**Web-based training**

Introduction (Alex Dmitrienko, Lilly)

Web-based training program
- http://www.amstat.org/sections/sbiop/webinarseries.html

Four-part web-based training series
- Geert Verbeke (Katholieke Universiteit Leuven, Belgium)
- Geert Molenberghs (Universiteit Hasselt, Belgium)

Presentation slides

Discussion thread

---

**Longitudinal and Incomplete Data**

**Part I: Linear Mixed Models**

**Geert Verbeke**
- Biostatistical Centre
- K.U.Leuven, Belgium
- geert.verbeke@med.kuleuven.be
- www.kuleuven.ac.be/biostat/

**Geert Molenberghs**
- Center for Statistics
- Universiteit Hasselt, Belgium
- geert.molenberghs@uhasselt.be
- www.censtat.uhasselt.be

American Statistical Association, Webinar, Spring 2007
Chapter 0
Related References


Part I

Continuous Longitudinal Data
Chapter 1
Introduction

Repeated Measures / Longitudinal data
Example

1.1 Repeated Measures / Longitudinal Data

Repeated measures are obtained when a response is measured repeatedly on a set of units

- Units:
  - Subjects, patients, participants, ...
  - Animals, plants, ...
  - Clusters: families, towns, branches of a company, ...
  - ...

- Special case: Longitudinal data
1.2 Rat Data

- Research question (Dentistry, K.U.Leuven):
  
  How does craniofacial growth depend on testosterone production?

- Randomized experiment in which 50 male Wistar rats are randomized to:
  - Control (15 rats)
  - Low dose of Decapeptyl (18 rats)
  - High dose of Decapeptyl (17 rats)

- Treatment starts at the age of 45 days; measurements taken every 10 days, from day 50 on.

- The responses are distances (pixels) between well defined points on x-ray pictures of the skull of each rat:
• Measurements with respect to the roof, base and height of the skull.

• Here, we consider only one response, reflecting the height of the skull.

• Individual profiles:

![Individual profiles diagram]

• Complication: Dropout due to anaesthesia (56%):

<table>
<thead>
<tr>
<th>Age (days)</th>
<th>Control</th>
<th>Low</th>
<th>High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>15</td>
<td>18</td>
<td>17</td>
<td>50</td>
</tr>
<tr>
<td>60</td>
<td>13</td>
<td>17</td>
<td>16</td>
<td>46</td>
</tr>
<tr>
<td>70</td>
<td>13</td>
<td>15</td>
<td>15</td>
<td>43</td>
</tr>
<tr>
<td>80</td>
<td>10</td>
<td>15</td>
<td>13</td>
<td>38</td>
</tr>
<tr>
<td>90</td>
<td>7</td>
<td>12</td>
<td>10</td>
<td>29</td>
</tr>
<tr>
<td>100</td>
<td>4</td>
<td>10</td>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>110</td>
<td>4</td>
<td>8</td>
<td>10</td>
<td>22</td>
</tr>
</tbody>
</table>

• Remarks:
  ▶ Much variability between rats, much less variability within rats
  ▶ Fixed number of measurements scheduled per subject, but not all measurements available due to dropout, for known reason.
  ▶ Measurements taken at fixed timepoints
Chapter 2  
A Model for Longitudinal Data  

▷ Introduction  
▷ Example: Rat data  
▷ The general linear mixed-effects model  

2.1 Introduction  

• In practice: often unbalanced data:  
  ▷ unequal number of measurements per subject  
  ▷ measurements not taken at fixed time points  

• Therefore, multivariate regression techniques are often not applicable  

• Often, subject-specific longitudinal profiles can be well approximated by linear regression functions  

• This leads to a 2-stage model formulation:  
  ▷ Stage 1: Linear regression model for each subject separately  
  ▷ Stage 2: Explain variability in the subject-specific regression coefficients using known covariates
2.2 Example: The Rat Data

- Individual profiles:

  ![Graphs showing individual profiles for different conditions.]

- Transformation of the time scale to linearize the profiles:

  $$\text{Age}_{ij} \rightarrow t_{ij} = \ln[1 + (\text{Age}_{ij} - 45)/10]$$

- Note that $t = 0$ corresponds to the start of the treatment (moment of randomization)

- **Stage 1 model:**

  $$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \varepsilon_{ij}, \quad j = 1, \ldots, n_i$$

- In the second stage, the subject-specific intercepts and time effects are related to the treatment of the rats

- **Stage 2 model:**

  $$\begin{cases} 
  \beta_{1i} = \beta_0 + b_{1i}, \\
  \beta_{2i} = \beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i},
  \end{cases}$$
• $L_i$, $H_i$, and $C_i$ are indicator variables:

\[
L_i = \begin{cases} 
1 & \text{if low dose} \\
0 & \text{otherwise}
\end{cases} \quad H_i = \begin{cases} 
1 & \text{if high dose} \\
0 & \text{otherwise}
\end{cases} \quad C_i = \begin{cases} 
1 & \text{if control} \\
0 & \text{otherwise}
\end{cases}
\]

• Parameter interpretation:
  ▶ $\beta_0$: average response at the start of the treatment (independent of treatment)
  ▶ $\beta_1$, $\beta_2$, and $\beta_3$: average time effect for each treatment group

### 2.3 The Linear Mixed-effects Model

#### 2.3.1 General idea

• A 2-stage approach can be performed explicitly in the analysis

• However, this is just another example of the use of summary statistics:
  ▶ $Y_i$ is summarized by $\hat{\beta}_i$
  ▶ summary statistics $\hat{\beta}_i$ analysed in second stage

• The associated drawbacks can be avoided by combining the two stages into one model
2.3.2 Example: The Rat Data

- **Stage 1 model:** \[ Y_{ij} = \beta_{1i} + \beta_{2i}t_{ij} + \epsilon_{ij}, \quad j = 1, \ldots, n_i \]

- **Stage 2 model:**
  \[
  \begin{align*}
  \beta_{1i} &= \beta_0 + b_{1i}, \\
  \beta_{2i} &= \beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i},
  \end{align*}
  \]

- **Combined:**
  \[ Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i})t_{ij} + \epsilon_{ij} \]

  \[
  \begin{cases}
  \beta_0 + b_{1i} + (\beta_1 + b_{2i})t_{ij} + \epsilon_{ij}, & \text{if low dose} \\
  \beta_0 + b_{1i} + (\beta_2 + b_{2i})t_{ij} + \epsilon_{ij}, & \text{if high dose} \\
  \beta_0 + b_{1i} + (\beta_3 + b_{2i})t_{ij} + \epsilon_{ij}, & \text{if control.}
  \end{cases}
  \]

Chapter 3
The General Linear Mixed Model

▷ The general model formulation
▷ Hierarchical versus marginal model formulation
▷ Examples
3.1 The General Linear Mixed Model

\[ Y_i = X_i \beta + Z_i b_i + \varepsilon_i \]

\[ b_i \sim N(0, D), \]

\[ \varepsilon_i \sim N(0, \Sigma_i), \]

\[ b_1, \ldots, b_N, \varepsilon_1, \ldots, \varepsilon_N \text{ independent} \]

Terminology:

- Fixed effects: \( \beta \)
- Random effects: \( b_i \)
- Variance components: elements in \( D \) and \( \Sigma_i \)

3.2 Hierarchical versus Marginal Model

- The general linear mixed model is given by:

\[ Y_i = X_i \beta + Z_i b_i + \varepsilon_i \]

\[
\begin{align*}
  b_i & \sim N(0, D), \\
  \varepsilon_i & \sim N(0, \Sigma_i), \\
  b_1, \ldots, b_N, \varepsilon_1, \ldots, \varepsilon_N & \text{ independent},
\end{align*}
\]

- It can be rewritten as:

\[ Y_i | b_i \sim N(X_i \beta + Z_i b_i, \Sigma_i) \]

\[ b_i \sim N(0, D) \]
• It is therefore also called a hierarchical model:
  ▶ A model for $Y_i$ given $b_i$
  ▶ A model for $b_i$

• Marginally, we have that $Y_i$ is distributed as: $Y_i \sim N(X_i \beta, Z_i D Z_i' + \Sigma_i)$

• Hence, very specific assumptions are made about the dependence of mean and covariance on the covariates $X_i$ and $Z_i$:
  ▶ **Implied mean:** $X_i \beta$
  ▶ **Implied covariance:** $V_i = Z_i D Z_i' + \Sigma_i$

• Note that the hierarchical model implies the marginal one, **not** vice versa

### 3.3 Example 1: The Rat Data

• The LMM was given by $Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i}) t_{ij} + \epsilon_{ij}$

• Implied marginal mean structure:
  ▶ Linear average evolution in each group
  ▶ Equal average intercepts
  ▶ Different average slopes

• Implied marginal covariance structure ($\Sigma_i = \sigma^2 I_{n_i}$):
  \[
  \text{Cov}(Y_i(t_1), Y_i(t_2)) = \begin{pmatrix} 1 & t_1 \\ t_1 & 1 \end{pmatrix} D \begin{pmatrix} 1 \\ t_2 \end{pmatrix} + \sigma^2 \delta_{\{t_1, t_2\}}
  \]
  \[
  = d_{22} t_1 t_2 + d_{12}(t_1 + t_2) + d_{11} + \sigma^2 \delta_{\{t_1, t_2\}}.
  \]
• Note that the model implicitly assumes that the variance function is quadratic over time, with curvature $d_{22}$.

• A negative estimate for $d_{22}$ indicates negative curvature in the variance function but cannot be interpreted under the hierarchical model.

• A model which assumes that all variability in subject-specific slopes can be ascribed to treatment differences can be obtained by omitting the random slopes $b_{2i}$ from the above model:

$$Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 L_i + \beta_2 H_i + \beta_3 C_i)t_{ij} + \varepsilon_{ij}$$

• This is the so-called random-intercepts model.

• The same marginal mean structure is obtained as under the model with random slopes.

• Implied marginal covariance structure ($\Sigma_i = \sigma^2 I_{n_i}$):

\[
\text{Cov}(Y_i(t_1), Y_i(t_2)) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} D \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \sigma^2 \delta_{(t_1, t_2)} = d_{11} + \sigma^2 \delta_{(t_1, t_2)}.
\]

• Hence, the implied covariance matrix is compound symmetry:

  ▶ constant variance $d_{11} + \sigma^2$
  ▶ constant correlation $\rho_i = d_{11}/(d_{11} + \sigma^2)$ between any two repeated measurements within the same rat.
3.4 Example 2: Bivariate Observations

- Balanced data, two measurements per subject ($n_i = 2$), two models:

<table>
<thead>
<tr>
<th>Model 1:</th>
<th>Model 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random intercepts</td>
<td>Uncorrelated intercepts and slopes</td>
</tr>
<tr>
<td>+ heterogeneous errors</td>
<td>+ measurement error</td>
</tr>
</tbody>
</table>

\[
V = \begin{pmatrix} 1 \\ 1 \end{pmatrix} d \begin{pmatrix} 1 & 1 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} d + \sigma_1^2 & d \\ d & d + \sigma_2^2 \end{pmatrix}
\]

\[
V = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} = \begin{pmatrix} d_1 + \sigma^2 & d_1 \\ d_1 & d_1 + d_2 + \sigma^2 \end{pmatrix}
\]

- Different hierarchical models can produce the same marginal model

- Hence, a good fit of the marginal model cannot be interpreted as evidence for any of the hierarchical models.

- A satisfactory treatment of the hierarchical model is only possible within a Bayesian context.
Chapter 4
Estimation and Inference in the Marginal Model

▷ ML and REML estimation
▷ Inference
▷ Fitting linear mixed models in SAS

4.1 ML and REML Estimation

• Recall that the general linear mixed model equals

\[ Y_i = X_i\beta + Z_i b_i + \varepsilon_i \]

\[
\begin{align*}
    b_i &\sim N(0, D) \\
    \varepsilon_i &\sim N(0, \Sigma_i) \\
\end{align*}
\]

\[
\text{independent}
\]

• The implied marginal model equals

\[ Y_i \sim N(X_i\beta, Z_iDZ'_i + \Sigma_i) \]

• Note that inferences based on the marginal model do not explicitly assume the presence of random effects representing the natural heterogeneity between subjects
• Notation:
  ▶ \(\beta\): vector of fixed effects (as before)
  ▶ \(\alpha\): vector of all variance components in \(D\) and \(\Sigma_i\)
  ▶ \(\theta = (\beta', \alpha')':\) vector of all parameters in marginal model

• Marginal likelihood function:

\[
L_{\text{ML}}(\theta) = \prod_{i=1}^{N} \left( \frac{(2\pi)^{-n_i/2}}{\sqrt{|V_i(\alpha)|}} \exp \left( -\frac{1}{2} (Y_i - X_i\beta)' V_i^{-1}(\alpha) (Y_i - X_i\beta) \right) \right)
\]

• If \(\alpha\) were known, MLE of \(\beta\) equals

\[
\hat{\beta}(\alpha) = \left( \sum_{i=1}^{N} X_i' W_i X_i \right)^{-1} \sum_{i=1}^{N} X_i' W_i y_i,
\]

where \(W_i\) equals \(V_i^{-1}\).

• In most cases, \(\alpha\) is not known, and needs to be replaced by an estimate \(\hat{\alpha}\)

• Two frequently used estimation methods for \(\alpha\):
  ▶ Maximum likelihood
  ▶ Restricted maximum likelihood
4.2 Inference

- Inference for $\beta$:
  - Wald tests, $t$- and $F$-tests
  - LR tests (not with REML)

- Inference for $\alpha$:
  - Wald tests
  - LR tests (even with REML)
  - Caution: Boundary problems!

- Inference for the random effects:
  - Empirical Bayes inference based on posterior density $f(b_i | y_i = y_i)$
  - ‘Empirical Bayes (EB) estimate’: Posterior mean

4.3 Fitting Linear Mixed Models in SAS

- A model for the rat data: $Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i})t_{ij} + \varepsilon_{ij}$

- SAS program:
  ```sas
  proc mixed data=rat method=reml;
  class id group;
  model y = t group*t / solution;
  random intercept t / type=un subject=id ;
  run;
  ```

- Fitted averages:
Chapter 5
Inference for the Random Effects

▷ Empirical Bayes inference
▷ SAS code

5.1 Empirical Bayes Inference

• Random effects $b_i$ reflect how the evolution for the $i$th subject deviates from the expected evolution $X_i\beta$.

• Estimation of the $b_i$ helpful for detecting outlying profiles

• This is only meaningful under the hierarchical model interpretation:

\[
Y_i|b_i \sim N(X_i\beta + Z_ib_i, \Sigma_i) \quad b_i \sim N(0, D)
\]

• Since the $b_i$ are random, it is most natural to use Bayesian methods

• Terminology: prior distribution $N(0, D)$ for $b_i$
• Posterior density:

\[
f(b_i|y_i) \equiv f(b_i|Y_i = y_i) = \frac{f(y_i|b_i) f(b_i)}{\int f(y_i|b_i) f(b_i) \, db_i}
\]

\[
\propto f(y_i|b_i) f(b_i)
\]

\[
\propto \ldots
\]

\[
\propto \exp \left\{ -\frac{1}{2} \left( b_i - DZ_i'W_i(y_i - X_i\beta) \right)'\Lambda_i^{-1} \left( b_i - DZ_i'W_i(y_i - X_i\beta) \right) \right\}
\]

for some positive definite matrix \( \Lambda_i \).

• Posterior distribution:

\[
b_i \mid y_i \sim N(DZ_i'W_i(y_i - X_i\beta), \Lambda_i)
\]

• Posterior mean as estimate for \( b_i \):

\[
\tilde{b}_i(\theta) = E[b_i \mid Y_i = y_i] = \int b_i f(b_i|y_i) \, db_i = DZ_i'W_i(\alpha)(y_i - X_i\beta)
\]

• Parameters in \( \theta \) are replaced by their ML or REML estimates, obtained from fitting the marginal model.

\( \tilde{b}_i = \tilde{b}_i(\theta) \) is called the **Empirical Bayes** estimate of \( b_i \).

• Approximate \( t-\) and \( F-\)tests to account for the variability introduced by replacing \( \theta \) by \( \tilde{\theta} \), similar to tests for fixed effects.
5.2 SAS code

• A model for the rat data: \[ Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i})t_{ij} + \varepsilon_{ij} \]

• In SAS the estimates can be obtained from adding the option ‘solution’ to the random statement:

```sas
random intercept t
   / type=un subject=id solution;
ods listing exclude solutionr;
ods output solutionr=out;
```

• The ODS statements are used to write the EB estimates into a SAS output data set, and to prevent SAS from printing them in the output window.

• In practice, histograms and scatterplots of certain components of \( \hat{b}_i \) are used to detect model deviations or subjects with ‘exceptional’ evolutions over time.