STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:3291–3315 (DOI: 10.1002/sim.1264)

TUTORIAL IN BIOSTATISTICS
Multilevel modelling of medical data

Harvey Goldstein*,†, William Browne and Jon Rasbash

Institute of Education, University of London, London, U.K.

SUMMARY

This tutorial presents an overview of multilevel or hierarchical data modelling and its applications in medicine. A description of the basic model for nested data is given and it is shown how this can be extended to fit flexible models for repeated measures data and more complex structures involving cross-classifications and multiple membership patterns within the software package MLwiN. A variety of response types are covered and both frequentist and Bayesian estimation methods are described. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: complex data structures; mixed model, multilevel model; random effects model; repeated measures

1. SCOPE OF TUTORIAL

The tutorial covers the following topics

1. The nature of multilevel models with examples.
2. Formal model specification for the basic Normal (nested structure) linear multilevel model with an example.
3. The MLwiN software.
4. More complex data structures: complex variance, multivariate models and cross-classified and multiple membership models.
5. Discrete response models, including Poisson, binomial and multinomial error distributions.
6. Specific application areas including survival models, repeated measures models, spatial models and meta analysis.
7. Estimation methods, including maximum and quasi likelihood, and MCMC.

Further information about multilevel modelling and software details can be obtained from the web site of the Multilevel Models Project, http://multilevel.ioe.ac.uk/.

* Correspondence to: Harvey Goldstein, Institute of Education, University of London, 20 Bedford Way, London WC1H OAL, U.K.
† E-mail: h.goldstein@ioe.ac.uk

Received June 2001
Accepted January 2002

Copyright © 2002 John Wiley & Sons, Ltd.

2. THE NATURE OF MULTILEVEL MODELS

Traditional statistical models were developed making certain assumptions about the nature of the dependency structure among the observed responses. Thus, in the simple regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ the standard assumption is that the y_i given x_i are independently identically distributed (i.i.d.), and the same assumption holds also for generalized linear models. In many real life situations, however, we have data structures, whether observed or by design, for which this assumption does not hold.

Suppose, for example, that the response variable is the birthweight of a baby and the predictor is, say, maternal age, and data are collected from a large number of maternity units located in different physical and social environments. We would expect that the maternity units would have different mean birthweights, so that knowledge of the maternity unit already conveys some information about the baby. A more suitable model for these data is now

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \tag{1}$$

where we have added another subscript to identify the maternity unit and included a unit-specific effect u_j to account for mean differences amongst units. If we assume that the maternity units are randomly sampled from a population of units, then the unit specific effect is a random variable and (1) becomes a simple example of a two-level model. Its complete specification, assuming Normality, can be written as follows:

$$\begin{gathered} y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \\ u_j \sim \mathrm{N}(0, \sigma_u^2), \;\; e_{ij} \sim \mathrm{N}(0, \sigma_e^2) \\ \mathrm{cov}(u_j, e_{ij}) = 0 \\ \mathrm{cov}(y_{i_1 j}, y_{i_2 j} \,|\, x_{ij}) = \sigma_u^2 \geqslant 0 \end{gathered} \tag{2}$$

where i_1, i_2 are two births in the same unit j with, in general, a positive covariance between the responses. This lack of independence, arising from two sources of variation at different levels of the data hierarchy (births and maternity units) contradicts the traditional linear model assumption and leads us to consider a new class of models. Model (2) can be elaborated in a number of directions, including the addition of further covariates or levels of nesting. An important direction is where the coefficient (and any further coefficients) is allowed to have a random distribution. Thus, for example the age relationship may vary across clinics and, with a slight generalization of notation, we may now write (2) as

$$\begin{gathered} y_{ij} = \beta_{0ij} x_{0ij} + \beta_{1j} x_{1ij} \\ \beta_{0ij} = \beta_0 + u_{0j} + e_{0ij} \\ \beta_{1j} = \beta_1 + u_{1j} \\ x_{0ij} = 1 \\ \mathrm{var}(u_{0j}) = \sigma_{u0}^2, \quad \mathrm{var}(u_{1j}) = \sigma_{u1}^2 \\ \mathrm{cov}(u_{0j} u_{1j}) = \sigma_{u01}, \quad \mathrm{var}(\mathrm{e}_{0ij}) = \sigma_{e0}^2 \end{gathered} \tag{3}$$

Copyright © 2002 John Wiley & Sons, Ltd.

and in later sections we shall introduce further elaborations. The regression coefficients β_0, β_1 are usually referred to as 'fixed parameters' of the model and the set of variances and covariances as the random parameters. Model (3) is often referred to as a 'random coefficient' or 'mixed' model.

At this point we note that we can introduce prior distributions for the parameters of (3), so allowing Bayesian models. We leave this topic, however, for a later section where we discuss MCMC estimation.

Another, instructive, example of a two-level data structure for which a multilevel model provides a powerful tool, is that of repeated measures data. If we measure the weight of a sample of babies after birth at successive times then the repeated occasion of measurement becomes the lowest level unit of a two-level hierarchy where the individual baby is the level-2 unit. In this case model (3) would provide a simple description with x_{1ij} being time or age. In practice linear growth will be an inadequate description and we would wish to fit at least a (spline) polynomial function, or perhaps a non-linear function where several coefficients varied randomly across individual babies, that is each baby has its own growth pattern. We shall return to this example in more detail later, but for now note that an important feature of such a characterization is that it makes no particular requirements for every baby to be measured at the same time points or for the time points to be equally spaced.

The development of techniques for specifying and fitting multilevel models since the mid-1980s has produced a very large class of useful models. These include models with discrete responses, multivariate models, survival models, time series models etc. In this tutorial we cannot cover the full range but will give references to existing and ongoing work that readers may find helpful. In addition the introductory book by Snijders and Bosker [1] and the edited collection of health applications by Leyland and Goldstein [2] may be found useful by readers.

A detailed introduction to the two-level model with worked examples and discussion of hypothesis tests and basic estimation techniques is given in an earlier tutorial [3] that also gives details of two computer packages, HLM and SAS, that can perform some of the analyses we describe in the present tutorial. The MLwiN software has been specifically developed for fitting very large and complex models, using both frequentist and Bayesian estimation and it is this particular set of features that we shall concentrate on.

3. MARGINAL VERSUS HIERARCHICAL MODELS

At this stage it is worth emphasizing the distinction between multilevel models and so-called 'marginal' models such as the GEE model [4, 5]. When dealing with hierarchical data these latter models typically start with a formulation for the covariance structure, for example, but not necessarily based upon a multilevel structure such as (3), and aim to provide estimates with acceptable properties only for the fixed parameters in the model, treating the existence of any random parameters as a necessary 'nuisance' and without providing explicit estimates for them. More specifically, the estimation procedures used in marginal models are known to have useful asymptotic properties in the case where the exact form of the random structure is unknown.

If interest lies only in the fixed parameters, marginal models may be useful. Even here, however, they may be inefficient if they utilize a covariance structure that is substantially incorrect. They are, however, generally more robust than multilevel models to serious mis-

Copyright © 2002 John Wiley & Sons, Ltd.

specification of the covariance structure [6]. Fundamentally, however, marginal models address different research questions. From a multilevel perspective, the failure explicitly to model the covariance structure of complex data is to ignore information about variability that, potentially, is as important as knowledge of the average or fixed effects. Thus, in the simple repeated measures example of baby weights, knowledge of how individual growth rates vary between babies, possibly differentially according to say demographic factors, will be important data and in a later section we will show how such information can be used to provide efficient predictions in the case of human growth.

When we discuss discrete response multilevel models we will show how to obtain information equivalent to that obtained from marginal models. Apart from that the remainder of this paper will be concerned with multilevel models. For a further discussion of the limitations of marginal models see the paper by Lindsey and Lambert [7].

4. ESTIMATION FOR THE MULTIVARIATE NORMAL MODEL

We write the general Normal two-level model as follows, with natural extensions to three or more levels:

$$\mathbf{Y}=\mathbf{X\beta}+\mathbf{E}$$

$$\mathbf{Y}=\{y_{ij}\},\quad \mathbf{X}=\{\mathbf{X_{ij}}\},\quad \mathbf{X_{ij}}=\{x_{0ij},x_{1ij},\ldots,x_{pij}\}$$

$$\mathbf{E}=\mathbf{E^{(2)}}+\mathbf{E^{(1)}}$$

$$\mathbf{E^{(2)}}=\{\mathbf{E_j^{(2)}}\},\quad \mathbf{E_j^{(2)}}=\mathbf{z_j^{(2)}e_j^{(2)}},\quad \mathbf{z_j^{(2)}}=\{\mathbf{z_{ij}^{(2)}}\}$$

$$\mathbf{z_{ij}^{(2)}}=\{z_{0j}^{(2)},z_{1j}^{(2)},\ldots,z_{q_2j}^{(2)}\},\quad \mathbf{e_j^{(2)}}=\{e_{0j}^{(2)},e_{1j}^{(2)},\ldots,e_{q_2j}^{(2)}\}^{\mathrm{T}}$$

$$\mathbf{E^{(1)}}=\{\mathbf{E_{ij}^{(1)}}\},\quad \mathbf{E_{ij}^{(1)}}=\mathbf{z_{ij}^{(1)}}\ \mathbf{e_{ij}^{(1)}}$$

$$\mathbf{z_{ij}^{(1)}}=\{z_{0j}^{(1)},z_{1j}^{(1)},\ldots,z_{q_1j}^{(1)}\},\quad \mathbf{e_{ij}^{(1)}}=\{e_{0ij}^{(1)},e_{1ij}^{(1)},\ldots,e_{q_1ij}^{(1)}\}^{\mathrm{T}}$$

$$\mathbf{e^{(2)}}=\{\mathbf{e_j^{(2)}}\},\quad \mathbf{e_j^{(1)}}=\{\mathbf{e_{ij}^{(1)}}\}$$

$$\mathbf{e^{(2)}}\sim\mathbf{N(0,\Omega_2)},\quad \mathbf{e_j^{(1)}}\sim\mathbf{N(0,\Omega_{1j})}$$

[Typically $\mathbf{\Omega_{1j}}=\mathbf{\Omega_1}$]

$$E(e_{hj}^{(2)}e_{h'j'}^{(2)})_{j\neq j'}=E(e_{hij}^{(1)}e_{h'i'j}^{(1)})_{i\neq i'}=E(e_{hj}^{(2)}e_{h'i'j'}^{(1)})=0$$

yields the block diagonal structure

$$\mathbf{V}=E(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^{\mathbf{T}})=\bigoplus_{\mathrm{j}}(\mathbf{V_{2j}}+\mathbf{V_{1j}})$$

$$\tilde{\mathbf{Y}}=\mathbf{Y}-\mathbf{X\beta} \tag{4}$$

$$\mathbf{V_{2j}}=\mathbf{z_j^{(2)}\Omega_2 z_j^{(2)^{\mathrm{T}}}},\quad \mathbf{V_{1j}}=\bigoplus_{i}\mathbf{z_{ij}^{(1)}\Omega_{1j}z_{ij}^{(1)^{\mathrm{T}}}}$$

Copyright © 2002 John Wiley & Sons, Ltd.

In this formulation we allow any number of random effects or coefficients at each level; we shall discuss the interpretation of multiple level-1 random coefficients in a later section.

A number of efficient algorithms are available for obtaining maximum likelihood (ML) estimates for (4). One [8] is an iterative generalized least squares procedure (IGLS) that will also produce restricted maximum likelihood estimates (RIGLS or REML) and is formally equivalent to a Fisher scoring algorithm [9]. Note that RIGLS or REML should be used in small samples to correct for the underestimation of IGLS variance estimates. The EM algorithm can also be used [10, 11]. Our examples use RIGLS (REML) estimates as implemented in the MlwiN software package [12] and we will also discuss Bayesian models. A simple description of the IGLS algorithm is as follows:

From (4) we have

$$\mathbf{V} = E(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^{\mathbf{T}}) = \bigoplus_{\mathrm{j}} (\mathbf{V_{2j}} + \mathbf{V_{1j}})$$

$$\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\beta$$

The IGLS algorithm proceeds by first carrying out a GLS estimation for the fixed parameters (β) using a working estimator of $\mathbf{V}$. The vectorized cross-product matrix of 'raw' residuals $\hat{\tilde{\mathbf{Y}}}\hat{\tilde{\mathbf{Y}}}^{\mathrm{T}}$ where $\hat{\tilde{\mathbf{Y}}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$, is then used as the response in a GLS estimation where the explanatory variable design matrix is determined by the last line of (4). This provides updated estimates for the Ω_{1j} and Ω_2 and hence V. The procedure is repeated until convergence. In the simple case we have been considering so far where the level 1 residuals are i.i.d., for a level-2 unit (individual) with just three level-1 units (occasions) there are just six distinct raw residual terms and the level-1 component $\mathbf{V_{1j}}$ is simply $\sigma_e^2\mathbf{I_3}$. Written as a vector of the lower triangle this becomes

$$\sigma_e^2 \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \tag{5}$$

and the vector of ones and zeroes becomes the level-1 explanatory variable for the GLS estimation, in this case providing the coefficient that is the estimator of σ_e^2. Similarly, for a model where there is a single variance term at level 2, the level-2 component $\mathbf{V_{2j}}$ written as a lower triangle vector is

$$\sigma_u^2 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Copyright © 2002 John Wiley & Sons, Ltd.

Goldstein [13] shows that this procedure produces maximum likelihood estimates under Normality.

5. THE MLwiN SOFTWARE

MLwiN has been under development since the late 1980s, first as a command-driven DOS based program, MLn, and since 1998 in a fully-fledged windows version, currently in release 1.10. It is produced by the Multilevel Models Project based within the Institute of Education, University of London, and supported largely by project funds from the U.K. Economic and Social Research Council. The software has been developed alongside advances in methodology and with the preparation of manuals and other training materials.

Procedures for fitting multilevel models are now available in several major software packages such as STATA, SAS and S-plus. In addition there are some special purpose packages, which are tailored to particular kinds of data or models. MIXOR provides ML estimation for multi-category responses and HLM is used widely for educational data. See Zhou *et al.* [14] for a recent review and Sullivan *et al.* [3] for a description of the use of HLM and SAS. Many of the models discussed here can also be fitted readily in the general purpose MCMC software package WinBUGS [15].

MLwiN has some particular advanced features that are not available in other packages and it also has a user interface designed for fully interactive use. In later sections we will illustrate some of the special features and models available in MLwiN but first give a simple illustration of the user interface. We shall assume that the user wishes to fit the simple two-level model given by (1).

In this tutorial we cannot describe all the features of MLwiN, but it does have general facilities for data editing, graphing, tabulation and simple statistical summaries, all of which can be accessed through drop-down menus. In addition it has a macro language, which can be used, for example, to run simulations or to carry out special purpose modelling. One of the main features is the method MLwiN uses to set up a model, via an 'equation window' in which the user specifies a model in more or less exactly the format it is usually written. Thus to specify model (1) the user would first open the equation window which, prior to any model being specified, would be as shown in Figure 1.

This is the default null model with a response that is Normal with fixed predictor represented by $X\beta$ and covariance matrix represented by Ω. Clicking on the **N** symbol delivers a drop

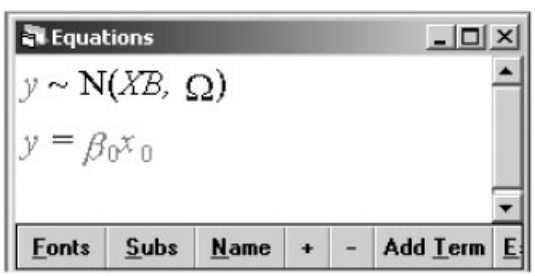

Figure 1. Default equation screen with model unspecified.

Copyright © 2002 John Wiley & Sons, Ltd.

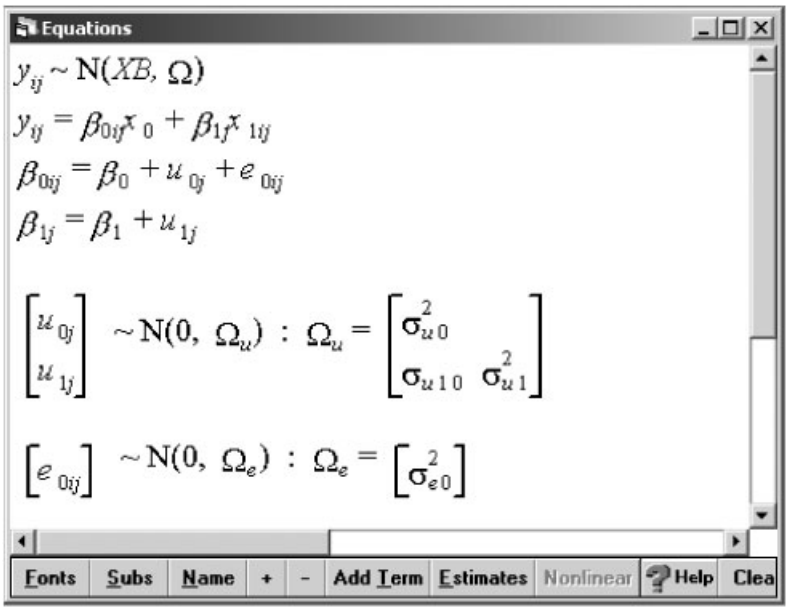

Figure 2. Equation screen with model display.

down menu, which allows the user to change the default distribution to binomial, Poisson or negative binomial. Clicking on the response y allows the user to identify the response variable from a list and also the number and identification for the hierarchical levels. Clicking on the x_0 term allows this to be selected from a list and also whether its coefficient β_0 is random at particular levels of the data hierarchy. Adding a further predictor term is also a simple matter of clicking an 'add term' button and selecting a variable. There are simple procedures for specifying general interaction terms.

Model (1), including a random coefficient for x_1 in its general form as given by (3), will be displayed in the equation window as shown in Figure 2. Clicking on the 'Estimates' button will toggle the parameters between their symbolic representations and the actual estimates after a run. Likewise, the 'Name' button will toggle actual variable names on and off. The 'Subs' button allows the user to specify the form of subscripts, for example giving them names such as in the screen shown in Figure 3, where we also show the estimates and standard errors from an iterative fit.

In the following sections we will show some further screen shots of models and results.

6. A GROWTH DATA EXAMPLE

We start with some simple repeated measures data and we shall use them to illustrate models of increasing complexity. The data set consists of nine measurements made on 26 boys between the ages of 11 and 13.5 years, approximately 3 months apart [16].

Copyright © 2002 John Wiley & Sons, Ltd.

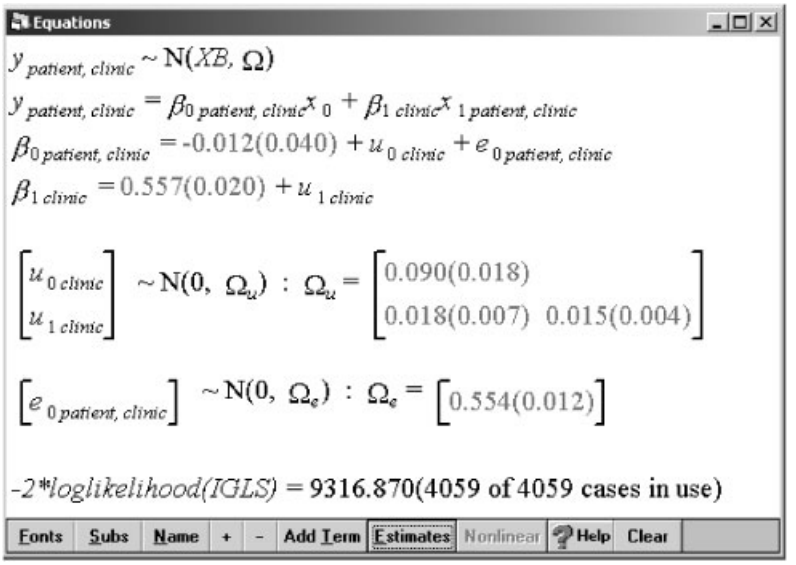

Figure 3. Equation screen with estimates.

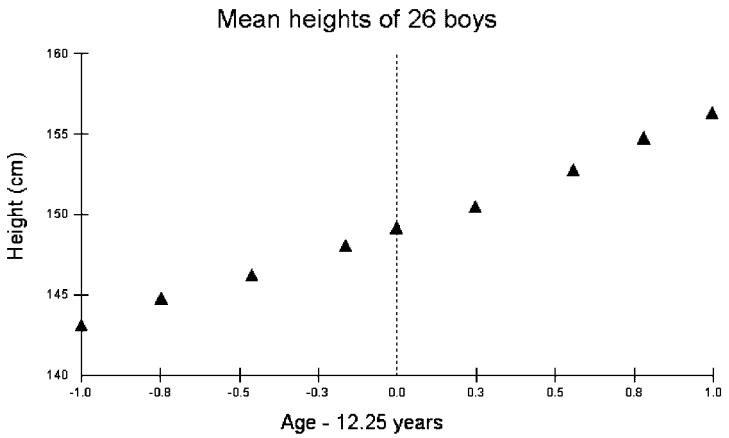

Figure 4.

Figure 4, produced by MLwiN, shows the mean heights by the mean age at each measurement occasion. We assume that growth can be represented by a polynomial function, whose coefficients vary from individual to individual. Other functions are possible, including fractional polynomials or non-linear functions, but for simplicity we confine ourselves to examining a fourth-order polynomial in age (t) centred at an origin of 12.25 years. In some applications of growth curve modelling transformations of the time scale may be useful, often

Copyright © 2002 John Wiley & Sons, Ltd.

to orthogonal polynomials. In the present case the use of ordinary polynomials provides an accessible interpretation and does not lead to computational problems, for example due to near-collinearities. The model we fit can be written as follows:

$$
\begin{aligned}
y_{ij} &= \sum_{h=0}^{4} \beta_{hj} t_{ij}^{h} + e_{ij} \\
\beta_{0j} &= \beta_0 + u_{0j} \\
\beta_{1j} &= \beta_1 + u_{1j} \\
\beta_{2j} &= \beta_2 + u_{2j} \\
\beta_{3j} &= \beta_3 \\
\beta_{4j} &= \beta_4 \\
\begin{pmatrix} u_0 \\ u_1 \\ u_2 \end{pmatrix} &\sim \mathrm{N}(0, \ \Omega_u) \\
\Omega_u &= \begin{pmatrix} \sigma_{u0}^2 & & \\ \sigma_{u01} & \sigma_{u1}^2 & \\ \sigma_{u02} & \sigma_{u12} & \sigma_{u2}^2 \end{pmatrix} \\
e &\sim \mathrm{N}(0, \sigma_e^2)
\end{aligned}
\tag{6}
$$

This is a two-level model with level-1 being 'measurement occasion' and level 2 'individual boy'. Note that we allow only the coefficients up to the second order to vary across individuals; in the present case this provides an acceptable fit. The level 1 residual term e_{ij} represents the unexplained variation within individuals about each individual's growth trajectory. Table I shows the restricted maximum likelihood (REML) parameter estimates for this model. The log-likelihood is calculated for the ML estimates since this is preferable for purposes of model comparison [17].

From this table we can compute various features of growth. For example, the average growth rate (by differentiation) at age 13.25 years ($t=1$) is $6.17 + 2\times 1.13 + 3\times 0.45 - 4\times 0.38 = 8.26$ cm/year. A particular advantage of this formulation is that, for each boy, we can also estimate his random effects or 'residuals', u_{0j}, u_{1j}, u_{2j}, and use these to predict their growth curve at each age [18]. Figure 5, from MLwiN, shows these predicted curves (these can be produced in different colours on the screen).

Goldstein *et al.* [16] show that growth over this period exhibits a seasonal pattern with growth in the summer being about 0.5 cm greater than growth in the winter. Since the period of the growth cycle is a year, this is modelled by including a simple cosine term, which could also have a random coefficient.

In our example we have a set of individuals all of whom have nine measurements. This restriction, however, is not necessary and (6) does not require either the same number of occasions per individual or that measurements are made at equal intervals, since time is modelled as a continuous function. In other words we can combine data from individuals with very different measurement patterns, some of whom may only have been measured once

Copyright © 2002 John Wiley & Sons, Ltd.

Table I. Height modelled as a fourth-degree polynomial on age. REML estimates.

Fixed effects	Estimate	Standard error
Intercept	149.0	1.57
t	6.17	0.36
t^2	1.13	0.35
t^3	0.45	0.16
t^4	−0.38	0.30

Random: level-2 (individual) correlation matrix, variances on diagonal

	Intercept	t	t^2
Intercept	64.0		
t	0.61	2.86	
t^2	0.22	0.66	0.67

Random: level-1 variance = 0.22.
−2 log-likelihood(ML) = 625.4.

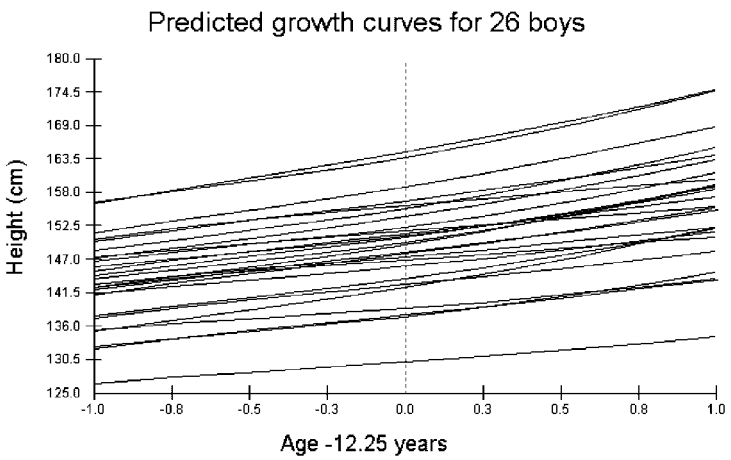

Figure 5.

and some who have been measured several times at irregular intervals. This flexibility, first noted by Laird and Ware [10], means that the multilevel approach to fitting repeated measures data is to be preferred to previous methods based upon a multivariate formulation assuming a common set of fixed occasions [19, 20].

In these models it is assumed that the level-1 residual terms are independently distributed. We may relax this assumption, however, and in the case of repeated measures data this may be necessary, for example where measurements are taken very close together in time. Suppose we wish to fit a model that allows for correlations between the level-1 residuals, and to start with for simplicity let us assume that these correlations are all equal. This is easily

Copyright © 2002 John Wiley & Sons, Ltd.

Table II. Height modelled as a fourth-degree polynomial on age, including a seasonal effect and serial correlation. REML estimates.

Fixed effects	Estimate	Standard error
Intercept	148.9	
t	6.19	0.36
t^2	2.16	0.45
t^3	0.39	0.17
t^4	−1.55	0.43
Cos(t)	−0.24	0.07

Random: level-2 (individual) correlation matrix, covariances on diagonal

	Intercept	t	t^2
Intercept	63.9		
t	0.61	2.78	
t^2	0.24	0.69	0.59

Random: level 1(SE in brackets)
σ_e^2 0.24(0.05)
α 6.59(1.90)
−2 log-likelihood (ML) = 611.5.

accomplished within the GLS step for the random parameters by modifying (5) to

$$\sigma_e^2 \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} + \delta \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \tag{7}$$

so that the parameter δ is the common level-1 covariance (between occasions). Goldstein *et al.* [16] show how to model quite general non-linear covariance functions and in particular those of the form $\mathrm{cov}(e_t e_{t-s}) = \sigma_e^2 \exp(-g(\alpha, s))$, where s is the time difference between occasions. This allows the correlation between occasions to vary smoothly as a function of their (continuous) time difference. A simple example is where $g = \alpha s$, which, in discrete time, produces an AR(1) model. The GLS step now involves non-linear estimation that is accomplished in a standard fashion using a Taylor series approximation within the overall iterative scheme. Pourahmadi [21, 22] considers similar models but restricted to a fixed set of discrete occasions.

Table II shows the results of fitting the model with $g = \alpha s$ together with a seasonal component. If this component has amplitude, say, α we can write it in the form $\alpha \cos(t^* + \gamma)$, where t^* is measured from the start of the calendar year. Rewriting this in the form $\alpha_1 \cos(t^*) - \alpha_2 \sin(t^*)$ we can incorporate the $\cos(t^*)$, $\sin(t^*)$ as two further predictor variables in the fixed part of the model. In the present case α_2 is small and non-significant and is omitted. The results

Copyright © 2002 John Wiley & Sons, Ltd.

show that for measurements made three months apart the serial correlation is estimated as 0.19 ($e^{-6.59/4}$) and as 0.04 ($e^{-6.59/2}$) for measurements taken at 6-monthly intervals. This suggests, therefore, that in practice, for such data when the intervals are no less than 6 months apart serial correlation can be ignored, but should be fitted when intervals are as small as 3 months. This will be particularly important in highly unbalanced designs where there are some individuals with many measurements taken close together in time; ignoring serial correlation will give too much weight to the observations from such individuals.

Finally, on this topic, there will typically need to be a trade-off between modelling more random coefficients at level 2 in order to simplify or eliminate a level-1 serial correlation structure, and modelling level 2 in a parsimonious fashion so that a relatively small number of random coefficients can be used to summarize each individual. An extreme example of the latter is given by Diggle [23] who fits only a random intercept at level 2 and serial correlation at level 1.

7. MULTIVARIATE RESPONSE DATA

We shall use an extension of the model for repeated measures data to illustrate how to model multivariate response data. Consider model (6) where we have data on successive occasions for each individual and in addition, for some or all individuals, we have a measure, say, of their final adult height $y_3^{(2)}$, and their (log) income at age 25 years, $y_4^{(2)}$, where the superscript denotes a measurement made at level 2. We can include these variables as further responses by extending (6) as follows:

$$
\begin{aligned}
y_{ij}^{(1)} &= \sum_{h=0}^{4} \beta_{hj} t_{ij}^{h} + e_{ij} \\
\beta_{0j} &= \beta_0 + u_{0j} \\
\beta_{1j} &= \beta_1 + u_{1j} \\
\beta_{2j} &= \beta_2 + u_{2j} \\
\beta_{3j} &= \beta_3 \\
\beta_{4j} &= \beta_4 \\
y_{3j}^{(2)} &= \alpha_3 + u_{3j} \\
y_{4j}^{(2)} &= \alpha_4 + u_{4j} \\
\begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} &\sim \mathrm{N}(0, \ \Omega_u) \\
e &\sim \mathrm{N}(0, \sigma_e^2)
\end{aligned}
\tag{8}
$$

Copyright © 2002 John Wiley & Sons, Ltd.

We now have a model where there are response variables defined at level 1 (with superscript (1)) and also at level 2 (with superscript (2)). For the level-2 variables we have specified only an intercept term in the fixed part, but quite general functions of individual level predictors, such as gender, are possible. The level-2 responses have no component of random variation at level 1 and their level-2 residuals covary with the polynomial random coefficients from the level-1 repeated measures response.

The results of fitting this model allow us to quantify the relationships between growth events, such as growth acceleration (differentiating twice) at $t=0$, age 12.25 years, $(2\beta_{2j})$ and adult height and also to use measurements taken during the growth period to make efficient predictions of adult height or income. We note that for individual j the estimated (posterior) residuals $\hat{u}_{3j}, \hat{u}_{4j}$ are the best linear unbiased predictors of the individual's adult values; where we have only a set of growth period measurements for an individual these therefore provide the required estimates. Given the set of model parameters, therefore, we immediately obtain a system for efficient adult measurement prediction given a set of growth measurements [24].

Suppose, now, that we have no growth period measurements and just the two adult measurements for each individual. Model (8) reduces to

$$\begin{aligned} y_{3j}^{(2)} &= \alpha_3 + u_{3j} \\ y_{4j}^{(2)} &= \alpha_4 + u_{4j} \\ \begin{pmatrix} u_3 \\ u_4 \end{pmatrix} &\sim \mathrm{N}(0,\ \Omega_u) \\ \mathbf{V_{1j}} = 0,\ \mathbf{V_{2j}} &= \begin{pmatrix} \sigma_{u3}^2 & \\ \sigma_{u34} & \sigma_{u4}^2 \end{pmatrix} \end{aligned} \tag{9}$$

Thus we can think of this as a two-level model with no level-1 variation and every level 2 unit containing just two level 1 units. The explanatory variables for the simple model given by (9) are just two dummy variables defining, alternately, the two responses. Thus we can write (9) in the more compact general form

$$\begin{aligned} y_{ij} &= \sum_{h=1}^{2} \beta_{0hj} x_{hij}, \quad x_{1ij} = \begin{cases} 1 & \text{if response 1} \\ 0 & \text{if response 2} \end{cases}, \quad x_{2ij} = 1 - x_{1ij} \\ \beta_{0hj} &= \beta_{0h} + u_{hj} \\ \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} &= \begin{pmatrix} \sigma_{u1}^2 & \\ \sigma_{u12} & \sigma_{u2}^2 \end{pmatrix} \end{aligned} \tag{10}$$

Note that there is no need for every individual to have both responses and so long as we can consider 'missing' responses as random, the IGLS algorithm will supply maximum likelihood estimates. We can add further covariates to the model in a straightforward manner by forming interactions between them and the dummy variables defining the separate response intercepts.

The ability to fit a multivariate linear model with randomly missing responses finds a number of applications, for example where matrix or rotation designs are involved (reference [18], Chapter 4), each unit being allocated, at random, a subset of responses. The possibility of

Copyright © 2002 John Wiley & Sons, Ltd.

having additionally level-1 responses allows this to be used as a very general model for meta analysis where there are several studies (level-2 units) for some of which responses are available only in summary form at level 2 and for others detailed level-1 responses are available. Goldstein *et al.* [25] provide a detailed example.

8. CROSS-CLASSIFIED AND MULTIPLE MEMBERSHIP STRUCTURES

Across a wide range of disciplines it is commonly the case that data have a structure that is not purely hierarchical. Individuals may be clustered not only into hierarchically ordered units (for example occasions nested within patients nested within clinics), but may also belong to more than one type of unit at a given level of a hierarchy. Consider the example of a livestock animal such as a cow where there are a large number of mothers, each producing several female offspring that are eventually reared for milk on different farms. Thus, an offspring might be classified as belonging to a particular combination of mother and farm, in which case they will be identified by a *cross-classification* of these.

Raudenbush [26] and Rasbash and Goldstein [27] present the general structure of a model for handling complex hierarchical structuring with random cross-classifications. For example, assuming that we wish to formulate a linear model for the milk yield of offspring taking into account both the mother and the farm, then we have a cross-classified structure, which can be modelled as follows:

$$\begin{aligned} y_{i(j_1,j_2)} &= (X\beta)_{i(j_1,j_2)} + u_{j_1} + u_{j_2} + e_{i(j_1 j_2)} \\ & j_1 = 1,\ldots,J_1, \quad j_2 = 1,\ldots,J_2, \quad i = 1,\ldots,N \end{aligned} \tag{11}$$

in which the yield of offspring i, belonging to the combination of mother j_1 and farm j_2, is predicted by a set of fixed coefficients $(X\beta)_{i(j_1,j_2)}$. The random part of the model is given by two level-2 residual terms, one for the mother (u_{j_1}) and one for the farm (u_{j_2}), together with the usual level-1 residual term for each offspring. Decomposing the variation in such a fashion allows us to see how much of it is due to the different classifications. This particular example is somewhat oversimplified, since we have ignored paternity and we would also wish to include factors such as age of mother, parity of offspring etc. An application of this kind of modelling to a more complex structure involving Salmonella infection in chickens is given by Rasbash and Browne [28].

Considering now just the farms, and ignoring the mothers, suppose that the offspring often change farms, some not at all and some several times. Suppose also that we know, for each offspring, the weight w_{ij_2}, associated with the j_2th farm for offspring i with $\sum_{j_2=1}^{J_2} w_{ij_2} = 1$. These weights, for example, may be proportional to the length of time an offspring stays in a particular farm during the course of our study. Note that we allow the possibility that for some (perhaps most) animals only one farm is involved so that one of these probabilities is one and the remainder are zero. Note that when all level-1 units have a single non-zero weight of 1 we obtain the usual purely hierarchical model. We can write for the special case of membership of up to two farms $\{1,2\}$:

$$\begin{aligned} y_{i(1,2)} &= (X\beta)_{i(1,2)} + w_{i1}u_1 + w_{i2}u_2 + e_{i(1,2)} \\ w_{i1} + w_{i2} &= 1 \end{aligned} \tag{12}$$

Copyright © 2002 John Wiley & Sons, Ltd.

and more generally

$$y_{i\{j\}} = (X\beta)_{i\{j\}} + \sum_{h\in\{j\}} w_{ih}u_h + e_{i\{j\}}$$

$$\sum_h w_{ih} = 1, \quad \mathrm{var}(u_h) = \sigma_u^2 \tag{13}$$

$$\mathrm{var}\left(\textstyle\sum_h w_{ih}u_h\right) = \sigma_u^2 \sum_h w_{ih}^2$$

Thus, in the particular case of membership of just two farms with equal weights we have

$$w_{i1} = w_{i2} = 0.5, \quad \mathrm{var}\left(\sum_h w_{ih}u_h\right) = \sigma_u^2/2$$

Further details of this model are given by Hill and Goldstein [29].

An extension of the multiple membership model is also possible and has important applications, for example in modelling spatial data. In this case we can write

$$y_{i\{j_1\}\{j_2\}} = (X\beta)_{i\{j\}} + \sum_{h\in\{j_1\}} w_{1ih}u_{1h} + \sum_{h\in\{j_2\}} w_{2ih}u_{2h} + e_{i\{j\}}$$

$$\sum_h w_{1ih} = W_1, \quad \sum_h w_{2ih} = W_2, \quad \mathrm{var}(u_{1h}) = \sigma_{u1}^2, \quad \mathrm{var}(u_{2h}) = \sigma_{u2}^2 \tag{14}$$

$$\mathrm{cov}(u_{1h}, u_{2h}) = \sigma_{u12}, \quad j = \{j_1, j_2\}$$

There are now two sets of higher level units that influence the response and in general we can have more than two such sets. In spatial models one of these sets is commonly taken to be the area where an individual (level 1) unit occurs and so does not have a multiple membership structure (since each individual belongs to just one area, that is we replace $\sum_h w_{1ih}u_{1h}$ by u_{1j_1}). The other set consists of those neighbouring units that are assumed to have an effect. The weights will need to be carefully chosen; in spatial models W_2 is typically chosen to be equal to 1 (see Langford *et al.* [30] for an example). Another application for a model such as (14) is for household data where households share facilities, for example an address. In this case the household that an individual resides in will belong to one set and the other households at the address will belong to the other set. Goldstein *et al.* [31] give an application of this model to complex household data.

9. META-ANALYSIS

Meta-analysis involves the pooling of information across studies in order to provide both greater efficiency for estimating treatment effects and also for investigating why treatments effects may vary. By formulating a general multilevel model we can do both of these efficiently within a single-model framework, as has already been indicated and was suggested by several authors [32, 33]. In addition we can combine data that are provided at either individual subject level or aggregate level or both. We shall look at a simple case but this generalizes readily [25].

Copyright © 2002 John Wiley & Sons, Ltd.

Consider an underlying model for individual level data where a pair of treatments are being compared and results from a number of studies or centres are available. We write a basic model, with a continuous response Y as

$$\begin{aligned} y_{ij} &= (X\beta)_{ij} + \beta_2 t_{ij} + u_j + e_{ij} \\ \mathrm{var}(u_j) &= \sigma_u^2, \quad \mathrm{var}(e_{ij}) = \sigma_e^2 \end{aligned} \tag{15}$$

with the usual assumptions of Normality etc. The covariate function is designed to adjust for initial clinic and subject conditions. The term t_{ij} is a dummy variable defining the treatment (0 for treatment A, 1 for treatment B). The random effect u_j is a study effect and the e_{ij} are individual-level residuals. Clearly this model can be elaborated in a number of ways, by including random coefficients at level 2 so that the effect of treatment varies across studies, and by allowing the level-1 variance to depend on other factors such as gender or age.

Suppose now that we do not have individual data available but only means at the study level. If we average (15) to the study level we obtain

$$y_{.j} = (X\beta)_{.j} + \beta_2 t_{.j} + u_j + e_{.j} \tag{16}$$

where $y_{.j}$ is the mean response for the jth study etc. The total residual variance for study j in this model is $\sigma_u^2 + \sigma_e^2/n_j$ where n_j is the size of the jth study. It is worth noting at this point that we are ignoring, for simplicity, levels of variation that might exist within studies, such as that between sites for a multi-site study. If we have the values of $y_{.j}, (X\beta)_{.j}, t_{.j}$ where the latter is simply the proportion of subjects with treatment B in the jth study, and also the value of n_j then we will be able to obtain estimates for the model parameters, so long as the n_j differ. Such estimates, however, may not be very precise and extra information, especially about the value of σ_e^2, will improve them.

Model (16) therefore forms the basis for the multilevel modelling of aggregate level data. In practice the results of studies will often be reported in non-standard form, for example with no estimate of σ_e^2 but it may be possible to estimate this from reported test statistics. In some cases, however, the reporting may be such that the study cannot be incorporated in a model such as (16). Goldstein *et al.* [25] set out a set of minimum reporting conventions for meta-analysis studies subsequently to be carried out.

While it is possible to perform a meta-analysis with only aggregate level data, it is clearly more efficient to utilize individual-level data where these are available. In general, therefore, we will need to consider models that have mixtures of individual and aggregate data, even perhaps within the same study. We can do this straightforwardly by specifying a model which is just the combination of (15) and (16), namely

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + \beta_2 t_{ij} + u_j + e_{ij} \\ y_{.j} &= \beta_0 + \beta_1 x_{.j} + \beta_2 t_{.j} + u_j + e_j z_j \\ z_j &= \sqrt{n_j^{-1}}, \quad e_j \equiv e_{ij} \end{aligned} \tag{17}$$

What we see is that the common level-1 and level-2 random terms link together the separate models and allow a joint analysis that makes fully efficient use of the data. Several issues immediately arise from (17). One is that the same covariates are involved. This is also a requirement for the separate models. If some covariate values are missing at either level then it is possible to use an imputation technique to obtain estimates, assuming a suitable random

Copyright © 2002 John Wiley & Sons, Ltd.

missingness mechanism. The paper by Goldstein *et al.* [25] discusses generalizations of (17) for several treatments and the procedure can be extended to generalized linear models.

10. GENERALIZED LINEAR MODELS

So far we have dealt with linear models, but all of those so far discussed can be modified using non-linear link functions to give generalized linear multilevel models. We shall not discuss these in detail (see Goldstein [18] for details and some applications) but for illustration we will describe a two-level model with a binary response.

Suppose the outcome of patients in intensive care is recorded simply as survived (0) or died (1) within 24 hours of admission. Given a sample of patients from a sample of intensive care units we can write one model for the probability of survival as

$$\begin{aligned} \text{logit}(\pi_{ij}) &= (X\beta)_{ij} + u_j \\ y_{ij} &\sim \text{Bin}(\pi_{ij}, 1) \end{aligned} \tag{18}$$

Equation (18) uses a standard logit link function assuming binomial (Bernoulli) error distribution for the (0,1) response y_{ij}. The level-2 random variation is described by the term u_j within the linear predictor. The general interpretation is similar to that for a continuous response model, except that the level-1 variation is now a function of the predicted value π_{ij}. While in (18) there is no separate estimate for the level-1 variance, we may wish to fit extrabinomial variation which will involve a further parameter.

We can modify (18) using alternative link functions, for example the logarithm if the response is a count, and can allow further random coefficients at level 2. The response can be a multi-category variable, either ordered or unordered, and this provides an analogue to the multivariate models for continuously distributed responses. As an example of an ordered categorical response, consider extending the previous outcome to three categories: survival without impairment (1); survival with impairment (2); death (3). If $\pi_{ij}^{(h)}$, $h=1,2,3$ are, respectively, the probabilities for each of these categories, we can write a *proportional odds* model using a logit link as

$$\begin{aligned} \text{logit}(\pi_{ij}^{(1)}) &= \alpha^{(1)} + (X\beta)_{ij} + u_j^{(1)} \\ \text{logit}(\pi_{ij}^{(1)} + \pi_{ij}^{(2)}) &= \alpha^{(2)} + (X\beta)_{ij} + u_j^{(2)} \end{aligned} \tag{19}$$

and the set of three (0,1) observed responses for each patient is assumed to have a multinomial distribution with mean vector given by $\pi_{ij}^{(h)}$, $h=1,2,3$. Since the probabilities add to 1 we require two lines in (19) which differ only in terms of the overall level given by the intercept term as well as allowing for these to vary across units.

Unlike in the continuous Normal response case, maximum likelihood estimation is not straightforward and beyond fairly simple two-level models involves a considerable computational load typically using numerical integration procedures [34]. For this reason approximate methods have been developed based upon series expansions and using quasi-likelihood approaches [18] which perform well under a wide range of circumstances but can break down in certain conditions, especially when data are sparse with binary responses. High-order Laplace

Copyright © 2002 John Wiley & Sons, Ltd.

approximations have been found to perform well [35] as have simulation-based procedures such as MCMC (see below).

It is worth mentioning one particular situation where care is required in using generalized linear multilevel models. In (15) and (16) we assume that the level-1 responses are independently distributed with probabilities distributed according to the specified model. In some circumstances such an assumption may not be sensible. Consider a repeated measures model where the health status (satisfactory/not satisfactory) of individuals is repeatedly assessed. Some individuals will typically always respond 'satisfactory' whereas some others can be expected to respond always 'not satisfactory'. For these individuals the underlying probabilities are either zero or one, which violates the model assumptions and what one finds if one tries to fit a model where there are non-negligible numbers of such individuals are noticeable amounts of underdispersion. Barbosa and Goldstein [36] discuss this problem and propose a solution based upon fitting a serial correlation structure.

We can also have multivariate multilevel models with mixtures of discrete and continuous responses. Certain of these can be fitted in MLwiN using quasi-likelihood procedures [12] and MCMC procedures for such models are currently being implemented. See also Olsen and Schafer [37] for an alternative approach.

11. SURVIVAL MODELS

Several different formulations for survival data modelling are available; to illustrate how these can be extended to multilevel structures, where they are often referred to as frailty models [38], we consider the proportional hazards (Cox) model and a piecewise discrete time model. Goldstein [18] gives other examples.

Consider a simple two-level model with, say, patients within hospitals or occasions within subjects. As in the standard single-level case we consider each time point in the data as defining a *block* indicated by l at which some observations come to the end of their duration due to either failure or censoring and some remain to the next time or block. At each block there is therefore a set of observations – the total *risk set*. To illustrate how the model is set up we can think of the data sorted so that each observation within a block is a level-1 unit, above which, in the repeated measures case, there are occasions at level 2 and subjects at level 3. The ratio of the hazard for the unit which experiences a failure at a given occasion referred to by (j,k) to the sum of the hazards of the remaining risk set units [39] is

$$\frac{\exp(\beta_1 x_{1ijk} + u_{jk})}{\sum_{j,k} \exp(\beta_1 x_{1ijk} + u_{jk})} \tag{20}$$

where j and k refer to the *real* levels 2 and 3, for example occasion and subject. At each block denoted by l the response variable may be defined *for each member of the risk set* as

$$y_{ijk(l)} = \begin{cases} 1 & \text{failed} \\ 0 & \text{not} \end{cases}$$

Copyright © 2002 John Wiley & Sons, Ltd.

Because of equivalence between the likelihood for the multinomial and Poisson distributions, the latter is used to fit model (20). This can be written as

$$E(y_{ijk(l)}) = \exp(\alpha_l + X_{jk}\beta_k) \tag{21}$$

Where there are ties within a block then more than one response will be non-zero. The terms α_l fit the underlying hazard function as a '*blocking factor*', and can be estimated by fitting either a set of parameters, one for each block, or a smoothed polynomial curve over the blocks numbered $1,\ldots,p$. Thus if the hth block is denoted by h, α_l is replaced by a low-order polynomial, order m, $\sum_{t=0}^{m}\gamma_t h^t$, where the γ_t are (nuisance) parameters to be estimated.

Having set up this model, the data are now sorted into the real two-level structure, for example in the repeated measures case by failure times within subjects with occasions within the failure times. This retains proportional hazards within subjects. In this formulation the Poisson variation is defined at level 1, there is no variation at level 2 and the between-subject variation is at level 3. Alternatively we may wish to preserve *overall* proportionality, in which case the failure times define level 3 with no variation at that level. See Goldstein [18] for a further discussion of this.

Consider now a piecewise survival model. Here the total time interval is divided into short intervals during which the probability of failure, given survival up to that point, is assumed constant. Denote these intervals by t $(1,2,\ldots,T)$ so that the hazard at time t is the probability that, given survival up to the end of time interval $t-1$, failure occurs in the next interval. At the start of each interval we have a 'risk set' n_t consisting of the survivors and during the interval r_t fail. If censoring occurs during interval t then this observation is removed from that interval (and subsequent ones) and does not form part of the risk set. A simple, *single-level*, model for the probability can be written as

$$\pi_{i(t)} = f[\alpha_t z_{it}, (\beta X)_{it}] \tag{22}$$

where $z_t = \{z_{it}\}$ is a dummy variable for the tth interval and α_t, as before, is a 'blocking factor' defining the underlying hazard function at time t. The second term is a function of covariates. A common formulation would be the logit model and a simple such model in which the first blocking factor has been absorbed into the intercept term could be written as

$$\text{logit}(\pi_{i(t)}) = \beta_0 + \alpha_t z_{it} + \beta_1 x_{1i}, \quad (z_2, z_3, \ldots, z_T) \tag{23}$$

Since the covariate varies across individuals, in general the data matrix will consist of one record for each individual within each interval, with a (0,1) response indicating survival or failure. The model can now be fitted using standard procedures, assuming a binomial error distribution. As before, instead of fitting $T-1$ blocking factors, we can fit a low-order polynomial to the sequentially numbered time indicator. The logit function can be replaced by, for example, the complementary log-log function that gives a proportional hazards model, or, say, the probit function and note that we can incorporate time-varying covariates such as age.

For the extension to a two-level model we write

$$\text{logit}(\pi_{ij(t)}) = \beta_0 + \sum_{h=1}^{p}\alpha_h^*(z_{it}^*)^h + \beta_1 x_{1ij} + u_j \tag{24}$$

Copyright © 2002 John Wiley & Sons, Ltd.

where u_j is the 'effect', for example, of the jth clinic, and is typically assumed to be distributed normally with zero mean and variance σ_u^2. We can elaborate (24) using random coefficients, resulting in a heterogeneous variance structure, further levels of nesting etc. This is just a two-level binary response model and can be fitted as described earlier. The data structure has two levels so that individuals will be grouped (sorted) within clinics. For a competing risks model with more than one outcome we can use the two-level formulation for a multi-category response described above. The model can be used with repeated measures data where there are repeated survival times within individuals, for example multiple pregnancy states.

12. BAYESIAN MODELLING

So far we have considered the classical approach to fitting multilevel models. If we add prior distributional assumptions to the parameters of the models so far considered we can fit the same range of models from a Bayesian perspective, and in most applications this will be based upon MCMC methods. A detailed comparison of Bayesian and likelihood procedures for fitting multilevel models is given in Browne and Draper [40]. A particular advantage of MCMC methods is that they yield inferences based upon samples from the full posterior distribution and allow exact inference in cases where, as mentioned above, the likelihood based methods yield approximations.

Owing also to their approach of generating a sample of points from the full posterior distributions, they can give accurate interval estimates for non-Gaussian parameter distributions.

In MCMC sampling we are interested in generating samples of values from the joint posterior distribution of all the unknown parameters rather than finding the maximum of this distribution. Generally it is not possible to generate directly from this joint distribution, so instead the parameters are split into groups and for each group in turn we generate a set of values from its conditional posterior distribution. This can be shown to be equivalent to sampling directly from the joint posterior distribution.

There are two main MCMC procedures that are used in practice: Gibbs sampling [41] and Metropolis–Hastings (MH) [42, 43] sampling. When the conditional posterior for a group of parameters has a standard form, for example a Normal distribution, then we can generate values from it directly and this is known as Gibbs sampling. When the distribution is not of standard form then it may still be possible to use Gibbs sampling by constructing the distribution using forms of Gibbs sampling such as adaptive rejection sampling [44].

The alternative approach is to use MH sampling where values are generated from another distribution called a proposal distribution rather than the conditional posterior distribution. These values are then either accepted or rejected in favour of the current values by comparing the posterior probabilities of the joint posterior at the current and proposed new values. The acceptance rule is designed so that MH is effectively sampling from the conditional posterior even though we have used an arbitrary proposal distribution; nevertheless, choice of this proposal distribution is important for efficiency of the algorithm.

MCMC algorithms produce chains of serially correlated parameter estimates and consequently often have to be run for many iterations to get accurate estimates. Many diagnostics are available to gauge approximately for how long to run the MCMC methods. The chains are also started from arbitrary parameter values and so it is common practice to ignore the

Copyright © 2002 John Wiley & Sons, Ltd.

first N iterations (known as a burn-in period) to allow the chains to move away from the starting value and settle at the parameters' equilibrium distribution.

We give here an outline of the Gibbs sampling procedure for fitting a general Normal two-level model

$$\begin{aligned} y_{ij} &= \mathbf{X}_{ij}\beta + \mathbf{Z}_{ij}u_j + e_{ij} \\ u_j &\sim \mathrm{N}(0, \Omega_u), e_{ij} \sim \mathrm{N}(0, \sigma_e^2), \\ i &= 1, \ldots, n_j, \quad j = 1, \ldots, J, \quad \sum_j n_j = N \end{aligned}$$

We will include generic conjugate prior distributions for the fixed effects and variance parameters as follows:

$$\beta \sim \mathrm{N}(\mu_p, S_p), \Omega_u \sim W^{-1}(\upsilon_u, S_u), \sigma_e^2 \sim SI\chi^2(\upsilon_e, s_e^2)$$

The Gibbs sampling algorithm then involves simulating from the following four sets of conditional distributions:

$$p(\beta \mid y, \sigma_e^2, u) \sim \mathrm{N}\left(\frac{\hat{D}}{\sigma_e^2}\left(\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{X}_{ij}^{\mathrm{T}}(y_{ij} - \mathbf{Z}_{ij}u_j) + S_p^{-1}\mu_p\right), \hat{D}\right)$$

$$p(u_j \mid y, \Omega_u, \sigma_e^2, \beta) \sim \mathrm{N}\left(\frac{\hat{D}_j}{\sigma_e^2}\sum_{i=1}^{n_j}\mathbf{Z}_{ij}^{\mathrm{T}}(y_{ij} - \mathbf{X}_{ij}\beta), \hat{D}_j\right)$$

$$p(\Omega_u \mid u) \sim W^{-1}\left(J + \upsilon_u, \sum_{j=1}^{J} u_j u_j^{\mathrm{T}} + S_u\right)$$

$$p(\sigma_e^2 \mid y, \beta, u) \sim \Gamma^{-1}\left(\frac{N + \upsilon_e}{2}, \frac{1}{2}\left(\upsilon_e s_e^2 + \sum_{j=1}^{J}\sum_{i=1}^{n_j} e_{ij}^2\right)\right)$$

where

$$\hat{D} = \left(\sum_{ij}\frac{\mathbf{X}_{ij}^{\mathrm{T}}\mathbf{X}_{ij}}{\sigma_e^2} + S_p^{-1}\right)^{-1}$$

and

$$\hat{D}_j = \left(\frac{1}{\sigma_e^2}\sum_{i=1}^{n_j}\mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{Z}_{ij} + \Omega_u^{-1}\right)^{-1}$$

Note that in this algorithm we have used generic prior distributions. This allows the incorporation of informative prior information but generally we will not have this information and so will use so-called 'diffuse' prior distributions that reflect our lack of knowledge. Since there are only 26 level-2 units from which we are estimating a 3×3 covariance matrix, the exact

Copyright © 2002 John Wiley & Sons, Ltd.

Table III. Height modelled as a fourth-degree polynomial on age using MCMC sampling.

Fixed effects	MCMC Estimate	Standard error
Intercept	149.2	1.58
t	6.21	0.38
t^2	1.14	0.36
t^3	0.45	0.17
t^4	−0.38	0.30

Random: level-2 (individual) correlation matrix, variances on diagonal, estimates are mean(mode)

	Intercept	t	t^2
Intercept	74.4 (68.6)		
t	0.61	3.34 (3.06)	
t^2	0.22	0.66	0.67 (0.70)

Random: level-1 variance = 0.23(0.22).

choice of prior is important. We here use the following set of priors for the child growth example considered in Section 6:

$$p(\beta_0) \propto 1, \quad p(\beta_1) \propto 1, \quad p(\beta_2) \propto 1, \quad p(\beta_3) \propto 1, \quad p(\beta_4) \propto 1, \quad p(\beta_5) \propto 1$$

$$p(\Omega_u) \sim \text{inverse Wishart}_3[3, 3 \times S_u], \quad S_u = \begin{bmatrix} 64.0 & & \\ 8.32 & 2.86 & \\ 1.42 & 0.92 & 0.67 \end{bmatrix}$$

$$p(\sigma^2_{e0}) \sim \text{inverse gamma}(0.001, 0.001)$$

The inverse Wishart prior matrix is based upon the REML estimates, chosen to be 'minimally informative', with degrees of freedom equal to the order of the covariance matrix.

The results of running model (5) for 50000 iterations after a burn-in of 500 can be seen in Table III. Here we see that the fixed effect estimates are very similar to the estimates obtained by the maximum likelihood method. The variance (chain mean) estimates are however inflated due to the skewness of the variance parameters. Modal estimates of the variance parameters, apart from that for the quadratic coefficient, are closer, as is to be expected. If we had used the uniform prior $p(\Omega_u) \propto 1$ for the covariance matrix, the estimates of the fixed coefficients are little changed but the covariance matrix estimates are noticeably different. For example, the variance associated with the intercept is now 93.0 and those for the linear and quadratic coefficients become 4.2 and 1.1, respectively.

Figure 6 shows the diagnostic screen produced by MLwiN following this MCMC run. The top left hand box shows the trace for this parameter, the level-2 intercept variance. This looks satisfactory and this is confirmed by the estimated autocorrelation function (ACF) and partial autocorrelation function (PACF) below. A kernel density estimate is given at the top right and

Copyright © 2002 John Wiley & Sons, Ltd.

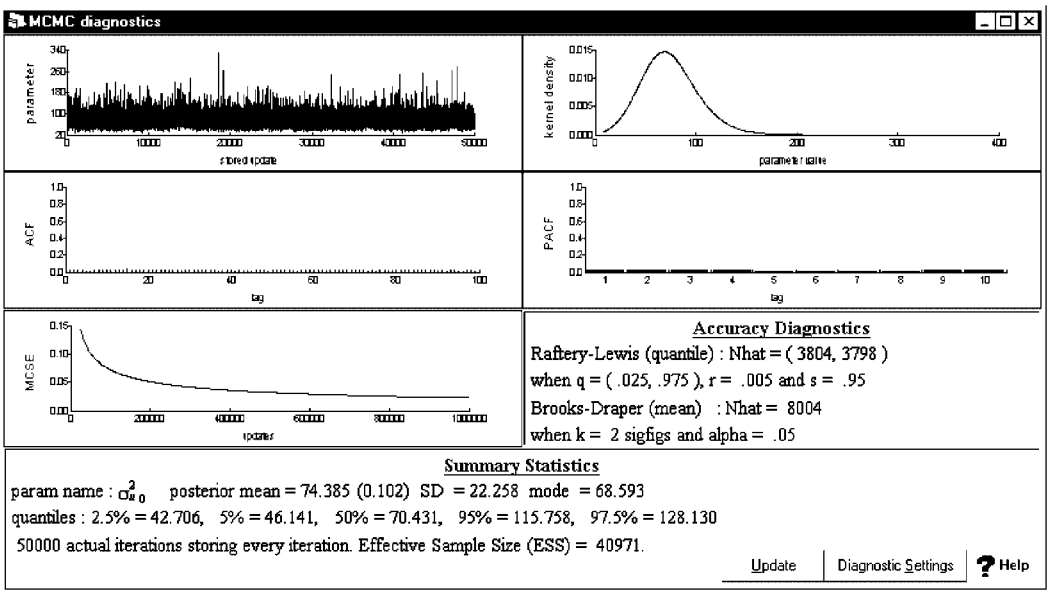

Figure 6. MCMC summary screen.

the bottom left box is a plot of the Monte Carlo standard error against number of iterations in the chain. The summary statistics give quantiles, mean and mode together with accuracy diagnostics that indicate the required chain length.

The MCMC methods are particularly useful in models like the cross-classified and multiple membership models discussed in Section 8. This is because whereas the maximum likelihood methods involve constructing large constrained variance matrices for these models, the MCMC methods simulate conditional distributions in turn and so do not have to adjust to the structure of the model.

For model fitting, one strategy (but not the only one) is to use the maximum or quasi-likelihood methods for performing model exploration and selection due to speed. Then MCMC methods could be used for inference on the final model to obtain accurate interval estimates.

13. BOOTSTRAPPING

Like MCMC the bootstrap allows inferences to be based upon (independent) chains of values and can be used to provide exact inferences and corrections for bias. Two forms of bootstrapping have been studied to date: parametric bootstrapping, especially for correcting biases in generalized linear models [45], and non-parametric bootstrapping based upon estimated residuals [46]. The fully parametric bootstrap for a multilevel model works as for a single-level model with simulated values generated from the estimated covariance matrices at each level. Thus, for example, for model (2) each bootstrap sample is created using a set of u_j, e_{ij} sampled from $\mathrm{N}(0, \sigma_u^2), \mathrm{N}(0, \sigma_e^2)$, respectively.

Copyright © 2002 John Wiley & Sons, Ltd.

In the non-parametric case, full 'unit resampling' is generally only possible by resampling units at the highest level. For generalized linear models, however, we can resample posterior residuals, once they have been adjusted to have the correct (estimated) covariance structures and this can be shown to possess particular advantages over a fully parametric bootstrap where asymmetric distributions are involved [46].

14. IN CONCLUSION

The models that we have described in this paper represent a powerful set of tools available to the data analyst for exploring complex data structures. They are being used in many areas, including health, with great success in providing insights that are unavailable with more conventional methods. There is a growing literature extending these models, for example to multilevel structural equation models, and especially to the application of the multiple membership models in areas such as population demography [31]. An active e-mail discussion group exists which welcomes new members (www.jiscmail.ac.uk/multilevel). The data set used for illustration in this paper is available on the Centre for Multilevel Modelling website (multilevel.ioe.ac.uk/download/macros.html) as an MLwiN worksheet.

REFERENCES

1. Snijders T, Bosker R. *Multilevel Analysis*. Sage: London, 1999.
2. Leyland AH, Goldstein H (eds). *Multilevel Modelling of Health Statistics*. Wiley: Chichester, 2001.
3. Sullivan LM, Dukes KA, Losina E. Tutorial in biostatistics: an introduction to hierarchical linear modelling. *Statistics in Medicine* 1999; **18**:855–888.
4. Zeger SL, Liang K, Albert P. Models for longitudinal data: a generalised estimating equation approach. *Biometrics* 1988; **44**:1049–1060.
5. Liang K-Y, Zeger SL, Qaqish B. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society*, *Series B* 1992; **54**:3–40.
6. Heagerty PJ, Zeger SL. Marginalized multilevel models and likelihood inference (with discussion). *Statistical Science* 2000; **15**:1–26.
7. Lindsey JK, Lambert P. On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine* 1998; **17**:447–469.
8. Goldstein H, Rasbash J. Efficient computational procedures for the estimation of parameters in multilevel models based on iterative generalised least squares. *Computational Statistics and Data Analysis* 1992; **13**:63–71.
9. Longford NT. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested effects. *Biometrika* 1987; **74**:812–827.
10. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
11. Bryk AS, Raudenbush SW. *Hierarchical Linear Models*. Sage: Newbury Park, California, 1992.
12. Rasbash J, Browne W, Goldstein H, Yang M, Plewis, I, Healy M, Woodhouse G, Draper D, Langford I, Lewis T. *A User's Guide to MlwiN*. (2nd edn.). Institute of Education: London, 2000.
13. Goldstein H. Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika* 1986; **73**:43–56.
14. Zhou X, Perkins AJ, Hui SL. Comparisons of software packages for generalized linear multilevel models. *American Statistician* 1999; **53**:282–290.
15. Spiegelhalter DJ, Thomas A, Best NG. *WinBUGS Version 1.3: User Manual*. MRC Biostatistics Research Unit: Cambridge, 2000.
16. Goldstein H, Healy MJR, Rasbash J. Multilevel time series models with applications to repeated measures data. *Statistics in Medicine* 1994; **13**:1643–1655.
17. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 1997; **53**:983–997.
18. Goldstein H. *Multilevel Statistical Models*. Arnold: London, 1995.
19. Grizzle JC, Allen DM. An analysis of growth and dose response curves. *Biometrics* 1969; **25**:357–361.
20. Albert PS. Longitudinal data analysis (repeated measures) in clinical trials. *Statistics in Medicine* 1993; **18**: 1707–1732.

Copyright © 2002 John Wiley & Sons, Ltd.

21. Pourahmadi M. Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika* 1999; **86**:677–690.
22. Pourahmadi M. Maximum likelihood estimation of generalised linear models for multivariate Normal covariance matrix. *Biometrika* 2000; **87**:425–436.
23. Diggle PJ. An approach to the analysis of repeated measurements. *Biometrics* 1988; **44**:959–971.
24. Goldstein H. Flexible models for the analysis of growth data with an application to height prediction. *Revue Epidemiologie et Sante Publique* 1989; **37**:477–484.
25. Goldstein H, Yang M, Omar R, Turner R, Thompson S. Meta analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society*, *Series C* 2000; **49**: 399–412.
26. Raudenbush, SW. A crossed random effects model for unbalanced data with applications in cross sectional and longitudinal research. *Journal of Educational Statistics* 1993; **18**:321–349.
27. Rasbash J, Goldstein H. Efficient analysis of mixed hierarchical and cross classified random structures using a multilevel model. *Journal of Educational and Behavioural Statistics* 1994; **19**:337–350.
28. Rasbash J, Browne W. Non-hierarchical multilevel models. In *Multilevel Modelling of Health Statistics*, Leyland A, Goldstein H (eds). Wiley: Chichester, 2001.
29. Hill PW, Goldstein H. Multilevel modelling of educational data with cross-classification and missing identification of units. *Journal of Educational and Behavioural Statistics* 1998; **23**:117–128.
30. Langford I, Leyland AH, Rasbash J, Goldstein H. Multilevel modelling of the geographical distributions of diseases. *Journal of the Royal Statistical Society*, *Series C* 1999; **48**:253–268.
31. Goldstein H, Rasbash J, Browne W, Woodhouse G, Poulain M. Multilevel models in the study of dynamic household structures. *European Journal of Population* 2000; **16**:373–387.
32. Hedges LV, Olkin IO. *Statistical Methods for Meta Analysis*. Academic Press: Orlando, Florida, 1985.
33. Raudenbush S, Bryk AS. Empirical Bayes meta-analysis. *Journal of Educational Statistics* 1985; **10**:75–98.
34. Hedeker D, Gibbons R. A random effects ordinal regression model for multilevel analysis. *Biometrics* 1994; **50**:933–944.
35. Raudenbush SW, Yang M, Yosef M. Maximum likelihood for generalised linear models with nested random effects via high-order multivariate Laplace approximation. *Journal of Computational and Graphical Statistics* 2000; **9**:141–157.
36. Barbosa MF, Goldstein H. Discrete response multilevel models for repeated measures; an application to voting intentions data. *Quality and Quantity* 2000; **34**:323–330.
37. Olsen MK, Schafer JL. A two-part random effects model for semi- continuous longitudinal data. *Journal of the American Statistical Association* 2001; **96**:730–745.
38. Clayton DG. A Monte Carlo method for Bayesian inference in frailty models. *Biometrics* 1991; **47**:467–485.
39. McCullagh P, Nelder J. *Generalised Linear Models*. Chapman and Hall: London, 1989.
40. Browne W, Draper D. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics* 2000; **15**:391–420.
41. Geman S, Geman D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern analysis and Machine Intelligence* 1984; **45**:721–741.
42. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 1953; **21**:1087–1092.
43. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; **57**:97–109.
44. Gilks WR, Wild P. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistics Society*, *Series C* 1992; **41**:337–348.
45. Goldstein H. Consistent estimators for multilevel generalised linear models using an iterated bootstrap. *Multilevel Modelling Newsletter* 1996; **8**(1):3–6.
46. Carpenter J, Goldstein H, Rasbash J. A nonparametric bootstrap for multilevel models. *Multilevel Modelling Newsletter* 1999; **11**(1):2–5.

Copyright © 2002 John Wiley & Sons, Ltd.