

Linear Mixed Models

Outline

- Motivation
- Linear Mixed Effects (LME) Model
- Multilevel Mixed Effects Models
- Estimation for LME
 - Inference for Fixed Effects
 - Inference for Variance Parameters
 - Inference about the Random Effects
 - Normality Assumption
- Extending Linear Mixed Models

Motivation

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

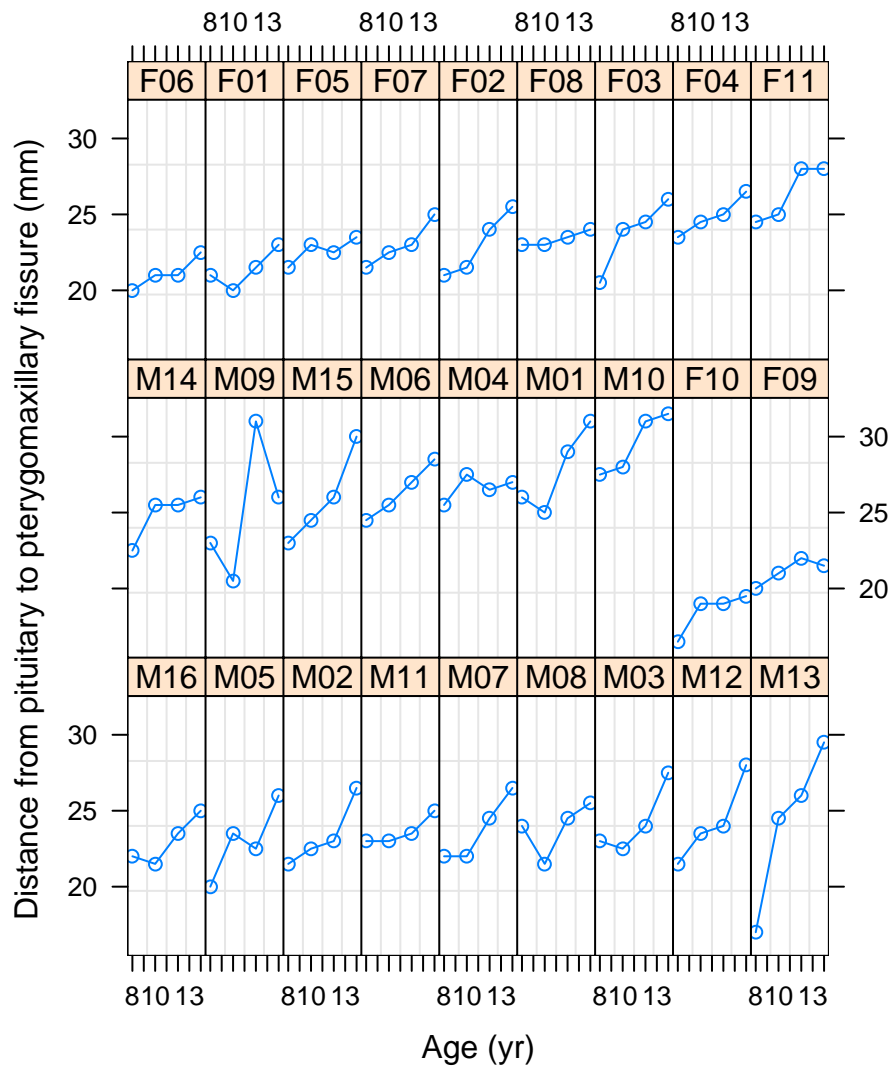


Figure 1: Orthodontic distance measurements

Three modeling strategies:

1. **Two-stage analysis:** fit a linear regression line to each subject and analyze the resulted slopes as responses in the second stage analysis.
2. **Fixed effect model:** include an indicator variable for subject id in the regression.
3. **Linear mixed model.**

Restriction of the Two-Stage Analysis

Let's consider only the girls for the moment.

```
> library (nlme)
> data (Orthodont)
> OrthFem <- subset (Orthodont, Sex == "Female")
> OrthFem[1:5,]
```

Grouped Data: distance ~ age | Subject

	distance	age	Subject	Sex
65	21.0	8	F01	Female
66	20.0	10	F01	Female
67	21.5	12	F01	Female
68	23.0	14	F01	Female
69	21.0	8	F02	Female

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

```
> of.lis <- lmList (distance ~ I(age - 11), data = OrthFem)
> coef (of.lis)
```

	(Intercept)	I(age - 11)
F10	18.500	0.450
F09	21.125	0.275
F06	21.125	0.375
F01	21.375	0.375
F05	22.625	0.275
F07	23.000	0.550
F02	23.000	0.800
F08	23.375	0.175
F03	23.750	0.850
F04	24.875	0.475
F11	26.375	0.675

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

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

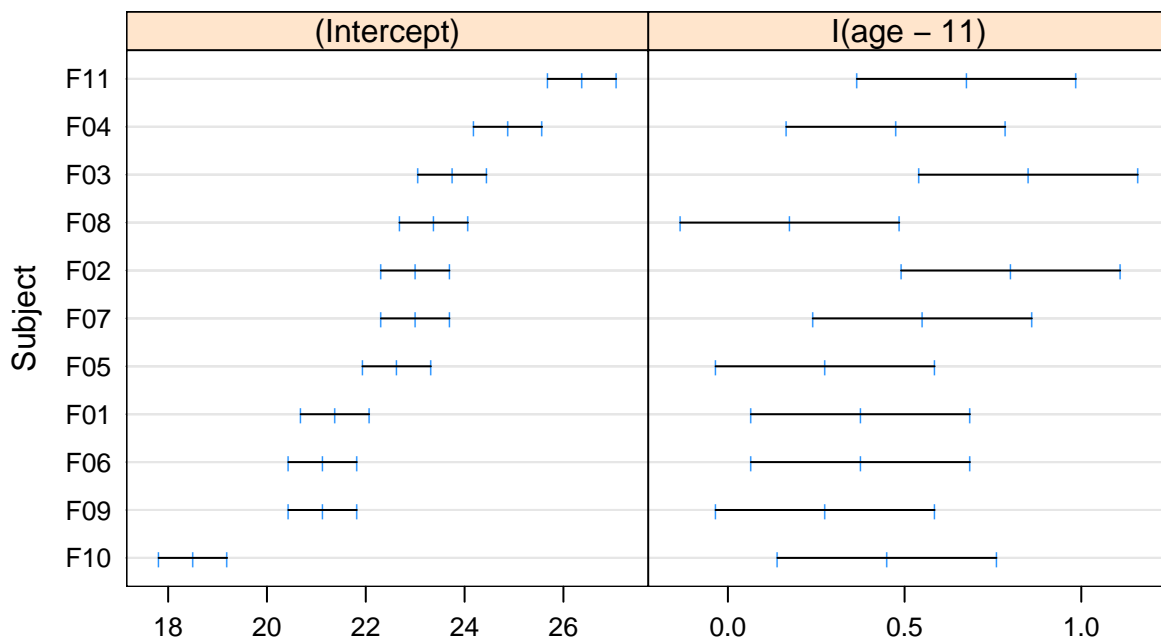


Figure 2: Confidence intervals (95%) for the coefficients of simple linear models.

- 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 is not balanced, then the individual growth curve parameters are estimated at different precision and we need weight them differently in the subsequent analysis.
- We already know the a forth method (which includes only 2 parameters) that ignores the correlation and uses OLS, is not very efficient.

Fixed Effects

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

```
> of.lm <- lm (distance ~ factor (Subject, ordered = FALSE) +
+           I(age - 11) - 1,
+           data = OrthFem)
> library (MASS)
> confint (of.lm)
              2.5 % 97.5 %
F10           17.71  19.29
F09           20.33  21.92
F06           20.33  21.92
F01           20.58  22.17
F05           21.83  23.42
F07           22.21  23.79
F02           22.21  23.79
F08           22.58  24.17
F03           22.96  24.54
F04           24.08  25.67
F11           25.58  27.17
I(age - 11)    0.37   0.59
> apply (intervals (of.lis)[, ,2], 2, mean)
      lower      est.      upper
0.1696451 0.4795455 0.7894458
```

- Here we are estimating 12 parameters (11 intercepts and one slope).
- The precision on the slope is substantially better (sd 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.

Linear Mixed Model

One solution is to use a random effect for subjects.

```
> of.lme <- lme (distance ~ I(age - 11),
+             random = ~ 1 | Subject, data = OrthFem)
> intervals (of.lme)
...
              lower    est.    upper
I(age - 11)  0.37242  0.47955  0.58667
...
> 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 CI for the slope are very close to the previous model.
- The std. dev. for the random effects (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.
- At first look, there is one 1 (variance) parameter for the intercepts instead of 11. But there is no free lunch.

Specification of Linear Mixed Models

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

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \quad (1)$$

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

\mathbf{Y}_i : ($n_i \times 1$) response vector

\mathbf{X}_i : ($n_i \times p$) design matrix for fixed effects

$\boldsymbol{\beta}$: ($p \times 1$) regression coefficients for fixed effects

\mathbf{Z}_i : ($n_i \times q$) design matrix for random effects

\mathbf{b}_i : ($q \times 1$) random effects

$\boldsymbol{\epsilon}_i$: ($n_i \times 1$) error vector

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

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, D)$$

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

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

$$\mathbf{Y}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, V(\boldsymbol{\alpha})) \quad (2)$$

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

- D must be symmetric and positive definite.
- Suppose that we have one covariate and $\mathbf{X}_i = \mathbf{Z}_i = (\mathbf{1}_i, \mathbf{X}_i)$ (random intercept and random slope model), then we can write:

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

or

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

where

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

and

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

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

- 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 \mathbf{Z}_i are typically a subset of the columns in \mathbf{X}_i . In particular, $\mathbf{Z}_i = \mathbf{1}_i$ corresponds to the random intercept model.
- The marginal mean for \mathbf{Y} is the same as in the marginal general linear model:

$$\begin{aligned} \mathbf{E}(\mathbf{Y}_i) &= \mathbf{E}(\mathbf{E}(\mathbf{Y}_i | \mathbf{b}_i)) \\ &= \mathbf{E}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i) \\ &= \mathbf{X}_i\boldsymbol{\beta}. \end{aligned}$$

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

Multilevel Mixed Effects Models

The hierarchical 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 $k = 1, \dots, K$ to index the group, and $i = 1, \dots, m_k$ for individuals in group k , $j = 1, \dots, n_i$ for observational times for individual i . The model can be written as:

$$\mathbf{Y}_{ki} = \mathbf{X}_{ki}\boldsymbol{\beta} + \mathbf{Z}_{k,i}\mathbf{b}_k + \mathbf{Z}_{ki}\mathbf{b}_{ki} + \boldsymbol{\epsilon}_{ki}, \quad (3)$$

where

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

- The first-level random effects \mathbf{b}_k , of length q_1 are assumed to be independent for different k (groups).
- The second-level random effects \mathbf{b}_{ki} , of length q_2 are assumed to be independent for different k (groups) or i (individuals), and of the \mathbf{b}_k .
- $\boldsymbol{\epsilon}_{ki}$ are independent for k , i , and independent of the random effects.
- Some people would call this a three-level model as in

$$\begin{aligned} \mathbf{Y}_i &\sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}_i, \sigma^2 I) \\ \boldsymbol{\beta}_i &\sim \mathcal{N}(\mathbf{Z}_{i1}\boldsymbol{\gamma}_{i1}, \tau_1^2 I) \\ \boldsymbol{\gamma}_{i1} &\sim \mathcal{N}(\mathbf{Z}_{i2}\boldsymbol{\gamma}_{i2}, \tau_2^2 I). \end{aligned}$$

Estimation for LME

- In the original paper, an EM 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).
- However, in general linear model specification, we only require $V(\boldsymbol{\alpha})$ to be symmetric and positive definite. So often the maximization is done in this slightly larger parameter space. Not very well specified (over-specified) models can result in instability in the estimates of the variance parameter.
- In particular, `lme()` uses a mixed EM (expectation-maximization) 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.

Inference for Fixed Effects in LME

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

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

has an approximate χ^2 distribution with $\text{rank}(L)$ degrees of freedom.

- Wald test tends to be anti-conservative. For small samples, the F test statistic:

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

has an approximate F distribution with numerator degrees of freedom $\text{rank}(L)$ and denominator degrees of freedom estimated from the data.

- Empirical variance estimates for $\hat{\boldsymbol{\beta}}$ can also be used (`PROC MIXED`, not supported in `lme`).
- The Wald test is *conditional* in that the parameter estimation is done once and the test for the certain coefficients is done conditional on the estimates for the other, nuisance, parameters in the model.
- For nested models, likelihood ratio tests can also be used (there the estimation has to be done using ML instead of REML). But Pinheiro and Bates (2000) showed that the LRT tests tend to be anticonservative (p -values too small) for small samples.
- Confidence intervals for regression coefficients can be constructed using the t distribution.
- Methods based on multivariate t -distribution and bootstrap are in development.

Inference for Variance Parameters in LME

- The MLEs for regression coefficients β and variance parameters α are asymptotically uncorrelated.
- Variance estimates for $\hat{\alpha}$ are computed using the inverse observed information matrix $(-\ddot{\ell})$.
- Confidence intervals for variance parameters can be tricky because it is often of interest to calculate CIs on the original parameters in (σ^2, D) and not the marginal parameters α .
- When the true parameter value is on the boundary of the parameter space (i.e., $\sigma = 0$), Wald test is not valid.
- Likelihood ratio test can be used to compare nested models with different variance parameters. Both ML and REML can be used.
- However, when one of the model sets some parameters at 0, which is 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_1^2 , or $\sim 0.5\chi_0^2 + 0.5\chi_1^2$.
 - A random intercept and slope model:

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

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_1^2 and χ_2^2 .

- 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_q^2 and χ_{q+1}^2 .
- 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.
- In Fitzmaurice, Laird and Ware (2004), they recommended using 0.1 significant level instead of 0.05 to compensate some of the conservativeness, that leads in overly simple 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:

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}} | \mathbf{y}) + 2n_{\text{par}}$$

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

- The REML versions of AIC and BIC replaces $\ell(\hat{\boldsymbol{\theta}} | \mathbf{y})$ with $\ell_R(\hat{\boldsymbol{\theta}} | \mathbf{y})$ 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 \mathbf{b}_i are *random variables*, not parameters. Technically, we *predict* the random effects, not estimate them.

BLUP

- For arbitrary vectors \mathbf{s} and \mathbf{t} , the *best linear unbiased predictors* (BLUP) $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{b}}$ minimize the prediction error:

$$E \left\{ \left(\mathbf{s}^T \mathbf{X} \tilde{\boldsymbol{\beta}} + \mathbf{t}^T \mathbf{Z} \tilde{\mathbf{b}} \right) - \left(\mathbf{s}^T \mathbf{X} \boldsymbol{\beta} + \mathbf{t}^T \mathbf{Z} \mathbf{b} \right) \right\}^2,$$

subject to the unbiasedness condition

$$E \left(\mathbf{s}^T \mathbf{X} \tilde{\boldsymbol{\beta}} + \mathbf{t}^T \mathbf{Z} \tilde{\mathbf{b}} \right) = E \left(\mathbf{s}^T \mathbf{X} \boldsymbol{\beta} + \mathbf{t}^T \mathbf{Z} \mathbf{b} \right).$$

It can be shown that the solutions are

$$\begin{aligned} \text{BLUP}(\boldsymbol{\beta}) &= \tilde{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}) \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y} \\ \text{BLUP}(\mathbf{b}) &= \tilde{\mathbf{b}} = \mathbf{D} \mathbf{Z}^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \tilde{\boldsymbol{\beta}}). \end{aligned}$$

- Note that $\tilde{\boldsymbol{\beta}}$ is identical to the general least squares estimator, which is also the *best linear unbiased estimator* (BLUE).
- The BLUP for \mathbf{b} is the *best linear predictor* (BLP)

$$\text{BLP}(\mathbf{b}) = \mathbf{D} \mathbf{Z}^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta})$$

with $\boldsymbol{\beta}$ replaced by $\tilde{\boldsymbol{\beta}}$.

- \mathbf{D} and \mathbf{V} have to be estimated, i.e., via ML or REML. In practice the BLUPs are replaced by the *empirical BLUPs* or EBLUPs

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X}) \mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{y} \\ \hat{\mathbf{b}} &= \hat{\mathbf{D}} \mathbf{Z}^T \hat{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}). \end{aligned}$$

- One simple (ad-hoc) justification of BLUP is due to Henderson (1959) that involves making the distributional assumption:

$$\begin{aligned} \mathbf{Y} | \mathbf{b} &\sim \mathcal{N}(\mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \mathbf{b}, \mathbf{R}) \\ \mathbf{b} &\sim \mathcal{N}(\mathbf{0}, \mathbf{D}), \end{aligned}$$

and maximize the likelihood of the (\mathbf{y}, \mathbf{b}) over the unknowns $\boldsymbol{\beta}$ and \mathbf{b} . This leads to the criterion

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})^T R(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}) + \mathbf{b}^T D\mathbf{b}.$$

This shows that BLUP estimation of $(\boldsymbol{\beta}, \mathbf{b})$ involves general least squares with a penalty term, hence is related to *ridge regression*.

Empirical Bayes and Shrinkage

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

$$f(\mathbf{b} | \mathbf{y}) = \frac{f(\mathbf{y} | \mathbf{b})f(\mathbf{b})}{\int f(\mathbf{y} | \mathbf{b})f(\mathbf{b})d\mathbf{b}},$$

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

$$\begin{aligned} \hat{\mathbf{b}}(\boldsymbol{\theta}) &= E(\mathbf{b} | \mathbf{y}) \\ &= \int \mathbf{b}f(\mathbf{b} | \mathbf{y})d\mathbf{b} \\ &= DZ^T V^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

In practice, unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are replaced by their ML or REML estimates. The resulting estimates for the random effects are called *empirical Bayes* (EB) estimates.

- Consider the prediction of the response for i ,

$$\begin{aligned} \hat{\mathbf{Y}}_i &= \mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_i \hat{\mathbf{b}} \\ &= \mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_i DZ_i^T V_i^{-1}(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}) \\ &= (\mathbf{I}_{n_i} - \mathbf{Z}_i DZ_i^T V_i^{-1}) \mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_i DZ_i^T V_i^{-1} \mathbf{y}_i \\ &= \Sigma_i V_i^{-1} \mathbf{X}_i \hat{\boldsymbol{\beta}} + (\mathbf{I}_{n_i} - \Sigma_i V_i^{-1}) \mathbf{y}_i \end{aligned}$$

where the residual variance

$$\Sigma_i = V_i - \mathbf{Z}_i DZ_i^T.$$

- It can be interpreted as a weighted average of the population mean $\mathbf{X}\hat{\boldsymbol{\beta}}$ and the observed data \mathbf{y}_i .
- Larger weights are given the overall mean if the residual variability Σ_i is large in comparison with the between-subject variability $\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T$.
- In Bayesian literature (Carlin and Louis, 1996) it is referred to as shrinkage. The observed data are shrunken toward the prior mean, which is $\mathbf{X}_i\boldsymbol{\beta}$ since the prior mean of the random effects was zero.

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 \mathbf{b} may not be preserved in the shrunken $\hat{\mathbf{b}}$.
- Checking the normality assumption is tricky. In particular, histograms of the $\hat{\mathbf{b}}_i$ are not useful, because the $\hat{\mathbf{b}}_i$ are not identically distributed and they have smaller variance than the population \mathbf{b}_i because of the shrinkage.

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.. In particular, the two-level random effects model can be written as a one-level random effects with block compound symmetry covariance matrix.
- The covariance matrix for the errors can be made general:

$$\mathbf{Y}_i | \mathbf{b}_i \sim \mathcal{N}(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z} \mathbf{b}_i, \sigma^2 \boldsymbol{\Lambda}_i),$$

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

- DHLZ further decomposed the error variance $\sigma^2 \boldsymbol{\Lambda}_i$ into a *serial correlation* and a *measurement error* components, the latter being $\tau^2 I$.
- Pinheiro and Bates (2000) decomposed

$$\boldsymbol{\Lambda}_i = \mathbf{B}_i \mathbf{C}_i \mathbf{B}_i,$$

where \mathbf{B}_i is diagonal and \mathbf{C}_i is a correlation matrix. To ensure uniqueness, all elements in \mathbf{B}_i are required to be positive.

Thus

$$\begin{aligned} \text{Var}(\epsilon_{ij}) &= \sigma^2 [\mathbf{B}_i]_{jj}^2 \\ \text{Corr}(\epsilon_{ij}, \epsilon_{ik}) &= [\mathbf{C}_i]_{jk}. \end{aligned}$$

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

Correlation Structure

- Compound symmetry:

$$\mathbf{C}_i = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \dots & \dots & 1 \end{pmatrix}_{n_i \times n_i} .$$

- Banded (width 1):

$$\mathbf{C}_i = \begin{pmatrix} 1 & \rho & 0 & \dots & 0 \\ \rho & 1 & \rho & \dots & 0 \\ 0 & \rho & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & \rho & 1 \end{pmatrix}_{n_i \times n_i} .$$

- Exponential (a special case of discrete time AR(1) model) and Gaussian correlated models.
- Autoregressive-moving average (ARMA) models (time series).
- Spatial correlation (e.g., CAR).

Variance Structure

A general variance function model is

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

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

For example,

$$\text{Var}(\epsilon_{ij} | \mathbf{b}_i) = \sigma^2 |v_{ij}|^{2\delta}.$$

- In practice μ_{ij} is replaced by $\hat{\mu}_{ij}$. The estimation is done by iterating the following steps until convergence.
 - given $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\lambda}^{(t)}$, estimate $\mu_{ij}^{(t)}$;
 - given $\mu_{ij}^{(t)}$, estimate $\boldsymbol{\beta}^{(t+1)}$ and $\boldsymbol{\lambda}^{(t+1)}$.

Further Reading

- Chapter 8 of Fitzmaurice, Laird and Ware (2004).
- Chapter 4 of Hedeker and Gibbons (2006).
- Chapters 6 and 7 of Verbeke and Molenberghs (2000).
- Laird, NM and Ware, JH (1982) Random effects models for longitudinal data, *Biometrics*, **38**:963-74.