Linear Mixed Models

Outline

- Motivation
- Random Intercept Model
- Random Intercept and Random Slope Model
- Linear Mixed Model (LMM)
- Multilevel Mixed Effects Models
- Optimization Algorithms
- Inference for Fixed Effects
- Inference for Variance Parameters
- Inference about the Random Effects
- Extending Linear Mixed Models

Motivation

Recall the orthodontic measurement data. One question of interest is the individual *growth curve*.

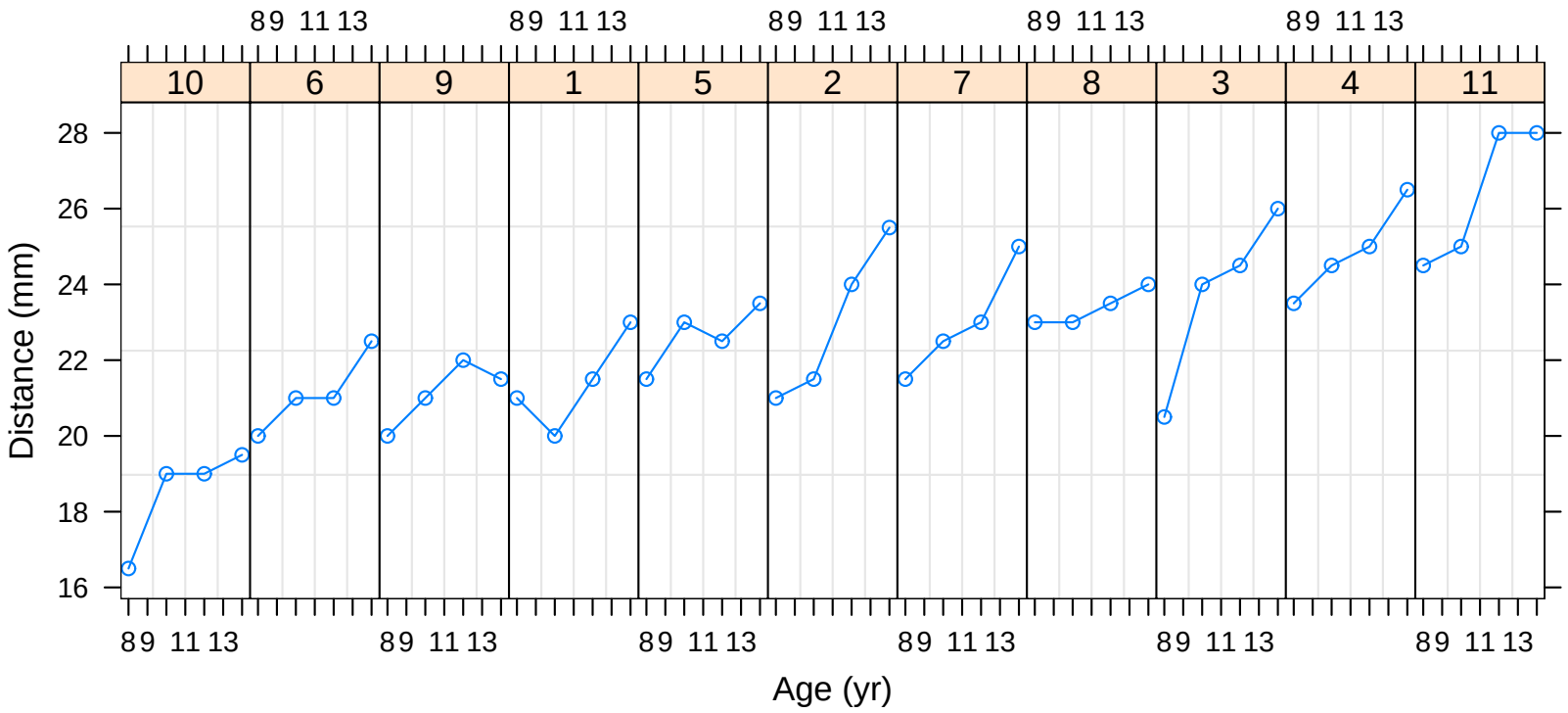

Figure 1: Orthodontic distance measurements for girls

Let's consider only the girls for the moment.

```
>Orthodont<- read.table ("orthodontic.dat",header=TRUE)
> library (nlme)
> Orth.new <- groupedData (distance ~ age | child,data = as.data.frame (Orthodont),
+                          FUN = mean,inner = ~ age, labels = list( x = "Age",
+                          y = "Distance" ),units = list( x = "(yr)", y = "(mm)") )
> OrthFem <- subset(Orth.new, male==0)
> OrthFem[1:5,]
Grouped Data: distance ~ age | child
  obs child age distance male
1   1     1   8     21.0    0
2   2     1  10     20.0    0
3   3     1  12     21.5    0
4   4     1  14     23.0    0
5   5     2   8     21.0    0

>library (lattice)
>plot (OrthFem )
```

`groupedData` is a special data class in R library `nlme` designed for describing clustered data. Many convenient functions are defined for it.

Let the random variable $\boldsymbol{Y}_i$ denote the outcomes for the ith individual, measured at time t_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Thus $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$.

Three modeling strategies to characterize the individual growth curve:

1. **Two-stage analysis**: fit a linear regression line to each subject and analyze the subject-specific regression coefficient as responses in the second stage analysis.

   - Stage 1:
   $$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \boldsymbol{\epsilon}_{ij} \text{ for each } i, \tag{1}$$
   - Stage 2:
   $$\beta_{0i} = \boldsymbol{z}_{0i}\beta_0 + \tau_i, \qquad \beta_{1i} = \boldsymbol{z}_{1i}\beta_1 + \upsilon_i, \tag{2}$$

   where $\boldsymbol{\epsilon}_i$, τ_i and υ_i assumed to be independent, following normal distribution.

2. **Fixed effects model**: include an indicator variable for subject id in the regression.
$$y_{ij} = \beta_0 + a_i + \beta_1 x_{ij} + \boldsymbol{\epsilon}_i, \tag{3}$$
   where a_i is fixed effect for the ith subject.

3. **Random effects model**: include random effect in the regression.
$$y_{ij} = \beta_0 + b_i + \beta_1 x_{ij} + \epsilon_{ij}, \tag{4}$$
   where b_i are random intercepts, both $\boldsymbol{\epsilon}_i$ and b_i are assumed to be independent, following normal distribution.

Two-Stage Analysis

```
> of.lis <- lmList (distance ~ I(age - 11), data = OrthFem)
> coef (of.lis)
   (Intercept) I(age - 11)
10      18.500       0.450
6       21.125       0.375
9       21.125       0.275
1       21.375       0.375
5       22.625       0.275
2       23.000       0.800
7       23.000       0.550
8       23.375       0.175
3       23.750       0.850
4       24.875       0.475
11      26.375       0.675

> sd(coef(of.lis)[,1])
[1] 2.104918

> sd(coef(of.lis)[,2])
[1] 0.2196071
```

Note: We centered the age variable so the intercept is interpretable.

```
> plot (intervals (of.lis))
```

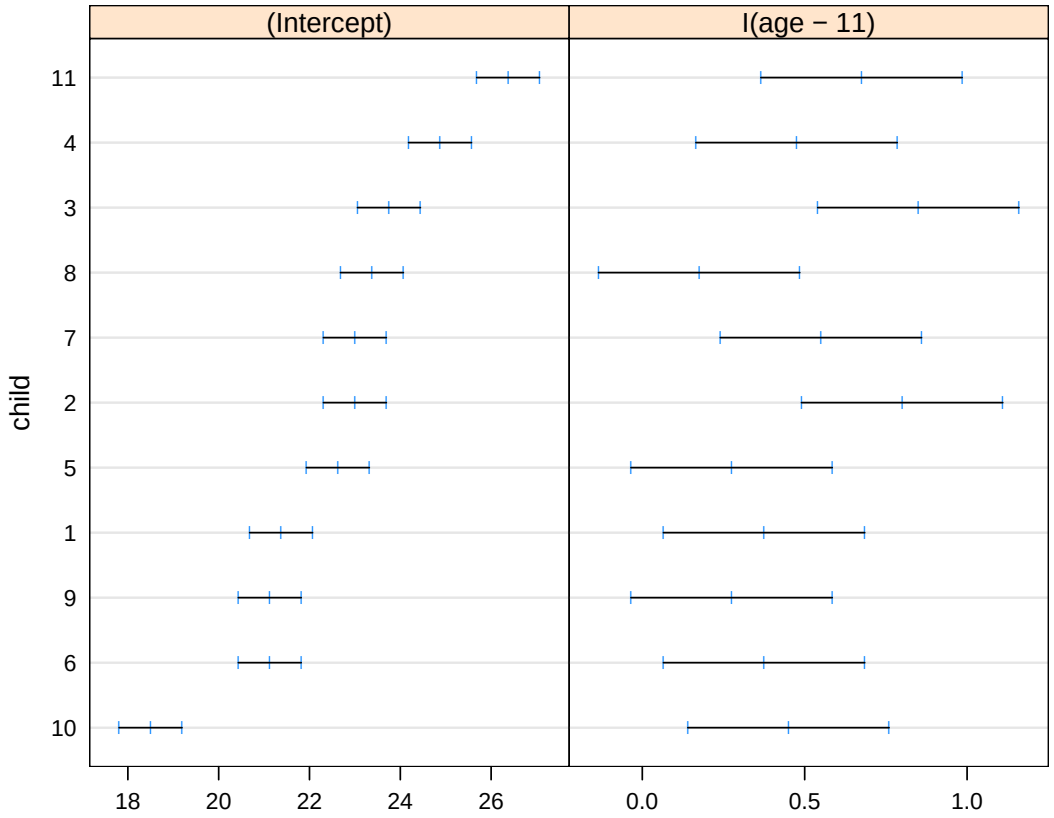

Figure 2: Confidence intervals (95%) for the coefficients of simple linear models.

Note: The s.e. and confidence intervals from lmList are wrong! So are those in Pinheiro and Bates' book.

- There is a lot of variation in the intercepts and the slopes are relatively comparable.
- Intuitively, we know this approach is not very efficient, since for every subjects we are estimating two parameters (not counting the standard errors), that are 22 parameters.
- If the data are not balanced, then the individual growth curve parameters are estimated at different precision and we need to weight them differently in the subsequent analysis.

Fixed Effects Model

We can include an indicator for subject, thus allow each to have a different intercept.

```
> of.lm <- lm (distance ~ factor (child, ordered = FALSE) +
+              I(age - 11) - 1,
+              data = OrthFem)
> summary(of.lm)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
factor(child)10 18.50000    0.39002  47.434  < 2e-16 ***
factor(child)6  21.12500    0.39002  54.164  < 2e-16 ***
factor(child)9  21.12500    0.39002  54.164  < 2e-16 ***
factor(child)1  21.37500    0.39002  54.805  < 2e-16 ***
factor(child)5  22.62500    0.39002  58.010  < 2e-16 ***
factor(child)2  23.00000    0.39002  58.972  < 2e-16 ***
factor(child)7  23.00000    0.39002  58.972  < 2e-16 ***
factor(child)8  23.37500    0.39002  59.933  < 2e-16 ***
factor(child)3  23.75000    0.39002  60.895  < 2e-16 ***
factor(child)4  24.87500    0.39002  63.779  < 2e-16 ***
factor(child)11 26.37500    0.39002  67.625  < 2e-16 ***
I(age - 11)      0.47955    0.05259   9.119 2.06e-10 ***
> sd(of.lm$coef[1:11])
[1] 2.104918
```

- Here we are estimating 12 parameters (11 intercepts and one slope).
- The precision on the slope is substantially better (0.05 vs 0.22).
- The intercepts (and CIs) are similar to those in separate regressions.
- However the intercepts in this model do not have the interpretation as population parameters.

Random Effects Model

One solution is to use a random intercept for subjects.

```
> of.lme <- lme (distance ~ I(age - 11),
+                random = ~ 1 | child, data = OrthFem)
> summary (of.lme)
Random effects:
 Formula: ~1 | child
        (Intercept)  Residual
StdDev:     2.06847 0.7800331
Fixed effects: distance ~ I(age - 11)
                Value Std.Error DF t-value p-value
(Intercept) 22.647727 0.6346568 32 35.6850       0
I(age - 11)  0.479545 0.0525898 32  9.1186       0

> orth.i <- cbind (two.stage = coef (of.lis)[,1],
+                  fixed = coef (of.lm)[1:11],
+                  random = coef (of.lme)[,1])
> rownames (orth.i) <- NULL
```

```
> orth.i
      two.stage  fixed random
 [1,]    18.500 18.500 18.642
 [2,]    21.125 21.125 21.177
 [3,]    21.125 21.125 21.177
 [4,]    21.375 21.375 21.419
 [5,]    22.625 22.625 22.626
 [6,]    23.000 23.000 22.988
 [7,]    23.000 23.000 22.988
 [8,]    23.375 23.375 23.350
 [9,]    23.750 23.750 23.712
[10,]    24.875 24.875 24.799
[11,]    26.375 26.375 26.247
```

- The estimate and se for the slope are very close to the previous model.
- The std. dev. for the random intercept (2.07) is slightly smaller than the std. dev. for the intercepts in the previous model (2.10).
- The intercepts are "shrunk" toward the mean.

Random Intercept Model

$$Y_{ij} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + b_i + \epsilon_{ij}, \tag{5}$$

where $b_i \sim N(0, \sigma_b^2)$ is the random subject effect, $\epsilon_{ij} \sim N(0, \sigma^2)$ are within-subject measurement errors, and b_i and ϵ_{ij} are assumed to be independent of each other.

Note the above random intercept model describes the mean response trajectory over time for the ith subject, i.e. the *conditional* mean of Y_{ij} given the subject-specific effect b_i:

$$E(Y_{ij}|b_i) = \boldsymbol{X}_{ij}^T\boldsymbol{\beta} + b_i;$$

and the mean response profile in the population, i.e. *marginal* mean of Y_{ij}:

$$E(Y_{ij}) = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}$$

The interpretation of the parameters:

- $\boldsymbol{\beta}$ describe patterns of change in the mean response over time in the population of interest.
- b_i represents the ith individual's deviation from the population mean intercept, after the effect of covariates have been accounted for.

For the random intercept model, the marginal variance of each response is given by

$$\begin{aligned}\text{Var}(Y_{ij}) &= \text{Var}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + b_i + \epsilon_{ij}) \\ &= \text{Var}(b_i + \epsilon_{ij}) \\ &= \sigma_b^2 + \sigma^2\end{aligned}$$

Similarly, the marginal covariance between any pair of response, Y_{ij} and Y_{ik}, is given by

$$\begin{aligned}\text{Cov}(Y_{ij}, Y_{ik}) &= \text{Cov}(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + b_i + \epsilon_{ij}, \boldsymbol{X}_{ik}^T\boldsymbol{\beta} + b_i + \epsilon_{ik}) \\ &= \text{Cov}(b_i + \epsilon_{ij}, b_i + \epsilon_{ik}) \\ &= \text{Cov}(b_i, b_i) \\ &= \sigma_b^2\end{aligned}$$

The correlation is

$$\text{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$$

Therefore, the introduction of a random intercept, b_i, induces correlation among the repeated measurements in longitudinal data.

Random Intercept and Random Slope Model

Considering the following random intercept and random slope model

$$Y_{ij} = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 t_{ij} + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij}, \tag{6}$$

where $b_{1i} \sim N(0, g_{11})$, and $b_{2i} \sim N(0, g_{22})$ are the random intercept and random slope, $\text{Cov}(b_{1i}, b_{2i}) = g_{12}$, $\epsilon_{ij} \sim N(0, \sigma^2)$ are within-subject measurement errors, and $\boldsymbol{b}_i = (b_{1i}, b_{2i})'$ and ϵ_{ij} are assumed to be independent.

The *conditional* mean of Y_{ij} given the subject-specific effect $\boldsymbol{b}_i$:

$$E(Y_{ij}|\boldsymbol{b}_i) = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 t_{ij} + b_{1i} + b_{2i} t_{ij};$$

and the mean response profile in the population, i.e. *marginal* mean of Y_{ij}:

$$E(Y_{ij}) = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 t_{ij}$$

The marginal variance of each response is given by

$$\begin{aligned} \text{Var}(Y_{ij}) &= \\ &= \\ &= \end{aligned}$$

The marginal covariance between any pair of response, Y_{ij} and Y_{ik}, is given by

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= \\ &= \\ &= \end{aligned}$$

- Subjects vary not only in their baseline level of response, but also in terms of the changes in their response over time.
- $\text{Var}(\boldsymbol{Y}_i)$ is more flexible – a function of time.

Linear Mixed Models (LMM)

Using the hierarchal notation of Laird and Ware (1982), we can express the linear mixed model as:

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \tag{7}$$

where $i = 1, \ldots, m$ and

$$\begin{aligned}
&\boldsymbol{Y}_i : (n_i \times 1) \text{ response vector} \\
&X_i : (n_i \times p) \text{ design matrix for fixed effects} \\
&\boldsymbol{\beta} : (p \times 1) \text{ regression coefficients for fixed effects} \\
&Z_i : (n_i \times q) \text{ design matrix for random effects} \\
&\boldsymbol{b}_i : (q \times 1) \text{ random effects} \\
&\boldsymbol{\epsilon}_i : (n_i \times 1) \text{ error vector}
\end{aligned}$$

Distributional assumptions: $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ are independent with

$$\begin{aligned}
\boldsymbol{b}_i &\sim \mathcal{N}(\mathbf{0}, D) \\
\boldsymbol{\epsilon}_i &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2 I\right)
\end{aligned}$$

Note that D is a $q \times q$ matrix that does not depend on i. Under these assumptions, $\boldsymbol{Y}_i$ has a multivariate normal distribution:

$$\boldsymbol{Y}_i \sim \mathcal{N}(X_i\boldsymbol{\beta}, V(\boldsymbol{\alpha})) \tag{8}$$

where $V(\boldsymbol{\alpha}) = Z_i D Z_i^T + \sigma^2 I$ and $\boldsymbol{\alpha}$ denotes the variance component parameters.

- D must be symmetric and positive definite.
- The errors $\boldsymbol{\epsilon}$ are assumed to be iid normally distributed with variance σ^2. This assumption will be relaxed later.
- The columns of matrix Z_i are typically a subset of the columns in X_i. In particular, $Z_i = \mathbf{1}_i$ corresponds to the random intercept model.
- Suppose that we have one covariate and $X_i = Z_i = (\mathbf{1}_i, \boldsymbol{X}_i)$ (random intercept and random slope model), then we can write:

$$\boldsymbol{Y}_i = X_i(\boldsymbol{\beta} + \boldsymbol{b}_i) + \boldsymbol{\epsilon}_i,$$

or

$$\boldsymbol{Y}_i = X_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i,$$

where

$$\boldsymbol{\beta}_i \sim \mathcal{N}\left(\boldsymbol{\beta}, D\right),$$

and

$$D = \begin{pmatrix} d_{00} & d_{01} \\ d_{10} & d_{11} \end{pmatrix},$$

with $d_{00} = \mathrm{Var}(\beta_{i0})$, $d_{11} = \mathrm{Var}(\beta_{i1})$ and $d_{01} = d_{10} = \mathrm{Cov}(\beta_{i0}, \beta_{i1})$.

- The *conditional* mean of $\boldsymbol{Y}$ given the subject-specific effect $\boldsymbol{b}_i$:

$$E(\boldsymbol{Y}_i|\boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i b_i$$

- The *marginal* mean for $\boldsymbol{Y}$ is the same as in the marginal general linear model:

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}_i) &= \mathrm{E}\left(\mathrm{E}(\boldsymbol{Y}_i \mid \boldsymbol{b}_i)\right) \\ &= \mathrm{E}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i) \\ &= \boldsymbol{X}_i\boldsymbol{\beta}. \end{aligned}$$

  Thus the regression coefficients $\boldsymbol{\beta}$ has the population mean interpretation. This will no longer hold for nonlinear models where $\mathrm{E}(\boldsymbol{Y}_i \mid \boldsymbol{b}_i)$ is not a linear function of $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$.

- The conditional variance of $\boldsymbol{Y}_i$ given the subject-specific effect $\boldsymbol{b}_i$:

$$\mathrm{Cov}(\boldsymbol{Y}_i|\boldsymbol{b}_i) = \mathrm{Cov}(\boldsymbol{\epsilon}_i) = \sigma^2 \boldsymbol{I}$$

  Note, it is often referred to as a *"conditional independence assumption"*. $\mathrm{Cov}(\boldsymbol{\epsilon}_i)$ can be relaxed to a more general covariance matrix, such that $\mathrm{Cov}(\boldsymbol{Y}_i|\boldsymbol{b}_i) = \boldsymbol{\Sigma}_i$, which describes the covariance among the longitudinal observations of a *specific* individual.

- The marginal variance of $\boldsymbol{Y}_i$, averaged over the distributions of $\boldsymbol{b}_i$, is

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{Y}_i) &= \mathrm{Cov}(\boldsymbol{Z}_i\boldsymbol{b}_i) + \mathrm{Cov}(\boldsymbol{\epsilon}_i) \\ &= \boldsymbol{Z}_i\,\mathrm{Cov}(\boldsymbol{b}_i)\boldsymbol{Z}_i^T + \mathrm{Cov}(\boldsymbol{\epsilon}_i) \\ &= \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{I} \end{aligned}$$

Multilevel Mixed Effects Models

The hierarchal specification of linear mixed model can be easily extended to multiple nested levels, to accommodate, for example, longitudinal measurements from subjects in the same clinical center, family or community.
Let $k = 1, \ldots, K$ to index the group, and $i = 1, \ldots, m_k$ for individuals in group k, $j = 1, \ldots, n_i$ for observational times for individual i. The model can be written as:

$$\boldsymbol{Y}_{ki} = \boldsymbol{X}_{ki}\boldsymbol{\beta} + \boldsymbol{Z}_{ki1}\boldsymbol{b}_k + \boldsymbol{Z}_{ki2}\boldsymbol{b}_{ki} + \boldsymbol{\epsilon}_{ki}, \tag{9}$$

where

$$\begin{aligned} \boldsymbol{b}_k &\sim \mathcal{N}(\boldsymbol{0}, D_1) \\ \boldsymbol{b}_{ki} &\sim \mathcal{N}(\boldsymbol{0}, D_2) \\ \boldsymbol{\epsilon}_{ki} &\sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 I\right). \end{aligned}$$

- The first-level random effects $\boldsymbol{b}_k$, of length q_1 are assumed to be independent for different k (groups).
- The second-level random effects $\boldsymbol{b}_{ki}$, of length q_2 are assumed to be independent for different k (groups) or i (individuals), and of the $\boldsymbol{b}_k$.
- $\boldsymbol{\epsilon}_{ki}$ are independent for k, i, and independent of the random effects.

Optimization Algorithms

- In the original paper (Laid and Ware, 1982; Lindstrom and Bates, 1988), an EM (expectation-maximization) algorithm was used for estimation. Nowadays, the estimation is often done based on the marginal model, using numeric optimization to maximize the likelihood (ML) or restricted likelihood (REML).
- In particular, `lme()` uses a mixed EM and Newton-Raphson iterations whereas SAS `PROC MIXED` uses Newton-Raphson. For small data sets with large models, it may be wise to monitor the convergence and change some optimization parameters when necessary. Better still, avoid fitting over-elaborated models.

Review of EM Algorithm and Newton Raphson Algorithm

- The EM algorithm (Dempster, Laird and Rubin, 1977)
  - A popular iterative algorithm for likelihood estimation in models with incomplete data and useful when maximization is simple for complete data.
  - The missing data are "imputed" by computing expectation of complete data log likelihood given the observed data. In LME, the random effects are treated as missing/unobserved data.
  - Individual iterations of the EM algorithm are quickly and easily computed and generally bring the parameters into the region of the optimum very quickly.
  - The rate of convergence of EM algorithm is slow and the progress toward the optimum tends to be slow when near the optimum.

- The Newton-Raphson algorithm
  - Newton-Raphson algorithm uses a first-order expansion of the score function around the current estimate of parameters to produce the next estimate of parameters.
  - Newton-Raphson iterations are individually more computationally intensive than the EM iterations because of the calculation of Hessian matrix.
  - Unstable when far from the optimum but converge quickly when close to the optimum.
  - The by-product of Newton-Raphson algorithm is the asymptotic variance-covariance matrix of the parameters obtained from the Hessian matrix in the last iteration.

- The optimization algorithm in LME models.
  - The function of lme() implements a hybrid approach with mixed EM and Newton-Raphson algorithms (Pinheiro and Bates, 2000).
    - A moderate number of EM iterations after the initial value of parameters are used to refine the parameter estimates and get near the optimum. By default 25 EM iterations are performed.
    - The Newton-Raphson iterations are performed afterwards to complete the convergence to the optimum.

An example to illustrate the hybrid approach for fitting LME using lme() and monitor the progress of the Newton-Raphson iterations.

```
> orth.lme.fit1 <- lme (distance ~ I(age - 11),random = ~ 1 | child,
+                       data = Orth.new, method = "ML", control=list(msVerbose=T))
  0:     321.28450: -0.376007
  1:     321.28450: -0.376007
>
> orth.lme.fit2 <- lme (distance ~ I(age - 11),random = ~ 1 | child,
+                       data = Orth.new, method = "ML", control=list(msVerbose=T, niterEM=0))
  0:     321.41166: -0.287682
  1:     321.28921: -0.393080
  2:     321.28450: -0.376159
  3:     321.28450: -0.376005
  4:     321.28450: -0.376007
```

Inference for Fixed Effects in LME

For a contrast matrix $\boldsymbol{L}$, to test the null hypothesis $H_0 : \boldsymbol{L\beta} = 0$ versus $H_1 : \boldsymbol{L\beta} \neq 0$.

Wald test

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{L}^T \left\{ \boldsymbol{L} \left(\sum_i \boldsymbol{X}_i^T \boldsymbol{V}_i(\hat{\boldsymbol{\alpha}})^{-1} \boldsymbol{X}_i \right) \boldsymbol{L}^T \right\} \boldsymbol{L}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

- Has an approximate χ^2 distribution with rank($\boldsymbol{L}$) degrees of freedom.
- Wald test tends to be anti-conservative (Dempster, Rubin and Tsutakawa, 1981). In practice, this downward bias of Wald test can be resoved by using apporximate t-/F-test.

F test

$$F = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{L}^T \left\{ \boldsymbol{L} \left(\sum_i \boldsymbol{X}_i^T \boldsymbol{V}_i(\hat{\boldsymbol{\alpha}})^{-1} \boldsymbol{X}_i \right) \boldsymbol{L}^T \right\} \boldsymbol{L}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\text{rank}(\boldsymbol{L})},$$

- Has an approximate F distribution with numerator degrees of freedom rank($\boldsymbol{L}$) and denominator degrees of freedom estimated from the data.
- In practice, several methods are available for estimating the appropriate number of degrees of freedom, such as the Satterthwaite-type approximation (Satterthwaite 1941).
- Robust versions of the above tests can be obtained by replacing the model-based covariance matrix with the robust one.

Likelihood ratio test

For nested models with different mean structures, likelihood ratio tests can also be used. For the null hypothesis $H_0 : \boldsymbol{\beta} \in \Theta_{\boldsymbol{\beta},0}$, where $\Theta_{\boldsymbol{\beta},0}$ is some subspace of the parameter space $\Theta_{\boldsymbol{\beta}}$ of the fixed effects $\boldsymbol{\beta}$, the LR test statistic is

$$-2 \ln \left[\frac{\mathcal{L}(\hat{\boldsymbol{\beta}}_{ML,0})}{\mathcal{L}(\hat{\boldsymbol{\beta}}_{ML})} \right]$$

- Under some regularity conditions, the LR test statistic follows asymptotic chi-squared distribution with degrees of freedom equal to the dimension difference between the two parameter spaces.
- The estimation has to be done using ML instead of REML for testing the fixed effects from nested models.
- LRT tests tend to be anticonservative (nominal p-value is smaller than true value). Hence, we prefer the F-tests for assessing the significance of terms in the fixed effects.

```
> LRTsim = simulate.lme(list(fixed = distance ~ male, data = Orth.new, random = ~ 1 | child),
+ m2 = list(fixed = distance ~ age * male),
+ method = "ML", nsim=1000)
> plot (LRTsim, df = c(2,3))
```

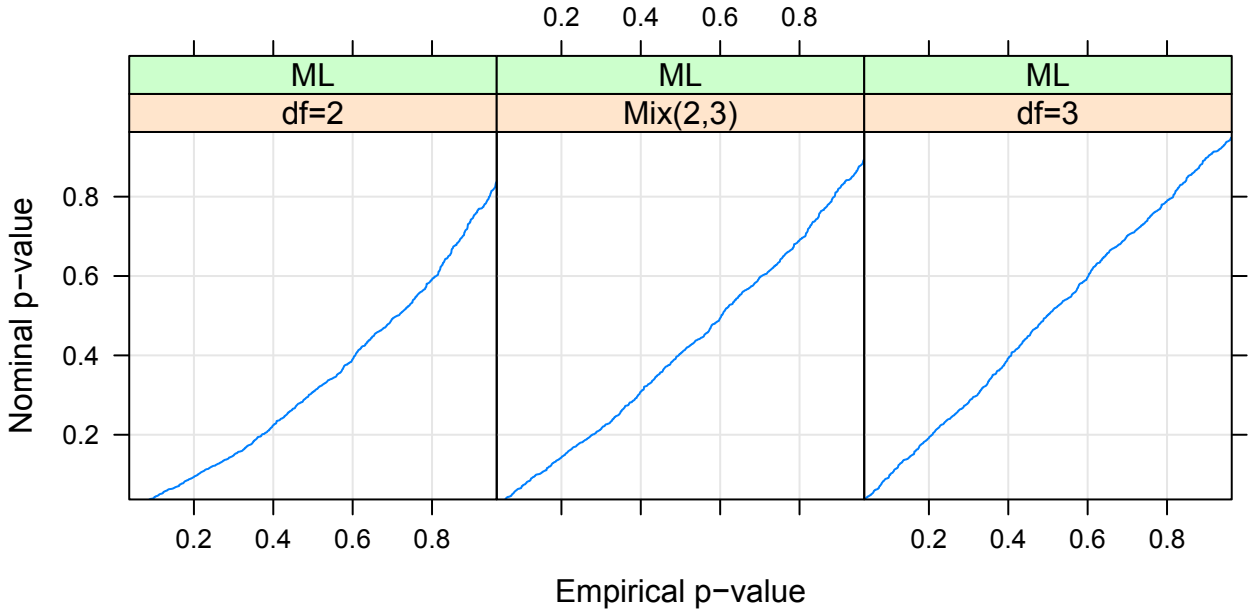

Figure 3: Plots of the nominal versus empirical p-values for the likelihood ratio test statistic comparing two nested models with different fixed effects

Inference for Variance Parameters in LME

In practice, usually the mean structure rather than the covariance model is of primary interest. However,

- adequate covariance is essential to obtain valid model-based inferences for the mean model parameters;
- useful for the interpretation of the random variation in the data;
- overparameterization of the the covariance structure leads to inefficient estimation.

Wald test

- The distribution of the ML as well as REML estimator $\hat{\boldsymbol{\alpha}}$ can be well approximated by a normal distribution with mean vector $\boldsymbol{\alpha}$ and with covariance matrix given by the inverse of the Fisher information matrix, similarly as for the fixed effects.

Likelihood ratio test

Likelihood ratio test can be used to compare nested models with different variance parameters.

- Both ML and REML can be used (note that REML should only be used for models with same fixed-effects specification).
- However, when one of the model sets some parameters at the boundary of the parameter space, the degrees of freedom for the LRT needs to be adjusted (Stram and Lee, 1994; Self and Liang, 1987). The unadjusted LRT tends to be too conservative.
  - A random intercept model: $\text{Var}(b_i) = \tau^2$. Testing the null hypothesis that the random intercept is not needed is equivalent to $H_0 : \tau = 0$ vs $H_1 : \tau > 0$. The LRT statistic has a mixture distribution that puts 0.5 mass on 0 and 0.5 mass at χ^2_1, or $\sim 0.5\chi^2_0 + 0.5\chi^2_1$.
  - A random intercept and slope model:

$$\text{Var}(\boldsymbol{b}_i) = \begin{pmatrix} \tau_0^2 & \rho\tau_0\tau_1 \\ \rho\tau_0\tau_1 & \tau_1^2. \end{pmatrix}.$$

    Under the null hypothesis $H_0 : \tau_1 = 0$, the correlation ρ degenerates. Therefore, the distribution of LRT statistic is a $1 : 1$ mixture of χ^2_1 and χ^2_2.

```
> LRTsim2 = simulate.lme(list(fixed = distance ~ age, data = OrthFem,
+                         random = ~ 1 | child), nsim = 1000,
+                  m2 = list(random = ~ age | child))
> plot(LRTsim2, df = c(1,2))
```

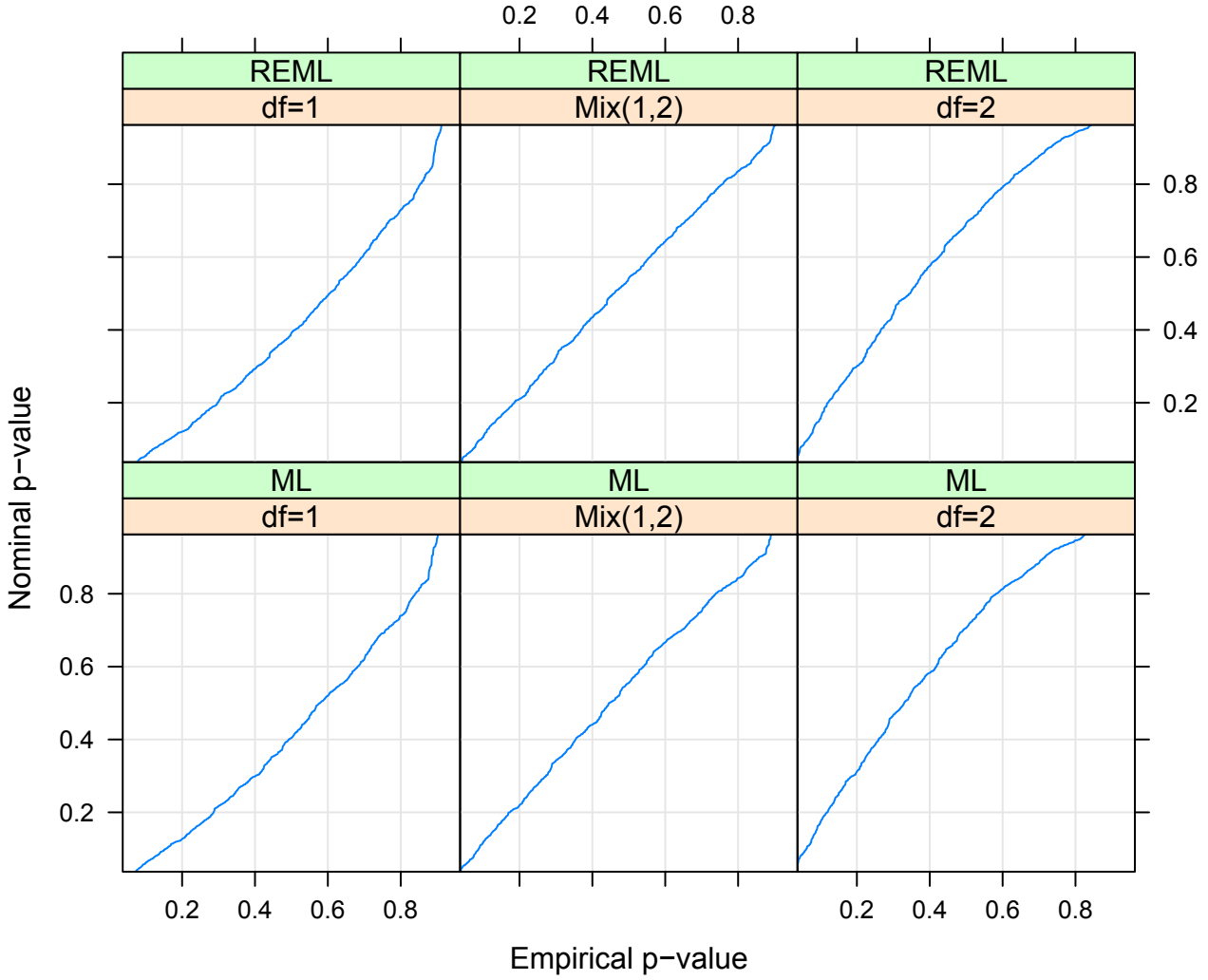

Figure 4: Plots of the nominal versus empirical p-values for the likelihood ratio test statistic comparing the random intercept only model and the random intercept

- More generally, when comparing a model with $q+1$ (correlated) random effects with the model with q (correlated) random effects, the distribution of the LRT statistic is a 1:1 mixture of χ^2_q and χ^2_{q+1}.
- When comparing models with $q+k$ and q correlated random effects where $k > 1$, the distribution of the LRT statistic is not well understood.
- There is no general rule to reliably come up with the distribution of LRT statistic.

Non-nested models

Information criteria can be used to compare non-nested models. They are based on the likelihoods with a penalty term that is larger for models with larger number of parameters.

- Let n_{par} denote the total number of parameters (fixed and random effects), and $N = \sum_{i=1}^{m} n_i$, then the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are defined as:

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) + 2n_{\text{par}}$$
$$\text{BIC} = -2\ell(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}) + n_{\text{par}} \log(N).$$

- For models with the same mean structure, the REML versions of AIC and BIC replaces $\ell(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y})$ with $\ell_R(\hat{\boldsymbol{\theta}} \mid \boldsymbol{y})$ and N with $N - p$ where p is the number of fixed effects parameters. (Again ML should be used to compare models with different fixed effects).
- Models with the *smaller* AIC or BIC are better.
- Information criteria are more flexible than likelihood ratio test but they only provide a "rule-of-thumb" and not a formal statistical significance test.
- Different criteria can lead to different conclusions.

Inference about the Random Effects

The random effects $\boldsymbol{b}_i$ are *random variables*, not parameters. Technically, we *predict* the random effects, not estimate them.

- Why do we want to estimate $\boldsymbol{b}_i$?
  - They reflect how much the subject-specific profiles (i.e. response trajectories $\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$) deviate from the overall average profile.
  - Helpful for detecting special profiles or groups of individuals evolving differently in time.

Prediction and Shrinkage

- From a Bayesian perspective, the posterior distribution of $\boldsymbol{b}_i$ given the data $\boldsymbol{y}_i$ is (dependence on the parameters $\boldsymbol{\theta}$ is suppressed)

$$f(\boldsymbol{b}_i \mid \boldsymbol{y}_i) = \frac{f(\boldsymbol{y}_i \mid \boldsymbol{b}_i) f(\boldsymbol{b}_i)}{\int f(\boldsymbol{y}_i \mid \boldsymbol{b}_i) f(\boldsymbol{b}_i) d\boldsymbol{b}_i},$$

  which can be shown to be a multivariate normal distribution. Thus $\boldsymbol{b}_i$ can be estimated using the posterior mean

$$\begin{aligned} \hat{\boldsymbol{b}}_i(\boldsymbol{\theta}) &= \mathrm{E}(\boldsymbol{b}_i \mid \boldsymbol{y}_i) \\ &= \int \boldsymbol{b}_i f(\boldsymbol{b}_i \mid \boldsymbol{y}_i) d\boldsymbol{b}_i \\ &= \boldsymbol{D}\boldsymbol{Z}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}). \end{aligned}$$

In practice, unknown parameters $\boldsymbol{\theta}$ are replaced by their ML or REML estimates. The resulting estimates for the random effects,

$$\hat{\boldsymbol{b}}_i(\hat{\boldsymbol{\theta}}) = \hat{D} Z_i^T \hat{V}_i^{-1}(\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}})$$

are called *empirical Bayes* (EB) estimates or the *empirical best linear unbiased predictors* (EBLUP).

- Consider the prediction of the response for i,

$$\begin{aligned}
\hat{\boldsymbol{Y}}_i &= X_i\hat{\boldsymbol{\beta}} + Z_i\hat{\boldsymbol{b}}_i \\
&= X_i\hat{\boldsymbol{\beta}} + Z_i D Z_i^T V_i^{-1}(\boldsymbol{y}_i - X_i\hat{\boldsymbol{\beta}}) \\
&= \left(I_{n_i} - Z_i D Z_i^T V_i^{-1}\right) X_i\hat{\boldsymbol{\beta}} + Z_i D Z_i^T V_i^{-1}\boldsymbol{y}_i \\
&= \hat{\Sigma}_i V_i^{-1} X_i\hat{\boldsymbol{\beta}} + \left(I_{n_i} - \Sigma_i V_i^{-1}\right)\boldsymbol{y}_i,
\end{aligned}$$

  where $\Sigma_i = V_i - Z_i D Z_i^T$ is the residual variance.

  - It can be interpreted as a weighted average of the population mean $X\hat{\boldsymbol{\beta}}$ and the observed data $\boldsymbol{y}_i$.
  - The empirical BLUP predictor "shrinks" the ith subject's response profile towards the population average mean response profile.
  - The amount of "shrinkage" depends on the relative magnitude of Σ_i and V_i. Or in other words, larger weights are given to the overall mean if the residual variability Σ_i is large in comparison with the between-subject variability $Z_i D Z_i^T$.

Example of shrinkage: the random-intercepts model

$$\hat{b}_i = 0 + \left(\frac{n_i\sigma_b^2}{\sigma^2 + n_i\sigma_b^2}\right)\left[\frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij} - X_{ij}^T\boldsymbol{\beta})\right]$$

- $\hat{b}_i$ is the weighted average of 0 (the prior mean of b_i) and the average residual $\bar{r}_{i\cdot}$.
- The larger the n_i, the more weight is put on $\bar{r}_{i\cdot}$.
- The larger the within-subject variability (σ^2) relative to the between-subject variability (σ_b^2), the more shrinkage toward zero (note, the weight on 0 can thought of as $\frac{\sigma^2}{\sigma^2+n_i\sigma_b^2}$).

Normality Assumption

- The random effects are assumed to be normally distributed. When that assumption is violated, the inference about marginal model, and especially the fixed effects, are still valid.
- However, the EB estimates of the random effects may be highly affected by their distributional assumption. In particular, heterogeneity in the population random effects $\boldsymbol{b}$ may not be preserved in the shrunken $\hat{\boldsymbol{b}}$.
- Checking the normality assumption is tricky. In particular, histograms of the $\hat{\boldsymbol{b}}_i$ are not useful, because the $\hat{\boldsymbol{b}}_i$ are not identically distributed and they have smaller variance than the population $\boldsymbol{b}_i$ because of the shrinkage.
- A simulation study shows that even when the true $\boldsymbol{b}_i$ follows a bi-modal (mixture of two normals), $\hat{\boldsymbol{b}}_i$ shows a uni-model (normal) shape in histogram.
- A suggestion for design: if one is interested in detecting subgroups in the random-effect population, one should spread out the measurements as widely as possible (take as many measurements as possible at the beginning and at the end of the study). By doing so, one can increase within-subject variability and decrease shrinkage effect, and hence more likely be able to detect heterogeneity in random effects.

Extending Linear Mixed Models

The covariance matrix for the random effects D is generally assumed to be simply a symmetric positive definite matrix. Sometime it is desirable to use other type of matrices such as an identity matrix, a diagonal matrix, etc.

The covariance matrix for the errors is usually assumed to be iid random variables with mean zero and constant variance ($\sigma^2 I_{n_i}$), but it can be made more general (heteroscedastic or correlated or both):

$$\boldsymbol{Y}_i \,|\, \boldsymbol{b}_i \sim \mathcal{N}\left(X_i\boldsymbol{\beta} + Z\boldsymbol{b_i}, \sigma^2\boldsymbol{\Lambda}_i\right),$$

where $\boldsymbol{\Lambda}_i$ are positive-definite matrices parametrized by fixed parameters $\boldsymbol{\lambda}$.

- The error variance $\sigma^2\boldsymbol{\Lambda}_i$ can be decomposed additively into a *serial correlation* and a *measurement error* components, the latter being $\tau^2 I$ (DHLZ, 2002).
- Pinheiro and Bates (2000) decomposed $\boldsymbol{\Lambda}_i = B_i C_i B_i$, where B_i is diagonal and C_i is a correlation matrix. To ensure uniqueness, all elements in B_i are required to be positive.

  Thus

$$\begin{aligned}\mathrm{Var}(\epsilon_{ij}) &= \sigma^2[B_i]^2_{jj}\\ \mathrm{Corr}(\epsilon_{ij}, \epsilon_{ik}) &= [C_i]_{jk}.\end{aligned}$$

  This decomposed $\boldsymbol{\Lambda}_i$ into a *variance structure* component and a *correlation structure* component.

- Variance functions for modeling heteroscedasticity.

  A general variance function model (Davidian and Giltinan, 1995) is

$$\mathrm{Var}(\epsilon_{ij} \mid \boldsymbol{b}_i) = \sigma^2 g^2(\mu_{ij}, \boldsymbol{v}_{ij}, \boldsymbol{\delta}),$$

  where $\mu_{ij} = \mathrm{E}(y_{ij} \mid \boldsymbol{b}_i)$, $\boldsymbol{v}_{ij}$ is a vector of *variance covariates*, $\boldsymbol{\delta}$ is a vector of variance parameters and g is the variance function, continuous in $\boldsymbol{\delta}$.

  For example,

$$\mathrm{Var}(\epsilon_{ij} \mid \boldsymbol{b}_i) = \sigma^2 |v_{ij}|^{2\delta}.$$

  – The covariate v_{ij} can be the expected value μ_{ij}. Very intuitive because it allows the within-group variance to depend on the fixed effects and the random effects through μ_{ij}.

  – In practice μ_{ij} is replaced by $\hat{\mu}_{ij}$.

$$\mathrm{Var}(\epsilon_{ij} \mid \boldsymbol{b}_i) \simeq \sigma^2 g^2(\hat{\mu}_{ij}, \boldsymbol{v}_{ij}, \boldsymbol{\delta}).$$

    The estimation is done by iterating the following steps until convergence.

    1. given $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\lambda}^{(t)}$, estimate $\mu_{ij}^{(t)}$;
    2. given $\mu_{ij}^{(t)}$, estimate $\boldsymbol{\beta}^{(t+1)}$ and $\boldsymbol{\lambda}^{(t+1)}$.

  – When the variance model does not involve μ_{ij}, the likelihood (or restricted likelihood) can be directly optimized, producing the exact ML (or REML) estimates.

- Correlation functions for modeling dependence.

  The general within-group correlation structure is assumed to depend on the corresponding positions t_{ij} and $t_{ij'}$ only through their distance $d(t_{ij}, t_{ij'})$:

$$\text{corr}(\epsilon_{ij}, \epsilon_{ij'}) = h[d(t_{ij}, t_{ij'}), \rho],$$

  where ρ is a correlation parameter (or a vector of parameters) and h is a correlation function taking values between -1 and 1, assumed continuous in ρ, and such that $h(0, \rho) = 1$.

Further Reading: optional

- Chapter 8 of Fitzmaurice, Laird and Ware (2004).
- Chapters 6 and 7 of Verbeke and Molenberghs (2000).
- Chapters 1-5 of Pinheiro and Bates (2000).
- Laird NM and Ware JH. (1982). Random effects models for longitudinal data, *Biometrics*, **38**:963-74.
- Lindstrom MJ and Bates DM (1988). Newton-Raphson and EM algorithm for linear mixed-effects models for repeated-measurements data, *JASA*, **83**:1014-1022.

Supplement: Introduction to SAS PROC MIXED

A sample SAS code using PROC MIXED (Verbeke and Molenberghs, 2000):

```
proc mixed data=prostate method=reml asycov asycorr covtest ic;
class id group timeclass;
model lnspa = group age group*time age*time/noint solution ddfm=satterth covb chisq;
id id time;
random intercept time/type=un subject=id g gcorr v vcorr solution;
repeated timeclass/type=simple subject=id r rcorr;
contrast'Final model' age*time 1,
                    group*time 1 0 0 0/chisq;
estimate 'Diff L/R-BPH, t=5yr' group 0 -4 4 0
                               group*time 0 -2 2 0/cl alpha=0.05;
run;
```

- The PROC MIXED statement
  - This statement calls the procedure MIXED and specified the data is "prostate". If no data is specified, then the most recently created data is used.
  - The MIXED procedure requires the data in long format.
  - The option 'method=' specify the estimation method, either 'method=ML' or 'method=REML'. By default, REML is used.
  - The options of 'asycov' and 'asycorr' request the asymptotic covariance and the associated correlation matrix for the estimators for the variance component in the marginal model.

  - The option of 'covtest' requests the asymptotic standard errors and the associated Wald tests for those variance components.
  - The option of 'ic' requests the information criteria.
  - Another option of 'EMPIRICAL' can be used to request the "sandwich" estimator of variance-covariance matrix of the fixed-effects parameters and PROC MIXED adjusts all standard errors and test statistics involving the fixed-effects parameters.

- The CLASS statement
  This statement specifies which variables should be considered as factors.

- The MODEL statement
  - This statement names the response variable and all fixed effects.
  - The 'solution' option is used to request the estimates for all fixed effects.
  - The option of 'covb' requests model based estimated covariance matrix for the fixed effects.
  - The option of 'ddfm=' specifies the method for the degree of freedom in the $t-$ and $F-$approximations for testing the fixed effects. Here it requests the Satterthwaite approximation.
  - The option of 'chisq' request additional Wald tests.

- The ID statement
  The values of the variables in the ID statement will be provided in the table when the options of 'predmeans' or 'predicted' in the MODEL statement.

- The RANDOM statement

- This statement defines the random effects, i.e., matrix of $\boldsymbol{Z}_i$.
- The 'subject=' option is used to identify the subjects in the data set.
- The 'type=' option specifies the covariance structure $\boldsymbol{D}$ for the random effects $\boldsymbol{b}_i$.
- The options of 'g' and 'gcorr' request the random effects covariance matrix $\boldsymbol{D}$ as well as the associated correlation matrix.
- The options of 'v' and 'vcorr' request the marginal covariance matrix $\boldsymbol{V}_i$ as well as the associated correlation matrix.
- The option of 'solution' requests the empirical Bayes estimates for the random effects $\boldsymbol{b}_i$.

• The REPEATED statement

- This statement is used to specify the residual variance Σ_i.
- Very often one selects 'type=simple' which corresponding to the most simple covariance structure $\Sigma_i = \sigma^2 I_{n_i}$. If no REPEATED statement is specified, PROC MIXED fits such a 'simple' covariance for the residual components.

• The CONTRAST statement

This statement allows general linear hypotheses of the fixed effects using $F-$tests. The option of 'chisq' requests the additional approximate Wald tests.

• The ESTIMATE statement

This statement allows estimation and testing of linear combinations of the fixed effects. It can provide point estimate as well as confidence interval.