CHAPTER 3

3 THE BASIC MULTILEVEL MODEL AND EXTENSIONS

In the previous Chapter we introduced a number of models and we cleared out the advantages of Multilevel Models in the analysis of hierarchically nested data. First of all, these models "respect" the hierarchy of the data and analyze data simultaneously in all levels. They allow for variables entry in all levels as well as cross-level interactions (interactions of variables measured in different levels). In Random Coefficient Models the lowest level regression coefficient are treated as random variables at the higher level, which explains further the variability of the model.

In this Chapter we first elaborate more on the development of a basic 2-level model. We reconsider alternative ways and notations of setting up and motivating the model and introduce procedures for estimating parameters, forming and testing functions of the parameters and constructing confidence intervals. Then we extend to the natural extensions of the basic 2-level model by introducing higher-level structure, as well as special cases. These are the cross-classified models, the generalized multilevel models for proportion as outcome and the multivariate multilevel model.

The scope of this Chapter is, therefore, to present in extent all theoretical aspects and advantages of a Multilevel Model and to show how this kind of analysis can be effective both in simple hierarchical data problems, as well as in even more complex theoretical statistical data structures.

3.1 The Basic Two-Level Model - The Formulas

3.1.1 The 2-level model and basic notation

We first consider a simple model for one group, relating the response variable to one simple explanatory variable. We write:

$$y_i = \alpha + \beta x_i + e_i \qquad (3.1)$$

where standard interpretations can be given to the intercept (α), slope (β) and residual (e_i). We follow the normal convention of using Greek letters for the regression coefficients and place a circumflex over any coefficient (parameter) which is a sample estimate. This is the formal model and describes a single-level relationship. To describe simultaneously the relationships for several groups we write, for group j

$$y_{ij} = \alpha_j + \beta_j x_{ij} + e_{ij} \qquad (3.2)$$

This is now the formal model where j refers to the level 2 unit and i to the level 1 unit. As it stands, (3.2) is still essentially a single level model, albeit describing a separate relationship for each group. In some situations, for example where there are few groups and interest centres on just those groups in the sample, we may analyze (3.2) by fitting all the $2n+1$parameters, namely

$$(\alpha_j, \beta_j) \quad j = 1,...,n \text{ and } \sigma_e^2$$

assuming a common 'within-group' residual variance and separate lines for each group.

If we wish to focus not just on these groups, but on a wider 'population' of groups then we need to regard the chosen groups as giving us information about the characteristics of all the groups in the population. Just as we choose random samples of individuals to provide estimates of population means etc., so a randomly chosen sample of groups can provide information about the characteristics of the population of groups. In particular, such a sample can provide estimates of the variation and covariation between groups in the slope and intercept parameters and will allow us to compare groups with different characteristics.

An important class of situations arises when we wish primarily to have information about each individual group in a sample, but where we have a large number of groups so that (3.2) would involve estimating a very large number of parameters. Furthermore, some groups may have rather small numbers of observations and application of (3.2) would result in imprecise estimates. In such cases, if we regard the groups as members of a population and then use our population estimates of the mean and between-group variation, we can utilize this information to obtain more precise estimates for each individual group. This will be discussed later in the section dealing with 'residuals'.

To make (3.2) into a genuine 2-level model we let α_j, β_j become random variables. For consistency of notation replace α_j by β_{0j} and β_j by β_{1j} and assume that

$$\beta_{0j} = \beta_0 + u_{0j} \qquad (3.3a)$$

$$\beta_{1j} = \beta_1 + u_{1j} \qquad (3.3b)$$

where u_{0j}, u_{1j} are random variables with parameters

$$E(u_{0j}) = E(u_{1j}) = 0 \qquad (3.4a)$$

$$\text{var}(u_{0j}) = \sigma_{u0}^2,\ \text{var}(u_{1j}) = \sigma_{u1}^2,\ \text{cov}(u_{0j}, u_{1j}) = \sigma_{u01} \qquad (3.4b)$$

We can now write (3.2) in the form

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + (u_{0j} + u_{1j} x_{ij} + e_{0ij}) \qquad (3.5)$$

where

$$\text{var}(e_{0ij}) = \sigma_{e0}^2 \qquad (3.6)$$

We shall require the extra suffix in the level 1 residual term for models with more complex residual term.

We have expressed the response variable y_{ij} as the sum of a fixed part and a random part within the brackets. We shall also generally write the fixed part of (3.5) in the matrix form

$$E(Y) = X\beta \text{ with } Y = \{y_{ij}\} \qquad (3.7)$$

$$E(y_{ij}) = X_{ij}\beta = (X\beta)_{ij},\ X = \{X_{ij}\} \qquad (3.8)$$

where $\{\}$ denotes a matrix, *X* is the design matrix for the explanatory variables and X_{ij} is the *ij*-th row of X. For model (3.5) we have $X = \{1 \ \ x_{ij}\}$. Note the alternative representation for the *i*-th row of the fixed part of the model.

The random variables are referred to as 'residuals' and in the case of a single level model the level 1 residual e_{0ij} becomes the usual linear model residual term. To make the model symmetrical so that each coefficient has an associated explanatory variable, we can define a further explanatory variable for the intercept

β_0 and its associated residual, u_{0j}, namely x_{0ij}, which takes the value 1.0. For simplicity this variable may often be omitted.

The feature of (3.5) which distinguishes it from standard linear models of the regression or analysis of variance type is the presence of more than one residual term and this implies that special procedures are required to obtain satisfactory parameter estimates. Note that it is the structure of the random part of the model, which is the key factor. In the fixed part the variables can be measured at any level. We can also include so called 'compositional' variables such as the average value of an explanatory variable for all individuals in each group. The presence of such variables does not alter the estimation procedure, although results will require careful interpretation. We will elaborate more on estimation procedures in the following section.

3.1.2 Parameter estimation for the variance components model

Equation (2.5) requires the estimation of two fixed coefficients, β_0, β_1, and four other parameters, $\sigma^2_{u0}, \sigma^2_{u1}, \sigma_{u01}$ and σ^2_{e0}. We refer to such variances and covariances as *random parameters*. We start, however, by considering the simplest 2-level model, which includes only the random parameters $\sigma^2_{u0}, \sigma^2_{e0}$. It is termed a variance components model because the variance of the response, about the fixed component, the *fixed predictor*, is

$$\mathrm{var}(y_{ij} | \beta_0, \beta_1, x_{ij}) = \mathrm{var}(u_0 + e_{0ij}) = \sigma^2_{u0} + \sigma^2_{e0} \qquad (3.9)$$

that is, the sum of a level 1 and a level 2 variance. This model implies that the total variance for each individual is constant and that the covariance between two individuals (denoted by i_1, i_2) in the same group is given by

$$\mathrm{cov}(u_{0j} + e_{0i_1 j}, u_{0j} + e_{i_2 j}) = \mathrm{cov}(u_{0j}, u_{0j}) = \sigma^2_{u0} \qquad (3.10)$$

since the level 1 residuals are assumed to be independent. The correlation between two such individuals is therefore

$$\rho = \frac{\sigma^2_{u0}}{(\sigma^2_{u0} + \sigma^2_{e0})} \qquad (3.11)$$

which is referred to as the ‘intra-level-2-unit correlation’ or the ‘intra-class’ correlation. This correlation measures the proportion of the total variance which is between-groups. In a model with 3 levels, we will have two such correlations; the ‘intra-level-3-unit correlation’ and the intra-level-2-unit correlation’, and so on.

The existence of a non-zero intra-unit correlation, resulting from the presence of more than one residual term in the model, means that traditional estimation procedures such as 'ordinary least squares' (OLS) which are used for example in multiple regression, are inapplicable. A later section illustrates how the application of OLS techniques leads to incorrect inferences. We now look in more detail at the structure of a 2-level data set, focusing on the covariance structure typified by Figure 3.1.

Figure 3.1: Covariance matrix of three first-level units in a single 2-level context for a variance components model

$$\begin{pmatrix} \sigma_{u0}^2+\sigma_{e0}^2 & \sigma_{u0}^2 & \sigma_{u0}^2 \\ \sigma_{u0}^2 & \sigma_{u0}^2+\sigma_{e0}^2 & \sigma_{u0}^2 \\ \sigma_{u0}^2 & \sigma_{u0}^2 & \sigma_{u0}^2+\sigma_{e0}^2 \end{pmatrix}$$

The matrix in figure 3.1 is the (3 x 3) covariance matrix for the scores of three individuals in a single group, derived from the above expressions. For two groups, one with three individuals and one with two, the overall covariance matrix is shown in Figure 3.2. This 'block-diagonal' structure reflects the fact that the covariance between individuals in different groups is zero, and clearly extends to any number of level 2 units.

Figure 3.2: The block-diagonal covariance matrix for the response vector Y for a 2-level variance components model with two level 2 units

$$\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} = \begin{bmatrix} \begin{pmatrix} \sigma_{u0}^2+\sigma_{e0}^2 & \sigma_{u0}^2 & \sigma_{u0}^2 \\ \sigma_{u0}^2 & \sigma_{u0}^2+\sigma_{e0}^2 & \sigma_{u0}^2 \\ \sigma_{u0}^2 & \sigma_{u0}^2 & \sigma_{u0}^2+\sigma_{e0}^2 \end{pmatrix} & 0 \\ 0 & \begin{pmatrix} \sigma_{u0}^2+\sigma_{e0}^2 & \sigma_{u0}^2 \\ \sigma_{u0}^2 & \sigma_{u0}^2+\sigma_{e0}^2 \end{pmatrix} \end{bmatrix}$$

A more compact way of presenting this matrix, which we shall use, again is given in figure 3.3.

Figure 3.3: Block-diagonal covariance matrix using general notation

$$V_2 = \begin{bmatrix} \sigma^2_{u0} J_{(3)} + \sigma^2_{e0} I_{(3)} & 0 \\ 0 & \sigma^2_{u0} J_{(2)} + \sigma^2_{e0} I_{(2)} \end{bmatrix}$$

where $I_{(n)}$ is the (n x n) identity matrix and $J_{(n)}$ is the (n x n) matrix of ones. The subscript 2 for V indicates a 2-level model. In single-level OLS models σ^2_{u0} is zero and this covariance matrix then reduces to the standard form $\sigma^2 I$ where σ^2 is the (single) residual variance.

3.1.3 The general 2-level model including random coefficients

We now extend (3.5) in the standard way to include further fixed explanatory variables

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \sum_{h=2}^{p} \beta_h x_{hij} + (u_{0j} + u_{1j} x_{1ij} + e_{0ij}) \qquad (3.12)$$

and more compactly as

$$y_{ij} = X_{ij}\beta + \sum_{h=0}^{1} u_{hj} z_{hij} + e_{0ij} z_{0ij} \qquad (3.13)$$

where we use new explanatory variables for the random part of the model and write these more generally as

$$Z = \{Z_0 \; Z_1\} \qquad (3.14)$$

where $Z_0 = \{1\}$ i.e a vector of 1s and $Z_1 = \{x_{1ij}\}$.

The explanatory variables for the random part of the model are often a subset of those in the fixed part, as here, but this is not necessary. Also, any of the explanatory variables may be measured at any of the levels; for example we may have individual characteristics at level 1 or group characteristics at level 2.

This model (3.13), with the coefficient of X_1 random at level 2, gives rise to the following typical block structure, for a level-two block with two level-one units. The matrix Ω_2 is the covariance matrix of the random intercept and slope at level 2. Note that we need to distinguish carefully between the covariance matrix of the responses given in the following structure and the covariance matrix of the random coefficients. We also refer to the intercept as a random coefficient. The matrix Ω_1 is the covariance matrix for the set of level-one random coefficients; in this case there is just a single variance term at level one. We also write $\Omega = \{\Omega_i\}$ for the set of these covariance matrices. More explicitly:

Figure 3.4: Response covariance matrix for a level 2 unit with two level 1 units for a 2-level model with a random intercept and random regression coefficient at level-2

$$\begin{pmatrix} A & B \\ B & C \end{pmatrix} = \begin{pmatrix} \sigma_{u0}^2 + 2\sigma_{u01}x_{1j} + \sigma_{u1}^2 x_{1j}^2 + \sigma_{e0}^2 & \sigma_{u0}^2 + \sigma_{u01}(x_{1j} + x_{2j}) + \sigma_{u1}^2 x_{1j}x_{2j} \\ \sigma_{u0}^2 + \sigma_{u01}(x_{1j} + x_{2j}) + \sigma_{u1}^2 x_{1j}.x_{2j} & \sigma_{u0}^2 + 2\sigma_{u01}x_{2j} + \sigma_{u1}^2 x_{2j}^2 + \sigma_{e0}^2 \end{pmatrix}$$

$$A = \sigma_{u0}^2 + 2\sigma_{u01}x_{1j} + \sigma_{u1}^2 x_{1j}^2 + \sigma_{e0}^2$$

$$B = \sigma_{u0}^2 + \sigma_{u01}(x_{1j} + x_{2j}) + \sigma_{u1}^2 x_{1j}.x_{2j}$$

$$C = \sigma_{u0}^2 + 2\sigma_{u01}x_{2j} + \sigma_{u1}^2 x_{2j}^2 + \sigma_{e0}^2$$

giving

$$\begin{pmatrix} A & B \\ B & C \end{pmatrix} = X_j.\Omega_2 X_j^T + \begin{pmatrix} \Omega_1 & \\ & \Omega_1 \end{pmatrix}$$

where

$$X_j = \begin{pmatrix} 1 & x_{1j} \\ 1 & x_{2j} \end{pmatrix},$$

$$\Omega_2 = \begin{pmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{pmatrix},$$

$$\Omega_1 = \sigma_{e0}^2$$

We also see here the general pattern for constructing the response covariance matrix which generalizes both to higher order models and to complex variation at level 1.

3.1.4 Parameter Estimates - Possible Approaches – Algorithms

It is obvious even from the previous discussion that the parameters estimation for both the fixed and the random part of the model is a crucial issue in Multilevel analysis, especially due to the large number of parameters that have to be estimated. For a model of P predictors for the lowest level and Q predictors for the highest level the number of estimates is shown in the following Table (taken by Hox (1995)):

Table 3.1: Number of parameters to be estimated in a "full" Multilevel model

Parameters	Number of Estimates
Intercept	1
Lowest level error variance	1
Slopes for the lowest level predictors	P
Highest level error variances for these slopes	P
Highest level covariances of the intercept with all slopes	P
Highest level covariances between all slopes	P(P-1)/2
Slopes for the highest level predictors	Q
Slopes for cross level interactions	P x Q

Several techniques and principles and their corresponding algorithms have been proposed in order to reach reliable estimates for both fixed and random part. As far as Maximum Likelihood technique is concerned, two different varieties of Maximum Likelihood estimation are used for multilevel regression analysis. One is called Full Maximum Likelihood (FML); in this method both the regression coefficients and the variance components are included in the likelihood function. The other method is called Restricted Maximum Likelihood (REML), here only the variance components are included in the likelihood function. The difference is that FML treats the estimates for the regression coefficients as known quantities when the variance components are estimated, while REML treats them as estimates that carry some amount of uncertainty (Bryk & Raudenbush, 1992). Since REML is more realistic, it should, in theory, lead to better estimates, especially when the number of groups is small (Bryk & Raudenbush, 1992). However, in practice, the differences between the two methods are not very important.

Computing the Maximum Likelihood estimates requires an iterative procedure. The most commons of the algorithms (The Iterative Generalized Least Square Method and the EM algorithm) are discussed in this chapter, as well as other techniques and procedures.

The Iterative Generalized Least Square (IGLS) Method

We now give an overview of the Iterative Generalized Least Squares (IGLS) method which also forms the basis for many of the developments in more complex analysis.

We consider the simple 2-level variance components model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + e_{0ij} \qquad (3.15)$$

Suppose that we knew the values of the variances, and so could construct immediately the block-diagonal matrix V_2, which we will refer to simply as V. We can then apply immediately the usual Generalized Least Squares (GLS) estimation procedure to obtain the estimator for the fixed coefficients

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \qquad (3.16)$$

where in this case

$$X = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n_m m} \end{pmatrix} \qquad Y = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_m m} \end{pmatrix} \qquad (3.17)$$

with m level 2 units and n_j level 1 units in the *j*-th level 2 unit. When the residuals have Normal distributions (3.16) also yields maximum likelihood estimates.

Our estimation procedure is iterative. We would usually start from 'reasonable' estimates of the fixed parameters. Typically these will be those from an initial OLS fit (that is assuming $\sigma_{u0}^2 = 0$), to give the OLS estimates of the fixed coefficients $\hat{\beta}_{(0)}$. From these we form the 'raw' residuals

$$\tilde{y}_{ij} = y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_{ij} \qquad (3.18)$$

The vector of raw residuals is written

$$\tilde{Y} = \{\tilde{y}_{ij}\} \qquad (3.19)$$

If we form the cross-product matrix $\tilde{Y}\tilde{Y}^T$ we see that the expected value of this is simply *V*. We can rearrange this cross product matrix as a vector by stacking the columns one on top of the other which is written as $vec(\tilde{Y}\tilde{Y}^T)$ and similarly we can construct the vector $vec(V)$. For the structure given in figure 3.2, these both have $3^2 + 2^2$= 13 elements. The relationship between these vectors can be expressed as the following linear model

$$\begin{pmatrix} \tilde{y}_{11}^2 \\ \tilde{y}_{21}\tilde{y}_{11} \\ \vdots \\ \tilde{y}_{22}^2 \end{pmatrix} = \begin{pmatrix} \sigma_{u0}^2 + \sigma_{e0}^2 \\ \sigma_{u0}^2 \\ \vdots \\ \sigma_{u0}^2 + \sigma_{e0}^2 \end{pmatrix} + R = \sigma_{u0}^2 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \sigma_{e0}^2 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{pmatrix} + R \qquad (3.20)$$

where R is a residual vector. The left hand side of (3.20) is the response vector in the linear model and the right hand side contains two explanatory variables, with coefficients σ_{u0}^2, σ_{e0}^2 which are to be estimated. The estimation involves an application of GLS using the estimated covariance matrix of $vec(\tilde{Y}\tilde{Y}^T)$, assuming Normality, namely $2(V^{-1} \otimes V^{-1})$ where $\otimes$ is the Kronecker product. The Normality assumption allows us to express this covariance matrix as a function of the random parameters. Even if the Normality assumption fails to hold, the resulting estimates are still consistent, although not fully efficient, but standard errors, estimated using the Normality assumption and, for example confidence intervals will generally not be consistent. For certain variance component models alternative distributional assumptions have been studied, especially for discrete response models of the kind discussed later in the thesis and maximum likelihood estimates obtained. For more general models, however, with several random coefficients, the assumption of multivariate Normality is a flexible one, which allows a convenient parameterization for complex covariance structures at several levels.

With the estimates obtained from applying GLS to (3.20) we return to (3.16) to obtain new estimates of the fixed effects and so alternate between the random and fixed parameter estimation until the procedure converges, that is the estimates for all the parameters do not change from one cycle to the next. At convergence, assuming multivariate Normality, the estimates are maximum likelihood. Essentially the same procedure can be used for the more complicated models discussed later on in the thesis. The maximum likelihood procedure produces biased estimates of the random parameters because it takes no account of the sampling variation of the fixed parameters. This may be important in small samples. Goldstein (1989a) shows how a simple modification leads to restricted iterative generalized least squares (RIGLS) or restricted maximum likelihood (REML) estimates which are unbiased. The IGLS algorithm is readily modified to produce these restricted estimates (RIGLS)

Full details of efficient computational procedures for carrying out all these calculations are given by Goldstein & Rasbash (1992).

The EM algorithm

To illustrate the procedure, consider the 2-level variance components model

$$y_{ij} = (X\beta\ )_{ij} + u_j + e_{ij}, \qquad \mathrm{var}(e_{ij}) = \sigma^2_{\mathrm{e}}, \quad \mathrm{var}(\mathrm{u}_{\mathrm{j}}) = \sigma^2_u \qquad (3.21)$$

The vector of level 2 residuals is treated as missing data and the 'complete' data therefore consists of the observed vector Y and the u_j treated as observations. The joint distribution of these, assuming Normality, and using our standard notation is

$$\begin{bmatrix} Y \\ u \end{bmatrix} = N\left\{ \begin{bmatrix} X\beta \\ 0 \end{bmatrix}, \begin{bmatrix} V & J^T\sigma^2_u \\ \sigma^2_u J & \sigma^2_u I \end{bmatrix} \right\} \qquad (3.22)$$

This generalizes readily to the case where there are several random coefficients. If we denote these by β_j we note that some of them may have zero variances. We can now derive the distribution of $\beta_j|Y$, and we can also write down the Normal log likelihood function for (3.22) with a general set of random coefficients, namely

$$\log(L) \propto -N\log(\sigma_e{}^2) - J\log|\Omega| - \sigma_e{}^{-2}\sum_{ij} e_{ij}^2 - \sum_j \beta_j^T \Omega_u{}^{-1}\beta_j \quad (3.23)$$

Maximizing the latter for the random parameters we obtain

$$\hat{\sigma}_e^2 = N^{-1}\sum_{ij} e_{ij}^2 \qquad (3.24)$$

$$\hat{\Omega}_u = m^{-1}\sum_j \beta_j \beta_j^T \qquad (3.25)$$

where m is the number of level 2 units. We do not know the values of the individual random variables. We require the expected values, conditional on the Y and the current parameters, of the terms under the summation signs, these being the sufficient statistics. We then substitute these expected values in (3.24) and (3.25) for the updated random parameters. These conditional values are based upon the 'shrunken' predicted values and their (conditional) covariance matrix. With these updated values of the random parameters we can form V and hence obtain the updated estimates for the fixed parameters using generalized least squares. We note that the expected values of the sufficient statistics can be obtained using the general result for a random parameter vector θ.

$$E(\theta\theta^T) = \text{cov}(\theta) + [E(\theta)][E(\theta)]^T \quad (3.26)$$

The prediction is known as the E (expectation) step of the algorithm and the computations in (3.25) and (3.26) the M (maximization) step. Given starting values, based upon OLS, these computations are iterated until convergence is obtained. Convenient computational formulae for computing these quantities at each iteration can be found in Bryk & Raudenbush (1992).

Markov Chain Monte Carlo estimation – The Gibbs Sampling

Markov Chain Monte Carlo algorithms exploit the properties of Markov chains where the probability of an event is conditionally dependent on a previous state. The procedure is iterative and at each stage from the full multivariate distribution the distribution of each component conditional on the remaining components is computed and used to generate a random variable. The components may be variates, regression coefficients, covariance matrices etc. After a suitable number of iterations, we obtain a sample of values from the distribution of any component, which we can then use to derive any desired characteristic such as the mean, covariance matrix, etc. The most common procedure is that of Gibbs Sampling and Gilks et al. (1993) provide a comprehensive discussion with applications and an application to a 2-level logit model is given by Zeger & Karim (1991). It allows the fitting of Bayesian models where prior distributions for the parameters are specified.

We outline a Gibbs Sampling procedure for a 2-level model. We write:

$$Y = X\beta + Z^{(2)}u + Z^{(1)}e \qquad (3.27)$$

We first consider the distribution $\beta | u^{(k)}, Y$ where k refers to the *k*-th iteration.

Given $u^{(k)}$, $Z^{(2)}u$ is just an offset so that we can regress y_{ij} on x_{ij} to estimate $\hat{\beta}^{(k)}$ and $\mathrm{var}(\hat{\beta}^{(k)})$.

We can then select a random vector from this distribution, assumed to be multivariate normal $(\hat{\beta}^{(k)}, \mathrm{var}(\hat{\beta}^{(k)}))$.

We now consider the distribution of $\Omega_2 | u^{(k)}$. We have (with a non-informative prior) that the (posterior) distribution of Ω_2^{-1} is a Wishart distribution with parameter (i.e. covariance) matrix

$$S^{(k)} = \sum_{j=1}^{J} u_j^{(k)} u_j^{(k)^T} \quad with \;\; d = J - q + 1 \text{ d.f.} \qquad (3.28)$$

where *J* is the number of level 2 units and *q* is the number of random coefficients.

A simple way of generating such a Wishart distribution is to generate *d* multivariate normal vectors from $N(0, S^{(k)})$ and form their SSP matrix. This provides $\hat{\Omega}_2^{(k)}$.

Finally we consider the distribution $u_j | \beta, \Omega_2, Y$. These are the usual level 2 residuals, for which we have standard expressions for their expected values and covariance matrix. We note that for a 2-level model (but not within a three level model) these are block-independent. Assuming Normality we can now generate a set of $u_j^{(k)}$ and this completes an iterative cycle.

There are some particular computational details to be noted. For example 'rejection sampling' at each cycle can be used and we can do several cycles for Ω_2, u_j for each β since the former tend to have higher autocorrelations across cycles.

The procedure can be applied to any existing models, e.g. logit models, where the conditional distributional assumptions are explicit. Gibbs Sampling tends to be computationally demanding, with hundreds if not thousands of iterations required and this can be particularly burdensome when several different models are being explored for their fit to the data. On the other hand, this approach has the advantage, in small samples, that it takes account of the uncertainty associated with the estimates of the random parameters and can provide exact measures of uncertainty. The maximum likelihood methods tend to overestimate precision because they ignore this uncertainty. In small samples this will be important especially when obtaining 'posterior' estimates for residuals, which will be discussed in the following section. Gibbs sampling approach is therefore useful for small and moderate sized samples and when used in conjunction with likelihood based EM or IGLS algorithms.

Other estimation procedures

A variation on IGLS is Expected Generalized Least Squares (EGLS) or the “Gauss-Newton method” as it is mentioned by other authors (Kreft & Leeuw, 1998). This focuses interest on the fixed part parameters and uses the estimate of V obtained after the first iteration merely to obtain a consistent estimator of the fixed part coefficients without further iterations. A variant of this separates the level 1 variance from V as a parameter to be estimated iteratively along with the fixed part coefficients.

Longford (1987) developed a procedure based upon a 'Fisher scoring' algorithm which can be seen that it is formally equivalent to IGLS. This algorithm can also incorporate certain extensions, for example to handle discrete response data.

We have already mentioned the full Bayesian approach, which has become computationally feasible with the development of 'Markov Chain Monte Carlo' (MCMC) methods, especially Gibbs Sampling (Zeger & Karim, 1991). An alternative to the full Bayes estimation, known as 'Empirical Bayes’, ignores the prior distributions of the random parameters, treating them as known for purposes of inference. When Normality is assumed, these estimates are the same as IGLS or RIGLS.

Another approach, which parallels all that was mentioned so far, is that of Generalized Estimating Equations (GEE) introduced by Liang & Zeger (1986). The principal difference is that GEE obtains the estimate of V using simple regression or 'moment' procedures based upon functions of the actual calculated raw residuals. It is concerned principally with modeling the fixed coefficients rather than exploring the structure of the random component of the model. While the resulting coefficient estimates are consistent they are not fully efficient. In some circumstances, however, GEE coefficient estimates may be preferable, since they will usually be quicker to obtain and they make weaker assumptions about the structure of *V*. The GEE procedure can be extended to handle most of the models dealt with more complex cases.

3.1.5 Estimating the residuals

In a single level model such as (3.1) the usual estimate of the single residual term e_i is just $\tilde{y}_i$ the raw residual. In a multilevel model, however, we shall generally have several residuals at different levels. In this chapter we consider estimating the individual residuals in all levels.

Given the parameter estimates, consider predicting a specific residual, say u_{0j} in a 2-level variance components model. Specifically we require for each level 2 unit

$$\hat{u}_{0j} = E(u_{0j}|Y,\hat{\beta},\hat{\Omega}) \qquad (3.29)$$

We shall refer to these as estimated or predicted residuals or, using Bayesian terminology, as posterior residual estimates. If we ignore the sampling variation attached to the parameter estimates in (3.29) we have

$$\mathrm{cov}(\tilde{y}_{ij},u_{0j}) = \mathrm{var}(u_{0j}) = \sigma^2_{u0} \qquad (3.30a)$$

$$\mathrm{cov}(\tilde{y}_{ij},e_{0ij}) = \sigma^2_{e0} \quad (3.30b)$$

$$\mathrm{var}(\tilde{y}_{ij}) = \sigma^2_{u0} + \sigma^2_{e0} \qquad (3.30c)$$

We regard (3.29) as a linear regression of u_{0j} on the set of $\{\tilde{y}_{ij}\}$ for the *j*-th level 2 unit and (3.13) defines the quantities required to estimate the regression coefficients and hence $\hat{u}_{0j}$. For the variance components model we obtain

$$\hat{u}_{0j} = \frac{n_j\sigma^2_u}{(n_j\sigma^2_u + \sigma^2_{e0})}\tilde{y}_j \qquad (3.31a)$$

$$\tilde{e}_{0ij} = \tilde{y}_{ij} - \hat{u}_{0j} \qquad (3.31b)$$

$$\tilde{y}_j = (\sum_i \tilde{y}_{ij})/n_j \quad (3.31c)$$

where n_j is the number of level 1 units in the *j*-th level 2 unit. The residual estimates are not, unconditionally, unbiased but they are consistent. The factor multiplying the mean $(\tilde{y}_j)$ of the raw residuals for the *j*-th unit is often referred to as a 'shrinkage factor' since it is always less than or equal to one. As n_j increases this factor tends to one, and as the number of level 1 units in a level 2 unit decreases the 'shrinkage estimator' of u_{0j} becomes closer to zero. In many applications the higher level residuals are of interest in their own right and the increased shrinkage for a small level 2 unit can be regarded as expressing the relative lack of information in the unit so that the best estimate places the predicted residual close to the overall population value as given by the fixed part.

These residuals therefore can have two roles. Their basic interpretation is as random variables with a distribution whose parameter values tell us about the variation among the level 2 units, and which provide efficient estimates for the fixed coefficients. A second interpretation is as individual estimates for each level 2 unit where we use the assumption that they belong to a population of units to predict their values. In particular, for units which have only a few level 1 units, we can obtain more precise estimates than if we were to ignore the population membership assumption and use only the information from those units. This becomes especially important for estimates of residuals for random coefficients, where in the extreme case of only one level-one unit in a level-two unit we lack information to form an independent estimate.

As in single level models we can use the estimated residuals to help check on the assumptions of the model. The two particular assumptions that can be studied readily are the assumption of Normality and that the variances in the model are constant. Because the variances of the residual estimates depends in general on the values of the fixed coefficients it is common to standardize the residuals by dividing by the appropriate standard errors, which are referred as 'diagnostic' or 'unconditional' standard errors (Goldstein, 1995).

When the residuals at higher levels are of interest in their own right, we need to be able to provide interval estimates and significance tests as well as point estimates for them or functions of them. For these purposes we require estimates of the standard errors (the so-called 'conditional' or 'comparative' standard errors) of the estimated residuals, where the sample estimate is viewed as a random realization from repeated sampling of the same higher-level units whose unknown true values are of interest.

The level 1 residuals are generally not of interest in their own right but are used rather for model checking, having first been standardized using the diagnostic standard errors. Checking the model assumptions in a multilevel model are used in an exactly analogous way as in simple regression models. In other words, we use plot of the standardized level 1 residuals against the fixed part predicted value to check the assumption of a constant level 1 variance ('homoscedasticity') and Normal score plots for level-one (and level-two) residuals to check the assumption of Normality.

3.1.6 Hypothesis testing and confidence intervals

In this section we deal with large sample procedures for constructing interval estimates for parameters or linear functions of parameters and for hypothesis testing. Hypothesis tests are used sparingly in multilevel analysis since the usual form of a null hypothesis, that a parameter value or a function of parameter values is zero, is usually implausible and also relatively uninteresting. Moreover, with large enough samples a null hypothesis will almost certainly be rejected. The exception to this is where we are interested in whether a difference is positive or negative, and this is discussed in the section on residuals below. Confidence intervals emphasize the uncertainty surrounding the parameter estimates and the importance of their substantive significance.

Fixed parameters

We have already presented parameter estimates techniques for the fixed part parameters together with their standard errors. These are adequate for hypothesis testing or confidence interval construction separately for each parameter. In many cases, however, we are interested in combinations of parameters. For hypothesis testing, this most often arises for grouped or categorized explanatory variables where n group effects are defined in terms of $n-1$ dummy variable contrasts and we wish simultaneously to test whether these contrasts are zero. We may also be interested in providing a pair of confidence intervals for the parameter estimates. We proceed as follows:

Define a (r x p) contrast matrix ***C***. This is used to form linearly independent functions of the p fixed parameters in the model of the form $f = C\beta$, so that each row of ***C*** defines a particular linear function. Parameters that are not involved have the corresponding elements set to zero. Suppose we wish to test the hypothesis that the coefficients of two variables each having two categories are jointly zero. We define

$$C = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad f = \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix}$$

and the general null hypothesis is

$$H_0\!: f = k, \quad k = \{0\} \text{ here}$$

We form

$$R = (\hat{f} - k)^T [C(X^T \hat{V}^{-1} X)^{-1} C^T]^{-1} (\hat{f} - k) \qquad (3.32a)$$

$$\hat{f} = C\hat{\beta} \qquad (3.32b)$$

If the null hypothesis is true this is distributed as approximately χ^2 with r degrees of freedom. Note that the term $(X^T \hat{V}^{-1} X)^{-1}$ is the estimated covariance matrix of the fixed coefficients.

If we find a statistically significant result we may wish to explore which particular linear combinations of the coefficients involved are significantly different from zero. The common instance of this is where we find that n groups differ and we wish to carry out all possible pairwise comparisons. A simultaneous comparisons procedure which maintains the overall type I error at the specified level involves carrying out the above procedure with either a subset of the rows of C or a set of (less than r) linearly independent contrasts. The value of R obtained is then judged against the critical values of the chi-squared distribution with r degrees of freedom.

We can also obtain an α% confidence region for the parameters by setting $\hat{R}$ equal to the α% tail region of the χ^2 distribution with r degrees of freedom in the expression

$$\hat{R} = (f - \hat{f})^T [C(X^T \hat{V}^{-1} X)^{-1} C^T]^{-1} (f - \hat{f}) \quad (3.33)$$

This yields a quadratic function of the estimated coefficients, giving an *r*-dimensional ellipsoidal region.

In some situations we may be interested in separate confidence intervals for all possible linear functions involving a subset of ***q*** parameters or ***q*** linearly independent functions of the parameters, while maintaining a fixed probability that all the intervals include the population value of these functions of the parameters. As before, this may arise when we have an explanatory variable with several categories and we are interested in intervals for sets of contrasts. For a $(1-\alpha)\%$ interval write C_i for the *i*-th row of *C,* then a simultaneous $(1-\alpha)\%$ interval for $C_i\beta$, for all C_i is given by

$$(C_i\hat{\beta} - d_i, C_i\hat{\beta} + d_i) \quad (3.34)$$

where

$$d_i = [C_i (X^T \hat{V}^{-1} X)^{-1} C_i^T \chi^2_{q,(\alpha)}]^{0.5} \quad (3.35)$$

where $\chi^2_{q,(\alpha)}$ is the $\alpha\%$ point of the χ^2_q distribution.

We can also use the likelihood ratio test criterion for testing hypotheses about the fixed parameters, although generally the results will be similar. The difference arises because the random parameter estimates used in (3.32a) and (3.32b) are those obtained for the full model rather than those under the null hypothesis assumption, although this modification can easily be made. We shall discuss the likelihood ratio test in the next section dealing with the random parameters.

Random parameters

In very large samples it is possible to use the same procedures for hypothesis testing and confidence intervals as for the fixed parameters. Generally, however, procedures based upon the likelihood statistic are preferable. To test a null hypothesis H_0 against an alternative H_1 involving the fitting of additional parameters we form the log likelihood ratio or deviance statistic

$$D_{01} = -2\log_e(\lambda_0 / \lambda_1) \quad (3.36)$$

where λ_0, λ_1 are the likelihoods for the null and alternative hypotheses and this is referred to tables of the chi-squared distribution with degrees of freedom equal to the difference (q) in the number of parameters fitted under the two models.

We can also use (3.36) as the basis for constructing a $(1-\alpha)\%$ confidence region for the additional parameters. If D_{01} is set to the value of the$\alpha\%$ point of the chi squared distribution with q degrees of freedom, then a region is constructed to satisfy (3.36), using a suitable search procedure. This is a computationally intensive task, however, since all the parameter estimates are recomputed for each search point.

An alternative is to use the 'profile likelihood' (McCullagh & Nelder, 1989). In this case the likelihood is computed for a suitable region containing values of the random parameters of interest, for fixed values of the remaining random parameters. Interval estimates can be provided also by bootstrap simulations.

Residuals

In studies of institutions (e.g. schools, hospitals etc) effectiveness (Goldstein & Spiegelhalter, 1996), one requirement is sometimes to try to identify institutions with residuals which are substantially different. From a significance testing standpoint, we will often be interested in the null hypothesis that institution (group) A has a smaller residual than institution (group) B against the alternative that the residual for institution (group) A is larger than that for institution (group) B (ignoring the vanishingly small probability that they are equal). In the case when a standard significance test accepts the alternative hypothesis (at a chosen level) of some difference against the null hypothesis of no difference, this is equivalent to accepting one of the alternatives (A > B, A < B) at the same level of significance and we shall use this interpretation.

Where we can identify two particular institutions (groups) then it is straightforward to construct a confidence interval for their difference or carry out a significance test. Often, however, the results are made available to a number of individuals, each of whom are interested in comparing their own institutions (e.g. schools) of interest. This may occur, for example where policy makers wish to select a few schools within a small geographical area for comparison, out of a much larger study. In the following discussion, we suppose that individuals wish to compare only pairs of institutions, although the procedure can be extended to multiple comparisons of three or more residuals. Further details are given by Goldstein & Healy (1994). When the sample size of a study is fairly large, we can assume that the estimated residuals together with their comparative standard errors estimates are uncorrelated.

First, we order the residuals from smallest to largest. We construct an interval about each residual so that the criterion for judging statistical significance at the $(1-\alpha)\%$level for any pair of residuals is whether their confidence intervals overlap. For example, if we consider a pair of residuals with a common standard error (se) and assuming Normality, the confidence interval width for judging a difference significant at the 5% level are given by $\pm 1.39(\mathrm{se})$. The general procedure defines a set of confidence intervals for each residual *i* as

$$\hat{u}_i \pm c(\mathrm{se})_i \quad (3.37)$$

For each possible pair of intervals, (3.37) there is a significance level associated with the overlap criterion, and the value ***c*** is determined so that the average, over all possible pairs is $(1-\alpha)\%$. A search procedure can be devised to determine ***c***. When the ratios of the standard errors do not vary appreciably, say by not more than 2:1, the value 1.4 can be used for ***c***. As this ratio increases so does the value of ***c***.

These kinds of residual analyses are useful for conveying the inherent uncertainty associated with estimates for individual level 2 (or higher) units, where the number of level 1 units per higher-level unit is not large. This uncertainty in turn places inherent limitations upon such comparisons.

3.2 Extensions of the 2-Level Linear Model

What we have discussed so far refers to notations, techniques and estimations for the two-level linear model, which is the most common case in the multilevel analysis theory. However, in order to examine more demanded applications presented in the next chapter, we need to present the logical extensions of the two-level linear model. In all cases discussed here, the extensions are straightforward and stem either from the hierarchy of the subjects or from the nature of the data that are being measured. The extensions discussed here are:

- The 3-Level linear Model
- Cross-Classification Models
- Models for Discrete response data – The Proportions as responses case

- Multivariate Multilevel Models – The basic 2-level Multivariate model
- Multilevel Structural Equation Models – Multilevel Factor Analysis case

3.2.1 The Three-Level Linear Model

The most profound, maybe, extension of a 2-Level linear model comes when we add more levels of hierarchy in the model. We focus on the 3-level model since higher-level cases are rarely of importance in practice. Some examples of 3-level hierarchical structures are students (Level-1) nested within schools (Level-2) nested within prefectures (Level-3). Or in another point of view repeated visits (Level-1) of patients (Level-2) in heath provider units (Level-3).

In the simplest case the basic linear 3-level model can be written as follows:

$$y_{ijk} = \beta x_{ijk} + (v_k + u_{jk} + e_{ijk}) \qquad (3.38)$$

where x_{ijk} is a vector of covariates and $\boldsymbol{\beta}$ a corresponding vector of parameter estimates. The vector of covariates includes a constant together with explanatory variables measured at any of the three levels. The error terms v_k, u_{jk} and e_{ijk} are considered are considered as random variables with mean zero and variances

$$\mathrm{var}(v_k) = \sigma_v^2 \quad (3.39a)$$

$$\mathrm{var}(u_{jk}) = \sigma_u^2 \quad (3.39b)$$

$$\mathrm{var}(e_{ijk}) = \sigma_e^2 \quad (3.39c)$$

If we now introduce Z explanatory variables in the random part of the model, in any of the three levels, we obtain the more general form of the 3-level model, as follows:

$$y_{ijk} = X_{ijk}\beta + \sum_{h=0}^{q_3} v_{hk} z_{hk}^{(3)} + \sum_{h=0}^{q_2} u_{hjk} z_{hjk}^{(2)} + \sum_{h=0}^{q_1} e_{hijk} z_{hijk}^{(1)} \qquad (3.40)$$

where x_{ijk} is again the vector of covariates, $\boldsymbol{\beta}$ the corresponding vector of parameter estimates, and $z_{hk}^{(3)}, z_{hjk}^{(2)}$ and $z_{hijk}^{(1)}$ the explanatory variables of the random part of the 3rd, 2nd and 1st level of hierarchy, respectively.

Although such models seem more complicated and demanding than the two-level models, the computations, estimation techniques and algorithms are totally analogous to the methods described before for the two-levels case.

3.2.2 Cross-Classified Models

So far we have considered only data where the units have a purely hierarchical or nested structure. In many cases, however, a unit may be classified along more than one dimension. An example is students classified both by the school they attend and by the neighbourhood where they live. This is diagrammatically represented as follows for three schools and four neighbourhoods with between one and six students per school/neighbourhood cell. The cross classification is at level 2 with students at level 1.

Table 3.2: A random cross-classification at level 2

	School 1	School 2	School 3
Neighbourhood 1	x x x x	x x	x
Neighbourhood 2	x	x x x x x x	x x x
Neighbourhood 3	x x	x	x x x x
Neighbourhood 4	x x x	x x	x x

Another example is in a repeated measures study where children are measured by different raters at different occasions. If each child has its own set of raters not shared with other children then the cross classification is at level 1, occasions by raters, nested within children at level 2. We note that, by definition, a level 1 cross classification has only one unit per cell.

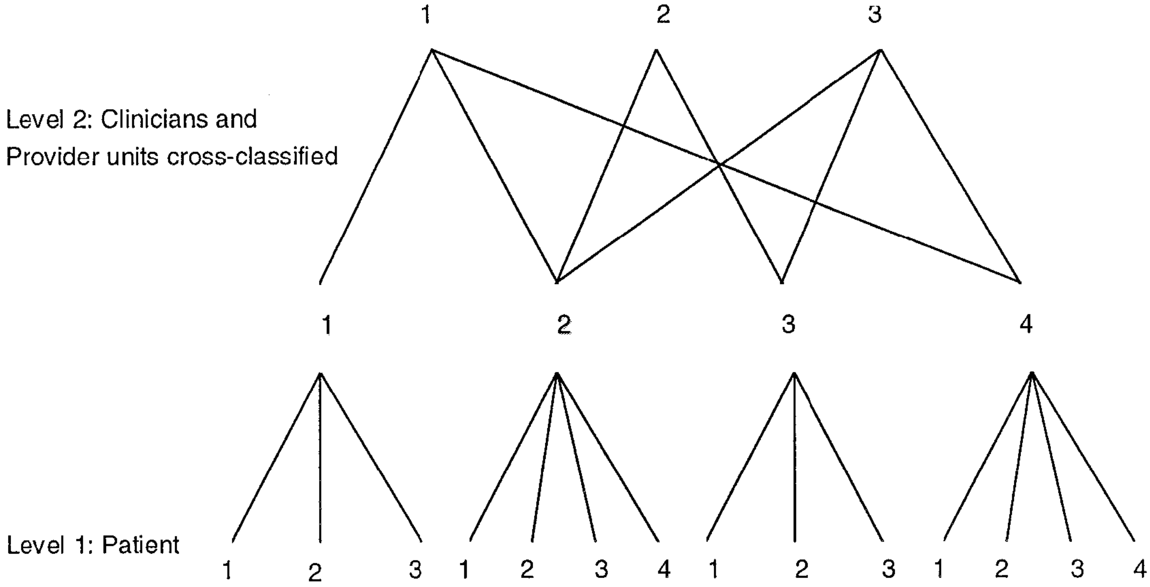
These basic cross-classifications occur commonly when a simple hierarchical structure breaks down in practice. Consider, for example, a repeated measures design, which follows a sample of students over time, say once a year, within a set of classes for a single school. If students change classes during the course, that is a cross classification at level 2 for classes by students. If we now include schools these will be classified as level 3 units, but if students also change schools during the course of the study then we obtain a level 3 cross classification of students by schools with classes nested at level 2 within schools and occasions as the level 1 units. The students have moved from being crossed with classes to being crossed with schools. Note that since students are crossed at level 3 with schools they are also automatically crossed with any units nested within schools and we do not need separately to specify the crossing of classes with students.

Such designs will occur also in panel or longitudinal studies of individuals who move from one locality to another, or workers who change their place of employment.

Other examples of such designs occur in panel studies of households where, over time, some households split up and form new households. The total set of all households is crossed with individual at level 2 with occasion at level 1. The households, which remain intact for more than one occasion, provide the information for estimating level 1 variation.

In health studies cross-classification occurs naturally in many cases. Consider for example the case where patients may be classified both by the hospitals they visit and by the clinicians the frequent, so that individuals within one hospital cluster are not grouped in the same way under clinicians. This type of cross-classification does not occur when clinicians operate within a single medical care, but this is not always the case. This kind of cross-classification is illustrated diagramatically in the following figure (Rice & Jones, 1997):

Figure 3.5: Patients within Cross-classified clinicians and Provider Units



In another example, patients may receive care from more that one medical centre during the year. This arrangement forms a multiple or cross-unit membership model, a special case of cross-classification (Carey, 2000).

We now set out the structure of the basic models described above and then go on to consider extensions and special cases of interest.

A basic cross-classified model

We consider first the simple model of Table 3.2 with variance components at level 2 and a single variance term at level 1.

We shall refer to the two classifications at level 2 using the subscripts j_1, j_2 and in general parentheses will group classifications at the same level. We write the model as

$$y_{i(j_1 j_2)} = X_{i(j_1 j_2)}\beta + u_{j_1} + u_{j_2} + e_{i(j_1 j_2)} \qquad (3.41)$$

The covariance structure at level 2 can be written in the following form

$$\operatorname{cov}(y_{i(j_1 j_2)} y_{i'(j_1 j'_2)}) = \sigma^2_{u_1} \qquad (3.42a)$$

$$\operatorname{cov}(y_{i(j_1 j_2)} y_{i'(j'_1 j_2)}) = \sigma^2_{u_2} \qquad (3.42b)$$

$$\operatorname{var}(y_{i(j_1 j_2)}) = \operatorname{cov}(y_{i(j_1 j_2)} y_{i'(j_1 j_2)}) = \sigma^2_{u_1} + \sigma^2_{u_2} \quad (3.42c)$$

Note that if there is no more than one unit per cell, then model (3.41) is still valid and can be used to specify a level 1 cross classification.

Thus the level-2 variance is the sum of the separate classification variances, the covariance for two level 1 units in the same classification is equal to the variance for that classification and the covariance for two level 1 units, which do not share either classification, is zero. If we have a model where random coefficients are included for either or both classifications, then analogous structures are obtained. We can also add further ways of classification with obvious extensions to the covariance structure.

We can now show how cross-classified models can be specified and estimated efficiently using a purely hierarchical formulation, including random cross-classified structures.

We illustrate the procedure using a 2-level model with crossing at level 2. The 2-level cross-classified model, using the same notations as in previous chapters for the basic model, can be written

$$y_{i(j_1 j_2)} = X_{i(j_1 j_2)}\beta + \sum_{h=1}^{q_1} z_{1hij_1} u_{1hj_1} + \sum_{h=1}^{q_2} z_{2hij_2} u_{2hj_2} + e_{i(j_1 j_2)} \qquad (3.43)$$

Parentheses group the ways of classification at each level. We have two sets of explanatory variables, type 1 and type 2, for the random components defined by the columns of $Z_1(n \times p_1 q_1)$, $\mathrm{Z}_2(n \times p_2 q_2)$ where p_1, p_2 are respectively the number of categories of each classification, i.e.

$$Z_1 = \{z_{1hij_1}\} \qquad (3.44a)$$

where $z_{1hij_1} = z_{1him} \quad if \quad j_1 = m, \;\; for \;\; m-th\, type\, 1\, level\, 2\, unit, 0\, otherwise$

and

$$Z_2 = \{z_{2hij_2}\} \qquad (3.44b)$$

where $z_{2hij_2} = z_{2him} \quad if \quad j_2 = m, \;\; for \;\; m-th\, type\, 2\, level\, 2\, unit, 0\, otherwise$

These variables are dummy variables where for each level 2 unit of type 1 we have q_1 random coefficients with covariance matrix $\Omega_{(1)2}$ and likewise for the type 2 units. To simplify the exposition we restrict ourselves to the variance component case where we have

$$\Omega_{(1)2} = \sigma^2_{(1)2}, \;\; \Omega_{(2)2} = \sigma^2_{(2)2} \qquad (3.45a)$$

$$E(\tilde{Y}\tilde{Y}^T) = V_1 + Z_1(\sigma^2_{(1)2} I_{(p_1)}) Z_1^T + Z_2(\sigma^2_{(2)2} I_{(p_2)}) Z_2^T \quad (3.45b)$$

It is clear that the second term in (3.45b) can be written as

$$Z_1(\sigma^2_{(1)2} I_{(p_1)}) Z_1^T = J \sigma^2_{(1)2} J^T \qquad (3.46)$$

where J is a (n x 1) vector of ones. The third term is of the general form $Z_3 \Omega_3 Z_3^T$, namely a level 3 contribution where in this case there is only a single level 3 unit and with no covariances between the random coefficients of the Z_{2h} and with the variance terms constrained to be equal to a single value, $\sigma^2_{(2)2}$.

More generally we can specify a level 2 cross classified variance components model by modeling one of the classifications as a standard hierarchical component and the second as a set of dummy explanatory variables, one for each category, with the random coefficients uncorrelated and with variances constrained to be equal. We can summarize the procedure using the simple model of (3.41). We specify one of the classifications, most efficiently the one with the larger number of units, as a standard hierarchical level 2 classification. For the other classification we define a dummy (0,1) variable for each unit, which is one if the observation belongs to that unit and zero if not. Then we specify that each of these dummy variables has a coefficient random at level 3 and in addition constrain the resulting set of level 3 variances to be equal. The variance estimate obtained is that required for this classification and the level-2 variance for the other classification is the one we require for that.

To extend this to further ways of classification we add levels. Thus, for a three way cross classification at level 2 we choose one classification, typically that with the largest number of categories, to model in standard hierarchical fashion at level 2, the second to model with coefficients random at level 3 as above and the third to model in a similar fashion with coefficients random at level 4. So we can obtain the third variance by defining a similar set of dummy variables with coefficients varying at level 4 and variances constrained to be equal. This procedure generalizes straightforwardly to sets of several random coefficients for each classification, with dummy variables defined as the products of the basic (0,1) dummy variables used in the variance components case and with corresponding variances and covariances constrained to be equal within classifications. In general a p-way cross classification at any level can be modeled by inserting sets of random variables at the next p-1 higher levels. Thus in a 2-level model with two crossed classifications at level 1 we would obtain a three level model with the original level 2 at level 3 and the level 1 cross classifications occupying levels 1 and 2.

Interactions in cross-classifications

If the second (type-2) classification has further explanatory variables with random coefficients as in (3.43) then we form extended dummy variable ‘interactions’ as the product of the basic dummy variables and the further explanatory variables with random coefficients, so that these coefficients have variances and covariances within the same type-2 level-2 unit but not across units. In addition the corresponding variances and covariances are constrained to be equal. We illustrate this case using the simple model with variance components at level-2 and a single variance term at level-1 (3.41). Consider the following extension of equation (3.41)

$$y_{i(j_1j_2)} = X_{i(j_1j_2)}\beta + u_{j_1} + u_{j_2} + u_{(j_1j_2)} + e_{i(j_1j_2)} \quad (3.47)$$

We have now added an 'interaction' term to the model which was previously an additive one for the two variances. The usual specification for such a random interaction term is that it has simple variance $\sigma^2_{u_{(12)}}$ across all the level 2 cells (Searle et al, 1992). To fit such a model we would define each cell of the cross classification as a level 2 unit with a between cell variance $\sigma^2_{u_{(12)}}$, a single level 3 unit with a variance σ^2_{u1} and a single level 4 unit with a variance σ^2_{u2}. The adequacy of such a model can be tested against an additive model using a likelihood ratio test criterion.

Extensions to this model are possible by adding random coefficients for the interaction component, just as random coefficients can be added to the additive components.

Level 1 cross classifications

Some interesting models occur when units are basically cross-classified at level 1. By definition we have a design with only one unit per cell, and we can also have a level 2 cross classification which is formally equivalent to a level 1 cross-classification where there is just one unit per cell. This case should be distinguished from the case where a level 2 cross classification happens to produce no more than 1 level 1 unit in a cell as a result of sampling, so that the confounding occurs by chance rather than by design.

A 2-level variance components model with a cross classification at level 1 can be written as

$$y_{(i_1 i_2)j} = X_{(i_1 i_2)j}\beta + u_j + e_{i_1 j} + e_{i_2 j} + e_{(i_1 i_2)j} \qquad (3.48)$$

where for level 1 we use a straightforward extension of the notation for a level 2 cross-classification. The term $e_{(i_1,i_2)j}$ is analogous to the interaction term in (3.47). To specify this model we would define the u_j as random at level 4, the $e_{i_1 j},\ e_{i_2 j}$ as random at levels 3 and 2, each with a single unit and the interaction term random across the cells of the cross classification at level 1, within the original level 2 units.

Suppose now that we were able to extend the design by replicating measurements for each cell of the level 1 cross-classification. Then (3.48) would refer to a 3-level model with replications as level 1 units, and which could be written as follows where the subscript *h* denotes replications

$$y_{h(i_1 i_2)j} = X_{h(i_1 i_2)j}\beta + u_j + e_{i_1 j} + e_{i_2 j} + e_{h(i_1 i_2)j} \qquad (3.49)$$

Since (3.48) is just model (3.49) with one unit per cell, we could interpret the 'interaction' variance in (3.48) as an estimate of the extent to which the additive variances of the cross-classification fail to account for the total level 1 variance.

So called 'generalisability theory' models (Cronbach & Webb, 1975) can be formulated as level-1 cross-classifications. The basic model is one where a test or other instrument consisting of a set of items, for example ratings or questions, is administered to a sample of individuals. The individuals are therefore cross-classified by the items at level 1 and may be further nested within schools etc. at higher levels. In educational test settings the item responses are often binary so that we would apply the methods discussed in next chapter (3.5.3.) to the present procedures in a straightforward way. Since each individual can only respond once to each item this an example of a genuine level 1 cross classification.

Another extension to what we have discussed so far is to allow simultaneous crossing at more than one level. However, the approach of such structures is totally analogous. Thus for example, if there is a 2-way cross classification at level 1 and a 3-way cross classification at level 2, we will require five levels, the first two describing the level 1 cross classification and the next three describing the level 2 cross classification.

Cross-unit membership models

In some circumstances units can be members of more than one higher-level unit at the same time. An example is friendship patterns where at any time individuals can be members of more than one friendship group. Another example is where children belong to more than one 'extended' family, which includes aunts and uncles as well as parents. In an educational system students may attend more than one institution. In all such cases we shall assume that for each higher level unit to which a lower level unit belongs there is a known weight, summing to 1.0 for each lower level unit, which represents, for example, the amount of time spent in that unit. We may also have data where, although there is no cross-unit membership, there is some uncertainty about which higher-level unit some lower level units belong to. For example, in a survey of students information about their neighbourhood of residence may only be available for a few students for larger geographical units. For these cases it may be possible to assign a weight for each of the constituent neighbourhoods, which is in effect a probability of belonging to each based upon available information.

Such a structure can be analyzed formally as a cross-unit membership model with most students having a single weight of 1.0 and the remainder zero.

Consider the 2-level variance components model (3.41) with each level 1 unit belonging to at most two level-2 units where the j_1, j_2 subscripts now refer to the same type of unit:

$$y_{i(j_1 j_2)} = X_{i(j_1 j_2)}\beta + w_{1ij_1}u_{j_1} + w_{2ij_2}u_{j_2} + e_{i(j_1 j_2)} \quad (3.50)$$

where $w_{1ij_1} + w_{2ij_2} = 1$.

The overall contribution at level 2 is therefore the weighted sum over the level 2 units to which each level 1 unit belongs. This leads to the following covariance structure

$$\mathrm{var}(y_{i(j_1 j_2)}) = (w_{1ij_1}^2 + w_{2ij_2}^2)\sigma_u^2 + \sigma_e^2 \qquad (3.51a)$$

$$\mathrm{cov}(y_{i(j_1 j_2)} y_{i'(j_1 j_2)}) = (w_{1ij_1} w_{1i'j_1} + w_{2ij_2} w_{2i'j_2})\sigma_u^2 \; (3.51b)$$

$$\mathrm{cov}(y_{i(j_1 j_2)} y_{i'(j_1' j_2)}) = w_{2ij_2} w_{2i'j_2}\sigma_u^2 \qquad (3.51c)$$

This has the structure of a standard 2-level cross-classified model with the additional constraint $\sigma_{u1}^2 = \sigma_{u2}^2 = \sigma_u^2$ and where the explanatory indicator variables Z_1, Z_2 described in (3.44a) and (3.44b) have the value 1 replaced by the relevant weights for each level 1 unit. As with the standard cross-classification this model can be extended to include random coefficients and general p-unit membership. In this case we need only in fact specify a single level 2 unit with explanatory variable design matrix *Z*, containing dummy weight vectors, and Ω_u as diagonal of order equal to the number of level 2 units, and elements equal to σ_u^2.

In the next Chapter we introduce examples where the cross-classified structure has to be taken seriously into account.

3.2.3 Models for Discrete response data – The Proportions as responses case

All the models of previous chapters have assumed that the response variable is continuously distributed. We now look at data where the response is essentially a count of events. This count may be the number of times an event occurs out of a fixed number of 'trials' in which case we usually deal with the resulting proportion as response: an example is the proportion of deaths in a population, classified by age.

We may have a vector of counts representing the numbers of events of different kinds which occur out of a total number of events: an example is the number of responses to each, ordered, category of a question on abortion attitudes. In all these cases the assumption of normality for the response variable is clearly violated as well as the assumption of homoscedastic error terms.

The approach to the problem of non-normally distributed variables is to include the necessary transformation and the choice of the appropriate error distribution (not necessarily a normal distribution) explicitly in the statistical model. Statistical models for such data are referred to as 'generalized linear models' (McCullagh & Nelder, 1989). A 2-level model can be written in the general form

$$\pi_{ij} = f(X_{ij}\beta_j) \quad (3.52)$$

where π_{ij} is the expected value of the response for the *ij*-th level 1 unit and *f* is a nonlinear function (called the 'link function) of the 'linear predictor' $X_{ij}\beta_j$. Note that we allow random coefficients at level 2. The model is completed by specifying a distribution for the *observed* response $y_{ij}|\pi_{ij}$. Where the response is a proportion this is typically taken to be binomial and where the response is a count taken to be Poisson. Equation (3.52) is a special case of the nonlinear models and the estimation methods for fixed and random part are extensions to those of linear models, using the appropriate transformations (e.g. the Taylor series expansion). However, what remains is to specify the nonlinear 'link' function *f*. Table 3.3 lists some of the standard choices, with logarithms chosen to base *e*. In addition to these we can also have the 'identity' function $f^{-1}(\pi) = \pi$, but this can create difficulties since it allows, in principle, predicted counts or proportions which are respectively less than zero or outside the range (0,1). Nevertheless, in many cases, using the identity function produces acceptable results, which may differ little from those obtained with the nonlinear functions. For the purpose of the thesis we consider only the type of model where responses are proportions in dichotomous variables.

Table 3.3: Some nonlinear link functions

Response	$f^{-1}(\pi)$	**Name**
Proportion	$\log\{(\pi)/(1-\pi)\}$	logit
Proportion	$\log\{-\log(1-\pi)\}$	complementary log log

Vector of proportions	$\log(\pi_s / \pi_t) \;\; (s = 1, \ldots, t-1)$	multivariate logit
Count	$\log(\pi)$	log

Proportions as responses

Consider the 2-level variance components model with a single explanatory variable where the expected proportion is modeled using a logit link function

$$\pi_{ij} = \{1 + \exp(-[\beta_0 + \beta_1 x_{1ij} + u_{0j}])\}^{-1} \qquad (3.53)$$

The observed responses y_{ij} are proportions with the standard assumption that they are binomially distributed

$$y_{ij} \sim Bin(\pi_{ij}, n_{ij})$$

where n_{ij} is the denominator for the proportion. We also have

$$\mathrm{var}(y_{ij} | \pi_{ij}) = \pi_{ij}(1 - \pi_{ij}) / n_{ij} \quad (3.54)$$

We now write the model in the standard way including the level 1 variation as

$$y_{ij} = \pi_{ij} + e_{ij} z_{ij}, \quad z_{ij} = \sqrt{\pi_{ij}\left(1 - \pi_{ij}\right) / n_{ij}}, \quad \sigma_e^2 = 1 \quad (3.55)$$

Using this explanatory variable Z and constraining the level 1 variance associated with this to be one we obtain the required binomial variance in equation (3.54). When fitting a model we can also allow the level-1 variance to be estimated and by comparing the estimated variance with the value 1.0 obtain a test for 'extra binomial' variation. Such variation may arise in a number of ways.

Estimation methods for both fixed and random parameters for a proportion as response model is a demanding procedure and therefore will not be discussed further. Goldstein (1995) refers to the distinction between 'predictive quasilikelihood' (PQL) and 'marginal quasilikelihood' (MQL) estimation procedures (Breslow & Clayton, 1993). In many applications the MQL procedure will tend to underestimate the values of both the fixed and random parameters, especially where n_{ij} is small. When the sample size is small the unbiased (RIGLS, REML) procedure should be used.

In the next Chapter we focus more on examples where proportions are used as the response variables and generalized linear models in the form discussed here are used.

3.2.4 Multivariate Multilevel Models – The basic 2-level Multivariate model

Multivariate Multilevel models

In the models discussed so far we have considered only a single response variable, either normal or not, measured at the first level of hierarchy. We now look at models where we wish simultaneously to model several responses as functions of explanatory variables. As we shall see, the ability to do this provides us with tools for tackling a very wide range of problems. These problems include missing data, rotation or matrix designs for surveys and prediction models.

We develop the model considering the case of two response variables measured at the individual level while explanatory variables are measured at all levels of hierarchy.

The basic 2-level multivariate model

To define a multivariate, in this case a 2-variate, model we treat the individuals as a level 2 unit and the 'within-individuals' measurements as level 1 units. Each level 1 measurement 'record' has a response, which is either the first or the second variable. 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 a dichotomous variable with values 1 and 0. The model is written as

$$y_{ij} = \beta_{01} z_{1ij} + \beta_{02} z_{2ij} + \beta_{11} z_{1ij} x_j + \beta_{12} z_{2ij} x_j + u_{01j} + u_{02j} \quad (3.56)$$

where

$$z_{1ij} = \begin{Bmatrix} 1 \text{ if 1st variable is present} \\ 0 \text{ if 2nd variable is present} \end{Bmatrix}, \quad z_{2ij} = 1 - z_{1ij}, \quad x_j = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix} \quad (3.57)$$

and

$$\text{var}(u_{01j}) = \sigma_{01}^2 \quad (3.58a)$$

$$\text{var}(u_{02j}) = \sigma_{02}^2 \quad (3.58b)$$

$$\text{cov}(u_{01j} u_{02j}) = \sigma_{012} \quad (3.58c)$$

There are several 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-individuals variances. In the case where only the intercept dummy variables are fitted, and since every individual has scores for both the response variables, the model estimates of these parameters become the usual between-subjects estimates of the variances and covariance. The multilevel estimates are statistically efficient even where some responses are missing, and in the case where the measurements have a multivariate Normal distribution they are maximum likelihood. Thus the formulation as a 2-level model allows for the efficient estimation of a covariance matrix with missing responses.

We can further allow the individuals to be grouped within level 3 units and therefore add more variability terms in the 3-rd level of the model. These can be variances and a covariance for the two components added at level 3 as well as additional variance terms for the second level explanatory variables.

Designs with subsets of responses – Missing cases

We have already seen that fully balanced multivariate designs are unnecessary and randomly missing responses are handled automatically. The basic 2-level formulation does not formally recognize that a response is missing, since we only record those present. We now look at designs where responses are effectively missing by design and we see how this can be useful in a number of circumstances.

In many kinds of surveys the amount of information required from respondents is so large that it is too onerous to expect each one to respond to all the questions or items. In education we may require achievement information covering a large number of areas, in surveys of businesses we may wish to have a large amount of detailed information, and in household questionnaires we may wish to obtain information on a wide range of topics. We consider only measurements that are used as responses in a model. If we denote the total set of responses as $\{N\}$ then we choose *p* subsets $\{N_i, i = 1, \dots p\}$ each of which is suitable for administering to a subject (level 1).

When choosing these subsets we can only estimate subject-level covariances between those responses that appear together in a subtest. It is therefore common in such designs to ensure that every possible pair of responses is present. If we wish to estimate covariances for higher-level units such as schools it is necessary only to ensure that the relevant pair of responses are assigned to the some schools - a large enough number to provide efficient estimates. The subjects are assigned at random to subtest and higher-level units are also assigned randomly, possibly with stratification. Each subset is viewed formally as a multivariate response vector with randomly missing values, although the missing observations are produced by design. We can then fit a multivariate response model for such data and obtain efficient estimates for the fixed part coefficients and covariance structures at any level. In this formulation, the variables to be used as explanatory variables should be measured for each level 1 unit.

Multivariate cross-classified models

We know consider the multivariate case of a type of models discussed previously – the cross-classified models. For multivariate models the responses may have different structures. Thus in a bivariate model one response may have a 2-level hierarchical structure and the other may have a cross classification at level 2. Suppose, for example that we measure the height and the mathematics attainment of a sample of students from a sample of schools. The mathematics attainment is assessed by a different set of teachers in each school and the heights are measured by a single anthropometrist. For the mathematics scores there is a level-1 cross-classification of students within each school whereas for height there is a 2-level hierarchy with students nested within schools. Height and mathematics attainment will be correlated at both the student and the school level and we can write a model for this structure as follows

$$y_{h(i_1i_2)j} = \delta_{1h}(X_{1(i_1i_2)j}\beta_1 + u_{1j} + e_{1i_1j} + e_{1i_2j}) + \delta_{2h}(X_{2i_1j}\beta_2 + u_{2j} + e_{2i_1j}) \quad (3.59)$$

where

$$\delta_{1h} = 1\, if \;\; mathematics,\; 0\, if \;\; height, \quad \delta_{2h} = 1 - \delta_{1h} \qquad (3.60)$$

and

$$\operatorname{cov}(u_{1j}u_{2j}) = \sigma_{u12} \qquad (3.61a)$$

$$\operatorname{cov}(e_{1i_1j}e_{2i_1j}) = \sigma_{e12} \qquad (3.61b)$$

and all other covariances are zero. This will therefore be specified as a 4-level model with the bivariate structure as level 1 and level 2 units being individual students. There will be a single level 3 unit with the coefficients of the dummy variables for teachers having variances random at this level, with level 4 being that of the school.

3.3 Conclusions of the Chapter

The conclusions drawn by the discussion of this Chapter can verify the theoretical advantage of Multilevel Models compared to other techniques. First of all, in simple situations they respect totally the hierarchy of the data and they end up to more precise estimations both for the model parameters and for the model variability. Secondly, all known statistical techniques and procedures (such as Maximum Likelihood, EM Algorithm, MCMC estimations etc) for statistical inferences (parameter estimates, testing functions, confidence intervals) can be easily used, making Multilevel Methods theoretically understandable and easily applicable for statisticians and researchers. Finally, the simple multilevel models can be readily extended to more sophisticated theoretical concepts, such as multivariate or generalized models, and therefore can be applied effectively to more complex situations. In the following Chapter we will discuss if the theoretical advantages of the Multilevel Techniques are also present in practice, or, in other words, why and how Multilevel Analysis is useful and effective in practical situations.