Linear Mixed Models for Correlated Data

There are many forms of correlated data:

- Multivariate measurements on different individuals, for example blood pressure, cholesterol, etc. Note that measurements on a particular individual are correlated.

- Clustered measurements, for example on a number of families we take cholesterol measurements. In this situation, the measurements on the different members of a family are correlated.

- Repeated measurements, for instance, we could measure cholesterol over time for several experimental conditions. Here, the measurements on each individual would be correlated.

- Longitudinal data, for example we could track the cholesterol levels for individuals over time. Again, the measurements on an individual would be correlated.

- Spatial data, for instance, we could measure the cholesterol level for individuals living in different counties. Note that the measurements which are near to one another in space would be correlated.

It is reasonable in all cases to assume that we would wish to include the effects of gender and age.
Linear Mixed Models

A general modelling framework for many of these problems is the linear mixed model (sometimes denoted LMM). These models use two components

- Fixed effects, like are found in regression. In our examples, these would be gender, age, experimental conditions, time, etc.

- Correlated errors which are modelled through random effects. These random effects account for the correlation in the data.

Motivation:

Consider a repeated measurements analysis where we let \( Y_{ij} \) be the response for the \( i \)th individual measured at the \( j \)th time, \( i = 1, \ldots, N \) and \( j = 1, \ldots, n_i \). Then, let

\[
Y_i = \begin{pmatrix}
Y_{i1} \\
\vdots \\
Y_{in_i}
\end{pmatrix} = \text{all measurements for subject } i.
\]

We will consider a two stage analysis.

Stage 1: (Regression Model)

\[
Y_i = Z_i \beta_i + \epsilon_i.
\]

\( Z_i \) is an \( n_i \times q \) matrix of known covariates, \( \beta_i \) is a \( q \times 1 \) vector of subject specific coefficients, and \( \epsilon_i \) are the errors, which are usually iid \( N(0, \sigma^2 I) \).

This models how the response evolves over time for the \( i \)th subject.
Linear Mixed Model Motivation, Cont.

Stage 2: (Parameter Model)

\[ \beta_i = K_i \beta + b_i. \]

\(K_i\) is a \(q \times p\) matrix of known covariates, \(\beta\) is a \(p \times 1\) vector of unknown regression parameters, and \(b_i\) are independent \(N(0, D)\).

This part of the model is used to explain the observed variability between subjects with respect to the subject-specific regression coefficients, \(\beta_i\).

Example:

Investigators are interested in the effect the inhibition of testosterone production in male Wistar rats on their craniofacial growth (Verdonck, et al 1998).

Male rats were randomized to one of three groups, control, low dose, or high dose of triptorelin, a testosterone inhibitor in rats. Triptorelin injections were performed at day 25, and again at day 45 for the high dose group. Measurements on all animals were taken every tenth day between 30 and 110 days.

The response of interest is the distance between two well defined points on the skull of each rat. The primary interest of the study is to estimate the changes over time, and determine if these changes are treatment dependent. Note that not all animals have measurements for the full 110 days.

Let’s discuss an analysis of these data using the two-stage analysis.
Rat Example, Cont.

Suppose that the model of interest for the first stage has the form

\[ Y_{ij} = \beta_1 + \beta_2 t_{ij} + \epsilon_{ij}, \quad j = 1, \cdots, n_i, \]

where \( t_{ij} = \log\left\{1 + \frac{(AGE_{ij} - 25)}{10}\right\} \), a logarithmic transformation of the original time scale.

The second stage model to be fit has the form

\[
\begin{cases}
\beta_{1i} = \beta_0 + b_{1i} \\
\beta_{2i} = \beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i}.
\end{cases}
\]

\( L_i, H_i, C_i \) are indicator variables defined to be 1 if the rat belongs to the low dose, high dose, and control groups, respectively. Otherwise, the variable is defined to be 0.

Notice that in this case the subject specific intercepts do not depend upon treatment. \( \beta_0 \) is the average response at the start of treatment, \( \beta_1, \beta_2, \beta_3 \) are the average time effects for each of the three treatment groups. These are the parameters of primary interest.

Let's find the following forms

\[ Z_i = \quad K_i = \quad \beta = \quad . \]

With other types of data, as with the rat example, it is often the regression parameters, \( \beta \) in the second stage which are of interest.
Two-stage Analysis, Cont.

There are some problems with the two-stage analysis:

- How to estimate the parameters?
- Suppose that the regression parameters from the second stage are the parameters of interest. There is no obvious relationship between these parameters and the covariates of interest.
- Not all time points were measured on all animals. This is not taken into account in the second stage.

We can address these issues by combining the two stages into a single model, the linear mixed model.

**The General Linear Mixed-Effects Model:**

If we substitute stage 2 into the model for stage 1, we get

$$Y_i = Z_i K_i \beta + Z_i b_i + \epsilon_i.$$  

Let $X_i = Z_i K_i$ be an $n_i \times p$ matrix. Then, we can write the model as

$$Y_i = X_i \beta + Z_i b_i + \epsilon_i,$$

where $\beta$ are the fixed effects, which are common to all subjects, while $b_i$ are the subject specific random effects. Additionally, $b_i \sim N_q(0, D)$, $\epsilon_i \sim N_{n_i}(0, \Sigma_i)$, the $b_i$'s and $\epsilon_i$'s are independent, and $Z_i(n_i \times q)$ and $X_i(n_i \times p)$ are matrices of known covariates.
Linear Mixed Model, Cont.

From the model

\[ Y_i = X_i \beta + Z_i b_i + \epsilon_i, \]

we can write

\[ Y_i | b_i \sim N(X_i \beta + Z_i b_i, \Sigma_i) \quad \text{and} \quad b_i \sim N(0, D), \quad (1) \]

for \( i = 1, \cdots, N \). We could then denote the respective density functions by \( f(y_i | b_i) \) and \( f(b_i) \).

Recall that

\[ f(A, B) = f(A | B) f(B), \quad \text{and} \]

\[ f(A) = \int f(A | B) dB = \int f(A | B) f(B) dB. \]

Thus, the marginal density of \( y \) can be written as

\[ f(y_i) = \int f(y_i | b_i) f(b_i) db_i. \]

This can easily be shown to be

\[ y_i \sim N(X_i \beta, Z_i D Z_i^T + \Sigma_i). \quad (2) \]

We call (1) the conditional or hierarchical formulation of the linear mixed model and (2) the marginal formulation of the model.

**Rat Example:**

We can form the combined model for this example as

\[ Y_{ij} = (\beta_0 + b_{1i}) + (\beta_1 L_i + \beta_2 H_i + \beta_3 C_i + b_{2i}) t_{ij} + \epsilon_{ij}. \]
Rat Example, Cont.

It is useful to write the model for each treatment group individually

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

- Notice that the intercepts and slopes are all subject specific.
- Different treatment groups have different slopes, but the same intercept.

We know that the hierarchical model has the form

\[
Y_i \mid b_i \sim N(X_i \beta + Z_i b_i, \Sigma_i) \quad \text{and} \quad b_i \sim N(0, D).
\]

Let’s find the form of \(X_i\).

We will assume that \(\Sigma_i = \sigma^2 I_{n_i}\). This is called conditional independence, and means that the responses on rat \(i\) are independent conditional on \(b_i\) and \(\beta\).
Rat Example, Cont.

Let’s consider the marginal model, which has the distribution

$$y_i \sim N(X_i \beta, Z_i D Z_i' + \Sigma_i).$$

We could also write this as

$$y_{ij} = \beta_0 + \beta_1 L_{ij} + \beta_2 H_{ij} + \beta_3 C_{ij} + \eta_{ij},$$

where $$\eta_i \sim N(X_{i0}, Z_i D Z_i' + \Sigma_i).$$

For the purpose of illustration, let’s consider the case where $$n_i \equiv 2.$$

Thus, in addition to the correlation in the errors, the marginal model implies that the variance function of the response is quadratic over time, with positive curvature, $$d_{22}.$$

What happens if we remove the random slopes from the rat model?
Random-Intercepts Model

Without random slopes,

- This assumes that all variability in subject-specific slopes can be attributed to treatment differences.

- Called the random-intercepts model. This has subject specific intercepts, but the same slopes within each treatment group.

What does the hierarchical model look like for this case?

Then, find the form of the marginal model, and specifically the marginal covariance matrix.
Random-Intercepts Model, Cont.

Thus, in this case

- Constant variance over time.
- Equal, positive correlation between any two measurements from the same animal.
- This covariance structure is called "compound symmetry", and $\rho$ is called the "intra-class correlation". When $\rho$ is large, the inter-subject variability ($d_{11}$) is large relative to the intra-subject variability ($\sigma^2$).

**Covariance Models:**

Often, the conditional independence assumption, ($\Sigma_i = \sigma^2 I_n$), is not realistic. For example, consider $\epsilon_i = \epsilon_{(1)i} + \epsilon_{(2)i}$, where

- $\epsilon_{(2)i}$ is the serial correlation component. That is, part of the individual’s profile is a response to time-varying stochastic processes.
- $\epsilon_{(1)i}$ is the measurement error component, and is independent of $\epsilon_{(2)i}$.

Then, $Y_i = X_i \beta + Z_i b_i + \epsilon_{(1)i} + \epsilon_{(2)i}$, where $b_i \sim N(0, D)$, $\epsilon_{(1)i} \sim N(0, \sigma^2 I_n)$, $\epsilon_{(2)i} \sim N(0, \tau^2 H_i)$, and the $b_i$’s and $\epsilon_i$’s are independent.
Covariance Models, Cont.

Typically, we must model the structure of the $n_i \times n_i$ correlation matrix $H_i$. Let the $(j,k)$th element of $H_i$ be $h_{ijk} = g(t_{ij}, t_{ik})$. That is, a function of the times $t_{ij}$ and $t_{ik}$.

Usually, this is assumed to be some function of the “distance” between the times. Often,

$$h_{ijk} = g(|t_{ij} - t_{ik}|),$$

for some decreasing function $g(\cdot)$ with $g(0) = 1$.

Some examples of this type of function are

- Exponential function: $g(|t_{ij} - t_{ik}|) = \exp(-\phi|t_{ij} - t_{ik}|)$.
- Gaussian function: $g(|t_{ij} - t_{ik}|) = \exp\{-\phi(t_{ij} - t_{ik})^2\}$.

Similar structures may be useful for the matrix $D$.

**Example: Autoregressive Covariance Structure.**

A first order Autoregressive Model (AR(1)) has the form

$$\alpha_t = \phi\alpha_{t-1} + \eta_t,$$

where $\eta_t \sim \text{iid } N(0, \sigma^2_{\eta})$.

Then, the covariance between two observations is

$$\text{cov}(\alpha_t, \alpha_{t+h}) = \frac{\sigma^2_{\eta} \phi^{|h|}}{1 - \phi^2}, \text{ for } h = 0, \pm 1, \pm 2, \cdots, |\phi| < 1.$$

From this, it can be shown that

$$\text{corr}(\alpha_t, \alpha_{t+h}) = \phi^{|h|}.$$
Autoregressive Model, Cont.

So, if we let \( \alpha_T = (\alpha_1, \ldots, \alpha_T)' \), then we can write

\[
\text{corr}(\alpha_T) = \begin{bmatrix}
1 & \phi^1 & \phi^2 & \cdots & \phi^T \\
\phi^1 & 1 & \phi^1 & \cdots & \phi^{T-1} \\
\phi^2 & \phi^1 & 1 & \cdots & \phi^{T-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\phi^T & \phi^{T-1} & \phi^{T-2} & \cdots & 1
\end{bmatrix}
\]

Notes:

- Much more complicated covariance structures are possible. In fact, this is a critical component of spatial random effects models and time series models.

- Often, we do not need both random effects \( b \) and \( \epsilon(2)i \).

**Estimation and Inference:**

The model

\[
Y_i = X_i \beta + Z_i b_i + \epsilon_i,
\]

contain the unknown quantities \( \beta, b_i, D, \Sigma_i \), which we must obtain from the data.
Estimation and Inference, Cont.

Notice that

- $\boldsymbol{\beta}, D, \Sigma$ are unknown, but fixed, parameters, and must be estimated from the data.

- $b_i$ is a random variable. Thus, it doesn’t make sense to estimate the value, but rather we need to predict it.

If we have $\hat{\beta}$ as an estimator of $\beta$ and $\hat{b}_i$ as a predictor of $b_i$, then,

- The population average prediction of $Y_i$ is: $\bar{Y}_i = X_i\hat{\beta}$.

- The subject-specific prediction is: $\hat{Y}_i = X_i\hat{\beta} + Z_i\hat{b}_i$.

There are several ways to get $\hat{\beta}$ and $\hat{b}_i$. Henderson (1950) derived estimating equations known as the mixed model equations:

$$
\begin{bmatrix}
\hat{\beta} \\
\hat{b}
\end{bmatrix} =
\begin{bmatrix}
X'\Sigma^{-1}X & X'\Sigma^{-1}Z \\
Z'\Sigma^{-1}X & Z'\Sigma^{-1}Z + B^{-1}
\end{bmatrix}^{-1}
\begin{bmatrix}
X'\Sigma^{-1}Y \\
Z'\Sigma^{-1}Y
\end{bmatrix},
$$

where

$$
\begin{bmatrix}
y_1 \\
\vdots \\
y_N
\end{bmatrix},
X =
\begin{bmatrix}
X_1 \\
\vdots \\
X_N
\end{bmatrix},
\begin{bmatrix}
b_1 \\
\vdots \\
b_N
\end{bmatrix},
\epsilon =
\begin{bmatrix}
\epsilon_1 \\
\vdots \\
\epsilon_N
\end{bmatrix},
$$

H.13
Estimation and Inference, Cont.

\[ \text{cov}(\epsilon) = \Sigma, \quad Z = \begin{bmatrix}
Z_1 & 0 & \cdots & 0 \\
0 & Z_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & Z_n
\end{bmatrix}, \]

and

\[ B = \begin{bmatrix}
D & 0 & \cdots & 0 \\
0 & D & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & D
\end{bmatrix}. \]

So the model has the form

\[ y = X\beta + Zb + \epsilon \text{ or } y \sim N(X\beta, ZDZ' + \Sigma). \]

If we let \( V = ZDZ' + \Sigma \), then the solutions to the estimating equations are

\[ \hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y \quad \text{and} \quad \hat{b} = BZ'V^{-1}(Y - X\hat{\beta}). \]

The estimate \( \hat{\beta} \) is a generalized least squares estimate. The predictor, \( \hat{b} \), is the best linear unbiased predictor (BLUP), for \( b \). Note that:

\[ E(\hat{\beta}) = \beta, \quad \text{var}(\hat{\beta}) = (X'V^{-1}X)^{-1}, \quad E(\hat{b}) = 0, \quad \text{and} \]

\[ \text{var}(\hat{b} - b) = B - BZ'V^{-1}ZB + BZ'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}B. \]
Estimation and Inference, Cont.

The previous variance is the variance of the prediction error (mean squared prediction error, MSPE). This is more meaningful than $\text{var}(\hat{b})$, since the appropriate measure of uncertainty for random variables is the mean square prediction error.

Another way to get these results: recall Bayes’ theorem

$$f(b|y) = \frac{f(y|b)f(b)}{\int f(y|b)f(b)db}.$$ 

So, in our case,

$$y|b \sim N(X\beta + Zb, \Sigma) \text{ and } b \sim N(0, D).$$

A little bit of manipulation will show that the posterior has the form

$$b|y \sim N(BZ'V^{-1}(y - X\beta), (Z'S^{-1}Z + D^{-1})^{-1}).$$

Therefore, the best predictor should be

$$E(b|y) = BZ'V^{-1}(y - X\beta).$$

**Estimating $V$:**

So far, all of these derivations have assumed that the variance-covariance matrix, $V$, and thus, $\Sigma_i$ and $D$, are known. This is virtually always not the case.
Estimating $V$, Cont.

One possible solution would be to obtain an estimate $\tilde{V}$ for $V$ and simply substitute this estimate into the forms for estimating $\hat{\beta}$ and $\hat{b}$:

$$
\hat{\beta} = (X'\tilde{V}^{-1}X)^{-1}X'\tilde{V}^{-1}Y \quad \text{and} \quad \hat{b} = BZ'\tilde{V}^{-1}(Y - X\hat{\beta}).
$$

This estimate of $\hat{b}$ is called the EBLUP or the estimated BLUP. It is also known as the empirical Bayes estimate.

We could use a similar philosophy to estimate the variance of $\hat{\beta}$, and the MSPE for $\hat{b}$. However, note:

- $\text{var}(\hat{\beta})$ is a consistent estimator of $\text{var}(\tilde{\beta})$ if $\tilde{V}$ is a consistent estimator of $V$.
- This estimator is biased, as the variability which arises from estimating $V$ is not accounted for in the estimate.
- Thus, $\text{var}(\hat{\beta})$ underestimates the true variability. However, in some situations this bias can be corrected.

How can we estimate $V$? i) Maximum Likelihood (MLE), ii) Restricted Maximum Likelihood (REML), or iii) Generalized Estimating Equations (GEE).
Maximum Likelihood Estimation

Consider collecting all of the unknown parameters in $\Sigma$ and $D$ into a parameter vector, $\theta$. Maximum likelihood estimates have the property that the values $\hat{\theta}$ and $\hat{\beta}$ maximize the likelihood of $y \sim N(X\beta, V(\theta))$.

That is, it minimizes $-2 \log L(y; \theta, \beta)$. Then the log-likelihood has the form

$$\ell(\beta, \theta, y) = \log |V(\theta)| + (y - X\beta)'V(\theta)^{-1}(y - X\beta) + N \log(2\pi).$$

First, we will “profile” $\beta$ out of this equation.

- To do this, we replace $\beta$ in the likelihood with it’s maximum likelihood estimate, $(X'V(\theta)^{-1}X)^{-1}X'V(\theta)^{-1}y$.

- Then, minimize the resulting equation with respect to $\theta$; this provides the estimator $\hat{\theta}_{MLE}$.

- Then, we can find

$$\hat{\beta}_{MLE} = (X'V(\hat{\theta}_{MLE})^{-1}X)^{-1}X'V(\hat{\theta}_{MLE})^{-1}y,$$

is the MLE of $\beta$.

- The MLE of $b$ is then

$$\hat{b}_{MLE} = B(\hat{\theta}_{MLE})Z'V(\hat{\theta}_{MLE})^{-1}(y - X\hat{\beta}_{MLE}).$$

- MLE estimates of $\theta$ are typically negatively biased (too small on average) since they do not take into account the number of fixed effects being estimated.
Restricted Maximum Likelihood Estimation

- Adjustments are made in the objective function to be minimized that account for the number of estimated mean parameters.

- This gives less biased estimates of $\theta$ by considering the likelihood of linear combinations of the elements of $y$.

- Consider $K'y$, where $K$ is any $N \times (N - p)$ full-rank contrast matrix, which has columns orthogonal to the $X$ matrix (that is $K'X = 0$). Then,

$$K'y \sim N(0, K'V(\theta)K)$$

Notice that there is no longer a $\beta$ in this distribution.

- Then, maximize the likelihood for the contrasts to gen the estimate $\hat{\theta}_{REML}$. Note, this does not depend upon the specific choice of $K$.

- This value can then be plugged into the appropriate forms to find $\hat{\beta}_{REML}$ and $\hat{b}_{REML}$.

- Note, SAS uses Newton-Raphson to estimate $\theta$ in both the MLE and REML cases. Then, estimates of $\hat{\beta}$ are given based upon $\hat{\theta}$. REML is the default in SAS.
REML vs. MLE

- Both methods are based upon the likelihood principle. This gives the estimates consistency, asymptotic normality, and efficiency.

- ML estimation provides estimates for fixed effects, whereas REML, by itself, cannot.

- In balanced models, REML provides estimates identical to those from ANOVA.

- REML takes into account the degrees of freedom for the fixed effects in the model. This is particularly important when the rank of $X$ is large relative to the sample size.

- Changing $\beta$ has no effect on the REML estimates of $\theta$.

- REML is not typically as sensitive to outliers as MLE.

- MLE is better than REML if we are comparing models (for example, using AIC).

- Most people prefer REML.
Estimated Generalized Least Squares

MLE and REML are numerically expensive and rely upon the Gaussian assumption. We could use generalized estimating equations according to the following derivation.

\[
Y_i = X_i\beta + Z_i b_i + \epsilon_i,
\]

where \(\epsilon_i \sim (0, \Sigma_i)\), \(b_i \sim (0, D)\), and \(\text{cov}(\epsilon_i, b_i) = 0\).

Then the generalized least squares estimator,

\[
\hat{\beta}_{GLS} = \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i' V_i(\theta)^{-1} X_i \right\}^{-1} \frac{1}{n} \sum_{i=1}^{n} X_i' V_i(\theta)^{-1} Y_i,
\]

which does not rely on the Gaussian assumption, just knowledge of the first two moments: \(\text{E}(Y_i) = X_i\beta\) and \(\text{var}(Y_i) = V_i\).

The idea of estimated GLS (called EGLS) is to substitute for \(V(\theta)\) a consistent estimator, \(\hat{V}\):

\[
\hat{\beta}_{GLS} = \left\{ X' \hat{V}^{-1} X \right\}^{-1} X' \hat{V}^{-1} Y.
\]

Thus, the fixed effects estimators for the MLE, REML, and EGLS are of the same form, except for the estimate of \(V\).

EGLS is appealing when \(V\) can be estimated “quickly” (for example with a non-iterative approach). Some argue that with a sufficient sample size and when interest lies primarily in \(\beta\), little efficiency is lost. For example, when \(\Sigma_i = \sigma^2 I\), one approach is to estimate \(D\) and \(\sigma^2\) by the method of moments to set \(\hat{V}\).

H.20
Inference for Mixed Models

Recall that
\[ \hat{\beta}_{GLS} = \{X'\hat{V}^{-1}X\}^{-1} X'\hat{V}^{-1}Y, \] and
\[ \text{var}(\hat{\beta}_{GLS}) = \{X'\hat{V}^{-1}X\}^{-1}. \]

In practice, substitute \( \theta \) with \( \hat{\theta} \) and consider the approximate Wald test, \( H_0 : A\beta = d \), with test statistic
\[ W = (A\hat{\beta} - d)' \left[ A \{X'\hat{V}^{-1}X\}^{-1} A' \right]^{-1} (A\hat{\beta} - d) \]

When the null hypothesis is true, \( W \sim \chi^2_{\text{rank}(A)} \).

As mentioned previously, there is variability from substituting \( \hat{\theta} \) for \( \theta \), which we have ignored. Thus, estimated standard errors underestimate the true variability.

Suppose that \( \text{var}(Y) = \sigma^2 V(\theta) \). Then, another test statistic for the hypothesis above is
\[ F^* = \frac{(A\hat{\beta} - d)' \left[ A \{X'\hat{V}^{-1}X\}^{-1} A' \right]^{-1} (A\hat{\beta} - d)}{\hat{\sigma}^2 \text{rank}(A)}. \]

When the null hypothesis is true, \( F^* \sim f_{\text{rank}(A), \text{den}(df)} \).

Note that \( \text{den}(df) \) needs to be approximated from the data. There are several ways to do this approximation (for example the Satterthwaite).

For small sample sizes, there two methods provide very different results (and different p-values). In certain balanced cases, the tests may be exact. Finally, both tests reduce to \( t \)-tests for a single \( \beta \) parameter.
Likelihood Ratio Tests

Consider the test of the hypothesis $H_0 : \beta \in \Theta_{\beta,0}$, where $\Theta_{\beta,0}$ is some subspace of the parameter space, $\Theta_\beta$ of the fixed effects $\beta$. Then,

$$-2 \log \lambda_N = -2 \log \left\{ \frac{\hat{L}_{ML,0}}{\hat{L}_{ML}} \right\},$$

where $\hat{L}_{ML,0}$ and $\hat{L}_{ML}$ are the maximized likelihoods obtained from maximizing over $\Theta_{\beta,0}$ and $\Theta_\beta$, respectively.

$$-2 \log \lambda_N \sim \chi^2_{df},$$

where $df$ is the difference in the dimension (or number of parameters) of $\Theta_\beta$ and $\Theta_{\beta,0}$.

Note that this is not valid in this case for REML.

Inference for Variance Components

- From classical theory, the ML and REML estimator, $\hat{\theta}$ is well approximated by a normal distribution with mean $\theta$ and covariance matrix given by the inverse Fisher information (for large samples).

- Approximate Wald tests can be developed, analogous to those for fixed effects. However,
  - Performance of the normal approximation depends strongly upon the true value of $\theta$. (The normal approximation fails for values of $\theta$ which are near the boundary of the parameter space $\Theta_\theta$. For example, $\sigma^2 \approx 0$.)
Inference for Variance Components, Cont.

- Wald tests, cont.
  - This works better for covariance parameters than for variance parameters.

- Likelihood ration tests can be developed using either ML or REML estimates. This has the same problems as the Wald tests when the parameters are near the boundary of the parameter space.

**Information Criteria**

Information Criteria can be used to suggest good models. Define Akaike’s Information Criteria (AIC) as

\[
AIC = \ell(\hat{\theta}, \hat{\beta}) - q,
\]

where \(\ell(\hat{\theta}, \hat{\beta})\) is the log-likelihood and \(q\) is the effective number of covariance parameters (those not estimated to be on a boundary constraint).

According to this definition, larger AIC values are better. Note that sometimes AIC is given by \(-2\ell + 2q\), and in this case, smaller is better.

Next, define the Bayesian Information Criteria (BIC) by

\[
BIC = \ell(\hat{\theta}, \hat{\beta}) - \frac{1}{2} q \log N^*,
\]

where \(N^*\) is the number of observations for the MLE \((n - p, \text{or the rank}(X) \text{for REML})\). Again, a larger BIC indicates a better model (unless we use the definition \(-2\ell + p \log N^*, \text{and then smaller is better}\).
Information Criteria, Cont.

These criteria can be used for both REML and MLE if the same mean structure is used. Otherwise we should on use the MLE.

Additionally, information criteria only provide a “rule of thumb”, but should not be taken as definite selections.

Pharmaceutical Stability Study

A pharmaceutical company was concerned about the shelf life of various drugs. There were 3 batches of products. On each batch, data are collected at 0, 1, 3, 6, 9, 12 months. Finally, at each time there are up to 6 repeated assay results taken from the batch. The response variable is the result of the assay, which is the relative potency of the drug compared to the claim on the label. See Section 6.2 in the SAS handouts.

Consider the model:

\[ y_{ijk} = \gamma_{0i} + \gamma_{1i}(age_{ik}) + \epsilon_{ijk}, \]

here \( i = 1, 2, 3 \) (batch), \( j = 1, 2, \ldots, n_{ik} \) (repeated measurements for batch \( i \) at time \( k \)), and \( k = 1, 2, \ldots, 6 \) (time). Here the intercept represents the initial potency and the slope reflects the degradation rate.

We will consider a random slope and intercept for each batch. Let

\[ \gamma_{0i} = \beta_0 + b_{0i} \quad \text{and} \quad \gamma_{1i} = \beta_1 + b_{1i}. \]
Example, Cont.

Here, $b_i = (b_{0i}, b_{1i})' \sim N(0, D_i)$ be the random effects and $\beta = (\beta_0, \beta_1)'$ be the fixed effects. Assume that $D_i \equiv D, i = 1, 2, 3$ and $\epsilon_{ik} \sim \sigma^2 I$. This is the default in PROC MIXED; if a more complicated model is needed, we should use the REPEATED statement.

**Split-Plot Designs:**

Split-plot designs originated from agricultural field trials. These often occur when there are two factors of interest and one factor requires larger experimental units than the other. For example, suppose that we are interested in studying irrigation amount and fertilizer type on the growth of a particular plant. Because of the equipment involved, different amounts of irrigation can only be done on a large scale, while different fertilizers can be applied much more locally.

Additionally, these types of designs are often used in other situations. For example an industrial experiment is used to study the freshness of milk. If the two factors are pasteurization process and type of container, we would need to pasteurize an entire batch, but we could use different types of containers in a particular batch.

Consider a typical split-plot design which has three replications, 4 levels of factor $A$, which must be applied to large units, and 3 levels of factor $B$, which may be applied to small units. Let’s draw a picture of this, assuming that the units are pieces of land.
Split-Plot Designs, Cont.

Why is this different from a randomized block design?

- Both have one occurrence of each of the twelve treatment combinations in each block.

- However, a randomized block design would have the twelve combinations assigned to the regions at random. A split-plot design does not have this amount of randomness.

- Also, a split-plot design has factor A applied only one time to each block, while factor B is applied multiple times. We need to account for the correlation which may be present due to some responses sharing an “application” of factor A.

- This is like a repeated measure design. Here the subplots correspond to occasions when measurements were taken.

Consider the model:

\[ Y_{ijk} = \mu + \rho_i + \alpha_j + (\rho\alpha)_{ij} + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}, \]

where \( i \) is the replicates (blocks or subjects), \( j \) is the level of factor A, and \( k \) is the level of factor B. Further, \( \mu \) is the overall mean, \( \rho_i \) is the effect of the block (random), \( \alpha_j \) is the main effect of factor A (fixed), \( (\rho\alpha)_{ij} \) is the block by factor A interaction (the whole plot error, random), \( \beta_k \) is the main effect of factor B (fixed), \( (\alpha\beta)_{jk} \) is the interaction between factors A and B (fixed), and \( \epsilon_{ijk} \) is the subplot error (random).
Split-Plot Designs, Cont.

A traditional ANOVA analysis looks at

- **Whole Plot comparisons:**
  - Compare factor A to the whole plot error ($\alpha_j$ to $(\rho\alpha)_{ij}$), and
  - Compare the block to the whole plot error ($\rho$ to $(\rho\alpha)_{ij}$).

- **Sub-plot comparisons:**
  - Compare factor B to the subplot error ($\beta$ to $\epsilon_{ijk}$), and
  - Compare the AB interaction to the subplot error ($((\alpha\beta)_{jk}$ to $\epsilon_{ijk}$).

Let’s consider this model from the mixed model perspective, and write it as

$$Y = X\beta + Zb + \epsilon.$$

To ensure that we understand what should make up the different elements of this model, let’s write it out in detail for the situation where $i = 1, 2$, $j = 1, 2, 3$, and $k = 1, 2$.

Let’s begin by finding the values for $Y$, $\beta$, $b$, and $\epsilon$. 
Split-Plot Designs, Cont.

Now, let’s find the forms of the design matrices, $X$ and $Z$.

In this setup, are all of the parameters estimable?

What does $\text{var}(b)$ look like? It is diagonal with variance components $\text{var}(\rho_i) = \sigma^2_{\rho}$ and $\text{var}\{(\rho_\alpha)_{ij}\} = \sigma^2_{\rho_\alpha}$.

Also, $\text{var}(\epsilon) = \sigma^2_{\epsilon}I$. Now, let’s look at a SAS example for this situation.

**Repeated Measures in Mixed Models:**

An equivalent form of the split-plot model can be written as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta_{ik} + \epsilon_{ijk},$$

where there are two random effects, $\delta_{ik}$ and $\epsilon_{ijk}$, ($i = 1, \ldots, n_A$, $j = 1, \ldots, n_B$, and $k = 1, \ldots, n_i$), corresponding to the whole-plot and subplot errors.
Repeated Measures in Mixed Models, Cont.

Under the usual assumptions ($\delta_{ik}$ and $\epsilon_{ijk}$ are iid Normal with zero means and variances $\sigma^2_\delta$ and $\sigma^2$, respectively), the leads to the variance-covariance matrix of the repeated observations on the $k$th subject of the $i$th groups ($y_{ik} = (y_{i1k}, \cdots, y_{in_Bk})'$) having the form

$$
\Sigma_{subject} = \begin{pmatrix}
\sigma^2_\delta + \sigma^2 & \sigma^2 & \cdots & \sigma^2 \\
\sigma^2 & \sigma^2_\delta + \sigma^2 & \cdots & \sigma^2 \\
\vdots & \vdots & \ddots & \vdots \\
\sigma^2 & \sigma^2 & \cdots & \sigma^2_\delta + \sigma^2
\end{pmatrix}
$$

$$
= \frac{1}{\sigma^2_\delta + \sigma^2} \begin{pmatrix}
1 & \rho & \cdots & \rho \\
\rho & 1 & \cdots & \rho \\
\vdots & \vdots & \ddots & \vdots \\
\rho & \rho & \cdots & 1
\end{pmatrix},
$$

where $\rho = \sigma^2_\delta / (\sigma^2_\delta + \sigma^2)$.

However, while this structure is appropriate for some agricultural experiments, it may not be realistic for repeated measurements on the same subject over time. For example, the structure might look more like an AR(1) structure

$$
\Sigma_{subject} = \sigma^2 \begin{pmatrix}
1 & \rho & \rho^2 & \cdots & \rho^{n_B-1} \\
\rho & 1 & \rho & \cdots & \rho^{n_B-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{n_B-1} & \rho^{n_B-2} & \rho^{n_B-3} & \cdots & 1
\end{pmatrix}.
$$
Repeated Measures in Mixed Models, Cont.

We could write the mixed model for a repeated measure as

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \]

where \( \epsilon_{ijk} \) combines random error of both the whole and subplots.

In general, \( Y = X\beta + \epsilon \), where \( \epsilon \sim N(0, \sigma^2\Sigma) \) and \( \Sigma \) is block diagonal if the random error covariance is the same for each subject.

Recall the heart rate data example, presented originally with MANOVA. In this data set, Spector (1987) desires to compare the effect of two drugs on human heart rate. To do this, 24 subjects were randomly assigned to one of three groups, representing the two drugs and control. Heart rate measurements were taken at four different time points, five minutes apart, after the administration of the drug.

Let’s look at performing this analysis under three different covariance structures using the repeated measures structure. Specifically, we will consider i) AR(1), ii) compound symmetry, and iii) unstructured covariance. Let’s look at the SAS Proc Mixed output for the situation.
Unbalanced or Unequally Spaced Data

Often, especially in clinical trials, patients miss their appointments or drop out of the study. (We will usually assume that this happens at random). Additionally, it is often not practical to take the measurements at evenly spaced time points.

For example, suppose that we have data which are the percentage of correct scores on a sentence hearing test administered to two groups of subjects wearing one of two different types of cochlear implant (A and B).

Suppose that there are 19 subjects in group A and 16 subjects in group B. Further, suppose that tests are administered 1, 9, 18, and 30 months after implantation.

The objectives are

1. to determine if there is any difference between the two implants, and

2. determine the average improvement curves as functions of the length of time since implantation.

Note that there are many missing values in the data.

The experimenters have reason to consider the model:

\[ y_{ikt} = \beta_0 + \beta_{0i} + \beta_1 t + \beta_{1i} t + \beta_2 t^2 + \beta_{2i} t^2 + \epsilon_{ikt}, \]

where \( i = 1, 2 \) (groups), \( k = 1, \ldots, n_i \) (individuals), and \( t = (t_1, t_2, t_3, t_4) = (1, 9, 18, 30) \) months (times).