumed to have a uniform prior, is proportional to

$$L_{00} = \exp\left(-\frac{(\tilde{y})^2}{2\sigma_e^2}\right) P_{obs}(0|0)$$

and for an observed zero where the true value is one we have the probability proportional to

$$L_{01} = \exp\left(-\frac{(\tilde{y} - \beta_1)^2}{2\sigma_e^2}\right) P_{obs}(1|0)$$

where \tilde{y} is the observed response minus predicted value of the response given the remaining parameters.

When a zero is observed, combining these probabilities with the priors, we select a new true value to be zero with probability

$$\frac{L_{00}P_{true}(0)}{L_{00}P_{true}(0) + L_{01}P_{true}(1)}$$

We have corresponding results when a one is observed, namely

$$L_{10} = \exp\left(-\frac{(\tilde{y})^2}{2\sigma_e^2}\right) P_{obs}(1|0)$$

$$L_{11} = \exp\left(-\frac{(\tilde{y} - \beta_1)^2}{2\sigma_e^2}\right) P_{obs}(1|1)$$

and we select a new true value of one with probability

$$\frac{L_{11}P_{true}(1)}{L_{11}P_{true}(1) + L_{10}P_{true}(0)}$$

Having sampled a new set of true values we then apply the standard steps in the MCMC algorithm for the remaining parameters. For generalised linear models the only change is in the expressions for the likelihoods and if we use, e.g. a probit link with binary data then there is no change except for the extra step generating a Normally distributed response from the binary response.

The extension to the multicategory, ordered or unordered, case requires us to evaluate the true priors for each category and then evaluating the corresponding probabilities. This, therefore, requires a misclassification matrix to be known, or a good estimate available.

5. Extensions

We may also consider models where the measurement error variances or misclassification probabilities are a function of further variables where the function parameters are to be estimated. Further work on this is planned. Missing responses can be handled by adding an imputation step for the missing data based on current parameter estimates. We can also introduce a prior distribution for the measurement error covariance matrix and introduce a further step that involves sampling from this prior.

We have assumed that there is no association between the Normal variable measurement errors and the misclassifications. One way to introduce an association is to allow the Normal measurement error covariance matrix to depend on the observed category so that for each such category, or combination of categories, we assume a known Ω_m^c where c denotes the category or category combination. In practice this is achieved by choosing corresponding Ω_{mij}^{-1} in (A.5).

A full description of the model and estimation procedure is given by Goldstein et al., (2007).

Appendix B1: Aggregating level 1 variables with measurement errors

1. Introduction

In a multilevel model where there is a level 2 (or higher level) predictor that is defined as an aggregation from the level 1 units within the cluster, we can distinguish two kinds of inferences. In the first we wish to condition on the underlying, but unknown, ‘true’ value of the variable. Thus, in educational data we may suppose that the average prior attainment of a school influences the subsequent attainment of individuals within it, where this average attainment is used as a proxy for the long-term intake characteristics of the school. It can then be argued that the observed attainment should be regarded as a variable measured with error where the analysis will attempt to correct for the measurement error. Alternatively, we may regard the actual average score itself as the influential variable so that, if it is measured accurately, there is no measurement error. We postpone a discussion of the role of level 1 measurement error until later. We shall also introduce below the common situation when the average is not available, but only an estimate of it.

In the first case above, for simplicity assume that the variable is Normally distributed, and that we have fitted a simple variance components (VC) model so that the total variance is

$$\sigma_T^2 = \sigma_u^2 + \sigma_e^2 \quad (\text{B.1})$$

Thus the variance of the mean of the N level 1 units in a level 2 unit is

$$\sigma_u^2 + \sigma_e^2 / N \quad (\text{B.2})$$

Since inference is with respect to the ‘true’ mean the measurement error variance is simply

$$\sigma_e^2 / N$$

with corresponding reliability (B.3)

$$\rho_T = N\sigma_u^2 / (N\sigma_u^2 + \sigma_e^2)$$

Which is just the ‘shrinkage’ factor. In many applications where N is very large, measurement error can be ignored, although attention needs to be paid to the value of the VPC ($\sigma_u^2 / (\sigma_u^2 + \sigma_e^2)$).

In the second case, where inference is with respect to the observed mean then the reliability is 1.0.

2. Sampling level 1 units

In the common situation where we only have a sample of n out of N level 1 units (B.2) becomes

$$\sigma_u^2 + \sigma_e^2 / n \quad (\text{B.4})$$

And the reliability becomes

$$\rho_T = n\sigma_u^2 / (n\sigma_u^2 + \sigma_e^2) \quad (\text{B.5})$$

Thus, for example with a Variance Partition Coefficient (VPC) (Goldstein et al., 2002) of 0.1 and $n=20$ we have $\rho_{T_1} = 0.69$. This essentially is the ‘true value’ definition adopted by Sampson et al. (Sampson et al., 1997).. In fact these authors fit a 3-level model where level 1 is the item level for the scale components. Level 2 is individual and level 3 is area. Their model can be formulated as a single factor model with scale item loadings equal to 1 (Rasch model). This formulation enables them to estimate individual level reliabilities also which can be incorporated if required. In practice n is typically large enough to ignore these when estimating the level 2 reliability (but see below).

Where inference is WRT the observed mean the reliability becomes

$$\rho_O = \sigma_u^2 + \sigma_e^2 / N(\sigma_u^2 + \sigma_e^2 / n)^{-1} = (n\sigma_u^2 + \binom{n}{N}\sigma_e^2) / (n\sigma_u^2 + \sigma_e^2) \quad (\text{B.6})$$

Which becomes 1.0 if the mean is computed from all the level 1 units with a cluster. If we write v for the VPC we have

$$\rho_O = (n + \binom{n}{N}\binom{1-v}{v}) / (n + \binom{1-v}{v}) \quad (\text{B.7})$$

As v tends to zero this tends to (n/N) as does (B.5). Now, the level 2 variance will often be sensitive to the population considered, or alternatively, the estimate of the VPC will depend on which other variables we adjust for in its estimation, especially if these are level 2 variables. In general the appropriate population will be the one that we intend to use in subsequent models where we adjust for the measurement error.

In the above example with a VPC of 0.1, $N=30$ and $n=5$, as we might have for educational data on classes we have $\rho_O = 0.46$. For survey data on small areas say with $N=200$, $n=20$ we have $\rho_O = 0.72$, which is not very different from the ‘true’ definition value given above.

If we now consider the (independent) measurement error reliability at level 1, say ρ_1 . Expression (7) becomes

$$\rho_{O_1} = (n + \binom{n}{N}\binom{1-v}{v} + \binom{n}{N^2}\binom{1-v}{v}\binom{1-\rho_1}{\rho_1}) / (n + \binom{1-v}{v} + \binom{1}{n}\binom{1-v}{v}\binom{1-\rho_1}{\rho_1}) \quad (\text{B.8})$$

So that this aggregated level 1 error term can typically be ignored.

3. Further considerations

The distinction between the ‘true’ and ‘observed’ definitions for reliability becomes important only when the actual cluster (level 2 unit) size is relatively small. This will usually be the case with certain kinds of data such as in education, but may also hold for certain kinds of survey data, especially in small area analysis.

For categorical variables, we are dealing with misclassification probabilities at level 1 but to a first approximation can assume Normality at level 2. Thus, for binary responses, we would substitute in the above formulae (B.5) and (B.7) corresponding terms based on the variance of a proportion. For ordered responses we can approximate by treating as a continuous variable and for multicategory responses we would use the corresponding multinomial variances and covariances – allowing for correlated measurement errors A further possibility is to assume a threshold model, but this adds further numerical complications concerned with estimating a

measurement error variance given just misclassification probabilities (see Appendix A).

Chapter 2. Modelling multilevel latent variable structures

1. Introduction

In the social and medical sciences it is common to incorporate into models ‘latent’ variables that are not directly measured, but whose existence is defined by their relationships to sets of observed variables or ‘indicators’. The simplest such models, known as factor analysis models have been in existence for many years (see e.g. Lawley and Maxwell (1971) for an introduction) and their generalisation to ‘structural equation models’ can largely be dated to a seminal paper of Joreskog (1969). Traditionally all such models assumed a single level data structure until McDonald and Goldstein (1989) introduced multilevel versions of the basic factor model, further developed by Muthen (2002), and Skrondal and Rabe-Hesketh (2004).

Existing software can handle several kinds of multilevel structural equation models. GLLAMM (Rabe-Hesketh et al., 2001) can fit general structural equation models to multilevel data with responses that are Normal, binary or ordered. MPLUS (Muthen and Muthen, 1998) can fit a similar range of models up to 2 levels and MLwiN can fit multilevel factor, but not structural equation, models. The MATLAB routines that have been developed extend existing models in the following ways. First, they allow certain constraints across parameters that are important for interpretation. Secondly, they allow different ways of specifying level 2 latent variables and thirdly they use MCMC estimation rather than maximum likelihood (ML). One problem with ML estimation is that it becomes very slow when the number of parameters in the model becomes large, typically increasing factorially with the number of parameters; MCMC estimation, however, avoids this kind of dependence on the number of parameters.

We shall present two examples, one from demography and one from education, that illustrate, for two levels, how to set up and analyse such models. First of all we set out some basic notation and describe some simple models.

2. One level factor models

A basic single level factor model with a single factor can be written as follows:

$$y_{ij} = \beta_{0i} + \lambda_i \eta_j + e_{ij} \tag{1}$$
$$\eta_j \sim N(0, \sigma_\eta^2), e_{ij} \sim N(0, \sigma_{ei}^2)$$

This is commonly referred to as a ‘measurement model’. Here i ($1, \dots, p$) indexes the response variable and j indexes the individual unit – in our analyses this will be a person. The y variables are observed, η is an unobservable or ‘latent’ factor and the remaining terms are the parameters to be estimated. Note that the parameters β_{0i} estimate the means of the responses. To ensure identifiability we need to fix either one (or more) of the loadings, the λ_i , to a known value (e.g. 1) or fix the factor variance, σ_η^2 , to a known value, typically 1. It is assumed that the responses are

jointly Normally distributed and if they have been standardised and σ_η^2 set to 1, then the loadings can be interpreted as the correlations between the underlying factor and each of the corresponding responses. Typically, the factor is interpreted in terms of the values of these correlations. Thus, if a subset of the responses has high correlations then the factor will be interpreted in terms of what these responses are assumed to measure. We shall return to this issue below.

Model (1) can be elaborated in several directions. One is by extending it to the multilevel case and we shall return to this in the next section. A second possibility is to include one or more further factors. We can write a 2-factor model as

$$\begin{aligned}
 y_{ij} &= \beta_{0i} + \lambda_{1i}\eta_{1j} + \lambda_{2i}\eta_{2j} + e_{ij} \\
 \begin{pmatrix} \eta_{1j} \\ \eta_{2j} \end{pmatrix} &\sim N \begin{pmatrix} \sigma_{\eta 1}^2 & \\ \sigma_{\eta 12} & \sigma_{\eta 2}^2 \end{pmatrix} \\
 e_{ij} &\sim N(0, \sigma_{ei}^2)
 \end{aligned} \tag{2}$$

where we now have two sets of loadings and a bivariate Normal distribution for the factors. For identifiability we now have to impose a further condition so that the two factors can be separated. One possibility is to set the factor covariance in (2) to zero and additionally constrain one loading to zero, for example $\lambda_{11} = 0$. Another possibility commonly used is to choose 2 subsets of the responses, say A and B. For subset A only the corresponding loadings on factor 1 are non-zero and for subset B only the corresponding loadings on factor 2 are non-zero. We now need to allow a non-zero covariance between the factors. This latter solution is popular and often known as ‘simple structure’. Prior to settling on a model specification, we may carry out an exploratory analysis in an attempt to assess what might be a reasonable model. We shall not discuss this issue further – a good introduction is Lawley and Maxwell (1971).

Finally we can define factors in terms of particular explanatory variables. Let us rewrite (1) as

$$\begin{aligned}
 y_{ij} &= \beta_{0i} + \lambda_i \eta_j z_j + e_{ij} \\
 \eta_j &\sim N(0, \sigma_\eta^2), \quad e_{ij} \sim N(0, \sigma_{ei}^2), \quad z_j = 1
 \end{aligned} \tag{3}$$

Now in fact z_i can be any known explanatory variable, or we can have several. Thus, suppose we wish to make the factor structure a function of, say, age we can write

$$\begin{aligned}
 y_{ij} &= \beta_{0i} + \lambda_{1i}\eta_{1j}z_{1j} + \lambda_{2i}\eta_{2j}z_{2j} + e_{ij} \\
 \eta_{1i} &\sim N(0, \sigma_{\eta 1}^2), \quad \eta_{2i} \sim N(0, \sigma_{\eta 2}^2), \quad e_{ij} \sim N(0, \sigma_i^2) \quad z_{1j} = 1 \quad z_{2j} = 'age'
 \end{aligned} \tag{4}$$

And the parameters of this 2-factor model will in general be identifiable. If age is centered then the interpretation of (4) will then be in terms of a general factor and one that is proportional to the departure of an individual from the mean age. For the models we shall be using we will restrict ourselves to just one such explanatory variable, the constant vector that we will refer to by the name ‘cons’.

3. Adding further explanatory variables and structural models

Returning to the one factor model for simplicity we can also extend this by adding further explanatory variables for the responses. For example we may wish to adjust for a gender difference in each response and so we can write

$$y_{ij} = \beta_{0i} + \beta_{1i}x_{1j} + \lambda_i \eta_j + e_{ij} \quad (5)$$

$$\eta_j \sim N(0, \sigma_\eta^2), e_{ij} \sim N(0, \sigma_{ei}^2)$$

Where x_1 is a dummy variable for gender. Alternatively we can introduce covariates of this kind by allowing the factor itself to depend on them, giving a ‘structural’ model

$$y_{ij} = \beta_{0i} + \lambda_i \eta_j + e_{ij}$$

$$\eta_j = \beta_1 x_{1j} + \eta_j^* \quad (6)$$

$$\theta_j^* \sim N(0, \sigma_\eta^{*2}), e_{ij} \sim N(0, \sigma_{ei}^2)$$

where the first line of (6) is the measurement part of the model and the second line describes the structural part of the model; together with the distributional assumptions (4) is a multilevel structural equation model (SEM). It reduces, on substituting the second line into the first, to

$$y_{ij} = \beta_{0i} + \lambda_i \beta_1 x_{1j} + \lambda_i \eta_j^* + e_{ij} \quad (7)$$

$$\theta_j^* \sim N(0, \sigma_\eta^{*2}), e_{ij} \sim N(0, \sigma_{ei}^2)$$

The major distinction between (5) and (7) is that the coefficients of x_1 are proportional to the loadings. We have economised on the number of parameters by incorporating the relationship with x_1 into the structural part of the model. For this reason, if (7) does not result in a significantly worse fit to the data we would generally prefer it for it’s generally simpler interpretation. One limitation of the present routines is that they cannot fit very general structural models, most notably those where we have linear relationships among different factors.

An important feature that we notice by inspecting (5) and (7) is that as further explanatory variables are added the values of the loadings in general will change. This in turn implies that the interpretation of the factor itself will change, possibly substantially. Thus, unlike ordinary models such as multiple regression with measured predictors, we cannot directly interpret the effect on factor loadings of introducing further variables into a model, since changes in these loadings changes our interpretation of what the factor means. One method of dealing with this is to fix the values of the loadings as estimated from some suitable ‘standard’ model and then use these values in all further analyses. An extreme version of this occurs in item response models, the so called Rasch model, where all the loadings are constrained to equal 1 from the outset (Goldstein et al., 2007). We shall be looking at this when we analyse our data set.

4. Two level factor and structural equation models

We now consider adding a further level to the basic model. We can think of the data structure as, say, pupils (j) nested within schools (k) and one possibility is to extend (1) as follows

$$y_{ijk} = \beta_{0i} + \lambda_i^{(1)}\eta_{jk}^{(1)} + \lambda_i^{(2)}\eta_k^{(2)} + u_{ik} + e_{ijk} \quad (8)$$

$$\eta_{jk}^{(1)} \sim N(0, \sigma_{\eta(1)}^2), \quad \eta_k^{(2)} \sim N(0, \sigma_{\eta(2)}^2), \quad e_{ijk} \sim N(0, \sigma_{ei}^2), \quad u_{ik} \sim N(0, \sigma_{ui}^2)$$

So that we have independent factor structures at each level. An alternative, corresponding to (6) is to allow the level 1 factor to depend on level 2 random effects as follows

$$y_{ijk} = \beta_{0i} + \lambda_i^{(1)}\eta_{jk}^{(1)} + e_{ijk}$$

$$\eta_{jk}^{(1)} = \eta_{jk}^{(1)*} + u_k^* \quad (9)$$

$$\eta_{jk}^{(1)*} \sim N(0, \sigma_{\eta(1)}^{*2}), \quad e_{ijk} \sim N(0, \sigma_{ei}^2), \quad u_k^* \sim N(0, \sigma_u^{*2})$$

In a similar fashion to the structural model (7) we have replaced the separate level 2 residuals u_{ik} with a single level 2 residual in the structural part for the level 1 factor. This again may lead to a simpler interpretation. We can add covariates and additional factors as before either to model (8) or (9).

5. Models for binary and ordered responses

In our educational example we in fact have responses that are binary (or ordered) test items and we therefore need to modify our expression of the model. A convenient way to do this is as follows. Consider the standard Normal variable

$$y_i \sim N(\beta_{0i} + \lambda_i\eta_j, 1), \quad \eta_j \sim N(0, \sigma_\eta^2)$$

Where we observe a positive (=1) response for our binary variable z if y is positive, that is

$$y_{ij} = \beta_{0i} + \lambda_i\eta_j + e_{ij} > 0 \quad \text{or}$$

$$e_{ij} > -(\beta_{0i} + \lambda_i\eta_j)$$

We have

$$\text{Pr ob}(z = 1) = \text{Pr ob}(e_{ij} > -(\beta_{0i} + \lambda_i\eta_j)) = \int_{-(\beta_{0i} + \lambda_i\eta_j)}^{\infty} \phi(t)dt = \int_{-\infty}^{(\beta_{0i} + \lambda_i\eta_j)} \phi(t)dt \quad (10)$$

Where $\phi(t)$ is the density function of the standard Normal distribution. This is the standard probit model. A common alternative is to use the logit 'link function'

$$\begin{aligned}
z_{ij} &\sim \text{binomial}(1, \pi_{ij}) \\
\pi_{ij} &= \text{prob}(z_{ij} = 1) \\
\text{logit}(\pi_{ij}) &= \beta_{0i} + \lambda_i \eta_j \\
\eta_j &\sim N(0, \sigma_\eta^2),
\end{aligned} \tag{11}$$

In practice these two formulations are very similar and we use (10) because it has computational advantages.

Where we have an ordered classification we can extend the binary probit model by considering the cumulative probability of being in one of the lowest $s+1$ categories as

$$\gamma_{ij}^s = \sum_{f=0}^s \pi_{ij}^f \tag{12}$$

Where the categories are numbered from 0 upwards. We now consider the underlying Normal variable

$$\begin{aligned}
y_{ij} &= \beta_{0i} + \alpha_s + \lambda_i \eta_j + e_{ij} \\
e_{ij} &\sim N(0,1), \quad \alpha_0 = 0
\end{aligned}$$

So that we now have the probability of an observation in category $s+1$ or higher is

$$\int_{-\infty}^{\alpha_s + (\beta_{0i} + \lambda_i \eta_j)} \phi(t) dt .$$

Thus we now have the extra ‘threshold’ parameters $\{\alpha_s\}$ to estimate in our model. In fact, for convenience the macros written for these models use the formulation

$$\int_{-\infty}^{\alpha_s - (\beta_{0i} + \lambda_i \eta_j)} \phi(t) dt \tag{13}$$

which is equivalent but with signs of the parameters reversed.

Full computational details are given in Appendix A of Chapter 3.

With these discrete response models the estimation proceeds first by choosing an underlying y value and then proceeding as with continuous Normal responses with the residual variance fixed at 1.

6. An educational example

We analyse data from the Programme for international Student Assessment (PISA) survey of reading performance, that represents a very ambitious and wide ranging attempt to measure and compare 15 year olds in 32 countries, and that utilises procedures and models used widely in analysing educational performance. Under the auspices of the Organisation for Economic Co-operation and Development (OECD) the testing for PISA was carried out in the first half of 2000, and the study was intended to be the first of a series. While concentrating on Reading it also has

Mathematics and Science components. The second study carried out in 2003 concentrates on Mathematics and the third in 2006 concentrated on Science. The sampling design selected schools as first stage units and sampled 15-year-old pupils within schools with a maximum of 35 students within each school. Extensive piloting of test items and general procedures, including translations, was carried out. The first comprehensive report (OECD, 2001) appeared in 2001 and an extensive (300 page) technical report (Adams and Wu, 2002) provides detail about the procedures used. In addition data are available for secondary analysis from the OECD web site (www.pisa.oecd.org/pisa/outcome.htm).

The PISA 2000 (OECD, 2001) analyses have concentrated on computing student proficiencies and country means for the three reading proficiency scales of ‘retrieving information’, ‘interpreting texts’ and reflection and evaluation’ as well as a combined scale. Each scale is defined by a different set of items. We will analyse data from the first subscale, ‘retrieving information’ containing 35 items. Details of this subscale can be found in the PISA 2000 technical report (Adams and Wu, 2002). The full scale contains 36 items, but one of these (R076Q03) was eliminated from the England file as ‘dodgy’ because it did not fit well using the 1-dimensional scaling procedure applied in the study. Most of the items are binary with 4 being 3 category ordered responses. Two countries, France and England have been chosen for this purpose.

The original data consisted of 326 schools, 141 in England and 185 in France, and 8299 students, 4070 in England and 4229 in France. We shall analyse a 30% random sample in order to speed up the analyses, but in the following tables we give the results from an analysis of the complete data and these can be compared with those for the 30% sample. It is worth mentioning that the interpretation of comparisons between the countries is both complicated and controversial and a discussion is given in Goldstein et al., (2007).

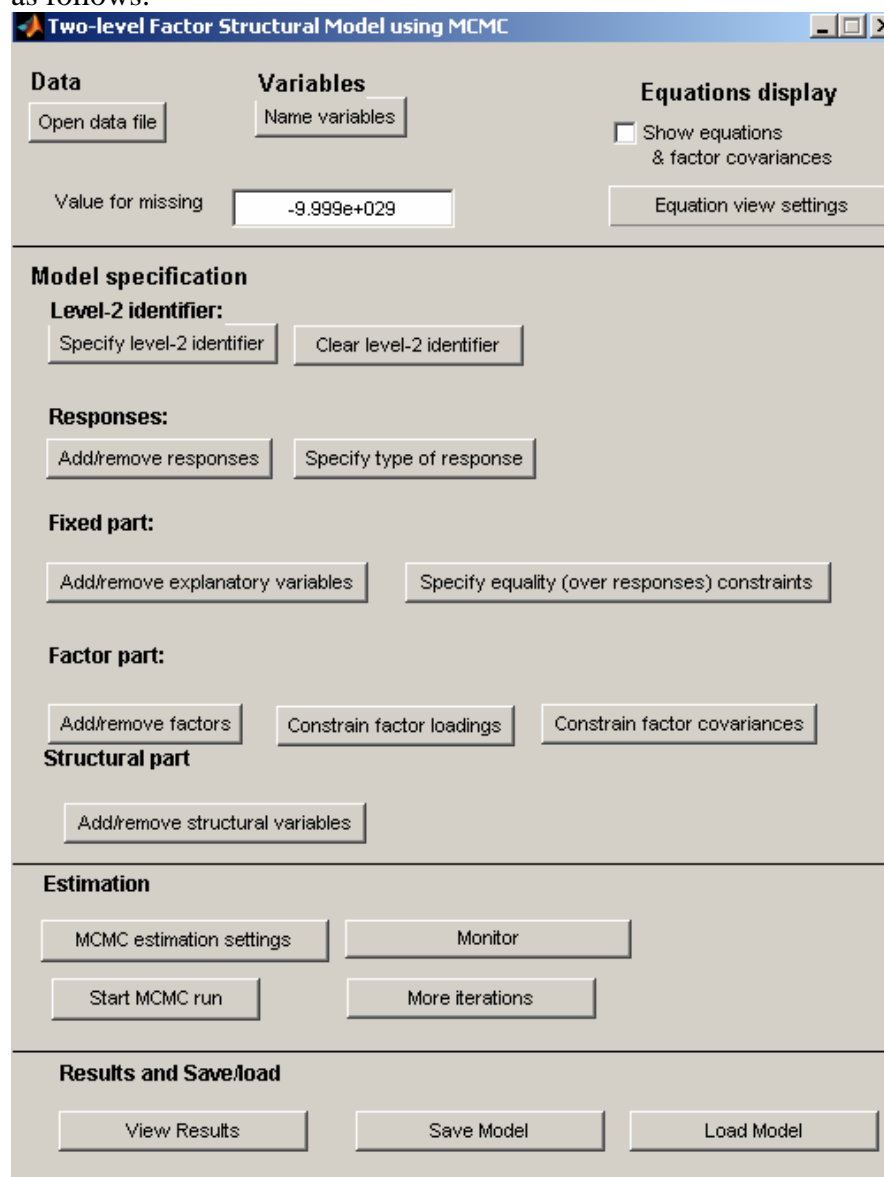
An important feature of this data set and many similar ones is that not every student has responses to all the items in the test. The items themselves are spread over booklets and each student takes a subset of booklets resulting in each pupil having many ‘missing’ responses. In our analysis the MCMC algorithm deals efficiently with a missing response by randomly ‘imputing’ it. An alternative is a ‘multiple imputation’ procedure that provides users with a small number (e.g. 5) ‘complete’ data sets where the missing values have been randomly imputed. The user then fits their model to each data set and combines the results in a prescribed fashion. This is the procedure used by OECD, but it does not take into account the multilevel structure of the data (pupils within schools) and is not generally as efficient as the analyses we shall be carrying out. For more details about multiple imputation procedures in multilevel models see www.missingdata.org.uk.

We begin by looking at some single level models and then will switch to 2-level models.

7. Fitting single level factor models

We begin by fitting model (1) to our data set with 31 binary responses and 4 3-category ordered responses, using the latent Normal variable approach described above. We start by running the file structural-equation.exe. The data are in a file

called 'pisadata'. On starting up the program we are confronted with a menu window as follows:



Click on 'Open data file' and load *pisadata*: click on this. The windows used to specify the model are similar in format to those used for the measurement error model and we shall only describe features specific to latent variable models. When naming the variables it is suggested that the first 35, which are the responses are left as C1, C2,...C35. C36 is the constant – name this 'cons', C37 is country (England =1) C38 is gender (Male =1) C39-c43 are 5 dummy variables for month of birth (march-April=1, May-June=2, July-Aug=3, Sept-Oct=4, Nov-Dec=5), C44 is the level 2 unit ID. Identify this by clicking on **specify level-2 identifier**. If you are fitting just a single level model click on **clear level-2 identifier**.

You can also set a missing value code for the data – the default value, as shown, is -999.9×10^{29} (code as -9.999e+29) and this is the value used in our data set.

We must now specify the response variables (**Add/remove responses**) and for each response we must specify the type – in our case either binary or ordered. The 4

ordered variables are C9, C14, C20, and C26, each with 3 categories. Note that you can select consecutive responses and give them the same type.

At this point let's click on **Show equations** and you will see that each response is listed as an additive function of the level 1 and level 2 residuals: we haven't yet specified any fixed predictors or factors. Click on **Equation view settings** and you will see several options. By **Response lines** click on **Collapse** and you will see just a single line with a generic expression for the model so far.

We can now add explanatory variables in the measurement part of the model – for now we shall just fit an intercept to take up the mean for each response. The next button allows you to constrain any coefficient to be equal across responses but we shall not use this. Now add a single factor at levels 1 and 2 and constrain their variances to be equal to 1.0.

If you now click on MCMC estimation settings you will see that you have to enter the 'burn in', 'number of iterations' and screen 'refresh rate' (n). The latter simply tells you how many MCMC cycles or iterations have been done – printing the number after every n-th. For more details about MCMC estimation see the section 'Introduction to MCMC estimation'. The model runs quite slowly so choose a small number for the burn-in, say 30, and run for 100 iterations.

If, after the program has finished you wish to add more iterations, click 'resume iterations'. You will see the results appear on the screen.

The full results for the model (in fact a simpler model with just a single level) are given in Table 1 that shows the results for the intercept parameters model. There are two models fitted here: the model where we allow loadings to be estimated and a model (the Rasch model) where the loadings are constrained to be equal (using the 'constrain factor loadings' button) to 1. Since the factor mean is zero the intercept parameter represents the mean for the item on the probit scale. Thus, a value of 0.64 means that the probability getting this item correct is the cumulative probability of the Standard Normal distribution for that value, i.e. 0.74.

Table 2 shows the factor loadings. Note that for the equal loading case where the factor loadings are constrained to 1 the factor variance is now estimated. We also see in this table a 'DIC value'. Here we see a difference of over 600 indicating that the model where loadings are estimated is a far better fit.

Finally, let us look at the issue we mentioned earlier, namely how to interpret the factors as further variables are added. One option is to approximate as in the Rasch model and keep the loadings equal to 1 so that the interpretation of the factor stays the same as further explanatory variables are added. The second is to fix the loadings as a result of a preliminary analysis and thereafter constrain them to be equal to these values. See what happens to the DIC when you do this. How are your interpretations affected?

You can then go on to look at a structural model where we introduce variables such as Country and gender.

We have not covered all possibilities in these notes, and if you want to take the analyses further see the Goldstein et al (2007) paper and also one by Steele and Goldstein (2006) that looks at structures for handling more than one factor.

8. Tables

Table 1. Comparisons of intercept parameter estimates – (threshold parameter estimates for four ordered category items in brackets). Burn-in=1000: sample=5000. Single level factor model with 8299 pupils.

<i>Question</i>	Equal Loadings (Rasch model)	Equal Loadings (Rasch model) standard errors	Unequal Loadings	Unequal Loadings standard errors
R040Q02	0.66	0.04	0.64	0.04
R040Q03A	0.34	0.03	0.36	0.04
R070Q02	0.33	0.03	0.34	0.03
R070Q03	0.93	0.03	0.90	0.03
R076Q05	0.05	0.03	0.07	0.03
R077Q02	0.83	0.04	0.84	0.04
R083Q02	1.33	0.03	1.24	0.04
R083Q03	1.13	0.03	1.15	0.04
R088Q03	1.00 (1.39)	0.066 (0.10)	1.00 (1.35)	0.08 (0.13)
R091Q05	2.18	0.06	2.00	0.07
R100Q04	0.31	0.03	0.31	0.03
R104Q01	1.31	0.03	1.56	0.07
R104Q02	-0.13	0.03	-0.10	0.03
R104Q05	-0.07 (1.73)	0.04 (0.15)	-0.11 (1.91)	0.03 (0.17)
R104Q06	0.93	0.03	0.89	0.03
R110Q04	1.30	0.04	1.45	0.06
R110Q05	1.35	0.04	1.44	0.05
R111Q04	1.01	0.03	0.98	0.04
R119Q06	1.36	0.04	1.31	0.04
R122Q03T	0.56 (0.84)	0.05 (0.09)	0.60 (0.89)	0.05 (0.08)
R216Q04	-0.05	0.04	-0.04	0.04
R219Q01E	1.38	0.03	1.29	0.04
R220Q01	0.34	0.03	0.33	0.03
R225Q03	1.78	0.05	1.82	0.06
R225Q04	1.12	0.03	1.16	0.04
R227Q02T	0.48 (0.91)	0.04 (0.07)	0.46 (0.87)	0.04 (0.07)
R227Q06	1.20	0.03	1.34	0.06
R234Q01	1.37	0.04	1.42	0.05
R234Q02	-0.76	0.03	-0.72	0.03
R237Q01	0.75	0.03	0.83	0.04
R238Q01	0.40	0.03	0.40	0.03
R239Q02	0.35	0.03	0.34	0.03
R245Q01	0.84	0.03	0.79	0.03
R246Q01	0.92	0.03	1.05	0.05
R246Q02	-0.21	0.03	-0.20	0.03

Table 2. Comparisons of parameter estimates – loadings (DIC is Deviance information criterion, PD is effective number of parameters)			
<i>Question</i>	Equal Loadings (Rasch model =1)	Unequal Loadings	Unequal Loadings standard errors
R040Q02	1	0.55	0.05
R040Q03A	1	0.69	0.05
R070Q02	1	0.73	0.05
R070Q03	1	0.60	0.05
R076Q05	1	0.82	0.06
R077Q02	1	0.62	0.05
R083Q02	1	0.43	0.04
R083Q03	1	0.63	0.05
R088Q03	1	0.63	0.06
R091Q05	1	0.32	0.08
R100Q04	1	0.77	0.05
R104Q01	1	1.10	0.08
R104Q02	1	0.35	0.03
R104Q05	1	0.92	0.09
R104Q06	1	0.54	0.04
R110Q04	1	0.90	0.07
R110Q05	1	0.82	0.06
R111Q04	1	0.54	0.05
R119Q06	1	0.56	0.05
R122Q03T	1	0.77	0.06
R216Q04	1	0.70	0.06
R219Q01E	1	0.47	0.05
R220Q01	1	0.53	0.05
R225Q03	1	0.70	0.06
R225Q04	1	0.76	0.05
R227Q02T	1	0.50	0.04
R227Q06	1	0.88	0.08
R234Q01	1	0.77	0.06
R234Q02	1	0.52	0.05
R237Q01	1	0.85	0.05
R238Q01	1	0.68	0.05
R239Q02	1	0.59	0.04
R245Q01	1	0.44	0.04
R246Q01	1	0.95	0.07
R246Q02	1	0.61	0.04
Factor variance (s.e.)	0.423 (0.012)	1.0	
DIC (PD)	89793.4 (5478)	89129.7 (5580)	

Chapter 3. Multilevel multivariate models with mixed response types at 2 levels

1. Introduction

Multivariate models, including those which incorporate a multilevel structure are traditionally confined to responses at the lowest level of the data hierarchy and usually also deal only with Normally distributed responses. One exception to the latter, and implemented in MLwiN, is where the responses are all binary or a mixture of binary and Normal. Browne (2004) discuss such models and gives examples. There are also some examples of the use of Normal responses jointly at levels 1 and 2; Steele et al (2007) model pupil and school level Normal responses in a multiprocess model for evaluating the impact of school resources on student achievement, and Goldstein (1989) fits a model with repeated measures on individuals during growth (level 1) jointly with their adult height (level 2) as the basis for an efficient prediction model. We now describe extensions that allow any of the responses additionally to be ordered or unordered categorical variables. As we will see this is particularly useful when we wish to carry out multiple imputation for missing data, where missingness may occur with continuous or discrete data.

2. Models for mixed multivariate responses at 2 levels

Appendix A sets out the general model and provides a detailed description of the MCMC estimation algorithm. To introduce the models we shall be using we consider the following simple version with Normal responses, written as

$$y_{ij}^{(1)} = X_{1ij}\beta^{(1)} + u_j^{(1)} + e_{ij}^{(1)}$$

$$y_j^{(2)} = X_{2j}\beta^{(2)} + u_j^{(2)}$$

$$e_{ij}^{(1)} \sim MVN(0, \Omega_1), u_j = (u_j^{(1)}, u_j^{(2)})^T, u_j \sim MVN(0, \Omega_2)$$

The superscripts denote the level at which a variable is measured or defined. Here $y_{ij}^{(1)}$ is a p_1 row vector containing the responses that are defined at level 1 for level 1 unit i nested in level 2 unit j . Also, $y_j^{(2)}$ is a p_2 row vector containing the remaining responses that are defined at the higher level. We assume the same set of predictors for each response and X_{1ij} is a $1 \times f_1$ matrix that contains the predictor variables for observation i nested in higher level unit j and $\beta^{(1)}$ is an $f_1 \times p_1$ matrix containing the fixed coefficients. X_{2j} is a $1 \times f_2$ vector that contains predictor variables for higher level unit j and $\beta^{(2)}$ is an $f_2 \times p_2$ matrix containing the fixed coefficients. The link between the level 1 and the level 2 responses is through the level 2 covariance matrix of the level 2 random effects for the level 1 responses, $u_j^{(1)}$ and the level 2 random effects for the level 2 responses, $u_j^{(2)}$, with covariance matrix Ω_2 .

To extend this model to handle binary responses we consider the threshold model that assumes that a response = 1 occurs when the value of an underlying ‘latent Normal’ variable is greater than 0 and a response = 0 when less than 0. The algorithm then has an extra step that draws a random value from this Normal distribution, given the data

and current parameter values. This is extended to the ordered category case by supposing that there are a series of thresholds along the distribution of the latent Normal variable defining the observed categories. For unordered categorical variables we use a ‘maximum indicant’ formulation whereby for a k -category variable we have a corresponding $k-1$ dimensional multivariate Normal distribution and the extra step involves drawing a sample from this. As a result of such steps we are then dealing with a set of multivariate normal responses and model (1) applies. Appendix A gives details for all these cases.

3. Growth data example

Our first example uses the growth data mentioned earlier, analysed by Goldstein (1989). This dataset consists of 108 children with height measured on up to six occasions around the age of 13 together with their adult heights, altogether 436 growth period measurements and 108 adult height measurements. We shall fit a cubic growth curve to the level 1 (within child) measures and a single intercept for the adult height measurement. We will also allow the age slope to vary at level 2 so that each child is allowed to grow at different rates. You might also wish, later, to allow the quadratic and/or cubic polynomial coefficients to vary across children. In fact we may well wish to introduce further covariates such as gender, but for purposes of illustration we shall fit the simplest explanatory model. The data set is ‘growthdata.txt’. The variables in order are: Child (level 2) ID, adult height (cm), height (cm), constant=1, age (years centered on 13), age squared, age cubed.

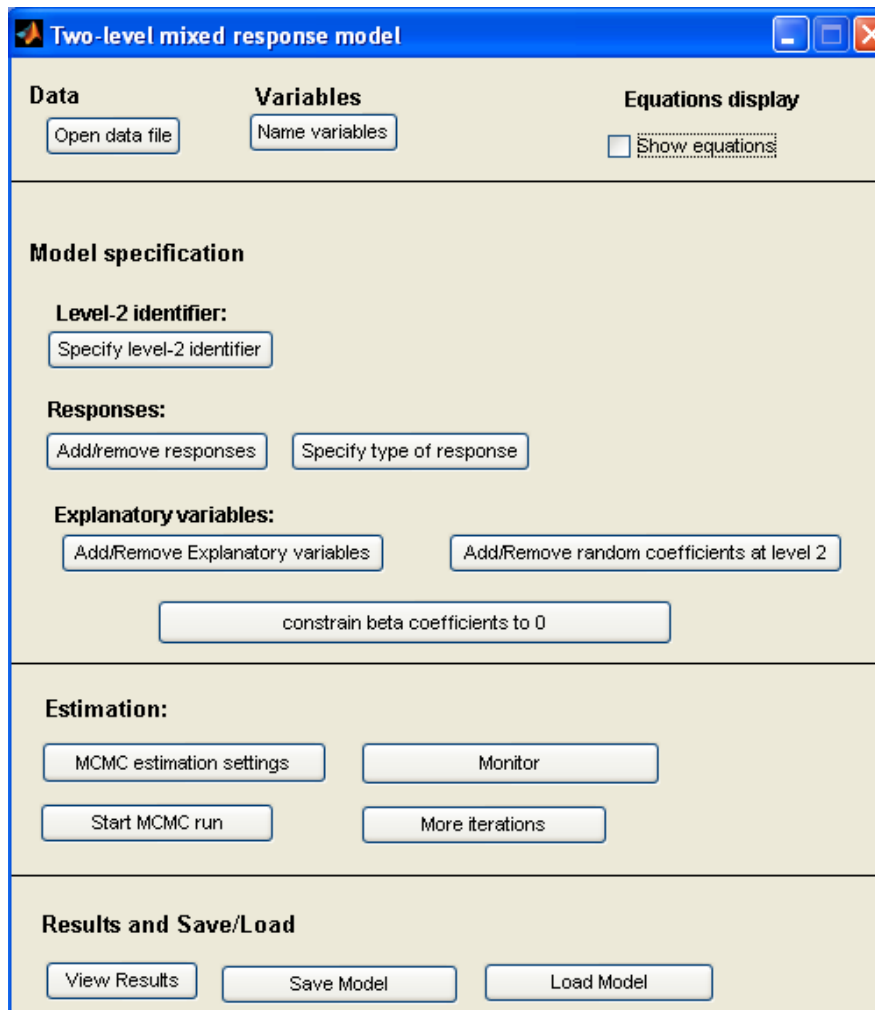
This model can be written as:

$$\begin{aligned}
 y_j^{(2)} &= \gamma_0 + u_{0j}^{(2)} \\
 y_{ij}^{(1)} &= \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + u_{0j}^{(1)} + u_{1j}^{(1)} t_{ij} + e_{ij} \quad (1)
 \end{aligned}$$

$$\begin{pmatrix} u_{0j}^{(1)} \\ u_{1j}^{(1)} \\ u_{0j}^{(2)} \end{pmatrix} \sim MVN(0, \Omega_2), \quad \Omega_2 = \begin{pmatrix} \sigma_{u0}^{(1)^2} & & \\ \sigma_{u01}^{(1,1)} & \sigma_{u1}^{(1)^2} & \\ \sigma_{u00}^{(1,2)} & \sigma_{u10}^{(1,2)} & \sigma_{u0}^{(2)^2} \end{pmatrix}, \quad e_{ij} \sim N(0, \sigma_e^2)$$

where $y_{ij}^{(1)}$ is the i -th measurement around the age of 13 for the j -th child, $y_j^{(2)}$ is the adult height of the j -th child and t_{ij} is age.

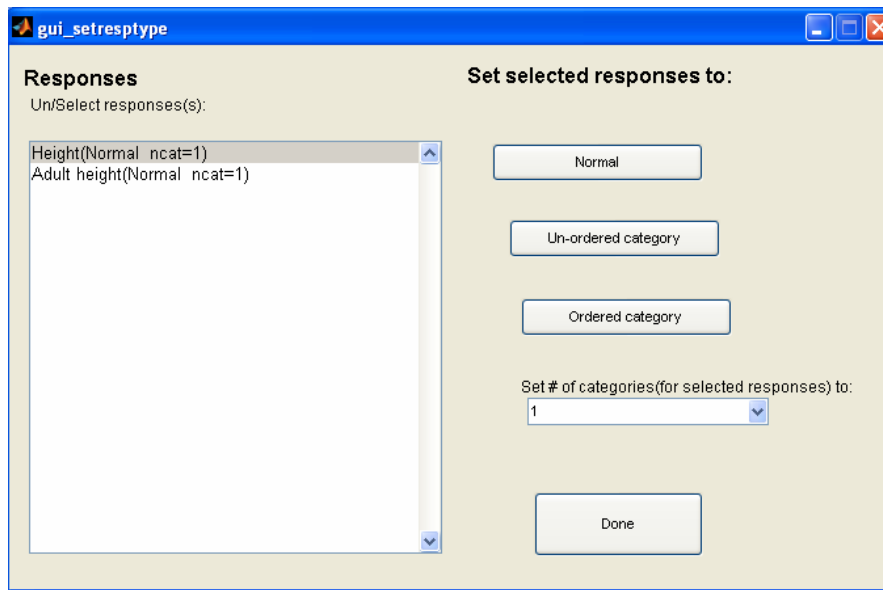
To start the program click on mixed-responses.exe and you will obtain the following screen



Open the data file 'growthdata.txt' and name the variables in the order given above.

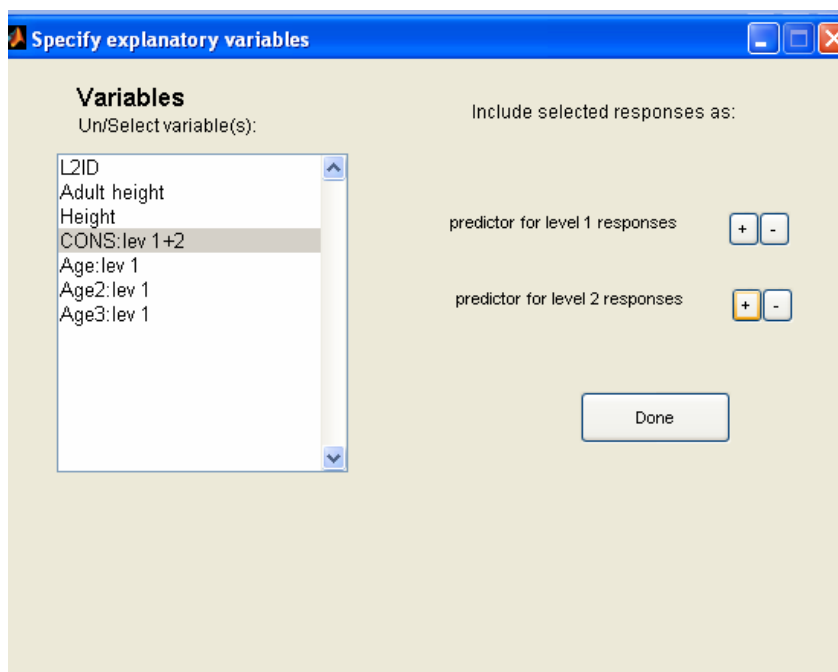
You can also set a missing value code for the data – the default value is $-9.999-999.9 \times 10^{29}$ (code as -9.999e+29) and this is the value used in our data set.

Select the level 2 identifier and add the adult height and height as explanatory variables. When you click on **type of response** you will see the following screen:



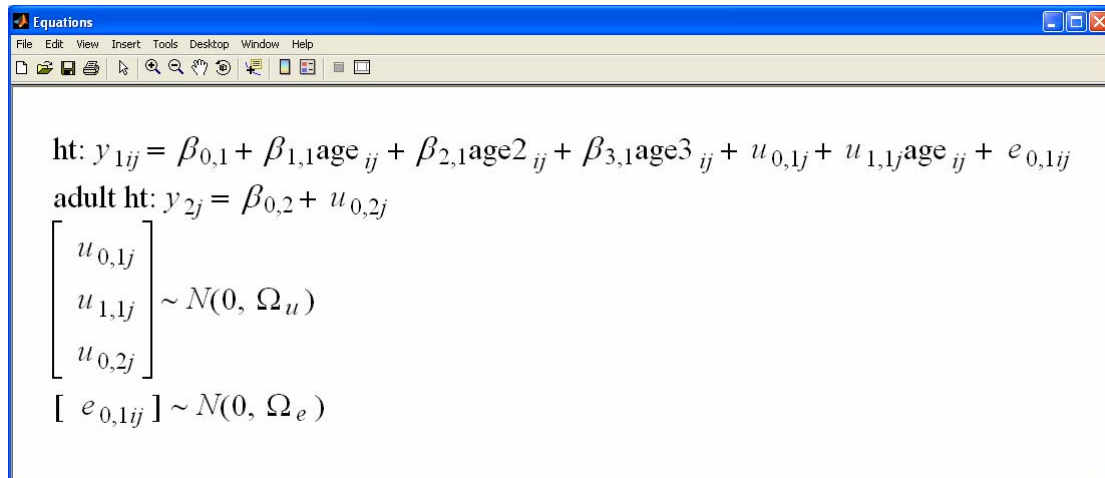
On the right hand side you can specify the type and if this is unordered or ordered you will also need to specify the number of categories. Note that the categories should be integer numbered starting from 0. In the present case both responses are Normal which is the default.

Now click on **Add/remove explanatory variables**, select the constant, age, age squared and age cubed as predictors for the level 1 response and the constant for the level 2 response. You will see a screen rather like the following when you have completed the selection:



The extension, lev 1 or lev 1+2, indicates the level for each predictor. Note that the software detects which are level 1 and level 2 responses by checking whether the variable varies within a level 2 unit – if it does then it is classified as a level 1 response. Now click on **Add/remove random coefficients at level 2**. Add the

intercept (cons) and age. The intercept is automatically included for level two responses. The equations screen will look similar to the following:



ht: $y_{1ij} = \beta_{0,1} + \beta_{1,1}age_{ij} + \beta_{2,1}age^2_{ij} + \beta_{3,1}age^3_{ij} + u_{0,1j} + u_{1,1j}age_{ij} + e_{0,1ij}$
 adult ht: $y_{2j} = \beta_{0,2} + u_{0,2j}$

$$\begin{bmatrix} u_{0,1j} \\ u_{1,1j} \\ u_{0,2j} \end{bmatrix} \sim N(0, \Omega_u)$$

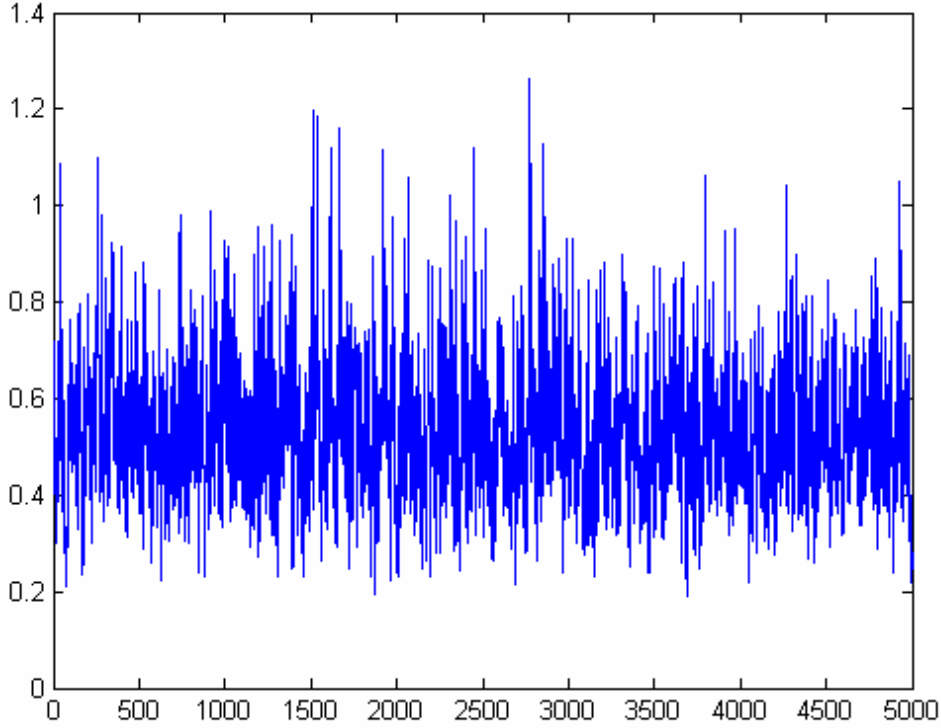
$$[e_{0,1ij}] \sim N(0, \Omega_e)$$

You are also able to constrain any of the fixed (**beta**) coefficients to zero. This may be useful where there are several responses at a level and you do not wish the same predictors to appear in each line of the model. It is suggested that you have a burn in of 100 and monitor 1000.

The results from fitting this model with a burn in of 500 and 5000 iterations are as follows:

Table 1. Two level growth model.		
Coefficient	Estimate	S.E.
Level 1 model intercept	153.05	0.69
Age (about age 13.0)	7.07	0.16
Age-squared	0.294	0.054
Age-cubed	-0.208	0.029
Level 2 model intercept	174.70	0.80
Level 2 covariance matrix		
$\begin{pmatrix} 55.77 & 1.29 & 50.01 \\ 1.30 & 0.53 & 1.24 \\ 50.01 & 1.24 & 69.42 \end{pmatrix}$		
Level 1 variance	3.21	

The chains are well behaved and that for the slope variance (at level 2) is as follows:



The 95% interval for the slope variance is from 0.30 to 0.86. The average height at age 13.0 years is 153.1 (standard error 0.69) and the average adult height is 174.7 (standard error 0.80).

For purposes of predicting the adult height of a child for whom we have any set of growth measurements, we require a prediction formula that we can derive from the parameters of our model. Thus, for example, if we have two growth measurements we will have a linear prediction of the form

$$\hat{y}_j = \gamma_0 + \alpha_1 \tilde{y}_{1j} + \alpha_2 \tilde{y}_{2j} \quad (2)$$

where, from (1)

$$\tilde{y}_{ij} = y_{ij}^{(1)} - (\beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2), \quad i = 1, 2$$

is the ‘raw’ residual for each measurement. The parameters of the prediction equation (2) can be derived from the covariance matrix of the response and predictors, namely

$$\begin{pmatrix} \tilde{y}_{1j} \\ \tilde{y}_{2j} \\ \tilde{y}_j^{(2)} \end{pmatrix} = MVN(0, \Omega), \quad \Omega = \begin{pmatrix} \sigma_{u0}^{(1)2} + \sigma_{u1}^{(1)2} t_{1j}^2 + 2\sigma_{u01}^{(0,1)} t_{1j} + \sigma_e^2 & & \\ \sigma_{u0}^{(1)2} + \sigma_{u01}^{(0,1)} t_{1j} t_{2j} & \sigma_{u0}^{(1)2} + \sigma_{u1}^{(1)2} t_{2j}^2 + 2\sigma_{u01}^{(0,1)} t_{2j} + \sigma_e^2 & \\ \sigma_{u00}^{(1,2)} + \sigma_{u10}^{(1,2)} t_{1j} & \sigma_{u00}^{(1,2)} + \sigma_{u10}^{(1,2)} t_{2j} & \sigma_{u0}^{(2)2} \end{pmatrix}$$

$$\tilde{y}_j^{(2)} = y_j^{(2)} - \gamma_0$$

So that we have

$$\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} \sigma_{u0}^{(1)2} + \sigma_{u1}^{(1)2} t_{1j}^2 + 2\sigma_{u01}^{(0,1)} t_{1j} + \sigma_e^2 & \\ \sigma_{u0}^{(1)2} + \sigma_{u01}^{(0,1)} t_{1j} t_{2j} & \sigma_{u0}^{(1)2} + \sigma_{u1}^{(1)2} t_{2j}^2 + 2\sigma_{u01}^{(0,1)} t_{2j} + \sigma_e^2 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{u00}^{(1,2)} + \sigma_{u10}^{(1,2)} t_{1j} \\ \sigma_{u00}^{(1,2)} + \sigma_{u10}^{(1,2)} t_{2j} \end{pmatrix}$$

so that a prediction, and confidence interval, can readily be computed for any set of measurements.

4. Class size data example

Our second example is concerned with multiple imputation for missing data using a study of the effect of class size on educational achievement. The data set is the same as used for studying measurement errors (see measurement error training materials). Briefly, a cohort of pupils was followed from entry to reception class until the end of the school year, with assessments at the start and end. The response variable is a normalised maths score (end of reception year) *postmaths*. The 5 explanatory variables are: *constant* (=1), *regcls-30* (regular class size centered at 30), normalised pretest maths *pre-maths*, gender *gend*, and free school meals eligibility *fsmn*. For present purposes we have selected a sample with 930 students and 52 classes, with no missing data so that we can introduce (randomly) missing data and compare results.

Before describing our analysis we briefly outline a general procedure for handling missing data known as (random) multiple imputation. Details and examples of how this works can be found by going to the web site www.missingdata.org.uk.

5. Multiple imputation

Suppose we have a model with a single, Normal, response, y and a single, Normal predictor, x . The model of interest is, say,

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad (3)$$

And we may have missing data in both x and y . We now set up an ‘imputation’ model that has all the variables as responses with just an intercept, i.e.

$$\begin{aligned} y_i &= \alpha_1 + e_{1i} \\ x_i &= \alpha_2 + e_{2i} \end{aligned} \quad (4)$$
$$\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \sim N(0, \Omega), \quad \Omega = \begin{pmatrix} \sigma_1^2 & \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

If we fit this model we will obtain intercept estimates and also estimates for the residuals \hat{e}_{1i} , \hat{e}_{2i} . If any of the y , x , are missing then we likewise obtain residual estimates where there is missing data (based upon the model parameter estimates and the observed non-missing data) that allows us to insert a predicted value for the missing value. We also have an estimate of the standard error of the predicted residual and this allows us to randomly sample a value from the (posterior) *distribution* of the residual. Random imputation does this for each missing value so producing a ‘completed’ data set. The procedure is then repeated, say, 5 times (although more may be needed in many circumstances especially when fitting a multilevel model) and the original model of interest (3) fitted this number of times yielding 5 sets of parameters. These are then combined as follows to provide the final estimates.

Let the estimate of any chosen parameter, for example β_1 , from the k th dataset be Q_k . Then, we average these to give

$$Q_{MI} = (1/K) \sum_{k=1}^K Q_k$$

We define the within imputation and between imputation components of variance by

$$\sigma_w^2 = (1/K) \sum_{k=1}^K \sigma_k^2 \quad \sigma_b^2 = (1/(K-1)) \sum_{k=1}^K (Q_k - Q_{MI})^2$$

and the variance estimate of Q_{MI} is then

$$(1 + 1/K) \sigma_b^2 + \sigma_w^2$$

and the square root of this is the standard error. This is extended straightforwardly to handle multiple parameters and covariances. The extension to a multilevel model simply involves adding random effects at higher levels into (3) and (4).

A computationally convenient way to sample the residuals is within an MCMC estimation where every n -th iteration we choose the residual estimates as a sample for a single completed data set. The value of n should be large enough (say 500) to ensure independent samples.

All of the above assumes, as in (4), multivariate Normality. However, many predictor (and response) variables are binary, ordered or nominal and treating these as Normal can lead to biases. This is where the ‘latent Normal’ variable approach is used as described in the introduction. To carry out the imputations we first estimate a Normal residual where the response is missing and then the imputation step is then based upon first imputing for the Normal distributions and then converting to the equivalent category (see Appendix A).

6. Class size data example

The following analyses are described to illustrate the multiple imputation procedure. You can explore the data set for yourself. It is in the file ‘classsize_impure’. The variables are: Classroom ID, constant=1, post-test maths score, pretest maths score, gender (girl =1), free school meals eligibility (yes =1), class size (measured about 30), We have chosen a subset of the original data set which has 930 student records and 52 classes. The original had 4570 complete student records and 246 classrooms (Blatchford et al., 2002).

The first analysis is for the complete data set. We present the results of the model of interest (3). We then randomly miss 20% of the class size data, 20% of the FSM data and 20% of the pre-test data, so that we have about 50% of our records with some missing data. We then carry out our multiple imputation procedure where all the variables are first modelled as responses. Class size is at level 2 and we have a mixture of Normal and binary responses. For present purposes it is suggested that you might like to choose just 1 or 2 imputed dataset to analyse, depending on the time available. In addition we look at the standard procedure of deleting all records with any missing data – the listwise deletion procedure.

Table 2 shows the results from fitting the 2-level model of interest to the full data set excluding records with missing data.

Table 2. No missing data. Response is Post Maths test. Maximum likelihood estimates.		
variable	estimate	Standard error
intercept	0.097	
Class size	-0.021	0.016
Pre-maths	0.560	0.025
Gender	-0.047	0.044
FSM	-0.186	0.062
Level 2 variance	0.122	0.030
Level 1 variance	0.428	0.020

Table 3 shows the results of carrying out the full multiple imputation procedure where we have randomly deleted values in the three predictor variables so that half the records have some missing data. Five completed data sets were imputed using the routines and then analysed and combined within MLwiN.

Table 3. Class size + FSM + pre-maths missing. 5 imputations		
variable	estimate	Standard error
intercept	0.087	
Class size	-0.026	0.018
Pre-maths	0.548	0.027
Gender	-0.056	0.046
FSM	-0.208	0.082
Level 2 var	0.114	0.029
Level 1 var	0.432	0.022
Missing data in 50% of records		

One of the issues in multiple imputation is the number of imputed data sets that should be used. For 2-level data five is often too small and as many as 50 may sometimes be needed to obtain accurate estimates, especially of standard errors. Table 4 shows the results from using 50 imputed data sets.

Table 4. Class size + FSM + pre-maths missing. 50 imputations		
variable	estimate	Standard error
intercept	0.106	
Class size	-0.021	0.020
Pre-maths	0.550	0.028
Gender	-0.055	0.046
FSM	-0.211	0.079
Level 2 var	0.119	0.031
Level 1 var	0.432	0.023
Missing data in 50% of records		

We note that the estimates in Table 4 are closer to those for the full data set and the standard errors somewhat higher. This illustrates the importance of having sufficient imputed data sets to obtain stable estimates.

We now look at the actual estimates obtained from running the multivariate model with every variable (except the constant term) as a response fitting just the intercept as a predictor. First, in Table 5, we show the results for the full dataset with no missing data.

Table 5. Complete data multivariate responses model.	
Variable	Intercept (s.e.)
Post maths	0.1341 (0.0699)
Pre Maths	0.0188 (0.0711)
Gender	0.0687 (0.0465)
FSM	-1.0643 (0.1286)
Class size (-30)	-4.2670 (0.5444)
Level 1 covariance matrix	
$\begin{pmatrix} 0.6955 & 0.4460 & -0.0964 & -0.1982 \\ 0.4460 & 0.7702 & -0.1230 & -0.1939 \\ -0.0964 & -0.1230 & 1.0000 & 0.0067 \\ -0.1982 & -0.1939 & 0.0067 & 1.0000 \end{pmatrix}$	
Level 2 covariance matrix	
$\begin{pmatrix} 0.2140 & 0.0974 & -0.0057 & -0.0938 & -0.1263 \\ 0.0974 & 0.1997 & 0.0219 & -0.1667 & 0.1983 \\ -0.0057 & 0.0219 & 0.0210 & -0.0322 & 0.0407 \\ -0.0938 & -0.1667 & -0.0322 & 0.6169 & -0.3327 \\ -0.1263 & 0.1983 & 0.0407 & -0.3327 & 15.4836 \end{pmatrix}$	

Table 6 fits the dataset with the missing responses.

Table 6. Missing 50% data multivariate responses model.	
Variable	Intercept (s.e.)
Post maths	0.1336 (0.0708)
Pre Maths	0.0321 (0.0713)
Gender	0.0734 (0.0474)
FSM	-1.0898 (0.1293)
Class size (-30)	-4.0494 (0.5968)
Level 1 covariance matrix	
$\begin{pmatrix} 0.6918 & 0.4440 & -0.0957 & -0.1956 \\ 0.4440 & 0.7836 & -0.1205 & -0.1742 \\ -0.0957 & -0.1205 & 1.0000 & -0.0119 \\ -0.1956 & -0.1742 & -0.0119 & 1.0000 \end{pmatrix}$	
Level 2 covariance matrix	
$\begin{pmatrix} 0.2147 & 0.1046 & -0.0057 & -0.0597 & -0.1930 \\ 0.1046 & 0.2141 & 0.0185 & -0.1404 & 0.0965 \\ -0.0057 & 0.0185 & 0.0242 & -0.0423 & 0.0151 \\ -0.0597 & -0.1404 & -0.0423 & 0.6005 & 0.0109 \\ -0.1930 & 0.0965 & 0.0151 & 0.0109 & 14.7433 \end{pmatrix}$	

Notice how all the standard errors have increased, especially for class size. The estimates themselves have changed little. Also note that at level 1 the variances for the two binary variables are fixed at 0. Finally in Table 7 we fit the same model but deleting all the records with any missing data, i.e. about half of them.

Table 7. Missing 50% data multivariate responses model with listwise deletion	
Variable	Intercept (s.e.)
Post maths	0.1017 (0.0881)
Pre Maths	0.0109 (0.0878)
Gender	0.0959 (0.0738)
FSM	-1.1241 (0.1588)
Class size (-30)	-4.0298 (0.6025)
Level 1 covariance matrix	
$\begin{pmatrix} 0.7197 & 0.4292 & -0.0863 & -0.2224 \\ 0.4292 & 0.7163 & -0.1065 & -0.1759 \\ -0.0863 & -0.1065 & 1.0000 & -0.0076 \\ -0.2224 & -0.1759 & -0.0076 & 1.0000 \end{pmatrix}$	
Level 2 covariance matrix	
$\begin{pmatrix} 0.2364 & 0.1585 & -0.0199 & -0.0194 & -0.1311 \\ 0.1585 & 0.2351 & -0.0292 & -0.0920 & 0.1334 \\ -0.0199 & -0.0292 & 0.0682 & -0.0319 & -0.1008 \\ -0.0194 & -0.0920 & -0.0319 & 0.7383 & 0.1142 \\ -0.1311 & 0.1334 & -0.1008 & 0.1142 & 14.7822 \end{pmatrix}$	

We see the large increase in standard errors here for the fixed coefficient intercepts.

7. Conclusions

We have seen how the ability to model variables simultaneously at more than one level leads to efficient predictions and flexible ways of handling missing data. You may like to extend the growth modelling by fitting higher order polynomial or fractional polynomial terms. In the class size case you may like to try deleting data informatively, for example selectively choosing more low scores for the pre-test and then seeing how much bias you can recover over a listwise deletion procedure, utilising the strong correlation between the post and pre test scores.

Appendix A3. An MCMC algorithm for estimating multivariate mixed response types at 2 levels

1. The model

The model structure we consider, for a 2 level model, is as follows

$$y_{ij}^{(1)} = X_{1ij}\beta^{(1)} + Z_{1ij}u_j^{(1)} + e_{ij}^{(1)}$$

$$y_j^{(2)} = X_{2j}\beta^{(2)} + Z_{2j}u_j^{(2)}$$

$$e_{ij}^{(1)} \sim MVN(0, \Omega_1), u_j = (u_j^{(1)}, u_j^{(2)})^T, u_j \sim MVN(0, \Omega_2)$$

The superscripts denote the level at which a variable is measured or defined. Here $y_{ij}^{(1)}$ is a p_1 row vector containing the (latent or actual) normal responses that are defined at level 1 for level 1 unit (observation) i nested in level 2 unit j . Also, $y_j^{(2)}$ is a p_2 row vector containing the remaining responses that are defined at the higher level. We assume the same set of predictors for each response and X_{1ij} is a $1 \times f_1$ matrix that contains the predictor variables for observation i nested in higher level unit j and $\beta^{(1)}$ is an $f_1 \times p_1$ matrix containing the fixed coefficients. Similarly Z_{1ij} is a $1 \times q_1$ matrix that contains predictor variables related to q_1 random effects for observation i nested in higher level unit j and $u_j^{(1)}$ is an $q_1 \times p_1$ matrix containing the random effects at level 2 for the level 1 responses. In the present paper we shall consider only the variance components case where $q_1=1$, but extensions to the general case are straightforward. Correspondingly, Z_{2j} is a $q_2 \times p_2$ matrix for the level 2 random effects for the level 2 responses. For the level 1 residuals $e_{ij}^{(1)}$ is a p_1 row vector (calculated by subtraction). X_{2j} is a $1 \times f_2$ vector that contains predictor variables for higher level unit j and $\beta^{(2)}$ is an $f_2 \times p_2$ matrix containing the fixed coefficients. The $u_j^{(2)}$ is an $q_2 \times p_2$ matrix of level 2 residuals (calculated by subtraction) and are correlated with the level 2 residuals for the level 1 responses. In this paper we assume $q_2=1$.

The first steps in the MCMC algorithm are concerned with how to generate the Normally distributed responses given the actual responses that may be binary, ordered, or unordered categorical. We will focus on the level 1 responses and consider each type of response in turn. The level 2 responses are generated via very similar steps.

We wish to sample Normal responses from any binary, ordered or general multicategory responses. Binary can be treated either as multicategory (*unordered*) with 2 categories or ordered with two categories. In the latter case we are effectively modelling the proportion of ‘1’ responses and in the former the proportion of ‘0’ responses. The latter is typically more computationally efficient.

2. Multicategory (unordered) responses:

We assume a ‘maximum indicant’ model (Aitchison and Bennet, 1970) defined as follows:

Consider the multinomial vector with p categories, where the response, y is (0,1) in each category. That is, we expand the actual response for level 1 unit i , (a categorical variable with values from 1 to p) into p (0,1) variables only one of which is 1.

Thus $y_{hi} = 1$ if response is in category h for individual i , 0 otherwise where h indexes the response. For each y_{hi} we assume an underlying latent variable v_{hi} exists and that we have the following model, where for now we omit the level 2 random effects:

$$v_{hi} = X_{1hi}\beta_{1h} + e_{hi}, \quad e_i \sim MVN(0, \Sigma)$$

Σ is a $p \times p$ correlation matrix, e_i mutually independent vectors (A.1)

$$X_{1hi} \text{ is } (1 \times s), \quad \beta_{1h} \text{ is } (s \times 1), \quad e_i \text{ is } (p \times 1), \quad \beta_1 = \{\beta_{11}^T, \dots, \beta_{1p}^T\}^T, \text{ is } (ps \times 1)$$

For identifiability purposes we will model only the first $p-1$ categories and assume that Σ is diagonal with variances equal to 1.

Let Y_{1hi}^* be the set of other responses, that is current residuals, adjusted for X_1 predictors (common to all responses) and (possible) random effects at higher levels. When sampling the v_{hi} we condition on this set so that (A.1) becomes

$$v_{hi} = X_{1hi}\beta_{1h} + Y_{1hi}^*\beta_{2h} + e_{hi} \quad (A.2)$$

Thus, if Ω_1 is the current residual covariance matrix for the full set of model responses, we write

$$\Omega_1 = \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix}, \quad \text{where } \Sigma_1 \text{ is the residual covariance matrix for the } Y_1^* \text{ and}$$

$$\Sigma_2 = I_{p-1}. \text{ We therefore have } \beta_2 = \Sigma_{12}\Sigma_1^{-1}.$$

While the same set of model predictors X_1 applies to each category, the coefficients in general are specific to each category. We therefore have

$$X_{1hi} = X_{1i}, \quad v_i = (X_{1i}^*\beta_1) + e_i, \quad v_i \text{ is } ((p-1) \times 1),$$

$$X_{1i}^* = I_{p-1} \otimes X_{1i} \text{ is } ((p-1) \times (p-1)s) \quad (A.3)$$

The maximum indicant model states that we observe category h for individual i iff $v_{hi} > v_{h^*i} \quad \forall h^* \neq h$. Thus the category probabilities are given by

$$\pi_{hi} = pr[X_{1hi}\beta_h + e_{hi} > X_{1hi}\beta_{h^*} + e_{h^*i}] \quad \forall h^* \neq h \quad (A.4)$$

If we now add level 2 random effects (j indexes level 2) (1) becomes $v_{hij} = X_{1hij}\beta_{1i} + z_{ij}u_{hj} + e_{hij}$ where u_{hj} is $(q \times 1)$ and we write $u_j = \{u_{hj}\}^T$ which is a $(q(p-1) \times 1)$ vector with $\Omega_u = \text{cov}(u_j)$. We also now write $z_{ij}^* = I_{p-1} \otimes z_{ij}$ which is $((p-1) \times q(p-1))$ and z_{ij} is $(1 \times q)$.

To sample the latent Normal responses $v_{ij} = \{v_{hij}\}$ we select a sample of $p-1$ values from $N(X_{1i}^*\beta_1 + Y_{1i}^*\beta_2 + z_{ij}^*u_j, \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12}^T)$ and accept this draw to replace the current set of $p-1$ values if and only if the maximum of these $p-1$ values actually occurs in the category where a response variable value of 1 is observed and if this maximum is greater than zero, or if the maximum is less than or equal to zero and a value of 1 is observed in the final category. If not, we select another sample.

3. Ordered responses

Suppose we have an ordered p -category response, ordered categories numbered $1, \dots, p$. We consider the probit link proportional odds model

$$\gamma_h = \int_{-\infty}^{\alpha_h - (X_1\beta_1 + Y_1^*\beta_2 + ZU)} \varphi(t) dt$$

$$\gamma_h = \sum_{g=1}^h \pi_g \quad \text{categories} \quad h = 1, \dots, p-1,$$

Where Y_1^*, β_2 are as before, and the underlying latent Normal variable is given by

$$Y^* = e^* + (X_1\beta_1 + Y_1^*\beta_2 + ZU), \quad e^* \sim N(0, 1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$$

Note that alternatively we could form $Y^* = e^* + (X_1\beta_1 + ZU)$, $e^* \sim N(0,1)$, but this will generally provide less efficient parameter estimates.

We assume that the intercept term is incorporated in the fixed part predictor so that $\alpha_1 = 0$.

We can convert this to a standard Normal model by sampling to obtain as follows (Albert and Chib, 1993).

For a category 1 response we sample from the standard Normal distribution $[-\infty, -(X_1\beta_1 + Y_1^*\beta_2 + ZU)]$

For a category p response we sample from the standard Normal distribution $[\alpha_{p-1} - (X_1\beta_1 + Y_1^*\beta_2 + ZU), \infty]$

For every other category h we sample from the standard Normal distribution $[\alpha_{h-1} - (X_1\beta_1 + Y_1^*\beta_2 + ZU), \alpha_h - (X_1\beta_1 + Y_1^*\beta_2 + ZU)]$

For the $\{\alpha_h\}$, conditional on current values of Y_1^* and other parameters we must select a new α_h ($h > 1$) and use MH sampling for these threshold parameters. Thus, the component of the likelihood associated with the ordered category is given by

$$P_\alpha = \prod_{i=1}^N \prod_{h=1}^p \pi_{\alpha,h}^{w_{i,h}}$$

for given α where $w_{i,h}$ is the observed (0,1) response for individual i in category h , and

$$\pi_h = \int_{\alpha_{h-1} - (X_1\beta_1 + ZU)}^{\alpha_h - (X_1\beta_1 + ZU)} \varphi(t) dt, \quad 1 < h < p \quad (p \geq 3)$$

$$\pi_1 = \int_{-\infty}^{- (X_1\beta_1 + ZU)} \varphi(t) dt,$$

$$\pi_p = \int_{\alpha_{p-1} - (X_1\beta_1 + ZU)}^{\infty} \varphi(t) dt,$$

We select a new set of values α^* (one at a time) using a suitable (Normal) proposal distribution (for example derived adaptively) and set new threshold parameters $= \alpha^*$ with probability $\min(1, P_{\alpha^*} / P_\alpha)$. In addition, the order relationships among the threshold parameters must be satisfied. If the selection results in an element of α that does not satisfy these relationships then that element is left at the current value. The α are sampled first followed by the Y^* .

The above two steps will yield Normally distributed responses, which together with any observed Normal responses produces a multivariate Normal set. Where level 1 responses are missing we sample new responses, omitting detailed subscripts, by

drawing from $MVN(X_2^* \beta_2^* + e_1^* \beta_1^* + z_2^* u_2^*, \Sigma_2 - \Sigma_{21} \Sigma_1^{-1} \Sigma_{12})$ where Σ_2 is the current covariance matrix of residuals for the missing responses, Σ_1 is the covariance matrix of residuals for the observed responses and Σ_{12} is the matrix of covariances between the observed and missing residuals. The $X_2^* \beta_2^*$ and $z_2^* u_2^*$ are the fixed predictor and level 2 residual contribution for the missing responses, $\beta_1^* = \Sigma_1^{-1} \Sigma_{12}$ and e_1^* are the level 1 residuals for the observed responses.

Following the above steps we have a new complete set of, say, p multivariate responses for each level 1 unit with the model

$$y_{hij} = X_{ij} \beta_h + z_{ij} u_{hij} + e_{hij}, \quad e_{ij} \sim MVN_p(0, \Omega_1) \quad (\text{A.5})$$

Where h indexes the response. Having sampled so that we have a set of Normal variables, we now have the following further steps. We consider first a model with only level 1 responses.

4. Sampling the fixed coefficients

To sample β we assume a uniform prior and sample from a Multivariate Normal distribution with mean

$$[\sum_{ij} (I_{(p_1 \times p_1)} \otimes X_{ij})^T \Omega_1^{-1} (I_{(p_1 \times p_1)} \otimes X_{ij})]^{-1} \sum_{ij} (I_{(p_1 \times p_1)} \otimes X_{ij})^T \Omega_1^{-1} \tilde{y}_{ij}^T, \quad \tilde{y}_{ij} = y_{ij} - z_{ij} u_j \text{ and}$$

covariance matrix $[\sum_{ij} (I_{(p_1 \times p_1)} \otimes X_{ij})^T \Omega_1^{-1} (I_{(p_1 \times p_1)} \otimes X_{ij})]^{-1}$ where z_{ij}, u_j are defined

with respect to the complete set of level 1 multivariate responses. That is

$$u_{hij} = (u_{h1j}, u_{h2j}, \dots, u_{hq_j})^T, \quad u_j = (u_{1j}, u_{2j}, \dots, u_{p_1j})^T$$

So that u_j is $q_1 p_1 \times 1$ with the random effects varying fastest.

5. Sampling the random effects

To sample the u_j with prior $N(0, \Omega_2)$ we note that the exponent of the likelihood for

$$\text{the } j\text{-th level 2 unit is } \sum_i (y_{ij} - X_{ij} \beta - z_{ij} u_j)^T \Omega_1^{-1} (y_{ij} - X_{ij} \beta - z_{ij} u_j) + u_j^T \Omega_2^{-1} u_j$$

Thus we sample u_j from the multivariate Normal distribution

$$MVN([\sum_i z_{ij}^T \Omega_1^{-1} z_{ij} + \Omega_2^{-1}]^{-1} [\sum_i z_{ij}^T \Omega_1^{-1} (y_{ij} - X_{ij} \beta)], [\sum_i z_{ij}^T \Omega_1^{-1} z_{ij} + \Omega_2^{-1}]^{-1})$$

6. Sampling the level 1 (multivariate) covariance matrix

For all the categorical responses the level 1 variances are fixed to 1.0, with zero correlations among the categories of each unordered categorical variable, but non-zero correlations between these categories and other categorical and continuous variables. Thus for this set of correlations and for the unconstrained variances we use an MH sampling procedure as follows. We assume uniform priors.

Let $\Omega_{1,lm}$ denote the l, m -th element of the covariance matrix. We update these covariance parameters using a Metropolis step and a Normal random walk proposal as follows.

At iteration t generate $\Omega_{1,lm}^* \sim N(\Omega_{1,lm}^{(t-1)}, \sigma_{plm}^2)$ where σ_{plm}^2 is a proposal distribution variance that has to be set for each covariance and variance. Then form a proposed

new matrix Ω_1^* by replacing the l, m th element of $\Omega_1^{(t-1)}$ by this proposed value unless Ω_1^* is not positive definite in which case set $\Omega_1^{(t)} = \Omega_1^{(t-1)}$. That is set

$\Omega_{1,lm}^{(t)} = \Omega_{1,lm}^*$ with probability $\min[1, p(\Omega_1^* | e_{ij}) / p(\Omega_1^{(t-1)} | e_{ij})]$ and $\Omega_{1,lm}^{(t)} = \Omega_{1,lm}^{(t-1)}$ otherwise. The components of the likelihood ratio are

$$p(\Omega_1^* | e_{ij}) = \prod_{ij} |\Omega_1^*|^{-1/2} \exp(-(e_{ij})^T (\Omega_1^*)^{-1} e_{ij} / 2) \text{ and}$$

$$p(\Omega_1^{(t-1)} | e_{ij}) = \prod_{ij} |\Omega_1^{(t-1)}|^{-1/2} \exp(-(e_{ij})^T (\Omega_1^{(t-1)})^{-1} e_{ij} / 2)$$

An adaptive procedure (Brown, 2004) can be used to select the proposal distribution parameters.

7. Sampling the level 2 covariance matrix

We sample a new level 2 covariance matrix

$$\Omega_2^{-1} \sim \text{Wishart}(v_u, S_u)$$

$$v_u = m + v_p, \quad S_u = \left(\sum_{j=1}^m u_j u_j^T + S_p \right)^{-1}$$

Where m is the number of level 2 units, u_j is the row vector of residuals for the j -th level 2 unit and the prior, $p(\Omega_2^{-1}) \sim \text{Wishart}(v_p, S_p)$, where v_u is the degrees of freedom – the sum of the number of level 2 units and degrees of freedom associated with the prior. One choice is $v_p = -3$, $S_p = 0$ which is equivalent to choosing a uniform prior for Ω_2 .

The level 1 residuals are obtained by subtraction.

8. Responses at both level 1 and level 2

We write the full multivariate model as follows with superscripts indicating the response level. The number of level 1 responses is p_1 and there are p_2 at level 2.

$$y_{hj}^{(1)} = X_{ij}^{(1)} \beta_h^{(1)} + z_{ij}^{(1)} u_{hj}^{(1)} + e_{hij} \tag{A.6}$$

$$y_{hj}^{(2)} = X_j^{(2)} \beta_h^{(2)} + z_j^{(2)} u_{hj}^{(2)}$$

Note that (6) allows complex level 2 variance by specifying several random effects (Goldstein, 2003, Chapter 2), but we shall assume here that $z_j^{(2)}$ is the constant vector =1, i.e. there are $q_2 = p_2$ level 2 random effects for the level 2 responses. The MCMC steps are now as follows.

Step 1: For non-Normal level 1 responses we sample as before.

Step 2: For non-Normal level 2 responses we sample as for level 1 conditioning on all the remaining level 2 responses.

Step 3: For the level 1 covariance matrix we sample using MH as before.

Step 4: For the level 2 covariance matrix we sample in similar fashion to before using the full level 2 covariance matrix if all the level 2 responses are Normal. If any are categorical then, because of constraints on variances and covariances, as in sampling the level 1 covariance matrix, we need to use MH sampling element by element.

The procedure is along the same lines as for the level 1 covariance matrix but now the components of the likelihood ratio for a particular level 2 covariance matrix Ω_2 are as follows:

$$p(\Omega_2^* | u_j^{(2)}) = \prod_{ij} |\Omega_2^*|^{-1/2} \exp(-(u_j^{(2)})^T (\Omega_2^*)^{-1} u_j^{(2)} / 2)$$

$$p(\Omega_2^{(t-1)} | u_j^{(2)}) = \prod_{ij} |\Omega_2^{(t-1)}|^{-1/2} \exp(-(u_j^{(2)})^T (\Omega_2^{(t-1)})^{-1} u_j^{(2)} / 2)$$

For this step, even more than for level 1, it is important to use good starting values for the variance terms. These can be obtained, for example, from univariate 2-level variance component models.

Step 5: The fixed effects for the level 1 responses are estimated, as before, using the multivariate model specified by the first line of (6).

Step 6: The level 2 response fixed effects are estimated using the multivariate (regression) model specified by the second line of (6).

Step 7: The level 2 random effects for the level 2 responses are obtained by subtraction. Where level 2 responses are missing we draw a sample from $MVN(0, \Omega_2)$, where Ω_2 now incorporates level 2 random effects from responses at both levels. We select the random effects corresponding to these missing responses from the drawn sample.

Step 8: For the level 1 response level 2 random effects we sample as before, ignoring the level 2 response residuals.

Where level 2 responses are missing we sample in similar fashion to the case where level 1 responses are missing that is from $MVN(X_2^* \beta_2^* + u_1^* \beta_1^*, \Sigma_2 - \Sigma_{21} \Sigma_1^{-1} \Sigma_{12})$ where Σ_2 is the current covariance matrix of level 2 residuals for the missing responses, Σ_1 is the covariance matrix of level 2 residuals for the non-missing level 2 responses and Σ_{12} is the matrix of covariances between the observed and missing residuals. The $X_2^* \beta_2^*$ is the fixed predictor for the missing responses, $\beta_1^* = \Sigma_1^{-1} \Sigma_{12}$ and u_1^* are the level 2 residuals for the observed responses.

9. Imputing categories

At any cycle of the MCMC algorithm we can sample a set of category responses given the current latent responses Y . For an ordered variable we use the current parameters to sample a residual on the Normal scale and assign it to the appropriate category. For an unordered variable we sample into the category indicated by the maximum from a draw from the associated multivariate Normal distribution.

References

Aitchison, J. and Bennett, J. A. (1970). Polychotomous quantal response by maximum indicant. *Biometrika* **57**: 253-262.

Albert, J. H. and Chibb, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*. **88**: 669-679.

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. J. (2004). MCMC estimation in MLwiN. Version 2.0. London, Institute of Education.

Browne, W., Goldstein, H., Woodhouse, G. and yang, M. (2001). An MCMC algorithm for adjusting for errors in variables in random slopes multilevel models. *Multilevel modelling newsletter* **13**(1): 4-9.

Ecob R & Goldstein H. (1983). Instrumental Variable Methods for the Estimation of Test Score Reliability. *Journal of Education Statistics* **8** (3) 223-241.

Fuller, W. A. (2006). *Measurement Error Models*. New York, Wiley:

Goldstein, H. (1989). Models for Multilevel Response variables with an application to Growth Curves. *Multilevel Analysis of Educational Data*. R. D. Bock. New York, Academic Press: 107-125.

Goldstein, H. (2003). *Multilevel Statistical Models. Third edition*. London, Edward Arnold:

Goldstein, H. and Browne, W. (2005). Multilevel factor analysis models for continuous and discrete data. *Contemporary Psychometrics. A Festschrift to Roderick P. McDonald*. A. Olivares and J. J. McArdle. Mahwah, NJ:, Lawrence Erlbaum.

Goldstein, H., Kounali, D. and Robinson, A. (2007). Modelling measurement errors and category misclassifications in multilevel models. Submitted for publication.

Goldstein, H., W. Browne, et al. (2002). "Partitioning variation in multilevel models." *Understanding Statistics* **1**: 223-232.

Hand, D. (2004). *Measurement theory and Practice*. London, Arnold:

Kounali, D, Robinson, A, Goldstein, H. (2006). The probity of free school meals as a measure of social deprivation. Paper given to BERA, 2006, University of Warwick.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor analysis as a statistical method*. London, Butterworth:

Mathworks (2004). Matlab. <http://www.mathworks.co.uk>.

McDonald, R. P. and Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of mathematical and statistical psychology* **42**: 215-232.

Muthen, B. O. (2002). Beyond SEM: General latent variable modelling. *Behaviormetrika* **29**:: 81-117.

Muthen, L. K. and Muthen, B. O. (1988). *MPLUS, Users Guide*. Los Angeles.

Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2001). GLLAMM: a general class of multilevel models and a STATA program. *Multilevel modelling newsletter* **13**(1): 17-23.

Sampson, R. J., Raudenbush, S. W. and Earls, F. (1997). Neighborhoods and violent crime: a multilevel study of collective efficacy. *Science* **277**, **5328**: 918-924.

Steele, F. and Goldstein, H. (2006). A multilevel factor model for mixed binary and ordinal indicators of womens status. *Sociological methods and research* **35**: 137-153.

Steele, F., Vignoles, A. and Jenkins, A. (2007). The Impact of School Resources on Pupil Attainment: A Multilevel Simultaneous Equation Modelling Approach. *Journal of the Royal Statistical Society, A*. **170**.