

TMA4315 Generalized linear models H2018

Module 7: Linear mixed effects models

Mette Langaas, Department of Mathematical Sciences, NTNU,
with contributions from Ingeborg G. Hem

01.11 and 08.11 [PL], 02.11 and 09.11 [IL]

(Latest changes: 04.11.2018 - adjustments for week 2 (typos etc)).

(Warning: some changes may occur before the second week.)

Overview

Aim: Present methods for analysing correlated responses in a (normal/Gaussian) regression setting.

We will only consider *two-level models* and in particular focus on *random intercept and random slope models*.

Learning material

- ▶ Textbook: Fahrmeir et al (2013): Chapter 2.4, 7.1-7.3, 7.7. In greater detail: pages 349-354 (not "Alternative view on the random intercept model"), 356-365 (not 7.1.5 "Stochastic Covariates""), 368-377 (not "Bayesian Covariance Matrix""), 379-380 (not "Testing Random Effects or Variance Parameters"" , only last part on page 383), 383 (middle), 401-409 (orange juice). Note: Bayesian solutions not on the reading list.
- ▶ Alternative readings: Zuur et al. (2009): "Mixed Effects Models and Extensions in Ecology with R", chapter 5 (pages 101-142). Available as free ebook from Springer for NTNU students. More explanations and less mathematics than Fahrmeir et al (2013), more focus on understanding. Link to ebook Chapter 5
- ▶ Classnotes 01.11.2018
- ▶ Classnotes 08.11.2018

Topics

First week

- ▶ correlated responses - when and why?
 - ▶ repeated measurements: clustered and longitudinal
 - ▶ example of clustered data from ecology: species richness
- ▶ notation
- ▶ random intercept models
 - ▶ intra class correlation (ICC)
- ▶ linear mixed effects models
 - ▶ measurement model and distributional assumptions
 - ▶ conditional and marginal formulation
- ▶ parameter estimation
 - ▶ with maximum likelihood for fixed effects
 - ▶ distribution of fixed parameter estimators

Jump to interactive (week 1)

Second week

- ▶ what did we do last week: beaches example
- ▶ parameter estimation (cont.)
 - ▶ (restricted) maximum likelihood for random effects
- ▶ predicting
 - ▶ random effects: method, formula, plots
 - ▶ random errors: two types of residuals
- ▶ random slope models
 - ▶ sleep study
 - ▶ interpretation of random effects
- ▶ hypothesis tests
- ▶ model selection
- ▶ fitting LMM with function `lmer` in package `lme4`
- ▶ what have we not covered?

Jump to interactive (week 2)

FIRST WEEK

Correlated responses

We may get correlated responses when we work with repeated measurements on a set of units. The units may be:

- ▶ subjects, patients, participants
- ▶ animal, plants
- ▶ families, towns, schools, classes, beaches

We will consider two types of repeated measurements: clustered and longitudinal

Repeated measurements

Clustered data

The data are nested in the sense that a lower level unit can only belong to one higher level unit (cluster), and there is in general no natural ordering of the units within each cluster.

Two-level clustered - examples

- ▶ patients in hospitals
- ▶ siblings in families
- ▶ pupils in schools

Longitudinal data

The data for each individual are observed at multiple points in time

Two-level longitudinal data - examples

- ▶ patients with drug A and drug B monitored over time
- ▶ metabolic rate for patients at fasting, then 15, 45, 75 and 135 minutes after a heavy meal

Q: Why not only analyse using linear models?

A: Assuming independent responses when they are correlated does not give the correct estimate for the standard error of the estimated parameters of interest.

Example from ecology: beaches and species

This example is taken from Zuur et al. (2009, chapter 5, pages 101-142), and data are referred to as RIKZ because they were collected by a Dutch institute with that name.

Data were collected on nine different beaches in the Netherlands with the aim to investigate if there is a relationship between the

- ▶ richness of species (number of species observed) and
- ▶ NAP: the height of the sampling station compared to mean tidal level.

Data: 45 observations, taken at 9 beaches:

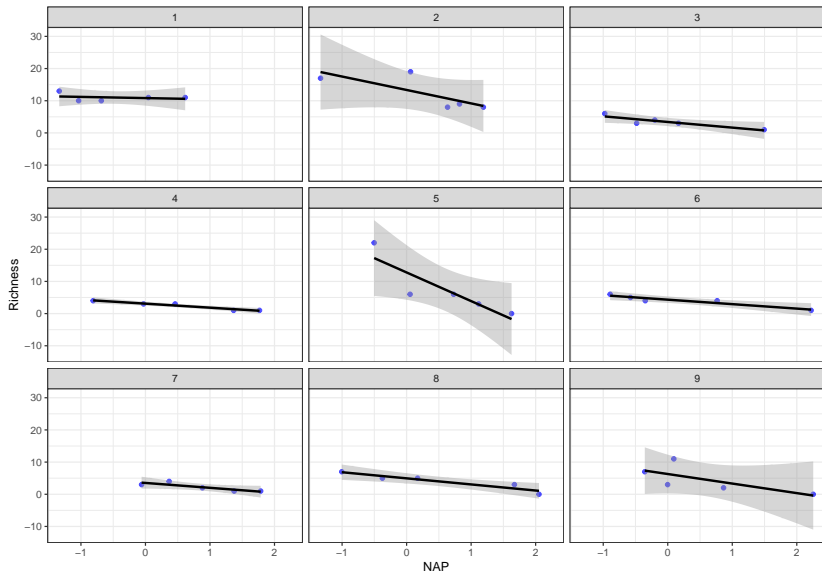
- ▶ beach: the beach that the samples were taken, for each beach 5 different samples were taken.
- ▶ exposure: an index composed of information about wave action, length of the surf zone, grain size, depth of anaerobic layer was measured, but we will not be used now.

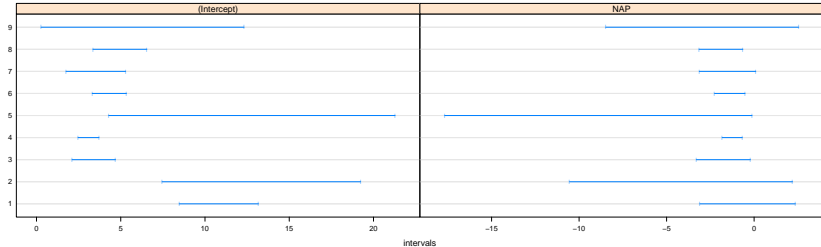
We want to use the richness of species as the response in a regression model. This is a count, so we could have used *Poisson regression*, but to make things simpler we assume that the counts are such that we instead can assume a *normal distribution* (in the next module on generalized linear mixed effects models we will use the Poisson distribution).

```
library("AED")
data(RIKZ)
summary(RIKZ)
```

```
##      Sample      Richness      Exposure      NAP
## Min.   : 1      Min.   : 0.000      Min.   : 8.00      Min.   : -1.3360
## 1st Qu.:12      1st Qu.: 3.000      1st Qu.:10.00     1st Qu.: -0.3750
## Median :23      Median : 4.000      Median :10.00     Median : 0.1670
## Mean   :23      Mean   : 5.689      Mean   :10.22     Mean   : 0.3477
## 3rd Qu.:34      3rd Qu.: 8.000      3rd Qu.:11.00     3rd Qu.: 1.1170
## Max.   :45      Max.   :22.000      Max.   :11.00     Max.   : 2.2550
##      Beach
## Min.   :1
## 1st Qu.:3
## Median :5
## Mean   :5
## 3rd Qu.:7
## Max.   :9
```

We first consider models with response Richness and one covariate NAP, and analyse data from each beach separately.





```
##           1           2           3           4           5
## (Intercept) 10.8218944 13.345694  3.400702  3.087716 12.782828  4.32
## NAP        -0.3718279 -4.175271 -1.755353 -1.248577 -8.900178 -1.38
##           7           8           9
## (Intercept)  3.520626  4.951455  6.295053
## NAP        -1.517613 -1.893066 -2.967530
```

The intercepts differ but the slopes are not that different.

Q: What if we now want to combine the regression models for the nine beaches into one model (for the population of beaches), so we can answer the question about relationship between species richness and NAP on the population level. What can we do then?

Possible solutions:

1. Use all 45 observations in a regression - with a common intercept and linear effect in NAP.

Problem: violation of assumption of independent observations lead to wrong estimates for variances of parameter estimates.

2. Add beach as covariate to regression - then we estimate one regression coefficient for each beach (intercepts) in addition to the linear effect in NAP.

Problem: why do we want to add 8 extra parameters to estimate (why 8?) values for the 9 beaches? Loss of power and what would we use the beach estimates for?

3. Add beach as a *random* covariate to the regression: this is called *random intercept models*.

Problem: new stuff - slightly complicated. We do this because beaches not of interest in themselves, only random sample from population of beaches, and therefore we only need to account for beaches, not estimate separate parameters.

Solution 1: all observations together - standard errors not correct

```
fitall = lm(Richness ~ NAP, data = RIKZ)
summary(fitall)
```

```
##
## Call:
## lm(formula = Richness ~ NAP, data = RIKZ)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.0675 -2.7607 -0.8029  1.3534 13.8723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6857      0.6578  10.164 5.25e-13 ***
## NAP          -2.8669      0.6307  -4.545 4.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.16 on 43 degrees of freedom
## Multiple R-squared:  0.3245, Adjusted R-squared:  0.3088
## F-statistic: 20.66 on 1 and 43 DF,  p-value: 4.418e-05
```

Solution 2: fixed effects for each beach - many estimates not so much of interest when population is in focus.

```
RIKZ$beachfactor = as.factor(RIKZ$Beach)
fitbeach = lm(Richness ~ NAP + beachfactor, data = RIKZ, contrasts = li
summary(fitbeach)
```

```
##
## Call:
## lm(formula = Richness ~ NAP + beachfactor, data = RIKZ, contrasts =
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.8518 -1.5188 -0.1376  0.7905 11.8384
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.5556     0.4885  13.421 2.30e-15 ***
## NAP           -2.4928     0.5023  -4.963 1.79e-05 ***
## beachfactor1    3.2503     1.3554   2.398  0.0220 *
## beachfactor2    6.3284     1.2908   4.903 2.15e-05 ***
## beachfactor3   -3.1546     1.3020  -2.423  0.0207 *
## beachfactor4   -2.7826     1.2943  -2.150  0.0386 *
## beachfactor5    2.3520     1.2967   1.814  0.0783 .
```

Notation

We have n_i repeated observations ($j = 1, \dots, n_i$) from each of $i = 1, \dots, m$ clusters (or individuals).

- ▶ Responses: $Y_{i1}, Y_{i2}, \dots, Y_{in_i}$ (e.g. species richness at beach i sample j)
- ▶ Covariates: $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{in_i}$ (e.g. NAP for beach i sample j)

The covariates \mathbf{x}_{ij} are $p \times 1$ vectors (as before, k covariates and one intercept so $p = k + 1$).

Random intercept models

Model

Simple linear regression

We start with a simple linear regression (one fixed effect).

First, only one covariate (in addition to the intercept), observed for cluster (beach) i on occasion j we have x_{ij}

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} \text{ where } \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma^2)$$

but, we know that Y_{i1} and Y_{i2} are observed for the same cluster and should not be independent (i.e. from same beach), to fix that we insert a *random intercept*.

Cluster-specific parameters γ_{0i}

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_{0i} + \varepsilon_{ij} \text{ where } \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma^2)$$

- ▶ β_0 : population intercept (fixed)
- ▶ γ_{0i} : deviation (for members of cluster i) from the population intercept β_0 - not a parameter but a random variable!
- ▶ $\beta_0 + \gamma_{0i}$: random intercept for cluster i
- ▶ β_1 : population slope (fixed), common to all clusters

The clusters are assumed to be random samples from a large population, and for the cluster deviation intercept we assume

$$\gamma_{0i} \sim N(0, \tau_0^2)$$

and that the γ_{0i} s and the ε_{ij} s are mutually independent random variables. So, we have now two error terms.

Q: What are now our unknown parameters?

A: $\gamma_0, \beta_1, \sigma^2, \tau_0^2$. Remark that the γ_{0i} s are not parameters but random variables.

Beach-example: Parameter estimates

(will talk about parameter estimation with ML and REML later)

Q: Try to identify $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2, \hat{\tau}_0^2$ in the print-out.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Richness ~ NAP + (1 | Beach)
## Data: RIKZ
##
## REML criterion at convergence: 239.5
##
## Scaled residuals:
##   Min      1Q  Median      3Q      Max
## -1.4227 -0.4848 -0.1576  0.2519  3.9794
##
## Random effects:
##   Groups Name          Variance Std.Dev.
##   Beach  (Intercept)  8.668   2.944
##   Residual                    9.362   3.060
## Number of obs: 45, groups: Beach, 9
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   6.5819    1.0958   6.007
## NAP          -2.5684    0.4947  -5.192
##
## Correlation of Fixed Effects:
##      (Intr)
## NAP -0.157
```

A: $\hat{\beta}_0 = 6.5818929$, $\hat{\beta}_1 = -2.5683996$, $\hat{\sigma}^2 = 9.3621916$, and $\hat{\tau}_0^2 = 8.6675181$

Intra class correlation (ICC)

The *conditional distribution* of Y_{ij} given the value of γ_{0i} (=regarding γ_{0i} as known) is

$$Y_{ij} \mid \gamma_{0i} \sim N(\beta_0 + \beta_1 x_{ij} + \gamma_{0i}, \sigma^2)$$

One motivation for inserting this new random intercept was to make sure that observations from the same cluster are dependent, but between clusters are independent. This means that we need to look at $\text{Cov}(Y_{ij}, Y_{kl})$ when $i = k$ and when $i \neq k$. To do that we need the (joint) *marginal distribution* of the responses.

What is the marginal distribution for Y_{ij} ?

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_{0i} + \varepsilon_{ij}$$

$$Y_{ij} \sim N(\beta_0 + \beta_1 x_{ij}, \sigma^2 + \tau_0^2)$$

We consider Y_{ij} , and Y_{kl} , where

$$Y_{kl} = \beta_0 + \beta_1 x_{kl} + \gamma_{0k} + \varepsilon_{kl} \sim N(\beta_0 + \beta_1 x_{kl}, \sigma^2 + \tau_0^2)$$

Now to the covariance between Y_{ij} and Y_{kl} .

$$\text{Cov}(Y_{ij}, Y_{kl}) = E[(Y_{ij} - \mu_{ij})(Y_{kl} - \mu_{kl})]$$

⊕ derivations on the board in class

$$\text{Cov}(Y_{ij}, Y_{kl}) = \begin{cases} \tau_0^2 + \sigma^2 = \text{Var}(Y_{ij}) & \text{for } i = k, j = l \\ \tau_0^2 & \text{for } i = k, j \neq l \\ 0 & \text{for } i \neq k, j \neq l \end{cases}$$

If we put this into a covariance matrix for the vector of responses for cluster i this type of structure is called *compound symmetry*.

⊕ write this out on the board in class

$$\text{Cov}(\mathbf{Y}_i) = \tau_0^2 \mathbf{1}\mathbf{1}^T + \sigma^2 \mathbf{I}$$

where as before $\mathbf{1}$ is a $n_i \times 1$ vector of 1s and \mathbf{I} is a $n_i \times n_i$ identity matrix.

The correlation between Y_{ij} and Y_{il} (two observations in the same cluster that is - same beach) is called the *within subject* or *within cluster* correlation coefficient, and is for our random intercept model

$$\text{Corr}(Y_{ij}, Y_{il}) = \frac{\text{Cov}(Y_{ij}, Y_{il})}{\sqrt{\text{Var}(Y_{ij})\text{Var}(Y_{il})}} = \frac{\tau_0^2}{\tau_0^2 + \sigma^2} \text{ for } j \neq l$$

Inserted parameter estimates this is called the *intra class correlation (ICC)* for the random intercept model.

Beach-example: ICC

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Richness ~ NAP + (1 | Beach)
## Data: RIKZ
##
## REML criterion at convergence: 239.5
##
## Scaled residuals:
##   Min       1Q   Median       3Q      Max
## -1.4227 -0.4848 -0.1576  0.2519  3.9794
##
## Random effects:
##   Groups Name      Variance Std.Dev.
##   Beach  (Intercept) 8.668    2.944
##   Residual                9.362    3.060
## Number of obs: 45, groups: Beach, 9
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   6.5819    1.0958   6.007
## NAP          -2.5684    0.4947  -5.192
##
## Correlation of Fixed Effects:
##   (Intr)
## NAP -0.157
```

Q: What is the ICC for our fit? Hint:

$$\text{Corr}(Y_{ij}, Y_{il}) = \frac{\tau_0^2}{\tau_0^2 + \sigma^2} \text{ for } j \neq l.$$

A: $8.688 / (8.688 + 9.362) = 0.48.$

Summing up - so far

We have now looked at models with a random intercept, which is a special case of linear mixed models (LMM). We have now see that we have used two components in our model:

- ▶ fixed effects - like we have used so far in our GLM-course. This can be gender, age, time, experimental condition.
- ▶ random effects - used to model correlated responses. Remember the correlation (ICC) in the previous example.

But, what if we also need the different clusters to have different slopes? The generalization of the random intercept model is the *random slope* model which can be written as:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_{0i} + \gamma_{1i} x_{ij} + \varepsilon_{ij}$$

and where the new parts are

- ▶ β_1 : population slope (fixed), common to all clusters
- ▶ γ_{1i} : deviation (for members of cluster i) from the population slope β_0 - not a parameter but a random variable!
- ▶ $\beta_1 + \gamma_{1i}$: random slope for cluster i .

More about this model in week 2.

We will now imagine that the fixed effects can be of form $\mathbf{X}_i \boldsymbol{\beta}$ - as before (in previous modules - but now remember that we have n_i observations for cluster i so therefore we write \mathbf{X}_i and not \mathbf{x}_i^T). But it is also possible to expand the random effects into a similar form $\mathbf{U}_i \boldsymbol{\gamma}$.

Linear mixed effects model (LMM)

We now define the LMM from a measurement model and distributional assumption for clusters $i = 1, \dots, m$.

Measurement model

for the i th cluster:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{U}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i$$

- ▶ \mathbf{Y}_i : $n_i \times 1$ random vector of responses
- ▶ \mathbf{X}_i : $n_i \times p$ design matrix
- ▶ \mathbf{U}_i : $n_i \times (q + 1)$ design matrix for random effects
- ▶ $\boldsymbol{\beta}$: $p \times 1$ vector of fixed coefficients (common for all clusters)
- ▶ $\boldsymbol{\gamma}_i$: $(q + 1) \times 1$ random vector
- ▶ $\boldsymbol{\varepsilon}_i$: $n_i \times 1$ random vector

Distributional assumptions

for the i th cluster:

$$\gamma_i \sim N(\mathbf{0}, \mathbf{Q})$$

$$\varepsilon_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

All elements of $\gamma_1, \gamma_2, \dots, \gamma_m$ and $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m$ are mutually independent. The dimension of \mathbf{I} is $n_i \times n_i$, and \mathbf{Q} is $(q+1) \times (q+1)$.

Remark: one possible generalization is to assume $\varepsilon_i \sim N(\mathbf{0}, \Sigma)$ where Σ is general, but we will not look into that in our course. This can be needed for example if time series structure (like AR1) is present.

Questions

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{U}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i$$

$$\boldsymbol{\gamma}_i \sim N(\mathbf{0}, \mathbf{Q})$$

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

Q:

- ▶ What is \mathbf{U}_i , $\boldsymbol{\gamma}_i$ and \mathbf{Q} for the random intercept model?
- ▶ General: what is the marginal distribution of \mathbf{Y}_i ?

A:

- ▶ We had $\mathbf{U}_i = \mathbf{1}$ ($n_i \times 1$) and scalar $\gamma_i = \gamma_{0i}$, and $\mathbf{Q} = \tau_0^2$.
- ▶ $\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{U}_i\mathbf{Q}\mathbf{U}_i^T + \sigma^2\mathbf{I})$.

Conditional formulation

Conditional Gaussian model for the response \mathbf{Y}_i given the random effect γ_i :

$$\mathbf{Y}_i | \gamma_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{U}_i\gamma_i, \sigma^2\mathbf{I})$$

Marginal Gaussian model

for the response for cluster i , \mathbf{Y}_i (Laird and Ware (1982) formulation)

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{U}_i\gamma_i + \boldsymbol{\varepsilon}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^*$$

$$\boldsymbol{\varepsilon}_i^* = \mathbf{U}_i\gamma_i + \boldsymbol{\varepsilon}_i$$

$$E(\boldsymbol{\varepsilon}_i^*) = \mathbf{0}$$

$$\mathbf{V}_i = \text{Cov}(\boldsymbol{\varepsilon}_i^*) = \text{Cov}(\mathbf{U}_i\gamma_i) + \text{Cov}(\boldsymbol{\varepsilon}_i) = \mathbf{U}_i\mathbf{Q}\mathbf{U}_i^T + \sigma^2\mathbf{I}$$

$$\boldsymbol{\varepsilon}_i^* \sim N(\mathbf{0}, \mathbf{V}_i)$$

Global model

From the cluster specific model:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{U}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i$$

into the global model for all clusters:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix}, \mathbf{U} = \begin{pmatrix} \mathbf{U}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_2 & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{U}_m \end{pmatrix}, \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \vdots \\ \boldsymbol{\gamma}_m \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_m \end{pmatrix}$$

Let $N = \sum_{i=1}^m n_i$, then dimensions are:

- ▶ $\mathbf{Y}, \boldsymbol{\varepsilon}: N \times 1$
- ▶ $\mathbf{X}: N \times p$
- ▶ $\boldsymbol{\beta}: p \times 1$
- ▶ $\mathbf{U}_-: N \times m(q+1)$
- ▶ $\boldsymbol{\gamma}: m(q+1) \times 1$

Conditional Gaussian model

for the response \mathbf{Y} given the random effect γ :

$$\mathbf{Y} \mid \gamma \sim N(\mathbf{X}\beta + \mathbf{U}\gamma, \sigma^2\mathbf{I})$$

Now \mathbf{I} is $N \times N$ where $N = \sum_{i=1}^m n_i$.

Marginal Gaussian model

for the response \mathbf{Y}

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$$

$$\boldsymbol{\varepsilon}^* = \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

$$\mathbf{V} = \text{Cov}(\boldsymbol{\varepsilon}^*) = \text{Cov}(\boldsymbol{\varepsilon}) + \text{Cov}(\mathbf{U}\boldsymbol{\gamma}) = \sigma^2\mathbf{I} + \mathbf{U}\mathbf{G}\mathbf{U}^T$$

$$\boldsymbol{\varepsilon}^* \sim N(\mathbf{0}, \mathbf{V})$$

Here \mathbf{G} is a $m(q+1)$ block-diagonal matrix with \mathbf{Q} m times on the diagonal (see below). This gives

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V} = \sigma^2\mathbf{I} + \mathbf{U}\mathbf{G}\mathbf{U}^T)$$

$$\mathbf{G} = \begin{pmatrix} \mathbf{Q} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{Q} \end{pmatrix}$$

Parameter estimation

- ▶ Fixed effects β : estimated using maximum likelihood
- ▶ Random effects parameters σ^2 and \mathbf{Q} (in \mathbf{V}): estimated using restricted maximum likelihood (REML).

For the random effects γ_i and ε_i we also provide predictions

- ▶ Predicted values for the random effects γ_i using *best linear unbiased predictors* (BLUP).
- ▶ Prediction values for the random effects ε_i are our *residuals*. Two types of residuals possible “response minus fixed effects”“, or”response minus fixed and predicted random effects“.

Parameter estimation with maximum likelihood for fixed effects

Cluster specific model:

$$\mathbf{Y}_i \sim N(\mu_i = \mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i = \sigma^2\mathbf{I} + \mathbf{U}_i\mathbf{Q}\mathbf{U}_i^T)$$

Global model:

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V} = \sigma^2\mathbf{I} + \mathbf{U}\mathbf{G}\mathbf{U}^T)$$

The log-likelihood function is then (\pm some constants)

$$l(\boldsymbol{\beta}, \mathbf{V}) = -\frac{1}{2} \ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

We assume that the parameters in \mathbf{V} are **known**, and transform our problem with $\mathbf{V}^{-1/2}$ to get the standard multiple linear regression model.

In TMA4267 Exam 2014 Problem 4, the weighted least squares estimator was derived from a similar situation. Now we have

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$$

where

$$\boldsymbol{\varepsilon}^* \sim N(\mathbf{0}, \mathbf{V})$$

We then define $\mathbf{V}^{-1/2}$ based on eigenvalue-vector decomposition of \mathbf{V} , and premultiply the above equation to get:

$$\mathbf{V}^{-1/2}\mathbf{Y} = \mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-1/2}\boldsymbol{\varepsilon}^*$$

$$\mathbf{Y}^\bullet = \mathbf{X}^\bullet\boldsymbol{\beta} + \boldsymbol{\varepsilon}^\bullet$$

where $\boldsymbol{\varepsilon}^\bullet \sim N(\mathbf{0}, \mathbf{I})$. We then know that the ML estimator for $\boldsymbol{\beta}$ is given as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\bullet T}\mathbf{X}^\bullet)^{-1}\mathbf{X}^{\bullet T}\mathbf{Y}^\bullet$$

Inserting then for $\mathbf{X}^\bullet T$ and \mathbf{Y}^\bullet leads to the weighted least squares solution for β

$$\hat{\beta} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y}$$

Since we have independence between clusters, we had block-diagonal \mathbf{V} and then

$$\hat{\beta} = \left(\sum_{i=1}^m \mathbf{x}_i^T \mathbf{v}_i^{-1} \mathbf{x}_i \right)^{-1} \sum_{i=1}^m \mathbf{x}_i^T \mathbf{v}_i^{-1} \mathbf{y}_i$$

But, we really don't know \mathbf{V} so an estimate must be inserted (week 2 with REML).

Beach-example: parameter estimation for fixed effects

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Richness ~ NAP + (1 | Beach)
## Data: RIKZ
##
## REML criterion at convergence: 239.5
##
## Scaled residuals:
##   Min      1Q  Median      3Q      Max
## -1.4227 -0.4848 -0.1576  0.2519  3.9794
##
## Random effects:
##   Groups Name          Variance Std.Dev.
##   Beach  (Intercept)  8.668   2.944
##   Residual                9.362   3.060
## Number of obs: 45, groups: Beach, 9
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   6.5819    1.0958   6.007
## NAP           -2.5684    0.4947  -5.192
##
## Correlation of Fixed Effects:
##   (Intr)
## NAP -0.157
```

Q: Where is the parameter estimate for effect of NAP, with standard deviation? What is the interpretation of “Correlation of Fixed Effects”?

A:“Fixed effects, NAP, Estimate and Std. Error”. Correlation of fixed effects is the off-diagonal elements of $\text{Cov}(\hat{\beta})$ which here is only $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$.

Properties of parameters estimators

Since $\hat{\beta}$ is a linear function of \mathbf{Y}_i , it has a multivariate normal distribution - let us write with

$$\hat{\beta} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y} = \mathbf{A} \mathbf{Y}$$

Mean:

$$E(\hat{\beta}) = \mathbf{A} E(\mathbf{Y}) = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \beta = \beta$$

so $\hat{\beta}$ is unbiased.

Variance-covariance matrix:

$$\begin{aligned}\text{Cov}(\hat{\beta}) &= \mathbf{A}\text{Cov}(\mathbf{Y})\mathbf{A}^T = \mathbf{A}\mathbf{V}\mathbf{A}^T \\ &= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\end{aligned}$$

Since \mathbf{V} is block diagonal this can be written as a sum with the clusters:

$$\text{Cov}(\hat{\beta}) = \left(\sum_{i=1}^m \mathbf{x}_i^T \mathbf{V}_i^{-1} \mathbf{x}_i\right)^{-1}$$

Also, if we insert $\hat{\mathbf{V}}$ as an estimate for \mathbf{V} then

$$\hat{\beta} = \left(\sum_{i=1}^m \mathbf{x}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{x}_i \right)^{-1} \sum_{i=1}^m \mathbf{x}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{y}_i$$

is asymptotic multivariate normal (under regularity conditions).

Remark: this is only asymptotically, so we need large samples for this to hold. And, we do not arrive at a t -distribution here - however approximations for t (and F) exists, but the problem is the number of degrees of freedom (in particular when we have time varying fixed effects).

In this course we will only use the asymptotic normality for confidence intervals (and when a Wald type test is desired). We will consider hypothesis testing with likelihood ratio test under “Model selection” in the end of this module.

Beach-example: confidence interval for fixed effects

```
##                2.5 %    97.5 %  
## .sig01         1.484204  5.100814  
## .sigma         2.435429  3.877572  
## (Intercept)   4.324560  8.824909  
## NAP           -3.566901 -1.599779
```

Q: Interpret! First the last two rows. The first two we have not covered yet - since we don't know how to estimate the parameters in \mathbf{V} but what do you think this is?

A: Estimate ± 1.96 times standard error of estimate.

Interactive session week 1

Plan:

- ▶ Problem 1abcd (not ef - covered in week 2).
- ▶ Problem 2abe (not cd - covered in week 2).
- ▶ If time: Exam 2017, Problem 3

Exercise 1: Taken from UiO, STK3100, 2011, problem 1

We will in this exercise look at the following model:

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \gamma_{0i} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\gamma_{0i} \sim N(0, \tau_0^2)$$

where $i \in \{1, \dots, m\}$ is the group index, and $j \in \{1, \dots, n_i\}$ is the index for repeated measurements in each group (cluster). We assume that all random variables are independent.

- a)** What is this model called? Discuss where and when such models are useful.
- b)** What is the *marginal* model for $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})$? Show how you arrive at your answer and write with vectors and matrices (for cluster i).

What are the advantages of having an expression for the marginal distribution of \mathbf{Y}_i when doing estimation?

In the rest of the exercise we will look at a dataset consisting of

Exercise 2: Taken from UiO, STK3100, 2013, problem 2

The data in this problem is part of a longitudinal study of income in the US, the Panel Study of Income Dynamics, begun in 1968. The subset consists of 42 heads of household who were aged 25-39 in 1968. The variables included are

- ▶ annual nominal income, which is the response variable
- ▶ age, age in 1968
- ▶ cyear, coded as -10 in 1968, 0 in 1978 and 10 in 1988
- ▶ educ, years of education in 1968
- ▶ sex, M = male, F = female
- ▶ person, containing information on person id

Below is an excerpt from the output from fitting a linear mixed model (LMM) with the function `lmer` in R. The log income (`lincm`) is the response,

$$\text{lincm}_{ij} = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{cyear}_{ij} + \beta_3 \text{educ}_i + \beta_4 \text{sex}_i + \gamma_{0i} + \varepsilon_{ij}$$

SECOND WEEK

Notation and LMM

- ▶ In the first week we started with models with a random intercept, using the beach-example - which we saw was a special case of linear mixed models (LMM).
- ▶ We looked at observations from clusters (family, beach) and will now also look at longitudinal data (repeated measurements on the same units).
- ▶ For cluster i we wrote the LMM model in a *measurement model* part and a *distributional assumptions* part:

Words to know: measurement model, distributional model, conditional model, marginal model, global model.

Measurement model

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{U}_i\gamma_i + \varepsilon_i$$

Distributional model

$$\gamma_i \sim N(\mathbf{0}, \mathbf{Q})$$

$$\varepsilon_i \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

This gave the **marginal model for each cluster**:

$$\mathbf{Y}_i \sim N(\mu_i = \mathbf{X}_i\beta, \mathbf{V}_i = \sigma^2\mathbf{I} + \mathbf{U}_i\mathbf{Q}\mathbf{U}_i^T)$$

Q: explain what the different parameters and random variables are, and what are their dimensions.

A:

- ▶ design matrix for fixed effects for cluster i : \mathbf{X}_i is $n_i \times p$ (intercept included)
- ▶ parameter vector for fixed effects: β $p \times 1$
- ▶ design matrix for random effects for cluster i : \mathbf{U}_i is $n_i \times (q + 1)$ (intercept included)
- ▶ random effects - used to model correlated responses: γ_i , $(q + 1) \times 1$
- ▶ random errors: ε_i $n_i \times 1$
- ▶ covariance matrix for the random effects: \mathbf{Q} , $(q + 1) \times (q + 1)$
- ▶ parameter σ^2 for variance of the random errors $\text{Cov}(\varepsilon_i) = \sigma^2 \mathbf{I}$ with dimension $n_i \times n_i$.

Beach-example: Parameter estimation

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Richness ~ NAP + (1 | Beach)
## Data: RIKZ
##
## REML criterion at convergence: 239.5
##
## Scaled residuals:
##   Min      1Q  Median      3Q      Max
## -1.4227 -0.4848 -0.1576  0.2519  3.9794
##
## Random effects:
##   Groups Name      Variance Std.Dev.
##   Beach  (Intercept) 8.668    2.944
##   Residual                9.362    3.060
## Number of obs: 45, groups: Beach, 9
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   6.5819    1.0958   6.007
## NAP          -2.5684    0.4947  -5.192
##
## Correlation of Fixed Effects:
##   (Intr)
## NAP -0.157
```

Global model

For all cluster together (even more letters now)

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}\gamma + \varepsilon$$

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix}, \mathbf{U} = \begin{pmatrix} \mathbf{U}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_2 & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{U}_m \end{pmatrix}$$

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$$

$$\boldsymbol{\varepsilon}^* = \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

$$\mathbf{V} = \text{Cov}(\boldsymbol{\varepsilon}^*) = \text{Cov}(\boldsymbol{\varepsilon}) + \text{Cov}(\mathbf{U}\boldsymbol{\gamma}) = \sigma^2\mathbf{I} + \mathbf{U}\mathbf{G}\mathbf{U}^T$$

$$\boldsymbol{\varepsilon}^* \sim N(\mathbf{0}, \mathbf{V})$$

Here \mathbf{G} is a $m(q+1)$ block-diagonal matrix with \mathbf{Q} m times on the diagonal, which gives

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V} = \sigma^2\mathbf{I} + \mathbf{U}\mathbf{G}\mathbf{U}^T)$$

Parameter estimation

Fixed effects β (repetition)

estimated using maximum likelihood, with the marginal distribution as starting point:

$$\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{V} = \sigma^2\mathbf{I} + \mathbf{U}\mathbf{G}\mathbf{U}^T)$$

We assume that the parameters in \mathbf{V} are **known**, then we get the weighted least squares solution for β .

$$\hat{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y} = \left(\sum_{i=1}^m \mathbf{x}_i^T\mathbf{V}_i^{-1}\mathbf{x}_i\right)^{-1} \sum_{i=1}^m \mathbf{x}_i^T\mathbf{V}_i^{-1}\mathbf{Y}_i$$

$$\hat{\beta} \sim N\left(\beta, \left(\sum_{i=1}^m \mathbf{x}_i^T\mathbf{V}_i^{-1}\mathbf{x}_i\right)^{-1}\right)$$

We insert estimates for \mathbf{V}_i (which we will find next), and the same distribution - but only asymptotically - to be used for inference for the fixed effects.

$$\hat{\beta} = \left(\sum_{i=1}^m \mathbf{x}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{x}_i \right)^{-1} \sum_{i=1}^m \mathbf{x}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{Y}_i$$
$$\approx N\left(\beta, \sum_{i=1}^m (\mathbf{x}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{x}_i)^{-1}\right)$$

Now follows:

- ▶ Random effects parameters σ^2 and \mathbf{Q} (in \mathbf{V}): estimated using restricted maximum likelihood (REML). We denote all parameters for random effect for ϑ . For the random intercept model this is $\vartheta = (\sigma^2, \tau_0^2)$.

Then, for the random effects γ_i and ε_i we also provide predictions

- ▶ Predicted values for the random effects γ_i using *best linear unbiased predictors* (BLUP).
- ▶ Prediction values for the random effects ε_i are our *residuals*.

Parameter estimation with restricted maximum likelihood (REML) for random effects

There are two ways to explain the REML - a transformation method (we start with this), and an integration method (to come next).

Transformation method

To aid in our understanding we start by looking at the REML solution for **multipele linear regression (Module 2)**,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ with } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

where \mathbf{X} is a $n \times p$ design matrix. Remember that $\text{SSE} = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}$ where the hat matrix is $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

- ▶ We found that the maximum likelihood estimator for σ^2 was $\hat{\sigma}^2 = \frac{\text{SSE}}{n}$, which is found from maximizing the likelihood inserted our estimate of $\hat{\boldsymbol{\beta}}$ (i.e. disregarding the uncertainty in the estimation).

- ▶ This estimator is biased, and has mean $E(\hat{\sigma}^2) = \frac{n-p}{n}\sigma^2$ (too small= biased downwards), where n is the number of observations and p the number of parameters estimated.
- ▶ It is possible to find an $n \times (n - p)$ matrix \mathbf{A} such that $\mathbf{A}^T\mathbf{Y}$ follows a $n - p$ -dimensional multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{A}^T\mathbf{A}\sigma^2$.
- ▶ This means that we have eliminated β as unknown parameter and we can proceed to use maximum likelihood on the $n - p$ -dimensional vector $\mathbf{A}^T\mathbf{Y}$ with σ^2 as the only unknown parameter, which will give the parameter estimator

$$\hat{\sigma}^2 = \text{SSE}/(n - p) = \mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}/(n - p)$$

This is called the REML estimate for σ^2 .

Remark: There are many solutions to \mathbf{A} but to get $E(\mathbf{A}^T \mathbf{Y}) = \mathbf{A}^T \mathbf{X} \beta = \mathbf{0}$ then \mathbf{A} can be chosen to have linearly independent columns orthogonal to column space of the design matrix.

Remark: We can not choose $\mathbf{A} = \mathbf{I} - \mathbf{H}$ since we need \mathbf{A} to have dimension $n \times n - p$. But we can for example choose an (orthogonal) basis with $n - p$ vectors for the column space of $\mathbf{I} - \mathbf{H}$.

Now, move to our **linear mixed effects model**. We have the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

with the marginal distribution

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V} = \sigma^2\mathbf{I} + \mathbf{U}\mathbf{G}\mathbf{U}^T)$$

- ▶ The REML estimator for the parameters in \mathbf{V} (called ϑ , and for the random intercept model that is σ^2 and τ_0^2) - and also then $\mathbf{V}(\vartheta)$ - are now
- ▶ found by maximizing the likelihood for $\mathbf{A}^T\mathbf{Y}$
- ▶ where \mathbf{A} is any $N \times (N - p)$ full-rank matrix with columns orthogonal to the columns of the design matrix \mathbf{X} .
- ▶ Again $\mathbf{A}^T\mathbf{Y}$ follows a multivariate normal distribution with mean vector $\mathbf{0}$ and now covariance matrix $\mathbf{A}^T\mathbf{V}(\vartheta)\mathbf{A}$, which is independent of $\boldsymbol{\beta}$.

- ▶ The maximization does not give a closed form solution, but we get a new $\hat{\mathbf{V}}$ - which now will be less biased (sadly only unbiased in “simple and balanced cases”).
- ▶ Even if β is not estimated in this optimization we already know that

$$\hat{\beta} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y}$$

and now we have a new $\hat{\mathbf{V}}$ which we insert in this equation, and thus get a new REML-estimator for β :

$$\hat{\beta} = (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{Y}$$

- ▶ This means, that when using REML-estimation for our linear mixed effects model this will influence both the fixed effects and the random effects parameters. However, asymptotically we will still have the same asymptotic distribution for the fixed effects as with ML estimation.

In addition the main justification for using REML is that in the absence of information on β then no information about the parameters in ϑ is lost when likelihood estimation is based on $\mathbf{A}^T\mathbf{Y}$ instead of on \mathbf{Y} . In statistical inference this is referred to as $\mathbf{A}^T\mathbf{Y}$ *is marginally sufficient for ϑ* (but this is way beyond the scope of this course).

In addition, according to Verbeke and Molenbergs (2000, page 46, Equation 5.8), the likelihood function is

$$L(\vartheta) = C \left| \sum_{i=1}^m \mathbf{x}_i^T \mathbf{V}(\vartheta)_i^{-1} \mathbf{x}_i \right|^{-1/2} L_{ML}(\hat{\beta}(\vartheta), \vartheta)$$

where C is a constant not depending on ϑ . Since also the term $\left| \sum_{i=1}^m \mathbf{x}_i^T \mathbf{V}(\vartheta)_i^{-1} \mathbf{x}_i \right|$ does not depend on β .

Therefore both β and ϑ can be found by maximizing what is referred to as the *REML likelihood function*:

$$L_{REML}(\vartheta, \beta) = \left| \sum_{i=1}^m \mathbf{x}_i^T \mathbf{V}(\vartheta)_i^{-1} \mathbf{x}_i \right|^{-1/2} L_{ML}(\beta, \vartheta)$$

Further reading: Theoretical explanation for REML (beyond the scope of this course) by Inge Helland, UiO and also by Verbeke and Molenberghs (2000), Section 5.3 (free ebook from Springer for NTNU students).

Comparing ML and REML estimation for the beaches example

```
fitREML = lmer(Richness ~ NAP + (1 | Beach), data = RIKZ)
fitML = lmer(Richness ~ NAP + (1 | Beach), data = RIKZ, REML = FALSE)
REMLest = c(fixef(fitREML), as.data.frame(VarCorr(fitREML))[, 4])
MLEst = c(fixef(fitML), as.data.frame(VarCorr(fitML))[, 4])
df = data.frame(REML = REMLest, ML = MLEst)
rownames(df) = c("$\\beta_0$", "$\\beta_1$", "$\\tau_0$", "$\\sigma$")
kable(df, digits = 4)
```

	REML	ML
β_0	6.5819	6.5844
β_1	-2.5684	-2.5757
τ_0	8.6675	7.5068
σ	9.3622	9.1110

Q: Comment on what you see.

Remark: the default for `lmer` is REML, and we need to write `REML=FALSE` to get ML.

Integration method

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\vartheta) = \sigma^2\mathbf{I} + \mathbf{UGU}^T)$$

For the fixed effects we started with the log-likelihood function and maximized to get estimator for $\boldsymbol{\beta}$ dependent on ϑ . If we now assume that we have found $\hat{\boldsymbol{\beta}}(\vartheta)$ and insert this estimate into the loglikelihood then the *profile log-likelihood* is (disregarding an additive constant)

$$l_P(\vartheta) = -\frac{1}{2} \ln|\mathbf{V}(\vartheta)| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\vartheta))^T \mathbf{V}(\vartheta)^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\vartheta))$$

The integration method (can be motivated from the Bayesian perspective by assuming a flat prior on β) constructs a *marginal or restricted log-likelihood* by integrating β out of the likelihood

$$l_{\text{REML}}(\vartheta) = \ln \int L(\beta, \vartheta) d\beta$$

It can be shown that the REML log-likelihood is

$$l_{\text{REML}}(\vartheta) = l_P(\vartheta) - \frac{1}{2} \ln \left| \sum_{i=1}^m \mathbf{x}_i^T \mathbf{V}(\vartheta)_i^{-1} \mathbf{x}_i \right|$$

Maximizing of $l_{\text{REML}}(\vartheta)$ provides the REML estimator for ϑ .

What do you need to know about REML?

- ▶ That REML is used to get a better estimator (less downwards biased) for the random effects parameters than using ML,
- ▶ so REML is the default choice in the `lmer` function for fitting LMMs in the `lme4`-package in R.
- ▶ Two ways of motivating this: by transformation or by integration.
- ▶ But sadly, for LMM this does not in general give unbiased estimates for the parameters ϑ in \mathbf{V} - but less biased.

REML estimation for the beaches example

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Richness ~ NAP + (1 | Beach)
## Data: RIKZ
##
## REML criterion at convergence: 239.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.4227 -0.4848 -0.1576  0.2519  3.9794
##
## Random effects:
## Groups   Name                Variance Std.Dev.
## Beach    (Intercept)  8.668    2.944
## Residual                    9.362    3.060
## Number of obs: 45, groups: Beach, 9
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   6.5819    1.0958   6.007
## NAP           -2.5684    0.4947  -5.192
##
## Correlation of Fixed Effects:
##      (Intr)
## NAP -0.157
```

Q: What have we covered so far, and what is missing? Explain the elements of the print-out!

ML estimation for the beaches example

```
## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: Richness ~ NAP + (1 | Beach)
## Data: RIKZ
##
##      AIC      BIC   logLik deviance df.resid
##  249.8    257.1   -120.9   241.8      41
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.4258 -0.5010 -0.1791  0.2452  4.0452
##
## Random effects:
##  Groups   Name                Variance Std.Dev.
##  Beach    (Intercept)  7.507    2.740
##  Residual                    9.111    3.018
## Number of obs: 45, groups: Beach, 9
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   6.5844    1.0321   6.380
## NAP           -2.5757    0.4873  -5.285
##
## Correlation of Fixed Effects:
##      (Intr)
## NAP -0.164
```

Q: Look for differences between the REML and ML output.

Prediction of random effects and random errors

Predicted values for random effects γ

Why do we want a prediction? To rank the beaches (or schools, patients?). Small-area-estimations.

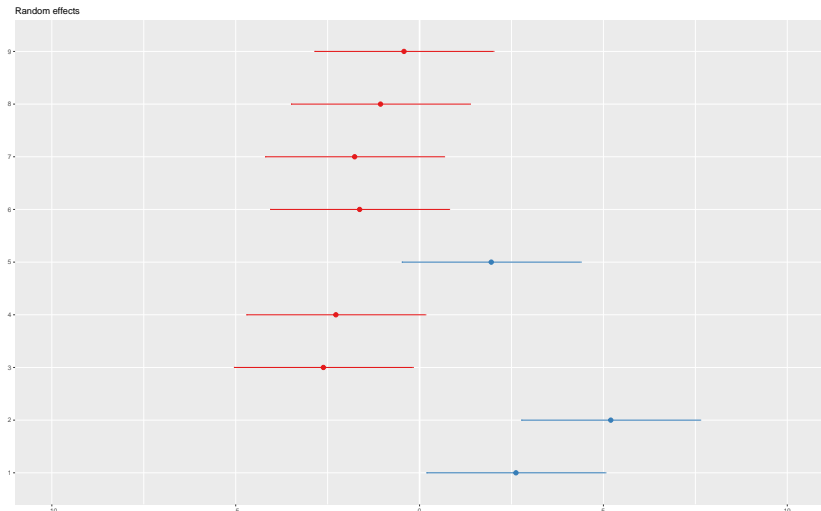
Model check can also use this to check that γ is normal is in agreement with our fitted model (same as when using residuals to check distribution of errors).

Best Linear Unbiased Predictor (BLUP) $\hat{\gamma}_i$

- ▶ linear function in \mathbf{Y} (linear)
- ▶ $E(\hat{\gamma}_i) = \mathbf{0}$ (unbiased)
- ▶ for any linear combination $\mathbf{a}^T \gamma_i$ of random effects $[E(\mathbf{a}^T \hat{\gamma}_i - \mathbf{a}^T \gamma_i)]^2$ is minimized among all such linear unbiased predictors (best)

Beaches random intercept - predicted intercept and estimated fixed effects

```
fit = lmer(Richness ~ NAP + (1 | Beach), data = RIKZ)  
plot_model(fit, type = "re")
```



Joint distribution of \mathbf{Y} and γ

The joint distribution of \mathbf{Y} and γ is:

$$\begin{pmatrix} \mathbf{Y} \\ \gamma \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{X}\beta \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V} = \mathbf{UGU}^T + \sigma^2\mathbf{I} & \mathbf{UG} \\ \mathbf{GU}^T & \mathbf{G} \end{pmatrix}\right)$$

Maximizing the likelihood based on the joint distribution

This is maximized with respect to β and γ , to give

$$\hat{\gamma} = \mathbf{G}\mathbf{U}^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

But, here the elements of \mathbf{G} and \mathbf{V} needs to be estimated, and we get:

$$\hat{\gamma} = \hat{\mathbf{G}}\mathbf{U}^T\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

$$\hat{\gamma}_i = \hat{\mathbf{Q}}\mathbf{U}_i^T\hat{\mathbf{V}}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\hat{\beta})$$

Remark: For details on this calculation - involving the Henderson's mixed model equations, see pages 371-372 of Fahrmeir et al (2013), or pages 98-99 in Verbeke and Molenberghs (2000), Section 5.3 (free ebook from Springer for NTNU students).

Conditional mean

Remember:

$$\begin{pmatrix} \mathbf{Y} \\ \gamma \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{X}\beta \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V} = \mathbf{UGU}^T + \sigma^2\mathbf{I} & \mathbf{UG} \\ \mathbf{GU}^T & \mathbf{G} \end{pmatrix}\right)$$

(Alternatively) The predicted random effects can be found as the mean of the conditional distribution of γ given \mathbf{Y} . If we also calculate the covariance of the estimated γ we can make approximate prediction intervals for the predicted random effects.

The general formula for the conditional multivariate normal \mathbf{X} (known from TMA4267) is:

$$\mathbf{X} \sim N(\mu, \Sigma)$$

$$\mathbf{X}_2 \mid (\mathbf{X}_1 = \mathbf{x}_1) \sim N(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$$

If we use the formula for the mean with $\mathbf{X}_1 = \mathbf{Y}$ and $\mathbf{X}_2 = \gamma$, then

$$E(\gamma | \mathbf{Y}) = \mathbf{0} + \mathbf{GU}^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\beta)$$

which can be used (inserted parameter estimates) to give our estimated random effects.

We may find the covariance matrix of $\mathbf{GU}^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\beta})$ (directly), and for each $\hat{\gamma}_i$ this is given as

$$\mathbf{QU}_i^T \left(\mathbf{V}_i^{-1} - \mathbf{V}_i^{-1}\mathbf{x}_i \left(\sum_{i=1}^m \mathbf{x}_i^T \mathbf{V}_i^{-1} \mathbf{x}_i \right)^{-1} \mathbf{x}_i^T \mathbf{V}_i^{-1} \right) \mathbf{U}_i \mathbf{Q}$$

(according to Verbeke and Molenberghs (2000), page 78). We insert estimates \mathbf{Q} and \mathbf{V}_i (thus underestimating the variability) and get the estimated covariance matrix for the random effect. Such an estimate is used in the caterpillar plot below.

Random intercept models: $\hat{\gamma}_i$

For $i = 1, \dots, m$:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{U}_i\gamma_{0i} + \boldsymbol{\varepsilon}_i$$

where

$$\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \text{ and } \gamma_{0i} \sim N(0, \tau_0^2)$$

and \mathbf{U}_i is a $n_i \times 1$ vector of ones. Further, the $n_i \times n_i$ marginal covariance matrix for \mathbf{Y}_i is

$$\mathbf{V}_i = \sigma^2\mathbf{I} + \tau_0^2\mathbf{1}\mathbf{1}^T \text{ with inverse } \mathbf{V}_i^{-1} = \frac{1}{\sigma^2} \left(\mathbf{I} - \frac{\tau_0^2}{\sigma^2 + n_i\tau_0^2} \mathbf{1}\mathbf{1}^T \right)$$

which means that the elements on the main diagonal for \mathbf{V}^{-1} are

$$\frac{1}{\sigma^2(\sigma^2 + n_i\tau_0^2)}$$

and the off-diagonal entries are $-\tau_0^2$.

The fixed effect estimate use this inverse matrix as the weighting matrix \mathbf{V}_i^{-1} in

$$\hat{\beta} = \left(\sum_{i=1}^m \mathbf{x}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{x}_i \right)^{-1} \sum_{i=1}^m \mathbf{x}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{Y}_i$$

The predicted random intercepts are

$$\hat{\gamma}_{0i} = \hat{\mathbf{Q}} \mathbf{U}_i^T \hat{\mathbf{V}}_i^{-1} (\mathbf{Y}_i - \mathbf{x}_i \hat{\beta}) = \dots = \frac{n_i \hat{\tau}_0^2}{\hat{\sigma}^2 + n_i \hat{\tau}_0^2} e_i$$

where e_i is the average (raw, level 0 - see below) residual

$$e_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij} - \mathbf{x}_{ij}^T \hat{\beta})$$

Interpretation

$$\hat{\gamma}_{0i} = \frac{n_i \hat{\tau}_0^2}{\hat{\sigma}^2 + n_i \hat{\tau}_0^2} e_i$$

Remember

$$E(\boldsymbol{\gamma} | \mathbf{Y}) = \mathbf{0} + \mathbf{GU}^T \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

The formula for $\hat{\gamma}_{i0}$ can be seen as a weighted sum between the conditional expectation 0 and the average residual e_i , with weighting factor $\frac{n_i \hat{\tau}_0^2}{\hat{\sigma}^2 + n_i \hat{\tau}_0^2}$ for the average residual (and 1-this for 0). The larger the n_i the closer the weight is to 1 and the smaller the shrinkage. Shrinkage is also high if the error variance σ^2 is large compared to the random effect variance τ_0^2 . The latter gives a very small ICC, so then it makes sense to have random effects close to 0.

Plotting predictions of random effects

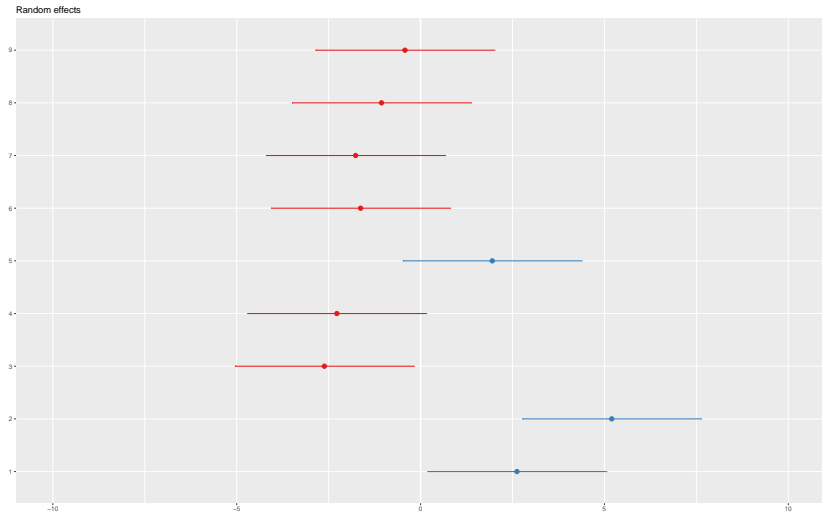
We can also use the R package `sjPlot` to produce plots and outputs from fitting linear mixed effects models with function `lmer` in the R package `lme4`. This plotting package can also be used to produce nice plots for `lm` and `glm`. The package uses `ggplot2` and other `tidyverse` packages.

For details see

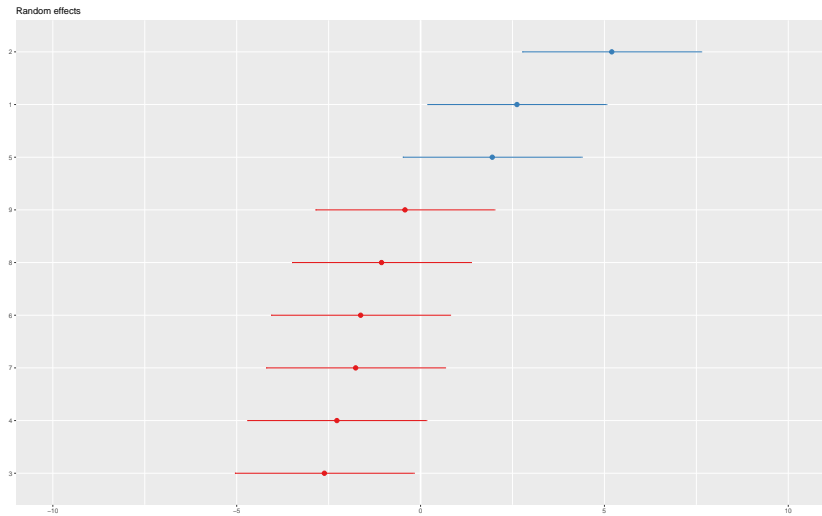
- ▶ `sjPlot` and more specifically
- ▶ `sjPlot` on CRAN - for vignettes and
- ▶ vignette on plotting random effects in LMM

For our beach-example: First the predicted values for the estimated random effect for each Beach - with confidence intervals. Unsorted and sorted version. (horizontal version of caterpillar plot). Then QQ-plots for the estimated random effects.

```
plot_model(fit, type = "re", y.offset = 0.4)
```



```
plot_model(fit, type = "re", sort.est = "(Intercept)", y.offset
```



Q: comment on what you see.

Predicted values for random errors ε_i (residuals)

Let μ_i denote $E(\mathbf{Y}_i)$. Fitted values for the LMM can be made on two levels:

$$\text{Level 0, marginal : } \hat{\mu}_i = \mathbf{X}_i \hat{\beta}$$

$$e_i = \mathbf{Y}_i - \mathbf{X}_i \hat{\beta}$$

$$\text{Level 1, conditional : } \tilde{\mu}_i = \mathbf{X}_i \hat{\beta} + \mathbf{U}_i \hat{\gamma}_i$$

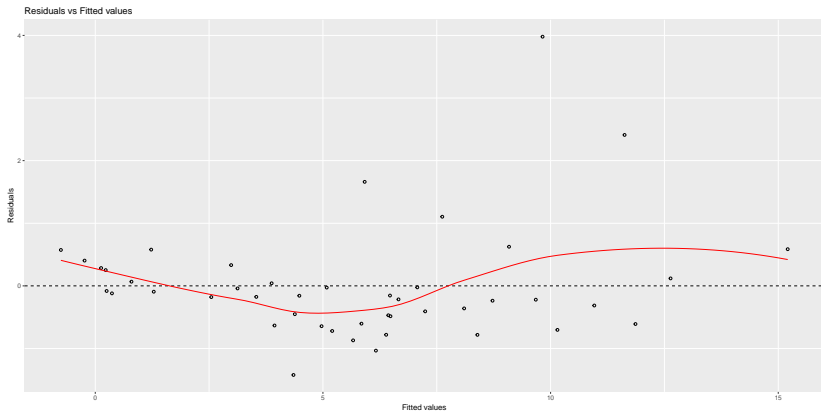
$$e_i = \mathbf{Y}_i - \mathbf{X}_i \hat{\beta} - \mathbf{U}_i \hat{\gamma}_i$$

For `lmer` the function `fitted` gives the level 1 fitted values (for our two-level models). This means that raw residuals can also be made on two levels, and the default is level 1 for `lmer`.

In addition to raw residuals, also Pearson residuals (standardized) are popular.

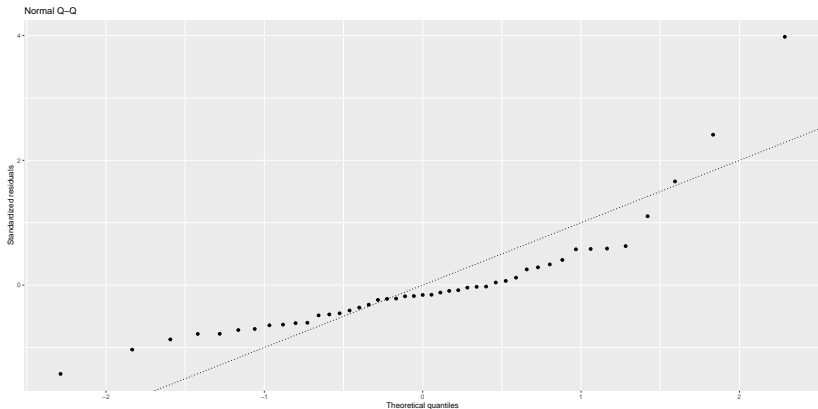
The residuals can be used in the same way as for the Multiple linear model (module 2).


```
fit = lmer(Richness ~ NAP + (1 | Beach), data = RIKZ)
df = data.frame(fitted = fitted(fit), resid = residuals(fit, scaled = T)
ggplot(df, aes(fitted, resid)) + geom_point(pch = 21) + geom_hline(yint
  linetype = "dashed") + geom_smooth(se = FALSE, col = "red", size =
  method = "loess") + labs(x = "Fitted values", y = "Residuals", titl
```



Q: any trend? homoscedastic?

```
ggplot(df, aes(sample = resid)) + stat_qq(pch = 19) + geom_abline(inter  
slope = 1, linetype = "dotted") + labs(x = "Theoretical quantiles",  
title = "Normal Q-Q")
```

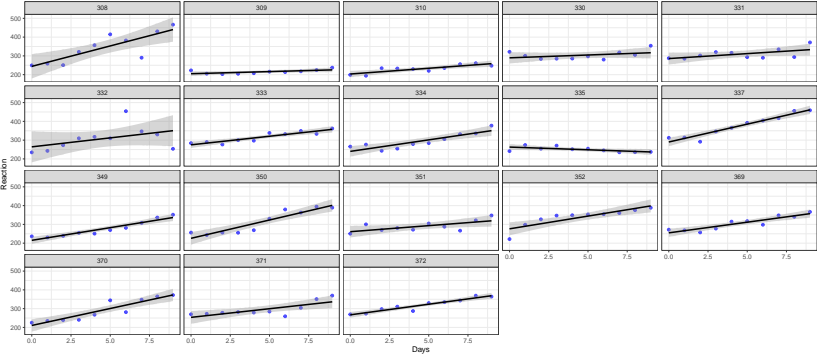


Q: normally distributed?

Random intercept and slope model

Example: Sleep deprivation study

In a study on the effect of sleep deprivation the average reaction time per day were measured. On day 0 the subjects had their normal amount of sleep. Starting that night they were restricted to 3 hours of sleep per night. The observations represent the average reaction time on a series of tests given each day to each subject. This was measured for 18 subjects for 10 days (days 0-9).



We observe that each subject's reaction time increases approximately linearly with the number of sleepdeprived days. But, it appears that subjects have different slopes and intercepts.

As a first model we may assume that there is a common intercept and slope for the population - called fixed effects, but allow for random deviations for the intercept and slope for each individual. This is called a *random intercept and slope model*.

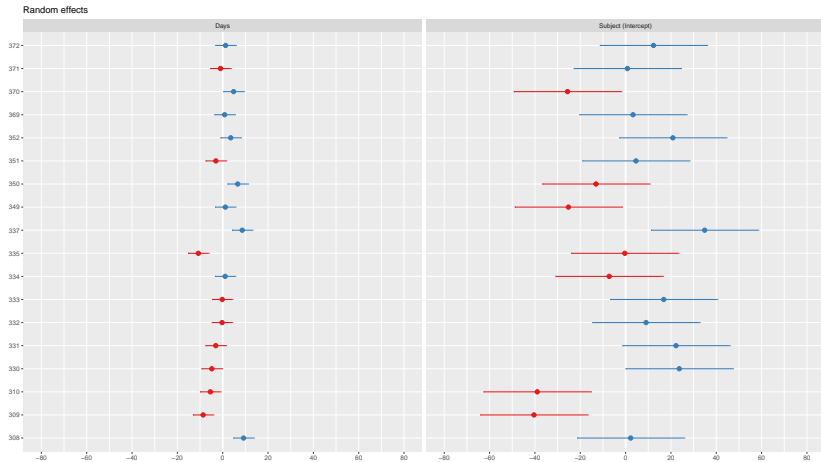
```
fm1 <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy)
summary(fm1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
## Data: sleepstudy
##
## REML criterion at convergence: 1743.6
##
## Scaled residuals:
##   Min       1Q   Median       3Q      Max
## -3.9536 -0.4634  0.0231  0.4634  5.1793
##
## Random effects:
##   Groups   Name      Variance Std.Dev. Corr
##   Subject (Intercept) 612.09   24.740
##   Subject   Days       35.07    5.922  0.07
## Residual              654.94   25.592
## Number of obs: 180, groups: Subject, 18
##
## Fixed effects:
##           Estimate Std. Error t value
## (Intercept) 251.405      6.825  36.838
## Days         10.467      1.546   6.771
##
## Correlation of Fixed Effects:
##   (Intr)
## Days -0.138
```

Q: What are our parameter estimates and their interpretation?

A: Here the population fixed effects estimates are an intercept of 251.4 ms and a slope of 10.47 ms/day. The random effects for the intercept and the slope have estimated standard deviations 24.74 ms and 5.92 ms/day.

```
library(sjPlot)
plot_model(fm1, type = "re")
```



Measurement model

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_{0i} + \gamma_{1i} x_{ij} + \varepsilon_{ij}$$

- ▶ β_0 : population intercept (fixed)
- ▶ γ_{0i} : deviation (for members of cluster i) from the population intercept β_0 - not a parameter but a random variable!
- ▶ $\beta_0 + \gamma_{0i}$: random intercept for cluster i
- ▶ β_1 : population slope (fixed), common to all clusters
- ▶ γ_{1i} : deviation (for members of cluster i) from the population slope β_1 - not a parameter but a random variable!
- ▶ $\beta_1 + \gamma_{1i}$: random slope for cluster i .

Distributional assumptions

$$\varepsilon_i \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$
$$\gamma_i = \begin{pmatrix} \gamma_{0i} \\ \gamma_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{Q} = \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{pmatrix}\right)$$

The parameter τ_{01} gives the covariance between the random intercept and random slope.

Marginal covariances for Y_i

$$\text{Cov}(Y_{ij}, Y_{kl}) = E[(Y_{ij} - \mu_{ij})(Y_{kl} - \mu_{kl})]$$

$$\text{Cov}(Y_{ij}, Y_{kl}) = \begin{cases} \tau_0^2 + 2\tau_{01}x_{ij} + \tau_1^2 x_{ij}^2 + \sigma^2 = \text{Var}(Y_{ij}) & \text{for } i = k, j = l \\ \tau_0^2 \tau_{01} x_{ij} + \tau_{01} x_{il} + \tau_i^2 x_{ij} x_{il} & \text{for } i = k, j \neq l \\ 0 & \text{for } i \neq k, j \neq l \end{cases}$$

The correlation between Y_{ij} and Y_{il} (two observations in the same cluster, that is, same beach) depends in a complicated way on the observed values for x and is rather difficult to interpret.

$$\text{Corr}(Y_{ij}, Y_{il}) = \frac{\text{Cov}(Y_{ij}, Y_{il})}{\sqrt{\text{Var}(Y_{ij})\text{Var}(Y_{il})}}$$

Hypothesis testing

Testing fixed effects

$$\hat{\beta} = \left(\sum_{i=1}^m \mathbf{x}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{x}_i \right)^{-1} \sum_{i=1}^m \mathbf{x}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{Y}_i \approx N(\beta, \left(\sum_{i=1}^m \mathbf{x}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{x}_i \right)^{-1})$$

Approximate Wald tests for fixed effects

$$\mathbf{C}\beta = \mathbf{d} \text{ vs. } \mathbf{C}\beta \neq \mathbf{d}$$

where \mathbf{C} is a $r \times p$ constant matrix and \mathbf{d} a $r \times 1$ constant vector.
Then:

$$(\hat{\beta} - \beta)^T \mathbf{C}^T \left[\mathbf{C} \left(\sum_{i=1}^m \mathbf{x}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{x}_i \right)^{-1} \mathbf{C}^T \right]^{-1} \mathbf{C} (\hat{\beta} - \beta)$$

asymptotically follows a χ^2 -distribution with r -degrees of freedom.

Beach-example: Hypothesis testing with normal approximation

```
fit = lmer(Richness ~ NAP + (1 | Beach), data = RIKZ, REML = FALSE)
summary(fit)
z = (summary(fit)$coefficients[2, 3])
1 - pchisq(z^2, 1)
```

```
## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: Richness ~ NAP + (1 | Beach)
## Data: RIKZ
##
##      AIC      BIC   logLik deviance df.resid
## 249.8    257.1  -120.9   241.8     41
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.4258 -0.5010 -0.1791  0.2452  4.0452
##
## Random effects:
## Groups Name Variance Std.Dev.
## Beach (Intercept) 7.507  2.740
## Residual          9.111  3.018
## Number of obs: 45, groups: Beach, 9
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  6.5844    1.0321   6.380
## NAP          -2.5757    0.4873  -5.285
##
## Correlation of Fixed Effects:
##      (Intr)
## NAP -0.164
## [1] 1.254807e-07
```

Likelihood ratio tests for fixed effects

Notation:

- ▶ A: the larger model and
- ▶ B: the smaller model (under H_0), and the smaller model is nested within the larger model (that is, B is a submodel of A).

The random effects parts of these models are assumed to be the same, while the changes are only to the fixed effects part.

The likelihood ratio statistic is defined as

$$-2 \ln \lambda = -2(\ln L(\hat{\beta}_B) - \ln L(\hat{\beta}_A))$$

which under the null is asymptotically χ^2 -distributed with degrees of freedom equal the difference in the number of parameters in the large and the small model. Again, p -values are calculated in the upper tail of the χ^2 -distribution.

Remark: this is the log-likelihood, not the REML version.

Remark: this result is **not valid** if the the models are fitted using REML instead of ML. The reason for this is that the mean structure of the model fitted under the null hypothesis is not the same mean structure under the alternative hypothesis, which leads to that different matrices **A** must be used for the REML method. Therefore these REML log-likelihoods are based on different observations and are therefore not comparable.

Beach-example: Hypothesis testing with likelihood ratio test

```
fit = lmer(Richness ~ NAP + (1 | Beach), data = RIKZ, REML = FALSE)
fit0 = lmer(Richness ~ 1 + (1 | Beach), data = RIKZ, REML = FALSE)
anova(fit0, fit)
```

```
## Data: RIKZ
## Models:
## fit0: Richness ~ 1 + (1 | Beach)
## fit: Richness ~ NAP + (1 | Beach)
##      Df    AIC    BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## fit0  3 269.30 274.72 -131.65   263.30
## fit   4 249.83 257.06 -120.92   241.83 21.474      1 3.586e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q: Which is model A (large) and model B (small)? What do we conclude? Compare to the Wald test result (above).

Testing parameters for random effects

In most situations the fixed effects model is of prime interest, however, a good choice of covariance structure is useful for interpreting the data and essential to be able to perform valid inference for the fixed effects.

- ▶ Overparameterization: gives inefficient estimation.
- ▶ Too restrictive specification: invalid inference about the fixed effects.

Wald test can also be used for random effects parameters ϑ in \mathbf{Q} and σ^2 , and

- ▶ asymptotically also $\hat{\vartheta}$ follows a multivariate normal distribution (under regularity conditions) with mean ϑ and covariance matrix given by the inverse of the Fisher information matrix.
- ▶ We may use the negative of the second order partial derivatives (Hessian) of the log-likelihood (ML or REML) wrt. ϑ .

But there is a problem: the performance of the normal approximation depends strongly on the true value of ϑ and large samples are needed for values of ϑ that are close to the boundary of the parameter space (for the hypothesis tested), and when on the boundary the normal approximation fails.

We will not dive deep into this matter in this course, but report that the solution to this - both for the Wald and the likelihood ratio test (preferably using the REML log-likelihood) is to use a mixture of χ^2 distributions in these cases.

Likelihood ratio test for random effects

- ▶ A: the larger model and
- ▶ B: the smaller model (under H_0), and the smaller model is nested within the larger model (that is, B is a submodel of A).

The fixed effects parts of these models are assumed to be the same, while the changes are only to the random effects part - and the changes gives nested models.

The likelihood ratio statistic is defined as

$$-2 \ln \lambda = -2(\ln L(\hat{\vartheta}_B) - \ln L(\hat{\vartheta}_A))$$

and the REML-likelihood is preferred.

For testing a random intercept model vs. no random intercept (need for this random effect) then

$$H_0 : \mathbf{Q} = 0 \text{ vs. } H_1 : \mathbf{Q} = \tau_0^2$$

asymptotically $-2 \ln \lambda$ is a mixture of χ_1^2 and χ_0^2 with equal weights. Here χ_0^2 is the distribution that gives probability mass 1 to the value 0.

If we instead had used the classical null distribution (χ_1^2) then the p -values would be too large and the null hypotheses kept to often.

For testing a random intercept versus a random slope (with \mathbf{Q} having three parameters) when p -values is found from 0.5% times a χ_1^2 and a χ_2^2 distribution.

Similar strategies for other situations - see Verbeke and Molenberghs (2000) pages 69-72.

Model selection methods

There are two main strategies:

- ▶ Hypothesis testing
 - ▶ asymptotic Wald tests for fixed effects
 - ▶ likelihood ratio test for fixed effects and parameters for random effects
- ▶ Information criteria: AIC and BIC

AIC and BIC for Maximum likelihood estimation (ML)

$$\text{AIC} = -2 \cdot l(\hat{\beta}, \hat{\vartheta}) + 2 \cdot r$$

$$\text{BIC} = -2 \cdot l(\hat{\beta}, \hat{\vartheta}) + \ln(N) \cdot r$$

- ▶ r : number of parameters in the model, both the β s and the parameters in the variance of the random effects, i.e. the σ^2 from our error and then all variances and covariances for the random effects in \mathbf{Q} .
- ▶ $N = \sum_{i=1}^m n_i$
- ▶ $l(\hat{\beta}, \hat{\vartheta})$ is the maximum log-likelihood inserted the parameter estimates

This can be used directly in the ML estimation, and as before BIC will give a smaller model than AIC.

AIC and BIC for Restricted Maximum likelihood estimation (REML)

$$\text{AIC} = -2 \cdot l(\hat{\beta}, \hat{\vartheta}) + 2 \cdot r$$

$$\text{BIC} = -2 \cdot l(\hat{\beta}, \hat{\vartheta}) + \ln(N - p) \cdot r$$

- ▶ r : number of parameters in the model, both the β s and the parameters in the variance of the random effects, i.e. the σ^2 from our error and then all variances and covariances for the random effects in \mathbf{Q} .
- ▶ $l(\hat{\beta}, \hat{\vartheta})$ is now the restricted maximum log-likelihood inserted the parameter estimates

Remember:

$$l_{\text{REML}}(\vartheta) = l_P(\vartheta) - \frac{1}{2} \ln \left| \sum_{i=1}^m \mathbf{x}_i^T \mathbf{V}(\vartheta)_i^{-1} \mathbf{x}_i \right|$$

Sleep study - comparing random effects models

```
fm1 = lmer(Reaction ~ Days + (Days | Subject), data = sleepstudy)
# random slope and intercept, correlated
fm2 = lmer(Reaction ~ Days + ((1 | Subject) + (0 + Days | Subject)), data = sleepstudy)
# random slope and intercept, uncorrelated
fm3 = lmer(Reaction ~ Days + (1 | Subject), data = sleepstudy)
# random intercept
AIC(fm1, fm2, fm3) #with REML
extractAIC(fm1)
extractAIC(fm2)
extractAIC(fm3) #refit and give ML
```

```
##      df      AIC
## fm1  6 1755.628
## fm2  5 1753.669
## fm3  4 1794.465
## [1]    6.000 1763.939
## [1]    5.000 1762.003
## [1]    4.000 1802.079
```

Q: Which model to choose?

A: The fm2 model gives the lowest AIC.

Top-down strategy for model selection

1. Start with model with all explanatory variables and possible interactions for the fixed effects - called a *beyond optimal* model. (Nothing is really done here, just decide on the largest possible fixed part).
2. With this beyond optimal fixed effects model we now focus on the random effects. The idea is that since we have many explanatory variables in the fixed effects the random component should not contain information that we would prefer to have in the fixed effect. To do this we may either use testing or AIC or BIC. Testing is problematic due to that the null hypotheses tested is on the boundary of the parameter values tested ($\tau^2 = 0$). REML must be used (to get as unbiased estimates as possible).

3. Now we have the optimal random effect, so we focus on the optimal fixed effect model. ML must be used because different fixed effects will give incomparable REML-log-likelihoods. Testing or AIC or BIC can be used.
4. The final model is then presented with REML estimates.

Top-down strategy for the beach-data

This example is taken from Zuur et al. (2009), pages 127-128.

1. We decide on fixed model with intercept, main effect of NAP and Exposure and the interaction thereof. (fExp is just a slight adjustment of Exposure by letting level 8 be 10 - just replicate what Zuur did, so fExp only has values 10 and 11.)
2. Fit the fixed model from 1 to “no random effect”, “random intercept” and “random intercept and slope for NAP”. Use REML.

```
B1 = gls(Richness ~ 1 + NAP * fExp, data = RIKZ, method = "REML")
B2 = lmer(Richness ~ 1 + NAP * fExp + (1 | Beach), data = RIKZ)
B3 = lmer(Richness ~ 1 + NAP * fExp + (1 + NAP | Beach), data = RIKZ)
AIC(B1, B2, B3)
```

```
##      df      AIC
## B1  5 238.5329
## B2  6 236.4925
## B3  8 237.1331
```

3. With the random intercept, now compare the different fixed effects models. Use ML.

```
F2 = lmer(Richness ~ 1 + NAP * fExp + (1 | Beach), data = RIKZ, REML = FALSE, confint(F2))
F2a = lmer(Richness ~ 1 + NAP + fExp + (1 | Beach), data = RIKZ, REML = FALSE, confint(F2a))
```

```
##              2.5 %    97.5 %
## .sig01        0.000000  3.145294
## .sigma        2.3114668  3.681773
## (Intercept)   6.9045813 10.804288
## NAP          -4.7299177 -2.275315
## fExp11       -8.1969707 -2.303772
## NAP:fExp11    0.1919491  3.877650
##              2.5 %    97.5 %
## .sig01        0.000000  3.297744
## .sigma        2.435935  3.879425
## (Intercept)   6.562780 10.630777
## NAP          -3.578864 -1.644351
## fExp11       -7.563508 -1.505626
```

```
F2b = lmer(Richness ~ 1 + NAP + (1 | Beach), data = RIKZ, REML = FALSE)
confint(F2b)
F2c = lmer(Richness ~ 1 + fExp + (1 | Beach), data = RIKZ, REML = FALSE)
confint(F2c)
AIC(F2, F2a, F2b, F2c)
```

```
##              2.5 %    97.5 %
## .sig01         1.484204  5.100814
## .sigma         2.435429  3.877572
## (Intercept)   4.324560  8.824909
## NAP          -3.566901 -1.599779
##              2.5 %    97.5 %
## .sig01         0.000000  3.95707
## .sigma         3.178126  5.05812
## (Intercept)   5.385254 10.29475
## fExp1         -8.522119 -1.15788
##      df      AIC
## F2     6 242.1135
## F2a    5 244.7589
## F2b    4 249.8291
## F2c    4 265.4332
```

Conclusion: keep the full model. No confidence intervals cover 0, and the AIC supports the full model.

Remark: In Zuur et al (2009), page 128, the conclusion was to use the additive model (model F2a above), based on asymptotic p -values (but this was smaller than 0.05) using the `nlme` package. We have instead used AIC for this selection.

```
summary(B2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Richness ~ 1 + NAP * fExp + (1 | Beach)
## Data: RIKZ
##
## REML criterion at convergence: 224.5
##
## Scaled residuals:
##   Min       1Q   Median       3Q      Max
## -1.4849 -0.4161 -0.0770  0.1521  3.7313
##
## Random effects:
## Groups Name Variance Std.Dev.
## Beach (Intercept) 3.307 1.819
## Residual 8.660 2.943
## Number of obs: 45, groups: Beach, 9
##
## Fixed effects:
## Estimate Std. Error t value
## (Intercept) 8.8611 1.0208 8.680
## NAP -3.4637 0.6279 -5.517
## fExp11 -5.2556 1.5452 -3.401
## NAP:fExp11 2.0005 0.9461 2.114
##
## Correlation of Fixed Effects:
## (Intr) NAP fExp11
## NAP -0.181
## fExp11 -0.661 0.120
## NAP:fExp11 0.120 -0.664 -0.221
```

Fitting LMM with function `lmer` in package `lme4`

This is based on the article *Fitting Linear Mixed-Effects Models Using lme4* by Bates, Bolker, Mächler and Walker (2015) in *Journal of Statistical Software*, and in particular pages 30 and onwards.

We use a data set called *the ergonometics experiment data set* `ergoStool` for illustration.

- ▶ `effort`: the effort required (socalled Borg scale) to arise from a stool (`krakk`) - this is our response
- ▶ `Type`: the type of stool - types T1, T2, T3 and T4 studied.
- ▶ `Subject`: each of nine different subjects tested the four different stools (in random order?). Subjects

Was there any clear winner among the stools, when the goal was to minimize effort?

The ergoStool data set

is found in the MEMSS package.

```
library(MEMSS)
summary(ergoStool)
table(ergoStool$Subject)
contrasts(ergoStool$Type) #default contrast used
```

```
##      effort      Type      Subject
##  Min.   : 7.00    T1:9    A       : 4
##  1st Qu.: 8.00    T2:9    B       : 4
##  Median :10.00    T3:9    C       : 4
##  Mean   :10.25    T4:9    D       : 4
##  3rd Qu.:12.00                    E       : 4
##  Max.   :15.00                    F       : 4
##                                     (Other):12
##
```

```
##  A B C D E F G H I
##  4 4 4 4 4 4 4 4 4
##      T2 T3 T4
## T1  0  0  0
## T2  1  0  0
```

p-values in lme4

Excerpt from *Fitting Linear Mixed-Effects Models Using lme4* by Bates, Bolker, Mächler and Walker (2015) in *Journal of Statistical Software* page 35:

Computing *p* values *One of the more controversial design decisions of lme4 has been to omit the output of *p* values associated with sequential ANOVA decompositions of fixed effects. The absence of analytical results for null distributions of parameter estimates in complex situations (e.g., unbalanced or partially crossed designs) is a long-standing problem in mixed-model inference. While the null distributions (and the sampling distributions of non-null estimates) are asymptotically normal, these distributions are not *t* distributed for finite size samples – nor are the corresponding null distributions of differences in scaled deviances *F* distributed. Thus approximate methods for computing the approximate degrees of freedom for *t* distributions, or the denominator degrees of freedom for *F* statistics (Satterthwaite 1946; Kenward and Roger 1997), are at best ad hoc solutions.*

However, computing finite-size-corrected p values is sometimes necessary. Therefore, although the package does not provide them (except via parametric bootstrapping, Section 5.1), we have provided a help page to guide users in finding appropriate methods:

```
R> help("pvalues")
```

pvalues {lme4} R Documentation Getting p-values for fitted models

Description

One of the most frequently asked questions about lme4 is “how do I calculate p-values for estimated parameters?” Previous versions of lme4 provided the `mcmc` function, which efficiently generated a Markov chain Monte Carlo sample from the posterior distribution of the parameters, assuming flat (scaled likelihood) priors. Due to difficulty in constructing a version of `mcmc` that was reliable even in cases where the estimated random effect variances were near zero (e.g. <https://stat.ethz.ch/pipermail/r-sig-mixed-models/2009q4/003115.html>), `mcmc` has been withdrawn (or more precisely, not updated to work with lme4 versions $\geq 1.0.0$).

Many users, including users of the `aov` function from the `languageR` package which relies on `mcmc`, will be deeply disappointed by this lacuna. Users who need p-values have a variety of options. In the list below, the methods marked MC provide explicit model comparisons; CI denotes confidence intervals; and P denotes parameter-level or sequential tests of all

What have we not covered?

- ▶ Multi-level models: we have only considered two levels
- ▶ Structures for the covariance matrix of ε_j : we have only considered $\sigma^2\mathbf{I}$.
- ▶ Effective sample size.
- ▶ Details about testing random effects with mixtures of χ^2 -distributions.
- ▶ External effects and the extended random intercept model to give different scales for the within and between cluster effects. See pages 353-354 in Fahrmeir et al (2013).
- ▶ Penalized least squares view.
- ▶ Bayesian view.

Interactive session week 2

Plan:

- ▶ First hour: Problem 1, go back to ILw1 to check what you skipped then.
- ▶ Second hour: Problem 2 and Team Kahoot!

Exercise 1: Taken from UiO, STK3100, 2016, problem 3

The data in this problem is based on four measurements of a particular bone for 20 boys. The measurements were taken at 8, 8.5, 9 and 9.5 years of age. The variables in the dataset are

- ▶ bone: length of bone in millimeters
- ▶ redage: centered age (i.e. age - 8.75)

Below is an excerpt from the output from fitting a linear mixed model (LMM) with the procedure `lmer` in R where the length of the bone, `bone`, is the response,

$$\text{bone}_{ij} = \beta_0 + \beta_1 \text{redage}_{ij} + \gamma_{0i} + \gamma_{1i} \text{redage}_{ij} + \varepsilon_{ij} \quad i = 1, \dots, 20, \quad j = 1, 2, 3, 4$$

where $\gamma_i = (\gamma_{0i}, \gamma_{1i})^T$, $i = 1, \dots, 20$ represent the random effects.

```
filepath <- "https://www.math.ntnu.no/emner/TMA4315/2018h/1  
bone <- read.table(filepath, header = TRUE)
```

Exercise 2: Taken from UiO, STK3100, 2015, problem 3

The data used in this problem concerns expenses in the the social security system Medicare in US. Average expenses per hospitalization, denoted as $ccpd$, were in six years recorded for 54 regions: the fifty US states, Puerto Rico, Virgin Islands, District of Columbia and an unspecified other. Thus there are $6 \times 54 = 324$ observations. The expenses are treated as response. The covariates are $j = \text{YEAR}$ which can take values $1, \dots, 6$ and a factor indicating the average length of stay at hospital, $AVETD$, in each region and year. This factor has tree levels, $1 = \text{six days or less}$, $2 = 7-9 \text{ days}$, $3 = 10 \text{ days or more}$. “Six days or less” is the reference level and the others are denoted as $AVETD2$ and $AVETD3$.

Below you find the output from fitting the linear mixed effects model

$$y_{ij} = \beta_0 + j \times \beta_1 + \beta_2 AVETD2_{ij} + \beta_3 AVETD3_{ij} + \gamma_{0i} + j \times \gamma_{1i} + \varepsilon_{ij}$$

Previous exams

TMA4315 December 2017, Problem 3: Random intercept linear mixed effects model

[20 points]

New on the reading list this year is how to handle correlated data in a regression setting, with the aid of the linear mixed effects model. The simplest version of such a model is the *random intercept model*.

Write a short introduction to the random intercept linear mixed effects model and its practical usage, for a student with a good background in multiple linear regression. The introduction should include an example and emphasis should be on:

- ▶ Model assumptions.
- ▶ The conditional and marginal model.
- ▶ What is the intra class correlation and how can it be calculated from a given model fit?
- ▶ How are the regression coefficients estimated?

R packages

```
# packages needed to run this module page  
install.packages("lme4")  
install.packages("devtools")  
library(devtools)  
install_github("romunov/AED")  
install.packages("lattice")  
install.packages("MEMSS")  
install.packages("sjPlot")  
install.packages("sjmisc")  
install.packages("ggpubr")  
install.packages("faraway")  
install.packages("kableExtra")  
install.packages("spcadjust")
```

Further reading

- ▶ Theoretical explanation for REML (beyond the scope of this course) by Inge Helland, UiO
- ▶ Verbeke and Molenberghs (2000), Section 5.3 (free ebook from Springer for NTNU students).
- ▶ *Fitting Linear Mixed-Effects Models Using lme4* by Bates, Bolker, Mächler and Walker (2015) in *Journal of Statistical Software*