Overview

The General Linear Mixed Model (LMM)

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon, \tag{1}$$

- $\boldsymbol{X}$ and $\boldsymbol{Z}$ known design matrices ($n \times p$ and $n \times r$)
- β vector of unknown fixed parameters
- $\boldsymbol{b}$ vector of random effects
- ε vector of unobservable random errors

Assumptions: Independence of $\boldsymbol{b}$ and ε, and

$$\mathrm{E}\begin{pmatrix} \boldsymbol{b} \\ \varepsilon \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix} \quad \text{and} \quad \mathrm{Cov}\begin{pmatrix} \boldsymbol{b} \\ \varepsilon \end{pmatrix} = \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{pmatrix} \tag{2}$$

Typically, also multivariate normality of $(\boldsymbol{b}', \varepsilon')'$ is assumed. (3)

The Longitudinal Linear Mixed Model

The longitudinal linear mixed model is typically given on the level of the subject.

$$\boldsymbol{Y_i} = \boldsymbol{X_i}\beta + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}, \quad \boldsymbol{b_i} \sim N(\boldsymbol{0}, \boldsymbol{D_0}), \;\; \varepsilon_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i) \tag{4}$$

where the subscript i indicates subject-specific quantities (Laird & Ware, 1982).

A typical longitudinal model is the growth-curve model

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, \;\; i = 1, \ldots, I,$$

with

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \overset{iid}{\sim} N\left(\boldsymbol{0}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right) \quad \text{independent of} \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

Here, $\boldsymbol{D_0} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ and $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}_{\boldsymbol{n_i}}$.

Example: A two-stage analysis → blackboard

The Longitudinal Linear Mixed Model

- Longitudinal linear mixed model (LLMM) special case of LMM (1).
  Let $\boldsymbol{D}$, $\boldsymbol{R}$ and $\boldsymbol{Z}$ be block-diagonal with blocks $\boldsymbol{D_0}$, $\boldsymbol{R}_i$ and $\boldsymbol{Z}_i$, respectively.
  Stack $\boldsymbol{Y}_i$, $\boldsymbol{X}_i$, $\boldsymbol{b}_i$ and ε_i to obtain $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{b}$ and ε.

- LLMM assumes that $\boldsymbol{Y}$ can be divided into independent subvectors $\boldsymbol{Y}_i$ for the ith subject.

- LMM (1) more general - allows for inclusion of other model components such as smooth functions modeled using penalized splines in the mixed model framework. $\longrightarrow$ Loss of this independence, but very flexible models.

- LMM (1) can also incorporate additional clusters (e.g. growth curves for children clustered in schools), or nested structures of random effects.

Excursus: Penalized Splines in a Nutshell

- Consider the nonparametric regression problem

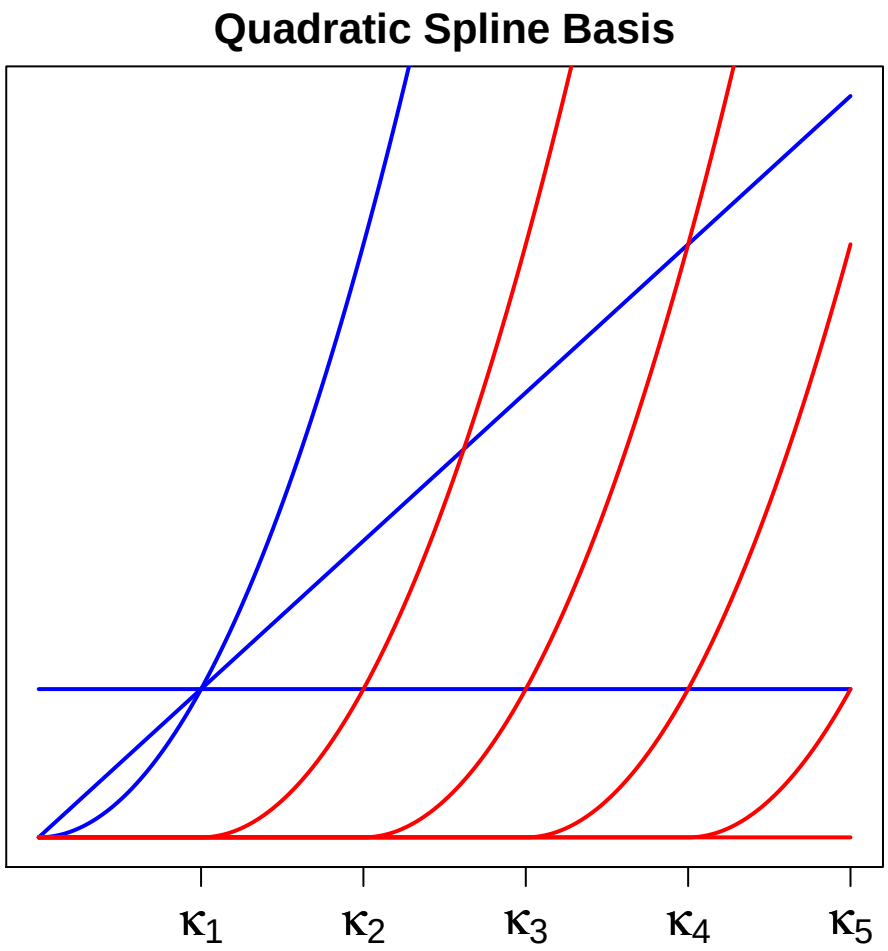
$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

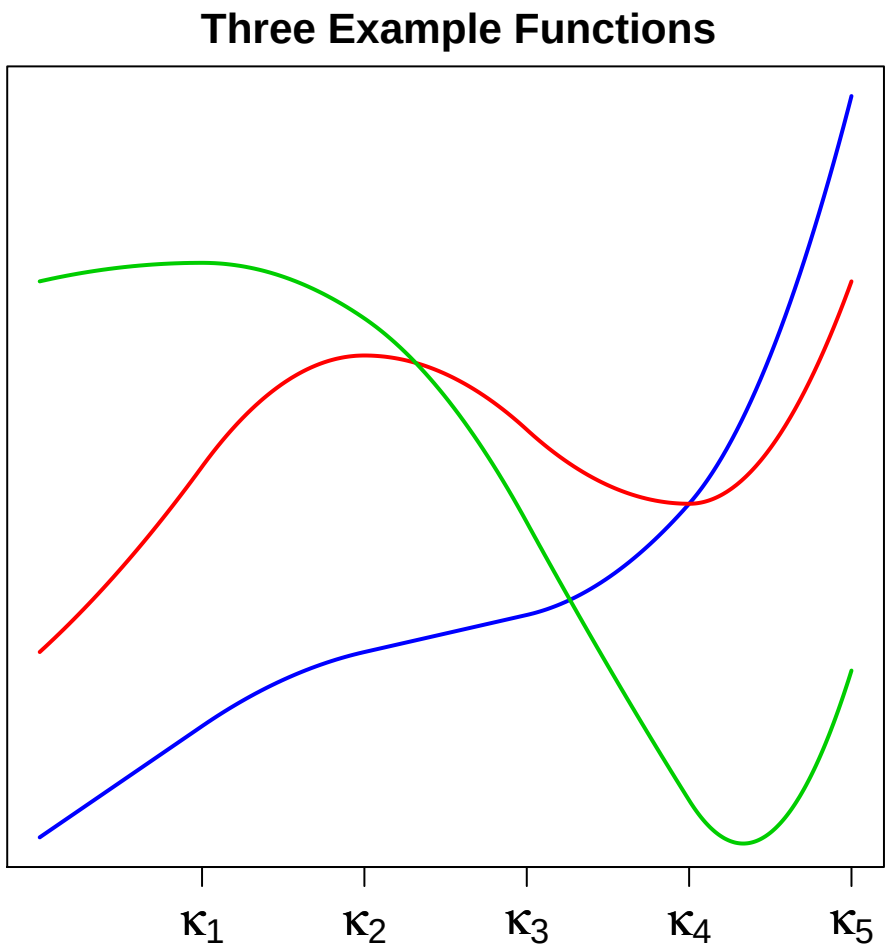
  with unknown smooth function $m(\cdot)$.
- Approximate $m(\cdot)$ by a linear combination of (many) spline basis functions, e.g. truncated polynomials,

$$m(x) \approx \sum_{j=0}^{q} \beta_j x^j + \sum_{j=1}^{K} b_j (x - \kappa_j)_+^q,$$

  where $\kappa_1, \ldots, \kappa_K$ is a sequence of knots, and $u_+^q = (\max\{0, u\})^q$.
- This approximation gives a piecewise polynomial function of degree q with certain smoothness properties (continuity of $(q-1)$th derivative).

Excursus: Penalized Splines in a Nutshell





Excursus: Penalized Splines in a Nutshell

- Estimation using a regularization penalty, to avoid overly wiggly function:

$$(\boldsymbol{Y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{X}\beta - \boldsymbol{Z}\boldsymbol{b}) + \frac{1}{\lambda}\boldsymbol{b}'\boldsymbol{b} \rightarrow \min_{\beta, \boldsymbol{b}},$$

  - $\boldsymbol{X}$ and $\boldsymbol{Z}$ design matrices for the spline basis functions x^j resp. $(x - \kappa_j)_+^q$
  - λ a smoothing parameter.

- Penalized least squares estimation equivalent to estimation in the LMM with
  - **fixed effects** $\beta_0, \ldots, \beta_q$: subspace of polynomial functions (degree q)
  - **random effects** $b_1, \ldots, b_K \overset{iid}{\sim} N(0, \sigma_b^2)$: deviations from subspace.

  Can estimate smoothing parameter $\lambda = \sigma_b^2/\sigma_\varepsilon^2$ data driven.

- Analogous for other basis choices (e.g. B-Splines) and for spatial effects, interaction surfaces, varying coefficients, ... $\rightarrow$ can combine with random effects for longitudinal data in a general LMM (additive mixed model).
- Examples: smooth dose-response functions for covariates, smooth profiles over time, time-varying effects, spatial surfaces (spatial information on subjects), ... More detailed information: Ruppert, Wand & Carroll, 2003.

D, R and V

- $\boldsymbol{D}$, $\boldsymbol{R}$ and $\boldsymbol{Z}$ together imply the correlation structure for $\boldsymbol{Y}$, as $\boldsymbol{V} = \text{Cov}(\boldsymbol{Y}) = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}' + \boldsymbol{R}$.
- $\boldsymbol{Z}\boldsymbol{b}$ with $\boldsymbol{D}$ models differences between subjects, e.g. in their average level and average evolution in time.
- $\boldsymbol{R}$ models the residual serial correlation not explained by $\boldsymbol{Z}\boldsymbol{b}$, as well as measurement error.

Some common models and implications for V

The random intercept model

$$Y_{ij} = \boldsymbol{x_{ij}}'\boldsymbol{\beta} + b_i + \varepsilon_{ij}, \quad b_i \overset{iid}{\sim} N(0, \sigma_b^2), \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

with $\boldsymbol{R} = \sigma_\varepsilon^2 \boldsymbol{I_n}$, $\boldsymbol{D} = \sigma_b^2 \boldsymbol{I_l}$, $\boldsymbol{D}_0 = \sigma_b^2$ and $\boldsymbol{Z}_i = (1, \ldots, 1)'$ implies $\boldsymbol{V}$ is block-diagonal, with the ith block for the ith subject equal to

$$\begin{pmatrix} \sigma_\varepsilon^2 + \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_\varepsilon^2 + \sigma_b^2 & \cdots & \sigma_b^2 \\ \vdots & \ddots & & \vdots \\ \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_\varepsilon^2 + \sigma_b^2 \end{pmatrix}.$$

This is called compound symmetry. In a marginal model, this is often written with σ^2 for $\sigma_\varepsilon^2 + \sigma_b^2$ and $\rho\sigma^2$ for σ_b^2. (Then, the correlation ρ could be negative.)

This implies that all observations on a subject are equally correlated, i.e. it assumes that observations on the same subject closer in time are not more similar than observations further away in time. (E.g. blood marker with short half-life.)

Some common models and implications for V

The growth-curve model

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}, \quad b_i \overset{iid}{\sim} N\left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right), \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

with $\boldsymbol{D}$ block-diagonal with blocks $\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$, $\boldsymbol{Z}_i = \begin{pmatrix} 1 & \dots & 1 \\ t_{11} & \dots & t_{1n_i} \end{pmatrix}'$ and $\boldsymbol{R} = \sigma_\varepsilon^2 \boldsymbol{I_n}$, implies

$$\begin{aligned} \text{Var}(Y_{ij}) &= \sigma_1^2 + 2\sigma_{12} t_{ij} + \sigma_2^2 t_{ij}^2 + \sigma_\varepsilon^2 \quad \text{and} \\ \text{Cov}(Y_{ij}, Y_{ik}) &= \sigma_1^2 + \sigma_{12} t_{ij} + \sigma_{12} t_{ik} + \sigma_2^2 t_{ij} t_{ik}, \quad j \neq k. \end{aligned}$$

In particular, the variance function is quadratic in time. Adding a quadratic term $b_{2i} t_{ij}^2$ would make it a forth-order polynomial in time. (Extrapolation??)

Residual Covariance Structure

Often, the conditional independence model $\boldsymbol{R} = \sigma^2_\varepsilon \boldsymbol{I_n}$ is chosen, which implies that the responses of a subject i are independent, given $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$. This may be unrealistically simplistic. If the covariance structure can not be meaningfully explained by random effects alone, more general forms of $\boldsymbol{R}$ should be chosen.

$\boldsymbol{R}$ is typically chosen to be block-diagonal with blocks depending on i only through n_i and the time points $t_{ij}, j = 1, \ldots, n_i$. Usually, it is modeled as

$$\text{Cov}(\varepsilon_{ij}, \varepsilon_{ik}) = \sigma^2 + \tau^2 g(|t_{ij} - t_{ik}|)$$

with a suitable decreasing function $g(\cdot)$ with $g(0) = 1$ and $\lim_{u \to \infty} g(u) = 0$.
Two frequently used functions are the exponential and Gaussian correlation functions $g(u) = \exp(-u)$ and $g(u) = \exp(-\phi u^2)$ (with $\phi > 0$).

In applications, however, the effect of serial correlation is often dominated by the combination of random effects and measurement error (also causing estimation problems when all are present in the model).

Conditional and Marginal Perspectives

Conditional perspective on the mixed model:

$$\boldsymbol{Y}|\boldsymbol{b} \sim N(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b}, \sigma^2 \boldsymbol{I_n}), \quad \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{D}). \tag{5}$$

Interpretation: Random effects are subject-specific effects on the mean that vary within the population and are estimated subject to a regularization constraint.

Marginal perspective on the mixed model:

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\beta, \boldsymbol{V}), \tag{6}$$

where $\boldsymbol{V} = \text{Cov}(\boldsymbol{Y}) = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}' + \boldsymbol{R}$.
Interpretation: The random effects induce a correlation structure and thus enable a proper statistical analysis of the correlated data.

Inference in the Linear Mixed Model

What is the likelihood for the linear mixed model (1)-(3)?

Inference is usually based on the marginal model (6) for $\boldsymbol{Y}$ (unless Bayesian methods are used).
Then, the unknown parameters are (β, θ), where θ contains the unknown parameters in $\boldsymbol{D}$ and $\boldsymbol{R}$ (and thus $\boldsymbol{V}$).

(5) implies (6), but they are not equivalent!

Example: random intercept + independent errors $\Rightarrow$ compound symmetry for $\boldsymbol{V}$
σ_b^2 covariance in (6) $\rightarrow$ could be negative
σ_b^2 variance in (5) $\rightarrow$ has to be non-negative

And (6) does not include $\boldsymbol{b}$. Not as much of a problem if interest is in (β, θ), but does not work if interest is in $\boldsymbol{b}$.

Estimation of β

For a given θ, (6) is simply a general linear model. We can estimate β using generalized least squares

$$(\boldsymbol{Y} - \boldsymbol{X}\beta)' \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\beta) \to \min_{\beta}$$

which gives

$$\widehat{\beta} = \left(\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{Y}. \tag{7}$$

(We assume that the inverses $\boldsymbol{V}^{-1}$ and $\left(\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1}$ exist. Generalizations use generalized inverses.)
This is also the maximum likelihood (ML) estimator when we have the normality assumption (3): The log-likelihood for β and θ from the marginal model (6) is

$$\ell(\beta, \theta) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{V}| - \frac{1}{2} (\boldsymbol{Y} - \boldsymbol{X}\beta)' \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\beta). \tag{8}$$

$$\frac{\partial}{\partial \beta} \ell(\beta, \theta) = \boldsymbol{X}' \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\beta) \stackrel{!}{=} \boldsymbol{0} \quad \Rightarrow \quad \widehat{\beta} = \left(\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{Y}.$$

Estimation of β

$\widehat{\beta}$ is also the BLUE, the best linear unbiased estimator (Gauss-Markov theorem):

Let $\widetilde{\beta} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{y}$ be another linear unbiased estimator of β.
$\Rightarrow \mathrm{E}(\widetilde{\beta}) = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}\beta = \beta \quad \Rightarrow \quad \boldsymbol{a} = \boldsymbol{0}$ and $\boldsymbol{B}\boldsymbol{X} = \boldsymbol{I}_p$.
$\Rightarrow$ For all $\boldsymbol{c} \in \mathbb{R}^p$

$$
\begin{aligned}
& \mathrm{Var}(\boldsymbol{c}'\widetilde{\beta}) - \mathrm{Var}(\boldsymbol{c}'\widehat{\beta}) \\
= \;& \boldsymbol{c}'\boldsymbol{B}\boldsymbol{V}\boldsymbol{B}'\boldsymbol{c} - \boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c} \\
= \;& \boldsymbol{c}'\boldsymbol{B}\boldsymbol{V}\boldsymbol{B}'\boldsymbol{c} - \boldsymbol{c}'\boldsymbol{B}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{B}'\boldsymbol{c} \\
= \;& \boldsymbol{c}'\boldsymbol{B}\boldsymbol{V}^{1/2}[\boldsymbol{I}_n - \boldsymbol{V}^{-1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1/2}]\boldsymbol{V}^{1/2}\boldsymbol{B}'\boldsymbol{c} \geq 0
\end{aligned}
$$

as the middle matrix is a projection matrix and thus positive semi-definite.
As $\widehat{\beta}$ is unbiased with smallest variance, it is the BLUE.

With the normality assumption (3), $\widehat{\beta}$ can even be shown to be the best unbiased estimator.

Estimation of θ

Maximum likelihood estimation

Substitute $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ into (8) $\rightarrow$ profile log-likelihood $\ell(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta})$.

Maximize to obtain ML estimate for $\boldsymbol{\theta}$.

No closed form solution in general $\rightarrow$ numerical maximization.

Restricted maximum likelihood estimation

Maximum likelihood estimates for variances are biased downward.

Example: $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, ML estimator $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$, but $\mathrm{E}(\widehat{\sigma}^2) = \frac{n-1}{n}\sigma^2$.

$\rightarrow$ Everyone uses $\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

This is the restricted maximum likelihood estimate.

It accounts for estimation of μ by $\bar{Y}$.

Restricted Maximum Likelihood Estimation of θ

Idea: Use likelihood not of $\boldsymbol{Y}$, but of $(n-p)$ linearly independent error contrasts $\boldsymbol{AY}$ with distribution independent of unknown β. (Patterson & Thompson, 1971).

Split the log-likelihood into two parts for $\boldsymbol{AY}$ (for θ) and $\boldsymbol{BY}$ (for β) with

1. $\boldsymbol{A}$ of rank $n-p$ and $\boldsymbol{B}$ of rank p.
2. The two parts are statistically independent, i.e. (here:) $\text{Cov}(\boldsymbol{AY}, \boldsymbol{BY}) = \boldsymbol{0}$.
3. $\boldsymbol{AY}$ are error contrasts, i.e. $\text{E}(\boldsymbol{AY}) = \boldsymbol{0}$.
4. $\boldsymbol{BX}$ is of rank p ($\Rightarrow \boldsymbol{BY}$ estimates one-to-one function of β).

Suitable matrices: $\boldsymbol{A} = \boldsymbol{I_n} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}$.
$\rightarrow$ Alternative name residual maximum likelihood estimation.

From the first two points, the log-likelihood is $\ell = \ell' + \ell^*$ (up to an additive constant), ℓ' for $\boldsymbol{AY}$ and ℓ^* for $\boldsymbol{BY}$.

Restricted Maximum Likelihood Estimation of θ

Maximizing ℓ^* (from $\boldsymbol{BY}$) with respect to β gives again the MLE/BLUE $\widehat{\beta}$ (depending on θ).

Maximizing ℓ' (from $\boldsymbol{AY}$) with respect to θ:
Harville (1974) showed that the restricted log-likelihood ℓ' is (up to an additive constant) independent of the precise error contrasts used.
He derived the restricted log-likelihood, which is (up to a constant)

$$\begin{aligned} l(\theta) &= -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\beta}) \\ &= \ell(\widehat{\beta}(\theta),\theta) - \frac{1}{2}\log|\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}| \end{aligned}$$

(Proof: blackboard)
Maximize $\longrightarrow$ restricted maximum likelihood (REML) estimate for θ.

Example: In a linear model, this gives the usual estimate for σ^2 (dividing by $n-p$). It takes into account the loss in degrees of freedom from estimation of β.

Restricted Maximum Likelihood Estimation of θ

The restricted likelihood is also the conditional likelihood for θ given the sufficient statistic $\boldsymbol{BY}$ for β and it can be argued that this entails no information loss (Smyth & Verbyla, 1996).

The concept that $\boldsymbol{BY}$ contains no information about θ absent knowledge about β has also been called marginal sufficiency (Sprott, 1975; Harville, 1977).

Implementation of ML or REML

Lindstrom & Bates, JASA, 1988 derive the first- and second-order derivatives of log-likelihood and restricted log-likelihood. They give details of implementation of a Newton-Raphson Algorithm for estimation.

Note that the requirements for $\boldsymbol{\theta}$ in the marginal model ($\boldsymbol{V}$ positive (semi-)definite) are different from those in the conditional model ($\boldsymbol{D}_0$ and $\boldsymbol{R}_i$ positive (semi-)definite). Thus, software packages maximize for $\boldsymbol{\theta}$ often over a larger parameter space than the conditional model implies. E.g. SAS `proc mixed` (at least v6.12) by default only requires the diagonal elements of $\boldsymbol{D}_0$ and $\boldsymbol{R}_i$ to be positive. Also, be aware of other numerical issues. For example, R `lme()` maximizes the (restricted) log-likelihood with respect to the scaled logarithm of the variances, and thus can never find a maximum at zero (Pinheiro & Bates, 2000, section 2.2).

An alternative is the EM algorithm, treating the random effects as missing data, which, however, can be slow to converge. R `lme()` uses a hybrid of EM and then Newton-Raphson.

EBLUE and EBLUP

After estimation of $\boldsymbol{\theta}$, we can substitute $\widehat{\boldsymbol{\theta}}$ into the formula $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ to obtain an estimate. This is the estimated BLUE (EBLUE).

Often it is also of interest to obtain predictions of $\boldsymbol{b}$ (e.g. find "unusual" subjects). Then, we have to go back to the conditional model formulation (5).
The best linear unbiased predictor (BLUP) for $\boldsymbol{b}$ is (see next slide)

$$\widehat{\boldsymbol{b}} = \boldsymbol{D}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}).$$

The combined BLUE and BLUP (or collectively referred to as BLUPs) can also be written as

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{b}} \end{pmatrix} = (\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{C} + \boldsymbol{G})^{-1}\boldsymbol{C}'\boldsymbol{R}^{-1}\boldsymbol{Y},$$

where $\boldsymbol{C} = (\boldsymbol{X}|\boldsymbol{Z})$ and $\boldsymbol{G} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}^{-1} \end{pmatrix}$.

The estimated BLUP (EBLUP) is obtained by substituting $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in $\boldsymbol{D}$, $\boldsymbol{V}$ and $\widehat{\boldsymbol{\beta}}$.

BLUE and BLUP

BLUE and BLUP (or collectively referred to as BLUP) $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{b}}$ are the solution to the simultaneous Henderson's mixed model equations (Henderson, 1950)

$$\begin{aligned} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\widehat{\boldsymbol{b}} &= \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}} + (\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{D}^{-1})\widehat{\boldsymbol{b}} &= \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Y} \end{aligned} \tag{9}$$

There are (at least) four different justifications for these equations (Robinson, 1991)

a) They arise when maximizing the joint density of $\boldsymbol{Y}$ and $\boldsymbol{b}$.

b) In a Bayesian approach, regard $\boldsymbol{\beta}$ as a parameter with a uniform improper prior distribution, independent of $\boldsymbol{b}$. (9) gives the posterior mode for $\boldsymbol{\beta}$ and $\boldsymbol{b}$. Thus, the resulting estimates are also called empirical Bayes estimates.

c) $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{b}}$ have minimum mean squared error within the class of linear unbiased estimates for $\boldsymbol{\beta}$ and $\boldsymbol{b}$. (Unbiasedness for $\widehat{\boldsymbol{b}}$: $\mathrm{E}(\widehat{\boldsymbol{b}}) = \mathrm{E}(\boldsymbol{b}) = \boldsymbol{0}$.)

d) Goldberger's derivation: For a future observation $\boldsymbol{Y}_* = \boldsymbol{x}_*'\boldsymbol{\beta} + \boldsymbol{z}_*'\boldsymbol{b} + \boldsymbol{\varepsilon}_*$, the best linear unbiased predictor is $\boldsymbol{x}_*'\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_*'\widehat{\boldsymbol{b}}$.

Shrinkage

$\widehat{\boldsymbol{b}}$ is a shrinkage estimator of $\boldsymbol{b}$, i.e. its components have less spread than the generalized least squares estimators would have when treating $\boldsymbol{b}$ as fixed effects.

This seems to contradict the unbiasedness property of the BLUP. However, as Robinson, 1991 points out, unbiasedness here means

$$\mathrm{E}(\widehat{\boldsymbol{b}}) = \mathrm{E}(\boldsymbol{b}) = \mathbf{0},$$

not

$$\mathrm{E}(\widehat{\boldsymbol{b}}|\boldsymbol{b}) = \boldsymbol{b} \text{ for all } \boldsymbol{b}.$$

Shrinkage

$$\begin{aligned}
\widehat{\boldsymbol{Y}_i} &= \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\widehat{\boldsymbol{b}_i} \\
&= \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}) \\
&= (\boldsymbol{V}_i - \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i')\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}\boldsymbol{Y} \\
&= \boldsymbol{R}_i\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}\boldsymbol{Y}_i
\end{aligned}$$

The BLUP for $\boldsymbol{Y_i}$ thus is a weighted average of the population-averaged profile $\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$ and the individual data $\boldsymbol{Y_i}$. The observed data is shrunk towards the population-averaged profile ("borrowing of strength").

Note that $\boldsymbol{V}_i = \boldsymbol{R}_i + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'$.

- If $\boldsymbol{R}_i\boldsymbol{V}_i^{-1}$ is large, i.e. the residual variability ($\boldsymbol{R}_i$) is large compared to the between-subject variability ($\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'$), much weight is given to the population-averaged profile $\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$.
- If $\boldsymbol{R}_i\boldsymbol{V}_i^{-1}$ is small, the opposite is the case.

Inference for β

For known $\boldsymbol{\theta}$, we have

$$\widehat{\boldsymbol{\beta}} \sim (\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}).$$

Given the normality assumption (3), or asymptotically, $\widehat{\boldsymbol{\beta}}$ is also normal.

In practice, for unknown $\boldsymbol{\theta}$, the covariance has to be estimated as $(\boldsymbol{X}'\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{V}(\widehat{\boldsymbol{\theta}})^{-1}\boldsymbol{X})^{-1}$.
Reported standard errors and confidence intervals are usually based on this estimated covariance matrix. Note that this ignores the uncertainty in $\boldsymbol{\theta}$.

This covariance estimate is model-based.

Robust Variance Estimation

What happens if our model for Cov($\boldsymbol{Y}$) is wrong?
For example, we could assume compound symmetry with constant variance over time, but in fact variances could increase over time, and correlations might decrease with time distance.
Suppose the true covariance is Cov($\boldsymbol{Y}$) $= \boldsymbol{V}$, but our model specifies instead $\boldsymbol{W}$ as the so-called working covariance matrix. Then,

$$\widehat{\boldsymbol{\beta}} \quad = \quad (\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{Y},$$

and $\widehat{\boldsymbol{\beta}}$ is still consistent. However, the covariance changes to

$$\text{Cov}(\widehat{\boldsymbol{\beta}}) \quad = \quad (\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}$$

and we could use $\widehat{\boldsymbol{\beta}}$ together with an estimated robust covariance matrix.

Robust Variance Estimation

A consistent estimator (for $I \to \infty$) for $\text{Cov}(\widehat{\boldsymbol{\beta}})$ is (Liang & Zeger, 1986)

$$\left(\sum_{i=1}^{I} \boldsymbol{X}_i' \widehat{\boldsymbol{W}}_i^{-1} \boldsymbol{X}_i\right)^{-1} \left\{\sum_{i=1}^{I} \boldsymbol{X}_i' \widehat{\boldsymbol{W}}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}})(\boldsymbol{Y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}})' \widehat{\boldsymbol{W}}_i^{-1} \boldsymbol{X}_i\right\} \left(\sum_{i=1}^{I} \boldsymbol{X}_i' \widehat{\boldsymbol{W}}_i^{-1} \boldsymbol{X}_i\right)^{-1}.$$

This is the sandwich estimator originally due to Huber, 1967 and White, 1980, also called robust or empirical variance estimator.

This estimator is valid under more general conditions, though the model-based estimator has less variance (and no asymptotic bias) when the variance model is correctly specified, i.e. the robust is less efficient than the model-based variance estimator when we do a good job of estimating the covariance.
It has also been shown that the empirical variance estimator can be highly biased when the number of clusters (here, I) is small.

Hypothesis Testing for β

We can use our variance estimators to construct Wald tests of

$$H_0 : \boldsymbol{L}\beta = \boldsymbol{0} \text{ versus } H_A : \boldsymbol{L}\beta \neq \boldsymbol{0} \tag{10}$$

or to construct confidence intervals. For this, we use that

$$T_W = (\widehat{\beta} - \beta)' \boldsymbol{L}' \left(\boldsymbol{L} (\boldsymbol{X}' \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{L}' \right)^{-1} \boldsymbol{L} (\widehat{\beta} - \beta)$$

asymptotically follows a chi-square distribution with rank(L) degrees of freedom. (Or use the corresponding robust covariance estimate.)

Note: The results given here for testing (of fixed effects or variance components) assume the longitudinal linear mixed model (4). Asymptotics are with regard to $I \to \infty$. Results for more general types of linear mixed models often remain elusive (Ruppert, Wand & Carroll, 2003, section 4.8).

Hypothesis Testing for β

The Wald test statistic is based on estimated standard errors which underestimate the true variability in $\widehat{\boldsymbol{\beta}}$, as they ignore the variability introduced by estimating $\boldsymbol{\theta}$.

This problem is often alleviated by using approximate *t*- or *F*-statistics instead:

- For a single parameter β_j in $\boldsymbol{\beta}$, the distribution of $(\widehat{\beta}_j - \beta_j)/\widehat{SE}(\widehat{\beta}_j)$ is approximated by a *t*-distribution.
- For general hypotheses (10), an *F*-approximation to the distribution of

$$F = \frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \boldsymbol{L}' \left(\boldsymbol{L}(\boldsymbol{X}' \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{L}' \right)^{-1} \boldsymbol{L}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\text{rank}(\mathbf{L})}$$

  is used, with numerator degrees of freedom rank($\mathbf{L}$).

Hypothesis Testing for β

The t degrees of freedom or the denominator degrees of freedom for the F-distribution are estimated from the data. Common methods include a so-called Satterthwaite-type approximation (Satterthwaite, Psychometrika,1941). (The general idea is to match the moments of the distributions.)

For small samples, the method by Kenward and Roger, Biometrics, 1997 gives better results, as it first estimates the amount by which the asymptotic covariance matrix underestimates (in a matrix sense) $\text{Cov}(\hat{\beta})$, and then inflates the variance-covariance matrix by this amount before using a Satterthwaite-type approach.

Computation of Satterthwaite-type degrees of freedom are computationally expensive and can add substantially to computation time.

Hypothesis Testing for β

We could also use a likelihood ratio test (LRT) based on large sample theory. When n is not that large, the LRT tends to be more reliable than the Wald test.

For an LRT, we use that the test statistic

$$T_{LRT} = 2 \sup_{H_A} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}) - 2 \sup_{H_0} \ell(\boldsymbol{\beta}, \boldsymbol{\theta})$$

under the null hypothesis converges in distribution to a $\chi^2_{\text{rank}(\boldsymbol{L})}$ variable.

We need

- nested models
- with the same covariance structure in each model.
- And don't use with REML estimation! (Due to the different fixed effects under null and alternative, the error contrasts are different and the two likelihoods thus not comparable.)

Also, be aware that the approximation can be very poor when $n - p$ is small.

Hypothesis Testing for D

Suppose now we want to test for certain variance/covariance parameters. Let K be the dimension of $\boldsymbol{D}_0$ and

$$\boldsymbol{D}_0 = \begin{pmatrix} d_{11} & \dots & d_{1K} \\ \vdots & \ddots & \vdots \\ d_{1K} & \dots & d_{KK} \end{pmatrix} = \left(\begin{array}{ccc} & \boldsymbol{D}_1 & \\ & & \\ \hline d_{1K} & \dots & d_{KK} \end{array}\right) \begin{array}{c} d_{1K} \\ \vdots \\ d_{KK} \end{array}.$$

We might be interested in testing

$$H_0: \quad \boldsymbol{D}_0 = \left(\begin{array}{ccc} & \boldsymbol{D}_1 & \\ & & \\ \hline 0 & \dots & 0 \end{array}\right) \begin{array}{c} 0 \\ \vdots \\ 0 \end{array} \text{ with } \boldsymbol{D}_1 \text{ positive-definite } (K-1 \times K-1) \text{ versus}$$

$$H_A: \quad \boldsymbol{D}_0 \text{ positive-semidefinite } (K \times K).$$

An example would be to test for a random intercept model versus a growth-curve model ($K = 2$), or for no random effect versus a random intercept ($K = 1$).

Hypothesis Testing for D

This case violates one of the standard regularity assumptions, that the tested parameters be in the interior of the parameter space. Instead, d_{KK} as a variance has to be non-negative and $d_{KK} = 0$ thus is on the boundary of the parameter space. (Actually, $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ have to fulfill more complicated restrictions to be positive-(semi)definite.)

One can show that for the above hypothesis, T_{LRT} under the null converges in distribution to a variable with a 0.5:0.5 mixture between a χ^2_{K-1} and a χ^2_K distribution (Self & Liang, JASA, 1987, Stram & Lee, Biometrics, 1994, 1995, Giampaoli & Singer, JSPI, 2009).

The asymptotics assume the longitudinal linear mixed model (4) with $I \to \infty$. Crainiceanu & Ruppert (JRSS-B, 2004) show that if $\boldsymbol{Y}$ cannot be subdivided into a large number of independent subvectors $\boldsymbol{Y}_i$, this mixture distribution is typically a bad approximation.

Hypothesis Testing for D

Crainiceanu & Ruppert (JRSS-B, 2004) also derive the exact distribution for the case $K = 1$ (for the general LMM (1)). This is implemented in the R-package `RLRsim`. Some extensions can be found in Greven et al (JCGS, 2008) and Scheipl, Greven & Küchenhoff (CSDA, 2008).

Note also that the $\chi^2_{K-1} : \chi^2_K$ mixture distribution given above is not valid, if a nuisance parameter is also on the boundary (e.g. if, when testing for a random slope, the random intercept is also (near) zero). Self & Liang, JASA, 1987 show that then the whole geometry of the problem changes.

For more complex hypotheses for $\boldsymbol{D}_0$, more complex chi-square mixtures can arise (Self & Liang, JASA, 1987, Stram & Lee, Biometrics, 1994, 1995).

When testing only for variance components, one can also use a restricted LRT (RLRT) after REML estimation. The asymptotics are the same.

Hypothesis Testing for D

Molenberghs & Verbeke, AmerStatistician, 2007 look at Score-Tests and Wald-Tests in addition to LRTs and discuss how the same asymptotics apply to these two tests. However, they are more difficult to compute in the constraint case, as they require additional derivatives as well as numerical constraint optimization.

Precision of the BLUPs

Remember

$$\begin{pmatrix} \widehat{\beta} \\ \widehat{b} \end{pmatrix} = (C'R^{-1}C + G)^{-1}C'R^{-1}Y,$$

where $C = (X|Z)$ and $G = \begin{pmatrix} 0 & 0 \\ 0 & D^{-1} \end{pmatrix}$.

The common variance-covariance matrix can be shown to be (blackboard)

$$\text{Cov}\begin{pmatrix} \widehat{\beta} \\ \widehat{b} - b \end{pmatrix} = (C'R^{-1}C + G)^{-1}.$$

Note that $\text{Cov}(\widehat{b} - b)$ is used rather than $\text{Cov}(\widehat{b})$, as the latter does not recognize the random variation in b.

Estimated versions can be used to construct approximate confidence/prediction intervals for expressions involving the BLUPs (such as Y_i or new observations Y_i^*).

Model Selection

Mixed model software also gives AIC and BIC values. These are based on the marginal likelihood, the likelihood from the marginal model (6). For example, the Akaike information criterion (AIC) is typically defined as

$$AIC = -2 \log \ell(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}) + 2k,$$

where $k = p + 1 + d$ is the number of unknown parameters (if d is the number of unknown parameters in $\boldsymbol{D}$). (The AIC using the marginal restricted log-likelihood is defined analogously.)

However, be aware that the derivation of the AIC also assumes a parameter space $\mathbb{R}^k$ and independent and identically distributed (i.i.d.) observations. The AIC thus suffers from similar problems to LRTs (boundary!) when it is used to select parameters in $\boldsymbol{D}$ / random effects. It will choose the smaller model more often than you think.

Model Selection

There are approaches to use an AIC based on the conditional log-likelihood for this case (Vaida & Blanchard, Biometrika, 2005; Liang, Wu & Zhou, Biometrika, 2008). However, they still suffer from problems and more research is needed: The first approach always chooses the larger model, and the second approach is computationally very demanding and not very robust (Greven & Kneib, TechReport, 2008).

Model Diagnostics

Fully covering model diagnostics would lead to far here. A few remarks:

- **Normality assumption for the random effects**: Due to the shrinkage effect, the BLUPs $\widehat{\boldsymbol{b}}$ can look normal even if the true distribution for $\boldsymbol{b}$ is not normal (e.g. bimodal). Verbeke & Molenberghs, 2000 give a discussion on fitting mixture distributions for $\boldsymbol{b}$ instead to check for normality.
- **Residual diagnostics**: Remember that $\text{Cov}(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta) = \boldsymbol{V}_i$. Thus, the residual vector $\boldsymbol{r}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta}$ will have mean zero, but will be correlated. Diagnostics are typically based on standardized residuals $\boldsymbol{r}_i^* = \boldsymbol{L}_i^{-1}\boldsymbol{r}_i$, where $\widehat{\boldsymbol{V}}_i = \boldsymbol{L}_i\boldsymbol{L}_i'$ is the Cholesky decomposition. (E.g. residual plots, Mahalanobis distance between observed and predicted responses, semi-variogram.) $\boldsymbol{r}_i^*$ are approximately uncorrelated with unit variance.
  One could also use the subject-specific residuals $\boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\beta} - \boldsymbol{Z}_i\widehat{\boldsymbol{b}}_i$ to e.g. look at serial correlation. However, $\widehat{\boldsymbol{b}}_i$ very much depends on the normality assumption for $\boldsymbol{b}$, and is also influenced by the assumed structure for $\boldsymbol{V}_i$. Thus, these residuals are not well-suited to check assumptions on $\boldsymbol{V}_i$.
- **Influence diagnostics**: We can remove one subject from the analysis and recompute the parameter estimates. There are several distance measures to measure the influence of a single subject.

Linear Mixed Models in R

See `examples_mixed_models.R` for example code for

- linear mixed models
- additive mixed models
- testing for variance components.

A reference for Linear Mixed Models using the `lme()` function in the `nlme` package in R: Pinheiro & Bates, 2000.

A reference for Additive Mixed Models or Generalized Additive Mixed Models using the `mgcv` package in R: Wood, 2006.

For many examples of Linear Mixed Models in SAS, see Verbeke & Molenberghs, 2000.