EPID 766: Analysis of Longitudinal Data from Epidemiologic Studies

Daowen Zhang

NC State University

zhang@stat.ncsu.edu

http://www4.stat.ncsu.edu/~dzhang2
# Contents

1 Review and introduction to longitudinal studies
   1.1 Review of 3 study designs ........................................ 5
   1.2 Introduction to longitudinal studies .............................. 11
   1.3 Data examples ......................................................... 12
   1.4 Features of longitudinal data ....................................... 22
   1.5 Why longitudinal studies? .......................................... 24
   1.6 Challenges in analyzing longitudinal data ....................... 27
   1.7 Methods for analyzing longitudinal data ......................... 32
   1.8 Two-stage method for analyzing longitudinal data ............... 33
   1.9 Analyzing Framingham data using two-stage method ............. 35

2 Linear mixed models for normal longitudinal data ................. 50
   2.1 What is a linear mixed (effects) model? ......................... 51
   2.2 Estimation and inference for linear mixed models ............... 68
   2.3 How to choose random effects and the error structure? ......... 70
### TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4 Analyze Framingham data using linear mixed models</td>
<td>71</td>
</tr>
<tr>
<td>2.5 GEE for linear mixed models</td>
<td>107</td>
</tr>
<tr>
<td>2.6 Missing data issues</td>
<td>111</td>
</tr>
<tr>
<td>3 Modeling and design issues</td>
<td>116</td>
</tr>
<tr>
<td>3.1 How to handle baseline response?</td>
<td>117</td>
</tr>
<tr>
<td>3.2 Do we model previous responses as covariates?</td>
<td>119</td>
</tr>
<tr>
<td>3.3 Modeling outcome vs. modeling the change of outcome</td>
<td>121</td>
</tr>
<tr>
<td>3.4 Design a longitudinal study: Sample size estimation</td>
<td>131</td>
</tr>
<tr>
<td>4 Modeling discrete longitudinal data</td>
<td>138</td>
</tr>
<tr>
<td>4.1 Generalized estimating equations (GEEs) for continuous and discrete</td>
<td>139</td>
</tr>
<tr>
<td>longitudinal data</td>
<td></td>
</tr>
<tr>
<td>4.1.1 Why GEEs?</td>
<td>139</td>
</tr>
<tr>
<td>4.1.2 Key features of GEEs for analyzing longitudinal data</td>
<td>143</td>
</tr>
<tr>
<td>4.1.3 Some popular GEE Models</td>
<td>145</td>
</tr>
<tr>
<td>4.1.4 Some basics of GEEs</td>
<td>147</td>
</tr>
<tr>
<td>4.1.5 Interpretation of regression coefficients in a GEE Model</td>
<td>153</td>
</tr>
<tr>
<td>Chapter</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>4.1.6</td>
<td>Analyze Infectious disease data using GEE</td>
</tr>
<tr>
<td>4.1.7</td>
<td>Analyze epileptic seizure count data using GEE</td>
</tr>
<tr>
<td>4.2</td>
<td>Generalized linear mixed models (GLMMs)</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Model specification and implementation</td>
</tr>
<tr>
<td>4.3</td>
<td>Analyze infectious disease data using a GLMM</td>
</tr>
<tr>
<td>4.4</td>
<td>Analyze epileptic count data using a GLMM</td>
</tr>
</tbody>
</table>

5 Summary: what we covered 205
1 Review and introduction to longitudinal studies

- Review of 3 study designs
- Introduction to longitudinal (panel) studies
- Data examples
- Features of longitudinal data
- Why longitudinal studies
- Challenges in analyzing longitudinal data
- Methods for analyzing longitudinal data: two-stage, linear mixed model, GEE, transition models
- Two-stage method for analyzing longitudinal data
- Analyzing Framingham data using two-stage method
1.1 Review of 3 study designs

1. Cross-sectional study:

- Information on the disease status ($Y$) and the exposure status ($X$) is obtained from a random sample at one time point. A snapshot of population.

- A single observation of each variable of interest is measured from each subject: ($Y_i, X_i$) ($i = 1, ..., n$). Regression such as logistic regression (if $Y_i$ is binary) can be used to assess the association between $Y$ and $X$:

  \[
  \log \left( \frac{P[Y_i = 1|X_i]}{1 - P[Y_i = 1|X_i]} \right) = \beta_0 + \beta_1 X_i, \\
  \beta_1 = \log \left( \frac{P[Y = 1|X = 1]/(1 - P[Y = 1|X = 1])}{P[Y = 1|X = 0]/(1 - P[Y = 1|X = 0])} \right).
  \]

$\beta_1 = \log$ odds-ratio between exposure population ($X = 1$) and non-exposure population ($X = 0$). $\beta_1 > 0 \implies$ the exposure population has a higher probability of getting the disease.
• Data \((Y_i, X_i)\) can be summarized as

\[
\begin{array}{c|cc}
\text{Y = 1} & \text{Y = 0} \\
\hline
X = 1 & n_{11} & n_{10} \\
X = 0 & n_{01} & n_{00}
\end{array}
\]

Then the MLE of \(\beta_1\) is given by

\[
\hat{\beta}_1 = \log \left( \frac{n_{11}n_{00}}{n_{10}n_{01}} \right).
\]

• Feature: All numbers \(n_{00}, n_{01}, n_{10}, n_{11}\) are random.

• No causal inference can be made! \(\hat{\beta}_1\) may not be stable (e.g., \(n_{11}\) may be too small). Useful public health information can be obtained, such as the proportion of people in the population with the disease, the proportion of people in the population under exposure.

• Can account for confounders in the model.
2. Prospective cohort study (follow-up study):

- A cohort with known exposure status \((X)\) is followed over time to obtain their disease status \((Y)\).
- A single observation of \((Y)\) may be observed (e.g., disease status or survival status in a survival study) or multiple observations of \((Y)\) may be observed (longitudinal study).
- Stronger evidence for causal inference. Causal inference can be made if \(X\) is assigned randomly (if \(X\) is a treatment indicator in the case of clinical trials).
- When single binary (0/1) \(Y\) (e.g., disease status) is obtained, we have

\[
\begin{array}{ccc}
D & \overline{D} \\
E & n_{11} & n_{10} & n_{1+} \\
\bar{E} & n_{01} & n_{00} & n_{0+}
\end{array}
\]

Here, \(n_{1+}\) and \(n_{0+}\) are fixed (sample sizes for the exposure and non-exposure groups).
3. Retrospective (case-control) study:

- A sample with known disease status ($D$) is drawn and their exposure history ($E$) is ascertained. Data can be summarized as

\[
\begin{array}{cc}
D & \bar{D} \\
E & n_{11} & n_{10} \\
\bar{E} & n_{01} & n_{00} \\
\hline
n_{+1} & n_{+0}
\end{array}
\]

where the margins $n_{+1}$ and $n_{+0}$ are fixed numbers.

- Assuming no bias in obtaining history information on $E$, association between $E$ and $D$ can be estimated.

\[
\begin{align*}
n_{11} & \sim Bin(n_{+1}, P[E|D]), \\
n_{10} & \sim Bin(n_{+0}, P[E|\bar{D}]).
\end{align*}
\]

**Odds ratio:** estimate from this study

\[
\hat{\theta} = \frac{n_{11}n_{00}}{n_{10}n_{01}}
\]
estimates the following quantity

\[ \theta = \frac{P[E|D]/(1 - P[E|D])}{P[E|\overline{D}]/(1 - P[E|\overline{D}]}) \]

\[ \approx \frac{P[D|E]}{P[D|\overline{E}]} \]

- If disease is rare, i.e., \( P[D|E] \approx 0 \), \( P[D|\overline{E}] \approx 0 \), relative risk of disease can be approximately obtained:

\[ \theta \approx \frac{P[D|E]}{P[D|\overline{E}]} = \text{relative risk} \]

More efficient than prospective cohort study in this case.

- **Problem**: recall bias! (it is difficult to ascertain exposure history \( E \).)
1.2 Introduction to longitudinal studies

A longitudinal study is a *prospective cohort* study where repeated measures are taken over time for each individual.

A longitudinal study is usually designed to answer the following questions:

1. How does the variable of interest change over time?
2. How is the (change of) variable of interest associated with treatment and other covariates?
3. How does the variable of interest relate to each other over time?
4. ...
1.3 Data examples

Example 1: Framingham study

In the Framingham study, each of 2634 participants was examined every 2 years for a 10 year period for his/her cholesterol level.

Study objectives:

1. How does cholesterol level change over time on average as people get older?

2. How is the change of cholesterol level associated with sex and baseline age?

3. Do males as a group have more stable (true) baseline cholesterol level and change rate than females?

A subset of 200 subjects’ data is used for illustrative purpose.
### A glimpse of the raw data

```
newid id cholst sex age time
1 1244 175 1 32 0
1 1244 198 1 32 2
1 1244 205 1 32 4
1 1244 228 1 32 6
1 1244 214 1 32 8
1 1244 214 1 32 10
2 835 299 0 34 0
2 835 328 0 34 4
2 835 374 0 34 6
2 835 362 0 34 8
2 835 370 0 34 10
3 176 250 0 41 0
3 176 277 0 41 2
3 176 265 0 41 4
3 176 254 0 41 6
3 176 263 0 41 8
3 176 268 0 41 10
4 901 243 0 44 0
4 901 211 0 44 2
4 901 204 0 44 4
4 901 196 0 44 6
4 901 246 0 44 8
```
Cholesterol level over time for a subset of 200 subjects from Framingham study
What we observed from this data set:

1. Cholesterol levels increase (linearly) over time for most individuals.

2. Each subject has his/her own trajectory line with a possibly different intercept and slope, implying two sources of variations: within and between subject variations.

3. Each subject has on average 5 observations (as opposed to one observation per subject for a cross-sectional study)

4. The data is not balanced. Some individuals have missing observations (e.g., subject 2’s Cholesterol is missing at \( time = 2 \))

5. The inference is NOT limited to these 200 individuals. Instead, the inference is for the target population and each subject is viewed as a random person drawn from the target population.
Example 2: Respiratory Infection Disease

Each of 275 Indonesian preschool children was examined up to six consecutive quarters for the presence of respiratory infection (yes/no). Information on age, sex, height for age, xerophthalmia (vitamin A deficiency) was also obtained.

Study objectives:

- Was the risk of respiratory infection related to vitamin A deficiency after adjusting for age, sex, and height for age, etc.?

Features of this data set:

1. Outcome is whether or not a child has respiratory infection, i.e., binary outcome.

2. Some covariates (age, vitamin A deficiency and height) are time-varying covariates and some are one-time covariates (e.g., gender).
A glimpse of the infection data

Print the first 20 observations

<table>
<thead>
<tr>
<th>Obs</th>
<th>id</th>
<th>infect</th>
<th>xero</th>
<th>sex</th>
<th>visit</th>
<th>season</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>121013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>121013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>121013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>121013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>121013</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>121013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>121113</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>121113</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>121113</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>121113</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>121113</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>121113</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>121114</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>121114</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>121114</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>121114</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>121114</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>121114</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>121140</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>121140</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Proportions of respiratory infection and vitamin A deficiency
Example 3: Epileptic seizure counts from the progabide trial

In the progabide trial, 59 epileptics were randomly assigned to receive the anti-epileptic treatment (progabide) or placebo. The number of seizure counts was recorded in 4 consecutive 2-week intervals. Age and baseline seizure counts (in an 8-week period prior to the treatment assignment) were also recorded.

Study objectives:

- Does the treatment work?
- What is the treatment effect adjusting for available covariates?

Features of this data set:

1. Outcome is count data, implying a Poisson regression model.
2. Baseline seizure counts were for 8 weeks, as opposed to 2 weeks for other seizure counts.
3. Randomization may be taken into account in the data analysis.
### A glimpse of the seizure data

Print the first 20 observations

<table>
<thead>
<tr>
<th>Obs</th>
<th>id</th>
<th>seizure</th>
<th>trt</th>
<th>visit</th>
<th>interval</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>101</td>
<td>76</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>101</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>101</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>101</td>
<td>9</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>101</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>102</td>
<td>38</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>7</td>
<td>102</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>8</td>
<td>102</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>9</td>
<td>102</td>
<td>9</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>102</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>11</td>
<td>103</td>
<td>19</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>103</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>13</td>
<td>103</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>14</td>
<td>103</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>15</td>
<td>103</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>104</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>31</td>
</tr>
<tr>
<td>17</td>
<td>104</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>18</td>
<td>104</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>19</td>
<td>104</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>20</td>
<td>104</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>31</td>
</tr>
</tbody>
</table>
Epileptic seizure counts from the progabide trial

Seizure counts for progabide arm

Seizure counts for control arm
1.4 Features of longitudinal data

Common features of all examples:

- Each subject has multiple time-ordered observations of response.
- Responses from the same subjects may be “more alike” than others.
- Inference is NOT in study subjects, but in population from which they are from.
- # of subjects >> # of observations/subject
- Source of variations – between and within subject variations.

Difference in the examples:

- Different types of responses (continuous, binary, count).
- Objectives depend on the type of study – “mean” behavior, etc.
Comparison of data structures:

<table>
<thead>
<tr>
<th></th>
<th>Classical study</th>
<th>Longitudinal study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject Data</td>
<td>Subject Data</td>
<td>Time</td>
</tr>
<tr>
<td>1   $x_1$</td>
<td>1 $x_{11}, x_{12}, \ldots, x_{15}$</td>
<td>$t_{11}, t_{12}, \ldots, t_{15}$</td>
</tr>
<tr>
<td></td>
<td>$y_1$</td>
<td>$y_{11}, y_{12}, \ldots, y_{15}$</td>
</tr>
<tr>
<td>2   $x_2$</td>
<td>2 $x_{21}, x_{22}, \ldots, x_{25}$</td>
<td>$t_{21}, t_{22}, \ldots, t_{25}$</td>
</tr>
<tr>
<td></td>
<td>$y_2$</td>
<td>$y_{21}, y_{22}, \ldots, y_{25}$</td>
</tr>
</tbody>
</table>

For simplicity, we consider one covariate case.
1.5 Why longitudinal studies?

1. A longitudinal study allows us to study the change of the variable of interest over time, either at population level or individual level.

2. A longitudinal study enables us to separately estimate the cross-sectional effect (e.g., cohort effect) and the longitudinal effect (e.g., aging effect):

Given \( y_{ij}, \text{age}_{ij} \) \((j = 1, 2, \cdots, n_i, j = 1 \text{ is the baseline})\). In a cross-sectional study, \( n_i = 1 \) and we are forced to fit the following model

\[
y_{i1} = \beta_0 + \beta_C \text{age}_{i1} + \varepsilon_{i1}.
\]

That is, \( \beta_C \) is the cross-sectional effect of age.

With longitudinal data \((n_i > 1)\), we can entertain the model

\[
y_{ij} = \beta_0 + \beta_C \text{age}_{i1} + \beta_L (\text{age}_{ij} - \text{age}_{i1}) + \varepsilon_{ij}.
\]
Then

\[ y_{i1} = \beta_0 + \beta_{C\text{age}} i_1 + \varepsilon_{i1} \quad (\text{let } j = 1), \]
\[ y_{ij} - y_{i1} = \beta_L (\text{age}_{ij} - \text{age}_{i1}) + \varepsilon_{ij} - \varepsilon_{i1}. \]

That is, \( \beta_L \) is the longitudinal effect of age and in general \( \beta_L \neq \beta_C \).

3. A longitudinal study is more powerful to detect an association of interest compared to a cross-sectional study, \( \implies \) more efficient, less sample size (number of subjects).

4. A longitudinal study allows us to study the within-subject and between-subject variations.

Suppose \( b \sim (\mu, \sigma_b^2) \) is the blood pressure for a patient population. However, what we observe is \( Y = b + \varepsilon \), where \( \varepsilon \sim (0, \sigma_\varepsilon^2) \) is the measurement error.

- \( \sigma_\varepsilon^2 = \) within-subject variation,
- \( \sigma_b^2 = \) between-subject variation.
If we have only one observation $Y_i$ for each subject from a sample of $n$ patients, then we can’t separate $\sigma^2_\varepsilon$ and $\sigma^2_b$. Although we can use data $Y_1, Y_2, ..., Y_n$ to make inference on $\mu$, we can’t make any inference on $\sigma^2_b$.

However, if we have repeated (or longitudinal) measurements $Y_{ij}$ of blood pressure for each subjects, then

$$Y_{ij} = b_i + \varepsilon_{ij}.$$ 

Now, it is possible to make inference about all quantities $\mu, \sigma^2_b$ and $\sigma^2_\varepsilon$.

5. A longitudinal study provides more evidence for possible causal interpretation.
1.6 Challenges in analyzing longitudinal data

Key assumptions in a classical regression model: There is only one observation of response per subject, \( \implies \) responses are independent to each other. For example, when \( y = \) cholesterol level,

\[
y_i = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{age}_i + \varepsilon_i.
\]

However, the observations from the same subject in a longitudinal study tend to be more similar to each other than those observations from other subjects, \( \implies \) responses (from the same subjects) are not independent any more. Although, the observations from different subjects are still independent.

What happens if we treat observations as independent (i.e., ignore the correlation)?

1. In general, the estimation of the associations (regression coefficients) of the outcome and covariates is valid.
2. However, the variability measures (e.g., the SEs from a classical regression analysis) are not right: sometimes smaller, sometimes bigger than the true variability.

3. Therefore, the inference is not valid (too significant than it should be if the SE is too small).

Sources of variation and correlation in longitudinal data:

1. Between-subject variation: For the blood pressure example, if each subject’s blood pressures were measured within a relatively short time, then the following model may be a reasonable one:

\[ y_{ij} = b_i + \varepsilon_{ij}, \]

where \( b_i \) is the true blood pressure of subject \( i \) with variance \( \sigma_b^2 \), \( \varepsilon_{ij} \) is the independent (random) measurement error with variance \( \sigma_\varepsilon^2 \), independent of \( b_i \).
For $j \neq k$, 

$$\text{corr}(y_{ij}, y_{ik}) = \frac{\text{cov}(y_{ij}, y_{ik})}{\sqrt{\text{var}(y_{ij})\text{var}(y_{ik})}}$$

$$= \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2_\varepsilon}.$$ 

Therefore, if the between-subject variation $\sigma_b^2 \neq 0$, then data from the same subjects are correlated.
The blood pressure example
2. Serial correlation: If the time intervals between blood pressure measurements are relatively large so it may not be reasonable to assume a constant blood pressure for each subject:

\[ y_{ij} = b_i + U_i(t_{ij}) + \varepsilon_{ij}, \]

where \( b_i \) = true long-term blood pressure, \( U_i(t_{ij}) \) = a stochastic process (like a time series) due to biological fluctuation of blood pressure, \( \varepsilon_{ij} \) is the independent (random) measurement error. Here the correlation is caused by both \( b_i \) and \( U_i(t_{ij}) \).

3. In a typical longitudinal study for human where \# of observations/subject is small to moderate, there may not be enough information for the serial correlation and most correlation can be accounted for by (possibly complicated) between-subject variation.
1.7 Methods for analyzing longitudinal data

1. Two-stage: summarize each subject’s outcome and regress the summary statistics on *one-time* covariates. Especially useful for continuous longitudinal data. However, this method is getting out-dated since the mixed model approach can do the same thing even better.

2. Mixed (effects) model approach: model fixed effects and random effects; use random effects to model correlation.

3. Generalized estimating equation (GEE) approach: model the dependence of marginal mean of the response on covariates. Correlation is not a main interest. Particularly good for discrete data.

1.8 Two-stage method for analyzing longitudinal data

- Outcome (usually continuous): $y_{i1}, ..., y_{in_i}$ measured at $t_{i1}, ..., t_{in_i}$; one-time covariates: $x_{i1}, ..., x_{ip}$.

- Two-stage analysis is conducted as follows:
  1. Stage 1: Get summary statistics from subject $i$’s data: $y_{i1}, ..., y_{in_i}$. For example, use mean $\bar{y}_i = (y_{i1} + \cdots + y_{in_i})/n_i$ or fit a linear regression for each subject:

    $$y_{ij} = b_{i0} + b_{i1}t_{ij} + \varepsilon_{ij},$$

    and get estimates $\hat{b}_{i0}, \hat{b}_{i1}$ of $b_{i0}$ and $b_{i1}$. Here we assume that subject $i$’s true response at time $t_{ij}$ is given by

    $$b_{i0} + b_{i1}t_{ij},$$

    a straight line. Suppose $t = 0$ is the baseline, then $b_{i0}$ is subject $i$’s true response at baseline and $b_{i1}$ is subject $i$’s change rate of the
true response ($y$ is the observed response). The error term $\varepsilon_{ij}$ can be regarded as measurement error.

2. Stage 2: Treat the summary statistics as new responses and regress the summary statistics on one-time covariates. For example, after we got $\hat{b}_{i0}$ and $\hat{b}_{i1}$, we can calculate the means of $\hat{b}_{i0}$ and $\hat{b}_{i1}$ and the standard errors of those means, compare $\hat{b}_{i0}$, $\hat{b}_{i1}$ among genders, or do the following regressions

$$
\hat{b}_{i0} = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p + e_{i0},
$$
$$
\hat{b}_{i1} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e_{i1}.
$$

Here, $\alpha_k$ is the effect of $x_k$ on the true baseline response (not $y$), $\beta_k$ is the effect of $x_k$ on the change rate of the true response.
1.9 Analyzing Framingham data using two-stage method

Example 1(a) The Framingham study:

- Stage I: For each subject, fit

\[ y_{ij} = b_{i0} + b_{i1}t_{ij} + \varepsilon_{ij}, \]

and get estimates \( \hat{b}_{i0} \) and \( \hat{b}_{i1} \).

SAS program for stage I:

```sas
options ls=80 ps=200;

data cholst;
  infile "cholst.dat";
  input newid id cholst sex age time;
run;

proc sort;
  by newid time;
run;

proc print data=cholst (obs=20);
  var newid cholst sex age time;
run;
```
title "First stage in two-stage analysis";
proc reg outest=out noprint;
   model cholst = time;
   by newid;
run;

data out; set out;
   b0hat = intercept;
   b1hat = time;
   keep newid b0hat b1hat;
run;

data main; merge cholst out;
   by newid;
   if first.newid=1;
run;

title "Summary statistics for intercepts and slopes";
proc means mean stderr var t probt;
   var b0hat b1hat;
run;

title "Correlation between intercepts and slopes";
proc corr;
   var b0hat b1hat;
run;
Part of output from above SAS program:

Summary statistics for intercepts and slopes

The MEANS Procedure

| Variable | Mean    | Std Error | Variance | t Value | Pr > |t| |
|----------|---------|-----------|----------|---------|------|---|
| b0hat    | 220.6893518 | 2.9478698 | 1737.99 | 74.86   | <.0001|
| b1hat    | 2.5502529    | 0.2566421 | 13.1730374 | 9.94   | <.0001|

Correlation between intercepts and slopes

The CORR Procedure

2 Variables: b0hat b1hat

Simple Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Sum</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>b0hat</td>
<td>200</td>
<td>220.68935</td>
<td>41.68917</td>
<td>44138</td>
<td>141.14286</td>
<td>360.16667</td>
</tr>
<tr>
<td>b1hat</td>
<td>200</td>
<td>2.55025</td>
<td>3.62947</td>
<td>510.05058</td>
<td>-14.00000</td>
<td>11.74286</td>
</tr>
</tbody>
</table>

Pearson Correlation Coefficients, N = 200

Prob > |r| under H0: Rho=0

<table>
<thead>
<tr>
<th></th>
<th>b0hat</th>
<th>b1hat</th>
</tr>
</thead>
<tbody>
<tr>
<td>b0hat</td>
<td>1.00000</td>
<td>-0.26939</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0001</td>
</tr>
<tr>
<td>b1hat</td>
<td>-0.26939</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0001</td>
</tr>
</tbody>
</table>
## Summary statistics from stage 1:

| Parameter | mean | SE  | $t$  | $P[T \geq |t|]$ |
|-----------|------|-----|------|-----------------|
| $\hat{b}_0$ | 221  | 3   | 75   | < .0001         |
| $\hat{b}_1$ | 2.55 | 0.257 | 10 | < .0001         |

$S^2_{b_0} = 1738, \quad S^2_{b_1} = 13.2.$

**Note:**

1. Similar to the blood pressure example, we can use the sample means of $\hat{b}_0$ and $\hat{b}_1$ to estimate the means of $b_0$ and $b_1$. Hence we can use sample mean of $\hat{b}_1$ (2.55) and its SE (0.257) to answer the first objective of this study.

2. However, since $\text{var}(\hat{b}_{i0})$ and $\text{var}(\hat{b}_{i1})$ contain variability due to estimating the **true** baseline response $b_{i0}$ and change rate $b_{i1}$ for individual $i$, so

$$\text{var}(\hat{b}_{i0}) > \text{var}(b_{i0}), \quad \text{var}(\hat{b}_{i1}) > \text{var}(b_{i1}).$$
Sample variances $S^2_{b_0}$ and $S^2_{b_1}$ are unbiased estimates of $\text{var}(\hat{b}_{i0})$ and $\text{var}(\hat{b}_{i1})$ and would overestimate $\text{var}(b_{i0})$ and $\text{var}(b_{i1})$.

3. Similarly,

$$\text{corr}(\hat{b}_0, \hat{b}_1) \neq \text{corr}(b_0, b_1).$$

Therefore, $\text{corr}(\hat{b}_0, \hat{b}_1) = -0.27$ cannot be used to estimate the correlation between the true baseline response $b_0$ and true change rate $b_1$.

4. We will use mixed model approach to address the above issues later.
• Stage II:
  1. Try to compare $E(b_0)$ and $E(b_1)$ between males and females.
  2. Try to compare $\text{var}(b_0)$ and $\text{var}(b_1)$ between males and females.
  3. Try to examine the effects of age and sex on $b_0$ using

\[
\hat{b}_0 = \alpha_0 + \alpha_1 \text{sex} + \alpha_2 \text{age} + e_0.
\]

Technically, we should use $b_0$ instead of $\hat{b}_0$. However, $\hat{b}_0$ is an unbiased estimate of $b_0$ (and $b_0$ is not observable), so using $\hat{b}_0$ is valid.

  4. Try to examine the effects of age and sex on $b_1$ using

\[
\hat{b}_1 = \beta_0 + \beta_1 \text{sex} + \beta_2 \text{age} + e_1.
\]

Similar to the above argument, using $\hat{b}_1$ here is valid.
SAS program for stage II:

```sas
title "Test equality of mean and variance of intercepts and slopes between sexes";
proc ttest;
    class sex;
    var b0hat b1hat;
run;

title "Regression to look at the association between intercept and sex, age";
proc reg data=main;
    model b0hat = sex age;
run;

title "Regression to look at the association between slope and sex, age";
proc reg data=main;
    model b1hat = sex age;
run;
```
Part of output from above SAS program:

Test equality of mean and variance of intercepts and slopes between sexes

The TTEST Procedure

Variable: b0hat

<table>
<thead>
<tr>
<th>sex</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Std Err</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>97</td>
<td>224.0</td>
<td>40.2259</td>
<td>4.0843</td>
<td>146.3</td>
<td>348.1</td>
</tr>
<tr>
<td>1</td>
<td>103</td>
<td>217.6</td>
<td>42.9885</td>
<td>4.2358</td>
<td>141.1</td>
<td>360.2</td>
</tr>
<tr>
<td>Diff (1-2)</td>
<td></td>
<td>6.3629</td>
<td>41.6719</td>
<td>5.8960</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sex</th>
<th>Method</th>
<th>Mean</th>
<th>95% CL Mean</th>
<th>Std Dev</th>
<th>95% CL Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>224.0</td>
<td>215.9</td>
<td>232.1</td>
<td>40.2259</td>
</tr>
<tr>
<td>1</td>
<td>Pooled</td>
<td>217.6</td>
<td>209.2</td>
<td>226.0</td>
<td>42.9885</td>
</tr>
<tr>
<td>Diff (1-2)</td>
<td>Satterthwaite</td>
<td>6.3629</td>
<td>-5.2408</td>
<td>17.9666</td>
<td>41.6719</td>
</tr>
</tbody>
</table>

| Method          | Variances | DF | t Value | Pr > |t| |
|-----------------|-----------|----|---------|------|---|
| Pooled          | Equal     | 198| 1.08    | 0.2818 |
| Satterthwaite   | Unequal   | 197.99 | 1.08 | 0.2809 |

Equality of Variances

<table>
<thead>
<tr>
<th>Method</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Folded F</td>
<td>102</td>
<td>96</td>
<td>1.14</td>
<td>0.5117</td>
</tr>
</tbody>
</table>
Variable: \( b_1 \hat{h} \)

<table>
<thead>
<tr>
<th>sex</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Std Err</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>97</td>
<td>1.7454</td>
<td>3.3567</td>
<td>0.3408</td>
<td>-14.0000</td>
<td>8.3000</td>
</tr>
<tr>
<td>1</td>
<td>103</td>
<td>3.3083</td>
<td>3.7282</td>
<td>0.3673</td>
<td>-11.3750</td>
<td>11.7429</td>
</tr>
<tr>
<td>Diff (1-2)</td>
<td></td>
<td>-1.5629</td>
<td>3.5529</td>
<td>0.5027</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sex</th>
<th>Method</th>
<th>Mean</th>
<th>95% CL Mean</th>
<th>Std Dev</th>
<th>95% CL Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>1.7454</td>
<td>1.0688</td>
<td>2.4219</td>
<td>3.3567</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.3083</td>
<td>2.5796</td>
<td>4.0369</td>
<td>3.7282</td>
</tr>
<tr>
<td>Diff (1-2)</td>
<td>Pooled</td>
<td>-1.5629</td>
<td>-2.5542</td>
<td>-0.5716</td>
<td>3.5529</td>
</tr>
<tr>
<td></td>
<td>Satterthwaite</td>
<td>-1.5629</td>
<td>-2.5511</td>
<td>-0.5747</td>
<td></td>
</tr>
</tbody>
</table>

| Method | Variances | DF  | t Value | Pr > |t| |
|--------|-----------|-----|---------|-------|
| Pooled | Equal     | 198 | -3.11   | 0.0022|
| Satterthwaite | Unequal    | 197.61 | -3.12 | 0.0021|

Equality of Variances

<table>
<thead>
<tr>
<th>Method</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Folded F</td>
<td>102</td>
<td>96</td>
<td>1.23</td>
<td>0.2996</td>
</tr>
</tbody>
</table>
Regression to look at the association between intercept and sex, age

The REG Procedure
Model: MODEL1
Dependent Variable: b0hat

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>53715</td>
<td>26857</td>
<td>18.11</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>197</td>
<td>292145</td>
<td>1482.96718</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>199</td>
<td>345859</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 38.50931  R-Square 0.1553
Dependent Mean 220.68935  Adj R-Sq 0.1467
Coeff Var 17.44956

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|------|---|
| Intercept| 1  | 138.21793          | 15.04083       | 9.19    | <.0001 |
| sex      | 1  | -9.75053           | 5.47862        | -1.78   | 0.0767|
| age      | 1  | 2.05576            | 0.34820        | 5.90    | <.0001|
Regression to look at the association between slope and sex, age

The REG Procedure
Model: MODEL1
Dependent Variable: b1hat

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>257.85057</td>
<td>128.92528</td>
<td>10.75</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>197</td>
<td>2363.58387</td>
<td>11.99789</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>199</td>
<td>2621.43443</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 3.46380  R-Square 0.0984
Dependent Mean 2.55025  Adj R-Sq 0.0892
Coeff Var 135.82170

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|-------|---|
| Intercept| 1  | 6.14089            | 1.35288        | 4.54    | <.0001|
| sex      | 1  | 1.73654            | 0.49279        | 3.52    | 0.0005|
| age      | 1  | -0.10538           | 0.03132        | -3.36   | 0.0009|
Summary from Stage II:

1. Comparison of $E(b_0)$ and $E(b_1)$ between males and females:

   $\hat{E}(\hat{b}_0) : 223.97$ (female), $217.6$ (male), $p$-value $= 0.28$,
   $\hat{E}(\hat{b}_1) : 1.75$ (female), $3.31$ (male), $p$-value $= 0.002$.

2. Comparison of $\text{var}(b_0)$ and $\text{var}(b_1)$ between males and females:

   $S_{{\hat{b}_0}}^2 : 1621$ (female), $1848$ (male), $p$-value $= 0.5$,
   $S_{{\hat{b}_1}}^2 : 11.3$ (female), $13.9$ (male), $p$-value $= 0.3$.

   However, the above tests do NOT compare $\text{var}(b_0)$ and $\text{var}(b_1)$
   between males and females. We will use mixed model approach to
   address this problem.

3. Model for true baseline response $b_0$:

   $\hat{b}_0 = \alpha_0 + \alpha_1 \text{sex} + \alpha_2 \text{age} + e_0$,
   $\hat{\alpha}_0 = 138.2(15.0)$, $\hat{\alpha}_1 = -9.75(5.5)$, $\hat{\alpha}_2 = 2.06(0.35)$. 
After adjusting for sex, one year increase in age corresponds to 2 unit increase in baseline cholesterol level. After adjusting for baseline age, on average males’ baseline cholesterol level is about 10 units less than females’.

4. Model for change rate of the true response $b_1$:

$$\hat{b}_1 = \beta_0 + \beta_1 \text{sex} + \beta_2 \text{age} + e_1,$$

$$\hat{\beta}_0 = 6.14(1.35), \quad \hat{\beta}_1 = 1.74(0.5), \quad \hat{\beta}_2 = -0.11(0.03).$$

After adjusting for sex, one year increase in age corresponds to 0.11 less in cholesterol level change rate. After adjusting for baseline age, males’ cholesterol level change rate is 1.74 greater than females’.
Some remarks on two-stage analysis:

1. What summary statistics should be obtained in the first stage analysis depends on what scientific questions you want to answer in the second stage analysis.

2. Two-stage analysis can only be used when the covariates considered are one-time covariates (fixed over time).

3. Summary statistics of a time-varying covariate cannot be used in the second stage analysis because of error in variable issue (the summary statistics have estimation errors in them).

4. When the covariates considered are time-varying covariates, two-stage analysis is not appropriate. Mixed effects modeling or GEE approach can be used.

5. Two-stage analysis can be applied to discrete response (binary or count data). However, mixed effect modeling or GEE approach can be more flexible.
6. Although two-stage approach can be used to make inference on the quantities of interest, it is less efficient compared to the mixed model approach. Therefore, mixed model approach should be used whenever possible.
2 Linear mixed models for normal longitudinal data

- What is a linear mixed model?
  1. Random intercept model
  2. Random intercept and slope model
  3. Other error structures
  4. General mixed models

- Estimation and inference
- Choose a variance matrix of the data
- Analyze Framingham data using linear mixed models
- GEE for mixed models, missing data issue
2.1 What is a linear mixed (effects) model?

A linear mixed model is an extension of a linear regression model to model longitudinal (correlated) data. It contains fixed effects and random effects where random effects are subject-specific and are used to model between-subject variation and the correlation induced by this variation.

What are fixed effects? Fixed effects are the covariate effects that are fixed across subjects in the study sample. These effects are the ones of our particular interest. E.g., the regression coefficients in usual regression models are fixed effects:

\[ y = \alpha + x\beta + \varepsilon. \]

What are random effects? Random effects are the covariate effects that vary among subjects. So these effects are subject-specific and hence are random (unobservable) since each subject is a random subject drawn from a population.
I. Random intercept only model:

Data from $m$ subjects:

<table>
<thead>
<tr>
<th>Subject</th>
<th>Outcome</th>
<th>Time</th>
<th>Random intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$y_{11}, y_{12}, \ldots, y_{1n_1}$</td>
<td>$t_{11}, t_{12}, \ldots, t_{1n_1}$</td>
<td>$b_1$</td>
</tr>
<tr>
<td>2</td>
<td>$y_{21}, y_{22}, \ldots, y_{2n_2}$</td>
<td>$t_{21}, t_{22}, \ldots, t_{2n_2}$</td>
<td>$b_2$</td>
</tr>
<tr>
<td></td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$i$</td>
<td>$y_{i1}, y_{i2}, \ldots, y_{in_i}$</td>
<td>$t_{i1}, t_{i2}, \ldots, t_{in_i}$</td>
<td>$b_i$</td>
</tr>
<tr>
<td></td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$m$</td>
<td>$y_{m1}, y_{m2}, \ldots, y_{mn_m}$</td>
<td>$t_{m1}, t_{m2}, \ldots, t_{mn_m}$</td>
<td>$b_m$</td>
</tr>
</tbody>
</table>

Other covariates: $x_{ij2}, \ldots, x_{ijp}, i = 1, \ldots, m, j = 1, \ldots, n_i.$

A random intercept model assumes:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + b_i + \varepsilon_{ij}.$$
Random intercept model:

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + b_i + \varepsilon_{ij} \]

where \( \beta \)'s are fixed effects of interest, \( b_i \sim N(0, \sigma_b^2) \) are random effects, \( \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \) are independent (measurement) errors.

Interpretation of the model components:

1. From the model,
   \[ \mathbb{E}[y_{ij}] = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp}. \]

2. \( \beta_k \): Average increase in \( y \) associated with one unit increase in \( x_k \), the \( k \)th covariate, while others are held fixed.

3. \( \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + b_i = \text{true response for subject } i \) at \( t_{ij} \).

4. \( \beta_0 + b_i \) is the intercept for subject \( i \) \( \implies b_i = \text{deviation of intercept of subject } i \text{ from population intercept } \beta_0. \)
5. $\sigma^2_b = \text{between-subject variance}$, $\sigma^2_\varepsilon = \text{within-subject variance}$.

6. Total variance of $y$: $\text{Var}(y_{ij}) = \sigma^2_b + \sigma^2_\varepsilon$, constant over time.

7. Correlation between $y_{ij}$ and $y_{ij'}$:

$$\text{corr}(y_{ij}, y_{ij'}) = \frac{\sigma^2_b}{\sigma^2_b + \sigma^2_\varepsilon} = \rho.$$  

8. Correlation is constant and positive.
Why treat $b_i$ as random

1. Treating $b_i$ as random enables us to make inference for the whole population from which the sample was drawn. Treating $b_i$ as fixed would only allow us to make inference for the study sample.

2. Usually $n_i$ is small for longitudinal studies. Therefore, as the number of total data points gets larger, the number of $b_i$ (which is $m$, the number of subjects) gets large proportionally. In this case, the standard properties (such as consistency) of the parameter estimates may not still hold if $b_i$ is treated as fixed.
When no $x$, random intercept only model reduces to

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_i + \varepsilon_{ij}.$$
II. Random intercept and slope model:

Data from \( m \) subjects:

<table>
<thead>
<tr>
<th>Subject</th>
<th>Outcome</th>
<th>Time</th>
<th>Random intercept</th>
<th>Random slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( y_{11}, \ldots, y_{1n_1} )</td>
<td>( t_{11}, \ldots, t_{1n_1} )</td>
<td>( b_{01} )</td>
<td>( b_{11} )</td>
</tr>
<tr>
<td>2</td>
<td>( y_{21}, \ldots, y_{2n_2} )</td>
<td>( t_{21}, \ldots, t_{2n_2} )</td>
<td>( b_{02} )</td>
<td>( b_{21} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( i )</td>
<td>( y_{i1}, \ldots, y_{in_i} )</td>
<td>( t_{i1}, \ldots, t_{in_i} )</td>
<td>( b_{i0} )</td>
<td>( b_{i1} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( m )</td>
<td>( y_{m1}, \ldots, y_{mn_m} )</td>
<td>( t_{m1}, \ldots, t_{mn_m} )</td>
<td>( b_{m0} )</td>
<td>( b_{m1} )</td>
</tr>
</tbody>
</table>

Other covariates: \( x_{ij2}, \ldots, x_{ijp}, i = 1, \ldots, m, j = 1, \ldots, n_i \).

A random intercept and slope model assumes:

\[
y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}.
\]
Random intercept and slope model:

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}, \]

\( \beta_k \) the same as before, random effects \( b_{i0}, b_{i1} \) are assumed to have a bivariate normal distribution

\[
\begin{pmatrix}
    b_{i0} \\
    b_{i1}
\end{pmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{01} & \sigma_{11} \end{bmatrix} \right).
\]

Usually, no constraint is imposed on \( \sigma_{ij}; \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \).

Interpretation of the model components:

1. Mean structure is the same as before:

\[ \mathbb{E}[y_{ij}] = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp}. \]

2. \( \beta_k \): Average increase in \( y \) associated with one unit increase in \( x_k \), the \( k \)th covariate, while others are held fixed.

3. \( \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + b_{i0} + b_{i1} t_{ij} = \text{true response for} \)

Graduate Summer Session in Epidemiology  Slide 58
subject \( i \) at \( t_{ij} \).

4. \( \beta_0 + b_i = \) the intercept for subject \( i \) \( \Rightarrow b_{i0} = \) deviation of intercept of subject \( i \) from population intercept \( \beta_0 \)

5. \( \beta_1 + b_{i1} = \) the slope for subject \( i \) \( \Rightarrow b_{i1} = \) deviation of slope of subject \( i \) from population slope \( \beta_1 \)

6. \( \text{Var}(b_{i0} + b_{i1}t_{ij}) = \sigma_00 + 2t_{ij}\sigma_{01} + t_{ij}^2\sigma_{11} = \) between-subject variance (varying over time).

7. \( \sigma_{\varepsilon}^2 = \) within-subject variance.

8. Total variance of \( y \): \( \text{Var}(y_{ij}) = \sigma_00 + 2t_{ij}\sigma_{01} + t_{ij}^2\sigma_{11} + \sigma_{\varepsilon}^2, \) not a constant over time.

9. Correlation between \( y_{ij} \) and \( y_{ij'} \): not a constant over time.
When no $x$, random intercept and slope model reduces to

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_i + b_1 t_{ij} + \varepsilon_{ij}.$$  

Graphical representation of data from random intercept and slope model
III. Other mixed models:

- A correlated error model

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + e_{ij}, \]

where \( e_{ij} \) are correlated (for the same subjects) normal errors (In the previous models, \( e_{ij} \) contains random effects and \( \varepsilon_{ij} \)).

For example,

1. Compound symmetric (exchangeable) variance matrix

\[
\begin{pmatrix}
  e_{i1} \\
  e_{i2} \\
  e_{i3}
\end{pmatrix} \sim N
\begin{bmatrix}
  0 \\
  0 \\
  0
\end{bmatrix}, \sigma^2
\begin{bmatrix}
  1 & \rho & \rho \\
  \rho & 1 & \rho \\
  \rho & \rho & 1
\end{bmatrix}.
\]

Here, \(-1 < \rho < 1\). A random intercept model is almost equivalent to this model.
2. AR(1) variance matrix

\[
\begin{pmatrix}
  e_{i1} \\
  e_{i2} \\
  e_{i3}
\end{pmatrix}
\sim N
\begin{pmatrix}
  0 \\
  0 \\
  0
\end{pmatrix},
\sigma^2
\begin{pmatrix}
  1 & \rho & \rho^2 \\
  \rho & 1 & \rho \\
  \rho^2 & \rho & 1
\end{pmatrix}.
\]

Here, \(-1 < \rho < 1\). It assumes that the error \((e_{i1}, e_{i2}, e_{i3})^T\) is an autoregressive process with order 1. This structure is more appropriate if \(y\) is measured at equally spaced time points.

3. Spatial power variance matrix

\[
\begin{pmatrix}
  e_{i1} \\
  e_{i2} \\
  e_{i3}
\end{pmatrix}
\sim N
\begin{pmatrix}
  0 \\
  0 \\
  0
\end{pmatrix},
\sigma^2
\begin{pmatrix}
  1 & \rho|t_2-t_1| & \rho|t_3-t_1| \\
  \rho|t_2-t_1| & 1 & \rho|t_3-t_2| \\
  \rho|t_3-t_1| & \rho|t_3-t_2| & 1
\end{pmatrix}.
\]

Here, \(0 < \rho < 1\). This error structure reduces to AR(1) when \(y\) is measured at equally spaced time points. This structure is appropriate if \(y\) is measured at unequally spaced time points.
4. Unstructured variance matrix

\[
\begin{pmatrix}
e_{i1} \\
e_{i2} \\
e_{i3}
\end{pmatrix} \sim \mathcal{N}
\begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}
,
\begin{bmatrix}
\sigma_{11} & \sigma_{12} & \sigma_{13} \\
\sigma_{12} & \sigma_{22} & \sigma_{23} \\
\sigma_{13} & \sigma_{23} & \sigma_{33}
\end{bmatrix}
\]

Here no restriction is imposed on \( \sigma_{ij} \). This structure may be used only if (potential) time points are the same for all subjects and the number is relatively small. We will have to estimate too many parameters if patients entered the study randomly.
IV. General linear mixed models

General model 1: fixed effects + random effects + pure measurement error:

For example,

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}, \]

where \( \varepsilon_{ij} \) is the pure measurement error (has an independent error structure with a constant variance).

Software to implement the above model: Proc Mixed in SAS:

```sas
Proc Mixed data= method=;
   class id;
   model y = t x / s; /* specify t x for fixed effects */
   random intercept t / subject=id type=un; /* specify the covariance matrix */
       /* for random effects */
   repeated / subject=id type=vc; /* specify the variance structure for error */
run;
```

\( vc \) = variance component (independent error).
General model 2: fixed effects + random effects + stochastic process

For example,
\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij} + b_{i0} + b_{i1} t_{ij} + U_i(t_{ij}), \]

where \( U_i(t) \) is a stochastic process with AR(1), a spatial power variance structure, or other variance structure.

Software to implement the above model: Proc Mixed in SAS:

```
Proc Mixed data= method=;
  class id;
  model y = t x / s; /* specify t x for fixed effects */
  random intercept t / subject=id type=un; /* specify the covariance matrix */
    /* for random effects */
  repeated / subject=id type=sp(pow)(t); /* specify the variance structure for error */
run;
```

If the time points are equally spaced, we can use type=ar(1) in the repeated statement for AR(1) variance structure for \( U_i(t) \):

```
repeated cat_t / subject=id type=ar(1); /* cat_t is class t */
```
**General model 3**: fixed effects + random effects + stochastic process + pure measurement error

For example,

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij} + b_{i0} + b_{i1} t_{ij} + U_i(t_{ij}) + \varepsilon_{ij}, \]

where \( U_i(t) \) is a stochastic process with some variance structure (e.g., a spatial power variance structure), \( \varepsilon_{ij} \) is the pure measurement error.

**Software** to implement the above model: Proc Mixed in SAS:

```
proc mixed data= method=;
   class id;
   model y = t x / s; /* specify t x for fixed effects */
   random intercept t / subject=id type=un; /* specify the covariance matrix */
       /* for random effects */
   repeated / subject=id type=sp(pow)(t) local; /* specify error variance structure */
   run;
```

If the time points are equally spaced, we can use `type=ar(1)` in the repeated statement if assuming AR(1) for \( U_i(t) \):

```
repeated cat_t / subject=id type=ar(1) local; /* cat_t is class t */
```
General model 4: fixed effects + un-structured error

For example,

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij} + e_{ij}, \]

where \( e_{ij} \) is the error with un-structured variance matrix.

Software to implement the above model: Proc Mixed in SAS:

```sas
Proc Mixed data= method=;
  class id;
  model y = t x / s; /* specify t x for fixed effects */
  repeated cat_t / subject=id type=un; /* specify error variance structure */
run;
```

Note that no “random” statement can be used in the above model. When the number of different time points is big, there will be too many parameters to estimate.
2.2 Estimation and inference for linear mixed models

Let $\theta$ consist of all parms in random effects (e.g., $b_{i0}, b_{i1}$) and errors ($\varepsilon_{ij}$). We want to make inference on $\beta$ and $\theta$. There are two approaches:

1. Maximum likelihood:

$$\ell(\beta, \theta; y) = \log L(\beta, \theta; y).$$

Maximize $\ell(\beta, \theta; y)$ jointly w.r.t. $\beta$ and $\theta$ to get their MLEs.

2. Restricted maximum likelihood (REML):

(a) Get REML estimate of $\theta$ from a REML likelihood $\ell_{REML}(\theta; y)$ (take into account the estimation of $\beta$). Leads to less biased $\hat{\theta}_{REML}$. For example, in a linear regression model

$$\hat{\sigma}^2_{REML} = \frac{\text{Residual Sum of Squares}}{n - p - 1}.$$ 

(b) Estimate $\beta$ by maximizing $\ell(\beta, \hat{\theta}_{REML}; y)$. 

Graduate Summer Session in Epidemiology  Slide 68
Hypothesis Testing

- After we fit a linear mixed model such as

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + b_i + b_1 t_{ij} + \varepsilon_{ij}, \]

SAS will output a test for each \( \beta_k \), including the estimate, SE, p-value (for testing \( H_0 : \beta_k = 0 \)), etc.

- If we want to test a contrast between \( \beta_k \) or estimate a linear combination of \( \beta_k \)'s, we can use estimate statement in Proc Mixed. Then SAS will output the estimate, its SE and the p-value for testing the linear combination is zero. See Programs 2 and 3 for Framingham data.
2.3 How to choose random effects and the error structure?

1. Use graphical representation to identify possible random effects.

2. Use biological knowledge to identify possible error structure.

3. Use information criteria to choose a final model:
   (a) Akaike's Information Criterion (AIC):
   \[
   AIC = -2\{\ell(\hat{\beta}, \hat{\theta}; y) - q\}
   \]
   where \( q = \# \) of elements in \( \theta \). Smaller AIC is preferred.

   (b) Bayesian Information Criterion (BIC):
   \[
   BIC = -2\{\ell(\hat{\beta}, \hat{\theta}; y) - 0.5 \times q \times \log(m)\}, \quad m = \# \ of \ subjects
   \]
   Again, smaller BIC is preferred.
2.4 Analyze Framingham data using linear mixed models

- Model to address objective 1: How does cholesterol level change over time on average as people get older?

- Consider the following basic model suggested by the data:

\[
y_{ij} = b_{i0}^* + b_{i1}^* t_{ij} + \varepsilon_{ij} \tag{2.1}
\]

where \( y_{ij} \) is the \( j \)th cholesterol level measurement from subject \( i \), \( t_{ij} \) is year from the beginning of the study (or baseline) and \( b_{i0}^* \), \( b_{i1}^* \) are random variables distributed as

\[
\begin{pmatrix}
b_{i0}^* \\
b_{i1}^*
\end{pmatrix} \sim N\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{01} & \sigma_{11} \end{bmatrix}\right),
\]

and \( \varepsilon_{ij} \) are independent errors distributed as \( N(0, \sigma_\varepsilon^2) \).
Model (2.1) assumes that

1. The true cholesterol level for each individual changes linearly over time with a different intercept and slope, which are both random (since the individual is a random subject drawn from the population).

2. Since $t = 0$ is the baseline, so $b_{i0}^*$ can be viewed as the true but unobserved cholesterol level for subject $i$ at the baseline, and $b_{i1}^*$ can be viewed as the change rate of the true cholesterol level for subject $i$.

3. $\beta_0$ is the population average of the true baseline cholesterol level of all individuals in the population, $\beta_1$ is the population average change rate of true cholesterol level and it tells us how cholesterol level changes on average as people get older. So $\beta_1$ is the longitudinal effect or aging effect on cholesterol level.

4. $\sigma_{00}$ is the variance of the true baseline cholesterol level $b_{i0}^*$; $\sigma_{11}$ is the variance of the change rate $b_{i1}^*$ of the true cholesterol level; and $\sigma_{01}$ is the covariance between true baseline cholesterol level
\( b_{i0}^* \) and the change rate \( b_{i1}^* \) of true cholesterol level.

\* The random variables \( b_{i0}^* \) and \( b_{i1}^* \) can be re-written as

\[ b_{i0}^* = \beta_0 + b_{i0}, \quad b_{i1}^* = \beta_1 + b_{i1}, \]

where \( b_{i0}, b_{i1} \) have the following distribution:

\[
\begin{pmatrix}
  b_{i0} \\
  b_{i1}
\end{pmatrix}
\sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix}
  \sigma_{00} & \sigma_{01} \\
  \sigma_{01} & \sigma_{11}
\end{bmatrix} \right).
\]

\* Model (2.1) then can be re-expressed as

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}. \quad (2.2) \]

Therefore, \( \beta_0, \beta_1 \) are fixed effects and \( b_{i0}, b_{i1} \) are random effects.
The following is the SAS program for fitting model (2.2):

```
title "Framingham data: mixed model without covariates";
proc mixed data=cholst;
    class newid;
    model cholst = time / s;
    random intercept time / type=un subject=newid g;
    repeated / type=vc subject=newid;
run;
```

The following is the output from the above program:

```
Framingham data: mixed model without covariates  1

The Mixed Procedure

Model Information

  Data Set                 WORK.CHOLST
  Dependent Variable      cholst
  Covariance Structures   Unstructured, Variance Components
  Subject Effects         newid, newid
  Estimation Method       REML
  Residual Variance Method Parameter
  Fixed Effects SE Method  Model-Based
  Degrees of Freedom Method  Containment
```
### Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>newid</td>
<td>200</td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 ...</td>
</tr>
</tbody>
</table>

### Dimensions

- **Covariance Parameters**: 4
- **Columns in X**: 2
- **Columns in Z Per Subject**: 2
- **Subjects**: 200
- **Max Obs Per Subject**: 6
- **Observations Used**: 1044
- **Observations Not Used**: 0
- **Total Observations**: 1044

### Iteration History

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Evaluations</th>
<th>-2 Res Log Like</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>10899.75433605</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>9960.12567386</td>
<td>0.00000120</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>9960.12082968</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

Convergence criteria met.
The Mixed Procedure

Estimated G Matrix

<table>
<thead>
<tr>
<th>Row</th>
<th>Effect</th>
<th>newid</th>
<th>Col1</th>
<th>Col2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>1</td>
<td>1467.30</td>
<td>-2.2259</td>
</tr>
<tr>
<td>2</td>
<td>time</td>
<td>1</td>
<td>-2.2259</td>
<td>3.8409</td>
</tr>
</tbody>
</table>

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>newid</td>
<td>1467.30</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>newid</td>
<td>-2.2259</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>newid</td>
<td>3.8409</td>
</tr>
<tr>
<td>Residual</td>
<td>newid</td>
<td>434.11</td>
</tr>
</tbody>
</table>

Fit Statistics

-2 Res Log Likelihood 9960.1
AIC (smaller is better) 9968.1
AICC (smaller is better) 9968.2
BIC (smaller is better) 9981.3

Null Model Likelihood Ratio Test

<table>
<thead>
<tr>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>939.63</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
Solution for Fixed Effects

| Effect   | Estimate | Standard Error | DF | t Value | Pr > |t| |
|----------|----------|----------------|----|---------|------|---|
| Intercept| 220.57   | 2.9305         | 199| 75.26   | <.0001|
| time     | 2.8170   | 0.2408         | 191| 11.70   | <.0001|

Type 3 Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>1</td>
<td>191</td>
<td>136.83</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

From this output, we see that:

1. \( \hat{\sigma}_{00} = 1467 \), as compared to \( \hat{\text{var}}(\hat{b}_0) = 1738 \) from the two-stage approach.

2. \( \hat{\sigma}_{11} = 3.84 \), as compared to \( \hat{\text{var}}(\hat{b}_1) = 13.2 \) from the two-stage approach.

3. \( \hat{\text{corr}}(\hat{b}_{i0}^*, \hat{b}_{i1}^*) = \hat{\text{corr}}(\hat{b}_{i0}, \hat{b}_{i1}) = -2.2259/\sqrt{1467 \times 3.84} = -0.03 \), as compared to \( \hat{\text{corr}}(\hat{b}_0, \hat{b}_1) = -0.27 \) from the two-stage approach.

4. The estimated mean of true baseline cholesterol level is \( \hat{\beta}_0 = 220.57 \) with SE=2.93, as compared to the sample mean
220.69 of $\hat{b}_0$ with SE = 2.94 from the two-stage approach.

5. The estimated change rate (longitudinal effect) $\hat{\beta}_1 = 2.82$ with SE=0.24, as compared to the sample mean 2.55 of $\hat{b}_1$ with SE = 0.26 from the two-stage approach.

6. $\hat{\sigma}_\varepsilon^2 = 434.11$.

- Q: Is it reasonable to assume $\varepsilon_{ij}$ in model (2.1) to be pure measurement error?
- We can consider a more general model such as AR(1) for $\varepsilon_{ij}$ and test this assumption.

```sas
data cholst; set cholst;
cat_time = time;
run;

title "Framingham data: mixed model without covariates + AR(1) error";
proc mixed data=cholst covtest;
class newid cat_time;
model cholst = time / s;
random intercept time / type=un subject=newid g;
repeated cat_time / type=ar(1) subject=newid;
run;
```
and the relevant output:

### Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Z</th>
<th>Pr Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>newid</td>
<td>1478.76</td>
<td>174.15</td>
<td>8.49</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>newid</td>
<td>-3.5618</td>
<td>10.7033</td>
<td>-0.33</td>
<td>0.7393</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>newid</td>
<td>4.1717</td>
<td>1.3186</td>
<td>3.16</td>
<td>0.0008</td>
</tr>
<tr>
<td>AR(1)</td>
<td>newid</td>
<td>-0.03193</td>
<td>0.06156</td>
<td>-0.52</td>
<td>0.6039</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>425.06</td>
<td>28.4010</td>
<td>14.97</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

### Fit Statistics

-2 Res Log Likelihood: 9959.9
AIC (smaller is better): 9969.9
AICC (smaller is better): 9969.9
BIC (smaller is better): 9986.3

* Note:

1. P-value for testing $H_0 : \rho = 0$ is 0.6039, no strong evidence against $H_0$.
2. All model selection criteria lead to iid error $\varepsilon_{ij}$.
3. We usually don’t use the above output to test variances because of the boundary issue.
• Model to investigate the cross sectional age effect and longitudinal age effect on cholesterol level:
  
  ★ Re-write the true baseline cholesterol level $b_{i0}^*$ and the change rate $b_{i1}^*$ in model (2.1) in terms of conditional distributions given age:

  $b_{i0}^* = \beta_0 + \beta_{C\text{age}}i + b_{i0}$ \hspace{1cm} (2.3)

  $b_{i1}^* = \beta_1 + \beta_{A\text{age}}i + b_{i1}$, \hspace{1cm} (2.4)

  where $\text{age}_i$ is individual $i$’s baseline age. Note that these $b_{i0}$ and $b_{i1}$ are different from those in the previous model.

  Then $\beta_C$ is the cross sectional age effect and $\beta_1 + \beta_{A\text{age}}i$ is the longitudinal effect for the population with baseline age equal to $\text{age}_i$.

  ★ The average longitudinal effect is

  $\beta_1 + \beta_A\mathbb{E}(\text{age})$,
which can be estimated by

\[ \hat{\beta}_1 + \hat{\beta}_A \overline{age}, \]

where \( \overline{age} \) is the sample average of age.

\( \star \) Suggest that we can center age and use the centered age (denoted by \( \text{cent} \_ \text{age}_i = age_i - \overline{age} \)) in (2.3). Then \( \beta_1 \) is the average longitudinal effect.

\( \star \) We are interested in testing \( H_0 : \beta_C = \beta_1. \)

\( \star \) Assume the usual distribution for \((b_{i0}, b_{i1})\):

\[
\begin{pmatrix}
  b_{i0} \\
  b_{i1}
\end{pmatrix}
\sim
\mathcal{N}
\left(
\begin{bmatrix}
  0 \\
  0
\end{bmatrix},
\begin{bmatrix}
  \sigma_{00} & \sigma_{01} \\
  \sigma_{01} & \sigma_{11}
\end{bmatrix}
\right).
\]

Here both \( \sigma_{00} \) and \( \sigma_{11} \) are the remaining variances in \( b_{i0}^* \) and \( b_{i1}^* \) after baseline age effect has been taken into account. So they should be smaller than those corresponding values in model (2.1).
Basic model (2.1) becomes

\[ y_{ij} = \beta_0 + \beta_{\text{cent\_age}_i} + \beta_1 t_{ij} + \beta_{A\text{cent\_age}_i} \times t_{ij} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}, \]  

where \( \varepsilon_{ij} \sim N(0, \sigma^2) \) are independent errors.

The following is the SAS program for fitting model (2.5):

```sas
data cholst; set cholst;
  cent_age = age - 42.56;
run;

title "Framingham data: longitudinal effect vs. cohort effect";
proc mixed data=cholst;
  class newid;
  model cholst = time cent_age cent_age*time / s;
  random intercept time / type=un subject=newid g;
  repeated / type=vc subject=newid;
  estimate "long-cross" time 1 cent_age -1;
run;
```
The relevant output of the above SAS program is

**Iteration History**

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Evaluations</th>
<th>-2 Res Log Like</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>10826.01576300</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>9929.74817925</td>
<td>0.00000516</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>9929.72729664</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

Convergence criteria met.

**Estimated G Matrix**

<table>
<thead>
<tr>
<th>Row</th>
<th>Effect</th>
<th>newid</th>
<th>Col1</th>
<th>Col2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>1</td>
<td>1226.69</td>
<td>9.7829</td>
</tr>
<tr>
<td>2</td>
<td>time</td>
<td>1</td>
<td>9.7829</td>
<td>3.2598</td>
</tr>
</tbody>
</table>

**Covariance Parameter Estimates**

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>newid</td>
<td>1226.69</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>newid</td>
<td>9.7829</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>newid</td>
<td>3.2598</td>
</tr>
<tr>
<td>Residual</td>
<td>newid</td>
<td>434.15</td>
</tr>
</tbody>
</table>

**Fit Statistics**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 Res Log Likelihood</td>
<td>9929.7</td>
</tr>
<tr>
<td>AIC (smaller is better)</td>
<td>9937.7</td>
</tr>
</tbody>
</table>
AICC (smaller is better) 9937.8  
BIC (smaller is better) 9950.9  

Null Model Likelihood Ratio Test  

<table>
<thead>
<tr>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>896.29</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects  

| Effect       | Estimate | Error   | DF  | t Value | Pr > |t| |
|--------------|----------|---------|-----|---------|-------|
| Intercept    | 220.57   | 2.7172  | 198 | 81.18   | <.0001|
| time         | 2.8157   | 0.2343  | 190 | 12.02   | <.0001|
| cent_age     | 1.9861   | 0.3455  | 652 | 5.75    | <.0001|
| time*cent_age| -0.1024  | 0.02930 | 652 | -3.50   | 0.0005|

Type 3 Tests of Fixed Effects  

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>1</td>
<td>190</td>
<td>144.42</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>cent_age</td>
<td>1</td>
<td>652</td>
<td>33.05</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>time*cent_age</td>
<td>1</td>
<td>652</td>
<td>12.22</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

Estimates  

| Label       | Estimate | Error  | DF  | t Value | Pr > |t| |
|-------------|----------|--------|-----|---------|-------|
| long-cross  | 0.8296   | 0.4174 | 652 | 1.99    | 0.0473|
What we learn from this output:

1. $\hat{\sigma}_{00} = 1226.7$, much smaller than the corresponding estimate 1467 from model (2.1) since baseline age was used to explain the variability in the true baseline cholesterol level.

2. $\hat{\sigma}_{11} = 3.26$, much smaller than the corresponding estimate 3.84 from model (2.1) since baseline age was used to explain the variability in the true baseline cholesterol change rate.

3. $\hat{\beta}_0 = 220.57$ is the estimate of mean true baseline cholesterol level for the individuals whose baseline age = 42.56 (the average age), which is the same as the one from model (2.1) but with a smaller SE (2.71 vs. 2.93).

4. The estimate of the longitudinal age effect is $\hat{\beta}_1 = 2.8157$ with SE = 0.2343, which is basically the same as $\hat{\beta}_1 = 2.8170$ with SE = 0.24 from model (2.1).

5. The estimate of the cross sectional age effect is $\hat{\beta}_C = 1.99$ with SE = 0.3455, which is very different from the estimate of the longitudinal age effect $\hat{\beta}_1 = 2.82$. 
6. The P-value for testing $H_0: \beta_L = \beta_C$ is 0.0473, significant at level 0.05!

7. $\hat{\sigma}_\varepsilon^2 = 434.15$ is basically the same as the corresponding estimate from model (2.1), which is 434.11.

8. Similarly, we can test $i.i.d. \varepsilon_{ij}$ by considering correlated errors such as AR(1) for $\varepsilon_{ij}$ and test to see if $\rho = 0$. 
Model to address **objective 2**: How is the change of cholesterol level associated with sex and baseline age?

Re-write the true baseline cholesterol level $b_{i0}^*$ and the change rate $b_{i1}^*$ in model (2.1) in terms of conditional distribution given gender and baseline age:

$$b_{i0}^* = \beta_0 + sex_i \beta_{0,sex} + age_i \beta_{0,age} + b_{i0}, \quad (2.6)$$

$$b_{i1}^* = \beta_1 + sex_i \beta_{1,sex} + age_i \beta_{1,age} + b_{i1}, \quad (2.7)$$

where we assume that $b_{i0}, b_{i1}$ have the following distribution

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{01} & \sigma_{11} \end{bmatrix} \right).$$

Then $\beta_{0,sex}, \beta_{0,age}$ are the sex effect and baseline age effect on the baseline cholesterol level. Of course, $\beta_0$ does **NOT** have a proper interpretation.
Similarly, \( \beta_{1,sex}, \beta_{1,age} \) are the sex effect and baseline age effect on the change rate of the true cholesterol level, and \( \beta_1 \) does **NOT** have a proper interpretation.

Substituting the above expressions into model (2.1), we got

\[
y_{ij} = \beta_0 + sex_i \beta_{0,sex} + age_i \beta_{0,age} + \beta_1 t_{ij} \\
+ sex_i t_{ij} \beta_{1,sex} + age_i t_{ij} \beta_{1,age} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}. \quad (2.8)
\]

Suppose we also want to test whether or not the change rates between 30 years old males and 40 years old females are the same using the above model.

From model (2.7), the (average) change rate of 30 years old males is

\[
\beta_1 + 1 \times \beta_{1,sex} + 30 \times \beta_{1,age} = \beta_1 + \beta_{1,sex} + 30 \beta_{1,age}.
\]
The (average) change rate of 40 years old females is

\[ \beta_1 + 0 \times \beta_{1,sex} + 40 \times \beta_{1,age} = \beta_1 + 40\beta_{1,age}. \]

The difference between these two rates is

\[ \beta_1 + \beta_{1,sex} + 30\beta_{1,age} - (\beta_1 + 40\beta_{1,age}) = \beta_{1,sex} - 10\beta_{1,age}. \]

Therefore, we need only to test \( H_0 : \beta_{1,sex} - 10\beta_{1,age} = 0. \)

⋆ We can use the following SAS program to answer our questions.

```sas
title "Framingham data: how baseline cholesterol level and";
title2" change rate depend on sex and baseline age";
proc mixed data=cholst;
  class newid;
  model cholst = sex age time sex*time age*time / s;
  random intercept time / type=un subject=newid g s;
  repeated / type=vc subject=newid;
  estimate "rate-diff" sex*time 1 age*time -10;
run;
```
Part of the relevant output from above program is

Framingham data: how baseline cholesterol level and change rate depend on sex and baseline age

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Evaluations</th>
<th>-2 Res Log Like</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>10813.99587154</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>9907.89014721</td>
<td>0.00000655</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>9907.86364103</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

Convergence criteria met.

Estimated G Matrix

<table>
<thead>
<tr>
<th>Row</th>
<th>Effect</th>
<th>newid</th>
<th>Col1</th>
<th>Col2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>1</td>
<td>1209.89</td>
<td>13.5502</td>
</tr>
<tr>
<td>2</td>
<td>time</td>
<td>1</td>
<td>13.5502</td>
<td>2.5211</td>
</tr>
</tbody>
</table>

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>newid</td>
<td>1209.89</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>newid</td>
<td>13.5502</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>newid</td>
<td>2.5211</td>
</tr>
<tr>
<td>Residual</td>
<td>newid</td>
<td>434.15</td>
</tr>
</tbody>
</table>
Fit Statistics

-2 Res Log Likelihood 9907.9
AIC (smaller is better) 9915.9
AICC (smaller is better) 9915.9
BIC (smaller is better) 9929.1

Null Model Likelihood Ratio Test

<table>
<thead>
<tr>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>906.13</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

| Effect    | Estimate | Standard Error | DF | t Value | Pr > |t| |
|-----------|----------|----------------|----|---------|-------|
| Intercept | 138.18   | 14.9148        | 197| 9.26    | <.0001|
| sex       | -9.6393  | 5.4352         | 652| -1.77   | 0.0766|
| age       | 2.0509   | 0.3454         | 652| 5.94    | <.0001|
| time      | 6.8003   | 1.2229         | 189| 5.56    | <.0001|
| sex*time  | 1.7995   | 0.4536         | 652| 3.97    | <.0001|
| age*time  | -0.1145  | 0.02835        | 652| -4.04   | <.0001|
Solution for Random Effects

| Effect   | newid | Estimate | Std Err  | DF  | t Value | Pr > |t| |
|----------|-------|----------|----------|-----|---------|------|--| |
| Intercept| 2     | 100.50   | 11.5761  | 651 | 8.68    | <.0001 | |
| time     | 2     | 2.7414   | 1.2643   | 651 | 2.17    | 0.0305 | |
| Intercept| 74    | 46.9844  | 11.0096  | 651 | 4.27    | <.0001 | |
| time     | 74    | 1.3579   | 1.2525   | 651 | 1.08    | 0.2787 | |
| Intercept| 171   | -51.5764 | 11.3046  | 651 | -4.56   | <.0001 | |
| time     | 171   | -0.6812  | 1.2583   | 651 | -0.54   | 0.5885 | |

Estimates

| Label    | Estimate | Std Error | DF | t Value | Pr > |t| |
|----------|----------|-----------|----|---------|------|--| |
| rate-diff| 2.9441   | 0.5606    | 651| 5.25    | <.0001 | |
What we learn from this output:

1. $\hat{\beta}_{0,sex} = -9.64$ (SE = 5.43), so after adjusting for baseline age, males’ baseline cholesterol level is about 10 units less than females’.

2. $\hat{\beta}_{0,age} = 2.05$ (SE = 0.35), so after adjusting for gender, one year older people’s baseline cholesterol level is about 2 units higher than that of one year younger people.

3. $\hat{\beta}_{1,sex} = 1.80$ (SE = 0.45), so after adjusting for baseline age, males’ change rate is 1.80 (cholesterol unit/year) greater than females’ change rate. Similar estimate from 2-stage analysis is 1.74 (SE=0.49).

4. $\hat{\beta}_{1,age} = -0.11$ (SE = 0.028), so after adjusting for sex, one year older people’s change rate is 0.11 less than one year younger people’s change rate. Similar estimate from 2-stage analysis is -0.11 (SE=0.031).

5. The change rate difference of interest is 2.94 (SE = 0.56). Significantly different!
6. $\hat{\sigma}_{00} = 1210$, which is smaller than the corresponding estimate from model (2.5) since we use both age and gender to explain the variability in baseline true cholesterol level.

7. $\hat{\sigma}_{11} = 2.52$, which is smaller than the corresponding estimate from model (2.5) since we use both age and gender to explain the variability in the cholesterol level change rate.

8. $\hat{\sigma}_{\varepsilon}^2 = 434.15$, basically the same as its estimates from models (2.1) and (2.5).

9. Similarly, we can test $iid \varepsilon_{ij}$ by considering correlated errors such as AR(1).

★ Note: The models (2.6) and (2.7) for $b_{i0}^*$ and $b_{i1}^*$ are basically the same as the second stage models in the two stage analysis for the Framingham data.

★ Compare results from this model to the results from the two-stage analysis:
(a) Effect on baseline cholesterol level:

Model (2.8) : \( \hat{\beta}_0 = 138.18 (SE = 14.9) , \)
\( \hat{\beta}_{0,sex} = -9.64 (SE = 5.43) , \hat{\beta}_{0,age} = 2.05 (SE = 0.35) ; \)

Two-stage : \( \hat{\alpha}_0 = 138.2 (SE = 15.0) , \)
\( \hat{\alpha}_1 = -9.75 (SE = 5.48) , \hat{\alpha}_2 = 2.06 (SE = 0.35) . \)

(b) Effect on change rate of cholesterol level:

Model (2.8) : \( \hat{\beta}_1 = 6.80 (SE = 1.22) , \)
\( \hat{\beta}_{1,sex} = 1.80 (SE = 0.45) , \hat{\beta}_{1,age} = -0.11 (SE = 0.03) ; \)

Two-stage : \( \hat{\beta}_0 = 6.14 (SE = 1.35) , \)
\( \hat{\beta}_1 = 1.74 (SE = 0.49) , \hat{\beta}_2 = -0.11 (SE = 0.03) . \)

\( \star \) We can also estimate the individual random effects and estimate their trajectory lines.
Estimated subject-specific lines from model (2.8):
• Model to address Objective 3: Do males have more stable (true) baseline cholesterol level and change rate than females?

★ From model (2.1), assume \( b_{i0}^*, b_{i1}^* \) have different distributions for males and females:

\[
\begin{align*}
\text{Males:} & \quad \left( \begin{array}{c} b_{i0}^* \\ b_{i1}^* \end{array} \right) \sim N \left( \begin{bmatrix} \mu_{m0} \\ \mu_{m1} \end{bmatrix}, \begin{bmatrix} \sigma_{m00} & \sigma_{m01} \\ \sigma_{m01} & \sigma_{m11} \end{bmatrix} \right), \\
\text{Females:} & \quad \left( \begin{array}{c} b_{i0}^* \\ b_{i1}^* \end{array} \right) \sim N \left( \begin{bmatrix} \mu_{f0} \\ \mu_{f1} \end{bmatrix}, \begin{bmatrix} \sigma_{f00} & \sigma_{f01} \\ \sigma_{f01} & \sigma_{f11} \end{bmatrix} \right). \tag{2.9}
\end{align*}
\]

★ We would like to test \( H_0 : \sigma_{m00} = \sigma_{f00}, \sigma_{m01} = \sigma_{f01}, \sigma_{m11} = \sigma_{f11} \) (i.e., the above two variance-covariance matrices are the same).
CHAPTER 2

The SAS program and its output for fitting above model are as follows:

```sas
data cholst; set cholst;
gender=sex;
run;

title "Framingham data: do males have more stable (true) baseline";
title2 "cholesterol level and change rate than females?";
proc mixed data=cholst;
class newid gender;
model cholst = sex time sex*time / s;
random intercept time / type=un subject=newid group=gender g;
repeated / type=vc subject=newid;
run;
```

Framingham data: do males have more stable (true) baseline cholesterol level and change rate than females?

The Mixed Procedure

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Evaluations</th>
<th>-2 Res Log Like</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>10889.09479529</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>9939.57691271</td>
<td>0.00000317</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>9939.56399905</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

The Mixed Procedure

Convergence criteria met.
### Estimated G Matrix

<table>
<thead>
<tr>
<th>Row</th>
<th>Effect</th>
<th>newid</th>
<th>gender</th>
<th>Col1</th>
<th>Col2</th>
<th>Col3</th>
<th>Col4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>1</td>
<td>0</td>
<td>1402.47</td>
<td>-4.7015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>time</td>
<td>1</td>
<td>0</td>
<td>-4.7015</td>
<td>1.8279</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Intercept</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1532.81</td>
<td>3.6119</td>
</tr>
<tr>
<td>4</td>
<td>time</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>3.6119</td>
<td>4.7970</td>
</tr>
</tbody>
</table>

### Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Group</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>newid</td>
<td>gender 0</td>
<td>1402.47</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>newid</td>
<td>gender 0</td>
<td>-4.7015</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>newid</td>
<td>gender 0</td>
<td>1.8279</td>
</tr>
<tr>
<td>UN(1,1)</td>
<td>newid</td>
<td>gender 1</td>
<td>1532.81</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>newid</td>
<td>gender 1</td>
<td>3.6119</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>newid</td>
<td>gender 1</td>
<td>4.7970</td>
</tr>
<tr>
<td>Residual</td>
<td>newid</td>
<td></td>
<td>433.71</td>
</tr>
</tbody>
</table>

### Fit Statistics

-2 Res Log Likelihood: 9939.6
AIC (smaller is better): 9953.6
AICC (smaller is better): 9953.7
BIC (smaller is better): 9976.7
In order to test $H_0$: the two variance matrices are the same using the likelihood ratio test (LRT), we need to fit a model with the same fixed and random effects but under $H_0$. The following is the SAS program and its output under $H_0$. This null model is called model (2.9).
Estimated G Matrix

<table>
<thead>
<tr>
<th>Row</th>
<th>Effect</th>
<th>newid</th>
<th>Col1</th>
<th>Col2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>1</td>
<td>1465.85</td>
<td>-0.2516</td>
</tr>
<tr>
<td>2</td>
<td>time</td>
<td>1</td>
<td>-0.2516</td>
<td>3.2618</td>
</tr>
</tbody>
</table>

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>newid</td>
<td>1465.85</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>newid</td>
<td>-0.2516</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>newid</td>
<td>3.2618</td>
</tr>
<tr>
<td>Residual</td>
<td>newid</td>
<td>434.17</td>
</tr>
</tbody>
</table>

Fit Statistics

-2 Res Log Likelihood   9943.0
AIC (smaller is better) 9951.0
AICC (smaller is better) 9951.1
BIC (smaller is better)  9964.2

* The difference of $-2$ residual log likelihood is $9943 - 9939.6 = 3.4$ (between models (2.9) and (2.90)) and the P-value = $P[\chi^2_3 \geq 3.4] = 0.33$. 
Note: We can also test $H_0$: whether or not males and females have the same variance matrices of true baseline cholesterol level and change rate of cholesterol level by adjusting for baseline age and sex. We already fit the model under $H_0$ (model (2.8)) and $-2$ residual log likelihood is 9907.9. The alternative model can be fit using the following SAS program (called model $(2.8_A)$).

```sas
title "Framingham data: do males have more stable (true) baseline cholesterol";
title2 "level and change rate than females adjusting for sex and baseline age";
proc mixed data=cholst;
   class newid gender;
   model cholst = sex age time sex*time age*time / s;
   random intercept time / type=un subject=newid group=gender g;
   repeated / type=vc subject=newid;
run;
```
Part of the output from above program is

Framingham data: do males have more stable (true) baseline cholesterol level and change rate than females adjusting for sex and baseline age

The Mixed Procedure
Model Information

Data Set WORK.CHOLST
Dependent Variable cholst
Covariance Structures Unstructured, Variance

The Mixed Procedure
Convergence criteria met.

Estimated G Matrix

<table>
<thead>
<tr>
<th>Row</th>
<th>Effect</th>
<th>newid</th>
<th>gender</th>
<th>Col1</th>
<th>Col2</th>
<th>Col3</th>
<th>Col4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intercept</td>
<td>1</td>
<td>0</td>
<td>1403.04</td>
<td>-2.6077</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>time</td>
<td>1</td>
<td>0</td>
<td>-2.6077</td>
<td>1.5955</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Intercept</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1021.77</td>
<td>30.7768</td>
</tr>
<tr>
<td>4</td>
<td>time</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>30.7768</td>
<td>3.3214</td>
</tr>
</tbody>
</table>

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Group</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>newid</td>
<td>gender 0</td>
<td>1403.04</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>newid</td>
<td>gender 0</td>
<td>-2.6077</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>newid</td>
<td>gender 0</td>
<td>1.5955</td>
</tr>
<tr>
<td>UN(1,1)</td>
<td>newid</td>
<td>gender 1</td>
<td>1021.77</td>
</tr>
</tbody>
</table>
UN(2,1)  newid  gender 1  30.7768
UN(2,2)  newid  gender 1  3.3214
Residual newid  gender 1  434.93

Fit Statistics

-2 Res Log Likelihood  9901.6
AIC (smaller is better)  9915.6
AICC (smaller is better)  9915.7
BIC (smaller is better)  9938.7

* The -2 residual log likelihood is 9901.6 so difference is 9907.9-9901.6 = 6.3. The P-value = $P[\chi^2_3 \geq 6.3] = 0.09$, more evidence against $H_0$. 
Comparison of fit statistics among models

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model (2.1)</td>
<td>9968.1</td>
<td>9981.3</td>
</tr>
<tr>
<td>Model (2.5)</td>
<td>9937.</td>
<td>9950.9</td>
</tr>
<tr>
<td>Model (2.8)</td>
<td>9915.9</td>
<td>9929.1</td>
</tr>
<tr>
<td>Model (2.9)</td>
<td>9953.6</td>
<td>9976.7</td>
</tr>
<tr>
<td>Model (2.9)_0</td>
<td>9951.0</td>
<td>9964.2</td>
</tr>
<tr>
<td>Model (2.8_A)</td>
<td>9915.6</td>
<td>9938.7</td>
</tr>
</tbody>
</table>
• Note:

1. The choice fixed effects terms depends on the questions we need to answer. After that is settled, we can use AIC or BIC to determine the random effects and the error structure (with default method=REML).

2. If we want to pick a (good) model among many models with different fixed effects and random effects (+ error structures), we can use AIC or BIC as a guide for model selection, but have to specify method=ML (in Proc Mixed).
2.5 GEE for linear mixed models

- When the variation pattern in data is so complicated that we don’t feel comfortable in the random effects and their variance structure we imposed, we can use the model we posed to estimate the fixed effects (β’s) and use the GEE approach to calculate the SEs for the fixed effect estimates. These SE estimates will be valid regardless of the validity of the random effects structure we put. So these SE estimates are robust (we will talk more on Thursday).

- For example, we can use the following model to estimate β’s:

\[
y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 \text{sex}_i + \beta_3 \text{age}_i + \beta_4 \text{sex}_i t_{ij} + \beta_5 \text{age}_i t_{ij} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}.
\]

If we specify empirical in Proc mixed, we will get robust SE estimates. See the following SAS program and output.
title "Using GEE to fit Framingham data";
proc mixed data=cholst empirical;
   class newid;
   model cholst = time sex age sex*time age*time / s;
   random intercept time / type=un subject=newid;
   repeated / type=vc subject=newid;
run;

Output of the above program

Using GEE to fit Framingham data 24

The Mixed Procedure

Model Information

Data Set WORK.CHOLST
Dependent Variable cholst
Covariance Structures Unstructured, Variance Components
Subject Effects newid, newid
Estimation Method REML
Residual Variance Method Parameter
Fixed Effects SE Method Empirical
Degrees of Freedom Method Containment

Iteration History

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Evaluations</th>
<th>-2 Res Log Like</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>10813.99587154</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>9907.89014721</td>
<td>0.00000655</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>9907.86364103</td>
<td>0.00000000</td>
</tr>
</tbody>
</table>

Convergence criteria met.
The Mixed Procedure

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UN(1,1)</td>
<td>newid</td>
<td>1209.89</td>
</tr>
<tr>
<td>UN(2,1)</td>
<td>newid</td>
<td>13.5502</td>
</tr>
<tr>
<td>UN(2,2)</td>
<td>newid</td>
<td>2.5211</td>
</tr>
<tr>
<td>Residual</td>
<td>newid</td>
<td>434.15</td>
</tr>
</tbody>
</table>

Fit Statistics

-2 Res Log Likelihood 9907.9
AIC (smaller is better) 9915.9
AICC (smaller is better) 9915.9
BIC (smaller is better) 9929.1

Null Model Likelihood Ratio Test

<table>
<thead>
<tr>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>906.13</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Solution for Fixed Effects

| Effect   | Estimate | Standard Error | DF | t Value | Pr > |t| |
|----------|----------|----------------|----|---------|-------|
| Intercept| 138.18   | 15.4017        | 197| 8.97    | <.0001|
| sex      | -9.6393  | 5.4588         | 651| -1.77   | 0.0779|
| age      | 2.0509   | 0.3749         | 651| 5.47    | <.0001|
| time     | 6.8003   | 1.2188         | 190| 5.58    | <.0001|
| time*sex | 1.7995   | 0.4524         | 651| 3.98    | <.0001|
| time*age | -0.1145  | 0.02868        | 651| -3.99   | <.0001|
What we observed:

1. Fixed effects estimates and variance-covariance parameter estimates are exactly the same as those from model (2.8).

2. The SEs for the fixed effects estimates are different from those from model (2.8). However, they are very close, indicating model (2.8) has a reasonably good fit to the data and we don’t have to use the GEE approach.
2.6 Missing data issues

However, GEE will be less efficient if a correct model can be specified; with missing data, the missing data mechanism has to be *missing completely at random* (MCAR) for the GEE inference to be valid.

Missing data mechanism:

1. *missing completely at random* (MCAR): The reason that the data are missing has nothing to do with anything, i.e., at each time point, the observed data can be viewed as a random sample from the population.

2. *missing at random* (MAR): The reason that a subject has missing data does not depend on his/her un-observed data. Mixed model inference is valid under this condition. MCAR implies MAR.

3. *missing not at random* (MNAR): The reason that a subject has missing data depends on his/her unobserved data. Special assumption (untestable) has to be made for inference.
Ways to assess MCAR

1. Suppose the missing data pattern (for \( y \)) looks like

\[
\begin{array}{ccc}
1 & 2 & 3 \\
\hline
? & ? & ? \\
? & ? & ?
\end{array}
\]

and assume \( x \) (such as age) is a completely observed variable.

2. Compare \( x \) for the two groups with observed \( y \) and missing \( y \) at times 2 and 3 (using, say, two-sample t-test). A significant difference indicates the violation of MCAR. Otherwise, we may feel comfortable about the MCAR assumption.

Remark: MAR cannot be tested.
Use age to test MCAR for Framingham data:

```plaintext
options ls=72 ps=72;

data cholst;
  infile "cholst.dat";
  input newid id cholst sex age time;
  if time = . then delete;
run;

data base; set cholst;
  if time=0;
  keep newid age;
run;

data time;
  do newid=1 to 200;
    do time=0 to 10 by 2;
      output;
    end;
  end;
run;

data cholst; merge cholst time;
  by newid time;
  if cholst= . then yobs=0;
  else yobs=1;
  drop age;
run;

data cholst; merge cholst base;
  by newid;
run;

proc sort;
  by time;
run;
```
title "Test equality of age between missing and non-missing groups";
proc ttest;
  var age;
  class yobs;
  by time;
run;

SAS output:
Test equality of age between missing and non-missing groups

-------------------------------- time=2 --------------------------------
The TTEST Procedure
T-Tests

| Variable | Method     | Variances | DF  | t Value | Pr > |t| |
|----------|------------|-----------|-----|---------|------|
| age      | Pooled     | Equal     | 198 | 0.35    | 0.7298 |
| age      | Satterthwaite | Unequal   | 29.6| 0.35    | 0.7325 |

-------------------------------- time=4 --------------------------------
The TTEST Procedure
T-Tests

| Variable | Method     | Variances | DF  | t Value | Pr > |t| |
|----------|------------|-----------|-----|---------|------|
| age      | Pooled     | Equal     | 198 | -0.23   | 0.8172 |
| age      | Satterthwaite | Unequal  | 39.5| -0.22   | 0.8304 |
| Variable | Method      | Variances | DF  | t Value | Pr > |t| |
|----------|-------------|-----------|-----|---------|------|---|
| age      | Pooled      | Equal     | 198 | 1.43    | 0.1536 |
| age      | Satterthwaite| Unequal   | 40.3| 1.37    | 0.1774 |
| age      | Pooled      | Equal     | 198 | 0.47    | 0.6418 |
| age      | Satterthwaite| Unequal   | 47  | 0.50    | 0.6179 |
| age      | Pooled      | Equal     | 198 | 0.24    | 0.8071 |
| age      | Satterthwaite| Unequal   | 63.3| 0.27    | 0.7879 |
3 Modeling and design issues

- How to handle baseline response?
- Do we model previous responses as covariates?
- Modeling response vs. modeling the change of response
- A simulation study comparing modeling response to modeling its change
- Design a longitudinal study. Sample size calculation
  1. Comparing time-averaged means
  2. Comparing slopes
3.1 How to handle baseline response?

- Model baseline outcome as part of the response. For example,

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, m, j = 1, 2, \ldots, n_i, \quad (3.1) \]

where the errors \( \epsilon_{ij} \) may include random effects and other errors, and hence are correlated. For example, \( \epsilon_{ij} = b_i + \epsilon_{ij} \) for a random intercept model.

- Model baseline outcome as a covariate. For example,

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 y_{i1} + e_{ij}, \quad i = 1, \ldots, m, j = 2, \ldots, n_i. \quad (3.2) \]

Comments:

1. There are some subtle difference between these two models. The regression parameters \( \beta_0, \beta_1 \) and the variance components have different interpretation and hence we will get different estimates from two models. \( \beta_1 \) in model (3.1) is the overall effect of \( x \) on \( y \), while \( \beta_1 \)
in model (3.2) is the adjusted covariate effect of \( x \) on \( y \) adjusting for the baseline response.

2. Model (3.2) is more convenient for prediction. Although one can also get a prediction model similar to model (3.2) by conditioning on the baseline response from model (3.1).

3. When baseline response \( y_{i1} \) is used as a covariate, it CANNOT be re-used in the outcome variable. For model (3.2), index \( j \) goes from 2 to \( n_i \). Because of this, the estimates from model (3.1) may be more efficient.

4. It is obvious that in the presence of missing data, the subjects with baseline measurements only will be deleted from the analysis if model (3.2) is used. In the case where missingness depends on the baseline measurements, inference using model (3.2) will be invalid. However, model (3.1) may still give valid inference. We will see a simulation study later.
3.2 Do we model previous responses as covariates?

One might consider an auto-regressive type of model like the following one instead of (3.1):

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 y_{i,j-1} + \epsilon_{ij}, \quad i = 1, \ldots, m, j = 2, \ldots, n_i. \]  

(3.3)

**Comments:**

1. This model is different from models (3.1) and (3.2). Here $\beta_1$ is the adjusted effect of $x$ on $y$ after adjusting for the previous response. Therefore, they have different interpretation.

2. Since we allow the current response depends on the previous response in this model, part of the correlation among responses is taken away by the coefficient $\beta_2$. Hence the errors may have a much simpler variance structure than the errors in model (3.1). In fact, people often assume $\epsilon_{ij}$ in (3.3) to be independent. This is an example of transition models. Consequently, the variance component parameters in this
model are different and have different interpretation from those in model (3.1).

3. We can obtain a similar model to this one if we assume the errors in model (3.1) have an AR(1) variance structure.

4. Similar to model (3.2), this model is more convenient for prediction.

5. Similar to model (3.2), subjects with baseline measurements only will be deleted from the analysis. If missingness of subsequent measurements depends on the baseline measurements, this model will give invalid inference on the parameters of interest.
3.3 Modeling outcome vs. modeling the change of outcome

Define change from baseline:

\[ D_{ij} = y_{ij} - y_{i1}, \quad i = 1, ..., m, j = 2, ..., n_i \]

and consider model

\[ D_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}, \quad i = 1, ..., m, j = 2, ..., n_i. \] (3.4)

Comments:

1. This model emphasizes the effect of \( x \) on the change (from baseline value) of outcome. Therefore, \( \beta_1 \) has different interpretation than the \( \beta_1 \)'s in previous models.

2. Since we are modeling the difference, part of the correlation in the responses due to among individual variation is removed. Therefore, the errors in this model will have a simpler variance structure than model
(3.1), and the parameters in the variance structures have different interpretation.

3. Baseline outcome $y_{i1}$ can be used as a covariate.

4. It cannot model how $x$ affects the overall mean of outcome.

5. Similar to models (3.2) and (3.3), subjects with baseline measurements only will be deleted from the analysis, and if missingness depends on the baseline measurements, the inference will be invalid.

6. If the data from different subjects were taken from different time points, the new response defined by the change from the baseline may not be very meaningful.

7. Which model to use depends on the scientific questions we want to address.
We generated data from the following model:

\[ y_{ij} = \beta_0 + \beta_1 t_j + b_i + \varepsilon_{ij}, \quad i = 1, \ldots, 50, \quad j = 1, 2, \]

where \( \beta_0 = 1, \beta_1 = 2, t_1 = 0, t_2 = 1, b_i \sim N(0, 1), \varepsilon_{ij} \sim N(0, 1). \)

1. \( y_{i1} \) can be viewed as pre-test (or baseline) score, \( y_{i2} \) can be viewed as post-test score for subject \( i. \)

2. In the simulation, we let \( y_{i2} \) be missing whenever the baseline measurement \( y_{i1} \) is negative.

3. \( \beta_1 = \mathbb{E}(y_{i2} - y_{i1}) = \mathbb{E}(y_{i2}) - \mathbb{E}(y_{i1}). \) We would like to make inference on \( \beta_1 \) in the presence of missing data.
### One simulated data set:

<table>
<thead>
<tr>
<th>Obs</th>
<th>id</th>
<th>score0</th>
<th>score1</th>
<th>scoredif</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1.33662</td>
<td>1.96479</td>
<td>0.62816</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.17404</td>
<td>1.93052</td>
<td>1.75648</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1.45672</td>
<td>5.07021</td>
<td>3.61349</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1.08229</td>
<td>3.71837</td>
<td>2.63608</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0.55392</td>
<td>2.51172</td>
<td>1.95780</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1.73579</td>
<td>3.43906</td>
<td>1.70327</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>-0.27640</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>0.78154</td>
<td>1.60275</td>
<td>0.82121</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>-0.33015</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>-1.11409</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>1.54039</td>
<td>2.02123</td>
<td>0.48084</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>1.20696</td>
<td>2.19839</td>
<td>0.99143</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>1.35767</td>
<td>2.33060</td>
<td>0.97293</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>0.68858</td>
<td>1.55404</td>
<td>0.86545</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>0.81951</td>
<td>3.78494</td>
<td>2.96542</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>0.49849</td>
<td>1.40747</td>
<td>0.90897</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>-1.68078</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>2.31063</td>
<td>3.70494</td>
<td>1.39431</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>1.05800</td>
<td>2.22613</td>
<td>1.16813</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>1.00388</td>
<td>4.72160</td>
<td>3.71773</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>4.45060</td>
<td>7.63933</td>
<td>3.18873</td>
</tr>
<tr>
<td>22</td>
<td>22</td>
<td>2.20755</td>
<td>2.18365</td>
<td>-0.02390</td>
</tr>
<tr>
<td>23</td>
<td>23</td>
<td>1.02019</td>
<td>1.81962</td>
<td>0.79943</td>
</tr>
<tr>
<td>24</td>
<td>24</td>
<td>2.30880</td>
<td>4.09571</td>
<td>1.78691</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>1.93793</td>
<td>3.26014</td>
<td>1.32222</td>
</tr>
<tr>
<td>26</td>
<td>26</td>
<td>-1.30937</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>27</td>
<td>27</td>
<td>-0.80651</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>28</td>
<td>28</td>
<td>0.65134</td>
<td>4.66953</td>
<td>4.01819</td>
</tr>
<tr>
<td>29</td>
<td>29</td>
<td>0.72529</td>
<td>0.77726</td>
<td>0.05197</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>1.00030</td>
<td>4.76540</td>
<td>3.76511</td>
</tr>
<tr>
<td>31</td>
<td>31</td>
<td>2.75257</td>
<td>5.03208</td>
<td>2.27951</td>
</tr>
<tr>
<td>32</td>
<td>32</td>
<td>-1.71925</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>33</td>
<td>33</td>
<td>0.65070</td>
<td>3.11335</td>
<td>2.46265</td>
</tr>
<tr>
<td>34</td>
<td>34</td>
<td>0.23703</td>
<td>2.03079</td>
<td>1.79376</td>
</tr>
<tr>
<td>35</td>
<td>35</td>
<td>-1.32099</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
1. If we take difference as we did in the previous model, then we would use the sample mean of the non-missing difference (only 38 differences) to estimate $\hat{\beta}_1$, which will give $\hat{\beta}_1 = 1.85$ (SE=0.20). Obviously, this estimate is biased (here it is biased towards zero). This is a special case of two-stage analyses.

2. Since we have a special case of longitudinal studies, we can use mixed model approach to estimate $\beta_1$. For this purpose, we need to re-arrange data in the right format for Proc mixed.
3. The data for the first 10 subjects are given below:

<table>
<thead>
<tr>
<th>Obs</th>
<th>id</th>
<th>score</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1.33662</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.96479</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.17404</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1.93052</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>1.45672</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>5.07021</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1.08229</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>3.71837</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>0.55392</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>2.51172</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>1.73579</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>3.43906</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
<td>-0.27640</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>0.78154</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>8</td>
<td>1.60275</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>9</td>
<td>-0.33015</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>9</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>10</td>
<td>-1.11409</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>.</td>
<td>1</td>
</tr>
</tbody>
</table>

4. We use the following SAS program for estimating $\beta_1$

```sas
proc mixed data=maindat;
  class id;
  model score = time / s;
  random int / subject=id type=vc;
  repeated / subject=id type=vc;
run;
```
5. Part of the output from the above SAS program:

The Mixed Procedure

Model Information

Data Set WORK.MAINDAT
Dependent Variable score
Covariance Structure Variance Components
Subject Effects id, id
Estimation Method REML
Residual Variance Method Parameter
Fixed Effects SE Method Model-Based
Degrees of Freedom Method Containment

Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>50</td>
<td>1 2 ..</td>
</tr>
</tbody>
</table>

Dimensions

- Covariance Parameters: 2
- Columns in X: 2
- Columns in Z Per Subject: 1
- Subjects: 50
- Max Obs Per Subject: 2
- Observations Used: 88
- Observations Not Used: 12
- Total Observations: 100

Convergence criteria met.

Covariance Parameter Estimates
Cov Parm | Subject | Estimate  
---|---|---
Intercept | id | 1.4573  
Residual  | id | 0.7828

Fit Statistics

-2 Res Log Likelihood | 300.6  
AIC (smaller is better) | 304.6  
AICC (smaller is better) | 304.8  
BIC (smaller is better) | 308.4

The SAS System

The Mixed Procedure

Solution for Fixed Effects

| Effect | Estimate | Standard Error | DF | t Value | Pr > |t| |
|---|---|---|---|---|---|
| Intercept | 0.9146 | 0.2117 | 49 | 4.32 | <.0001 |
| time | 2.0503 | 0.1987 | 37 | 10.32 | <.0001 |

Type 3 Tests of Fixed Effects

<table>
<thead>
<tr>
<th>Effect</th>
<th>Num DF</th>
<th>Den DF</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>1</td>
<td>37</td>
<td>106.50</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
The following table gives the simulation results comparing the longitudinal approach modeling all responses simultaneously and the two-stage approach modeling the difference based on 1000 simulation runs:

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>SE</th>
<th>SD</th>
<th>Cov. prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longitudinal approach</td>
<td>2.002</td>
<td>0.222</td>
<td>0.257</td>
<td>0.91</td>
</tr>
<tr>
<td>Two-stage approach</td>
<td>1.712</td>
<td>0.214</td>
<td>0.217</td>
<td>0.72</td>
</tr>
</tbody>
</table>

where **Mean** is the sample mean of 1000 $\hat{\beta}_1$’s from both approaches; **SE** is the sample mean of 1000 estimated SEs of $\hat{\beta}_1$; **SD** is the sample standard deviation of 1000 $\hat{\beta}_1$’s; **Cov. prob.** is the empirical coverage probability of 95% CI of $\beta_1$. 
What we see from this table:

1. The estimate $\hat{\beta}_1$ using longitudinal approach by modeling all responses simultaneously is unbiased; however, if we take difference of the responses (here we are forced to delete all subjects with missing measurements), the estimate $\hat{\beta}_1$ is biased.

2. Although the estimate $\hat{\beta}_1$ from the two-stage approach has slightly smaller SE or SD, since the estimate itself is biased, the coverage probability of the 95% CI of $\beta_1$ is too low, making invalid inference on $\beta_1$. However, the coverage probability of the 95% CI of $\beta_1$ from the longitudinal approach is almost right at the nominal level (0.95).

3. With mixed model approach, we can estimate other quantities.
3.4 Design a longitudinal study: Sample size estimation

Recall that in the classical setting, sample size estimation is posed as a hypothesis testing problem such as the following one

\[ H_0 : \mu_1 = \mu_2 \quad vs \quad H_A : \mu_1 \neq \mu_2. \]

Assume \( y_{1k}, \ldots, y_{mk} \sim \mathcal{N}(\mu_k, \sigma^2), k = 1, 2 \). Given significance level \( \alpha \), power \( \gamma \), and the difference \( \Delta = (\mu_1 - \mu_2)/\sigma \) we wish to detect, the required total sample size (number of subjects) in each group should be

\[
m = 2 \left[ \frac{z_{\alpha/2} + z_{\gamma}}{\Delta} \right]^2.
\]
Design a longitudinal study (cont’d):

I: Compare time-averaged means between two groups.

Assume model for the data to be collected:

Group A: \( y_{ij} = \mu_A + \varepsilon_{ij}, i = 1, \ldots, m, j = 1, \ldots, n, \)

Group B: \( y_{ij} = \mu_B + \varepsilon_{ij}, i = 1, \ldots, m, j = 1, \ldots, n, \)

\( m = \# \) of subjects, \( n = \# \) of observations/subject, \( \varepsilon_{ij} \) normally distributed errors with mean zero, variance \( \sigma^2 \) and correlation \( \rho \).

We want to test

\[ H_0 : \mu_A = \mu_B \quad \text{vs} \quad H_A : \mu_A \neq \mu_B \]

at level \( \alpha \) with power \( \gamma \) to detect difference \( \Delta = (\mu_A - \mu_B)/\sigma \). The quantities \( m \) and \( n \) have to satisfy

\[ m = 2(1 + (n - 1)\rho) \left( \frac{z_{\alpha/2} + z_{1-\gamma}}{n\Delta^2} \right)^2. \]
Comments:

1. When \( n = 1 \), the study reduces to a cross-sectional study and the sample size formula reduces to the classical one.

2. When \( \rho = 0 \) (responses are independent), the required sample size is \( 1/n \) of that for classical study.

3. When \( \rho = 1 \), required sample size is the same as that of the classical study.

4. For fixed \( n \), smaller \( \rho \) gives smaller sample size.

5. If correlation is high, use more subjects and less obs/subject; if correlation is low, use less subjects and more obs/subject.

6. The sample size formula depends on information on \( \sigma^2 \) and \( \rho \).

7. One can choose a combination of \( m \) and \( n \) to meet one’s specific needs.

8. The above formula is for two-sided test.
• An example: If \( n = 3, \alpha = 0.05, \gamma = 0.8 \), then the number of subjects \((m)\) per group is

\[
m = 2(1 + 2\rho) \frac{(1.96 + 0.84)^2}{3\Delta^2}.
\]

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( \Delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>0.2</td>
<td>733</td>
</tr>
<tr>
<td>0.3</td>
<td>838</td>
</tr>
<tr>
<td>0.4</td>
<td>942</td>
</tr>
<tr>
<td>0.5</td>
<td>1047</td>
</tr>
<tr>
<td>0.6</td>
<td>1152</td>
</tr>
<tr>
<td>0.7</td>
<td>1256</td>
</tr>
<tr>
<td>0.8</td>
<td>1361</td>
</tr>
</tbody>
</table>
Design a longitudinal study (cont’d):

II: Compare slopes between two groups.

Model for the data to be collected:

Group A: \( y_{ij} = \beta_{0A} + \beta_{1A}t_j + \varepsilon_{ij}, i = 1, \ldots, m, j = 1, \ldots, n, \)

Group B: \( y_{ij} = \beta_{0B} + \beta_{1B}t_j + \varepsilon_{ij}, i = 1, \ldots, m, j = 1, \ldots, n, \)

\( m = \# \text{ of subjects}, \ n = \# \text{ of observations/subject}, \ \varepsilon_{ij} \text{ are normal errors with mean zero, variance } \sigma^2 \text{ and correlation } \rho. \)

We are interested in testing

\[ H_0 : \beta_{1A} = \beta_{1B} \quad \text{vs} \quad H_A : \beta_{1A} \neq \beta_{1B} \]

at level \( \alpha \) with power \( \gamma \) to detect difference \( \Delta = (\beta_{1A} - \beta_{1B})/\sigma. \) The quantities \( m \) and \( n \) have to satisfy

\[
m = \frac{2(1 - \rho)(z_{\alpha/2} + z_{1-\gamma})^2}{n \Delta^2 s_t^2}, \quad s_t^2 = \frac{\sum_{j=1}^n (t_j - \bar{t})^2}{n}.
\]
Comments:

1. For fixed time points $t_j$, larger $\rho$ gives smaller sample size $m$.
2. If $\rho = 1$, one subject from each group is enough.
3. $\rho = 0$ will require maximum sample size $m$.
4. If correlation is low, use more subjects and less obs/subject; if correlation is high, use less subjects and more obs/subject.
5. The sample size formula depends on information on $\sigma^2$ and $\rho$ and the placement of time points $t_j$’s.
6. One can choose a combination of $m$ and $n$ to meet one’s specific needs.
7. The above formula is for two-sided test.
• An example: If \( n = 3, \alpha = 0.05, \gamma = 0.8, t = (0, 2, 5) \) so \( s_t^2 = 4.222 \), then the number of subjects \((m)\) per group is

\[
m = \frac{2(1 - \rho)(1.96 + 0.84)^2}{3 \times 4.222 \Delta^2}.
\]

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
<th>0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>2479</td>
<td>1102</td>
<td>620</td>
<td>397</td>
<td>276</td>
<td>203</td>
<td>155</td>
<td>123</td>
<td>100</td>
</tr>
<tr>
<td>0.3</td>
<td>2169</td>
<td>964</td>
<td>543</td>
<td>348</td>
<td>241</td>
<td>178</td>
<td>136</td>
<td>108</td>
<td>87</td>
</tr>
<tr>
<td>0.4</td>
<td>1859</td>
<td>827</td>
<td>465</td>
<td>298</td>
<td>207</td>
<td>152</td>
<td>117</td>
<td>92</td>
<td>75</td>
</tr>
<tr>
<td>0.5</td>
<td>1550</td>
<td>689</td>
<td>388</td>
<td>248</td>
<td>173</td>
<td>127</td>
<td>97</td>
<td>77</td>
<td>62</td>
</tr>
<tr>
<td>0.6</td>
<td>1240</td>
<td>551</td>
<td>310</td>
<td>199</td>
<td>138</td>
<td>102</td>
<td>78</td>
<td>62</td>
<td>50</td>
</tr>
<tr>
<td>0.7</td>
<td>930</td>
<td>414</td>
<td>233</td>
<td>149</td>
<td>104</td>
<td>76</td>
<td>59</td>
<td>46</td>
<td>38</td>
</tr>
<tr>
<td>0.8</td>
<td>620</td>
<td>276</td>
<td>155</td>
<td>100</td>
<td>69</td>
<td>51</td>
<td>39</td>
<td>31</td>
<td>25</td>
</tr>
</tbody>
</table>
4 Modeling discrete longitudinal data

- Generalized estimating equations (GEEs)
  1. Why GEEs?
  2. Key features of GEEs
  3. Some popular GEE models
  4. Some basics of GEEs
  5. Interpretation of GEEs
  6. Analyze infectious disease data using GEE
  7. Analyze epileptic data using GEE

- Generalized linear mixed models (GLMMs)
  1. Model specification & implementation
  2. Analyze infectious disease data using a GLMM
  3. Analyze epileptic data using a GLMM
4.1 Generalized estimating equations (GEEs) for continuous and discrete longitudinal data

4.1.1 Why GEEs?

• Recall that a linear mixed model for longitudinal data may take the form:

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}. \]

• Key features:
  1. Outcome \( y_{ij} \) is continuous and normally distributed.
  2. Correlation in outcome observations from the same individuals is directly modeled using random effects (e.g., random intercept and slope).

• However,
  1. in many biomedical studies, the outcome variables are discrete (not
continuous). For example, the outcome is binary (yes/no) in Indonesian children study, and the outcome is count in the Epileptic clinical trial.

2. sometimes, we are mainly interested in the covariate effects, not in correlation among the outcome observations from the same subject. A partial reason is that it is much harder to know how the discrete observations are correlated to each other over time than continuous outcomes.

3. we might also want to model the correlation in a natural way jointly with the estimation of covariate effects of interest.
What is wrong with the classical regression approach such as the logistic regression for binary outcomes?

- Classical logistic regression model:

\[ y_i \sim \text{Binomial}(1, \pi_i(x_i)), \quad y_i = 1/0, \quad \pi_i(x_i) = \mathbb{E}[y_i|x_i], \]

\[ \text{logit}\{\pi_i(x_i)\} = \log\left\{ \frac{\pi_i(x_i)}{1 - \pi_i(x_i)} \right\} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}. \]

- Key features:
  1. Each subject contributes only one binary observation.
  2. It is reasonable to assume that the outcomes from different subjects are independent.

- However, in a longitudinal study,
  1. Each subject has multiple binary (1/0) responses over time.
  2. The subjects with higher probability to get disease will tend to have more 1’s, resulting a correlation.
  3. Even though a classical regression by ignoring correlation will give
us correct and meaningful regression coefficient estimates, their SEs are often too small, resulting invalid inference.

4. The correlation has to be taken into account for valid inference (to get correct standard errors of the regression coefficient estimates).

- **Generalized estimating equations (GEEs)** is an approach that allows us to make valid inference by implicitly taken into account the correlation.
4.1.2 Key features of GEEs for analyzing longitudinal data

1. We only need to correctly specify how the mean of the outcome variable is related to the covariates of interest. For example, for the infection disease study,

\[
y_{ij} \sim \text{Binomial}\{1, \pi_{ij}(x_{ij})\},
\]

\[
\logit\{\pi_{ij}(x_{ij})\} = \beta_0 + \beta_1 X_{ij} + \beta_2 \text{sin}v_{ij} + \beta_3 \text{age}_i + \beta_4 \text{time}_{ij} + \beta_5 \text{sex}_i + \beta_6 \text{height}_{ij},
\]

\[
\pi_{ij}(x_{ij}) = P[y_{ij} = 1|x_{ij}] = \text{E}(y_{ij}|x_{ij}) \text{ is the population probability of respiratory infection for the population defined by the specific covariate values (i.e., } x_{ij}).
\]

2. The correlation among the observations from the same subject over time is not the major interest and is treated as nuisance.

3. We can specify a correlation structure. The validity of the inference does not depend on the whether or not the specification of the
correlation structure is correct. GEE will give us a robust inference on the regression coefficients, which is valid regardless whether or not the correlation structure we specified is right.

4. GEE calculates correct SEs for the regression coefficient estimates using \textit{sandwich} estimates that will take into account the possibility that the correlation structure is misspecified.

5. The regression coefficients in GEE have a population-average interpretation.

6. A fundamental assumption on missing data is that missing data mechanism has to be MCAR, while a likelihood-based approach only requires MAR. The GEE approach will also be less efficient than a likelihood-based approach if the likelihood can be correctly specified.
4.1.3 Some popular GEE Models

- Continuous (Normal):

\[ \mu(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \]

where \( \mu(x) = \mathbb{E}(y|x) \) is the mean of outcome variable at \( x = (x_1, \ldots, x_p) \), such as mean of cholesterol level.

- Proportion (Binomial, Binary):

\[ \logit\{\pi(x)\} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \]

\( \pi(x) = P[y = 1|x] = \mathbb{E}(y|x) \) such as disease risk.

\[ \logit(\pi) = \log\{\pi/(1 - \pi)\} \] is the logit link function. Other link functions are possible.
• Count or rate (Poisson-type)

\[
\log\{\lambda(x)\} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,
\]

\(\lambda(x)\) is the rate (e.g. \(\lambda(x)\) is the incidence rate of a disease) for the count data (number of events) \(y\) over a (time, space) region \(T\) such that

\[
y|x \sim \text{Poisson}\{\lambda(x)T\}
\]

Here \(\log(.)\) link is used. Other link functions are possible.

**Note:** For count data, we have to be concerned about the possible over-dispersion in the data. That is

\[
\text{var}(y|x) > \text{E}(y|x).
\]

One way to model this phenomenon is to use an over-dispersion parameter \(\phi\) and model the variance-mean relationship as follows:

\[
\text{var}(y|x) = \phi \text{E}(y|x).
\]
### 4.1.4 Some basics of GEEs

- **Data:** $y_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$ with mean

  $$\mu_{ij} = \mathbb{E}(y_{ij} | x_{ij}).$$

  Denote

  $$y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix}, \quad \mu_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{in_i} \end{pmatrix}.$$

- **Suppose we correctly specify the mean structure for data** $y_{ij}$:

  $$g(\mu_{ij}) = \beta_0 + x_{1ij} \beta_1 + \ldots + x_{pij} \beta_p,$$

  where $g(\mu)$ is the link function such as the logit function for binary response and the log link for count data.
A GEE solves the following generalized estimating equation for $\beta$ (Liang and Zeger, 1986):

$$S_\beta(\alpha, \beta) = \sum_{i=1}^{m} \left( \frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} (y_i - \mu_i) = 0,$$

(4.1)

where $V_i$ is some matrix (intended to specify for $\text{var}(y_i|x_i)$) and $\alpha$ is the possible parameters in the correlation structure.

The above estimating equation is unbiased no matter what matrix $V_i$ we use as long as the mean structure is right. That is

$$\mathbb{E}[S_\beta(\alpha, \beta)] = 0.$$

Under some regularity conditions, the solution $\hat{\beta}$ from the GEE equation (4.1) has asymptotic distribution

$$\hat{\beta} \overset{a}{\sim} \mathcal{N}(\beta, \Sigma),$$
where

\[
\Sigma = I_0^{-1} I_1 I_0^{-1},
\]

\[
I_0 = \sum_{i=1}^{m} D_i^T V_i^{-1} D_i
\]

\[
I_1 = \sum_{i=1}^{m} D_i^T V_i^{-1} \text{var}(y_i|x_i)V_i^{-1} D_i,
\]

\[
= \sum_{i=1}^{m} D_i^T V_i^{-1} (y_i - \mu_i(\hat{\beta}))(y_i - \mu_i(\hat{\beta}))^T V_i^{-1} D_i.
\]

\(\Sigma\) is called the empirical, robust or sandwich variance estimate.

- If \(V_i\) is correctly specified, then \(I_1 \approx I_0\) and \(\Sigma \approx I_0^{-1}\) (model based). In this case, \(\hat{\beta}\) is the most efficient estimate. Otherwise, \(\Sigma \neq I_0^{-1}\).
• $V_i$, the working variance matrix for $y_i$ (at $x_i$), can be decomposed as

$$V_i = A_i^{1/2} R_i A_i^{1/2},$$

where

$$A_i = \begin{pmatrix}
\text{var}(y_{i1}|x_{i1}) & 0 & \ldots & 0 \\
0 & \text{var}(y_{i2}|x_{i2}) & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & \ldots & 0 & \text{var}(y_{in_i}|x_{in_i})
\end{pmatrix},$$

and $R_i$ is the correlation structure.

• We may try to specify $R_i$ so that it is close to the “true”. This $R_i$ is called the working correlation matrix and may be mis-specified.
Some working correlation structures

1. **Independent**: \( R_i(\alpha) = I_{n_i \times n_i} \). No \( \alpha \) needs to be estimated.

2. **Exchangeable (compound symmetric)**:

\[
R_i = \begin{bmatrix}
1 & \alpha & \cdots & \alpha \\
\alpha & 1 & \cdots & \alpha \\
\vdots & \vdots & \ddots & \vdots \\
\alpha & \alpha & \cdots & 1
\end{bmatrix}
\]

Let \( e_{ij} = y_{ij} - \hat{\mu}_{ij} \). Since \( E(e_{ij}e_{ik}) = \phi \alpha \) (at true \( \beta \)), \[\Rightarrow\]

\[
\hat{\alpha} = \frac{1}{(N^* - p - 1)\hat{\phi}} \sum_{i=1}^{m} \sum_{j<k} e_{ij}e_{ik},
\]

where \( N^* = \sum_{i=1}^{m} n_i(n_i - 1)/2 \) (total \# of pairs), \( \phi \) is usually estimated using the Pearson \( \chi^2 \).
3. **AR(1):**

\[
R_i = \begin{bmatrix}
1 & \alpha & \cdots & \alpha^{n_i-1} \\
\alpha & 1 & \cdots & \alpha^{n_i-2} \\
\vdots & \vdots & \ddots & \vdots \\
\alpha^{n_i-1} & \alpha^{n_i-2} & \cdots & 1
\end{bmatrix}.
\]

Since \( E(e_{ij}e_{i,j+1}) = \phi \alpha \) (at true \( \beta \)), \( \implies \)

\[\hat{\alpha} = \frac{1}{(N^{**} - p - 1)\hat{\phi}} \sum_{i=1}^{m} \sum_{j=1}^{n_i-1} e_{ij}e_{i,j+1},\]

where \( N^{**} = \sum_{i=1}^{m} (n_i - 1) \) (total \# of adjacent pairs).

4. Many more can be found in SAS.

- **Software:** Proc Genmod in SAS.
4.1.5 Interpretation of regression coefficients in a GEE Model

- A classical logistic model: \( y = \text{indicator of lung cancer} \sim \text{Bin}(1, \pi) \)

\[
\text{logit}(\pi) = \beta_0 + \beta_1 X_E + \beta_2 X_C,
\]

where

\[
X_E = \begin{cases} 
1 & \text{exposure} = \text{yes}, \\
0 & \text{exposure} = \text{no}, 
\end{cases} \quad X_C = \begin{cases} 
1 & \text{confounder} = \text{yes}, \\
0 & \text{confounder} = \text{no}. 
\end{cases}
\]

For example, \( X_E = \text{smoking (yes/no)} \), \( X_C = \text{Age (> 50 vs. \leq 50)} \).
Then

\[
\beta_1 = \text{age-adjusted log(OR) (\approx log(RR)) of lung cancer comparing the population of smokers and the population of non-smokers.}
\]
• In general, $\beta_k$ in a logistic regression can be interpreted as

$$\beta_k = \log(\text{OR}) \text{ of disease under consideration for two populations with covariate values } x_k + 1 \text{ and } x_k \text{ while other covariates are held fixed.}$$

• The regression coefficients in a GEE logistic model have the same population-averaged interpretation as those in a classical logistic model.

• GEE combines information from a sample of subjects to estimate these population-averaged estimates. These will be contrasted with subject-specific regression coefficients later.
4.1.6 Analyze Infectious disease data using GEE

- **Data:**
  - 275 Indonesian preschool children.
  - Each was followed over 6 consecutive quarters.
  - Outcome = respiratory infection (yes/no)
  - Covariates: Xero (xerophthalmia (yes/no)), season, age, sex, height (height for age)

- **GEE logistic model:** \( y_{ij} \) (1/0) = infection indicator \( \sim \) Bin(1, \( \pi_{ij} \)),

\[
\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 \text{Xero}_{ij} + \beta_2 \text{sinv}_{ij} + \beta_3 \text{age}_i \\
+ \beta_4 \text{time}_{ij} + \beta_5 \text{sex}_i + \beta_6 \text{height}_{ij}.
\]

See the SAS program `indon_gee.sas` and its output `indon_gee.lst` for details.
SAS program: indon_gee.sas

options ls=72 ps=72;
/*------------------------------------------------------*/
/*
/* Proc Genmod to fit population average (marginal) */
/* model using GEE approach for the Indonesia children */
/* infection disease data */
/*
/*------------------------------------------------------*/

data indon;
   infile "indon.dat";
   input id infect intercept age xero cosv sinv sex height stunted
       visit baseage season visitsq;
   time = age-baseage;
   /* sinv = sin(season*8*atan(1)/4); */
run;

data indon; set indon;
   nobsize=_n_;
run;

title "Print the first 20 observations";
proc print data=indon (obs=20);
   var id infect xero sex visit season;
run;

title "Model 1: Use exchangeable working correlation"; 
proc genmod descending;
   class id;
   model infect = xero sinv baseage time sex height
      / dist=bin link=logit;
   repeated subject=id / type=exch corrw;
run;
SAS first uses indep. corr to obtain reg parm est:

### Analysis Of Initial Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald 95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-2.4708</td>
<td>0.2122</td>
<td>-2.8867 -2.0549</td>
<td>135.59</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>xero</td>
<td>1</td>
<td>0.6661</td>
<td>0.4343</td>
<td>-0.1852 1.5174</td>
<td>2.35</td>
<td>0.1251</td>
</tr>
<tr>
<td>sinv</td>
<td>1</td>
<td>-0.0640</td>
<td>0.1517</td>
<td>-0.3613 0.2333</td>
<td>0.18</td>
<td>0.6730</td>
</tr>
<tr>
<td>baseage</td>
<td>1</td>
<td>-0.0335</td>
<td>0.0064</td>
<td>-0.0461 -0.0209</td>
<td>27.22</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>time</td>
<td>1</td>
<td>0.0000</td>
<td>0.0200</td>
<td>-0.0392 0.0392</td>
<td>0.00</td>
<td>0.9999</td>
</tr>
<tr>
<td>sex</td>
<td>1</td>
<td>-0.3841</td>
<td>0.2188</td>
<td>-0.8129 0.0448</td>
<td>3.08</td>
<td>0.0792</td>
</tr>
<tr>
<td>height</td>
<td>1</td>
<td>-0.0476</td>
<td>0.0207</td>
<td>-0.0882 -0.0070</td>
<td>5.29</td>
<td>0.0215</td>
</tr>
</tbody>
</table>

### Actual SAS output from the above program:

Model 1: Use exchangeable working correlation

The GENMOD Procedure

Model Information

<table>
<thead>
<tr>
<th>Data Set</th>
<th>WORK.INDON</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution</td>
<td>Binomial</td>
</tr>
<tr>
<td>Link Function</td>
<td>Logit</td>
</tr>
<tr>
<td>Dependent Variable</td>
<td>infect</td>
</tr>
</tbody>
</table>

Number of Observations Read: 1200
Number of Observations Used: 1200
Number of Events: 107
Number of Trials: 1200
Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
</table>
| id    | 275    | 121013 121113 121114 121140 121215 121315 121316 121413 121414 121514 121616 121617 121713 121813 121913 122013 122113 122114 122170 122214 122315 124150 124160 125140 127130 128130 129130 129140 131015 131016 131116 131140 131215 131216 131315 131417 ...

Response Profile

<table>
<thead>
<tr>
<th>Ordered Value</th>
<th>infect</th>
<th>Total Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>107</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1093</td>
</tr>
</tbody>
</table>

PROC GENMOD is modeling the probability that infect='1'.

Parameter Information

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prm1</td>
<td>Intercept</td>
</tr>
<tr>
<td>Prm2</td>
<td>xero</td>
</tr>
<tr>
<td>Prm3</td>
<td>season</td>
</tr>
<tr>
<td>Prm4</td>
<td>baseage</td>
</tr>
<tr>
<td>Prm5</td>
<td>time</td>
</tr>
<tr>
<td>Prm6</td>
<td>sex</td>
</tr>
<tr>
<td>Prm7</td>
<td>height</td>
</tr>
</tbody>
</table>

Algorithm converged.
GEE Model Information

Correlation Structure: Exchangeable
Subject Effect: id (275 levels)
Number of Clusters: 275
Correlation Matrix Dimension: 6
Maximum Cluster Size: 6
Minimum Cluster Size: 1

Algorithm converged.

Working Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>Col1</th>
<th>Col2</th>
<th>Col3</th>
<th>Col4</th>
<th>Col5</th>
<th>Col6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row1</td>
<td>1.0000</td>
<td>0.0460</td>
<td>0.0460</td>
<td>0.0460</td>
<td>0.0460</td>
<td>0.0460</td>
</tr>
<tr>
<td>Row2</td>
<td>0.0460</td>
<td>1.0000</td>
<td>0.0460</td>
<td>0.0460</td>
<td>0.0460</td>
<td>0.0460</td>
</tr>
<tr>
<td>Row3</td>
<td>0.0460</td>
<td>0.0460</td>
<td>1.0000</td>
<td>0.0460</td>
<td>0.0460</td>
<td>0.0460</td>
</tr>
<tr>
<td>Row4</td>
<td>0.0460</td>
<td>0.0460</td>
<td>0.0460</td>
<td>1.0000</td>
<td>0.0460</td>
<td>0.0460</td>
</tr>
<tr>
<td>Row5</td>
<td>0.0460</td>
<td>0.0460</td>
<td>0.0460</td>
<td>0.0460</td>
<td>1.0000</td>
<td>0.0460</td>
</tr>
<tr>
<td>Row6</td>
<td>0.0460</td>
<td>0.0460</td>
<td>0.0460</td>
<td>0.0460</td>
<td>0.0460</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Exchangeable Working Correlation

Correlation: 0.0459822403

GEE Fit Criteria

QIC: 700.6732
QICu: 699.4715
### Analysis Of GEE Parameter Estimates

**Empirical Standard Error Estimates**

| Parameter | Estimate | Standard Error | 95% Confidence Limits | Z  | Pr > |Z| |
|-----------|----------|---------------|---------------------|----|------|---|
| Intercept | -2.4592  | 0.2250        | -2.9002 -2.0181     | -10.93 | <.0001 |
| xero      | 0.5529   | 0.4479        | -0.3249 1.4308      | 1.23 | 0.2170 |
| sinv      | -0.0575  | 0.1146        | -0.2822 0.1672      | -0.50 | 0.6158 |
| baseage   | -0.0339  | 0.0062        | -0.0460 -0.0218     | -5.50 | <.0001 |
| time      | 0.0010   | 0.0210        | -0.0401 0.0421      | 0.05 | 0.9614 |
| sex       | -0.4000  | 0.2374        | -0.8654 0.0653      | -1.68 | 0.0920 |
| height    | -0.0509  | 0.0262        | -0.1024 0.0005      | -1.94 | 0.0524 |
Some remarks:

- **Proc Genmod** in SAS fits the model using independence correlation structure to get initial parameter estimate and get the estimate of over-dispersion parameter (SAS does not output the initial estimates now). We should read the output under “Analysis Of GEE Parameter Estimates”, which is valid even if the correlation structure we specified (it is exchangeable here) may not be true.

- Given other characteristics, the odds-ratio of getting respiratory infection between two populations with or without Vitamin A deficiency is estimated to be $e^{0.5529} = 1.74$. If respiratory infection could be viewed as a rare disease, kids with Vitamin A deficiency would be 74% more likely to develop respiratory infection. However, p-value=0.2170 indicates that there is no significant difference in infection risk for these two populations.
4.1.7 Analyze epileptic seizure count data using GEE

- Data:
  - 59 patients, 28 in control group, 31 in treatment (progabide) group.
  - 5 seizure counts (including baseline) were obtained.
  - Covariates: treatment (covariate of interest), age.

- GEE Poisson model: \( y_{ij} \) = seizure counts obtained at the \( j \)th \((j = 1, 2, ..., 5)\) time point for patient \( i \), \( y_{ij} \sim \) over-dispersed Poisson(\( \mu_{ij} \)), \( \mu_{ij} = E(y_{ij}) = t_{ij} \lambda_{ij} \), where \( t_{ij} \) is the length of the time window from which the seizure count \( y_{ij} \) was observed, \( \lambda_{ij} \) is hence the rate to have a seizure. First consider model

\[
\begin{align*}
\log(\lambda_{ij}) &= \beta_0 + \beta_1 I(j > 1) + \beta_2 \text{trt}_i + \beta_3 \text{trt}_i I(j > 1) \\
\log(\mu_{ij}) &= \log(t_{ij}) + \beta_0 + \beta_1 I(j > 1) + \beta_2 \text{trt}_i + \beta_3 \text{trt}_i I(j > 1)
\end{align*}
\]

Note that \( \log(t_{ij}) \) is often called an offset.
• Interpretation of $\beta$’s:

<table>
<thead>
<tr>
<th>Group</th>
<th>log of seizure rate $\lambda$ Before RAND ($j = 1$)</th>
<th>After RAND ($j &gt; 1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control (trt=0)</td>
<td>$\beta_0$</td>
<td>$\beta_0 + \beta_1$</td>
</tr>
<tr>
<td>Treatment (trt=1)</td>
<td>$\beta_0 + \beta_2$</td>
<td>$\beta_0 + \beta_1 + \beta_2 + \beta_3$</td>
</tr>
</tbody>
</table>

Therefore, $\beta_1 =$ time and placebo effect, $\beta_2 =$ difference in seizure rates at baseline between two groups, $\beta_3 =$ treatment effect of interest (accounting for time & placebo effect).

If randomization is taken into account ($\beta_2 = 0$), we can consider the following model

$$\log(\mu_{ij}) = \log(t_{ij}) + \beta_0 + \beta_1 I(j > 1) + \beta_2 trt_i I(j > 1).$$

• See the SAS program `seize_gee.sas` and its output `seize_gee.lst` for details.
First part of seize.gee.sas

options ls=80 ps=1000 nodate;

/*------------------------------------------------------*/
/*                      */
/* Proc Genmod to fit population average (marginal) */
/* model using GEE approach for the epileptic seizure */
/* count data */
/*                      */
/*------------------------------------------------------*/

data seizure;
  infile "seize.dat";
  input id seize visit trt age;
  nobs=_n_;  
  interval = 2;
  if visit=0 then interval=8;
  logtime = log(interval);
  assign = (visit>0);
run;

title "Model 1: overall effect of the treatment";
proc genmod data=seizure;
  class id;
  model seize = assign trt assign*trt
    / dist=poisson link=log offset=logtime;
  repeated subject=id / type=exch corrw;
run;
SAS first uses indep corr to get reg. coeff est.

Analysis Of Initial Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald 95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>1.3476</td>
<td>0.0341</td>
<td>1.2809 1.4144</td>
<td>1565.44</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>assign</td>
<td>1</td>
<td>0.1108</td>
<td>0.0469</td>
<td>0.0189 0.2027</td>
<td>5.58</td>
<td>0.0181</td>
</tr>
<tr>
<td>trt</td>
<td>1</td>
<td>0.0265</td>
<td>0.0467</td>
<td>-0.0650 0.1180</td>
<td>0.32</td>
<td>0.5702</td>
</tr>
<tr>
<td>assign*trt</td>
<td>1</td>
<td>-0.1037</td>
<td>0.0651</td>
<td>-0.2312 0.0238</td>
<td>2.54</td>
<td>0.1110</td>
</tr>
</tbody>
</table>

Actual SAS output from the above program:

Model 1: overall effect of the treatment

The GENMOD Procedure

Model Information

Data Set WORK.SEIZURE
Distribution Poisson
Link Function Log
Dependent Variable seize
Offset Variable logtime

Number of Observations Read 295
Number of Observations Used 295

Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>59</td>
<td>101 102 103 104 106 107 108 110 111 112 113 114 116 117 118 121 122 123 124 126 128 129 130 135</td>
</tr>
</tbody>
</table>
Parameter Information

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prm1</td>
<td>Intercept</td>
</tr>
<tr>
<td>Prm2</td>
<td>assign</td>
</tr>
<tr>
<td>Prm3</td>
<td>trt</td>
</tr>
<tr>
<td>Prm4</td>
<td>assign*trt</td>
</tr>
</tbody>
</table>

Algorithm converged.

GEE Model Information

Correlation Structure   Exchangeable
Subject Effect         id (59 levels)
Number of Clusters      59
Correlation Matrix Dimension  5
Maximum Cluster Size    5
Minimum Cluster Size    5

Algorithm converged.

Working Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>Col1</th>
<th>Col2</th>
<th>Col3</th>
<th>Col4</th>
<th>Col5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row1</td>
<td>1.0000</td>
<td>0.7716</td>
<td>0.7716</td>
<td>0.7716</td>
<td>0.7716</td>
</tr>
<tr>
<td>Row2</td>
<td>0.7716</td>
<td>1.0000</td>
<td>0.7716</td>
<td>0.7716</td>
<td>0.7716</td>
</tr>
<tr>
<td>Row3</td>
<td>0.7716</td>
<td>0.7716</td>
<td>1.0000</td>
<td>0.7716</td>
<td>0.7716</td>
</tr>
<tr>
<td>Row4</td>
<td>0.7716</td>
<td>0.7716</td>
<td>0.7716</td>
<td>1.0000</td>
<td>0.7716</td>
</tr>
<tr>
<td>Row5</td>
<td>0.7716</td>
<td>0.7716</td>
<td>0.7716</td>
<td>0.7716</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Exchangeable Working Correlation

Correlation 0.7715879669

GEE Fit Criteria

QIC -659.5869
QICu -668.6570

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

| Parameter     | Estimate | Standard Error | 95% Confidence Limits | Z  | Pr > |Z| |
|---------------|----------|----------------|-----------------------|----|------|---|
| Intercept     | 1.3476   | 0.1574         | 1.0392 1.6560         | 8.56 | <.0001 |
| assign        | 0.1108   | 0.1161         | -0.1168 0.3383        | 0.95 | 0.3399 |
| trt           | 0.0265   | 0.2219         | -0.4083 0.4613        | 0.12 | 0.9049 |
| assign*trt    | -0.1037  | 0.2136         | -0.5223 0.3150        | -0.49 | 0.6274 |
Second part of seize.gge.sas

title "Model 2: take randomization into account";
proc genmod data=seizure;
   class id;
   model seize = assign assign*trt
      / dist=poisson link=log offset=logtime scale=pearson aggregate=nobs;
   repeated subject=id / type=exch corrw;
run;

Output from the above program:

Model 2: take randomization into account

Algorithm converged.

Working Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>Col1</th>
<th>Col2</th>
<th>Col3</th>
<th>Col4</th>
<th>Col5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row1</td>
<td>1.0000</td>
<td>0.7750</td>
<td>0.7750</td>
<td>0.7750</td>
<td>0.7750</td>
</tr>
<tr>
<td>Row2</td>
<td>0.7750</td>
<td>1.0000</td>
<td>0.7750</td>
<td>0.7750</td>
<td>0.7750</td>
</tr>
<tr>
<td>Row3</td>
<td>0.7750</td>
<td>0.7750</td>
<td>1.0000</td>
<td>0.7750</td>
<td>0.7750</td>
</tr>
<tr>
<td>Row4</td>
<td>0.7750</td>
<td>0.7750</td>
<td>0.7750</td>
<td>1.0000</td>
<td>0.7750</td>
</tr>
<tr>
<td>Row5</td>
<td>0.7750</td>
<td>0.7750</td>
<td>0.7750</td>
<td>0.7750</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Exchangeable Working Correlation

Correlation 0.7749556223

GEE Fit Criteria

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>QIC</td>
<td>-667.9355</td>
</tr>
<tr>
<td>QICu</td>
<td>-672.7671</td>
</tr>
</tbody>
</table>

Graduate Summer Session in Epidemiology Slide 168
### Analysis Of GEE Parameter Estimates

#### Empirical Standard Error Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>95% Confidence Limits</th>
<th>Z</th>
<th>Pr &gt;</th>
<th>Z</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.3616</td>
<td>0.1111</td>
<td>1.1438</td>
<td>1.5794</td>
<td>12.25</td>
<td>&lt; .0001</td>
<td></td>
</tr>
<tr>
<td>assign</td>
<td>0.1173</td>
<td>0.1283</td>
<td>-0.1341</td>
<td>0.3688</td>
<td>0.91</td>
<td>0.3604</td>
<td></td>
</tr>
<tr>
<td>assign*trt</td>
<td>-0.1170</td>
<td>0.2076</td>
<td>-0.5240</td>
<td>0.2900</td>
<td>-0.56</td>
<td>0.5731</td>
<td></td>
</tr>
</tbody>
</table>
A program to adjust for age

title "Model 3: adjusting for other covariates (age)";
proc genmod data=seizure;
   class id;
   model seize = assign trt assign*trt age
       / dist=poisson link=log offset=logtime scale=pearson;
   repeated subject=id / type=exch corrw;
run;

Output of the program to adjust for all covariates
Model 3: adjusting for other covariates

Algorithm converged.

Working Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>Col1</th>
<th>Col2</th>
<th>Col3</th>
<th>Col4</th>
<th>Col5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row1</td>
<td>1.0000</td>
<td>0.7620</td>
<td>0.7620</td>
<td>0.7620</td>
<td>0.7620</td>
</tr>
<tr>
<td>Row2</td>
<td>0.7620</td>
<td>1.0000</td>
<td>0.7620</td>
<td>0.7620</td>
<td>0.7620</td>
</tr>
<tr>
<td>Row3</td>
<td>0.7620</td>
<td>0.7620</td>
<td>1.0000</td>
<td>0.7620</td>
<td>0.7620</td>
</tr>
<tr>
<td>Row4</td>
<td>0.7620</td>
<td>0.7620</td>
<td>0.7620</td>
<td>1.0000</td>
<td>0.7620</td>
</tr>
<tr>
<td>Row5</td>
<td>0.7620</td>
<td>0.7620</td>
<td>0.7620</td>
<td>0.7620</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Exchangeable Working Correlation

Correlation 0.7620401547

GEE Fit Criteria

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>QIC</td>
<td>-686.4053</td>
</tr>
<tr>
<td>QICu</td>
<td>-697.2349</td>
</tr>
</tbody>
</table>
### Analysis Of GEE Parameter Estimates
#### Empirical Standard Error Estimates

| Parameter  | Estimate | Standard Error | 95% Confidence Limits | Z  | Pr > |Z| |
|------------|----------|----------------|-----------------------|----|------|---|
| Intercept  | 2.2601   | 0.4330         | 1.4113                | 3.1088 | 5.22 | <.0001 |
| assign     | 0.1108   | 0.1161         | -0.0168               | 0.3383 | 0.95 | 0.3399 |
| trt        | -0.0175  | 0.2141         | -0.4371               | 0.4020 | -0.08 | 0.9348 |
| assign*trt | -0.1037  | 0.2136         | -0.5223               | 0.3150 | -0.49 | 0.6274 |
| age        | -0.0321  | 0.0147         | -0.0610               | -0.0032 | -2.17 | 0.0296 |
4.2 Generalized linear mixed models (GLMMs)

4.2.1 Model specification and implementation

- Generalized linear mixed models (GLMMs) are an extension of
  1. linear mixed models (continuous ⇒ discrete)
  2. logistic (Poisson) models (independent discrete data ⇒ discrete longitudinal data)

- Consider a special linear mixed model:

\[
y_{ij} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + b_i + \varepsilon_{ij},
\]

where \( b_i \sim N(0, \sigma_b^2) \) and \( \varepsilon_{ij} \sim iid N(0, \sigma_\varepsilon^2) \).

Let \( \mu_{ij}^b = E[y_{ij} \mid b_i] \). Then the above model is equivalent to

\[
y_{ij} \mid b_i, x_i \sim iid N(\mu_{ij}^b, \sigma_\varepsilon^2),
\]

\[
\mu_{ij}^b = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + b_i.
\]
• Extend above model (4.2) to logistic model for longitudinal binary data:

\[ y_{ij} | b_i, x_i \overset{i.i.d.}{\sim} \text{Binomial}\left\{1, \pi_{ij}^b(x_i)\right\}, \]

\[
\text{logit}\left\{\pi_{ij}^b(x_i)\right\} = \log\left\{\frac{\pi_{ij}^b(x_i)}{1-\pi_{ij}^b(x_i)}\right\} \\
= \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + b_i,
\]

\[ b_i \sim N(0, \sigma_b^2), \]

where \( b_i \) are the (normal) subject-specific random effects. This is a special GLMM (logistic-normal).

• Remarks:

1. In this model the correlation is modeled through random effects \( b_i \). A subject with higher \( b_i \) will have higher disease probability \( \pi_{ij}^b \) (if other covariate values are kept the same).

2. Random effects \( b_i \) vary from subject to subject and are assumed to be independent. Hence the data \( \{y_{ij}\} \) from the same individuals
are correlated.

3. The random effects $b_i$ are usually assumed to have a normal distribution $N(0, \sigma^2_b)$. The variance $\sigma^2_b$ measures the between-subject variation, and also measures the strength of the correlation. If $\sigma^2_b = 0$, no correlation. When $\sigma^2_b$ increases, the correlation increases.

4. The success probability $\pi_{ij}^b$ is subject-specific, so the parameters $\beta$’s in (4.3) have a subject-specific interpretation (more detail in the infectious disease example).

5. For given $x$, $\pi(x)$ (the success probability for the population with covariate $x$) can be obtained through

$$
\pi(x) = \mathbb{E}[\pi^b(x)] = \int \pi^b(x)f(b)db.
$$
6. Even though \( \pi^b(x) \) has a logistic form in model (4.3), \( \pi(x) \) does NOT have a logistic form. In particular:

\[
\logit\{\pi(x)\} \neq \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.
\]

7. However, approximately we have

\[
\logit\{\pi(x)\} \approx (1 + 0.346 \sigma_b^2)^{-1/2} \times (\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p). \tag{4.4}
\]

That is, \((1 + 0.346 \sigma_b^2)^{-1/2} \beta_k\) has an approximate population-level interpretation in terms of log odds-ratio.
• Extend above model (4.2) to log-linear model for longitudinal Poisson (count) data:

\[
\begin{align*}
  y_{ij} | b_i \sim^\text{ind} \text{Poisson}(\mu_{ij}^b = T_{ij} \lambda_{ij}^b), \\
  \log(\lambda_{ij}^b) &= \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + b_i, \\
  \log(\mu_{ij}^b) &= \log(T_{ij}) + \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + b_i, \\
  b_i &\sim \text{N}(0, \sigma_b^2),
\end{align*}
\]

(4.5)

where \( b_i \) are the (normal) subject-specific random effects. This is a special GLMM (Poisson-normal).

• Remarks:

1. In this model, the correlation is modeled through random effects \( b_i \). A subject with higher \( b_i \) will have larger rate \( \lambda_{ij}^b \) (if other covariate values are kept the same), and tend to have larger responses.

2. Random effects \( b_i \) vary from subject to subject and are assumed to be independent. Hence the data \( \{y_{ij}\} \) from the same individuals are correlated.
3. The random effects $b_i$ are usually assumed to have a normal distribution $N(0, \sigma_b^2)$. The variance $\sigma_b^2$ measures the between-subject variation, and also measures the strength of the correlation. If $\sigma_b^2 = 0$, no correlation. When $\sigma_b^2$ increases, the correlation increases.

4. The event rate $\lambda_{ij}^b$ is subject-specific, so the parameters $\beta$'s in (4.5) have a subject-specific interpretation (more detail in the Epileptic seizure count example).

5. There still may be overdispersion for $y_{ij}\mid b_i$. That is $\text{var}(y_{ij}\mid b_i) > \text{E}(y_{ij}\mid b_i)$. So we may take the over-dispersion into account by assuming

$$\text{var}(y_{ij}\mid b_i) = \phi\text{E}(y_{ij}\mid b_i).$$

**Note:** This $\phi$ is different from the $\phi$ in GEE.
One way to account for overdispersion is to use statement `random _residual_ in Proc Glimmix.

The other way is to assume $y_{ij}|b_i$ has the following log quasi-likelihood function:

$$
\ell_q(y_{ij}, \mu_{ij}) = \frac{y_{ij}(\log \mu_{ij} - \log y_{ij}) - (\mu_{ij} - y_{ij})}{\phi} - \frac{1}{2} \log \phi.
$$

Or to assume $y_{ij}|b_i$ has a generalized Poisson distribution:

$$
f(y_{ij}|b_i) = \frac{(1 - \xi)\mu_{ij}^b \{(1 - \xi)\mu_{ij}^b + \xi y_{ij}\}y_{ij}^{-1}e^{-(1-\xi)\mu_{ij}^b - \xi y_{ij}}}{y_{ij}!}.
$$

In this case,

$$
E(y_{ij}|b_i) = \mu_{ij}^b, \quad \text{var}(y_{ij}|b_i) = \mu_{ij}^b/(1 - \xi)^2,
$$

so over-dispersion parameter is $1/(1 - \xi)^2$. 
6. For given $x$, the population event rate $\lambda(x)$ (the event rate for the population with covariate $x$) can be obtained through

$$
\lambda(x) = \mathbb{E}[\lambda^b(x)] = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + b} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p} \mathbb{E}(e^b) = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p} e^{\sigma^2_b / 2} = e^{\sigma^2_b / 2 + \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p} = e^{\tilde{\beta}_0 + \beta_1 x_1 + \cdots + \beta_p x_p},
$$

$$
\implies \log\{\lambda(x)\} = \tilde{\beta}_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \quad (4.6)
$$

therefore, the regression coefficients $\beta$’s (except $\beta_0$) in model (4.5) also have population average interpretation.
• For a liner mixed model like the following (random intercept only model)

\[ y_{ij} = \beta^T x_{ij} + b_i + \epsilon_{ij}, \]

where \( \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_\epsilon) \), we have

\[ \mathbb{E}(y_{ij}|b_i) = \beta^T x_{ij} + b_i \quad \text{and} \quad \mathbb{E}(y_{ij}) = \beta^T x_{ij}. \]

So the \( \beta \)'s (except the intercept \( \beta_0 \)) always have population-averaged interpretation as well as subject-specific interpretation.
• Why GLMMs?

1. We are interested in how the outcome variable is related to the independent variables (covariates).

2. We are also interested in how individuals’ data vary from subject to subject (between-subject variation). This can be modeled through the use of random effects. The random effects have a natural interpretation.

3. A GLMM is a likelihood-based model. So it requires much less strong assumption for missing data mechanism. Only MAR mechanism is required for a GLMM to make valid inference, compared to MCAR for GEE approach.

4. The regression coefficients have a subject-specific interpretation, and for some special GLMMs we can still (approximately) make population level inference.
• **Implementation**: Proc Glimmix for GLMMs in SAS where approximate integration is used for approximate or exact maximum quasi-likelihood/likelihood estimation. Or Proc Nlmixed (non-linear mixed model) in SAS where numerical integration is used for maximum likelihood estimation.
4.3 Analyze infectious disease data using a GLMM

- Assume infection indicator $y_{ij}$ ($1 =$ infection, $0 =$ no infection):

$$
y_{ij} | b_i \sim \text{Binomial}(1, \pi_{ij}^b),
$$

$$
\logit(\pi_{ij}^b) = \beta_0 + \beta_1 Xero_{ij} + \beta_2 \text{sinv}_{ij} + \beta_3 \text{age}_i
$$

$$
+ \beta_4 \text{time}_{ij} + \beta_5 \text{sex}_i + \beta_6 \text{height}_{ij} + b_i,
$$

where $b_i \sim N(0, \sigma^2_b)$.

- **Interpretation** of $\beta_1$ (coefficient of a time-varying covariate $Xero$):

Let $\pi_{1}^b, \pi_{0}^b$ be the infection probability for any subject $i$ (the same kid) when $Xero$ is 1 and 0 (while other covariate values are fixed). Then

$$
\logit(\pi_{1}^b) - \logit(\pi_{0}^b) = \beta_1,
$$

that is

$$
\beta_1 = \log \left( \frac{\pi_{1}^b/(1 - \pi_{1}^b)}{\pi_{0}^b/(1 - \pi_{0}^b)} \right).
$$
That is, $\beta_1$ is the log odds-ratio of getting respiratory infection if a subject becomes Vitamin A deficiency (from Vitamin A sufficiency). Similar interpretation holds for continuous time-varying covariates.

- **Interpretation** of $\beta_5$ (coefficient of a one-time covariate sex): Let $\pi_{1i}^b$ be the infection probability for subject $i$ who is a boy and $\pi_{0j}^b$ be the infection probability for subject $j$ who is a girl. Assume they have the same covariate values (except sex). Then

$$\logit(\pi_{1i}^b) - \logit(\pi_{0j}^b) = \beta_5 + (b_i - b_j).$$

If $b_i \approx b_j$, then

$$\logit(\pi_{1i}^b) - \logit(\pi_{0j}^b) \approx \beta_5,$$

$$\beta_5 \approx \log \left[ \frac{\pi_{1i}^b / (1 - \pi_{1i}^b)}{\pi_{0j}^b / (1 - \pi_{0j}^b)} \right].$$

That is, $\beta_5$ is the log odds-ratio of getting respiratory infection comparing a boy and a girl who are similar in other subject characteristics except gender. Similar interpretation holds for
continuous one-time covariates.

See the SAS program indon_mix.sas and its output indon_mix.lst for details.
Remark 1: As indicated by (4.4), \((1 + 0.346\sigma_b^2)^{-1/2}\beta\) have population log odds-ratio interpretation:

\[
\logit(\pi_{ij}) \approx (1 + 0.346\sigma_b^2)^{-1/2}\beta^T x_{ij} = \tilde{\beta}^T x_{ij},
\]

where \(\tilde{\beta} = (1 + 0.346\sigma_b^2)^{-1/2}\beta\). Therefore, \(\tilde{\beta}\) has population-average interpretation. That is, we can use \(\tilde{\beta}\) to compare two populations.
**SAS program indon_mix.sas**

```sas
options ls=80 ps=1000 nodate;
/*---------------------------------------------*/
/*  Proc Glimmix to fit subject-specific (random effect) */
/*  model for the Indonesian children infection disease */
/*  data */
/*---------------------------------------------*/

data indon;
  infile "indon.dat";
  input id infect intercept age xero cosv sinv sex height stunted visit baseage season visitsq;
  time = age - baseage;
run;

title "Random intercept model for infection disease data";
proc glimmix data=indon method=quad;
  class id;
  model infect = xero sinv age time sex height / dist=bin link=logit s;
  random int / subject=id type=vc;
run;
```
**SAS output** indon_mix.lst

Random intercept model for infection disease data

The GLIMMIX Procedure

Model Information

<table>
<thead>
<tr>
<th>Data Set</th>
<th>WORK.INDON</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Variable</td>
<td>infect</td>
</tr>
<tr>
<td>Response Distribution</td>
<td>Binomial</td>
</tr>
<tr>
<td>Link Function</td>
<td>Logit</td>
</tr>
<tr>
<td>Variance Function</td>
<td>Default</td>
</tr>
<tr>
<td>Variance Matrix Blocked By</td>
<td>id</td>
</tr>
<tr>
<td>Estimation Technique</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>Likelihood Approximation</td>
<td>Gauss-Hermite Quadrature</td>
</tr>
<tr>
<td>Degrees of Freedom Method</td>
<td>Containment</td>
</tr>
</tbody>
</table>

Class Level Information

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>275</td>
<td>121013 121113 121114 121140 121215 121315 121316 ...</td>
</tr>
</tbody>
</table>

Number of Observations Read 1200
Number of Observations Used 1200
Dimensions

G-side Cov. Parameters: 1
Columns in X: 7
Columns in Z per Subject: 1
Subjects (Blocks in V): 275
Max Obs per Subject: 6

Optimization Information

Optimization Technique: Dual Quasi-Newton
Parameters in Optimization: 8
Lower Boundaries: 1
Upper Boundaries: 0
Fixed Effects: Not Profiled
Starting From: GLM estimates
Quadrature Points: 9
<table>
<thead>
<tr>
<th>Iteration</th>
<th>Restarts</th>
<th>Evaluations</th>
<th>Objective Function</th>
<th>Change</th>
<th>Max Gradient</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>4</td>
<td>711.86096381</td>
<td>.</td>
<td>370.9219</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>706.20769988</td>
<td>5.65326394</td>
<td>299.1176</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>4</td>
<td>703.19922822</td>
<td>3.00847166</td>
<td>39.35369</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>3</td>
<td>702.16517109</td>
<td>1.03405713</td>
<td>35.20026</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>4</td>
<td>687.59198873</td>
<td>14.57318236</td>
<td>16.85617</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>4</td>
<td>685.75840614</td>
<td>1.83358259</td>
<td>6.4086</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>4</td>
<td>681.30002966</td>
<td>4.45837648</td>
<td>42.48172</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>3</td>
<td>680.48333006</td>
<td>0.81669960</td>
<td>24.84155</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>2</td>
<td>679.3625642</td>
<td>1.12076585</td>
<td>25.03619</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>3</td>
<td>678.98158225</td>
<td>0.38098196</td>
<td>9.369015</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>3</td>
<td>678.94174122</td>
<td>0.03984103</td>
<td>7.16441</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>3</td>
<td>678.92777026</td>
<td>0.01397095</td>
<td>9.372961</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>2</td>
<td>678.90767998</td>
<td>0.02009328</td>
<td>3.875721</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>3</td>
<td>678.9028667</td>
<td>0.00481028</td>
<td>0.136479</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>3</td>
<td>678.90284743</td>
<td>0.00001927</td>
<td>0.034321</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>3</td>
<td>678.90284698</td>
<td>0.00000045</td>
<td>0.009377</td>
</tr>
</tbody>
</table>

Convergence criterion (GCONV=1E-8) satisfied.

Fit Statistics

-2 Log Likelihood 678.90
AIC (smaller is better) 694.90
AICC (smaller is better) 695.02
BIC (smaller is better) 723.84
CAIC (smaller is better) 731.84
HQIC (smaller is better) 706.51
Fit Statistics

-2 Log Likelihood 678.90
AIC (smaller is better) 694.90
AICC (smaller is better) 695.02
BIC (smaller is better) 723.84
CAIC (smaller is better) 731.84
HQIC (smaller is better) 706.51

Fit Statistics for Conditional Distribution

-2 log L(infect | r. effects) 579.67
Pearson Chi-Square 880.12
Pearson Chi-Square / DF 0.73

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>id</td>
<td>0.7141</td>
<td>0.3650</td>
</tr>
</tbody>
</table>

Solutions for Fixed Effects

| Effect   | Estimate | Standard Error | DF | t Value | Pr > |t|  |
|----------|----------|----------------|----|---------|-------|------|
| Intercept| -2.7427  | 0.2667         | 273| -10.29  | <.0001|      |
| xero     | 0.5042   | 0.4860         | 920| 1.04    | 0.2998|      |
| sinv     | -0.05165 | 0.1588         | 920| -0.33   | 0.7451|      |
| age      | -0.03727 | 0.007755       | 920| -4.81   | <.0001|      |
| time     | 0.03999  | 0.02173        | 920| 1.84    | 0.0660|      |
| sex      | -0.4368  | 0.2613         | 920| -1.67   | 0.0949|      |
| height   | -0.05349 | 0.02335        | 920| -2.29   | 0.0222|      |
• **Remark 1 (subject-specific interpretation):** Since \( \hat{\beta}_1 = 0.5042 \), so if a child becomes Vitamin A deficiency from Vitamin A sufficiency, his/her odds-ratio of getting respiratory infection will be \( e^{0.5042} = 1.66 \), that is, about 66% increase in risk.

• **Remark 2 (approximate population-average interpretation):** \( \hat{\sigma}_b^2 = 0.7141 \), so \( (1 + 0.346\hat{\sigma}_b^2)^{-1/2} = 0.9 \). So the population-averaged effect of Vitamin A deficiency is \( 0.9 \times 0.5042 = 0.454 \). That is, given other covariates, the population of children with Vitamin A deficiency will be 57% (odds-ratio \( e^{0.454} = 1.57 \approx \) relative risk if respiratory infection can be viewed as a rare event) more likely to have respiratory infection than the population of children without Vitamin A deficiency.
The population-average effect of sex is \(0.9 \times (-0.4368) = -0.39\) (odds-ratio = 0.65). So boys are less likely to have respiratory infection than girls. Other effects can be obtained similarly.

- **Remark 3**: When the response \(y_{ij}\) is binary, we don't have to worry about over-dispersion for the conditional distribution of \(y_{ij} | b_i\).
4.4 Analyze epileptic count data using a GLMM

- Assume seizure counts

\[ y_{ij} | b_i \sim \text{Overdispersed - Poisson}(\mu_{ij}^b), \]

where

\[ \mu_{ij}^b = E(y_{ij} | b_i) = t_{ij} \lambda_{ij}^b, \]

\[ \lambda_{ij}^b \] is the rate to have a seizure for subject \( i \). Consider model

\[
\begin{align*}
\log(\lambda_{ij}^b) &= \beta_0 + \beta_1 I(j > 1) + \beta_2 \text{trt}_i I(j > 1) + b_i \\
\log(\mu_{ij}^b) &= \log(t_{ij}) + \beta_0 + \beta_1 I(j > 1) + \beta_2 \text{trt}_i I(j > 1) + b_i,
\end{align*}
\]

where \( b_i \sim N(0, \sigma_b^2) \) is a random intercept describing the between-subject variation.
• Interpretation of $\beta$’s:

<table>
<thead>
<tr>
<th>Group</th>
<th>$\log(\lambda^b)$ for random subject $i$</th>
<th>Before RAND ($j = 1$)</th>
<th>After RAND ($j &gt; 1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control (trt=0)</td>
<td>$\beta_0 + b_i$</td>
<td>$\beta_0$</td>
<td>$\beta_0 + \beta_1 + b_i$</td>
</tr>
<tr>
<td>Treatment (trt=1)</td>
<td>$\beta_0 + b_i$</td>
<td>$\beta_0 + \beta_1 + \beta_2 + b_i$</td>
<td></td>
</tr>
</tbody>
</table>

$\beta_1$: difference in log of rate of seizure counts comparing after randomization and before randomization for a random subject in control group (time + placebo effect).

$\beta_2$: difference in log of rate of seizure counts for a treated subject compared to if he/she received a placebo (treatment effect).

• For more details of the result, see SAS program `seize_mix.sas` and its output `seize_mix.lst`

• Remark: Since here we used the Poisson GLMM with log link and a random intercept, so the regression coefficients (except the intercept) also have a population-averaged interpretation.
SAS program seize_mix.sas

```sas
options ls=80 ps=1000 nodate;

/*------------------------------------------------------*/
/* */
/* Proc Glimmix to fit random intercept model to the */
/* epileptic seizure count data */
/* */
/* */
/*------------------------------------------------------*/

data seizure;
  infile "seize.dat";
  input id seize visit trt age;
  nobs=_n_;
  interval = 2;
  if visit=0 then interval=8;
  logtime = log(interval);
  assign = (visit>0);
  agn_trt = assign*trt;
run;

title "Random intercept model for seizure data with conditional overdispersion";
proc glimmix data=seizure;
  class id;
  model seize = assign agn_trt / dist=poisson link=log offset=logtime s;
  random int / subject=id type=vc;
  random _residual_; *for conditional overdispersion;
run;
```
Random intercept model for seizure data with conditional overdispersion

The GLIMMIX Procedure

Model Information

Data Set WORK.SEIZURE
Response Variable seize
Response Distribution Poisson
Link Function Log
Variance Function Default
Offset Variable logtime
Variance Matrix Blocked By id
Estimation Technique Residual PL
Degrees of Freedom Method Containment

Class Level Information

Class  Levels  Values
id     59        101 102 103 104 106 107 108 110 111 112 113 114 ...

Number of Observations Read 295
Number of Observations Used 295
Dimensions

G-side Cov. Parameters  1
R-side Cov. Parameters  1
Columns in X  3
Columns in Z per Subject  1
Subjects (Blocks in V)  59
Max Obs per Subject  5

Optimization Information

Optimization Technique  Dual Quasi-Newton
Parameters in Optimization  1
Lower Boundaries  1
Upper Boundaries  0
Fixed Effects  Profiled
Residual Variance  Profiled
Starting From  Data

Iteration History

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Restarts</th>
<th>Subiterations</th>
<th>Objective</th>
<th>Function Change</th>
<th>Gradient Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>4</td>
<td>609.19264304</td>
<td>0.49414053</td>
<td>0.000205</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>5</td>
<td>671.59595217</td>
<td>0.14411653</td>
<td>3.061E-6</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>3</td>
<td>675.96769701</td>
<td>0.01612221</td>
<td>0.000016</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>2</td>
<td>675.86073055</td>
<td>0.00032842</td>
<td>1.901E-8</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>675.85749753</td>
<td>0.00000336</td>
<td>3.111E-8</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>675.85746125</td>
<td>0.00000000</td>
<td>5.906E-6</td>
</tr>
</tbody>
</table>

Convergence criterion (PCONV=1.11022E-8) satisfied.
Fit Statistics

-2 Res Log Pseudo-Likelihood 675.86
Generalized Chi-Square 822.08
Gener. Chi-Square / DF 2.82

Covariance Parameter Estimates

<table>
<thead>
<tr>
<th>Cov Parm</th>
<th>Subject</th>
<th>Estimate</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>id</td>
<td>0.5704</td>
<td>0.1169</td>
</tr>
<tr>
<td>Residual (VC)</td>
<td></td>
<td>2.8154</td>
<td>0.2591</td>
</tr>
</tbody>
</table>

Solutions for Fixed Effects

| Effect      | Estimate | Standard Error | DF | t Value | Pr > |t| |
|-------------|----------|----------------|----|---------|-------|
| Intercept   | 1.0655   | 0.1079         | 58 | 9.88    | <.0001|
| assign      | 0.1122   | 0.07723        | 234| 1.45    | 0.1477|
| agn_trt     | -0.1063  | 0.1054         | 234| -1.01   | 0.3144|
• **Remark:** There is considerable amount of over-dispersion for $y_{ij} | b_i$. It is estimated that

$$\text{var}(y_{ij} | b_i) = 2.82 \text{E}(y_{ij} | b_i).$$

• There is considerable between-patient variance in log-seizure rate. That variation is estimated to be 0.57.

• The regression coefficient estimates (except the intercept) have population-average interpretation, and they are almost the same as those from the GEE model.

For example, $\hat{\beta}_2 = -0.1063$ with SE = 0.1054. Then if a subject switches from control to treatment, the rate of having seizure will decrease by 10% (since $e^{-0.1063} = 0.9$). The same rate deduction can also be used to compare treatment and control groups (i.e., population interpretation).
If we would like to fit the data using the conditional quasi-likelihood approach, we need to use Proc Nlmixed:

```plaintext
proc nlmixed qpoints=15;
  parms beta0=-1.4 beta1=0.12 beta2=-0.12 theta=0.1 phi=1;
  eta = beta0 + beta1*assign + beta2*agn_trt + b;
  mu = interval*exp(eta);
  if seize=0 then
    l = -(mu - seize)/phi - log(phi)/2;
  else
    l = (seize*(log(mu) - log(seize)) - (mu - seize))/phi - log(phi)/2;
  model seize ~ general(l);
  random b ~ normal(0, theta) subject=id;
run;
```
The relevant output is

| Parameter | Estimate | Standard Error | DF | t Value | Pr > |t| | Alpha | Lower |
|-----------|----------|----------------|----|---------|-------|-------|-------|-------|
| beta0     | 1.0350   | 0.1100         | 58 | 9.41    | <.0001| 0.05  | 0.8148|
| beta1     | 0.1123   | 0.07898        | 58 | 1.42    | 0.1603| 0.05  | -0.04577|
| beta2     | -0.1065  | 0.1077         | 58 | -0.99   | 0.3269| 0.05  | -0.3222|
| theta     | 0.5835   | 0.1204         | 58 | 4.85    | <.0001| 0.05  | 0.3426|
| phi       | 2.9456   | 0.2684         | 58 | 10.98   | <.0001| 0.05  | 2.4084|

Therefore, the between-patient variance is estimated to be 0.5835 and the conditional over-dispersion parameter estimated to be $\hat{\phi} = 2.9$. The inference on the treatment effect $\beta_2$ is similar.
• If we would like to fit a generalized Poisson distribution for the
conditional distribution, we can use the following Proc Nlmixed
program

```plaintext
proc nlmixed; * qpoints=15;
   parms beta0=-1.4 beta1=0.12 beta2=-0.12 theta=0.1 xi=0.5;
   bound theta>0, xi>-1, xi<1;

   eta = beta0 + beta1*assign + beta2*agn_trt + b;
   mu = interval*exp(eta);
   mu1 = (1-xi)*mu;
   mu2 = mu1 + xi*seize;

   l = log(mu1) + (seize-1)*log(mu2) - mu2 - lgamma(seize+1);

   model seize ~ general(l);
   random b ~ normal(0, theta) subject=id;
run;
```
The relevant output is

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
<th>Alpha</th>
<th>Lower</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>1.0635</td>
<td>0.1061</td>
<td>58</td>
<td>10.02</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td></td>
<td>0.8510</td>
<td></td>
</tr>
<tr>
<td>beta1</td>
<td>0.1256</td>
<td>0.08190</td>
<td>58</td>
<td>1.53</td>
<td>0.1307</td>
<td>0.05</td>
<td>-0.03837</td>
<td></td>
<td></td>
</tr>
<tr>
<td>beta2</td>
<td>-0.1150</td>
<td>0.1110</td>
<td>58</td>
<td>-1.04</td>
<td>0.3043</td>
<td>0.05</td>
<td>-0.3372</td>
<td></td>
<td></td>
</tr>
<tr>
<td>theta</td>
<td>0.5175</td>
<td>0.1076</td>
<td>58</td>
<td>4.81</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>0.3020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>xi</td>
<td>0.4516</td>
<td>0.03048</td>
<td>58</td>
<td>14.81</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>0.3906</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The estimated between-patient variance is $\hat{\theta} = 0.52$ and the conditional over-dispersion is $1/(1 - \hat{\xi})^2 = 1/(1 - 0.4516)^2 = 3.3$. The inference on the treatment effect $\beta_2$ is again similar.
5 Summary: what we covered

1. Advantages of longitudinal studies over other classical studies.

2. Challenge in analyzing data from longitudinal studies: correlation, within-subject and between-subject variation.

3. Linear mixed models for analyzing continuous longitudinal data: random effects are explicitly used to model the between-subject variation.

4. Generalized estimating equations (GEEs) for analyzing discrete longitudinal data when the correlation is not of major interest. Population-average interpretation.

5. Generalized linear mixed model for analyzing discrete longitudinal data where random effects are used to model the correlation. Subject-specific interpretation.