MULTILEVEL MODELLING NEWSLETTER

Vol. 18   **September, 2006**

Editorial

The first Multilevel Modelling Newsletter appeared in 1989 when the technique was not widely known and there was felt to be a need to have a relatively informal forum for discussion of applications and methodological developments.

The Centre for Multilevel Modelling moved from the Institute of Education to Bristol University in 2005 under the joint direction of Jon Rasbash and Fiona Steele. With this move and all the exciting developments it has opened up, we have been reviewing all our activities and after various discussions feel that the time has come to close down this newsletter.

First of all, the techniques themselves are now relatively well known and have been incorporated into the major statistical software packages. Thus there is far less need for the newsletter.

Secondly, we have noticed a reduction in the number of articles sent to us for publication and for referencing, partly at least because of the existence of many other outlets.

Thirdly, the internet has changed the way in which materials can be accessed. Most of the functions of the newsletter can be implemented rather better through web based materials: Jon Rasbash describes some of our Centre plans to do this elsewhere in this newsletter and Hilary Browne describes how our revamped website will look.

So this is the last edition. In signing off, as the final editor, I would like to pay tribute to all of those who directly edited, produced and contributed to the Newsletter and to the

Harvey Goldstein

h.goldstein@bristol.ac.uk

In this issue

New version of *MLwiN*   **(Jon Rasbash)**
The new Centre web site   **http://www.cmm.bristol.ac.uk**
Bivariate longitudinal models with dropout   **(Nikos Pantazis & Giota Touloumi)**
Simulated maximum likelihood estimation for multilevel models   **(Edmond Ng)**

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

Centre for Multilevel Modelling, University of Bristol
Hilary Browne
hilary.browne@bristol.ac.uk.
new website - http://www.cmm.bristol.ac.uk

Multilevel Modelling centre of excellence

Our new web site which supersedes www.*MLwiN*.com is now on-line and fully-functioning

My brief as the Multilevel Modelling web developer was to construct a site that reflects on-line the reality of the Centre for Multilevel modelling as the global authority on multilevel modelling that we have become. The centre is the home for our multilevel modelling research and software development as well as production and sales of *MLwiN* software.

The new site has been constructed using the latest CSS layout techniques and adherence to international web-accessibility guidelines. Every page from *MLwiN*.com has been examined and the content fully-updated by the team, and will now be easier to find as more logical and numerous links have been put in to improve navigation. We hope you find the nested navigations bars helpful and logical.

New features

There is plenty of previously unseen new material, eg in the research section as well as information about the UK free-download offer and the new *MLwiN* version 2.1.

Search Bar
You can use the new search bar on the top right hand side which word-searches the whole of this site including the hundreds of PDF documents, or you may find our web site map helpful.

Harvey Goldstein will continue to run his own personal pages independently on the site: http://www.cmm.bristol.ac.uk/team/HG_Personal/index.shtml,. The search bar also searches all his web pages and PDFs.

Four new Forms
There are now four on-line forms where you can
1. order *MLwiN* software
2. apply for the new free *MLwiN* download if you are a UK-based academic
3. request technical support if you are a *MLwiN* customer
4. reserve multilevel modeling workshop places

Web Site Plans

We plan in the next few months to improve the on-line purchasing experience, and also produce on-line learning materials in several forms, such as podcasts and multilingual audio-visual presentations. Also interactive question-answer projects, many projects are currently being developed by the team.

_____

Do take a cyber-tour around the new web site, and let me know your feedback – email: hilary.browne@bristol.ac.uk.

New features in version 2.1 of *MLwiN*
Jon Rasbash
## Centre for Multilevel Modelling, University of Bristol
j.rasbash@bristol.ac.uk

The beta version of 2.1 will be available by December 2006. It will be announced on the MlwiN web site. Version 2.1 of *MLwiN* has three sets of new features:

- Improved model specification interface
- A new "bespoke predictions" window
- Surface plotting routines

Improved Model Specification Term

When variables and interactions were created in previous versions, recoding main effects(changing the centering or reference categories) did not update the interactions involving these changed variables. Interactions + main effects had to be removed from the model , main effects recoded and then main effects and interactions re-entered. This can be a time consuming and error prone process. In the new version, polynomials are created by specifying an optional order. If polynomial order, reference category (for a categorical variable) or centering value for a main effect is changed all associated interactions are updated.

The Specify Term window used to add terms and interactions into the model has been updated to include this extra functionaliy
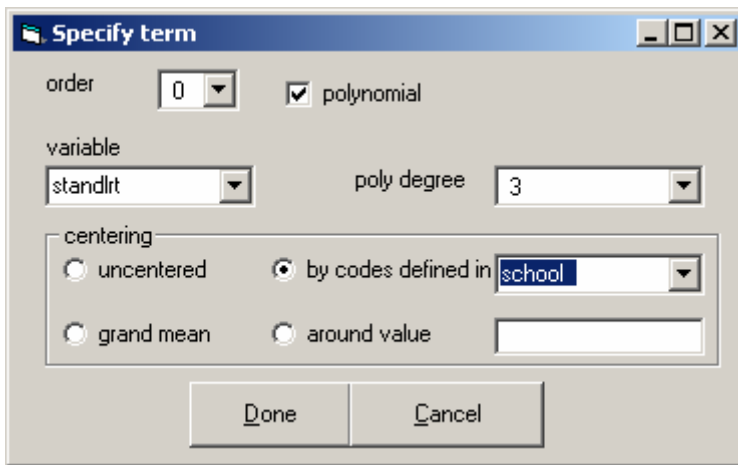
Bespoke Predictions Window

The current predictions window only predicts for values within the range of the observed data. Often we want to generate predictions for a specific ("bespoke") set of values for the explanatory variables to explore a multidimensional model predictions space.

The "bespoke" predictions window allows users to specify a set of values for each predictor variable and a prediction data set is created containing all combinations of these variables. Values of all interaction terms in the model can of course be automatically derived and need not be specified. Model predictions (and confidence intervals) are then calculated for the generated prediction data set. Another plot-specification window then facilitates specification of visulations of the prediction data set in up to 4 dimensions of predictor variable using colour coding and trellis plots.

Where the response is discrete, predictions can be on the link function scale for example log, logit, complementary-log-log, or the natural probability or count scale. This automation is particularly useful for multicategory response models where calculating probablities for each response category as a function of explanatory variables along with confidence intervals is less than straightforward. Predictions may be cluster specific or population average allowing users the ability to visualise and explore the consequences of these two different model assumptions.

 For example, below is a 2 level multinomial model from the MlwiN user guide. The model compares the odds of using different types of contraception: sterilised, modern or traditional against a common reference category of no contraception. These odds are modelled as functions of the number of previous children(lc0, lc1, lc2, lc3+) the age of the woman(age) and the interaction of these two variables.

Equations [_] [□] [×]

$\text{resp}_{ijk} \sim \text{Multinomial}(\text{cons}_{jk}, \pi_{ijk})$

$\log(\pi_{1jk} / \pi_{4jk}) = \beta_{0k}\text{cons.ster}_{ijk} + 0.076(0.033)\text{age.ster}_{ijk} + 1.959(0.365)\text{lc1.ster}_{ijk} + 2.158(0.357)\text{lc2.ster}_{ijk} +$
$\qquad 2.376(0.348)\text{lc3plus.ster}_{ijk} + 0.015(0.040)\text{age.lc1.ster}_{ijk} + -0.023(0.038)\text{age.lc2.ster}_{ijk} +$
$\qquad -0.099(0.035)\text{age.lc3plus.ster}_{ijk}$

$\beta_{0k} = -3.547(0.338) + v_{0k}$

$\log(\pi_{2jk} / \pi_{4jk}) = \beta_{1k}\text{cons.mod}_{ijk} + -0.033(0.019)\text{age.mod}_{ijk} + 0.953(0.240)\text{lc1.mod}_{ijk} + 0.896(0.231)\text{lc2.mod}_{ijk} +$
$\qquad 1.020(0.222)\text{lc3plus.mod}_{ijk} + 0.026(0.027)\text{age.lc1.mod}_{ijk} + -0.028(0.028)\text{age.lc2.mod}_{ijk} +$
$\qquad -0.051(0.023)\text{age.lc3plus.mod}_{ijk}$

$\beta_{1k} = -1.879(0.210) + v_{1k}$

$\log(\pi_{3jk} / \pi_{4jk}) = \beta_{2k}\text{cons.trad}_{ijk} + 0.007(0.024)\text{age.trad}_{ijk} + 0.713(0.311)\text{lc1.trad}_{ijk} + 0.999(0.282)\text{lc2.trad}_{ijk} +$
$\qquad 1.146(0.273)\text{lc3plus.trad}_{ijk} + 0.002(0.036)\text{age.lc1.trad}_{ijk} + 0.005(0.033)\text{age.lc2.trad}_{ijk} +$
$\qquad -0.016(0.027)\text{age.lc3plus.trad}_{ijk}$

$\beta_{2k} = -2.529(0.251) + v_{2k}$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \\ v_{2k} \end{bmatrix} \sim N(0, \ \Omega_v) \ : \ \Omega_v = \begin{bmatrix} 0.369(0.117) & & \\ 0.136(0.073) & 0.306(0.088) & \\ 0.018(0.074) & -0.030(0.065) & 0.257(0.093) \end{bmatrix}$$

| Name | Fonts | + | - | Add **T**erm | **E**stimates | Nonlinear | Clear | Notation | Responses | Help |

The bespoke predictions window for this model looks like this:

The top panel labelled **setup** allows specification of the explanatory variable values for the bespoke prediction grid. The lower screen allows specification of a percentile value for confidence and coverage intervals around predictions; whether predictions should be on the logit or probability scale; whether predictions are to be cluster specific or population average, and the final portion of the screen also allows coverage intervals based on between-district variation to be included.

A plotting window allows flexible specification of how the prediction results are to be plotted

For example, here we ask for the x variable to be age, the 4 response types to be grouped together on each graph and the different levels of lc(#of children) to be placed on a vertical trellis.

Pressing **Apply** creates:
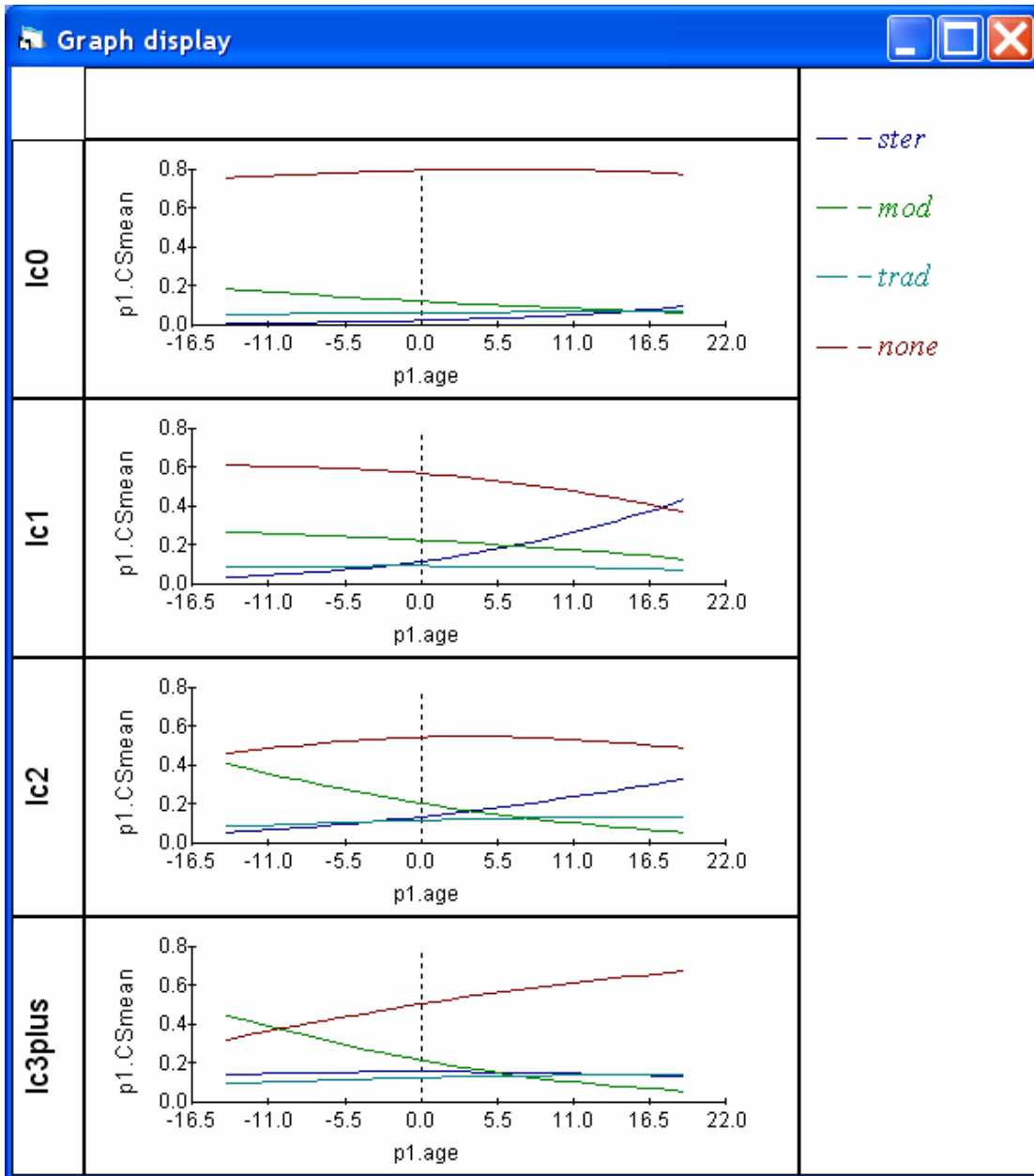
Similar bespoke prediction and visualisation functionality exists for Normal, Poisson, Binomial and negative binomial models.

Surface Plotting

Some basic surface plotting and rotation functions have been added. For example, below is a rotatable surface plot of the interaction between 2 continuous variables

---
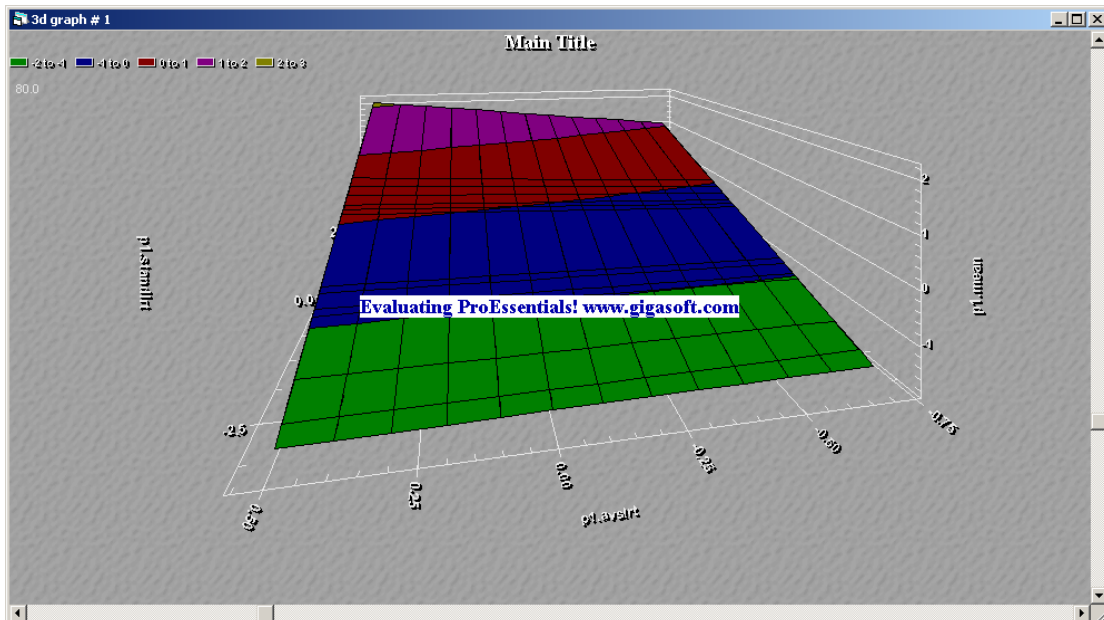
The new Multilevel Centre web site

In September 2006 the new web site for the Centre for Multilevel Modelling went live. Located at the University of Bristol this is a complete redesign by Hilary Browne (Hilary.Browne@bristol.ac.uk) which updates the software ordering, publications etc. The launch coincided with the availability of a free version for the UK academic community.

The new address is *****.

Fitting bivariate models for longitudinal data with informative drop-outs using MLwiN

Nikos Pantazis & Giota Touloumi

Department of Hygiene & Epidemiology, Athens Medical School

npantaz@med.uoa.gr

Introduction

In many research areas it is common for longitudinal data on one or more continuous response variables and event history data for a terminating event to be both collected for each individual participating in a study. A well-known example, which was also the motivation for this study, derives from the epidemiology of HIV-1 infection: two of the most commonly used markers of disease progression (CD4 cell count and HIV-RNA viral load) are routinely measured at various time points for each study participant along with the time of serious clinical events such as AIDS onset or death. One major statistical problem in such studies, where the main objective is to estimate the rate of change in continuous response variables, is that the series of repeated measurements are prematurely terminated for some individuals due to the occurrence of the terminating event. Several methods to analyze such unbalanced data are available provided the probability of the terminating event is not related to the response variable or is related only to the observed part of the vector of repeated measurements i.e. missing data are at random. According to the terminology introduced by Little and Rubin (1987), missing data can be classified as: completely at random (MCAR) when the probability of non-response does not depend on the response vector $\mathbf{Y}$ or at random (MAR) when the probability of non-response does not depend on the unobserved part of the responses ($\mathbf{Y_m}$) given the observed part ($\mathbf{Y_o}$). Finally, when the probability of non-response depends on unobserved outcomes, missing data are called non-ignorable (MNI). Wu and Carroll (1988) introduced the term "informative right censoring" for the special case of non-ignorable missing data where the probability of an individual terminating the study early is related to that subject's underlying marker's trend. Laird (1988) gives a detailed discussion of the impact of various missingness processes on inference about the response variable,

concluding that likelihood based approaches can provide valid inferences ignoring the missingness process when data are MCAR or MAR. However, serious biases can occur when missingness is non-ignorable. In such a case one needs to apply methods that simultaneously model the observed data and the missingness process.

Little (1995) summarized methods proposed to jointly model the measurement and drop-out process and outlined possible extensions. Wu and Caroll (1988) and Wu and Bailey (1989) studied the parametric conditional linear model which was later extended by Schluchter (1992) who proposed a log-normal model for the drop-out process and a linear random effects model for the longitudinal measurements. Maximum likelihood estimation in this model was accomplished using an EM algorithm. Faucett and Thomas (1996) and Touloumi *et al.* (1999) proposed similar parametric models where the estimation procedure was based on Bayesian methods (Gibbs sampling) or a combination of the restricted iterative generalized least squares method (RIGLS) with a nested EM algorithm respectively.

Most of the proposed methods that jointly model the repeated response measurements and the drop-out mechanism are dealing with data on a single marker. We have recently extended the Joint Multivariate Random Effects (JMRE) model proposed by Touloumi *et al.* (1999), in order to simultaneously model longitudinal data of two correlated continuous markers in the presence of informative drop-outs (Pantazis *et al.*, 2005). The bivariate JMRE model allowed also for non-linear trends in one of the two modelled continuous markers.

In this paper we briefly review the bivariate JMRE model and its estimation procedure, we focus on the data preparation and the use of the MLwiN statistical software to fit the model and present the most interesting results of the application of the model on simulated data. Concluding remarks and references to some

applications of the JMRE models are given in the final section.

Review of the bivariate JMRE model

Following Little's (1995) terminology, the bivariate JMRE model belongs to the class of "non-ignorable random-coefficient based drop-out" models. The model combines a bivariate linear mixed model for the underlying patterns of the markers' evolution with an accelerated failure time survival model for the drop-out process. The key feature of the model is that survival model's residuals are allowed to be correlated with both markers' random effects.

Bivariate markers' model

We assume that two continuous markers are measured repeatedly for each individual in the study. Superscripts c and r are used for these two markers, while superscript t is used to denote survival parts of the model. Subscripts j and i are used to denote data of the jth measurement on the ith individual. Assuming that the markers' trends can be described by two linear mixed models with correlated random-effects, a bivariate model for the longitudinal markers' measurements can be written as follows:

$$y_{ij}^c = \mathbf{x}_{ij}^c \mathbf{b}^c + \mathbf{z}_{ij}^c \boldsymbol{\beta}_i^c + e_{ij}^c$$
$$\text{and}$$
$$y_{ij}^r = \mathbf{x}_{ij}^r \mathbf{b}^r + \mathbf{z}_{ij}^r \boldsymbol{\beta}_i^r + e_{ij}^r$$

Vectors $\boldsymbol{\beta}_i^c$ (k^c x 1) and $\boldsymbol{\beta}_i^r$ (k^r x 1) contain random (subject-specific) regression coefficients for k^c and k^r explanatory variables included in the corresponding design vectors $\mathbf{z}_{ij}^c$ (1 x k^c) and $\mathbf{z}_{ij}^r$ (1 x k^r).

The joint distribution of $\boldsymbol{\beta}_i^c$ and $\boldsymbol{\beta}_i^r$ is assumed to be multivariate normal with zero mean and variance-covariance matrix:

$$\Sigma^{c,r} = \begin{bmatrix} \Sigma^c & \\ \Sigma^{cr} & \Sigma^r \end{bmatrix}$$

Matrices Σ^c (k^c x k^c) and Σ^r (k^r x k^r) include variances and covariances of the random effects for the first and second marker respectively, while Σ^{cr} (k^r x k^c) includes covariances between random effects of each marker.

Vectors $\mathbf{b}^c$ (p^c x 1) and $\mathbf{b}^r$ (p^r x 1) contain fixed regression coefficients for the two markers whereas $\mathbf{x}_{ij}^c$ (1 x p^c) and $\mathbf{x}_{ij}^r$ (1 x p^r)

are the corresponding design vectors containing the values of p^c and p^r explanatory variables respectively. Finally e_{ij}^c and e_{ij}^r are level-1 residuals for the two continuous markers.

The bivariate markers' model can be written in matrix form as follows:

$$\begin{bmatrix} \mathbf{Y}_i^c \\ \mathbf{Y}_i^r \end{bmatrix} = \begin{bmatrix} X_i^c & 0 \\ 0 & X_i^r \end{bmatrix} \begin{bmatrix} \mathbf{b}^c \\ \mathbf{b}^r \end{bmatrix} + \begin{bmatrix} Z_i^c & 0 \\ 0 & Z_i^r \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_i^c \\ \boldsymbol{\beta}_i^r \end{bmatrix} + \begin{bmatrix} \mathbf{e}_i^c \\ \mathbf{e}_i^r \end{bmatrix}$$

Survival model

Censoring of survival times due to loss to follow-up or due to study termination is assumed to be a totally random process. Let T_i be the observed survival time (*i.e.* time from some well defined time origin to the terminating event T_i^s or censored survival time C_i). Thus $T_i = \min(T_i^s, C_i)$. Survival time observations include also an indicator variable δ_i taking the value of 1 if $T_i = T_i^s$ and 0 otherwise. The survival model is a log-normal one with the following form:

$$\log(T_i^s) = \mathbf{x}_i^t \mathbf{b}^t + \beta_i^t$$

where $\mathbf{b}^t$ (v x 1) is the vector of regression coefficients and $\mathbf{x}_i^t$ (1 x v) is the design vector of v explanatory variables for the survival model. Finally β_i^t is the survival residual for the ith individual. Survival residuals are assumed to follow the normal distribution with zero mean and variance σ^{t^2}.

JMRE model

We assume that the joint distribution of the survival residuals β_i^t and the between-subjects markers' residuals $\boldsymbol{\beta}_i^c$, $\boldsymbol{\beta}_i^r$ is multivariate normal:

$$\boldsymbol{\beta}_i^{joint} = \begin{bmatrix} \boldsymbol{\beta}_i^c \\ \boldsymbol{\beta}_i^r \\ \beta_i^t \end{bmatrix} \square \ N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ 0 \end{bmatrix}, \Sigma^{joint} \right)$$

where $\Sigma^{jo\,\mathrm{int}} = \begin{bmatrix} \Sigma^c & & \\ \Sigma^{cr} & \Sigma^r & \\ \boldsymbol{\sigma}^{ct} & \boldsymbol{\sigma}^{rt} & \sigma^{t^2} \end{bmatrix}$.

Vectors $\boldsymbol{\sigma}^{ct}$ (1 x k^c) and $\boldsymbol{\sigma}^{rt}$ (1 x k^r) include covariances of markers' random effects and survival residuals.

The joint model can now be written in the usual mixed models formulation:

$$\mathbf{Y}_i^{\mathrm{joint}} = X_i^{\mathrm{joint}}\mathbf{b}^{\mathrm{joint}} + Z_i^{\mathrm{joint}}\boldsymbol{\beta}_i^{\mathrm{joint}} + \mathbf{e}_i^{\mathrm{joint}}$$

where

$$\mathbf{Y}_i^{jo\,\mathrm{int}} = \left[\mathbf{Y}_i^{c^T}, \mathbf{Y}_i^{r^T}, \ln(T_i^s) \right]^T$$

and

$$\mathbf{b}^{jo\,\mathrm{int}} = \left[\mathbf{b}^{c^T}, \mathbf{b}^{r^T}, \mathbf{b}^{t^T} \right]^T.$$

It is assumed that level-1 residuals are uncorrelated given the random effects and have constant variances σ^{c^2} and σ^{r^2} for the first and second marker respectively.

Fitting Procedure

For uncensored survival data (i.e. $T_i = T_i^s$ for all individuals) all the model parameters can be estimated using the Restricted Iterative Generalized Least Squares (RIGLS) method or the Restricted Maximum Likelihood (REML) method in any statistical software package with reasonable flexibility in the definition of the variance matrix. The RIGLS method, assuming multivariate Normality, yields estimates which are equivalent to the REML ones and has been described in detail elsewhere (Goldstein, 1995).

With censored survival data, however, the standard RIGLS method has to be modified in order to obtain unbiased estimates of the model parameters. To deal with censored survival data a version of the EM algorithm is applied, considering censored survival observations as missing data. Thus at each iteration and only for individuals with censored survival observations, the survival part of the response variable (i.e. the logarithm of censored survival times) and the corresponding survival components of the residuals cross-product matrix are replaced by their conditional expectations given the observed data and the

current parameter estimates (E-step). New parameters' estimates are obtained via RIGLS (M-step) and these two steps are repeated until convergence. The calculation of these conditional expectations is based on multivariate normal theory (Johnson and Wichern, 1998) and properties of the truncated normal distribution (Johnson and Kotz, 1970). The null hypothesis $[\boldsymbol{\sigma}^{ct}, \boldsymbol{\sigma}^{rt}] = [\mathbf{0}, \mathbf{0}]$ (that is, non-informative drop-out) can be tested by the likelihood ratio (LR) test, estimating the likelihood of the joint model and the model in which $[\boldsymbol{\sigma}^{ct}, \boldsymbol{\sigma}^{rt}]$ is constrained to $[\mathbf{0}, \mathbf{0}]$. However it should be noted that the validity of the LR test for non-ignorability depends critically on the model's assumptions (Kenward, 1998).

Due to the replacement of censored survival times by their conditional expectations during the fitting procedure, without taking into account the uncertainty of these expectations, standard errors of the fixed effects parameters can be underestimated. This effect is of minimal magnitude especially on standard errors of markers' parameters but if the focus is on the survival model one can use the Louis (1982) method or alternatively a computationally simpler approach based on a modified version of the multiple imputations method to adjust the standard errors (Rubin, 1978; Touloumi *et al.*, 2003). Therefore hypotheses testing based on Wald tests should be performed after the adjustment of standard errors, especially for factors affecting the survival part of the model. Alternatively one can use likelihood ratio tests to compare nested models, provided that the IGLS algorithm is used for estimation or the fixed part of the models under consideration is the same, in case of estimation through the restricted version of the algorithm (RIGLS).

The basic model, as described above, can be used for linear or piecewise linear average markers' trends. However in our motivation example one of the two continuous markers was characterized by a fast initial exponential like decline followed by a slow, almost linear, increase thereafter. In order to describe this trend with reasonable flexibility we used a nonlinear model. The model used is a mixture of a linear model describing the ultimate trend and an exponential one to capture the initial decline. More specifically the model for this marker has the following form:

$$Y_{ij}^r = b_0^r + \beta_{0i}^r + (b_1^r + \beta_{1i}^r)t_{ij} + b_2^r\,\mathrm{e}^{-b_3^r t_{ij}} + e_{ij}^r$$

The specific model is relatively simple, allows the estimation of various quantities of clinical interest (e.g. time from infection until minimum virus load levels) and has easy interpretable parameters.

In order to estimate all model's parameters we used a "linearization" method based on a first order Taylor approximation (McCullagh and Nelder, 1989; Goldstein, 1995) incorporated in the main JMRE fitting algorithm.

Implementation in MLwiN

As noted in the previous section if the survival time is observed for all individuals the model can be easily fitted in MLwiN requiring only some appropriate data preparation. More specifically repeated measurements for the first and second marker along with the logarithm of the survival time should be stacked underneath each other for each individual to form the response vector $\mathbf{Y}_i^{joint}$. A similar strategy should be followed for the design matrices for fixed and random effects. In the following table, data for the first individual are shown; id is a sequential identifier for the individuals and yvar is the combined response vector. consc, consr and conss are indicator variables for the first marker, the second marker and the survival measurements respectively but will also be used as parts of the appropriate design matrices. timec and timer are the actual times for the measurements of the first and second marker respectively.

id	yvar	consc	timec	consr	timer	conss
1	27.6	1	6.26	0	0	0
1	30.8	1	6.53	0	0	0
1	25.6	1	6.81	0	0	0
1	28.7	1	7.01	0	0	0
1	27.3	1	7.32	0	0	0
1	26.9	1	7.60	0	0	0
1	27.5	1	7.85	0	0	0
1	2.4	0	0	1	6.53	0
1	2.3	0	0	1	6.81	0
1	2.2	0	0	0	0	1

Data should also contain a constant variable of ones (cons). Any available covariates that may be included in the joint design matrices should be arranged in similar way as the consc, consr, conss, timer, times variables (for example if the age of an individual is considered it should be entered as

three different variables agec, ager, ages which will contain the actual age or zero according to the corresponding indicator variable). The following code declares the model without any additional covariates (linear average trend, random intercept and random slope for both markers) and starts its estimation:

```
resp "yvar"

expl "consc" "timec" "consr" "timer"
"conss"

iden 1 "cons"

iden 2 "id"

setv 1 "consc" "consr"

clre 1 "consc" "consr"

setv 2 "consc"  "timec" "consr" "timer"
"conss"

batc

start
```

With censored survival data for some of the individuals, the aforementioned procedure must be adjusted to incorporate the nested EM algorithm. More specifically we have used the PREFile MLwiN command to invoke a macro implementing the EM algorithm prior to each RIGLS iteration. Alternatively one can use the "Options -> Directories -> Pre file" dialog to invoke the macro. An additional indicator variable taking the value of one if the survival time is right censored and zero otherwise should be included in the data in order to discriminate between censored and actual survival times respectively. The macro called by the PREFile command estimates the expected survival time for each individual with censored survival time and the corresponding part of the residuals cross-product matrix conditional on current estimates and the data and substitutes these observed quantities with the expected ones, prior to each RIGLS iteration until convergence. Initial estimates can be obtained running a few RIGLS iterations treating censored survival times as observed (*i.e.* not invoking the EM algorithm). Finally additional macros have been developed to implement the "linearization" method for the non-linear trend for one of the two markers and the multiple imputation adjustments of the standard errors. All additional macros used in the fitting of the bivariate JMRE model are available upon request from the authors.

Simulations

We performed various simulation studies to evaluate the performance of the bivariate JMRE model in terms of bias and precision. The performance of the proposed model was compared also against simpler models such as two independent univariate JMRE models, a bivariate random effects model and two independent univariate random effects models. Finally, the robustness of the model was investigated by examining the effects of violating some of its distributional assumptions.

500 data sets were generated under the following assumptions: (a) a linear trend for the first marker (CD4 cell counts on the square-root scale) allowing for individual level random deviations in both the intercept and the slope, (b) a non-linear model, as described in the fitting procedure section, for the second marker (HIV– RNA viral load levels on the $\log_{10}$-scale), (c) survival times follow a log-normal distribution, (d) random effects of marker trends and survival residuals have a joint distribution which is a multivariate normal distribution and (e) missing markers' data can be due to death or onset of AIDS (informative censoring) or due to withdrawal at 5% per year (assumed completely at random). The CD4 cell data were truncated when the value of 0 was reached. The choice of the parameters was based on results from real examples. More specifically, correlation between marker intercepts was assumed to be -0.39 and between CD4 cell count slope and HIV–RNA ultimate slope -0.52 and the survival time was assumed to be correlated mainly with the HIV–RNA intercept (r=-0.37) and CD4 cell count slope (r=0.42). Each data set contained 200 subjects and the study duration was assumed to be 20 years with assessment for both markers every 3 months.

Table 1 shows the results of the simulation study for the fixed effects parameters. Under the label 'true' are the parameter values that were used to generate the data. As shown in Table 1 the parameters which are most affected by ignoring informative censoring are the average rate of CD4 cell decline and the ultimate slope of the viral load. The bias in the slopes is in the direction expected: the rates of CD4 cell loss and viral load increase are underestimated. Subjects with steeper CD4 cell declines or faster viral load ultimate increases tend to have shorter survival times, resulting in higher proportions of informatively missing markers' data and thus these subjects are downweighted during the random-effects estimation procedure, resulting in overoptimistic estimates.

Regarding the bias in the random-effects parameters, the bivariate JMRE model yields the least biased estimates compared with the estimates from the other three approaches. The bias in the estimated variances and covariances ranged from -3.6% to 10.8% under the bivariate JMRE model, from -30.0% to 30.6% for the two independent JMRE models and -15.2% to 26.3% under the other two approaches. For the two independent JMRE models the biases were particularly large for the covariances between survival time and CD4 cell count intercept and slope and with HIV–RNA ultimate slope. Of course, the two independent JMRE models approaches do not provide estimates of the correlation between the two markers, whereas the other two approaches, which also ignore informative censoring, do not provide any information about the relationship between survival time and marker values.

Empirical coverage probabilities for the marker intercepts were between 94.2% and 95.2% in all four approaches. In contrast the bias in the estimation of CD4 cell counts and HIV–RNA slopes resulted in low values of empirical coverage probabilities in the two univariate random effects models (57.8% and 67.4% respectively) and the bivariate random-effects model (75.2% and 77.6% respectively) whereas the two independent JMRE models (91.6% and 91.8% respectively) and the bivariate JMRE model (95.0% and 94.2% respectively) performed much better.

Table 1. Simulation results: estimates, percentage bias and mean-squared errors by the four approaches; the bivariate JMRE model, two independent random-effects models, the bivariate random-effect model and two independent JMRE models

Parameter	TRUE	Bivariate JMRE model Estimate (%Bias) (MSE)	Two independent RE models Estimate (%Bias) (MSE)	Bivariate RE model Estimate (%Bias) (MSE)	Two independent JMRE models Estimate (%Bias) (MSE)
CD4 intercept $\sqrt{\text{cells}/\mu l}$)	*25.72*	25.702 -0.07 0.131	25.605 -0.45 0.131	25.631 -0.34 0.131	25.671 -0.19 0.131
CD4 slope ($\sqrt{\text{cells}/\mu l}$ per year)	*-1.131*	-1.119 1.06 0.00672	-0.998 11.73 0.00591	-1.032 8.74 0.00584	-1.09 3.66 0.00662
HIV-RNA intercept ($\log_{10}$ copies/ml)	*3.874*	3.876 0.06 0.00363	3.89 0.41 0.00361	3.885 0.28 0.00357	3.88 0.16 0.00363
HIV-RNA ultimate slope ($\log_{10}$ copies/ml per year)	*0.077*	0.075 -2.77 0.000136	0.06 -21.57 0.000117	0.063 -17.88 0.000117	0.071 -7.39 0.000131
HIV-RNA Nonlinear parameter b_2 ($\log_{10}$ copies/ml)	*1.13*	1.129 -0.08 0.00137	1.116 -1.24 0.00135	1.121 -0.84 0.00136	1.125 -0.42 0.00137
HIV-RNA Nonlinear parameter b_3 (years^{-1})	*4*	4.008 0.2 0.104	4.073 1.83 0.109	4.038 0.94 0.105	4.029 0.73 0.107
Log Survival time (years)	*2.3*	2.305 0.23 0.01046			2.248 -2.27 0.00934

Further simulations were performed in order to asses the performance of the bivariate JMRE model when data deviate from its assumptions. Various scenarios with level-2 (random effects) or level-1 residuals' distributions being heavy tailed or skewed (generated from multivariate t or Gumbel distributions respectively) were evaluated. Moreover, we assessed the effects of misspecifying the drop-out model and violating the assumption of conditional independence of the two markers level-1 residuals. Results, based on 500 replications for each scenario, showed that although variance components estimation was affected by heavy-tailed or skewed random effects distributions, fixed effects parameters were only marginally affected (bias 0.0% - 1.1% and 0.1% - 5.0% respectively). However 95% coverage probabilities were lower than expected especially when random effects distributions were skewed (93.8% - 95.2% and

83.2% – 95.6% respectively). Deviations from the assumed drop-out model or from level-1 residuals' distribution had smaller effects on both fixed effects and variance components.

Discussion

The analysis of repeated measurements data has been an active methodological research area for several years. The wide availability of appropriate statistical software has also facilitated the analysis of such data in many applications. However, incompleteness of longitudinal data is a common issue which should be taken into account at least as part of a sensitivity analysis. During the last two decades many statistical methods have been proposed for the analysis of repeated measurements of a continuous variable in presence of some non-ignorable missingness process. We have recently proposed a model

for simultaneously analysing longitudinal data of two correlated continuous variables subject to informative drop-out (Pantazis *et al.*, 2005). The fitting procedure of this model was based on the RIGLS method combined with a nested EM algorithm and allowed also for a non-linear average trend for one of the two modelled variables. A similar bivariate model was also recently proposed (Thiebaut *et al.*, 2005) and its estimation algorithm was based on direct maximization of the likelihood using Monte Carlo methods for the integration of random effects. Roy and Lin (2002) also have presented a latent variable approach for the analysis of such multivariate longitudinal data where the drop-out model that was used is similar to the model that was proposed by Diggle and Kenward (1994).

The proposed model could be in principle fitted using a Bayesian approach. It should be noted that for the single marker case, Faucett and Thomas (1996) have proposed a similar model where a Markov chain Monte Carlo technique (Gibbs sampling) is used to estimate the joint posterior distribution of the unknown parameters of the model.

The proposed bivariate JMRE model can be fitted in a computationally easy way within the MLwiN statistical software package requiring some minimal data preparation and an additional macro for the implementation of the nested EM algorithm. As shown in the simulation study the bivariate JMRE model achieves the greater bias reduction in all model's parameters compared to the two independent JMRE models and especially the simple random effects models which ignore the informativeness of the missing data. Moreover, empirical coverage probabilities for the fixed effects were close to the nominal 95% level. Moderate violations of its distributional or drop-out assumptions had negligible effect on fixed effects estimation whereas their standard errors were marginally affected especially with non-skewed random effects distributions.

The JMRE models have thus far been applied to data derived from a big collaborative multicohort study within the framework of the natural history of the HIV-1 infection. For a comparison of results obtained by the univariate JMRE model and by a simple random effects model see CASCADE Collaboration (2003). For an application of the bivariate JMRE model where one of the two variables is modelled through a piecewise linear random effects model see (Touloumi *et al.,* 2004). Finally for a comparison of the univariate and bivariate JMRE models allowing for non-linear trends in one of the two modelled markers applied to real data see (Pantazis *et al.*, 2005.)

A drawback of the macros used to fit the bivariate or univariate JMRE model, in their current state, is their dependence on the structure of the model under consideration. Slight modifications are needed to fit models with different random effects structures or different explanatory covariates in the survival part of the model.

References

CASCADE Collaboration. Differences in CD4 cell counts at seroconversion and decline among 5739 HIV-1-infected individuals with well-estimated dates of seroconversion (2003) *J Acquir Immune Defic Syndr*, 34, 76-83.

Diggle, P. and Kenward, M. G. (1994) Informative drop-out in longitudinal data analysis (with discussion). *J R Statist Soc C*, 43, 49–93.

Faucett, C. L. and Thomas, D. C. (1996) simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Stat Med*, 15, 663-85.

Goldstein, H. (2003) *Multilevel Statistical Models*, 3rd edn, London: Arnold.

Johnson, N. L. and Kotz, S. (1970) *Continuous Univariate Distributions*, vol. 1. New York: Wiley.

Johnson, R. A. and Wichern, D. W. (1998) *Applied Multivariate Statistical Analysis*. Englewood Cliffs: Prentice Hall.

Kenward M.G. (1998) Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Stat Med*, 1998 17, 2723-32.

Laird, N. M. (1988) Missing data in longitudinal-studies. *Stat Med*, 7, 305–315.

Little, R. J. A. (1995) Modeling the drop-out mechanism in repeated-measures studies. *J Am Statist Ass*, 90, 1112–1121.

Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J R Statist Soc B*, 44, 223-233

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn., pp. 379–380. London: Chapman and Hall.

Pantazis, N., Touloumi, G., Walker, A. S. and Babiker, A. G. (2005) Bivariate modelling of longitudinal measurements of two human immunodeficiency type 1 disease progression markers in the presence of informative drop-outs. *J R Statist Soc C*, 54, 405-423

Roy, J. and Lin, X. H. (2002) Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: changes in methadone treatment practices. *J Am Statist Ass*, 97, 40–52.

Rubin, D. (1978) Multiple imputations in sample surveys—a phenomenological bayesian approach to nonresponse. *Proc Surv Res Meth Sect Am Statist Ass*, 20–34.

Schluchter, M. D. (1992) Methods for the analysis of informatively censored longitudinal data. *Stat Med*, 11, 1861–1870.

Thiebaut R, Jacqmin-Gadda H, Babiker A, Commenges D; CASCADE Collaboration (2005) Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of CD4+ cell count and HIV RNA viral load in response to treatment of HIV infection. *Stat Med*, 24, 65-82.

Touloumi, G., Babiker, A. G., Kenward, M. G., Pocock, S. J. and Darbyshire, J. H. (2003) A comparison of two methods for the estimation of precision with incomplete longitudinal data, jointly modeled with a time-to event outcome. *Stat Med*, 22, 3161–3175.

Touloumi, G., Pocock, S. J., Babiker, A. G. and Darbyshire, J. H. (1999) Estimation and comparision of rates of change in longitudinal studies with informative drop-outs. *Stat. Med*, 18, 1215–1233.

Touloumi G, Pantazis N, Babiker AG, Walker SA, Katsarou O, Karafoulidou A, Hatzakis A, Porter K; CASCADE Collaboration (2004) Differences in HIV RNA levels before the initiation of antiretroviral therapy among 1864 individuals with known HIV-1 seroconversion dates. *AIDS*, 18, 1697-705.

Wu, M. C. and Bailey, K. R. (1989) Estimation and comparison of changes in the presence of informative right censoring—conditional linear-model. *Biometrics*, 45, 939–955.

Wu, M. C. and Carroll, R. J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics*, 44, 175–188.

Simulated maximum likelihood estimation for multilevel models with binary outcomes

Edmond Ng
University of Bristol

Introduction

Multilevel models for discrete data are extremely useful in a wide range of social science and medical settings, where the principal outcome is discrete and the data are naturally clustered. In our example, we shall consider data on the use of prenatal care during pregnancies in a developing country where children are nested within families.

However, unlike the case where the response in Normally distributed, when (restricted) maximum likelihood estimates can be readily obtained, fitting such models is not straightforward. This is because the discrete distribution of the response means that the likelihood is computed by integrating a product of discrete and normal densities, which has no analytical solution (see Rabe-Hesketh et al., 2005) and requires numerical techniques that become very time-consuming when the number of random parameters becomes large.

One option to tackle this problem is to use Monte-Carlo methods. In this article we aim to demonstrate the use of the technique of Simulated Maximum Likelihood (SML) (McCulloch, 1997) to fit multilevel models with binary responses.

Background

Consider two-level data consisting, for example, of the same number of repeated observations on individual subjects. Let y_i denote the ith response from a subject, $(i = 1, \ldots, I)$, and assume that responses from different subjects are independent. For the sake of simplicity the subscript, j, for subject, is omitted throughout. Let u denote a p-vector of subject-specific random effects, assumed to follow a multivariate normal distribution $N(0, \Sigma)$, with density $f(u, \Sigma)$. Conditional on subject-specific random effects and covariates, the two-level multilevel model assumes individual repeated observations are independent. They follow a discrete distribution, denoted by $g(y_i \mid u, \beta, x_i, z_i)$, $(i = 1, \ldots, I)$, where

$$E[y_i \mid u, x_i] = \lambda^{-1}(x_i^T \beta + z_i^T u),$$

x_i is a (qx1) vector of covariates at the ith response with (qx1) coefficient vector β, z_i is a (px1) vector

of covariates for the random effects and λ is the link function (e.g. logistic, probit, log).

The likelihood for the I observations on an individual is then

$$L(y; \beta, \Sigma) = \int \prod_{i=1}^{I} \{g(y_i \mid u, \beta, x_i, z_i)\} f(u, \Sigma) du, \quad (1)$$

where the integral is over the p-dimensional normal distribution of the random effects. As already discussed, when the response model, *g*, is normal (and λ is the identity link), the integral is tractable and (restricted) maximum likelihood estimates can be found. However, when the response is discrete, there is no closed form expression for the integral. One possibility is to obtain approximate maximum likelihood estimates. We replace the likelihood above by a quasi-likelihood approximation (marginal or penalised), which can be maximised using the algorithm for normal models (Breslow and Clayton, 1993; Goldstein and Rasbash, 1996). However, the quasi-likelihood parameter estimates are known to be downwardly biased towards zero when the data are 'sparse' and/or the fitted probabilities are close to zero or one (Rodríguez and Goldman, 1995; Lin and Breslow, 1996). The other is numerical or Monte-Carlo integration, which we will describe below.

Simulated Maximum Likelihood (SML)

In SML Monte-Carlo integration is used to repeatedly evaluate the likelihood given by (1) and a search is carried out over plausible values of β and Σ to find those which give the maximum.

We now describe how this can be done using importance sampling (see, e.g., Ripley, 1987, p 122). Briefly, if a random variable X has a probability density function $p(x)$, and $q(x)$ is a probability density function defined on the same support, then, for any function k,

$$\int k(x) p(x) dx$$
$$\cong \frac{1}{H} \sum_{h=1}^{H} k(x_h) p(x_h) / q(x_h), \quad (2)$$

where $(x_1, \ldots x_H)$ is an independent identically distributed sample from the density $q(x)$, and the approximation becomes exact as $H \rightarrow \infty$. The

probability density $q(x)$ is known as the *importance density*.

We can apply this idea to (1) as follows. Let $\widetilde{\Sigma}$ be the quasi-likelihood estimate of Σ. Choose as the importance density the multivariate normal density with mean 0 and covariance matrix $\widetilde{\Sigma}$, which we write as $f(u, \widetilde{\Sigma})$. Simulate a large number, H, of draws from this distribution, $(u_1, \ldots, u_H)$. Then applying the importance principle (2),

$$L(y; \beta, \Sigma)$$

$$= \int \prod_{i=1}^{I} \{g(y_i \mid u, \beta, x_i, z_i)\} f(u, \Sigma) du$$

$$\cong \frac{1}{H} \sum_{h=1}^{H} \prod_{i=1}^{I} \frac{\{g(y_i \mid u_h, \beta, x_i, z_i)\} f(u_h, \Sigma)}{f(u_h, \widetilde{\Sigma})}. \quad (3)$$

This approximation can be used in two ways. First, we can obtain an estimate of the log-likelihood at the quasi-likelihood parameter estimates by setting (β, Σ) equal to the quasi-likelihood estimates $(\widetilde{\beta}, \widetilde{\Sigma})$. In this case the importance ratio will be equal to 1.

Secondly, by repeatedly calculating the right hand side of this expression at various values of (β, Σ), we can search for the parameter values that maximise the simulated likelihood. Note that we do not simulate a new sample, $(u_1, \ldots, u_H)$, as the search progresses. Rather, we simulate a sample at the beginning, and, conditional on this, find (β, Σ) to maximise the Monte-Carlo estimate. If a new sample is drawn, then the estimate of the log-likelihood will be slightly different at the same parameter values, and this generally will lead to the optimising software reporting an error.

In order to search for the maximum, the optimising software requires the derivative of (3) with respect to (β, Σ). While this can be worked out numerically, this is prohibitively slow for practical use. We have, therefore, calculated the derivatives of the right hand side of (3), with respect to (β, Σ), for a sample of size H. Expressions for these are given in Appendix A in Ng *et al.* (2006).

Once the maximum has been found, the analytic expressions for the derivative can be used to compute the asymptotic variance matrix of the parameter estimates.

Two-stage algorithm approach to obtain maximum likelihood estimates

McCulloch (1997) reports that SML only works well if the importance distribution is fairly accurate. If this is not the case, then one option is to use the iterative procedure described in Appendix B in Ng *et al.* (2006). Although, given time, this should lead to the maximum likelihood estimates, it is too slow for practical use; further it ignores the fact that good starting values are available from the quasi-likelihood estimates. The most accurate of these are the 2nd order Penalised Quasi-Likelihood (PQL) estimates (Goldstein and Rasbash, 1996). Accordingly, we propose the following algorithm:

Two-stage algorithm:
1. Fit the model using second-order PQL to give estimates denoted $(\widetilde{\beta}, \widetilde{\Sigma})$.

2. Set the importance density $\Sigma_I = \widetilde{\Sigma}$ and simulate $(u_1, \ldots, u_H)$ from the multivariate normal distribution with covariance matrix Σ_I;

3. Fixing $(u_1, \ldots, u_H)$, find the values of (β, Σ) that maximise the Monte-Carlo estimate of the likelihood (3), using the derivatives in Appendix A in Ng *et al.* (2006). Report these as the maximum likelihood estimates.

4. Use the analytical form of the first derivatives to obtain a numerical estimate of the second derivative of the log-likelihood function at the maximum. Invert this for an estimate of the variance matrix of the parameter estimates.

Example

In the following example we give the results of applying SML to the analysis of a data set and compare them against those from other alternatives such as EM-Laplace2 (implemented in HLM version 6.02) (Raudenbush *et al.*, 2000), MCMC (implemented in MLwiN 2.02) (Browne, 2002) and numerical (adaptive) quadrature methods (implemented in GLLAMM (Generalised Linear Latent And Mixed Models) (Rabe-Hesketh *et al.*, 2005) in Stata 8.2 and proc NLMIXED in SAS release 8.

Data

In these data there is a binary response on 737 children (level 1) from 479 families (level 2), and three covariates. Fifty six percent of families have only one child, making this a 'sparse' data set for which quasi-likelihood estimates are most likely to suffer noticeable bias.

We fit the following model, where the binary response from child i in family j is denoted Y_{ij} :

$$\text{logit } \{\Pr (Y_{ij} = 1)\}$$
$$= (\beta_0 + u_j) + \beta_1 X1_{ij} + \beta_2 X2_{ij} + \beta_3 X3_{ij},$$

where u_j is the family specific random effect with variance σ_u^2. Using (i) 2nd order PQL, and (ii) second-order PQL followed by simulated maximum likelihood, as in our two-stage algorithm above, gives the results in Table 1.

The 2nd order PQL estimates of random and fixed parameters are consistent with the downward biases found in Ng *et al.* (2006, Table 1). This is most marked for the variance, σ_u^2, which is 14% smaller than the SML estimate. Further, the standard errors appear to be underestimated by 2nd order PQL. Again, this is most marked for the standard error of the estimate of σ_u^2.

Additionally, SML, like the methods based on numerical integration, gives an estimate of the log-likelihood at the maximum, shown in the bottom row of the table above. This is useful for comparing models. For example, refitting the model without *X1* gives $-2 \times$ log-likelihood = 793.0 (1 dp). The likelihood ratio test for the importance of *X1* is thus 793.0 – 791.4 = 1.6, which comparing with χ_1^2 is not significant at the 5% level. As expected, this agrees, to 1 dp, with the Wald test $(0.581/0.466)^2$ =1.6. Also the SML estimates for the fixed and the random parameters are very close to those from the two numerical integration methods, proc NLMIXED and GLLAMM.

Relative to SML and NLMIXED, EM-Laplace2 parameter estimates appear too small; in particular the random parameter, σ_u^2, is 45% less than the SML estimate. The reported deviance from EM-Laplace2 is more than twice those from SML and the other two numerical methods, presumably because it includes an additional constant term from the likelihood. Changes in the deviance are similar, though.

The MCMC estimate for σ_u^2 is about 27% higher than the SML estimate. The Gibbs sampling with Gamma diffuse priors (α =0.001 and β =0.001) used in the MCMC run in MLwiN is based on a burn-in of 5,000 iterations and a chain-length of 100,000. It was run for another 400,000 iterations to confirm the relatively large level 2 variance. The mean (SD), median and effective sample size (assuming successive values in chain independent) of the resulting chain are 3.511 (1.339), 3.126 and 2,014. Our results are probably due to the fact that, although the prior is diffuse, it gives zero weight to a zero variance, so that when there is relatively little information in the data, both the posterior mean and median are pulled away from zero relative to the maximum likelihood estimate. Interestingly, the Bayesian Deviance Information Criterion, DIC, is fairly close to those from SML and the numerical methods.

Table 1: Parameter estimates for the model for the data from PQL(2), SML, EM-Laplace2, MCMC, proc NLMIXED and GLLAMM

	2nd order PQL	2nd order PQL followed by SML(H=3000)	EM-Laplace2	MCMC	Proc NLMIXED	GLLAMM
Estimation method						
Fixed parameters:						
β_0	0.905 (0.244)	0.954 (0.275)	0.862 (0.209)	1.020 (0.299)	0.962 (0.276)	0.960 (0.274)
β_1	0.573 (0.445)	0.581 (0.466)	0.487 (0.418)	0.644 (0.491)	0.596 (0.471)	0.599 (0.470)
β_2	1.431 (0.218)	1.501 (0.258)	1.344 (0.208)	1.594 (0.282)	1.515 (0.259)	1.514 (0.259)
β_3	1.798 (0.351)	1.883 (0.408)	1.712 (0.316)	2.014 (0.449)	1.905 (0.410)	1.901 (0.408)
Random parameters:						
σ_u^2	2.402 (0.473)	2.800 (1.110)	1.546 (NA)	3.569 (1.399)	2.894 (1.087)	2.874 (1.073)
–2 log-like.	NA	792.0	2146.4	(DIC) 754.7	791.5	791.5

Note: NA=Not available; DIC=Deviance Information Criterion; NLMIXED used adaptive quadrature; GLLAMM: 8 quadrature points were used.

Comparison with numerical integration methods in SAS

To conclude, we present a further comparison of PQL followed by SML with the numerical integration method implemented in SAS. We found the latter considerably faster than Stata's GLLAMM (see below for some timings) especially when there are many random effects in the model.

We simulated data from a two-level logistic model,

$$\text{logit}\ \{\Pr\ (Y_{ij} = 1)\}$$
$$= (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})X1_{ij}$$
$$+ (\beta_2 + u_{2j})X2_{ij} + (\beta_3 + u_{3j})X3_{ij},$$

where

$$\beta = [1111]',\ X1 \text{ to } X3 \sim N(0,1)\ \text{and}$$

$$\begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \end{bmatrix} \sim N\left(0, \begin{bmatrix} 0.5 & & & \\ 0.05 & 0.5 & & \\ 0.05 & 0.05 & 0.5 & \\ 0.05 & 0.05 & 0.05 & 0.5 \end{bmatrix}\right).$$

There are 200 level 2 units with 1 to 19 level 1 units each.

In the comparison, models with increasing numbers of random effects are fitted. Results from our 'two-stage algorithm' are compared with those from proc NLMIXED. The optimisation technique used in proc NLMIXED is quasi-Newton and adaptive Gaussian quadrature is used to integrate over the random effects numerically. Parameter estimates and standard errors were very similar for both methods (results not shown). Table 2 shows the relative time taken to converge, the log-likelihood value, and absolute value of the largest gradient at convergence.

The log-likelihood values are comparable between the two methods. The gradients at convergence are lower in NLMIXED. However, imposing a more stringent convergence criterion on SML would only be sensible if H were increased. Elapsed time goes up at an exponential rate in NLMIXED while we were able to achieve comparable accuracy with a more linear increase in time using SML.

The similar parameter estimates, standard errors and log-likelihoods are encouraging. Since the relative increase in computational time for quadrature with increasing numbers of random effects is much faster, SML would appear to have the edge for models with a large number of random effects. We have not optimised our code in a low-level language, so it is plausible that further optimisation could bring the baseline times for the two methods closer. In addition, as discussed below, SML has the potential to extend to non-normal random effects and more than two levels.

Table 2: Results from Simulated ML (H=5000) and SAS proc NLMIXED

Random parameters in model	Simulated Maximum Likelihood			SAS proc NLMIXED		
	Time (minutes)	Max. |gradient|	Log-likelihood	Time (minutes)	Max. |gradient|	Log-likelihood
1	3.4	0.004121	-1033.097	0.05	1.723e-6	-1033.021
3	4.9	0.017214	-1019.218	0.4	0.000719	-1019.143
6	30.0	0.006752	-1006.041	4.65	0.000437	-1005.950
10	86.1	0.004275	-995.822	32.35	0.00323	-997.042

Finally, note that GLLAMM took 22 seconds, 4.5 minutes and 93 minutes for the models with 1, 3 and 6 random parameters. The model with 10 random parameters had not converged after 15 hours.

Discussion

We have demonstrated the application of SML to fit two-level models for binary responses. Our results show that SML performs comparably to the other methods, and has important advantages when the number of random parameters becomes large.

Our experience confirms that of McCulloch (1997), that SML needs good starting values to converge to the maximum likelihood estimates. Using SML after 2nd order PQL therefore seems sensible. Our code, written in Matlab, can be generalised to include other distributions (Poisson, negative binomial) and link functions. The code is available from the website of the

journal, Statistical Modelling (http://stat.uibk.ac.at/SMIJ).

Finally, note that the SML approach is complementary to fitting similar models using a Bayesian approach, using MCMC methods in MLwiN, WinBUGS (Spiegelhalter *et al.* 1999) or other Bayesian software. Advantages of the SML approach are that it is unnecessary to specify priors for variance terms, which even if notionally 'uninformative' may turn out to have unforeseen effects (see Table 1); that a likelihood is available for testing, and that Monte-Carlo error can be relatively straightforwardly addressed (e.g. Booth and Hobert, 1999).

This article is an abridged version of Ng *et al.* (2006).

References

Booth, J. G. and Hobert, J. P. (1999) Maximising generalised linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal statistical Society, Series B*, **61**, 265–285.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Browne, W. J. (2002). *MCMC estimation in MLwiN*. London: Institute of Education.

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

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

Goldstein, H. and Rasbash, J. (1996) Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, **159**, 505–513.

Matlab (2003). Release 13, The Mathworks Inc. (www.mathworks.com).

Lin, X. and Breslow, N. E. (1996) Bias correction in generalised linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, **91**, 1007–1016.

McCulloch, C. E. (1997) Maximum likelihood algorithms for generalised linear mixed models. *Journal of the American Statistical Association*, **92**, 162–170.

Ng, E. S. W., Carpenter, J. R., Goldstein, H. and Rasbash, J. (2006) Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling*, **6**, 23-42.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, **128**, 301-323.

Raudenbush, S. W., Yang, M. L., Yosef, M. (2000) Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal Of Computational And Graphical Statistics*, **9**: 141-157.

Rodríguez, G. and Goldman, N (1995) An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, **158**, 73-89.

Spiegelhalter, D. J., Thomas, A. and Best, N. G. (1999). *WinBUGS version 1.2 user manual*. Cambridge: MRC Biostatistics Unit.