Applied multilevel modelling — an introduction

James Carpenter
Email: jrc@imbi.uni-freiburg.de
www.missingdata.org.uk

London School of Hygiene & Tropical Medicine

and

Institute for Medical Biometry, Freiburg, Germany

Friday, 16th December 2005
Acknowledgements

- Harvey Goldstein
- Mike Kenward
- Bill Browne
- ESRC funding
Objectives

By the end of the morning I hope you will:

- understand the key concepts of multilevel modelling;
- be able to relate multilevel modelling to a range of applications;
- have an overview of the research area, and
- want to try it yourself!
I: Multilevel models for continuous data

Outline

- Multilevel structure
- Terminology
- Examples: (i) school effects (ii) asthma trial
- Multilevel models: random effects
- Analysis of schools data
- Modelling stationary correlation: the variogram
- Analysis of asthma data
- Estimation
- Software
- Summary
II: Extensions

- Multilevel models for discrete data
  - Marginal vs Conditional models
  - Example: clinical trial
  - Estimation issues
  - Software

- Models for cross-classified data
  - Notation
  - Example: Scottish educational data

- Models for multiple membership cross-classified data
  - Notation
  - Example: Danish poultry farming
Multilevel data

Multilevel data arise when some observations are related to each other. For example:

- clustering
  - of children in classes in schools in education authorities
  - of patients in cluster randomised clinical trials
  - of survey respondents in households
  - ...

- repeated measures (another form of clustering)
  - of babies’ weights in their first year;
  - of patients in a clinical trial;
  - ...

The more you think about it, the more you see multilevel structures.
Terminology

Multilevel models have been used in a variety of research areas, each with their own emphasis.

Thus, a variety of names are used to describe the same broad class of models:

- Hierarchical models
- Random effect models
- Mixed models
- Longitudinal data models
- ...

Most of these are special cases of what I will call *multilevel models.*
Example 1

Effect of school on educational achievement.

Exam scores of 4059 children age 16.

Children belong to one of 65 schools; between 8 and 198 students from each school.

We also have a reading test score at age 11, amongst other variables.

Interest focuses on the ‘value added’ by schools.

Level 2: Schools

Level 1: Exam results, 16 years
Data

![Scatter plot showing the relationship between standardised literacy score at age 11 years and standardised exam score at age 16 years. The plot is a scatter plot with two axes: the x-axis represents standardised literacy score at age 11 years, ranging from -3 to 3, and the y-axis represents standardised exam score at age 16 years, ranging from -2 to 2. The data points are scattered around the plot, indicating a positive correlation between the two scores.]
Plot of results age 16, 11 for each school

Mean literacy of students entering a school age 11

Mean school exam score, age 16
Example 2

471 patients with severe asthma randomised to receive one of 4 levels of an active drug or placebo.

Each patient was allocated to one of 27 investigators.

Patients were followed up 2, 4, 8 & 12 weeks after randomisation; in addition they kept daily diaries.

Level 3: Investigators

Level 2: Patients

Level 1: FEV readings
**Completion by treatment arm**

<table>
<thead>
<tr>
<th>Treatment group</th>
<th>No. randomised</th>
<th>No. completing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>91</td>
<td>38</td>
</tr>
<tr>
<td>A, 100 mcg</td>
<td>91</td>
<td>72</td>
</tr>
<tr>
<td>A, 200 mcg</td>
<td>92</td>
<td>84</td>
</tr>
<tr>
<td>A, 400 mcg</td>
<td>99</td>
<td>88</td>
</tr>
<tr>
<td>A, 800 mcg</td>
<td>98</td>
<td>90</td>
</tr>
</tbody>
</table>
Data

![Graph showing Time from randomisation vs FEV1 (litres)]
Mean profiles

- Placebo
- A, 100 mcg
- A, 200 mcg
- A, 400 mcg
- A, 800 mcg

Weeks since randomisation

FEV1 (litres)
Some patient profiles

- Patient: 5035
- Patient: 5106
- Patient: 5119
- Patient: 5120
- Patient: 5155
- Patient: 5221
- Patient: 5224
- Patient: 5280
- Patient: 5333
- Patient: 5362
- Patient: 5386
- Patient: 5402
- Patient: 5412
- Patient: 5432
- Patient: 5474
- Patient: 5481
- Patient: 5574
- Patient: 5580
- Patient: 5609
- Patient: 5806
Notation

Focus on two level data for now.

Let $i \in (1, \ldots I)$ index level 1 units (e.g. students), and $j \in (1, \ldots J)$ index level 2 units (e.g. schools).

Let $y_{ij}$ be the response, with covariate $x_{ij}$.

We will use the (multivariate) normal distribution for modelling.

This is characterised by its mean and variance; our focus will be on modelling the variance.
Standard linear regression

A model for a straight line is

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + e_{ij}, \quad e_{ij} \sim iid N(0, \sigma^2), \]

which assumes all level 1 units independent.

In other words, if we have \( I \) level 1 units for each of \( J \) level two units, then writing

\[ y' = (y_{11}, y_{21}, \ldots, y_{I1}, y_{12}, \ldots, y_{I1}, \ldots, y_{IJ}, \ldots, y_{IJ})', \]

\[ V_y = \sigma^2 \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & \ddots & \ldots & 0 \\ 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}, \quad \text{a } IJ \text{ by } IJ \text{ matrix.} \]
Problems when data are multilevel:

1. Standard errors are wrong

Hence so are hypothesis tests, p-values and confidence intervals.

Subjects within a cluster (level 2 unit) are often similar to each other, i.e. not independent. They therefore convey less information about the value of a parameter than an independent (unclustered) sample of the same size (Goldstein, 2003, p. 23).

Further, we would like to understand the sources of variability by modelling the variance. This is not possible with standard regression.
2. Effect estimates can be misleading

Suppose we wish to relate literacy at age 11 to exam results at age 16.

An analysis which does not recognise the likely correlation between outcomes from students from the same school will give equal weight to results from all students.

However, given that such correlation exists, and data are unbalanced (different numbers of students in different schools) it will be better to give relatively more weight to children from smaller schools than larger schools.
Multilevel model

Recall i—level 1 (e.g. student), j—level 2 (e.g. school).

Model: \( y_{ij} = x_{ij} \beta + error_{ij} \).

So, \( \text{Var } y_{ij} = \text{Var } error_{ij} \), as mean \( x'_{ij} \beta \) constant.

Let \( y'_j = (y_{1j}, y_{2j}, \ldots, y_{Ij}) \) — the data for level 2 unit \( j \).

Suppose this has \( I \) by \( I \) covariance matrix \( \Sigma \).

The diagonal elements are \( \text{Var}[error_{ij}] \),

The off diagonal elements are \( \text{Cov}[error_{ij}, error_{i'j}] \).
Block diagonal covariance matrix

The full data can then be written as a $IJ$ length column vector

\[ y' = (y'_1, y'_2, \ldots y'_J). \]

Key assumption: data from different level 2 units independent

This means the full covariance matrix of $y$ is:

\[
\Sigma_{\text{full}} = \begin{pmatrix}
\Sigma & 0 & 0 & \ldots & 0 \\
0 & \Sigma & 0 & \ldots & 0 \\
0 & 0 & \ddots & \ldots & 0 \\
0 & 0 & \ldots & \Sigma & 0 \\
0 & 0 & \ldots & 0 & \Sigma
\end{pmatrix}.
\]
Unstructured model

We attempt to estimate every one of the $I(I + 1)/2$ parameters in $\Sigma$.

For obvious reasons, this is known as an *unstructured covariance matrix*.

Provided we have enough data, this can work well for designed studies (where everyone is observed at the same times) such as the asthma trial.

For the educational data (and generally in epidemiology and social sciences) where the children in each school have very different literacy at age 11, it is a non-starter.

We therefore need to consider modelling the variance.
Variance model

Key idea: Decompose $error_{ij}$ into three independent components:

1. Person specific Random effects, denoted $U_j$
   — captures differences between individuals.

2. Stationary correlation $w_{ij}$
   — captures systematic local variation

3. Residual error
   — everything that’s left over.

So model becomes

$$y_{ij} = x_{ij}' \beta + z_{ij}' U_j + w_{ij} + e_{ij},$$

and

$$\Sigma = \Sigma_u + \Sigma_w + \Sigma_e.$$
Residual error

Denote the residual error by $e_{ij}$.

Model $e_{ij} \sim \mathcal{N}_1(0, \Omega_e = \sigma^2_e)$, — ie all independent of each other.

Write $e_j' = (e_{ij}, e_{2j}, \ldots, e_{IJ})$, then $e_j \sim \mathcal{N}_I(0, \Sigma_e)$, where

$$
\Sigma_e = \sigma^2_e \begin{pmatrix}
1 & 0 & 0 & \ldots & 0 \\
0 & 1 & 0 & \ddots & 0 \\
0 & 0 & \ddots & 1 & 0 \\
0 & 0 & \ldots & 0 & 1
\end{pmatrix}
$$

is an $I$ by $I$ matrix.
Random effects

$U_j$ is $d$ length row vector of level 2 (e.g. school) specific random effects.

Pre-multiply by $d$-length column vector $z_{ij}$ to give impact of level-2 specific effect on observation $i$.

Example 1: Random intercepts:

$U_j = u_{0j} \sim N(0, \sigma_{u0}^2)$, $z_{ij} = 1$ for all $i$.

$\text{Var}(z_{ij}'u_{0j}) = \text{Var}(u_{0j}) = \sigma_{u0}^2$; $\text{Cov}(z_{ij}'u_{0j}, z_{ij}'u_{0j}) = \text{Cov}(u_{0j}, u_{0j}) = \sigma_{u0}^2$.

Thus

$$\Sigma_u = \begin{pmatrix}
\sigma_u^2 & \cdots & \sigma_u^2 \\
\vdots & \ddots & \vdots \\
\sigma_u^2 & \cdots & \sigma_u^2
\end{pmatrix}.$$
Example 2: random intercepts and slopes

\[ U_j = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u0}^2 & \sigma_{u0,u1} \\ \sigma_{u0,u1} & \sigma_{u1}^2 \end{pmatrix} \right). \]

Then \( z_{ij} = (1, x_{ij}) \), so effect of random terms on school \( j \) at student \( i \) is

\[ u_{0j} + x_{ij}u_{1j}. \]

Now,
\[
\text{Var}(z'_{ij}u_{0j}) = \text{Var}(u_{0j} + x_{ij}u_{1j}) = \sigma_{u0}^2 + x_{ij}^2\sigma_{u1}^2 + 2x_{ij}\sigma_{u0,u1};
\]
\[
\text{Cov}(z'_{ij}u_{0j}, z'_{i'j}u_{0j}) = \text{Cov}(u_{0j} + x_{ij}u_{1j}, u_{0j} + x_{i'j}u_{1j}) = \\
\sigma_{u0}^2 + x_{ij}x_{i'j}\sigma_{u1}^2 + (x_{ij} + x_{i'j})\sigma_{u0,u1}.
\]

Hence \( \Sigma_u \), which is rather messy!
Putting this together

Putting the random intercepts and error model together gives

\[ Y_{ij} = (\beta_0 + u_{0j}) + \beta_1 x_{ij} + e_{ij} \]

\[ u_{0j} \sim N(0, \sigma_u^2) \]

\[ e_{0j} \sim N(0, \sigma_e^2) \]

\[ \text{Var}Y_{ij} = \sigma_u^2 + \sigma_e^2 \quad \text{(gives diagonal elements of } \Sigma) \]

\[ \text{Cov}(Y_{ij}, Y_{i'j}) = \sigma_{u0}^2 \quad \text{(gives off-diagonal elements of } \Sigma) \]

Hence, correlation between results of students in the same school is the same, no matter how far apart their initial literacy scores.
Random intercepts: schematic

\[ Y = (\alpha + u_2) + \beta x \]
\[ Y = (\alpha + u_1) + \beta x \]
\[ Y = \alpha + \beta x \]
\[ Y = (\alpha + u_3) + \beta x \]
\[ Y = (\alpha + u_4) + \beta x \]
\[ Y = (\alpha + u_5) + \beta x \]

Response

literacy (11 years), x
And for random intercepts and slopes:

\[ Y_{ij} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j})x_{ij} + e_{ij} \]

\[ \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u0}^2 & \sigma_{u0u1} \\ \sigma_{u0u1} & \sigma_{u1}^2 \end{pmatrix} \right] \]

\[ e_{0j} \sim N(0, \sigma_e^2) \]

\[ \text{Var} Y_{ij} = \sigma_{u0}^2 + x_{ij}^2 \sigma_{u1}^2 + 2\sigma_{u0u1}x_{ij} + \sigma_e^2 \]

(gives diagonal elements of \( \Sigma \))

\[ \text{Cov}(Y_{ij}, Y_{i'j}) = \sigma_{u0}^2 + x_{ij}x_{i'j} \sigma_{u1}^2 + (x_{ij} + x_{i'j})\sigma_{u0u1} \]

(gives off-diagonal elements of \( \Sigma \))

Now correlation between results of students in the same school can decline as distance between their literacy scores increases.
Random intercepts & slopes: schematic

\[ Y = (\alpha + u) + (\beta + v)x \]

(Note: \( u_{0j} \leftrightarrow u_j; \quad u_{1j} \leftrightarrow v_j \).)
Fitting random intercepts model

\[ \text{normexam}_{ij} \sim N(XB, \Omega) \]

\[ \text{normexam}_{ij} = \beta_{0ij}\text{cons} + 0.563(0.012)\text{standlrt}_{ij} \]

\[ \beta_{0ij} = 0.002(0.040) + u_{0j} + e_{0ij} \]

\[ \begin{bmatrix} u_{0j} \\ e_{0ij} \end{bmatrix} \sim N(0, \Omega_u): \Omega_u = \begin{bmatrix} 0.094(0.018) \end{bmatrix} \]

\[ \begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e): \Omega_e = \begin{bmatrix} 0.566(0.013) \end{bmatrix} \]

\[-2*\text{loglikelihood (IGLS Deviance)} = 9357.251 (4059 of 4059 cases in use)\]
Plotting school level residuals

(Plot shows $\hat{u}_{0j} \pm \sqrt{2\text{Var} \hat{u}_{0j}}$)
Interpretation

Can we deduce that schools at the right hand end are better - that they give students with the same 11 year literacy a better education? (jargon term: \textit{value added})

\begin{center}
\textbf{The UK department of Education does!}
\end{center}

\textbf{BUT}

Our model assumed each school has the same slope: this is equivalent to the correlation between students’ 16 year results in a school being the same, \textit{no matter how far apart their 11 year literacy}.

Let’s fit a random intercepts and slopes model to test this.
Random intercept & slope model

\begin{align*}
normexam_{ij} & \sim N(XB, \Omega) \\
normexam_{ij} & = \beta_{0ij}\text{cons} + \beta_{1ij}\text{standlrt}_{ij} \\
\beta_{0ij} & = -0.012(0.040) + u_{0j} + e_{0ij} \\
\beta_{1ij} & = 0.557(0.020) + u_{1j}
\end{align*}

\[
\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.092(0.018) \\ 0.018(0.007) & 0.015(0.004) \end{bmatrix}
\]

\[
\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.554(0.012) \end{bmatrix}
\]

\(-2*\text{loglikelihood(IGLS Deviance)} = 9316.888\) (4059 of 4059 cases in use)

Log likelihood ratio test: \(9357.3 - 9316.9 = 40\); cf \(\chi^2_2\),

\[p = 2.1 \times 10^{-9}.\]
School specific slopes

(Note school lines do not extend out of range of their students’ 11 year literacy intake)
Interpretation

Picture is complex!

Some schools appear good for students with high literacy at 11, but less good for students with low literacy at 11. For some schools it is the reverse.

This may reflect specific teaching strategies.

Some schools are poor overall.

Use of a random intercepts model alone for these data is misleading for parents and a disservice to schools.

Back to variance model

We decomposed $\Sigma = \Sigma_u + \Sigma_w + \Sigma_u$. Now consider $\Sigma_w$.

If $\text{Var} y_{ij}$ increases (say with time), such as in growth data, the variance/covariance is *non-stationary*.

However, if the variance is constant over time and the covariance depends only on the time between observations, it is *stationary*.

We have seen that random effects (e.g. random intercepts and slopes) often give good models for non-stationary processes.

However, they are not so good for stationary processes. Even if the variance is non-stationary, there may be a stationary component.
Stationary correlation structure

Again have \( w'_j = (w_{ij}, w_{2j}, \ldots, w_{Ij}) \).

Recall trials example; see individuals repeatedly over times \( i = 1, 2, \ldots, I \). A stationary covariance process has \( I \) by \( I \) matrix \( \Sigma_w = \)

\[
\begin{pmatrix}
1 & \rho(|1 - 2| = 1) & \cdots & \rho(|1 - I| = I - 1) \\
\rho(|2 - 1| = 1) & 1 & \cdots & \rho(|2 - I| = I - 2) \\
\vdots & \vdots & \ddots & \vdots \\
\rho(|(I - 1) - 1| = I - 2) & \cdots & 1 & \rho(|(I - 1) - I| = 1) \\
\rho(|I - 1| = I - 1) & \cdots & \rho(|I - (I - 1)| = 1) & 1
\end{pmatrix}
\]

where \( \rho(0) = 1 \), and \( \rho(|\omega|) \to 0 \) as \( |\omega| \to \infty \).

Forms for \( \rho(\cdot) \) include AR(1), \( \exp(-\alpha|\omega|) \), \( \cdots \).
Drawing it together

Recall trials example.
For an individual with $I$ measurements, their variance/covariance matrix is the $I$ by $I$ matrix

$$
\Sigma = \Sigma_u + \Sigma_w + \Sigma_e,
$$

Do we need these three components? Depends on data set

- Some problems (typically growth) observations sufficiently spaced out that $\Sigma_w = 0$.
- Or, perhaps because of a standard manufacturing process association will be dominated by $\Sigma_w$ and $\Sigma_u = 0$.

Remaining error, $\Sigma_e \neq 0$!
The variogram

The sample variogram gives a graphical guide to the correlation structure. It is less prone to local fluctuations than the sample correlation matrix.

It is valid for limited non-stationary data, provided the increments $Y_{(i+\omega)} - Y_i$ do not depend on $i$.

If the variance process is stationary, let $\omega$ be the time between observations, and $\rho(\omega)$ the correlation between them.

Then the variogram is

$$\gamma(\omega) = \sigma^2(1 - \rho(\omega)),$$

where $\sigma^2$ is the ‘process’ variance.

The variogram goes to 0 as $\omega \rightarrow \infty$.  

slide 40/61
Estimating the variogram

Fit a ‘full’ model to the data, i.e. a model with the most general mean and variance structure you are interested in.

Calculate

1. the residuals, $e_{ij}$;
2. the half-squared differences, $\nu_{ijk} = \frac{1}{2}(e_{ij} - e_{ik})^2$.
3. the corresponding time differences, $\delta_{ijk} = t_{ij} - t_{ik}$.

Then the sample variogram, $\gamma(\omega)$, is the average of all the $\nu_{ijk}$ corresponding for which $\delta_{ijk} = \omega$.

Finally, estimate the ‘process variance’ by the average of all

$$\frac{1}{2}(y_{ij} - y_{lk}), \text{ for which } i \neq l.$$ 

For more details see Diggle et al. (2002).
Interpreting the variogram

\[ \gamma(\omega) \]

- **Process variance**
- **Random intercept variance** $\sigma_u^2$
- **Serial correlation** $\tau^2$
- **Error variance** $\sigma_e^2$
Application to asthma study

We estimate the variogram to guide the choice of variance model for the asthma study.

In designed studies, such as this, the choice ‘full model’ or ‘unstructured model’ is fairly clear.

We fit a different mean for each treatment group at each time, and an unstructured covariance matrix.

We also fit a random term at a 3rd level — investigator.
Asthma study: model

Let $\phi_{il} = 1$ if patient $i$ has treatment $l = 1, \ldots, 5$

We have observations at 2, 4, 8 & 12 weeks, corresponding to $j = 1, 2, 3, 4$.

Let $k$ index investigator. The model is

$$y_{ijk} = \phi_{il} \beta_{jl} + u_j + \nu_k,$$

$$\nu_k \sim N(0, \sigma^2_\nu)$$

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} \sim N(0, \Sigma), \text{ (unstructured; 10 parameters})$$
Unstructured model for asthma data

\[
\text{fev}_{ijk} \sim N(XB, \Omega)
\]

\[
\text{fev}_{ijk} = 2.156(0.087)\text{bd100v1}_{ijk} + 2.197(0.088)\text{bd100v2}_{ijk} + 2.204(0.091)\text{bd100v3}_{ijk} + 2.185(0.090)\text{bd100v4}_{ijk} + 2.315(0.087)\text{bd200v1}_{ijk} + 2.290(0.087)\text{bd200v2}_{ijk} + 2.344(0.090)\text{bd200v3}_{ijk} + 2.317(0.088)\text{bd200v4}_{ijk} + 2.219(0.085)\text{bd400v1}_{ijk} + 2.258(0.085)\text{bd400v2}_{ijk} + 2.310(0.088)\text{bd400v3}_{ijk} + 2.312(0.086)\text{bd400v4}_{ijk} + 2.172(0.085)\text{bd800v1}_{ijk} + 2.228(0.085)\text{bd800v2}_{ijk} + 2.255(0.088)\text{bd800v3}_{ijk} + 2.275(0.086)\text{bd800v4}_{ijk} + 1.951(0.087)\text{placv1}_{ijk} + 1.909(0.089)\text{placv2}_{ijk} + 1.888(0.095)\text{placv3}_{ijk} + 1.907(0.096)\text{placv4}_{ijk} + \mu_{20k}2\text{weeks}_{ijk} + \mu_{21k}4\text{weeks}_{ijk} + \mu_{22k}8\text{weeks}_{ijk} + \mu_{23k}12\text{weeks}_{ijk} + \nu_{24k}\text{cons}
\]

\[
\begin{bmatrix}
\nu_{24k}
\end{bmatrix} \sim N(0, \Omega_{v}) : \Omega_{v} = \begin{bmatrix}
0.047(0.021)
\end{bmatrix}
\]

\[
\begin{bmatrix}
\mu_{20k} \\
\mu_{21k} \\
\mu_{22k} \\
\mu_{23k}
\end{bmatrix} \sim N(0, \Omega_{u}) : \Omega_{u} = \begin{bmatrix}
0.529(0.035) \\
0.462(0.033) & 0.534(0.036) \\
0.475(0.035) & 0.485(0.035) & 0.570(0.039) \\
0.470(0.034) & 0.459(0.034) & 0.490(0.036) & 0.540(0.038)
\end{bmatrix}
\]

\(-2\text{loglikelihood(IGLS Deviance)} = 1812.285\) (1696 of 1696 cases in use)
Correlation virtually constant over time.
Error variance $\approx 0.1$.
Random intercept variance $\approx 0.5$. 
Simpler variance model

Variogram suggests random intercepts model adequate.

Fit this model and compare with unstructured model:

<table>
<thead>
<tr>
<th>Model</th>
<th>No. variance parameters</th>
<th>$-2 \times \ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstruc. var</td>
<td>10</td>
<td>1831.2</td>
</tr>
<tr>
<td>Rand. ints</td>
<td>2</td>
<td>1845.9</td>
</tr>
</tbody>
</table>

Difference in $-2 \times \log$ likelihood is 14.7.

Compare with $\chi^2_8$, p=0.064.

The variogram can

- greatly simplify choosing the variance model, and
- be used to compare the sample and model correlation.

Non-stationary processes must be made approximately stationary first.
Asthma study: fitted vs observed means

FEV1 (litres) vs Weeks since randomisation

- **raw data**
- **model**

- **Active, 100mcg**
- **Placebo**

Estimated means valid if data MAR.
Borderline investigator effect vanishes after baseline adjustment.
(Restricted) maximum likelihood

Our model is $y \sim MN(X\beta, \Sigma)$, with log-likelihood

$$-0.5\{\log |\Sigma| + (y - X\beta)'\Sigma^{-1}(y - X\beta)\}.$$  

Maximise by direct Newton-Raphson search (e.g. Raudenbush and Bryk (2002), ch. 14). Often, iterative generalised least squares is faster (see below).

To correct the downward bias of ML estimators, use REML log-likelihood

$$-0.5\{\log |\Sigma| + (y - X\beta)'\Sigma^{-1}(y - X\beta) + \log |X'\Sigma^{-1}X|\}.$$  


For moderately large data sets the results are similar, though in uncommon situations with many fixed parameters the two can give quite different answers (Verbeke and Molenberghs, 2000, p. 198).
Testing

Changes in REML log-likelihoods cannot generally be used to compare nested models unless $X$ is unchanged. So maximum likelihood may be preferred for model building (Goldstein, 2003, p. 36).

**MLwiN** always gives non-REML likelihood

Often Wald, or F-tests used for fixed effects, and likelihood ratio tests for random effects.

Asymptotically, fixed effects estimates ($\hat{\beta}$) and variance term estimates independent; however, standard errors can be too small in small data sets.

Kenward and Roger (1997) give an adjustment, implemented in SAS, for this.
IGLS estimation

Goldstein (1986) showed how to maximise the multivariate normal likelihood iteratively using two regressions.

Recall:

1. If \( y_i = x_i' \beta + \epsilon_i \), \( \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \) then

   \[
   \hat{\beta} = (X'X)^{-1}X'Y.
   \]

2. If \( y_i = x_i' \beta + \epsilon_i \), \( \epsilon_i \not\overset{iid}{\sim} N(0, \sigma^2) \) but Cov\((Y) = \Sigma\), known

   \[
   \hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y.
   \]
Model:  \( Y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + e_{ij} \)

\[
\mathbf{E}Y = \begin{pmatrix}
  1 & x_{11} \\
  1 & x_{21} \\
  \vdots & \vdots \\
  1 & x_{IJ}
\end{pmatrix}
\begin{pmatrix}
  \beta_0 \\
  \beta_1
\end{pmatrix}
\]

1. Guess \( \Sigma \), ie guess \( \text{Cov}(Y) \), set \( \hat{\beta}^1 = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y \)

2. Calculate residuals, \( R = Y - X\hat{\beta} \)

3. **IDEA** Since \( \mathbf{E}(RR') = \Sigma \), the covariance matrix of \( Y \), create a regression model to estimate \( \Sigma \).
   After estimating \( \Sigma \), update \( \beta \) is step 1.

Iterate till convergence
How do we estimate $\Sigma$?

$$\mathbf{E}(R' R) = \begin{pmatrix}
  r_{11}^2 & r_{12} r_{11} & \cdots \\
  r_{21} r_{11} & r_{22}^2 & \cdots \\
  \vdots & \ddots & \ddots
\end{pmatrix}
= \begin{pmatrix}
  \sigma_{u0}^2 + \sigma_e^2 & \sigma_{u0}^2 & \cdots \\
  \sigma_{u0}^2 & \sigma_{u0}^2 + \sigma_e^2 & \cdots \\
  \vdots & \ddots & \ddots
\end{pmatrix}$$

So $\text{vec}(R' R) = \begin{pmatrix}
  r_{11}^2 \\
  r_{21} r_{11} \\
  \vdots
\end{pmatrix}
= \begin{pmatrix}
  1 & 1 \\
  0 & 1 \\
  1 & 1 \\
  \vdots & \vdots
\end{pmatrix}
\begin{pmatrix}
  \sigma_e^2 \\
  \sigma_{u0}^2
\end{pmatrix}
+ \omega$

$\uparrow \gamma$

$\leftarrow \mathbf{Z}$
Estimation of $\sum$ (ctd)

A regression, BUT $r$’s not iid; turns out

$$\text{Cov}(\text{vec}(RR)) = \Sigma^* = \Sigma \otimes \Sigma$$

Thus

$$\hat{\Gamma} = (Z'\Sigma^{*-1}Z)Z'\Sigma^{*-1}R$$

Full details in Goldstein and Rasbash (1992).

For discussion of residuals in multilevel modelling see, e.g. Robinson (1991).
Software

Most packages offer software for continuous data.

Some comments on packages I have tried to use (no implied judgment on other packages):

**SAS, proc MIXED** — the commercial airliner (expensive)

Excellent for fitting a wide range of standard models; provides a wide range of (stationary) covariance models.

Kenward-Roger adjustment to standard errors/degrees of freedom for small samples implemented (ddfm=kr). Not intuitive for beginners

Probably the best for regulatory pharma-industry analyses
Software (ctd)

**MLwin** — the light aircraft (free to UK academics)

Fast, flexible multilevel modelling package — but you may crash!

Fits an extremely wide range of models.

Very strong on random effects models; weak on stationary covariance models.

Intuitive for teaching.

**Stata** (GLLAMM) — overland travel (≈ 200 USD)

Has a very general, but very slow, multilevel/latent variable modelling package. A faster package for continuous data with Stata 9.0.
Software (ctd)

**WinBUGS** (free)

Very flexible package for model fitting using MCMC, with R/S+ like syntax.

I like to have a good idea what the answer is *before* I use it.

**R** (free)

Non-linear estimation package limited compared with SAS and MLwiN;

Trellis graphics and data manipulation awkward.

Syntax hard for teaching

[For up-to-date reviews (re)-visit www.mlwin.com]
Summary

- Once you look, multilevel structure is everywhere.
- This is an important aspect of the data; ignoring it leads to misleading results.
- Distinguish between the mean model and the correlation model; understand the impact of random effects models on the correlation.
- Estimation: REML is usually preferable. Don’t use changes in REML likelihood for fixed effects inference, though.
- Get started using the tutorial material at http://tramss.data-archive.ac.uk/Software/MLwiN.asp
- Further references: see Carpenter (2005).
References


II: Extensions

- Multilevel models for discrete data
  - Marginal vs Conditional models
  - Example: clinical trial
  - Estimation issues
  - Software

- Models for cross-classified data
  - Notation
  - Example: Scottish educational data

- Models for multiple membership cross-classified data
  - Notation
  - Example: Danish poultry farming
Review: random intercepts

Suppose \( j \) indexes subjects (level 2) and \( i \) indexes repeated observations (level 1).

Recall the random intercepts model:

\[
y_{ij} = (\beta_0 + u_j) + \beta_1 x_{ij} + e_{ij}
\]
\[
u_j \sim N(0, \sigma_u^2)
\]
\[
e_{ij} \sim N(0, \sigma_e^2).
\]

If \( y_{ij} \) is now binary, a natural generalisation is

\[
\logit \Pr(y_{ij} = 1) = (\beta_0 + u_j) + \beta_1 x_{ij}
\]
\[
u_{0i} \sim N(0, \sigma_u^2)
\]
Interpretation

In the previous model logit \( \Pr(y_{ij} = 1) \) is really logit \( \Pr(y_{ij} = 1 | x_{ij}, u_j) \).

Thus, this is a \textit{Subject Specific} (SS) model; \( \beta \) is the log-odds ratio of the effect of \( x \) for a specific \( u_j \).

We can also construct \textit{Population Averaged} (PA) models

\[
\logit \Pr(y_{ij} = 1 | x_{ij}) = \beta_0^p + \beta_1^p x_{ij}
\]

Define expit as the inverse logit function. Then

\[
E_u \expit(\beta_0 + \beta_1 x_{ij} + u_j) \neq \expit(\beta_0 + \beta_1 x_{ij}).
\]

So PA estimates do not equal SS estimates.

The exception is the normal model, where expit is replaced by the identity.
Interpretation (ctd)

Let $\beta$ denote SS estimates, $\beta^p$ PA estimates. In general

- $|\beta^p| < |\beta|$, with equality if $\sigma_u^2 = 0$.
- If $\sigma_u^2$ is large, the two will be very different.

Example

$y_{ij}$ indicates coronary heart disease for subject $j$ at time $i$

$x_{1ij}$ indicates the subject smokes

$x_{2ij}$ indicates one of their parents had CHD

$\beta^p_1$: effect of smoking on log-odds of CHD in population

$\beta_1$: effect of stopping smoking for a given subject

$\beta^p_2$: Effect of parental CHD in the population

$\beta_2$: Effect of change in parental CHD for given individual
Choosing between PA and SS

It depends on the question:
SS more appropriate when population averaged effects are important:

- In epidemiology & trials, where we are interested in the average effect of treatment on a population
- In education, when we are interested in the average effects of interventions across populations

BUT: the same process in two populations with different heterogeneity will lead to different $\beta_p$.

SS preferable for

- Estimating effects on individuals, schools
- Modelling sources of variability.
Example: Clinical trial

241 patients randomised to receive either a placebo or active treatment.

Three treatment periods.

In each period, each patient undergoes a series of tests (the number varying between patients).

Each test has a binary response (1=success).

<table>
<thead>
<tr>
<th>Dropout</th>
<th>Treatment group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Placebo</td>
</tr>
<tr>
<td>Period 1</td>
<td>0</td>
</tr>
<tr>
<td>Period 2</td>
<td>10</td>
</tr>
<tr>
<td>Period 3</td>
<td>25</td>
</tr>
<tr>
<td>Completers</td>
<td>82</td>
</tr>
</tbody>
</table>
No. of tests by period and intervention

Period 1 | Period 2 | Period 3 | Period 1 | Period 2 | Period 3
---|---|---|---|---|---
Placebo Treatment
Active Treatment

slide 7/53
SS models

Let $k$ index patient, $j$ period and $i$ test, and $\delta_k = 1$ if patient $k$ has active treatment. A general model is

$$E(y_{ijk}) = \mu_{jk},$$

$$\text{expit}(\mu_{jk}) = \alpha_j + \delta_k \beta_j + \text{base}_k \gamma_j + \omega_{jk},$$

where either:

(i) $\omega_{jk} = u_k, \ u_k \sim N(0, \sigma_u^2)$ — random intercepts

(ii) $\omega_{jk} = (u_{0k} + u_{1k} t_j), \ (u_{0k}, u_{1k}) \sim N_2(0, \Omega)$ — random intercepts and slopes

(iii) $\omega_{jk} = u_{jk}$ and

$$\begin{pmatrix} u_{1k} \\ u_{2k} \\ u_{3k} \end{pmatrix} \sim N \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u_1} & 0 \\ 0 & \sigma^2_{u_2} \end{pmatrix}.$$
Results: correlation matrices

We are able to fit these models because of the repeated observations on patients in periods. Correlation matrices are:

(i) Random intercepts: $-2\ell = 4941$; var: 5.9, corr: 1.

(ii) Random intercepts and slopes: $-2\ell = 4671$ (+2 paras)

$$
\begin{pmatrix}
9.99 \\
0.86 & 6.76 \\
0.46 & 0.83 & 8.29
\end{pmatrix}.
$$

(iii) Unstructured: $-2\ell = 4580$. (+3 paras)

$$
\begin{pmatrix}
9.97 \\
0.72 & 8.40 \\
0.45 & 0.66 & 7.26
\end{pmatrix}
$$
# Results: treatment effects

Baseline adjusted treatment effects:

<table>
<thead>
<tr>
<th>Model</th>
<th>period 1</th>
<th>period 2</th>
<th>period 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>RI</td>
<td>1.0 (0.22)</td>
<td>0.5 (0.23)</td>
<td>1.2 (0.40)</td>
</tr>
<tr>
<td>RI+S</td>
<td>1.6 (0.65)</td>
<td>0.4 (0.41)</td>
<td>1.3 (0.53)</td>
</tr>
<tr>
<td>Unstruc</td>
<td>1.4 (0.68)</td>
<td>0.7 (0.58)</td>
<td>1.2 (0.54)</td>
</tr>
</tbody>
</table>

Note how these estimates depend on the random effects model.

This is generally true.

It’s worth taking time over the random effects model.
Getting PA coefficients from SS ones

Sometimes it is useful to obtain PA estimates from SS ones (NB can’t go the other way!).

This is particularly easy when we have a designed study, so that the coefficients for the random terms in the model are the same for each patient at each time.

For simplicity we need to estimate the mean at each time, rather than fitting a slope across time. As before the model is:

\[ E(y_{ijk}) = \mu_{jk}, \]
\[ h(\mu_{jk}) = \alpha_j + \delta_k\beta_j + \text{base}_k\gamma_j + \gamma' \psi_k \]
\[ \psi_k \sim N(0, \Sigma_{\psi}). \]

The table below, derived from Zeger et al. (1988), shows how to obtain the PA coefficients.
## Table for obtaining PA from SS effects

<table>
<thead>
<tr>
<th>Link function, $h(.)$</th>
<th>Random intercepts, $z_j = 1$</th>
<th>General random structure $z_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>log</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha^p_j = \alpha_j + \sigma^2_u / 2$</td>
<td>$\alpha^p_j = \beta_j + z_j' \Sigma_u z_j / 2$</td>
<td></td>
</tr>
<tr>
<td>$\beta^p_j = \beta_j$</td>
<td>$\beta^p_j = \beta_j$</td>
<td></td>
</tr>
<tr>
<td>$\gamma^p_j = \gamma_j$</td>
<td>$\gamma^p_j = \gamma_j$</td>
<td></td>
</tr>
<tr>
<td><strong>probit</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha^p_j = \alpha_j / \sqrt{1 + \sigma^2_u}$</td>
<td>$\alpha^p_j = \alpha_j / \sqrt{</td>
<td>I + \Sigma z_j z_j'</td>
</tr>
<tr>
<td>$\beta^p_j = \beta_j / \sqrt{1 + \sigma^2_u}$</td>
<td>$\beta^p_j = \beta_j / \sqrt{</td>
<td>I + \Sigma z_j z_j'</td>
</tr>
<tr>
<td>$\gamma^p_j = \gamma_j / \sqrt{1 + \sigma^2_u}$</td>
<td>$\gamma^p_j = \gamma_j / \sqrt{</td>
<td>I + \Sigma z_j z_j'</td>
</tr>
<tr>
<td><strong>logistic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha^p_j \approx \alpha_j / \sqrt{1 + 0.34584 \sigma^2_u}$</td>
<td>$\alpha^p_j \approx \alpha_j / \sqrt{</td>
<td>I + 0.34584 \Sigma z_j z_j'</td>
</tr>
<tr>
<td>$\beta^p_j \approx \beta_j / \sqrt{1 + 0.34584 \sigma^2_u}$</td>
<td>$\beta^p_j \approx \beta_j / \sqrt{</td>
<td>I + 0.34584 \Sigma z_j z_j'</td>
</tr>
<tr>
<td>$\gamma^p_j \approx \gamma_j / \sqrt{1 + 0.34584 \sigma^2_u}$</td>
<td>$\gamma^p_j \approx \gamma_j / \sqrt{</td>
<td>I + 0.34584 \Sigma z_j z_j'</td>
</tr>
</tbody>
</table>
Accuracy of logistic approximation

\[ \eta \left( 1 + 0.34584 \times \sigma^2 \right)^{0.5} \]

\[ \sigma^2 = 1 \]
\[ \sigma^2 = 8 \]
\[ \sigma^2 = 16 \]
Example

We use data from period 3 of the trial described above, and fit a random intercepts model with (i) probit link and (ii) logistic link.

We compare using the transformation with fitting an independence GEE (=GLM here) with robust SEs.

<table>
<thead>
<tr>
<th>Link</th>
<th>GLM estimates</th>
<th>Transformed SS estimates</th>
<th>SS estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>probit</td>
<td>0.323 (0.190)</td>
<td>0.318 (0.167)</td>
<td>0.544 (0.285)</td>
</tr>
<tr>
<td>logit</td>
<td>0.636 (0.407)</td>
<td>0.577 (0.301)</td>
<td>1.100 (0.573)</td>
</tr>
</tbody>
</table>

For the logistic model $\sigma_u^2 = 7.6$.

Agreement improves as sample size increases.
Using simulation to obtain PA estimates

Sometimes it is easier to use simulation. Consider the random intercepts model:

1. Fit the model. Draw $M$ values of the SS random effect from $N(0, \hat{\sigma}_u^2)$: $u^1, \ldots, u^M$.

2. For $m = 1, \ldots, M$ and for a particular $x$, compute

$$\pi^m = \expit(\beta_0 + \beta_1 x + z' u^m)$$

3. Mean (population averaged) probability is

$$\hat{\pi}^p = \frac{1}{M} \sum_{m=1}^{M} \pi^m.$$
Estimation: PA models

If \( \text{Var}(y_j) \) were known, we could use the score equation for \( \beta \) when the data follow a log-linear model:

\[
s(\beta) = \sum_{j=1}^{J} \left( \frac{\partial \mu_j}{\partial \beta} \right) \text{Var}(y_j)^{-1} (y_j - \mu_j) = 0.
\]

Liang and Zeger (1986) showed that if we write \( \text{Var}(y_j) \) as \( \text{Var}(y_j, \beta, \alpha) \), then if we use any \( \sqrt{J} \)-consistent estimator for \( \alpha \), the estimates of \( \beta \) obtained by solving this are asymptotically as efficient as those we would get were \( \alpha \) known.

In practice, usually an estimating equation for \( \alpha \) is formed and the resulting Generalised Estimating Equations (GEEs) solved simultaneously (e.g. Prentice (1988)).
GEEs: notes on estimation

- $\hat{\beta}$ is nearly efficient relative to ML estimates provided $\text{Var}[E y_j]$ is reasonably approximated.
- $\hat{\beta}$ is consistent as $J \to \infty$ even if the covariance structure of $y_j$ is incorrect.
- Once the mean model is chosen, the robustness of inferences can be checked by trying various covariance structures, and comparing the parameter estimates and their robust standard errors.
- As GEEs are moment-based estimators, they are invalid with missing data, unless it is missing completely randomly.

Further details in Diggle et al. (2002).
Estimation: SS models

For illustration, consider the random intercepts model:

\[
\text{logit } \Pr(y_{ij} = 1) = \beta_0 + u_j, \quad u_j \sim N(0, \sigma_u^2).
\]

The likelihood is

\[
\prod_{j=1}^J \int_{-\infty}^{\infty} \left\{ \prod_{i=1}^I \left( \frac{1}{1 + e^{-(\beta_0 + u_j)}} \right)^{Y_{ij}} \left( \frac{1}{1 + e^{(\beta_0 + u_j)}} \right)^{(1-Y_{ij})} \right\} \times \frac{1}{\sqrt{2\pi\sigma_u^2}} e^{-\frac{u_j^2}{2\sigma_u^2}} du_j
\]

\[
= \prod_{j=1}^J \left\{ \int f(Y_j; \beta, u_j) g(u_j, \Sigma) du_j \right\}.
\]
Obtaining parameter estimates

The likelihood for individual $j$ is

$$L(\beta, \sigma_u^2|Y_j) = \int f(Y_j|u_j, \beta)g(u_j, \sigma_u^2) \, du_j.$$ 

When $f, g$ normal this integral is the multivariate normal distribution of the data, maximised as described in session 1.

Otherwise, the integral is intractable. Options are

- Numerical integration (e.g. SAS NLMIXED) (slow for many random effects)
- Quasi likelihood methods
Penalised Quasi-likelihood

Consider observation \( i,j \), and drop subscripts:

\[
y = \mu(\eta) + \epsilon = \expit(X\beta + Zu) + \epsilon.
\]

After update \( t \), have \( \beta_t, u_t \), say. Expand about true \( \beta, u \):

\[
y_{ij} \approx \mu(X\beta_t + Zu_t) + \frac{\partial \mu}{\partial \eta} X(\beta - \beta_t) + \frac{\partial \mu}{\partial \eta} Z(u - u_t) + \epsilon.
\]

Re-arrange:

\[
y - \mu(X\beta_t + Zu_t) + \frac{\partial \mu}{\partial \eta} X\beta_t + \frac{\partial \mu}{\partial \eta} Zu_t = \frac{\partial \mu}{\partial \eta} X\beta + \frac{\partial \mu}{\partial \eta} Zu + \epsilon.
\]

I.e. \( y^* = X^*\beta + Z^*u + \epsilon \).

Constrain \( \text{Var} \epsilon = \mu(1 - \mu) \), and obtain new estimates with 1 step of fitting routine for normal data.
Comments on quasi-likelihood methods

- Estimation at almost the same speed as for normal models.
- No estimate of the log-likelihood.
- Estimates can be badly biased if
  - fitted values close to 0 or 1
  - some individuals have few observations
- Obvious solution is to try the 2nd order Taylor expansion — called PQL(2)
  - bias is substantially reduced, but can’t always be fitted
- There have been various proposals for simulation-based bias correction.

Ng et al. (2005) consider these, and conclude Simulated Maximum Likelihood (SML) is best.
SML method for bias correction

Recall likelihood: \( L(\beta, \Sigma | y) = \int f(y | u, \beta) g(u, \Sigma) \, du \).

Obtain initial estimates \( \hat{\beta}, \hat{\Sigma} \), from PQL(2). Set \( \tilde{\Sigma} = \hat{\Sigma} \). Then

\[
L(\beta, \Sigma) = \int \frac{f(y | u, \beta) g(u, \Sigma) g(u, \tilde{\Sigma})}{g(u, \tilde{\Sigma})} \, du \\
\approx \frac{1}{H} \sum_{h=1}^{H} \frac{f(y | u_h, \beta) g(u_h, \Sigma)}{g(u_h, \tilde{\Sigma})}
\]

where \( u_1, \ldots, u_h \overset{iid}{\sim} g(u, \tilde{\Sigma}) \).

Search for \( \beta, \Sigma \) that maximise this Monte-Carlo likelihood estimate (keeping \( \tilde{\Sigma} \) fixed).
Simulation study

Compare PQL (2) followed by SML with SAS proc NLMIXED.

Use example of Kuk (1995). Simulate 100 data sets from

\[
Y_{ij} \sim \text{bin}(6, \pi_{ij})
\]

\[
\logit(\pi_{ij}) = 0.2 + u_j + 0.1x_{ij}
\]

\[
u_j \sim N(0, 1)
\]

\[J = 15\] level 2 units with \[I = 2\] level 1 units each.

\[x = -15, \ldots, 14\].
### Results

Values are average estimates (Mean Squared Error)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True values</td>
<td>0.2</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>SML (H=500)</td>
<td>0.203</td>
<td>0.099</td>
<td>0.939</td>
</tr>
<tr>
<td></td>
<td>(0.136)</td>
<td>(0.00134)</td>
<td>(0.561)</td>
</tr>
<tr>
<td>SAS NLMIXED</td>
<td>0.190</td>
<td>0.097</td>
<td>0.927</td>
</tr>
<tr>
<td></td>
<td>(0.135)</td>
<td>(0.00134)</td>
<td>(0.478)</td>
</tr>
</tbody>
</table>

NLMIXED needs starting values: gave 0 for $\beta_0$, $\beta_1$ and 1 for $\sigma^2$. NLMIXED had convergence problems on 17/100 data sets: these are excluded.

**Conclude**

(i) SAS NLMIXED larger bias, smaller variance; BUT we only used H=500

(ii) PQL(2) + SML seems to work well.
Example: Bangladesh data

A sub-sample of the 1988 Bangladesh fertility survey **Huq and Cleland (1990)**.

All the women had been married at some time.

\[ i = 1, \ldots, 1934 \text{ women (level 1)} \text{ from } j = 1, \ldots, 60 \text{ districts (level 2).} \]

2–118 women in each district.

Response: \[ y_{ij} = \begin{cases} 1 & \text{if the women reported contraceptive use} \\ 0 & \text{otherwise} \end{cases} \]

Covariates: urban resident; age; no. of children
Model

Let $1[\cdot]$ is an indicator for the event in brackets.

The model is:

$$\text{logit}\{\Pr(Y_{ij} = 1)\} = (\beta_0 + u_{0j}) + (\beta_1 + u_{1j}) \times 1[\text{urban resident}]$$
$$+ \beta_2 \times (\text{centred age}) + \beta_3 \times 1[1 \text{ child}]$$
$$+ \beta_4 \times 1[2 \text{ children}] + \beta_5 \times 1[\geq 3 \text{ children}],$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0} & \sigma_{u0u1} \\ \sigma_{u0u1} & \sigma^2_{u1} \end{pmatrix} \right\}.$$
Results

Broadly, odds of contraception use higher for urban residents and increases with no. of children.

Used 6 different estimation routines. Most variation in variance estimates:

<table>
<thead>
<tr>
<th>Method</th>
<th>2nd order</th>
<th>2nd order PQL + SML (H=3000)</th>
<th>EM-Laplace2</th>
<th>MCMC</th>
<th>NLMIXED</th>
<th>GLLAMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Package</td>
<td>MLwiN</td>
<td>(+ matlab)</td>
<td>HLM</td>
<td>MLwiN</td>
<td>SAS</td>
<td>Stata</td>
</tr>
<tr>
<td>$\sigma^2_{u0}$</td>
<td>0.396</td>
<td>0.398</td>
<td>0.379</td>
<td>0.435</td>
<td>0.388</td>
<td>0.390</td>
</tr>
<tr>
<td>(0.118)</td>
<td>(0.132)</td>
<td>(NA)</td>
<td>(0.144)</td>
<td>(0.129)</td>
<td>(0.129)</td>
<td></td>
</tr>
<tr>
<td>$\sigma_{u0u1}$</td>
<td>-0.414</td>
<td>-0.407</td>
<td>-0.391</td>
<td>-0.455</td>
<td>-0.404</td>
<td>-0.406</td>
</tr>
<tr>
<td>(0.160)</td>
<td>(0.177)</td>
<td>(NA)</td>
<td>(0.191)</td>
<td>(0.175)</td>
<td>(0.176)</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2_{u1}$</td>
<td>0.686</td>
<td>0.661</td>
<td>0.627</td>
<td>0.770</td>
<td>0.664</td>
<td>0.666</td>
</tr>
<tr>
<td>(0.284)</td>
<td>(0.339)</td>
<td>(NA)</td>
<td>(0.340)</td>
<td>(0.322)</td>
<td>(0.322)</td>
<td></td>
</tr>
<tr>
<td>$-2\ell$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RI+S</td>
<td>NA</td>
<td>2398.9</td>
<td>5953.1</td>
<td>(DIC) 2386.4</td>
<td>2398.7</td>
<td>2398.6</td>
</tr>
<tr>
<td>RI only</td>
<td>NA</td>
<td>2413.0</td>
<td>5968.1</td>
<td>(DIC) 2408.7</td>
<td>2413.7</td>
<td>2413.7</td>
</tr>
<tr>
<td>Change</td>
<td>NA</td>
<td>14.1</td>
<td>15.0</td>
<td>(21.3)</td>
<td>15.0</td>
<td>15.0</td>
</tr>
</tbody>
</table>
Implications: software

- HLM v6: EM-Laplace2: No SE estimated for random components. In another problem, estimated random components are smaller than estimates from SML/NLMIXED.
- HLM v5: Laplace6: convergence problems.
- GLLAMM is slow
- MCMC (Gamma diffuse priors on variances) gives inflated estimates, relative to ML.
- NLMIXED needs starting values (here from PQL(2))
- SML appears to work well

Conclude: PQL(2) plus SML or NLMIXED for ‘bias correction’ usually gives best answers.
Cross classified data

So far we have considered only hierarchical structures.

However, social structures are *not always hierarchical*. People often belong to more than one grouping at a given hierarchical level.

E.g. neighbourhood and school may both have effects on educational outcomes:

- a school may contain children from several neighbourhoods;
- children from one neighbourhood may attend different schools

Children are nested in a *cross classification of neighbourhood and school*. 
Classification diagrams

School

Children

Neighbourhood

Pupil

Nested structure

Neighbourhood

School

Pupil

Cross classified structure

slide 30/53
Standard notation

$Y_{i(j_1, j_2)}$ is the response for pupil $i$ in neighbourhood $j_1$ and school $j_2$.

The subscripts $j_1$ and $j_2$ are bracketed together to indicate that these classifications are at the same level, i.e. pupils are nested within a cross-classification of neighbourhoods and schools.

A basic cross-classified model may be written:

$$y_{i(j_1, j_2)} = \beta' x_{i(j_1, j_2)} + u_{1j_1} + u_{2j_2} + e_{i(j_1, j_2)}.$$

$u_{1j_1}$ is the random neighbourhood effect

$u_{2j_2}$ is the random school effect
Alternative notation

For hierarchical models we have *one subscript per level*, and nesting is implied by reading from left to right.

E.g. $ijk$ denotes the $i$th level 1 unit within the $j$th level 2 unit within the $k$th level 3 unit.

For cross-classified models, we can group together indices for classifications at the same level using parentheses (see previous slide). However, having one subscript per classification becomes cumbersome.

Using an alternative notation, we have a *single subscript*, no matter how many classifications there are.
Data matrix for cross-classification

Single subscript notation.

Let $i$ index children.

<table>
<thead>
<tr>
<th>$i$</th>
<th>Neighbourhood$(i)$</th>
<th>School$(i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>
Cross classified model

Using the single-subscript notation, $y_i$ is the outcome for child $i$.

Classification 1 is child, 2 is neighbourhood and 3 is school.

$$y_i = \beta' x_i + u_{nbhd(i)}^{(2)} + u_{sch(i)}^{(3)} + e_i$$

$u_{nbhd(i)}^{(2)} \sim N(0, \Omega_u^{(2)})$  Random departure due to neighbourhood

$u_{sch(i)}^{(3)} \sim N(0, \Omega_u^{(3)})$  random departure due to school

$e_i \sim N(0, \Omega_e)$  individual-level residual

Covariates may be defined for any of the 3 classifications

Coefficients may be allowed to vary across neighbourhood or school.
Cross-classified model: notation

From the previous slide, the model is:

\[ y_i = \beta' x_i + u_{\text{nbhd}(i)}^{(2)} + u_{\text{sch}(i)}^{(3)} + e_i. \]

Thus, for pupils 1 and 10 in the data set 2 slides back,

\[ y_1 = \beta' x_1 + u_1^{(2)} + u_1^{(3)} + e_1; \]

\[ y_{10} = \beta' x_{10} + u_2^{(2)} + u_4^{(3)} + e_{10}. \]
Other cross-classified structures

- Pupils within primary schools by secondary schools
- Patients within GPs by hospitals
- Survey respondents within sampling clusters by interviewers
- Repeated measures within raters by individuals (e.g. patients by nurses)
- Students following modular degree courses, e.g. Simonite and Browne (2003)
Example: Scottish school children

3435 children who attended 148 primary schools and 19 secondary schools in Fife, Scotland.

Classifications: 1 – student; 2 – primary school; 3 – secondary school

\( y_i \) – overall achievement at age 16 for student \( i \)

\( x_i \) – verbal reasoning at age 12 (mean centred)

2-level cross classified model:

\[
y_i = \beta_0 + \beta_1 x_i + u_{\text{prim}(i)}^{(2)} + u_{\text{sec}(i)}^{(3)} + e_i
\]
### Results

<table>
<thead>
<tr>
<th></th>
<th>Hierarchical model</th>
<th>Cross-classified model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>5.99</td>
<td>5.98</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.16 (0.003)</td>
<td>0.16 (0.003)</td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_u^{(2)}$ (primary)</td>
<td>—</td>
<td>0.27 (0.06)</td>
</tr>
<tr>
<td>$\sigma_u^{(3)}$ (secondary)</td>
<td>0.28 (0.06)</td>
<td>0.01 (0.02)</td>
</tr>
<tr>
<td>$\sigma_e$ (student residual)</td>
<td>4.26 (0.10)</td>
<td>4.25 (0.10)</td>
</tr>
</tbody>
</table>

Most of the variation in results at 16 years can be attributed to primary schools — an intriguing result for educational researchers overlooked without a cross-classified model.
Cross-classification: covariance matrix

Recall in our hierarchical model, the covariance matrix was block diagonal:

\[
\Sigma_{\text{full}} = \begin{pmatrix}
\Sigma & 0 & 0 & \ldots & 0 \\
0 & \Sigma & 0 & \ldots & 0 \\
0 & 0 & \ddots & \ldots & 0 \\
0 & 0 & \ldots & \Sigma & 0 \\
0 & 0 & \ldots & 0 & \Sigma \\
\end{pmatrix}.
\]

With a cross classified model this is no longer true.

E.g. for our example, suppose we first classify pupils within secondary schools.

For every student who shares a primary school with a student from another secondary school, there is an off diagonal term — \( \Omega^{(2)} \) (primary).
Estimation

This makes ML estimation *much* harder.

Provided the cross-classification is not too widespread, careful ordering of the data can cause the covariance matrix to consist of —much larger— diagonal blocks.

Consequently, estimation is much more memory intensive, and much slower.

MCMC estimation in MLwiN is much more efficient.

See the manual *MCMC Estimation in MLwiN* by W Browne, downloadable from [www.mlwin.com](http://www.mlwin.com)
The multiple membership model

Suppose now a student changes school during study, and so belongs to more than one school.

A model with membership of two schools (1,2) is:

\[ y_{i(1,2)} = (X\beta)_{i(1,2)} + w_{i1}u_1 + w_{i2}u_2 + e_{i(1,2)} \]

\[ w_{i1} + w_{i2} = 1 \]

More generally:

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

\[ \sum_{h\in\{j\}} w_{ih} = 1, \quad u_h \overset{iid}{\sim} N(0, \sigma_u^2) \]

\[ \operatorname{Var}(\sum_{h\in\{j\}} w_{ih}u_h) = \sigma_u^2 \sum_{h\in\{j\}} w_{ih}^2 \]

Usually, if \( h \in \{1, 2\} \), \( w_{i1} + w_{i2} = 0.5 \)

\[ \operatorname{Var}(\sum_h w_{ih}u_h) = \sigma_u^2/2. \]
Example: Danish poultry farming

Interested in understanding variation in salmonella outbreaks in chicken flocks.

Data from salmonella outbreaks in flocks of chickens in Danish poultry farms between 1995 & 1997.

Response is whether or not a ‘child flock’ was infected.

There are two hierarchies in the data: a production hierarchy and a breeding hierarchy.

Full details:
Browne et al. (2001)
Production hierarchy

Level 1 units are child flocks of chickens.

Child flocks live for only a short time (~35 days) before they are slaughtered for consumption.

Child flocks are kept in houses; in a year a house may have a throughput of 10–12 flocks.

Houses are grouped in farms.

Data from 10,127 child flocks, 725 houses, 304 farms.
Breeding hierarchy

There are 200 parent flocks.

Eggs are taken from parent flocks to 4 hatcheries.

After hatching, chicks are transported to the farms in the production hierarchy (previous slide).

Child flocks draw chicks from up to six parent flocks.
Each child flock connected to multiple parent flocks, so child flocks are multiple members of parent flocks. Parental membership information is known.

Parent flocks are also cross-classified with the house/farm production hierarchy.
Questions

To what extent is variability in child flock infection attributable to

- production processes (hygiene on houses and farms)?
- hatcheries processes?
- parent flock processes
  - genetic predisposition to salmonella?
  - poor parent flock hygiene introducing infected eggs into the system?
Model: notation

Let $\pi_I = \Pr(\text{flock } i \text{ has salmonella})$.


Let $1[\text{hatch1}], \ldots, 1[\text{hatch4}]$ indicate the four hatcheries in which all the eggs form the parent flocks are hatched.

Let $\{p.flock(i)\}$ be the set of parent flocks for child flock $i$.

We know the exact makeup of each child flock (in terms of parent flocks). Define weights for each child flock $i$, $w_{ij}$, to represent this makeup.

For each child flock, $i$, these satisfy

$$\sum_{j \in \{p.flock(i)\}} w_{ij} = 1$$
Model: components of variance

\[
\text{logit} \pi_i = \beta_0 + \beta_1 \times 1[96] + \beta_2 \times 1[97] \\
+ \beta_3 \times 1[\text{hatch2}] + \beta_4 \times 1[\text{hatch3}] + \beta_5 \times 1[\text{hatch4}] \\
+ u^{(2)}_{\text{house}(i)} + u^{(3)}_{\text{farm}(i)} + \sum_{j \in \{\text{p.flock}(i)\}} w_{ij} u^{(4)}_j
\]

\[u^{(2)}_{\text{house}(i)} \sim N(0, \sigma^2_{(2)})\]
\[u^{(3)}_{\text{farm}(i)} \sim N(0, \sigma^2_{(3)})\]
\[u_j^{(4)} \sim N(0, \sigma^2_{(2)})\]
\[u^{(2)} \perp u^{(3)} \perp u^{(4)}\]
## Results

**Estimation:** PQL unstable; used MCMC

<table>
<thead>
<tr>
<th>Description</th>
<th>Estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-1.86</td>
<td>0.187</td>
</tr>
<tr>
<td>1996</td>
<td>-1.04</td>
<td>0.131</td>
</tr>
<tr>
<td>1997</td>
<td>-0.89</td>
<td>0.151</td>
</tr>
<tr>
<td>hatchery 2</td>
<td>-1.47</td>
<td>0.22</td>
</tr>
<tr>
<td>hatchery 3</td>
<td>-0.17</td>
<td>0.21</td>
</tr>
<tr>
<td>hatchery 4</td>
<td>-0.92</td>
<td>0.29</td>
</tr>
<tr>
<td>parent flock variance</td>
<td>1.02</td>
<td>0.22</td>
</tr>
<tr>
<td>farm variance</td>
<td>0.59</td>
<td>0.11</td>
</tr>
<tr>
<td>house variance</td>
<td>0.19</td>
<td>0.09</td>
</tr>
</tbody>
</table>

**Conclude**

Some hatcheries better than others; variability dominated by parent flock.
Summary

- Extended ideas of session 1 to multilevel discrete data.
- Contrasted subject-specific and population-averaged models.
- Described random-effects models for discrete data.
- Discussed estimation & software issues for SS models.
- Extended ideas to cross-classified and multiple membership models.
- Such models can be fitted in MLwiN, with care.
- More examples and documentation at www.mlwin.com
References


References


Ng, E. S. W., Carpenter, J. R., Goldstein, H. and Rasbash, J. (2005) Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood.
