

Multilevel Modelling –

W.J Browne and J. Rasbash, Institute of Education, University of London.

1 Introduction

In the social, medical and biological sciences multilevel or hierarchically structured populations are the norm. For example, school education provides a clear case of a system in which individuals are subject to the influences of grouping. Pupils or students learn in classes; classes are taught within schools; and schools may be administered within local authorities or school boards. The **units** in such a system lie at four different levels of a hierarchy. Pupils are assigned to level-1, classes to level-2, schools to level-3 and authorities or boards to level-4. Units at one level are recognised as being grouped or nested within units at the higher level. Other examples of hierarchical populations are people within households, within areas; patients in wards within hospitals or animals within herds within farms. Such hierarchies are often described in terms of **clusters** of level 1 units within each level 2 unit etc, and the term **clustered population** is used.

A common criticism of using statistical models to analyse quantitative data in the social sciences is that these methods place too much attention on the individual, and too little on the social and institutional contexts in which the individuals are located. Multilevel modelling redresses this imbalance by simultaneously modelling processes at all levels of the population hierarchy. By focussing attention on the levels of the hierarchy in the population, multilevel modelling enables the researcher to understand where and how effects are occurring.

Fitting a model which does not recognise the existence of clustering creates serious technical problems. For example, ignoring clustering will generally cause standard errors of regression coefficients to be underestimated. Consider a population where voters are clustered into wards and wards into constituencies. If the data is analysed ignoring the hierarchical structure of the population then the resulting underestimated standard errors might lead to the inference that there was a preference for one party when in fact the difference could be ascribed to chance. Correct standard errors would only be estimated if variation at the ward and constituency level were allowed for in the analysis. Multilevel modelling provides an efficient way of doing this. It also makes it possible to model and investigate the relative sizes and effects of ward and constituency characteristics on electoral behavior, as well as that of individual characteristics such as gender and age.

In this chapter we develop the basic two-level model in section 2. In section 3 we describe the use of contextual variables. In section 4 we describe how repeated measures or growth curves models can be fitted in a multilevel framework. In section 5 we describe multilevel multivariate response models. In section 6 we describe multilevel models with discrete responses. In section 7 we describe how multilevel models can be fitted in situations where the population structure is complex but not necessarily purely hierarchical. In section 8 we describe briefly the various estimation methods available for estimating multilevel models. Section 9 briefly outlines other types of multilevel models that are not covered in previous sections. Section 10 lists software that is currently available for fitting multilevel models. Finally section 11 suggests some further reading and other useful resources.

2 Random Intercept and Random Slope Models

In this section we develop the basic two level random intercept and random slope models. We locate the discussion in the context of an educational example where we have exam scores on pupils nested within schools. However, the ideas apply to any two level structure, patients in hospitals, people in households, people in geographical areas and so on. These basic models for two level structures readily extend to structures with three or more levels, for example pupils within classes within schools, people within families within geographical areas.

In our educational example we have data on 4059 students. The response is the exam score at age sixteen and one of the main explanatory variables is a measure of prior-ability of the students when they entered secondary school. Both these variables are standardised Normal variables and their plot is shown in figure 1.

{Figure 1 near here}

We can write the standard regression model relating exam score to prior ability as :

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \{1\}$$

Which gives us a single prediction line. Once we make this a multilevel model we can have a different prediction line for each school. Two models are possible:

A random intercept model :

Here schools differ in terms of their intercept only, which gives rise to a set of parallel lines.

A random slopes model :

Here schools vary in terms of both their slopes and their intercepts, which gives rise to a set of, potentially, crossing lines.

In the next two sub-sections we deal with these two models in turn.

2.1 A Random intercept model

We can extend equation 1 to represent the random intercepts model :

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + e_{0ij} \quad \{2\}$$

y_{ij} is the exam score for the i th child in the j th school

β_{0j} is the intercept for the j th school

β_1 is the slope coefficient for the prior ability variable

x_{ij} the prior ability value for the i th child in the j th school

e_{0ij} is the departure of the i th child in the j th school from its school's predicted line

The intercept for the j th school (β_{0j}) is expressed as

$$\beta_{0j} = \beta_0 + u_{0j} \quad \{3\}$$

Where β_0 is the average intercept for all the schools in the sample, u_{0j} is a random departure for the j th school. Substituting {3} into {1} we have

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + e_{0ij} \quad \{4\}$$

In the basic multilevel model we assume :

$$u_{0j} \sim N(0, \sigma_{u_0}^2)$$

$$e_{0ij} \sim N(0, \sigma_{e_0}^2)$$

We now have hit one of the key differences between multilevel models and standard multiple regression. This model has two random variables, e_{0ij} , a pupil level random variable and u_{0j} , a school level random variable. Standard multiple regression only ever has one random variable, often called the error term. As multilevel models become more complex they often contain many random variables.

When we estimate this model we estimate four parameters, β_0 and β_1 which are like standard multiple regression coefficients, they give the average prediction line from which the j th school's line is offset by a random departure u_{0j} . These regression coefficients are called **fixed parameters**. We also estimate $\sigma_{u_0}^2$ the variance of the school level intercept departures and $\sigma_{e_0}^2$ the variance of pupils' exam scores around

their school's summary line. The variances of these random departures at the pupil and school levels are known as the **random parameters**.

In this model we have two levels, pupils are the lowest level units, generically the lowest level is called level-1. Pupils are nested within the higher level units schools, generically the higher level is called level-2. If we had a three level nested population structure, for example pupils within classes within schools then pupils would be level-1, classes would be level-2 and schools would be level-3.

The correlation between two students in the same school, which is referred to as the "intra-level-2-unit correlation" is given by

$$\rho = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_{e0}^2},$$

that is the between school variance over the total variance. The higher the value of this correlation, the more similar two students from the same school are, compared to two students picked at random from the population. That is the "clustering" effect of the higher level units, in this case schools, is stronger. As the effect of clustering increases it becomes more important from both technical and substantive considerations to use multilevel techniques.

The technical issue is that standard multiple regression assumes that the observations are independent. Clearly, in the presence of clustering, this assumption is false. This

results in the standard errors of the regression coefficients produced by multiple regression being underestimated, which can lead to incorrect inferences. You will be inferring relationships are more significant than they actually are.

The substantive issue is that in the presence of high amounts of clustering much of the total variability is between higher level units and therefore it becomes important to explore the nature of this variability. Multilevel models provide an excellent framework for exploring differences between higher level units.

The results for a random intercepts model on our example educational data set with 4059 pupils from a sample of 65 schools are shown in table 1.

{table 1 about here}

We see the overall intercept, β_0 , is close to zero, which we would expect since both the exam score and prior ability variables have been Normalised. There is a strong positive association between prior ability and exam score, β_1 . We also see significant between school and between student variability. The intra-school correlation is $0.092/(0.092+0.057) = 0.14$.

2.1.2 Residuals in a random intercepts model

We may wish to explore the school and student random departures known as residuals. In ordinary multiple regression, we can estimate the residuals simply by subtracting the predictions for each individual from the observed values. In

multilevel models with residuals at each of several levels, a more complex procedure is needed. The true values of the level-2 residuals are unknown, but we will often require to obtain estimates of them. We can in fact predict the values of the residuals given the observed data and the estimated parameters of the model (see Goldstein, 1995, Appendix 2.2)

The prediction procedure for residuals in a random intercepts model is as follows.

Suppose that y_{ij} is the observed value for the i th student in the j th school and that \hat{y}_{ij} is the predicted value from the average regression line. Then the *raw residual* for this subject is $r_{ij} = y_{ij} - \hat{y}_{ij}$. The raw residual for the j th school is the mean of these over the students in the school. Write this as r_{+j} . Then the predicted level-2 residual for this school is obtained by multiplying r_{+j} by a factor as follows –

$$\hat{u}_{0j} = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_{e0}^2 / n_j} r_{+j}$$

where n_j is the number of students in this school.

The multiplier in the above formula is always less than or equal to 1 so that the estimated residual is usually smaller in magnitude than the raw residual. We say that the raw residual has been multiplied by a *shrinkage factor* and the estimated residual is sometimes called a shrunken residual. The shrinkage factor will be noticeably less than 1 when σ_{e0}^2 is large compared to σ_{u0}^2 or when n_j is small (or both). In either case we have relatively little information about the school (its students are very variable or few in number) and the raw residual is pulled in towards the population average line.

From here on ‘residual’ will mean shrunken residual. Note that we can now estimate the level-1 residuals simply by the formula

$$\hat{e}_{0ij} = r_{ij} - \hat{u}_{0j}$$

Residuals are typically used for two purposes. The first is diagnostic, for example to check that they are Normally distributed, typically by constructing Normal plots of standardised residuals against their Normal scores or finding outlying units.

The second is to compare units. Figure 2 shows a plot of the estimated school level residuals (\hat{u}_{0j}) from the model in table 1, with their associated confidence intervals for the 65 schools in the example data set. The residuals are ranked on the plot.

Recall that the school level residuals (\hat{u}_{0j}) are the intercept departures of the j th school from the average line for all schools defined by $\hat{\beta}_0 + \hat{\beta}_1 x_{ij}$. Figure 2 shows that there are a group of fifteen or so schools that are significantly above the average line (confidence intervals do not descend through the $y = 0$), another group of fifteen or so schools that are significantly below the average line (confidence intervals do not ascend through the line $y = 0$) and a middle group of thirty or so schools that show no significant difference from the average line.

{Figure 2 near here}

The estimated school level intercept residuals can be used to form sixty-five separate prediction equations one for each school :

$$\hat{y}_{ij} = \hat{\beta}_{0j} + \hat{\beta}_{1j}x_{ij}$$

This prediction equation when applied to each schools data produces the sixty-five parallel lines shown in figure 3

{figure 3 near here}

2.2 A random slopes model

We can extend the random intercept model to allow for the possibility of schools having different slopes by allowing the slope coefficient, β_1 , to vary randomly at the school level. That is

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j}x_{1ij} + e_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 + u_{1j} \end{aligned} \quad \{5\}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) \quad \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix}$$

$$e_{0ij} \sim N(0, \sigma_{e0}^2)$$

u_{0j} and u_{1j} are random departures at the school level from β_0 and β_1 . They allow the j th school's summary line to differ from the average line in both its slope and intercept. u_{0j} and u_{1j} follow a multivariate Normal distribution with mean 0 and covariance matrix Ω_u . In this model we have two random variables at level 2 so Ω_u is a 2 by 2 covariance matrix. The elements of Ω_u are

$\text{var}(u_{0j}) = \sigma_{u0}^2$ (the variation across the school summary lines in their intercepts)

$\text{var}(u_{1j}) = \sigma_{u1}^2$ (the variation across the school summary lines in their slopes)

$\text{cov}(u_{0j}, u_{1j}) = \sigma_{u01}$ (the school level intercept/slope covariance).

Students scores depart from their schools summary line by an amount e_{0ij} .

The results from this model are shown in table 2. The significant value for σ_{u1}^2 is evidence that the slope coefficient of prior-ability varies across schools. In addition there is a significant positive covariance between intercepts and slopes estimated as 0.018, suggesting that schools with higher intercepts tend to have steeper slopes. This will lead to a fanning out pattern when we plot the schools predicted lines.

{table 2 near here}

Figure 4 shows the 65 predicted school lines from the random slope model. We see that schools line cross. So questions such as "which school is better?", now have to be qualified : "which school is better, for students of a given prior-ability score?".

{figure 4 near here}

We can go on and add further pupil level explanatory variables to our model such as gender, socio-economic status or ethnicity, just as we would in multiple regression. As well as estimating an average effect, we can make the coefficients of these variables random at the school level to see if these average effects vary across

schools. All the examples of explanatory variables listed are student or level-1 explanatory variables. In the next section we look at adding level-2 explanatory variables.

3 Contextual effects

Many interesting questions in social science are of the form; how are individuals effected by their social contexts? For example,

Do girls learn more effectively in a girls school or a mixed sex school?

Do low ability pupils fare better when they are educated alongside higher ability pupils or are they discouraged and fare worse?

In this section we extend the random slopes model of equation {5} to address the second question. For each school we calculate the average prior-ability of its pupils, based on these averages the bottom 25% of schools are coded 1(**low**), the middle 50 % coded 2(**mid**) and the top 25% coded 3(**high**). Note: when contextual variables are formed from a function of lower level variables in this way the resulting variables are referred to as compositional variables by some writers.

This categorical school level contextual variable can be included in the model by adding explanatory variables for the **mid** and **high** groups which are contrasted against the reference group, **low**. The results are shown in model B of table 3.

Children attending **mid** and **high** ability schools score 0.084 and 0.231 points more than children attending **low** ability schools. The effects are of borderline statistical significance. This model assumes the contextual effects of school ability are the same across the intake ability spectrum because these contextual effects are modifying the intercept term. For example, the effect of being in a **high** ability school is the same for low ability and high ability pupils. To relax this assumption we need to include the interaction between **prior-ability** and the school ability contextual variables.

This is done in model C where the slope coefficient for **prior-ability** for pupils from **low** intake ability schools is 0.460. For pupils from **mid** ability schools the slope is steeper $0.460+0.149$ and for pupils from **high** ability schools the slope is steeper still $0.460+0.324$. These two interaction terms have explained variability in the slope of **prior-ability** in terms of a school level variable therefore the between school variability of the **prior-ability** slope (σ_{μ}^2) has been substantially reduced (from 0.015 to 0.011). We now have three different linear relationships between **exam score** and **prior-ability** for pupils from **low**, **mid** and **high** ability schools.

{table 3 near here}

The prediction line for **low** ability schools is

$$\hat{\beta}_0 \text{cons} + \hat{\beta}_1 \text{standlrt}_{ij}$$

and the prediction line for the **high** ability schools is

$$\hat{\beta}_0 \text{cons} + \hat{\beta}_1 \text{standlrt}_{ij} + \hat{\beta}_6 \text{high}_j + \hat{\beta}_8 \text{high.standlrt}_{ij}.$$

The difference between these two lines, that is the effect of being in a **high** ability school is

$$\hat{\beta}_6 \text{high}_j + \hat{\beta}_8 \text{high.standlrt}_{ij}$$

The graph of this last function along with its confidence envelope is shown in Figure 5. This graph shows how the effect of pupils being in a **high** ability school as compared to a low ability school changes across the spectrum of pupil prior-ability.

{figure 5 near here}

4. Multilevel repeated measures models

Repeated measures data occur in many application areas, in particular in growth studies, where child or animal growth is measured over time. Repeated measurements also occur in medical applications and clinical trials where health outcomes, for example blood pressure or cell counts are measured at different occasions, often before and after treatment with a particular drug. The important idea is that the measurements are nested within a subject and measure the same quantity but at different times. Note that repeated measurements can occur at a higher level of the hierarchy for example in education at the school level, different cohorts of children may be measured over time and so the repeated measures are made on the higher level units, schools.

There are two possible types of timing patterns for the measurements. Firstly the timings may be systematic, which often happens in clinical trials and other designed experiments. Here each individual is measured at designated time intervals or ages. Secondly the timings may be random which occurs in observational studies. Here for

example a growth study of wild animals may rely on the animals being trapped and measured which will happen at random times. As we generally fit the age/time of observation as a predictor variable both these types of measurements will be fitted by exactly the same method. Note that if the observations occur at designated time points the data could be alternatively fitted as a multivariate response model as described in section 5.

4.1 Fitting repeated measures using a multilevel model

To fit repeated measures data using a multilevel model we consider the individual measurements as the level 1 units and the individuals as level-2 units. Measurements are regarded as exchangeable within an individual. This means that in growth studies it would be as acceptable to observe individuals who shrink over time as individuals who grow and even individuals whose measurement fluctuates. Even though this may not make sense we would generally fit the age or time of measurement as a predictor which now means that instead of the measurements being exchangeable the fluctuations (residuals) from a growth versus age regression line are exchangeable. Even this may not make sense and models that implicitly model correlation between measurements are discussed in section 9.4. We will now illustrate repeated measures modelling in the following example.

4.2 An example

Our example concerns a longitudinal study of a cohort of 407 students from 33 infant schools who entered school in 1982. Each students reading attainment was tested on

up to 6 occasions between 1982 and 1989. For a student we have both their reading score and age at each occasion they were tested. More details of the study are available in Tizard et al. (1988).

The reading score is based on a different test depending on the age of the children. These tests have then been scaled so that the scores are comparable. The test scores have been adjusted so that the mean score for the tests at an age group is equal to the average age at that age group and the variance has also been adjusted. The age variable has then been centred.

So we have a dataset that has three levels. As the data is repeated measures we have individual reading test scores as the lowest level with up to 6 level-1 units in each level-2 unit, the individual. Then the individuals are nested within level-3 units, schools although in our analysis for simplicity this level will be ignored.

Some typical questions we may want to answer are:

How does reading attainment change as students get older?

Does this pattern vary from student to student?

The most basic model we could fit would be a simple variance components model with response y_{ij} being reading score

$$\begin{aligned}y_{ij} &= \beta_0 + u_{0j} + e_{0ij} \\u_{0j} &\sim N(0, \sigma_{u0}^2) \quad \{6\} \\e_{0ij} &\sim N(0, \sigma_{e0}^2)\end{aligned}$$

The estimates for this model are given in the middle column of table 4. In this model we have not adjusted for age of the individuals and so the variance at level 1 is much greater than the variance at level 2. We can now add the age variable as a fixed effect as in model {7}

$$\begin{aligned}
 y_{ij} &= \beta_0 + \beta_1 x_{1ij} + u_{0j} + e_{0ij} \\
 u_{0j} &\sim N(0, \sigma_{u0}^2) \\
 e_{0ij} &\sim N(0, \sigma_{e0}^2)
 \end{aligned}
 \quad \{7\}$$

Here x_{1ij} is the age of individual j at time i . The estimates for model {7} are given in the right hand column of table 4. Here we see that the level 1 variance has shrunk dramatically with the addition of the age parameter. The age parameter estimate is approximately 1 which is an artifact of how the reading variable was created. The likelihood has been dramatically reduced and so this is a significantly better model for the data. We can also see that now the level 2 variance (between individuals) is double the level 1 variance. It is often the case in repeated measures models that the majority of the variation is at level 2 rather than level 1 which is different from most multilevel models. The current model is fitting a simple common regression line for the relationship between reading and age with residuals to allow parallel regression lines for each individual.

{table 4 near here}

We will now increase the complexity of the model by allowing regression lines with different slopes for each individual. The model is a standard random slopes regression model.

$$\begin{aligned}
 y_{ij} &= \beta_{0j} + \beta_{1j}x_{1ij} + e_{ij} \\
 \beta_{0j} &= \beta_0 + u_{0j} \\
 \beta_{1j} &= \beta_1 + u_{1j}
 \end{aligned} \tag{8}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) \quad \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix}$$

$$e_{0ij} \sim N(0, \sigma_{e0}^2)$$

The estimates for model {8} are given in the middle column of table 5. Here we see that allowing the slopes of the individual regression lines to vary has again reduced the level 1 variance. There is significant variation between the slopes of the lines and the likelihood has again been reduced by a large amount suggesting this is a significantly better model than model {7}.

We can elaborate this model further by allowing each regression between reading and age to be a quadratic curve rather than a straight line. The equation for this model is given in equation {9} and the resulting estimates in the right hand column of table 5.

$$\begin{aligned}
y_{ij} &= \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{1ij}^2 + e_{ij} \\
\beta_{0j} &= \beta_0 + u_{0j} \\
\beta_{1j} &= \beta_1 + u_{1j} \\
\beta_{2j} &= \beta_2 + u_{2j} \\
\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{bmatrix} &\sim N(0, \Omega_u) \quad \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & & \\ \sigma_{u01} & \sigma_{u1}^2 & \\ \sigma_{u02} & \sigma_{u12} & \sigma_{u2}^2 \end{bmatrix} \\
e_{0ij} &\sim N(0, \sigma_{e0}^2)
\end{aligned} \tag{9}$$

{table 5 near here}

Here we again see a decrease in likelihood and a reduction of level 1 variance. The age-squared fixed effect estimate is very small but there is some variability around this estimate. We can plot the resulting curves for the individuals as shown in figure 6. Often in growth studies it is common to fit a polynomial relationship between height or weight and age as we have done here for reading ability.

{figure 6 near here}

5 Multivariate models

In previous sections we have considered only a single response variable. In this section we look at models where we wish simultaneously to model several responses as functions of explanatory variables. The ability to do this provides us with tools for tackling a wide range of problems. These problems include missing data and rotation or matrix designs for surveys.

We shall be using data consisting of scores on two components of a science examination taken in 1989 by 1905 students in 73 schools in England to illustrate the multilevel multivariate response model. The examination is the General Certificate of Secondary Education (GCSE) taken at the end of compulsory schooling, normally when students are 16 years of age. The first component is a traditional written question paper (marked out of a total score of 160), and the second consists of coursework (marked out of a total score of 108). This coursework includes projects undertaken during the course and marked by each student's own teacher but 'moderated', i.e. a sample checked, by external examiners. Interest in these data centres on the relationship between the component marks at both the school and student level, whether there are gender differences in this relationship and whether the variability differs for the two components.

5.1 Specifying a multivariate model

To define a multivariate (in the case of our example a bi-variate) model we treat the individual student as a level 2 unit and the 'within-student' measurements (in our case written and coursework responses) as level 1 units. Each level 1 measurement 'record' has a response, which is either the written paper score or the coursework score. The basic explanatory variables are a set of dummy variables that indicate which response variable is present. Further explanatory variables are defined by multiplying these dummy variables by individual level explanatory variables, for example gender.

Omitting school identification, the data matrix for three students, two of whom have both measurements and the third who has only the written paper score, is displayed in table 6. The first and third students are female (1) and the second is male (0).

{table 6 near here}

The statistical model for the two level model ignoring the school level, is written as follows:

$$y_{ij} = \beta_0 z_{1ij} + \beta_1 z_{2ij} + \beta_2 z_{1ij} x_j + \beta_3 z_{2ij} x_j + u_{0j} z_{1ij} + u_{1j} z_{2ij}$$
$$z_{1ij} = \begin{cases} 1 & \text{if written} \\ 0 & \text{if coursework} \end{cases}, \quad z_{2ij} = 1 - z_{1ij}, \quad x_j = \begin{cases} 1 & \text{if girl} \\ 0 & \text{if boy} \end{cases} \quad \{10\}$$
$$\text{var}(u_{0j}) = \sigma_{u0}^2, \quad \text{var}(u_{1j}) = \sigma_{u1}^2, \quad \text{cov}(u_{0j}, u_{1j}) = \sigma_{u01}$$

There are several interesting features of this model. There is no level 1 variation specified because level 1 exists solely to define the multivariate structure. The level 2 variances and covariance are the (residual) between-student variances. In the case where only the intercept dummy variables are fitted, and in the case where every student has both scores, the model estimates of these parameters become the usual between-student estimates of the variances and covariance. The multilevel estimates are statistically efficient even where some responses are missing.

Thus, the formulation as a 2-level model allows for the efficient estimation of a covariance matrix with missing responses, where the missingness is at random. This means, in particular, that studies can be designed in such a way that not every individual has every measurement, with measurements randomly allocated to individuals. Such 'rotation' or 'matrix' designs are common in many areas and may be efficiently modelled in this way. A more detailed discussion is given by Goldstein (1995, Chapter 4) and the ability to provide estimates of covariance matrices at each

higher level of a data hierarchy enables further models such as multilevel factor analyses to be fitted (see Rowe and Hill, 1997).

A third, school, level can be incorporated and this is specified by inserting a third subscript, k , and two associated random intercept terms :

$$\begin{aligned}
 y_{ijk} &= \beta_0 z_{1ijk} + \beta_1 z_{2ijk} + \beta_2 z_{1ijk} x_{jk} + \beta_3 z_{2ijk} x_{jk} + v_{0k} z_{1ijk} + v_{1k} z_{2ijk} + u_{0jk} z_{1ijk} + u_{1jk} z_{2ijk} \\
 z_{1ijk} &= \begin{cases} 1 & \text{if written} \\ 0 & \text{if coursework} \end{cases}, \quad z_{2ijk} = 1 - z_{1ij}, \quad x_{jk} = \begin{cases} 1 & \text{if girl} \\ 0 & \text{if boy} \end{cases} \\
 \begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} &\sim N(0, \Omega_v) \quad \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & \\ \sigma_{v01} & \sigma_{v1}^2 \end{bmatrix} \\
 \begin{bmatrix} u_{0k} \\ u_{1k} \end{bmatrix} &\sim N(0, \Omega_u) \quad \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix}
 \end{aligned} \tag{11}$$

The 2 by 2 covariance matrix between written and coursework responses is partitioned into a between student component Ω_u and a between school component Ω_v . The mean of the written paper responses for boys is estimated by β_0 , the mean of coursework responses for boys is estimated β_1 , the girl-boy difference for the written paper responses is estimated by β_2 and the girl-boy difference for the coursework responses is estimated by β_3 . The results are shown in table 7. The girls do somewhat worse on the written paper (-2.5) but considerably better than the boys on the coursework component of the examination (6.8). The coursework component also has a larger variance at both student and school level with correlations between coursework and written 0.42 and 0.49 at school and student level respectively. The intra-school correlation is 0.27 for the written paper and 0.29 for the coursework.

{table 7 near here}

This model could be extended further, by allowing the girl-boy differences for each response to vary across schools. Further explanatory variables can be added and their coefficients can vary randomly at either level.

Another, interesting example of multilevel multivariate modelling is given in Duncan et al (1999). Here individuals have two responses, the first is a binary response indicating whether or not an individual smokes. The second response is only present for those individuals who smoke and is the number of cigarettes smoked. This model has two interesting features. Firstly, if we were to model the number smoked as a continuous univariate response, there would be a large spike at zero, which would violate any simple Normal theory and is tricky to model correctly. However, in the multivariate framework, these individuals are properly included by the first binary response. Secondly, the covariance between the two responses at higher levels can be very informative. In Duncan et al the individuals were nested within neighbourhoods. A positive covariance at the neighbourhood level means that smokers who are in an area where the probability of smoking is high will tend to smoke more cigarettes than smokers in an area where the probability of smoking is low. In other words if you are a smoker and a lot people around you are smoking you will smoke greater numbers of cigarettes than if you are not surrounded by smokers.

6 Multilevel models for discrete response data.

6.1 Introduction to modelling discrete responses

In this chapter so far we have assumed that the response variable of interest was a continuous variable. In the last section we showed how to combine several continuous variables via a multivariate model. We also mentioned an example that involved a smoking response that was not regarded as continuous but took values that were either 0 or 1. In this section we will look at multilevel models that fit data like this where the response variable is discrete.

There are two main types of discrete data variables:

Firstly proportional data where the response can take values $0, 1/N, 2/N, \dots, 1$ (with the special case of binary 0,1 data) where N is the size of the population for example the proportion of people who pass a maths test, are in favour of abortion, own red cars etc. Secondly count data where the response can take any positive integer value for example the number of instances of a particular disease, the number of children born on a particular day, the number of cars that travel through a road junction in a 10 minute period etc.

It is common practise to use the Binomial distribution to fit models to proportional data and the Poisson distribution to fit models to count data. In the rest of this section we will concentrate mainly on Binomial models and mention Poisson models only briefly at the end of the section. Other more complex models with discrete responses, such as multicategory data (ordered and unordered) and event history data will be discussed briefly with further references in section 9.5 of this chapter.

6.2 Binomial data

When we consider proportional data we have a response of the form x out of y observations have a particular property, for example 8 out of 10 people questioned eat meat. Here our response is whether people eat meat and any individual person either eats meat or doesn't. So we assume that in the population in general there is an underlying probability π , such that $0 \leq \pi \leq 1$ that a person eats meat. We have taken 10 people at random and found that 8 of them eat meat so an estimate of π based on our 10 people is $8/10 = 0.8$.

Of course we may expect the probability that people eat meat to be different depending on characteristics of the individual. For example if an individual's parents are vegetarian we might expect them to be more likely to be vegetarian, gender and religion may also influence whether a person eats meat.

We could of course fit a model to this dataset assuming the variable to be Normally distributed but this may give us problems. For example say we fitted a model for the probability of eating meat with predictors age and gender and got the following fixed effects estimates :

$$\pi_i = 0.8 + 0.05 * \text{gender}_i + 0.003 * \text{age}_i$$

where $\text{gender}_i = 1$ for male, 0 for female. Then for a male aged 52 the estimated probability of eating meat is $1.006 > 1.0$! A probability that does not lie between 0 and 1 is clearly a problem so we generally model the probabilities with a Binomial distribution rather than a Normal distribution.

6.2.1 Link functions

We have seen that there is a problem with fitting the probabilities directly to the prediction equation that we get from the Normal model. This is because probabilities lie in the range $[0,1]$ whilst the prediction equation can theoretically generate values in the range $(-\infty, \infty)$. Consequently we need to transform via a function the probability, π to a value that lies on the whole real line. These functions are known as link functions and there are 3 common link functions for the Binomial distribution as shown in table 8.

{table 8 near here}

We will use the logit function in the examples that follow but the other two functions can be used in an analogous way. Generally modelling Binomial data with a logit function is known as logistic regression and is not in itself a multilevel technique. To translate logistic regression to multilevel logistic regression is analogous to moving from linear modelling to normal response multilevel modelling. As we will see in the example that follows we have a higher level classification of the data across which we believe the probability response varies.

6.3 An example : A data set on political voting intentions

This example comes from the longitudinal component of the British Election Study (See Heath et al. 1996). It consists of a sub-sample of 800 voters grouped within 110 voting constituencies who were asked how they voted in the 1983 British general election. For our purposes the responses are classified simply as to whether or not the individual voted Conservative. We are interested in establishing what factors influence whether a voter votes conservative while accounting for different underlying probabilities due to the constituencies. We have as predictor variables 4 attitudinal variables that describe the individual voters attitude (on a 21 point scale) to important issues of the day, namely defence, unemployment, taxes and privatisation.

To start our analysis we will fit a simple variance components model as follows

$$\begin{aligned}y_{ij} &\sim \text{Binomial}(1, \pi_{ij}) \\ \text{logit}(\pi_{ij}) &= \beta_0 + u_{0j} \quad \{12\} \\ u_{0j} &\sim N(0, \sigma_{u0}^2)\end{aligned}$$

Unlike the normal models that we have fitted so far the variance at level 1 is constrained by the Binomial assumption and is not estimated. In fact when we use the logistic link function the level 1 variance is approximately $\pi^2/3 = 3.29$. When we fit the above model to the voting dataset we get the estimates given in the middle column of table 9.

{table 9 near here}

Here we see that the variance between constituencies is a lot smaller than the variance within constituencies (3.29).

To interpret the fixed effect intercept value, β_0 we need to use the anti-logit function to transform the value back to the probability scale. Here we have

$[1+\exp(0.248)]^{-1} = 0.438$ which is an estimate of the median proportion of people voting Conservative.

To answer our questions on the effects of the attitudinal variables we need to expand our model as follows:

$$\begin{aligned} y_{ij} &\sim \text{Binomial}(1, \pi_{ij}) \\ \text{logit}(\pi_{ij}) &= \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + u_{0j} \quad \{13\} \\ u_{0j} &\sim N(0, \sigma_{u0}^2) \end{aligned}$$

Here x_{1ij} is opinion on defence, x_{2ij} is opinion on unemployment, x_{3ij} is opinion on taxes and x_{4ij} is opinion on privatisation.

The results of fitting model 13 can be seen in the right hand column of table 9. Here we see that all attitudinal variables have positive predictors and this is due to the way the scales were created. We can conclude that a person is more likely to vote conservative if they have the following views:

- They are in favour of Britain possessing nuclear weapons.
- They prefer more unemployment if it in turn leads to low inflation.

- They are against paying more taxes to pay for more government spending.
- They are in favour of privatisation of public services

All the predictors were centred so we can now interpret the person with 0 for all attitudinal variables as a person with average views. Such a person has an estimated $[1+\exp(0.367)]^{-1} = 0.409$ probability of voting conservative.

To interpret the effect of the predictor variables due to the non-linear relationship between the response and the predictor variables it is best to consider specific cases in isolation. For example a person with a 5 point above average view on possessing nuclear weapons (assuming all other variables are 0) will have an estimated $[1+\exp(0.367 - 5*0.093)]^{-1} = 0.524$ probability of voting conservative. As all the predictors are on the same scale we can compare their parameter estimates and this shows that a persons view on privatisation has greatest effect on their underlying probability of voting conservative.

6.4 Count data

As with the proportional data that we studied in section 6.2, count data has restrictions on the values it takes. Count data must take positive integer values (or zero) and so if we were to fit the count response as a normal response we could get predicted counts that were negative. This is clearly a problem and so instead the Poisson distribution is used. The Poisson distribution has a parameter that represents the rate that events occur in the underlying population. As a link function for the Poisson distribution we

need to convert our prediction equation values that can lie in the range $(-\infty, \infty)$ to a rate that lies in the range $(0, \infty)$ and consequently a log link function is used.

The fitting of Poisson models is essentially similar to binomial models. A health-based example of a multilevel Poisson model involving counts of deaths due to malignant melanoma in the European community is given in Langford, Bentham and McDonald (1998).

7 Non-hierarchical multilevel models

In the models discussed in this chapter so far we have assumed the populations from which data has been drawn are hierarchical. This assumption is not always justified. Two main types of non-hierarchical model will be considered in this section, cross-classified models and multiple membership models.

7.1 Two way cross-classifications – a basic model.

Suppose, we have data on a large number of patients, attending many hospitals and we also know the neighbourhood in which each patient lives. Suppose that we regard patient, neighbourhood and hospital as important sources of variation for the patient level outcome measure we wish to study. Now, typically hospitals will draw patients from many different neighbourhoods and the inhabitants of a neighbourhood will go to many hospitals. No pure hierarchy can be found and patients are said to be contained within a cross-classification of hospitals by neighbourhoods. This is

represented diagrammatically in table 10, for the case of twenty patients contained within a cross-classification of three neighbourhoods by five hospitals :

{Table 10 near here}

There are many other examples of two level cross-classifications: pupils grouped within a cross-classification of primary school by secondary school, see Goldstein and Sammons(1997); patients grouped within a cross classification of primary by secondary health care units; in survey analysis we can have individuals grouped within a cross-classification of interviewers by areas, see O’Muircheartaigh and Campanelli(1999).

The basic two-way, Normal response, cross-classified variance components model can be written as

$$\begin{aligned}
 y_{i(j_1, j_2)} &= (X\beta)_{i(j_1, j_2)} + u_{j_1} + u_{j_2} + e_{i(j_1, j_2)} \\
 u_{j_1} &\sim N(0, \sigma_{u_1}^2) \\
 u_{j_2} &\sim N(0, \sigma_{u_2}^2) \\
 e_{i(j_1, j_2)} &\sim N(0, \sigma_e^2)
 \end{aligned}
 \tag{14}$$

Where $(X\beta)_{i(j_1, j_2)}$ is the linear predictor. Using the example of students within a cross-classification of primary by secondary schools: $y_{i(j_1, j_2)}$ is the exam score at age 16 of the i th student, contained in the cell defined by primary school j_1 and secondary school j_2 . u_{j_1} is the random effect for primary school j_1 , u_{j_2} is the random effect for

secondary school j_2 , and $e_{i(j_1, j_2)}$ is a level 1 residual for the i th student contained in the cell defined by primary school j_1 and secondary school j_2 .

The results for this model fitted to an Educational data set with 3,435 students who attended 148 primary schools and 19 secondary schools in Fife, Scotland are shown in Table 11. Model A fits students within primary schools and ignores secondary school, model B fits students within secondary schools and ignores primary school and model C fits the cross-classification. The response is an attainment score at age 16, the explanatory variable, **vrq**, is a verbal reasoning measure taken at age 11. Notice that in all three models the sum of the variance components is 4.53. When one side of the cross-classification is ignored, the released variance is split between the classification left in the model and the pupil level variance, inflating both estimates. This has the most drastic effect when the primary school hierarchy is ignored, in this case (model B) the inflated estimate of the between secondary school variance is 2.5 times its standard error as opposed to 0.5 times its standard error in the full model.

A variety of more complex patterns of cross-classification are possible. A fuller account is given in Rasbash & Browne(2000).

{table 11 near here}

7.2 Multiple membership models

Where lower level units are influenced by more than one higher level unit from the same classification we have a multiple membership model. For example, if patients

are treated by several nurses, then patients are “multiple members of “ nurses. Each of the nurses treating a patient contributes to the treatment outcome.

A two level multiple membership model can be written as

$$\begin{aligned}
 y_{i\{j\}} &= XB + \sum_{h \in \{j\}} u_h \pi_{ih} + e_{i\{j\}} \\
 u_h &\sim N(0, \sigma_u^2) \\
 e_{i\{j\}} &\sim N(0, \sigma_e^2) \qquad \{15\} \\
 \sum_h \pi_{ih} &= 1
 \end{aligned}$$

Where $\{j\}$ is the full set of level 2 units, in this case nurses. The level 1 units, patients, are indexed uniquely by i and may be a “member of” more than one nurse. The index h uniquely indexes nurses and π_{ih} is a predetermined weight declaring the proportion of membership of patient i to nurse h . For example, if we knew that a quarter of patient i 's treatment was administered by nurse h then a weight of 0.25 might be reasonable. Often we will not have information at this level of detail in which case we assume equal weights.

To clarify this, consider a simple example. Suppose we have four patients (P1,...,P4) treated by up to two of three nurses (n1, n2, n3). The weighted membership matrix, π might look like that shown in table 12.

{table 12 near here}

Here patient 1 was seen by nurses 1 and 3 but not nurse 2 and so on. If we substitute the values of π_{ih} , i and h . from the above table into model {15} we get the following series of equations :

$$y_{1\{j\}} = XB + 0.5u_1 + 0.5u_3 + e_{1\{j\}}$$

$$y_{2\{j\}} = XB + 1u_1 + e_{2\{j\}}$$

$$y_{3\{j\}} = XB + 0.5u_2 + 0.5u_3 + e_{3\{j\}}$$

$$y_{4\{j\}} = XB + 0.5u_1 + 0.5u_2 + e_{4\{j\}}$$

A fuller account of multiple membership models along with an example analysis is given in Rasbash & Browne(2000).

8 Estimation Methods for multilevel models

In this chapter so far we have given estimates for parameters in many multilevel models without giving any formulae for calculating these estimates. Unlike standard regression models, in multilevel modelling there is no simple formula that can be used to directly calculate parameter estimates. Instead there are two possible types of approach that are used, iterative procedures and simulation procedures.

In iterative approaches we start with initial estimates of the parameters of interest. We split the parameters into groups and proceed by estimating a group of parameters conditional on the estimates of the other parameters being true. Through estimating the groups of parameters in turn iterative procedures converge to a point estimate for each parameter and a standard error. With modern computer speeds iterative procedures are generally very quick and give good estimates for most problems.

However for some problems including multilevel discrete response models iterative methods produce approximate estimates which can be biased. Several iterative procedures are described in sections 8.1-8.3.

In simulation procedures the aim is not to converge to a point estimate for each parameter. Instead the aim is to generate a sample of estimates from the distribution of each parameter of interest. Given this sample of values from the distribution of the parameter we can construct many different summary statistics, including plots of its distribution function and accurate confidence intervals. Simulation methods construct the samples of parameter estimates by many different techniques and these will be described briefly in sections 8.4-8.5. Simulation methods often involve generating many thousands of simulated draws and so are consequently often much slower than iterative procedures and require more computer storage. They do however give better estimates for some problems and can be applied to more complicated models where there are at present no equivalent iterative procedures.

We will now briefly describe the methods giving references for the interested reader to find more details. Goldstein (1995) gives more details on most methods described in this section.

8.1 Maximum Likelihood Methods (ML)

The maximum likelihood estimate (MLE) for a parameter θ , is defined as the estimate that maximises the likelihood function, $\text{lh}(\theta; y_1, y_2, \dots, y_n)$ where assuming we have n IID random variables y_i with pdf $f(y_i|\theta)$, then

$$\text{lh}(\theta; y_1, y_2, \dots, y_n) = f(y_1|\theta)f(y_2|\theta)\dots f(y_n|\theta).$$

In linear regression modelling the least squares estimates that are used are maximum likelihood estimates and this property is generally perceived to be a desirable one. In multilevel modelling there are several techniques that produce maximum likelihood estimates and these will be described briefly here.

8.1.1 Iterative generalised least squares (IGLS)

In IGLS (Goldstein 1986) the parameters are split into two groups, the fixed effects form one group and the variances the other. The variances form a general V matrix when the problem is thought of as a multivariate response model. Then assuming that the variances and hence the V matrix are known then the fixed effects can be estimated by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y.$$

Assuming the fixed effects are then known the raw residuals can be calculated and a similar weighted regression for the variance terms can be found based on the cross products of these raw residuals. These two steps generate new estimates for all the parameters and are repeated until consecutive estimates are close enough (based on a desired tolerance).

8.1.2 Fisher Scoring

Longford (1987) developed a procedure that is formally equivalent to the IGLS algorithm based on the 'Fisher scoring' technique.

8.2 Restricted maximum likelihood methods (REML)

In multilevel modelling the variance estimates produced from maximum likelihood based techniques are biased. This is because the maximum likelihood estimates fail to take account of the sampling variation in the fixed effects when producing variance estimates. REML techniques involve modifying the maximum likelihood estimation procedures to account for this additional variation.

8.2.1 RIGLS

RIGLS (Goldstein 1989) is the REML equivalent of the IGLS algorithm.

8.2.2 EM Algorithm

The EM algorithm (Dempster, Laird and Rubin 1977) can be used to fit many models including multilevel models. It consists of two steps, an Expectation step and a maximisation step, hence its name. In a multilevel context unlike the maximum likelihood methods, the higher level residuals are treated as missing data and estimated at each iteration. The resulting log likelihood can be maximised for the variance parameters and then these variance estimates can be used to construct the V

matrix. Then the fixed effects and residuals can be calculated via generalized least squares in a similar way to IGLS. See Bryk and Raudenbush (1992) for more details.

8.3 Quasi-likelihood methods

The above methods can all be used for normal linear response models. For some models that do not have normal responses these procedures cannot be used to find maximum likelihood estimates. Instead quasi-likelihood methods are used that use Taylor series expansions to approximate the non-linear problem by a normal linear problem. There are two types of quasi-likelihood, marginal quasi-likelihood (Goldstein 1991) and penalised quasi-likelihood (Laird 1978) which involves the addition of estimates for the residuals in the linearisation to obtain an improved approximation. See Goldstein (1995 chapter 5) for more information.

8.4 Markov chain Monte Carlo (MCMC) methods

There are two main motivations behind the use of MCMC methods instead of ML methods for multilevel models. Firstly they can be used to produce accurate interval estimates and secondly they can be used to fit models in a Bayesian framework and hence can incorporate prior information into the analysis. Before describing two MCMC methods we will give a very brief description of Bayesian statistics. Both MCMC methodology and Bayesian statistics are huge fields in statistics and we will only touch the tip of the iceberg here. For more details see for example Gilks, Richardson and Spiegelhalter (1996)

8.4.1 Bayesian statistics

The main result in Bayesian statistics is (in words) the following:

Posterior distribution \propto Prior distribution * Likelihood function

In the frequentist (non-Bayesian) world we work with the likelihood function and construct estimates based on this function alone, often ML estimates. Bayesians however view the world differently and believe that every unknown parameter, θ should have a 'prior distribution', $p(\theta)$. This is a distribution that represents our ideas on the value of the parameter prior to collecting the data. Often we have no prior information about θ and so express this by making $p(\theta)$ a 'diffuse' prior distribution, for example $p(\theta) \propto 1$. It could be argued that frequentist statistics is similar to Bayesian statistics with all the unknown parameters having 'diffuse' priors, although both groups interpret many things differently. In Bayesian statistics the posterior distribution is used to make inferences about the parameters of interest rather than the likelihood. The MCMC methods we now discuss aim to estimate the joint posterior distribution of all parameters of interest.

8.4.2 Gibbs sampling

Consider a two level variance components model

$$\begin{aligned}
y_{ij} &= \beta_0 + u_{0j} + e_{0ij} \\
u_{0j} &\sim N(0, \sigma_{u0}^2) \\
e_{0ij} &\sim N(0, \sigma_{e0}^2)
\end{aligned}
\quad \{16\}$$

To evaluate the joint posterior distribution $p(\beta_0, u_0, \sigma_{u0}^2, \sigma_{e0}^2 | y)$ would involve integrating over many parameters and for all but the simplest examples proves intractable. Fortunately an alternative approach is available as although the joint distribution is difficult to simulate from the conditional posterior distributions are not as they have ‘nice forms’. Simulating from these conditional posterior distributions is known as Gibbs sampling and the algorithm for the above problem is as follows:

Split the parameters into four subsets, $\beta_0, u_0, \sigma_{u0}^2$ and σ_{e0}^2 , and for each parameter choose a starting value $\beta_0(0), u_0(0), \sigma_{u0}^2(0)$ and $\sigma_{e0}^2(0)$. The estimates from an iterative estimation procedure could be used as ‘good’ starting values. Then the following four steps will be performed :

- Generate a random value $\beta_0(1)$ from $p(\beta_0 | y, u_0(0), \sigma_{u0}^2(0), \sigma_{e0}^2(0))$.
- Generate $u_0(1)$ from $p(u_0 | y, \beta_0(1), \sigma_{u0}^2(0), \sigma_{e0}^2(0))$.
- Generate $\sigma_{u0}^2(1)$ from $p(\sigma_{u0}^2 | y, \beta_0(1), u_0(1), \sigma_{e0}^2(0))$.
- Generate $\sigma_{e0}^2(1)$ from $p(\sigma_{e0}^2 | y, \beta_0(1), u_0(1), \sigma_{u0}^2(1))$.

So having performed these 4 steps we have now updated all the parameters of interest.

This process is now repeated many times using the previously generated set of

parameters to generate the next set. The chain of values generated by this sampling procedure is known as a Markov chain as every new value generated for a parameter depends on its previous values only through the last value generated.

8.4.3 Metropolis-Hastings (MH) sampling

Gibbs sampling works well when the conditional posterior distributions have a form that is easily simulated from. If this is not the case (as for example in multilevel logistic regression models) then a different procedure known as Metropolis Hastings sampling can be used. Often a conditional distribution for a parameter θ is difficult to simulate from but we can still evaluate the value of the distribution function for a specific value θ .

Consequently for such a parameter we generate a potential new estimate from a 'proposal distribution' and then either accept this new value or stick with the current value depending on a condition that ensures this method is equivalent to simulating from the conditional distribution. One special case of MH sampling is random walk Metropolis sampling.

As an example of how this method works the updating procedure for the parameter β_0 at time step t in the Normal variance components model {16} is as follows:

- Draw β_0^* from the proposal distribution $\beta_0(t) \sim N(\beta_0(t-1), \sigma_p^2)$ where σ_p^2 is the proposal distribution variance.

- Define $r_t = p(\beta_0^*, u_0, \sigma_{u_0}^2, \sigma_{e_0}^2 | y) / p(\beta_0(t-1), u_0, \sigma_{u_0}^2, \sigma_{e_0}^2 | y)$ as the posterior ratio and let $a_t = \min(1, r_t)$ be the acceptance probability.
- Accept the proposal $\beta_0(t) = \beta_0^*$ with probability a_t , otherwise let $\beta_0(t) = \beta_0(t-1)$.

As MH sampling often involves rejecting the proposed value and sticking with the current value, it tends to produce chains that have larger auto-correlation and consequently have to be run for longer.

8.4.4 Convergence Issues and run lengths

When running an MCMC procedure there are two important considerations. Firstly as an MCMC procedure starts from arbitrary parameter starting values it takes time before the Markov chain is sampling from the true joint posterior distribution. To counter this when a Markov chain is run it is common practice to throw away the first n estimates to allow the chain to converge to its stationary distribution. This period of n iterations is commonly known as a ‘burn-in’ period.

The second consideration is how long do we need to run the chain to get ‘good’ estimates. This is important as the Markov chain is autocorrelated and the more correlated the chain is the longer we will need to run to get ‘good’ estimates. The field of MCMC diagnostics is large and there are many diagnostics that will give estimated expected run lengths. Figure 7 contains several such MCMC diagnostics along with auto-correlation functions that show how correlated the chain is.

8.4.5 Summary measures

The chain of estimates produced by MCMC (and bootstrap see later) methods for a parameter can be thought of as a sample drawn from this parameter's posterior distribution. As with any sample of values there are many possible summary statistics. The best summary from (a Bayesian viewpoint) is the distribution itself that can be approximated by a kernel density plot.

For point estimation we can construct the mean and median from the chain of values, and the mode (equivalent to the MLE) can be estimated from the kernel density plot. To construct interval estimates we can calculate the appropriate quantiles from the (sorted) chain of values. These will give a more accurate interval estimate as they do not rely on any distributional assumptions. Figure 7 contains many of these summary statistics.

{Figure 7 near here}

8.5 Bootstrap methods

Bootstrapping (Efron and Tibshirani 1993) is a resampling technique that involves generating many simulated datasets based on the current data set and/or parameter estimates and finding estimates for these 'new' simulated datasets consequently giving a sample of estimates for each parameter. Given this sample of estimates we can then use the same summary statistics as described in the MCMC section above.

After fitting a multilevel model using an iterative technique to produce starting values there are two possible techniques to generate ‘new’ datasets. A ‘new’ dataset in this context means generating a new response variable. Given this new dataset the standard ML, REML or QL techniques are used to produce parameter estimates.

8.5.1 Parametric bootstrap

Here the ‘new’ response variable is constructed by adding a random draw from the random part of the model to the fixed predictor, for example consider again the simple variance components model:

$$\begin{aligned}y_{ij} &= \beta_0 + u_{0j} + e_{0ij} \\u_{0j} &\sim N(0, \sigma_{u_0}^2) \quad \{16\} \\e_{0ij} &\sim N(0, \sigma_{e_0}^2)\end{aligned}$$

Here we would generate a set of u_{0j} 's and a set of e_{0ij} 's from the respective Normal distributions and add these to the value β_0 to give a new response for each individual.

8.5.2 Non-parametric bootstrap

Here instead of generating new sets of residuals, we simply resample (with replacement) from the existing estimated residuals and then add these to fixed predictor to construct the ‘new’ response variable. In fact in multilevel modelling this procedure is more complicated as the residuals that are estimated in a multilevel model are ‘shrunk’ residuals. Consequently the residuals are ‘reflated’ before

resampling occurs. Non-parametric bootstrapping has the advantage of not relying on the assumption that the residuals come from a Normal distribution.

8.5.3 Iterative bootstrap

Bootstrapping does not actually produce less biased estimates than the iterative ML/QL techniques but an iterative bootstrap can be used to correct the bias. As an example of how the iterative bootstrap works assume that the true value of a parameter is 1.0 and a ML technique gives an estimate of 0.8. We will assume that the bias is proportional to the true value. Then running the bootstrap which is based on the same procedure will give an estimate of $0.8 \times 0.8 = 0.64$. We could then produce a bias corrected estimate of $0.8 + (0.8 - 0.64) = 0.96$ which is closer to the true value. Now we use this bias corrected estimate as our starting value for the next set of bootstrap runs. As we iterate through more sets of bootstrapped estimates the bias corrected estimate will quickly converge to the true value.

9 Other topics in multilevel modelling

In this section we will explain briefly (with references) other areas of multilevel modelling which we have not covered in the earlier sections. Most of these areas are at the edge of current research.

9.1 Outliers and diagnostics in multilevel modelling

The detection of outliers and other diagnostics such as the identification of data points with a large influence or leverage are as important in multilevel modelling as in any

other statistical modelling. Often significant fixed effects or significant higher level variance may be due to a data point or points that are outlying. However in complex data structures such as multilevel modelling the concept of an outlier is not so easily defined. This is particularly true as an observation can be an outlier but a higher level unit could also be an outlier. The statistical techniques used for exploring outliers in complex data structures are discussed in Langford and Lewis (1998).

9.2 Weighting in multilevel modelling

Individual units, may have differential weights attached to them, e.g. as a result of varying sample selection probabilities from a survey. Thus in a 2 level model we may have differential weights attached to both the level 2 and level 1 units. Weighting for differential selection probabilities in multilevel models is discussed in Pfefferman *et al* (1998).

9.3 Multilevel survival models

This class of models, also known as event duration models, have as their response variable the length of time between 'events'. Such events may be, for example, birth and death, or the beginning and end of a period of employment with corresponding times being length of life or duration of employment.

The multilevel structure of such models arises in two general ways. The first is where we have repeated durations within individuals. Thus, individuals may have repeated spells of various kinds of employment of which unemployment is one, or women may

have repeated spells of pregnancy. In this case we have a 2-level model with individuals at level 2, often referred to as a renewal process. We can include explanatory dummy variables to distinguish different kinds or 'states' of employment or pregnancy, such as the sequence number. The second kind of model is where we have a single duration for each individual, but the individuals are grouped into level 2 units. In the case of employment duration the level 2 units would be firms or employers. If we had repeated measures on individuals within firms then this would give rise to a 3-level structure.

A characteristic of duration data is that for some observations we may not know the exact duration but only that it occurred within a certain interval. This is known as interval censored data, if less than a known value, left censored data, if greater than a known value, right censored data. For example, if we know at the time of a study, that someone began her pregnancy before a certain date then the information available is only that the duration is longer than a known value. Such data are known as right censored. In another case we may know that someone entered and then left employment between two measurement occasions, in which case we know only that the duration lies in a known interval. For a description of multilevel survival models see Chapter 12 of Goldstein (1995).

9.4 Multilevel time series or auto-correlation models

The standard assumption in multilevel models is that the level 1 residuals are independent. In some situations this assumption is false. For growth measurements the specification of level 2 variation serves to model a separate curve for each

individual. If measurements on an individual are obtained very close together in time, they will tend to have similar departures from an individual's underlying growth curve. That is, there is an "auto-correlation" between the level 1 residuals. A detailed discussion of multilevel autocorrelation models is given in Goldstein *et al* (1994).

9.5 Multilevel categorical response models

In section 6 we dealt with discrete response models where the response is a proportion for example the voting example or a count. In fact the response in the voting example could have been altered from the binary response "whether people voted conservative or not" to a categorical response as to which party they voted for. This response would then have a fixed number of categories, 1 for each party and each voter i would have a probability of voting for each party c where the probabilities for a single voter will sum to 1 ie $\sum_c p_{ic} = 1$. This type of response is fitted with a multinomial model.

Another type of categorical response, common in market research data is an ordered categorical response for example people are often asked to choose one of {very good, good, OK, poor, very poor} to describe their opinion of an item. Again each individual will have probabilities of choosing each category that sum to 1 but the categories have a definite order so the probabilities of being in particular categories are correlated. This type of response is fitted using an ordered multinomial model. As with the binomial and Poisson models, multinomial models can be extended to multilevel multinomial models, where individuals in a particular higher level unit share the same underlying category probabilities. See chapter 7.4 of Goldstein 1995 for more details.

9.6 Complex level 1 variation

In the models discussed in this chapter we have fitted a single constant term for the level 1 variation. Often the level 1 variation will be non-constant, this situation is often referred to heteroscedasticity. In multilevel models it is straight-forward to directly model level 1 variation as a function of explanatory variables.

9.7 Meta-analysis

The purpose of meta-analysis is to provide an overall summary of results when information from several studies on the same topic is available. These ‘studies’ may be centres in a single clinical trial, distinct experimental studies, distinct (or possibly overlapping) observational surveys, or mixtures of these. Meta-analysis can therefore be regarded as a special case of the general hierarchical data model, where individual observations are nested within studies or centres.

In applied work, it is often assumed that the effect of interest is constant across the component studies (Thompson & Pocock, 1991), yielding the so-called ‘fixed effect’ model. The assumption of homogeneity can, however, be relaxed to allow for random variation between studies of the effects, yielding the so-called ‘random effects’ model (DerSimonian & Laird, 1986). Statistical models for this case can be fitted using a variance components multilevel model formulation. A general multilevel formulation (Goldstein 1995), however, allows more general random coefficient models to be studied, and we describe this in more detail below. A straightforward extension is to

include covariates in such a model and to observe the extent to which they account for between-study variation. An additional problem is when some studies provide individual level data, while for others only summary results (such as means) are available and methods of meta-analysis which can combine such results efficiently are now available (Goldstein et al., 2000).

10 Available software for multilevel modelling

Multilevel modelling is a fairly new technique and so the software available can be broadly split into three categories: general-purpose statistics packages, special purpose multilevel modelling packages and other special purpose software with some multilevel modelling facilities. These packages vary in terms of estimation methods used, size and type of models allowed and speed of computation. There have been several papers that have compared the performance of packages on particular multilevel problems for example Kreft, de Leeuw and van der Leeden (1994) and Zhou, Perkins and Hui (1999).

We will not discuss this here but simply include a list of packages along with web sites that contain more information.

10.1 General statistic packages

BMDP – <http://www.statsol.ie/prbm.htm>

GENSTAT – <http://www.nag.co.uk/stats/tt-soff.asp>

SAS (PROC MIXED) – <http://www.sas.com>

Stata – <http://www.stata.com>

S-plus – <http://www.splus.mathsoft.com>

10.2 Multilevel modelling packages

HLM – <http://www.ssicentral.com/hlm/hlm.htm>

MIXOR/MIXREG – <http://www.uic.edu/~hedecker/mix.html>

MLwiN – <http://www.ioe.ac.uk/mlwin/>

VARCL - <http://www.gamma.rug.nl/>

10.3 Other special-purpose packages

BUGS (WinBUGS) – <http://www.mrc.bsu.cam.ac.uk/bugs/>

EGRET – <http://www.cytel.com/products/egret/egret1.html>

LISREL – <http://www.ssicentral.com/lisrel/mainlis.htm>

M-PLUS – <http://www.statmodel.com>

SABRE – <http://www.cas.lancs.ac.uk/software/sabre3.1/sabre.html>

11 Further Reading and Useful Resources

11.1 Books on Multilevel modelling

There are many books written on multilevel modelling at both an introductory and more advanced level. For the beginner the following are recommended:

Hox, J.J. (1994). *Applied Multilevel Analysis*. Amsterdam : TT-Publikaties.

Kreft, I and De Leeuw, J. (1998). *Introducing Multilevel Modelling*. London : Sage.

For the reader who wants a more technical text the following are recommended:

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models*. Newbury Park, Ca: Sage.

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

Longford, N. (1993). *Random Coefficient Models*. OUP: Oxford.

11.2 The MLwiN and TRaMSS web sites

This chapter has been written with no emphasis on any particular software package.

The authors however are both members of the team that produce the software package *MLwiN* and a lot of the material in this chapter is covered in greater detail in the User manual for *MLwiN* (Rasbash, Browne, Goldstein, Yang et al. 2000). This user manual can be downloaded for free from the MLwiN web site <http://www.ioe.ac.uk/mlwin> which also contains more information on the package.

There is also a web site TRaMSS which stands for Teaching Resources and Materials for Social Scientists. This web site includes a free training version of the *MLwiN* software along with a set of downloadable tutorials. These tutorials take the user through a series of worked examples that cover the basic concepts of multilevel modelling. These materials can be found at

www.tramss.data-archive.ac.uk

11.3 References from the chapter

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models*. Newbury Park, Ca: Sage.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum Likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

Dersimonian R. and Laird N. (1986) Meta-analysis in Clinical-Trials. *Journal of Controlled Clinical Trials*. **7**, 177-188

Duncan, C., Jones, K. and Moon, G. (1999). Smoking and deprivation: are there neighbourhood effects? *Social Science and Medicine*, **48**, p.497-506

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman and Hall.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika* **73**, 43-56.

Goldstein, H. (1989). Restricted unbiased iterative generalised least squares estimation. *Biometrika* **76**, 622-623.

Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika* **78**, 45-51.

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

Goldstein, H., Healy, M.J.R. and Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine* **13**: 1643-55.

Goldstein, H. and Sammons, P. (1997). The influence of secondary and junior schools on sixteen year examination performance: a cross-classified multilevel analysis. *School effectiveness and school improvement*. **8**: 219-230.

Goldstein H., Yang M., Omar R., Turner R. and Thompson S. (2000) Meta-analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society, Series C*, **49**: 1-14.

Heath , A., Yang, M. and Goldstein, H. (1996). Multilevel analysis of the changing relationship between class and party in Britain 1964-1992. *Quality and Quantity* **30**: 389-404.

Kreft, I.G.G., De Leeuw, J. and van der Leeden, R. (1994) Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML2, and VARCL. *American Statistician* **48**, 324-335.

Laird, N.M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika* **65**, 581-590.

Langford, I.H., Bentham, G. and McDonald, A. (1998). Multilevel modelling of geographically aggregated health data: a case study on malignant melanoma mortality and UV exposure in the European community. *Statistics in Medicine* **17**: 41-58.

Langford, I.H. and Lewis, T. (1998). Outliers in multilevel models (with discussion). *Journal of the Royal Statistical Society, Series A*. **161**: 121-160.

Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika* **74**, 817-827.

O'Muircheartaigh C. and Campanelli P. (1999). A Multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, Series A*, **162**, 437-446

Pfeffermann D., Skinner C. J., Holmes D. J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, **60**: 23-40.

Rasbash J. and Browne W.J. (2000). *Non-hierarchical multilevel models*. In *Multilevel Modelling of Health Statistics*. Ed: A. Leyland and H. Goldstein. Wiley : Chichester.

Rowe , K. J. and Hill, P. W. (1994). *Simultaneous estimation of multilevel structural equations to model students' educational progress*. Tenth International Congress for School effectiveness and Improvement, Memphis, Tennessee.

Tizard, B., Blatchford, P., Burke, J., Farquhar, C. and Plewis, I. (1988). *Young Children at school in the Inner City*. Hove, Sussex: Lawrence Erlbaum.

Zhou, X, Perkins, A.J. and Hui, S.L. (1999). Comparisons of software packages for generalized linear multilevel models. *The American Statistician*. **53**: 282-290.

| Parameter | Estimate (s.e.) |
|------------------------------------|-----------------|
| Fixed : | |
| β_0 (intercept) | 0.002(0.040) |
| β_1 (prior-ability) | 0.563(0.013) |
| Random: | |
| σ_{u0}^2 (between schools) | 0.092(0.018) |
| σ_{e0}^2 (between students) | 0.57(0.013) |

Table 1: Estimates for a random intercepts model

| Parameter | Estimate (s.e.) |
|---------------------------------|-----------------|
| Fixed : | |
| β_0 (intercept) | -0.012(0.040) |
| β_1 (prior-ability) | 0.566(0.020) |
| Random: | |
| Level 2 | |
| σ_{u0}^2 (intercept) | 0.090(0.018) |
| σ_{u01} (covariance) | 0.018(0.006) |
| σ_{u1}^2 (prior-ability) | 0.015(0.004) |
| Level 1 | |
| σ_{e0}^2 | 0.55(0.012) |

Table 2: Estimates for a random slopes model

| Parameter | Estimate (s.e.) A | Estimate (s.e.) B | Estimate (s.e.) C |
|---------------------------------|----------------------|----------------------|----------------------|
| Fixed : | | | |
| β_0 (intercept) | -0.012(0.040) | -0.117(0.08) | -0.184(0.084) |
| β_1 (prior-ability) | 0.566(0.020) | 0.554(0.020) | 0.460((0.042) |
| β_2 (mid) | | 0.084(0.092) | 0.149(0.098) |
| β_3 (high) | | 0.231(0.105) | 0.324(0.101) |
| β_4 (prior-ability.mid) | | | 0.089(0.049) |
| β_5 (prior-ability.high) | | | 0.177(0.055) |
| Random: | | | |
| Level 2 | | | |
| σ_{u0}^2 (intercept) | 0.090(0.018) | 0.078(0.016) | 0.078(0.016) |
| σ_{u01} (covariance) | 0.018(0.006) | 0.013(0.006) | 0.011(0.005) |
| σ_{u1}^2 (prior-ability) | 0.015(0.004) | 0.015(0.004) | 0.011(0.004) |
| Level 1 | | | |
| σ_{e0}^2 | 0.553(0.012) | 0.553(0.012) | 0.55(0.012) |
| Deviance | 9316.9 | 9312.6 | 9303.76 |

Table 3: Effect of adding contextual variables modelling average school ability

| Parameter | Model 4.1 Estimate (s.e.) | Model 4.2 Estimate (s.e.) |
|---------------------------------------|---------------------------|---------------------------|
| Fixed : | | |
| β_0 (intercept) | 7.115 (0.053) | 7.117 (0.041) |
| β_1 (age) | - | 0.997 (0.007) |
| Random : | | |
| σ_{u0}^2 (between individuals) | 0.078 (0.083) | 0.603 (0.048) |
| σ_{e0}^2 (within individuals) | 4.562 (0.172) | 0.307 (0.012) |
| Likelihood | 7685.736 | 3795.588 |

Table 4: Results for variance components models {6} and {7}

| Parameter | Model 4.3 Estimate (s.e.) | Model 4.4 Estimate (s.e.) |
|-------------------------------|---------------------------|---------------------------|
| Fixed : | | |
| β_0 (intercept) | 7.117 (0.043) | 7.115 (0.046) |
| β_1 (age) | 0.995 (0.012) | 0.995 (0.007) |
| β_2 (age squared) | - | 0.001 (0.003) |
| Random : | | |
| Level 2 | | |
| σ_{u0}^2 (intercept) | 0.683 (0.053) | 0.765 (0.060) |
| σ_{u01} | 0.123 (0.012) | 0.139 (0.014) |
| σ_{u1}^2 (age) | 0.037 (0.004) | 0.039 (0.004) |
| σ_{u02} | - | -0.014 (0.003) |
| σ_{u12} | - | -0.002 (0.001) |
| σ_{u2}^2 (age squared) | - | 0.001 (0.000) |
| Level 1 | | |
| σ_{e0}^2 | 0.161 (0.007) | 0.134 (0.007) |
| Likelihood | 3209.392 | 3137.620 |

Table 5: Results for random slopes models {8} and {9}

| Student | Response y_{ij} | Intercepts | | Gender | |
|------------|----------------------|----------------------|-------------------------|-------------------------|----------------------------|
| | | Written z_{1ij} | Coursework z_{2ij} | Written $z_{1ij}x_j$ | Coursework $z_{2ij}x_j$ |
| 1 (female) | y_{11} | 1 | 0 | 1 | 0 |
| 1 | y_{12} | 0 | 1 | 0 | 1 |
| 2 (male) | y_{21} | 1 | 0 | 0 | 0 |
| 2 | y_{22} | 0 | 1 | 0 | 0 |
| 3 (female) | y_{31} | 1 | 0 | 1 | 0 |

Table 6: Data matrix for the examination data.

| Parameter | Estimate(s.e.) |
|------------------------------|--------------------------|
| <i>Fixed :</i> | |
| β_0 (written) | 49.5(0.90) |
| β_1 (coursework) | 69.7(1.2) |
| β_2 (written.girl) | -2.5(0.6) |
| β_3 (coursework.girl) | 6.8(0.7) |
| <i>Random :</i> | |
| Level 3(school) | |
| σ_{v0}^2 (written) | 46.8(9.2) |
| σ_{v01} (covariance) | 24.9(8.9), $\rho = 0.42$ |
| σ_{v1}^2 (coursework) | 75.2(14.6) |
| Level 2(student) | |
| σ_{u0}^2 (written) | 124.6(4.3) |
| σ_{u01} (covariance) | 73.0(4.2), $\rho = 0.49$ |
| σ_{u1}^2 (coursework) | 180.1(6.2) |

Table 7: Results from a multivariate response model of the GCSE data.

| Function name | Formula |
|-----------------------|---------------------------|
| Logit | $\log_e [\pi/(1-\pi)]$ |
| Probit | $\Phi^{-1}(\pi)$ |
| Complementary log-log | $\log_e [-\log_e(1-\pi)]$ |

Table 8: Common link functions for the Binomial distribution

| Parameter | Model 6.1 Estimate (s.e.) | Model 6.2 Estimate (s.e.) |
|--|---------------------------|---------------------------|
| <i>Fixed :</i> | | |
| β_0 (intercept) | -0.248 (0.081) | -0.367 (0.095) |
| β_1 (defence) | - | 0.093 (0.018) |
| β_2 (unemployment) | - | 0.069 (0.014) |
| β_3 (taxes) | - | 0.046 (0.019) |
| β_4 (privatisation) | - | 0.143 (0.018) |
| <i>Random :</i> | | |
| σ_{u0}^2 (between constituencies) | 0.134 (0.091) | 0.167 (0.119) |

Table 9: Estimates for two variance components models fitted to the voting dataset.

| | Neighbourhood 1 | Neighbourhood 2 | neighbourhood 3 |
|------------|-----------------|-----------------|-----------------|
| Hospital 1 | x | xx | |
| Hospital 2 | xx | xx | x |
| Hospital 3 | xxx | | xx |
| Hospital 4 | x | xx | xx |
| Hospital 5 | xx | | |

Table 10: Patients cross classified by hospital and neighbourhood

| | Pupils within ps A | Pupils within ss B | crossed model C |
|---|-----------------------|-----------------------|--------------------|
| <i>Fixed:</i> | | | |
| β_0 (intercept) | 5.97(0.07) | 6.023 | 5.98(0.07) |
| β_1 (vrq) | 0.16(0.003) | 0.16(0.003) | 0.16(0.003) |
| <i>Random:</i> | | | |
| σ_{u1}^2 primary school variance | 0.28(0.06) | | 0.27((0.06) |
| σ_{u2}^2 (secondary school variance) | | 0.05(0.02) | 0.01(0.02) |
| σ_e^2 (pupil variance) | 4.25(0.10) | 4.48(0.11) | 4.25(0.10) |
| Deviance | 14845.9 | 14918.16 | 14845.6 |

Table 11: Comparison of results from incomplete nested model versus full cross-classified model for the Fife, Educational Data.

| | n1($h=1$) | n2($h=2$) | n3($h=3$) |
|-------------|-------------|-------------|-------------|
| P1($i=1$) | 0.5 | 0 | 0.5 |
| P2($i=2$) | 1 | 0 | 0 |
| P3($i=3$) | 0 | 0.5 | 0.5 |
| P4($i=4$) | 0.5 | 0.5 | 0 |

Table 12: An example weighted membership matrix for patients and nurses.

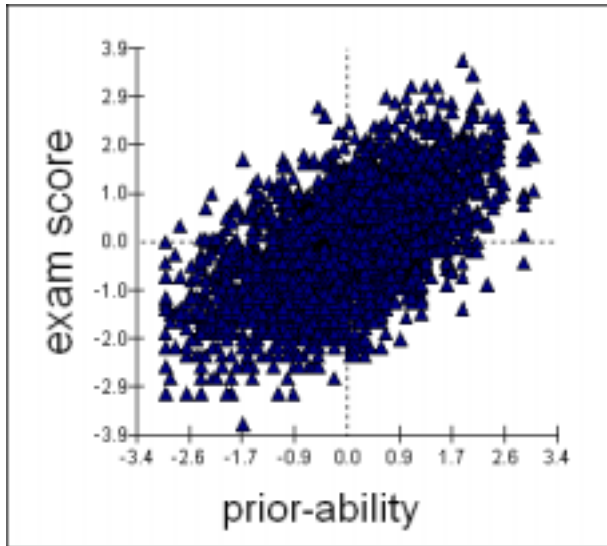


Figure 1: A plot of exam score against prior ability for an educational data set.

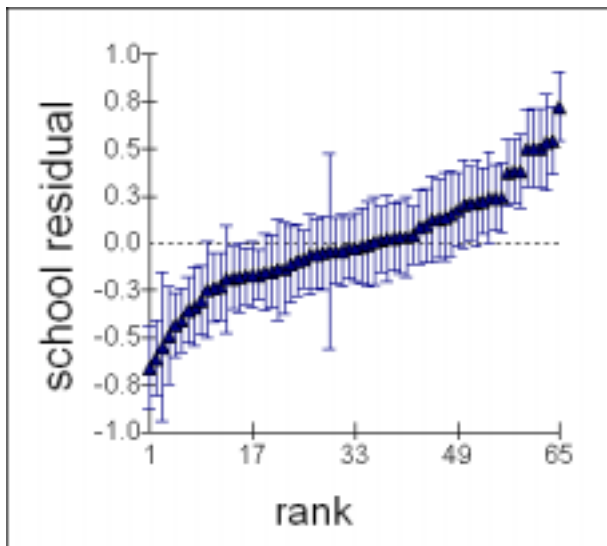


Figure 2: School residuals against their rank with associated 95% confidence intervals

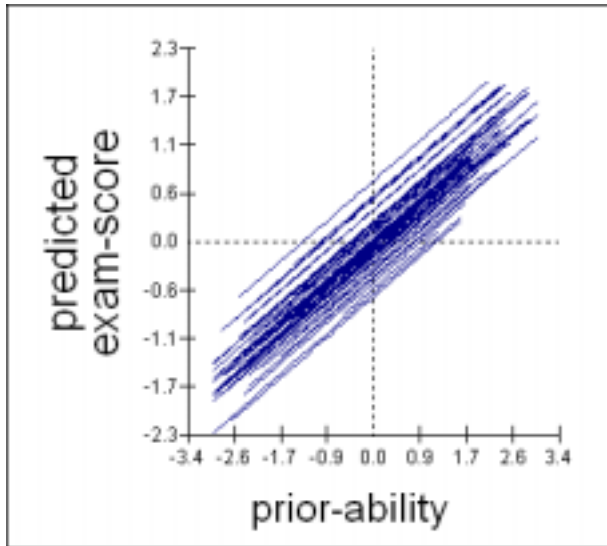


Figure 3: 65 predicted school lines from the random intercept model

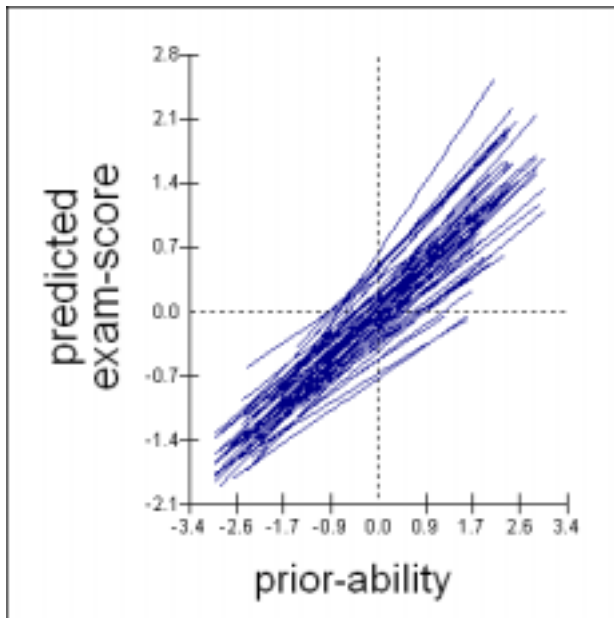


Figure 4: School prediction lines from a random slopes model.

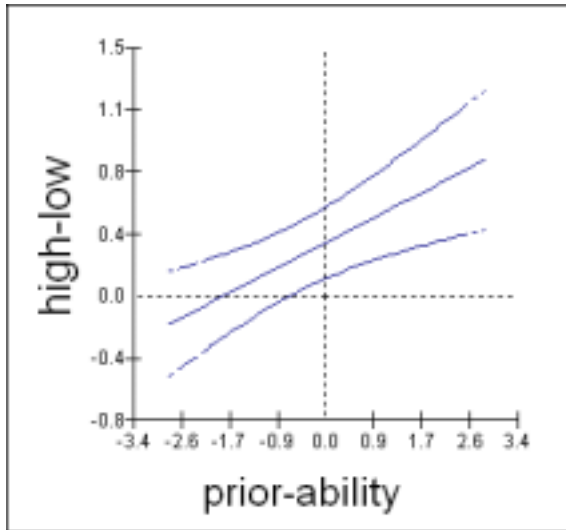


Figure 5: High-low school ability difference as a function of pupil level prior-ability.

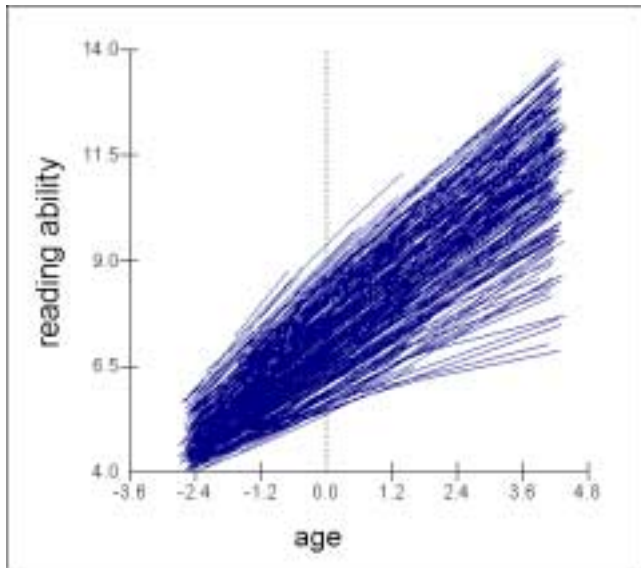


Figure 6: Plot of individual reading versus age regression curves for model {9}

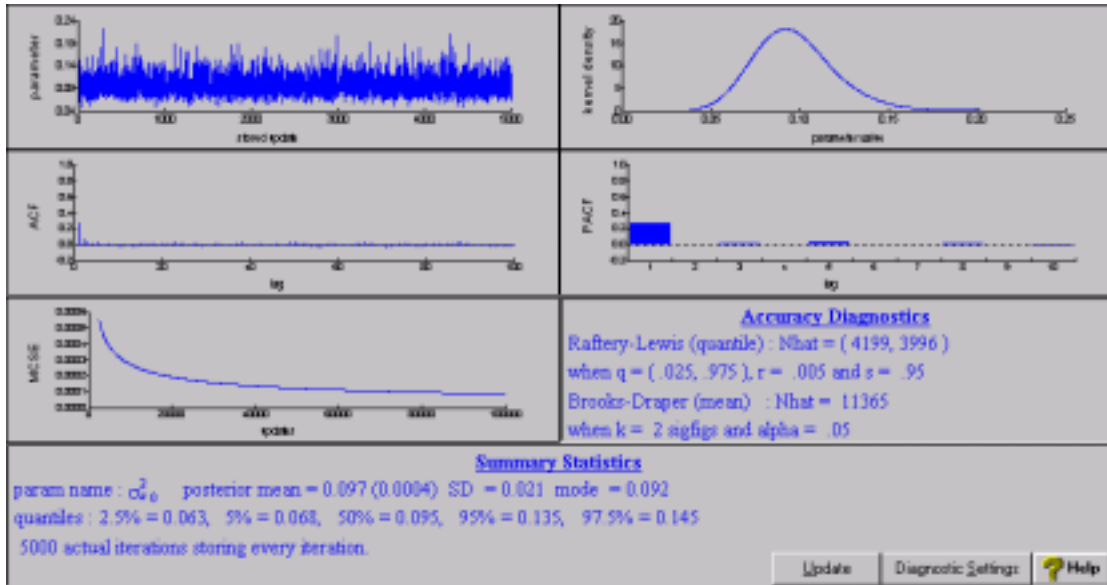


Figure 7: summary information for a MCMC chain for a variance parameter as given in the software package MLwiN.