4. Estimation in the longitudinal LMM

Sonja Greven

Summer Term 2017

Overview Chapter 4 - Estimation in the longitudinal LMM

4.1 The marginal model

4.2 Estimation of the fixed effects

4.3 Estimation of the covariance parameters

4.4 Numerical calculation of the estimates

4.5 Prediction of the random effects

The marginal model

Estimation is usually based on the marginal model. The longitudinal linear mixed model (3.5)

$$\begin{cases} \mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \\ \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}_q, \mathbf{D}) \\ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}_{n_i}, \boldsymbol{\Sigma}_i) \\ \mathbf{b}_1, \ldots, \mathbf{b}_N, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N \text{ independent} \end{cases}$$

implies the marginal model (3.8)

$$\mathbf{Y}_i \ \sim \ \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \boldsymbol{\Sigma}_i), \text{ for } i = 1, \ldots, N.$$

Remember that (3.5) and (3.8) are not equivalent. For example, in the random intercept model with independent errors (see slides 34-35 in Chapter 3),

- d^2 is a variance in $\boldsymbol{D}$ for (3.5) and should be non-negative
- in the marginal model (3.8), Cov$(Y_{ij}, Y_{ik}) = d^2 + \sigma^2 I(j = k)$. Thus, d^2 is a covariance and can be negative as long as $\boldsymbol{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \boldsymbol{\Sigma}_i = (d^2 + \sigma^2 I(j = k))_{j,k=1,\dots,n_i}$ is positive definite.

This has implications for the parameter space. More later.

The covariance parameters

Let $\boldsymbol{\alpha}$ be the vector of variance and covariance parameters defining

$$\mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^T + \boldsymbol{\Sigma}_i,$$

i.e.

- the different elements in $\mathbf{D}$ ($q(q+1)/2$ if there are no further assumptions)
- all parameters in $\boldsymbol{\Sigma}_i$, e.g.
  - σ^2 if $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}$
  - ϕ and τ^2 if $\boldsymbol{\Sigma}_i = \tau^2(\exp(-\phi|t_{ij} - t_{ik}|))_{j,k=1,\ldots,n_i}$.

We then write $\mathbf{V}_i(\boldsymbol{\alpha})$ to stress the dependence on $\boldsymbol{\alpha}$ and $\mathbf{V}(\boldsymbol{\alpha}) = \text{diag}(\mathbf{V}_1(\boldsymbol{\alpha}), \ldots, \mathbf{V}_N(\boldsymbol{\alpha}))$ for the covariance matrix of $\boldsymbol{Y}$.

Notation

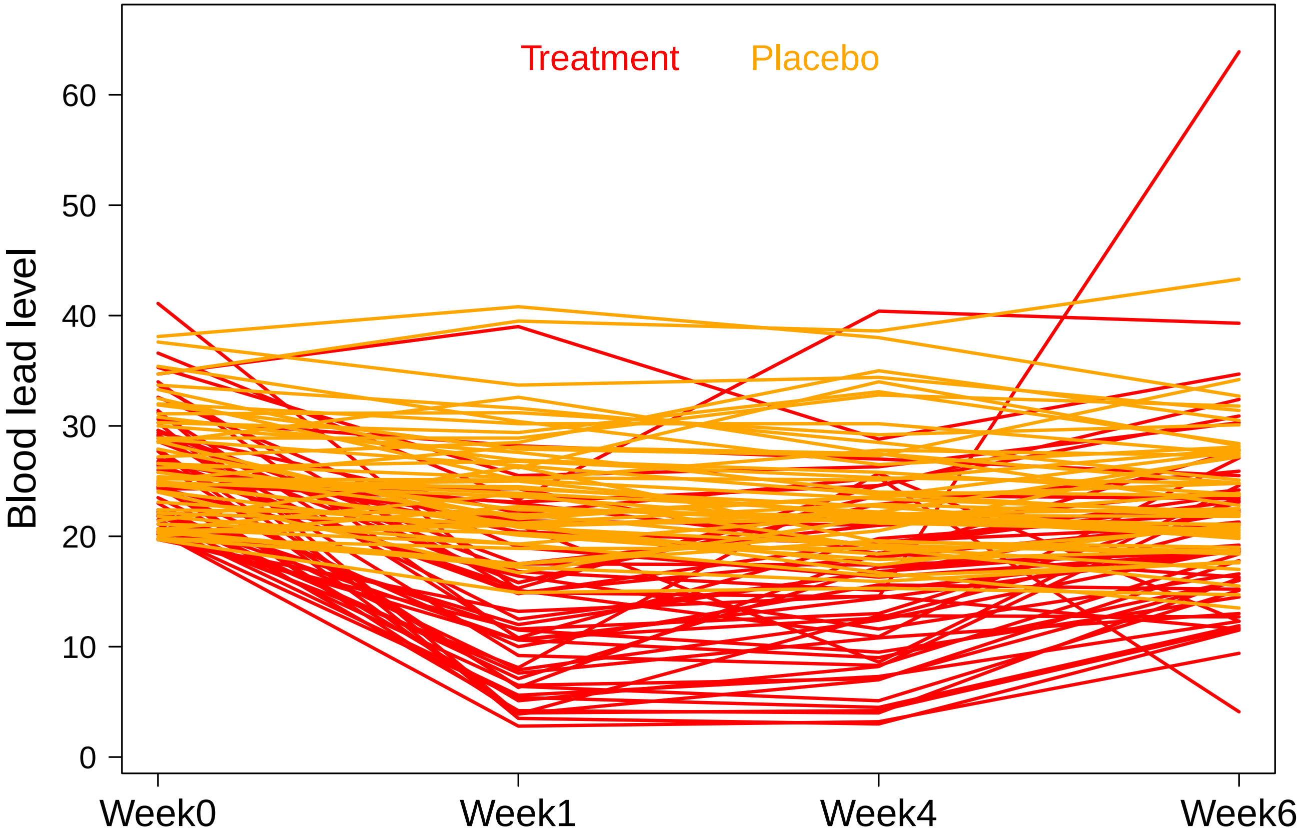
We denote the vector of length s containing all parameters in the marginal model as $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix}.$$

Let $\Theta = \Theta_{\boldsymbol{\beta}} \times \Theta_{\boldsymbol{\alpha}}$ be the parameter space for $\boldsymbol{\theta}$ with

- $\Theta_{\boldsymbol{\beta}} = \mathbf{R}^p$: parameter space for $\boldsymbol{\beta}$
- $\Theta_{\boldsymbol{\alpha}}$: The set of $\boldsymbol{\alpha}$ values resulting in a positive semi-definite matrix $\mathbf{D}$ and positive definite matrices $\boldsymbol{\Sigma}_i$ $(i = 1, \ldots, N)$ for the linear mixed model (3.5) or, for the marginal model (3.8), those resulting in positive-definite matrices $\mathbf{V}_i$.

Example: The TLC trial



Example: The TLC trial

For the TLC trial, consider the following model

$$\begin{aligned} Y_{ij} = &\quad \beta_0 + \quad \beta_1 I(t_j = 1) + \quad \beta_2 I(t_j = 4) + \quad \beta_3 I(t_j = 6) \\ &+ \beta_4 g_i + \beta_5 g_i I(t_j = 1) + \beta_6 g_i I(t_j = 4) + \beta_7 g_i I(t_j = 6) + b_i + \epsilon_{ij} \end{aligned}$$

where

- Y_{ij} is the lead blood level for child i in week t_j
- $t_j \in \{0, 1, 4, 6\}$ indicates the week
- $g_i = 1$ if child i is in the succimer group and $= 0$ if in the placebo group

Interpretation?

Example: The TLC trial

Then, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7) \in \Theta_\beta = \mathbb{R}^8$ and if we assume

$$b_i \overset{i.i.d.}{\sim} \mathcal{N}(0, d^2) \quad \text{independent of} \quad \epsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

we have $\boldsymbol{\alpha} = (d^2, \sigma^2)$ with $d^2 \geq 0$, $\sigma^2 > 0$ and $\Theta_\alpha = [0, \infty) \times (0, \infty)$ in the hierarchical model or

$$\begin{pmatrix} d^2 + \sigma^2 & d^2 & \dots & d^2 \\ d^2 & d^2 + \sigma^2 & \dots & d^2 \\ \vdots & \vdots & \ddots & \vdots \\ d^2 & d^2 & \dots & d^2 + \sigma^2 \end{pmatrix}$$

positive definite in the marginal model with $\Theta_\alpha = \{(d^2, \sigma^2) : \boldsymbol{V} > \mathbf{0}\}$.

Overview

4.1 The marginal model

4.2 Estimation of the fixed effects

4.3 Estimation of the covariance parameters

4.4 Numerical calculation of the estimates

4.5 Prediction of the random effects

Estimation of the fixed effects

For the marginal model $\mathbf{Y}_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i(\boldsymbol{\alpha}))$, the **marginal likelihood** and log-likelihood are given by

$$\begin{aligned} L_{ML}(\boldsymbol{\theta}) &= \prod_{i=1}^{N} \Big\{ (2\pi)^{-n_i/2} |\mathbf{V}_i(\boldsymbol{\alpha})|^{-\frac{1}{2}} \\ &\quad \times \exp\left(-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i(\boldsymbol{\alpha})^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \right) \Big\} \\ \ell_{ML}(\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N} \log|\mathbf{V}_i(\boldsymbol{\alpha})| \\ &\quad -\frac{1}{2}\left\{ \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T \mathbf{V}_i(\boldsymbol{\alpha})^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \right\} \end{aligned}$$

or

$$\ell_{ML}(\boldsymbol{\theta}) \quad = \quad -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}(\boldsymbol{\alpha})| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}(\boldsymbol{\alpha})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Now consider first estimation of the fixed effects. Taking the derivative with respect to (w.r.t.) $\boldsymbol{\beta}$ yields

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\beta}}\ell_{ML}(\boldsymbol{\theta}) &= \boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\alpha})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \overset{!}{=} \mathbf{0} \\
\Rightarrow \widehat{\boldsymbol{\beta}}_{ML}(\boldsymbol{\alpha}) &= \left(\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\alpha})^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{\alpha})^{-1}\boldsymbol{y} \\
&= \left\{\sum_{i=1}^{N}\left(\mathbf{X}_i^T\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\mathbf{X}_i\right)\right\}^{-1}\sum_{i=1}^{N}\left(\mathbf{X}_i^T\mathbf{V}_i(\boldsymbol{\alpha})^{-1}\boldsymbol{y}_i\right).
\end{aligned}$$

Estimation of the fixed effects

- Thus, for known $\boldsymbol{\alpha}$, the maximum likelihood (ML) estimator for $\boldsymbol{\beta}$ corresponds to the generalized least squares (GLS) estimator. It minimizes the weighted least squares criterion

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

  with $\mathbf{W} = \mathbf{V}(\boldsymbol{\alpha})^{-1}$.

- We here assumed that the inverses $\mathbf{V}_i(\boldsymbol{\alpha})^{-1}$ and $\left\{\sum_{i=1}^{N} \left(\mathbf{X}_i^T \mathbf{V}_i(\boldsymbol{\alpha})^{-1} \mathbf{X}_i\right)\right\}^{-1}$ exist. Generalizations exist using generalized inverses.

- $\widehat{\boldsymbol{\beta}}_{ML}(\boldsymbol{\alpha})$ is the best linear unbiased estimator (BLUE).

- $\widehat{\boldsymbol{\beta}}_{ML}(\boldsymbol{\alpha})$ still depends on $\boldsymbol{\alpha}$, which needs to be estimated.

Overview

4.1 The marginal model

4.2 Estimation of the fixed effects

4.3 Estimation of the covariance parameters

4.4 Numerical calculation of the estimates

4.5 Prediction of the random effects

ML estimation of α

Substituting $\widehat{\boldsymbol{\beta}}_{ML}(\boldsymbol{\alpha})$ into the log-likelihood $\ell_{ML}(\boldsymbol{\theta}) = \ell_{ML}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ gives the profile log-likelihood $\ell_{ML}(\widehat{\boldsymbol{\beta}}_{ML}(\boldsymbol{\alpha}), \boldsymbol{\alpha})$, which depends only on $\boldsymbol{\alpha}$.

This can be maximized (numerically) to obtain the ML estimate $\widehat{\boldsymbol{\alpha}}_{ML}$ for $\boldsymbol{\alpha}$. The maximum likelihood estimate for $\boldsymbol{\theta}$ then is $\widehat{\boldsymbol{\theta}}_{ML} = (\widehat{\boldsymbol{\beta}}_{ML}(\widehat{\boldsymbol{\alpha}}_{ML}), \widehat{\boldsymbol{\alpha}}_{ML})$.

ML estimators are known to be biased downwards for variances $\rightarrow$ **restricted** or **residual maximum likelihood (REML)** estimation.

REML estimation - Motivation

- We want to estimate the variance of a normal distribution from $Y_1, \ldots, Y_N$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$ variables.

- For known mean μ the ML estimator is

$$\widehat{\sigma}^2_{ML} = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \mu)^2 .$$

- If for unknown mean μ is replaced by $\widehat{\mu} = \overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$ in this formula, the resulting ML estimator $\widehat{\sigma}^2$ is biased downwards:

$$E(\widehat{\sigma}^2_{ML}) = \frac{N-1}{N} \sigma^2 < \sigma^2 .$$

REML estimation: Error contrasts

- For this reason the unbiased estimator $\widehat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \widehat{\mu})^2$ is usually used. We will see that this corresponds to the REML estimator.

- As the Y_i are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, the vector $\mathbf{Y} = (Y_1, \ldots, Y_N)^T$ has a multivariate normal distribution

$$\mathbf{Y} \sim \mathcal{N}(\mu \mathbf{1}_N, \sigma^2 \mathbf{I}_N),$$

  where $\mathbf{1}_N$ is a vector of ones of length N.

- **Idea:** We want to estimate σ^2 "directly", without estimating μ first.

- How can we do this? By using a transformation of $Y_1, \ldots, Y_N$ that "eliminates" μ.

REML estimation: Error contrasts

- A linear transformation $$\mathbf{U} = \mathbf{A}^T\mathbf{Y}$$ is called an **error contrast** if $\mathbf{A}$ is an $N \times (N-1)$ matrix with linearly independent columns that are orthogonal to $\mathbf{1}_N$.

- The random vector $\mathbf{U}$ has a normal distribution with mean $$\mathsf{E}(\mathbf{U}) = \mathbf{A}^T\mathsf{E}(\boldsymbol{Y}) = \mathbf{A}^T(\mu\mathbf{1}_N) = \mu\mathbf{A}^T\mathbf{1}_N = \mathbf{0}_{N-1}$$ by construction and with variance $\mathsf{Var}(\mathbf{U}) = \sigma^2\mathbf{A}^T\mathbf{A}$. This distribution does not involve μ and the likelihood based on $\mathbf{U}$ can thus be directly maximized to obtain an estimate for σ^2.

REML estimation: Error contrasts

- $$\mathbf{U} = \mathbf{Y} - \mathbf{1}_N(\frac{1}{N}\mathbf{1}_N^T\mathbf{Y}) = (\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T)\mathbf{Y} =: \mathbf{A}^T\boldsymbol{Y}$$
  corresponding to the centered variables $Y_1 - \hat{\mu}, \ldots, Y_N - \hat{\mu}$ fulfills the requirements in our $\mathcal{N}(\mu, \sigma^2)$ example besides being of size $N \times N$.

- Only $N - 1$ columns of $\mathbf{A}$ can be linearly independent due to the loss in degrees of freedom from estimating $\hat{\mu}$. Deleting any column of $\mathbf{A}$ then defines $N - 1$ linearly independent error contrasts.

- One can show that maximizing the likelihood based on $\mathbf{U} \sim \mathcal{N}(\mathbf{0}_N, \sigma^2\mathbf{A}^T\mathbf{A})$ yields $\hat{\sigma}^2_{REML} = \frac{N}{N-1}\hat{\sigma}^2_{ML} = \frac{1}{N-1}\sum_{i=1}^N(Y_i - \widehat{\mu})^2$ and that this unbiased estimator is independent of the particular error contrast chosen.

REML estimation for the linear regression model

- In the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ the ML estimator for σ^2 is given by

$$\widehat{\sigma}^2_{ML} = \frac{1}{N}(\mathbf{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y})^T(\mathbf{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}).$$

- This estimator is biased downwards as the p regression coefficients $\beta_1, \ldots, \beta_p$ are unknown and replaced by estimates.

- Defining error contrast $\mathbf{U} = \mathbf{A}^T\mathbf{Y}$, where the $N \times (N - p)$ matrix $\mathbf{A}$ has $N - p$ linearly independent columns that are orthogonal to $\mathbf{X}$, and maximizing the likelihood based on $\mathbf{U}$ yields the commonly used and unbiased REML estimator $\widehat{\sigma}^2_{REML} = \frac{N}{N-p}\widehat{\sigma}^2_{ML}$ for σ^2.

REML estimation for the linear mixed model

In linear mixed models the likelihood based on error contrasts (with columns in $\mathbf{A}$ again orthogonal to $\mathbf{X}$) can be written as

$$\begin{aligned} L_{REML}(\boldsymbol{\alpha}) &= (2\pi)^{-(n-p)/2} \left| \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{X}_i \right|^{1/2} \\ &\quad \times \left| \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i(\boldsymbol{\alpha})^{-1} \mathbf{X}_i \right|^{-1/2} \prod_{i=1}^{N} |\mathbf{V}_i(\boldsymbol{\alpha})|^{-1/2} \\ &\quad \times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha}))^T \mathbf{V}_i(\boldsymbol{\alpha})^{-1} (\mathbf{y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha})) \right\}. \end{aligned}$$

This result is independent of the particular choice of the error contrasts.

REML estimation for the linear mixed model

- Thus,

$$L_{REML}(\boldsymbol{\alpha}) = const \left| \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i(\boldsymbol{\alpha})^{-1} \mathbf{X}_i \right|^{-1/2} L_{ML}(\widehat{\beta}(\boldsymbol{\alpha}), \boldsymbol{\alpha}),$$

  with $L_{ML}(\widehat{\beta}(\boldsymbol{\alpha}), \boldsymbol{\alpha})$ the profile likelihood and $const$ a constant not depending on $\boldsymbol{\alpha}$.

- REML can be used to estimate $\boldsymbol{\alpha}$ by $\widehat{\boldsymbol{\alpha}}_{REML}$, but not $\boldsymbol{\beta}$. (The whole point of REML is that $L_{REML}(\boldsymbol{\alpha})$ does not depend on $\boldsymbol{\beta}$!)

REML estimation for the linear mixed model

- REML is mainly used to reduce the downwards bias when estimating $\boldsymbol{\alpha}$ compared to ML estimation. It is however not guaranteed that the mean squared error (MSE) is also reduced.

- As the estimator for the fixed effects $\boldsymbol{\beta}$ depends on $\boldsymbol{\alpha}$, the two estimators $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\alpha}}_{REML})$ and $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\alpha}}_{ML})$ are in general not identical.

- $L_{REML}(\boldsymbol{\alpha})$ is the likelihood of the error contrasts $\mathbf{U} = \mathbf{A}^T\mathbf{Y}$, where $\mathbf{A}$ is chosen to be orthogonal to $\boldsymbol{X}$. Thus, for two models with different design matrices $\boldsymbol{X}$, the resulting error contrasts $\mathbf{U}$ are different and the REML likelihoods $L_{REML}(\boldsymbol{\alpha})$ are not comparable ("comparing apples and oranges"). This will be important for inference, see Chapter 5.

Estimation of β and α in the linear mixed model: Summary

- For a given value $\boldsymbol{\alpha}$, $\widehat{\boldsymbol{\beta}}$ is given by the ML (or GLS) estimate

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \left\{ \sum_{i=1}^{N} \mathbf{X}_i^T \mathbf{V}_i(\boldsymbol{\alpha})^{-1} \mathbf{X}_i \right\}^{-1} \sum_{i=1}^{N} \left(\mathbf{X}_i^T \mathbf{V}_i(\boldsymbol{\alpha})^{-1} \mathbf{y}_i \right).$$

- $\boldsymbol{\alpha}$ is estimated by maximizing the profile likelihood $L_{ML}(\widehat{\beta}(\boldsymbol{\alpha}), \boldsymbol{\alpha})$ (ML) or restricted likelihood $L_{REML}(\boldsymbol{\alpha})$ (REML).

- The estimates for $\boldsymbol{\beta}$ are obtained by plugging in the estimates for $\boldsymbol{\alpha}$. This yields the so-called **empirical BLUEs** $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\alpha}}_{REML})$ or $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\alpha}}_{ML})$.

Example: The TLC trial

For the TLC trial, we can fit the linear mixed model from 4.1

$$\begin{aligned} Y_{ij} = {} & \beta_0 + \beta_1 I(t_j = 1) + \beta_2 I(t_j = 4) + \beta_3 I(t_j = 6) \\ + {} & \beta_4 g_i + \beta_5 g_i I(t_j = 1) + \beta_6 g_i I(t_j = 4) + \beta_7 g_i I(t_j = 6) + b_i + \epsilon_{ij} \end{aligned}$$

using

```
lmeREML <- lme(lead ~ group * week, random = ~ 1 | id,
                          data = lead)  # REML is the default
lmeML <- lme(lead ~ group * week, random = ~ 1 | id,
                          data = lead, method = "ML")
```

where `group` and `week` are factors for the treatment groups and weeks, respectively.

Example: The TLC trial - REML results

```
Random effects:
 Formula: ~1 | id
        (Intercept) Residual
StdDev:    5.112717 4.214287

Fixed effects: lead ~ group * week
               Value Std.Error  DF   t-value p-value
(Intercept)   26.272 0.9370175 294 28.037898  0.0000
groupS         0.268 1.3251428  98  0.202242  0.8401
week1         -1.612 0.8428574 294 -1.912542  0.0568
week4         -2.202 0.8428574 294 -2.612542  0.0094
week6         -2.626 0.8428574 294 -3.115592  0.0020
groupS:week1 -11.406 1.1919804 294 -9.568950  0.0000
groupS:week4  -8.824 1.1919804 294 -7.402807  0.0000
groupS:week6  -3.152 1.1919804 294 -2.644339  0.0086
```

Example: The TLC trial - ML results

```
Random effects:
 Formula: ~1 | id
        (Intercept) Residual
StdDev:    5.061331 4.171931

Fixed effects: lead ~ group * week
                Value Std.Error  DF   t-value p-value
(Intercept)    26.272 0.9370175 294 28.037898  0.0000
groupS          0.268 1.3251428  98  0.202242  0.8401
week1          -1.612 0.8428574 294 -1.912542  0.0568
week4          -2.202 0.8428574 294 -2.612542  0.0094
week6          -2.626 0.8428574 294 -3.115592  0.0020
groupS:week1  -11.406 1.1919804 294 -9.568950  0.0000
groupS:week4   -8.824 1.1919804 294 -7.402807  0.0000
groupS:week6   -3.152 1.1919804 294 -2.644339  0.0086
```

Overview

4.1 The marginal model

4.2 Estimation of the fixed effects

4.3 Estimation of the covariance parameters

4.4 Numerical calculation of the estimates

4.5 Prediction of the random effects

Numerical calculation of the estimates

How are the estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ numerically computed?

- The historically first approach was the expectation maximization (EM, Dempster, Laird und Rubin (1977)) algorithm, treating the random effects as missing data. This can be slow to converge near the optimum.

- The Newton-Raphson algorithm can be used to directly optimize $L_{ML}(\widehat{\beta}(\boldsymbol{\alpha}), \boldsymbol{\alpha})$ or $L_{REML}(\boldsymbol{\alpha})$, but is most stable near the optimum and computationally more expensive (computation of derivatives).

- `lme` uses a hybrid of first EM and then Newton-Raphson.

  `lmer` offers several options for constrained nonlinear optimizers of the (RE)ML criterion that do not use derivatives.

Numerical calculation of the estimates

The Newton-Raphson algorithm cannot take into account restrictions on the parameters such as we discussed on Slide 5 for $\boldsymbol{\Theta}_{\boldsymbol{\alpha}}$ to guarantee positive (semi)-definite covariance matrices.

Whether negative estimates for variances in $\boldsymbol{D}$ can occur depends on the restrictions and parameterizations the software package uses for $\boldsymbol{\alpha}$.

Users should be aware of the different approaches that software packages take to these restrictions (and of the version of restrictions used).

Numerical calculation of the estimates in R

- To avoid negative variance estimates, `lme` maximizes the (restricted) log-likelihood with respect to the log-variances. This means that a maximum in zero (the corresponding random effect vanishes) cannot be found.

- `lmer` assumes a diagonal $\boldsymbol{\Sigma}_i$. It uses a Cholesky decomposition for $\boldsymbol{D}$ with constraints on the diagonal elements. This ensures a positive semi-definite $\boldsymbol{D}$ but allows for singular $\boldsymbol{D}$ corresponding to zero variances. For more details, see Bates et al, 2014.

Numerical issues

Consider for example the model

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + b_{i1} + b_{i2} t_{ij} + b_{i3} t_{ij}^2 + \epsilon_{ij},$$

with subject-specific quadratic trends. Note that t_{ij}^2 can become very large depending on the time scale for t_{ij} (e.g. months in a decade-long study).

It can be helpful in such cases to rescale time e.g. from months to decades.

Otherwise, the variance of b_{i3} will be very small and close to the boundary 0 and this can lead to numerical problems during maximization.

$$\mathsf{Var}(b_{i3} t_{ij}^2) = d_{33} t_{ij}^2 \quad \Rightarrow \quad \mathsf{Var}(b_{i3} 120^2 \left(\frac{t_{ij}}{120}\right)^2) = (14400\ d_{33}) \left(\frac{t_{ij}}{120}\right)^2.$$

Overview

4.1 The marginal model

4.2 Estimation of the fixed effects

4.3 Estimation of the covariance parameters

4.4 Numerical calculation of the estimates

4.5 Prediction of the random effects

Prediction of the random effects

- Often one is mainly interested in the population effects $\boldsymbol{\beta}$.
- It can however also be of interest to look at 'estimates' of the random effects $\mathbf{b}_i$, e.g. to obtain individual prognoses or to detect unusual or extreme subjects.
- As the $\boldsymbol{b}_i$ are random variables, they cannot be estimated, strictly speaking. We thus usually talk of prediction of the random effects.
- To obtain predictions for $\boldsymbol{b}_i$, we need the hierarchical model formulation (3.5), as the random effects do not occur in the marginal model (3.8) (which was used to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$).

Prediction of the random effects

Remember the longitudinal linear mixed model (3.5)

$$\begin{cases} \mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \\ \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}) \\ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_i) \\ \mathbf{b}_1, \ldots, \mathbf{b}_N, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N \text{ are independent} \end{cases}$$

- Marginal distribution of $\mathbf{b}_i$:

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

- Conditional distribution of $\mathbf{Y}_i$:

$$\mathbf{Y}_i | \mathbf{b}_i \sim \mathcal{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{\Sigma}_i)$$

A Bayesian approach for prediction

Bayes theorem:

$$f(\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i) = \frac{f(\mathbf{y}_i|\mathbf{b}_i)f(\mathbf{b}_i)}{\int f(\mathbf{y}_i|\mathbf{b}_i)f(\mathbf{b}_i)d\mathbf{b}_i}.$$

Usually, $\mathbf{b}_i$ is predicted as the mean of the posterior distribution

$$\begin{aligned} \widehat{\mathbf{b}}_i(\boldsymbol{\theta}) &= \mathsf{E}[\mathbf{b}_i|\mathbf{Y}_i = \mathbf{y}_i] \\ &= \mathbf{D}(\boldsymbol{\alpha})\mathbf{Z}_i^T\mathbf{V}_i(\boldsymbol{\alpha})^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}). \end{aligned}$$

This results from rules for the conditional expectation for normally distributed vectors with two blocks and

$$\begin{pmatrix} \mathbf{Y}_i \\ \mathbf{b}_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{X}_i\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V}_i & \mathbf{Z}_i\mathbf{D} \\ \mathbf{D}\mathbf{Z}_i^T & \mathbf{D} \end{pmatrix}\right).$$

Prediction of the random effects

- In practice, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are replace by their ML or REML estimates $\rightarrow$ "Empirical Bayes" estimation.

- $\widehat{\mathbf{b}}_i(\boldsymbol{\theta})$ also corresponds to the **best linear unbiased predictor (BLUP)** of $\boldsymbol{b}_i$. Unbiased here means

$$\mathsf{E}(\widehat{\mathbf{b}}_i(\boldsymbol{\theta})) = \mathsf{E}(\boldsymbol{b}_i) = \mathbf{0}$$

  and not $\mathsf{E}(\widehat{\mathbf{b}}_i(\boldsymbol{\theta})|\boldsymbol{b}_i) = \boldsymbol{b}_i$ for all $\boldsymbol{b}_i$, i.e. $\widehat{\mathbf{b}}_i(\boldsymbol{\theta})$ is not centered around $\boldsymbol{b}_i$. Best means that the BLUP minimizes $\mathsf{E}[(\widehat{\boldsymbol{b}}_i - \boldsymbol{b}_i)^T(\widehat{\boldsymbol{b}}_i - \boldsymbol{b}_i)]$ among all linear unbiased predictors $\widehat{\boldsymbol{b}}_i$.

  If $\widehat{\boldsymbol{\theta}}$ is estimated, $\widehat{\mathbf{b}}_i(\widehat{\boldsymbol{\theta}})$ is called the empirical BLUP (eBLUP).

Shrinkage: Example random intercept model

Consider the model with only a **random intercept**:

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

with

$$\mathbf{Z}_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \ b_i \sim \mathcal{N}(0, d^2).$$

We additionally assume $\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}_{n_i}, \sigma^2 \boldsymbol{I}_{n_i})$.

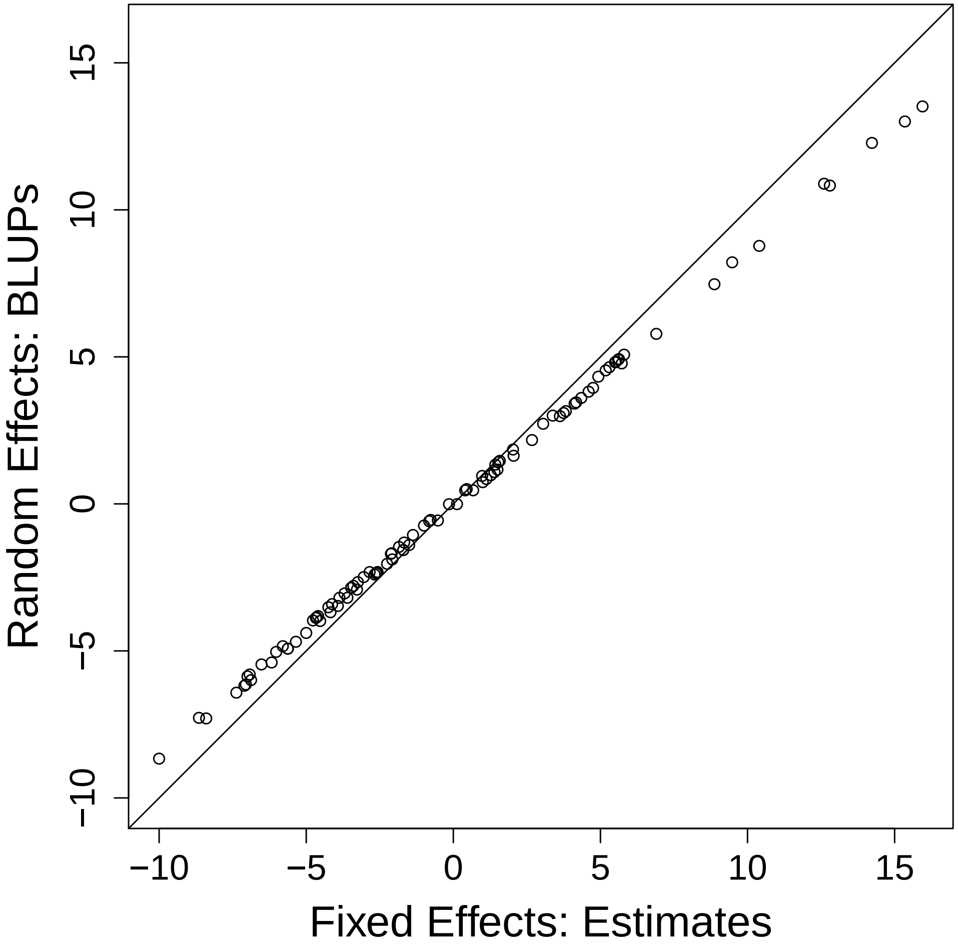
Shrinkage: Example random intercept model

Empirical Bayes estimate / BLUP for $\widehat{b}_i$:

$$
\begin{aligned}
\widehat{b}_i &= \mathbf{D}(\boldsymbol{\alpha})\mathbf{Z}_i^T \mathbf{V}_i(\boldsymbol{\alpha})^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \\
&= d^2 \mathbf{1}_{n_i}^T \left(d^2 \mathbf{1}_{n_i \times n_i} + \sigma^2 \mathbf{I}_{n_i}\right)^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \\
&= \frac{d^2}{\sigma^2} \mathbf{1}_{n_i}^T \left(\mathbf{I}_{n_i} - \frac{d^2}{n_i d^2 + \sigma^2} \mathbf{1}_{n_i \times n_i}\right) (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \\
&= \frac{d^2}{\sigma^2}(1 - n_i \frac{d^2}{n_i d^2 + \sigma^2}) \mathbf{1}_{n_i}^T (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \\
&= \frac{n_i d^2}{\sigma^2 + n_i d^2} \, \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}).
\end{aligned}
$$

Interpretation?

Shrinkage: Example TLC trial



Shrinkage effekt

Prediction of $\mathbf{Y}_i$:

$$\begin{aligned}
\widehat{\mathbf{Y}}_i &= \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{b}}_i \\
&= \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\mathbf{D}(\widehat{\boldsymbol{\alpha}})\mathbf{Z}_i^T\mathbf{V}_i(\widehat{\boldsymbol{\alpha}})^{-1}(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}) \\
&= (\mathbf{I}_{n_i} - \mathbf{Z}_i\mathbf{D}(\widehat{\boldsymbol{\alpha}})\mathbf{Z}_i^T\mathbf{V}_i(\widehat{\boldsymbol{\alpha}})^{-1})\mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\mathbf{D}(\widehat{\boldsymbol{\alpha}})\mathbf{Z}_i^T\mathbf{V}_i(\widehat{\boldsymbol{\alpha}})^{-1}\mathbf{y}_i \\
&= \boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\alpha}})\mathbf{V}_i(\widehat{\boldsymbol{\alpha}})^{-1}\mathbf{X}_i\widehat{\boldsymbol{\beta}} + (\mathbf{I}_{n_i} - \boldsymbol{\Sigma}_i(\widehat{\boldsymbol{\alpha}})\mathbf{V}_i(\widehat{\boldsymbol{\alpha}})^{-1})\mathbf{y}_i
\end{aligned}$$

This is a weighted average between the population mean $\mathbf{X}_i\widehat{\boldsymbol{\beta}}$ and the observations $\mathbf{y}_i$.

This is called **shrinkage** (towards the population-averaged profile $\mathbf{X}_i\widehat{\boldsymbol{\beta}}$) and reflects the “borrowing of strength” in LMMs. Predictions (also for $\boldsymbol{b}_i$) have less spread than if treating $\boldsymbol{b}_i$ as fixed effects.

Henderson's mixed model equations

BLUE $\widehat{\boldsymbol{\beta}}$ and BLUP $\widehat{\boldsymbol{b}} = (\widehat{\boldsymbol{b}}_1^T, \ldots, \widehat{\boldsymbol{b}}_N^T)^T$ are the solution to the simultaneous Henderson's mixed model equations (Henderson, 1950)

$$\begin{aligned} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\widehat{\boldsymbol{b}} &= \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y} \qquad (4.1) \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\widehat{\boldsymbol{b}} &= \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} \end{aligned}$$

There are different justifications for these equations (Robinson, 1991) including that the estimates are **empirical Bayes estimates** (with uniform improper prior for $\boldsymbol{\beta}$).

Henderson's mixed model equations

Equations (4.1) also arise when maximizing the log-likelihood based on the joint density of $\boldsymbol{Y}$ and $\boldsymbol{b}$ over $\boldsymbol{\beta}$ and $\boldsymbol{b}$ ("penalized log-likelihood"):

$$\begin{aligned}\ell_{pen}(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\alpha}) &= \log f(\boldsymbol{y}|\boldsymbol{b}) - \log f(\boldsymbol{b}) \qquad (4.2)\\ &= const - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zb})^T \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zb}) - \frac{1}{2}\boldsymbol{b}^T \boldsymbol{G}^{-1}\boldsymbol{b}\end{aligned}$$

- For $\boldsymbol{G} \quad \to \boldsymbol{0}$ this yields $\widehat{\boldsymbol{b}} = \boldsymbol{0}$.

- For $\boldsymbol{G}^{-1} \to \boldsymbol{0}$ we have $\boldsymbol{b}^T \boldsymbol{G}^{-1} \boldsymbol{b} \to 0$ and the estimate of $\boldsymbol{b}$ converges to the estimate treating $\boldsymbol{b}$ as fixed effects.

Henderson's mixed model equations

Solving (4.1) jointly for $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}})$ yields a compact way to express the solutions (equivalent to the separate formulas for BLUE and BLUP) as

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{b}} \end{pmatrix} = (\boldsymbol{C}^T \boldsymbol{R}^{-1} \boldsymbol{C} + \mathrm{diag}(\mathbf{0}_{p \times p}, \boldsymbol{G}^{-1}))^{-1} \boldsymbol{C}^T \boldsymbol{R}^{-1} \boldsymbol{y} \tag{4.3}$$

with $\boldsymbol{C} = (\boldsymbol{X} | \boldsymbol{Z})$ and blockdiagonal matrix $\mathrm{diag}(\mathbf{0}_{p \times p}, \boldsymbol{G}^{-1})$.

In this form the close relationship to ridge estimation becomes apparent.