7. Model building and model choice

Sonja Greven

Summer Term 2017

General recommendations

- As $\mathrm{E}(\boldsymbol{b}_i) = \mathbf{0}$, all covariates in $\boldsymbol{Z}_i$ should be linear transformations of covariates in $\boldsymbol{X}_i$.
- If $\boldsymbol{Z}_i$ contains x^p, it should also contain $x^0, x^1, \ldots, x^{(p-1)}$.
- The more complex the structure for the fixed and random effects is, the simpler the covariance structure in $\boldsymbol{\Sigma}_i$ should be.

Overview Chapter 7 - Model building and model choice

7.1 Model diagnostics

7.2 Model selection

7.3 Assumptions and confounding

Residual diagnostics 1

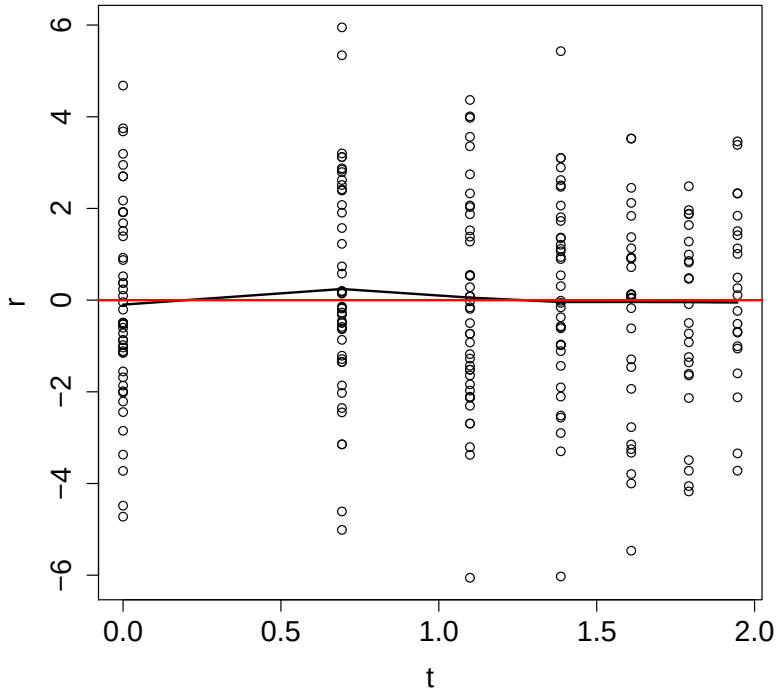
Plotting the residuals $r_{ij} = y_{ij} - \mathbf{x}_{ij}^T \widehat{\boldsymbol{\beta}}$ against covariates can help in diagnosing a misspecified mean structure, e.g. an omitted variable or a missing quadratic term. There should be no systematic trend!

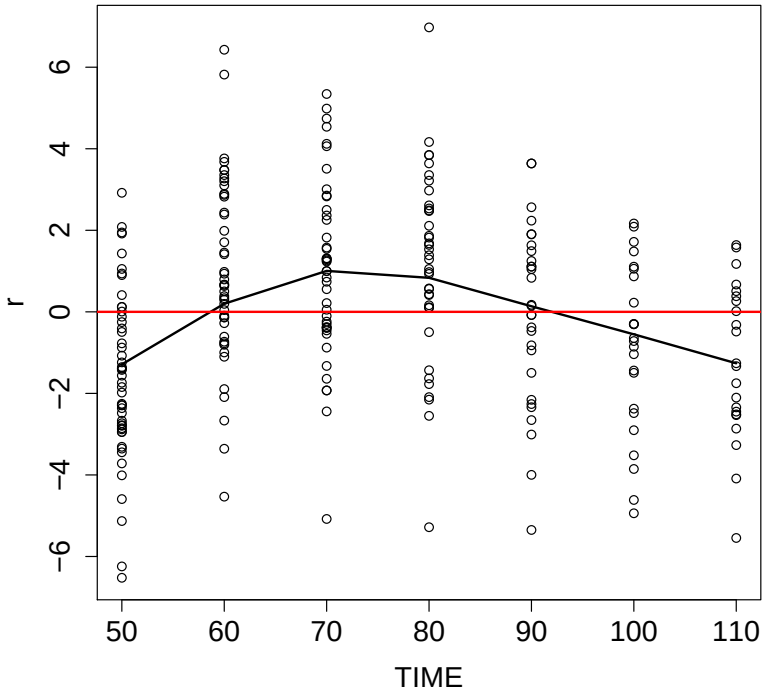
Example rat data, random intercept model with linear trend in transformed time variable $t = \log(1 + (TIME - 50)/10)$:

```
> lme1 <- lme(RESPONSE ~ group * t - group,
              random = ~ 1 | SUBJECT, data = rats)
> r <- resid(lme1, level = 0)  # 0 - without random effects
> plot(rats$t, r, xlab = "t")
> lines(lowess(rats$t, r))
```

Analogously for the original untransformed time variable $TIME$.

Residual diagnostics 1





Residual diagnostics 2

When plotting the residuals against the estimated mean, there should be no systematic trend.

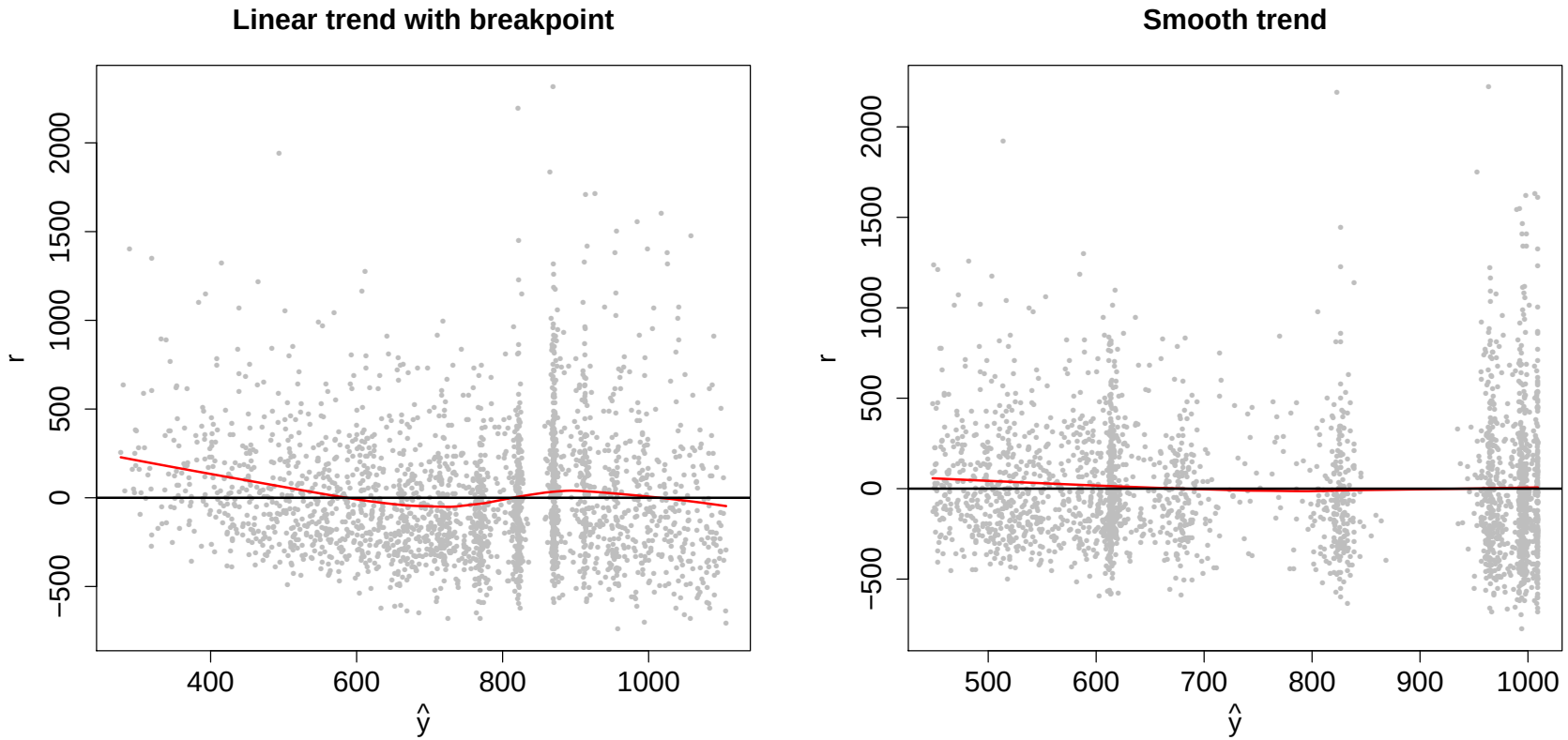
CD4 example, random intercept, linear time trend with breakpoint in 0:

```
> cd4$Timesc <- cd4$Time * (cd4$Time > 0) # for breakpoint
> lme1 <- lme(CD4 ~ Time + Timesc, data = cd4, random = ~ 1|ID)
> yhat <- predict(lme1, level = 0)        # 0 - predictions with-
> r    <-   resid(lme1, level = 0)        # out random effects
> plot(yhat, r)
> lines(lowess(yhat, r, iter = 0))
> abline(h = 0)
```

For comparison, random intercept model with smooth time trend:

```
> mygamm <- gamm(CD4 ~ s(Time), random = list(ID = ~ 1),
              data = cd4, method = "REML")
> r <- resid(mygamm$lme, level = 1) # 1 - include random
                    # effects for smooth, not for subjects
> yhat <- predict(mygamm$lme, level = 1)
> plot(yhat, r)
> lines(lowess(yhat, r, iter = 0))
> abline(h = 0)
```

Residual diagnostics 2



Residual diagnostics 3

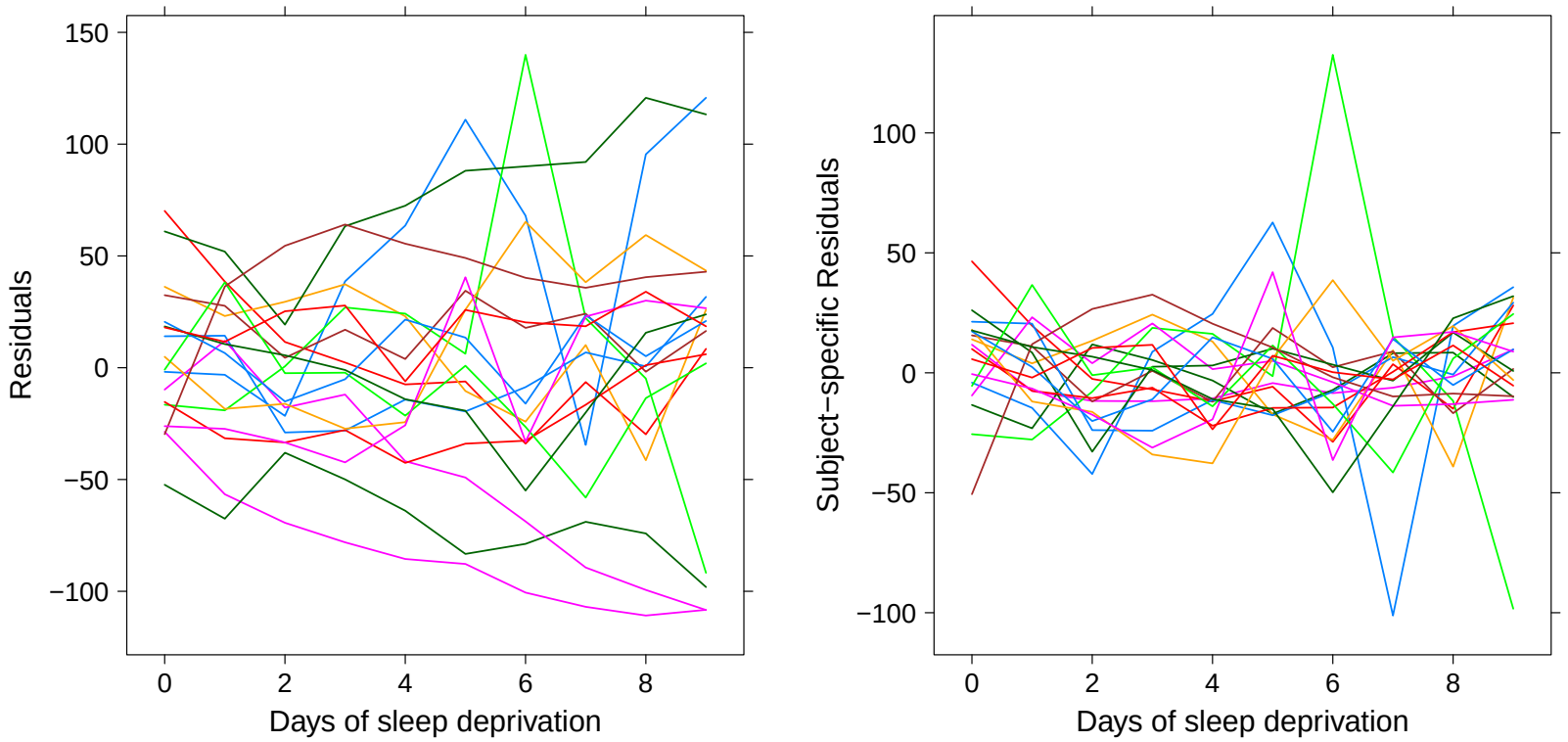
Plotting the residuals $r_{ij} = y_{ij} - \mathbf{x}_{ij}^T \widehat{\boldsymbol{\beta}}$ against covariates, e.g. time, can also indicate a missing random slope.

Example sleepstudy data, models without and with random slope:

```
> lme1 <- lme(Reaction ~ Days, random = ~    1 | Subject)
> r <- resid(lme1, level = 0) # 0: residuals w/o random effects
> xyplot(r ~ Days, groups = Subject, type = "l")

> lme2 <- lme(Reaction ~ Days, random = ~ Days | Subject)
> r <- resid(lme2, level = 1) # 1: residuals with random effects
                # to see difference when including random slope
> xyplot(r ~ Days, groups = Subject, type = "l")
```

Residual diagnostics 3



Transformed residuals

Remember that

$$\text{Cov}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) = \boldsymbol{V}_i.$$

Thus, the residual vector $\boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$ will have zero mean, but will be correlated and heteroscedastic. We need to keep this in mind for diagnostics.

One could consider the subject-specific residuals $\boldsymbol{y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}_i\widehat{\boldsymbol{b}}_i$. However, $\widehat{\boldsymbol{b}}_i$ very much depends on the normality assumption for $\boldsymbol{b}_i$, and is also influenced by the assumed structure for $\boldsymbol{V}_i$.

Diagnostics are thus often based on transformed residuals $\boldsymbol{r}_i^* = \boldsymbol{L}_i^{-1}\boldsymbol{r}_i$, where $\widehat{\boldsymbol{V}}_i = \boldsymbol{L}_i\boldsymbol{L}_i^T$ is the Cholesky decomposition with lower triangular matrix $\boldsymbol{L}_i$. $\boldsymbol{r}_i^*$ are approximately uncorrelated with unit variance.

Transformed residuals

The transformed residuals $\boldsymbol{r}_i^*$ have the following interpretation:

- The first element is the standardized residual for y_{i1}.
- The jth element is an estimate of

$$\frac{Y_{ij} - \mathsf{E}(Y_{ij}|Y_{i1}, \ldots, Y_{i(j-1)})}{\mathsf{Var}(Y_{ij}|Y_{i1}, \ldots, Y_{i(j-1)})},$$

i.e. the standardized deviation from the conditional mean given all previous observations.

Transformed residuals

After the transformation, the residuals can be used for the same kind of diagnostics as in the linear model, e.g.

- to identify **outlying observations**
- to identify skewness
- to plot the transformed residuals $\boldsymbol{r}_{ij}^*$ against the transformed predicted values $\widehat{\mu}_{ij}^*$ with

$$\widehat{\boldsymbol{\mu}}_i^* = \boldsymbol{L}_i^{-1}\widehat{\boldsymbol{\mu}}_i = \boldsymbol{L}_i^{-1}\boldsymbol{X}_i\widehat{\boldsymbol{\beta}},$$

  or against a selected transformed covariate (such as e.g. time).

Outlier diagnostics

Define the squared **Mahalanobis distance**

$$d_i = \mathbf{r}_i^{*T} \mathbf{r}_i^*.$$

as a summary measure of multivariate distance between observed and fitted values for individual i. If the model is correctly specified, we have the approximate distribution

$$d_i \sim \chi^2_{n_i}, \quad \text{for } i = 1, \ldots, N.$$

This can be used to identify **outlying individuals**: p-values can be computed for each subject and used to compare subjects, keeping in mind that p-values smaller α are expected to occur αN times.

Transformed residuals in R

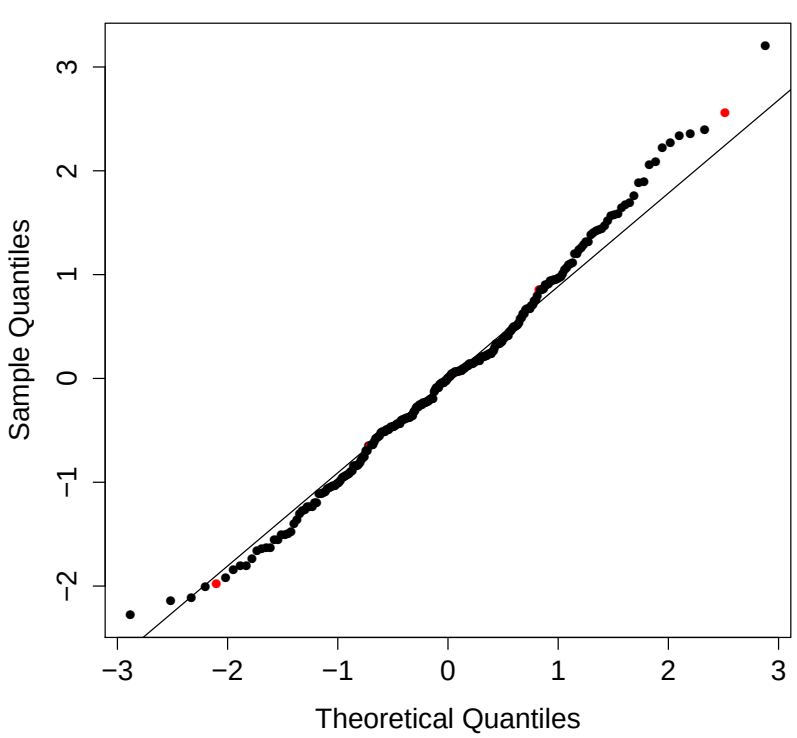
```
> library(RLRsim) # useful to extract lme model components
> r.star <- function(m){ # takes an lme object
+   design <- extract.lmeDesign(m)
+   Z <- design$Z
+   D <- design$Vr * design$sigmasq
+   R <- design$sigmasq * diag(nrow(Z))
+   V <- Z %*% D %*% t(Z) + R
+   L <- t(chol(V))
+   r.star <- solve(L, resid(m, level = 0))
+   return(r.star) # returns the transformed residuals
+ }
```

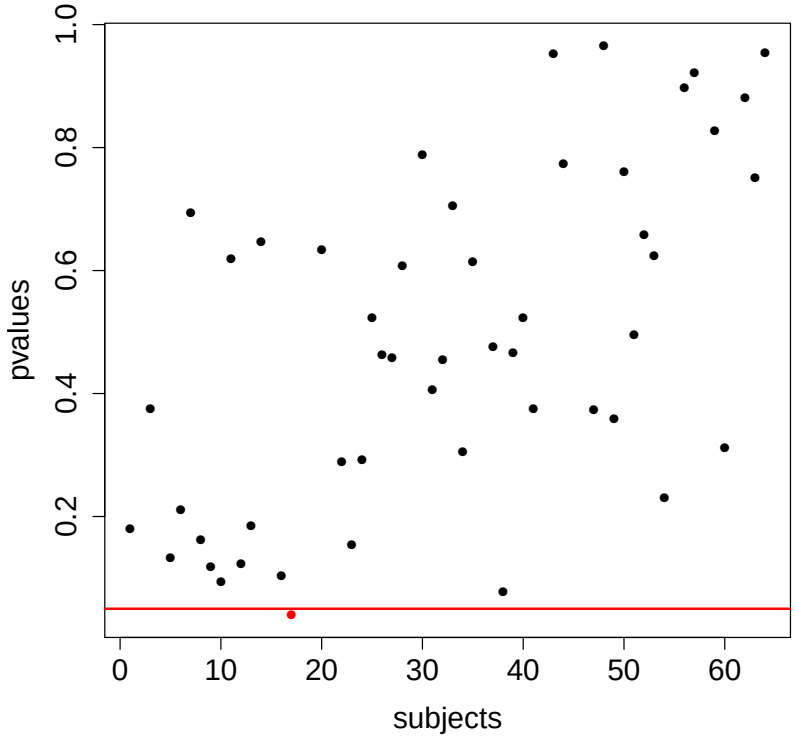
Example rat data

Random intercept model:

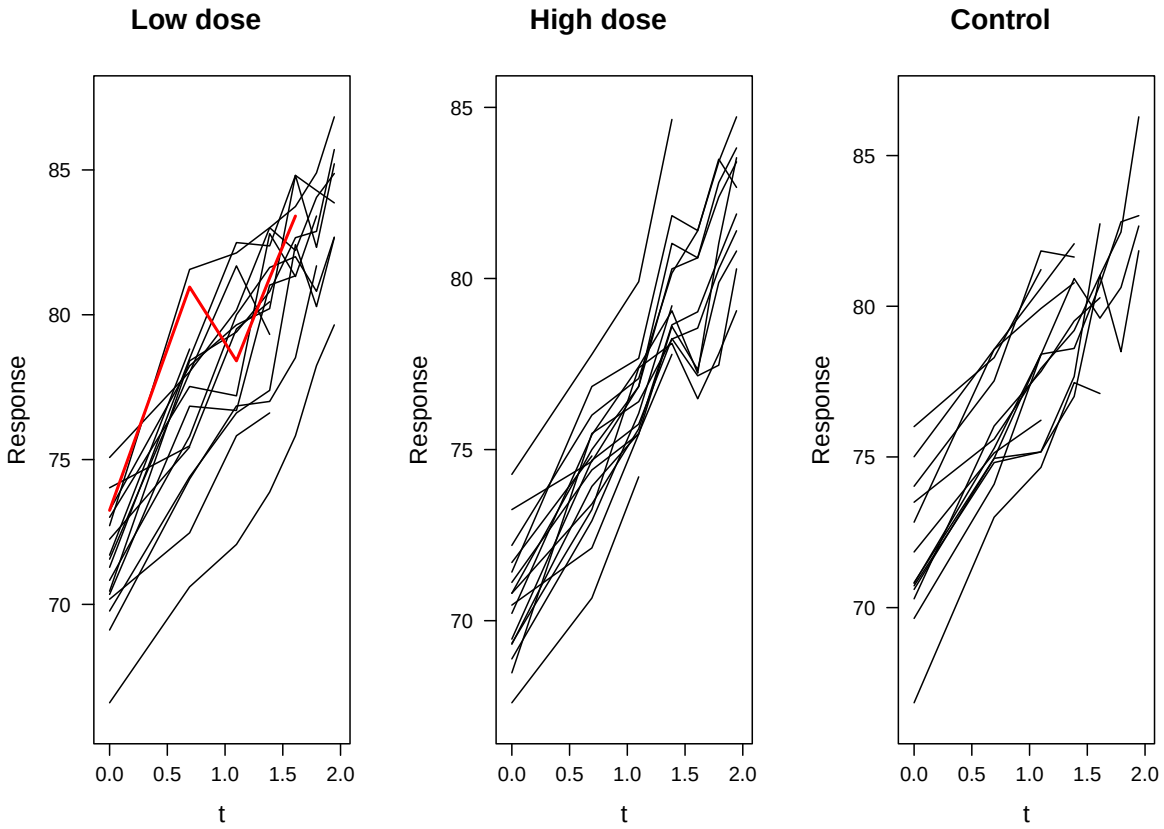
```
### Transformed residuals ###
> lme1 <- lme(RESPONSE ~ group * t - group,
              random = ~ 1 | SUBJECT, data = rats)
> r.star1 <- r.star(lme1) # transformed model residuals
### QQ-Plot ###
> qqnorm(r.star1)
> qqline(r.star1)
### Outlier Diagnostics ###
> subjects <- unique(sort(rats$SUBJECT)) # for each subject
> di <- sapply(subjects, FUN = function(subj)
       crossprod(r.star2[(rats$SUBJECT == subj)])) # compute d_i
> ni <- sapply(subjects, FUN = function(subj)
       sum(rats$SUBJECT == subj)) # and n_i
```

```
> pvalues <- pchisq(di, ni, lower = FALSE) # chi^2_{n_i} p-values
> plot(subjects, pvalues); abline(h = 0.05)
```

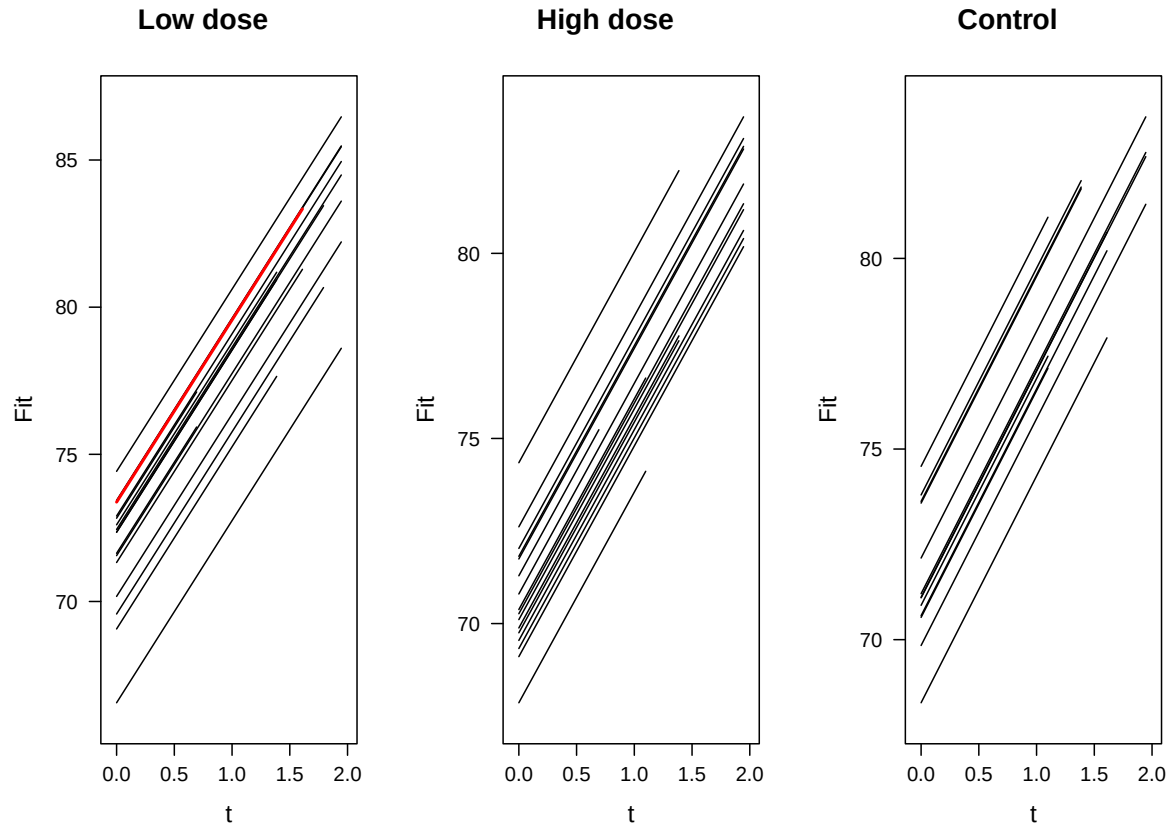




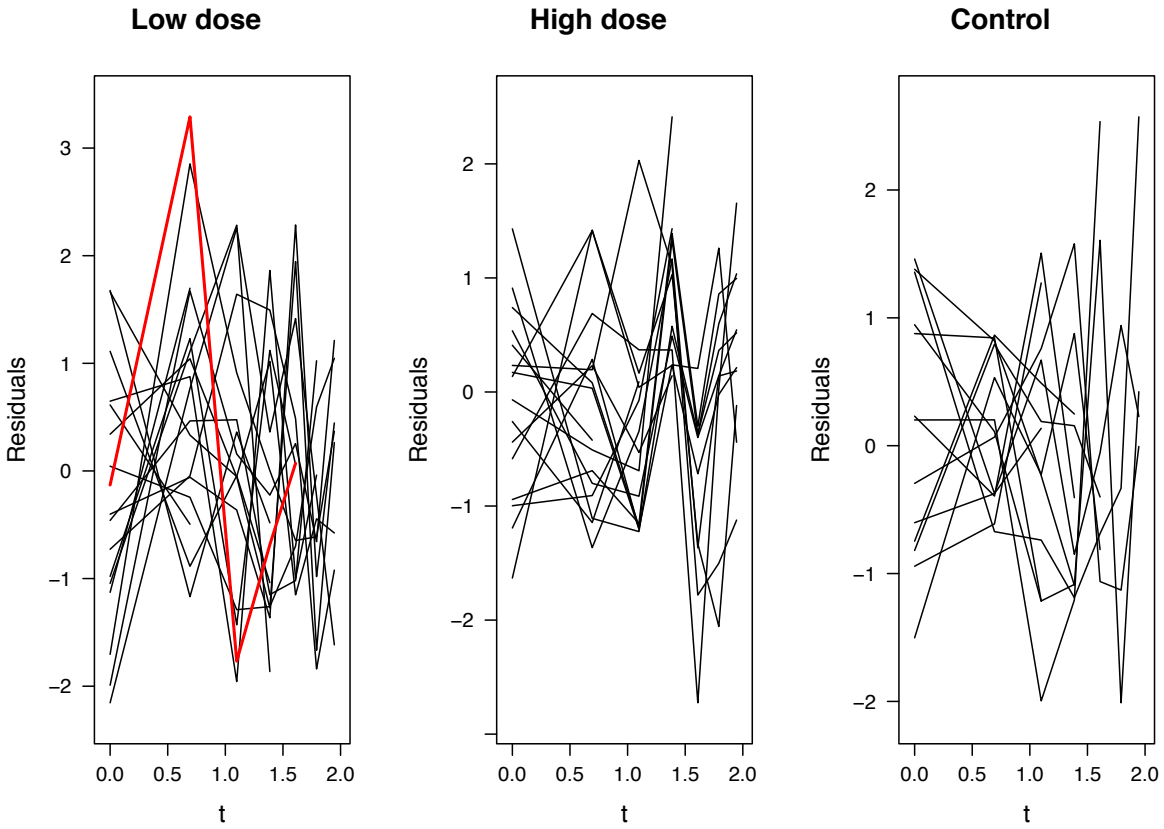
Example rat data - Data



Example rat data - Fit



Example rat data - Residuals



The choice of the covariance structure

A good model for the covariance structure is important for inference on the fixed effects, interpretation and prediction.

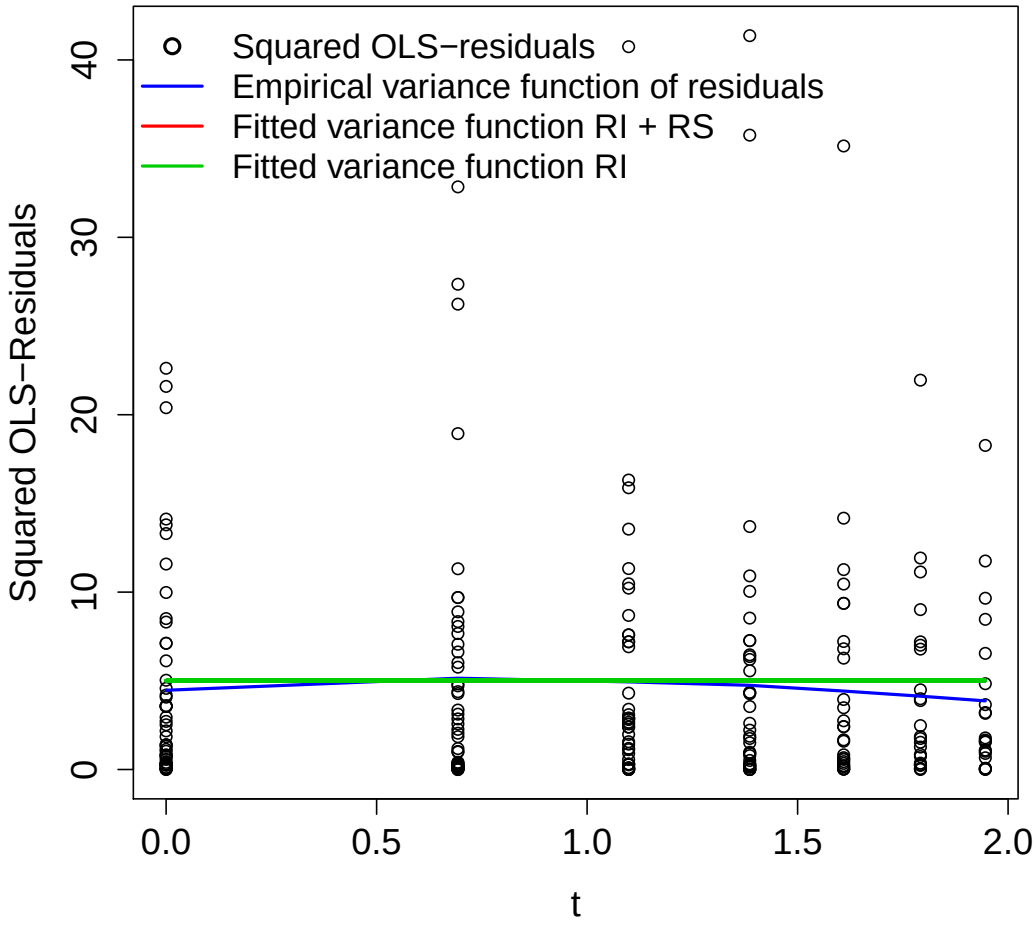
An informal check is to plot the squared OLS residuals

$$\mathbf{r}_{OLS,i} = \mathbf{y}_i - \mathbf{X}_i \widehat{\boldsymbol{\beta}}_{OLS}$$

and the fitted variance function against t. The fitted variance function corresponds to the diagonal entries of $\widehat{\boldsymbol{V}} = \boldsymbol{Z}\widehat{\boldsymbol{D}}\boldsymbol{Z}^T + \widehat{\boldsymbol{R}}$.

Example rat data with random intercept and slope: The fitted variance function is

$$(1 \quad t\,)\,\widehat{\mathbf{D}} \begin{pmatrix} 1 \\ t \end{pmatrix} + \widehat{\sigma}^2 = \widehat{d}_{11} + 2\widehat{d}_{12}t + \widehat{d}_{22}t^2 + \widehat{\sigma}^2.$$


Squared OLS−residuals
Empirical variance function of residuals
Fitted variance function RI + RS
Fitted variance function RI
Squared OLS−Residuals
0
10
20
30
40
0.0
0.5
1.0
1.5
2.0
t

The semi-variogram revisited

A more comprehensive check for the covariance structure is the following.

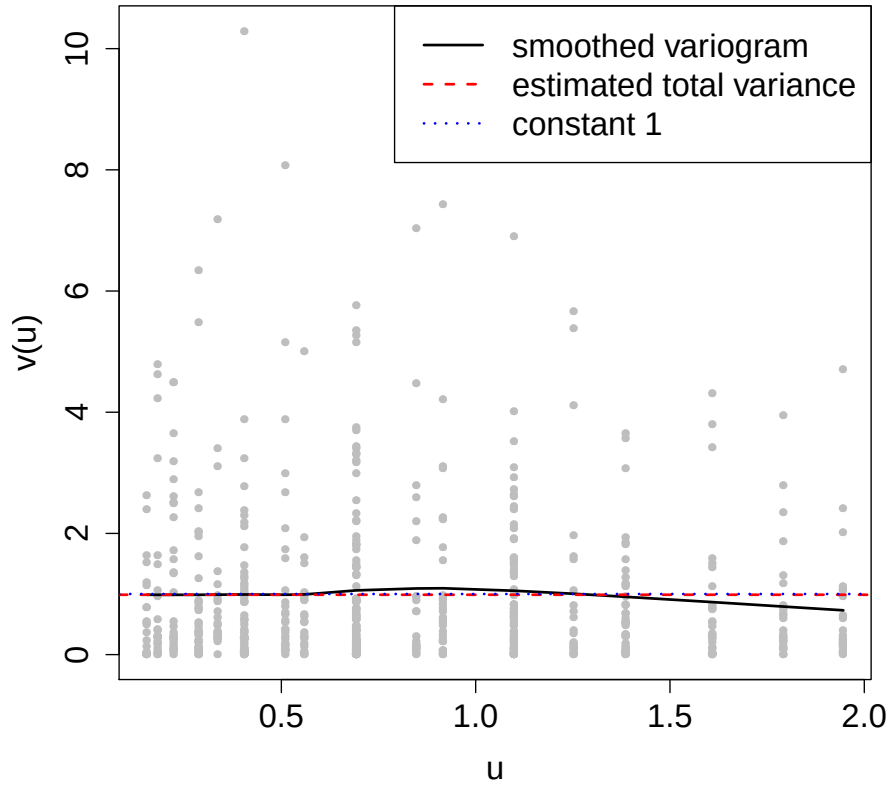
As the transformed residuals are approximately uncorrelated with mean zero and variance one, we have

$$\frac{1}{2}\mathsf{E}[(r^*_{ij} - r^*_{ik})^2] = \frac{1}{2}\left[\mathsf{Var}(r^*_{ij}) + \mathsf{Var}(r^*_{ik}) - 2\mathsf{Cov}(r^*_{ij}, r^*_{ik})\right]$$
$$= \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 1 - 0 = 1.$$

Thus, if the model for the covariance structure is correct, the empirical semi-variogram for the transformed residuals should randomly fluctuate around the constant 1.

Example rat data

Semi-variogram for the transformed residuals, random intercept model:



The normality assumption for the random effects

It would be of interest to look at the distribution of the $\boldsymbol{b}_i$ a) to check the normality assumption and b) to find outlying individuals. However, the $\widehat{\boldsymbol{b}}_i$

- all have different distributions unless all $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are equal.
- can look normal even if the true distribution of $\boldsymbol{b}_i$ is not normal (e.g. bi-modal). This is due to the shrinkage effect.

Fitting a model with a mixture distribution for the random effects (see Section 6.3) allows to check for normality of the random effects.

Overview Chapter 7 - Model building and model choice

7.1 Model diagnostics

7.2 Model choice

7.3 Assumptions and confounding

Model choice

Often, there are several possible model specifications. To compare two models M_1 and M_2, one can

- directly compare the likelihood if the numbers of parameters in M_1 and M_2 are the same.

  Examples:

  – Gaussian vs. exponential serial correlation
  – different transformations of a covariate in the fixed effects

- conduct a test if M_1 and M_2 are nested, see Chapter 5.

- use information criteria for model selection.

Information criteria

- **Goal:** Comparison of models M_1 and M_2 with potentially different numbers of parameters (potentially non-nested).
- Denote by l_1 and l_2 the maximized log-likelihood for models M_1 and M_2 and by df_1 and df_2 the number of parameters for models M_1 and M_2.
- Select model M_2 if for a function $\mathcal{F}$ specific to the information criterion

$$-2l_1 + \mathcal{F}(df_1) > -2l_2 + \mathcal{F}(df_2).$$

- If M_1 is nested in M_2, a likelihood ratio test corresponds to

$$\mathcal{F}(df_2) - \mathcal{F}(df_1) = \chi^2_{df_2 - df_1; 1-\alpha},$$

where $\chi^2_{d;1-\alpha}$ is the $(1-\alpha)$-Quantile of the χ^2_d distribution.

The Akaike information criterion (AIC) - Background

- The AIC uses $\mathcal{F}(df) = 2df$, with $df = \dim(\boldsymbol{\Theta})$ the number of parameters.
- Suppose data $\boldsymbol{y}$ is generated from a true underlying model with density $g(\cdot)$. We approximate $g(\cdot)$ by a parametric class of models $f_{\boldsymbol{\theta}}(\cdot) = f(\cdot|\boldsymbol{\theta})$.
- Under regularity conditions, minimizing the AIC over a set of models minimizes (an unbiased estimator of) the expected Kullback-Leibler distance between an approximating model $f_{\hat{\boldsymbol{\theta}}}$ and the underlying truth g.
- For the linear mixed model, the question is: which are the correct log-likelihood and number of parameters to use?

The marginal AIC

The first option is to base the AIC in the linear mixed model on the marginal log-likelihood for the marginal model (3.5),

$$\begin{aligned}\log f(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\alpha}) = \ell_{ML}(\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log|\mathbf{V}_i(\boldsymbol{\alpha})| \\ &\quad -\frac{1}{2}\left\{\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{V}_i(\boldsymbol{\alpha})^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}.\end{aligned}$$

Statistical software (e.g. `lme`) often returns a marginal AIC using $\ell_{ML}(\widehat{\boldsymbol{\theta}}_{ML})$ and with df set to the total number of parameters in $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$.

The marginal AIC

- The marginal AIC as predictive quantity assumes that two independent replications $\boldsymbol{z}$ and $\boldsymbol{y}$ come from the same marginal distribution, but **do not share the same random effects**. It is thus appropriate when the focus is on the population-level **fixed effects**.

- The parameter space $\boldsymbol{\Theta}$ for $\boldsymbol{\theta}$ is not open (e.g. $d_{kk} \geq 0$), violating the usual regularity assumptions for the AIC.

- This induces a preference for models with fewer random effects (Greven & Kneib, 2010). The selection of fixed effects is likely not or not much affected.

The marginal AIC

For REML estimation, an AIC based on $\ell_{REML}(\widehat{\boldsymbol{\alpha}}_{REML})$ is often returned by statistical software (e.g. `lme`). The marginal AIC should not be used with REML estimation to select fixed effects as

a) the REML-likelihoods for different fixed effects are not comparable

b) the fixed effects do no even occur in the REML-likelihood

c) additionally, the used degrees of freedom often incorrectly still include the number of fixed effects.

The conditional AIC

An alternative is to base the AIC on the conditional log-likelihood

$$\begin{aligned}\log f(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\boldsymbol{\alpha}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log|\boldsymbol{\Sigma}_i(\boldsymbol{\alpha})| \\ &\quad -\frac{1}{2}\left\{\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)^T\boldsymbol{\Sigma}_i(\boldsymbol{\alpha})^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i)\right\}.\end{aligned}$$

The conditional AIC uses $\log f(\boldsymbol{y}|\widehat{\boldsymbol{b}},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\alpha}})$, where the predicted or estimated quantities can be based on ML or REML estimation. The conditional log-likelihood is always based on $\boldsymbol{Y}$ and valid with ML or REML estimation.

The conditional AIC

- The conditional AIC as a predictive quantity assumes that two independent replications $\boldsymbol{z}$ and $\boldsymbol{y}$ come from the same conditional distribution and **share the same random effects**. Vaida & Blanchard (2005) argue that it is appropriate when the focus is on the **random effects**.

- Greven & Kneib (2010) propose an unbiased estimator for the degrees of freedom in the conditional AIC (when $\boldsymbol{R} = \sigma^2 \boldsymbol{I}_n$), implemented in R-package `cAIC4` for models fitted with `lme4` or `gamm4`. The random effects, due to shrinkage, contribute between 0 and Nq df.

Example rat data

Consider again the random intercept model for the rat data

$$Y_{ij} = \beta_0 + b_{1i} + \beta_{g_i} t_j + \epsilon_{ij}$$

with transformed time t_j and compare with the untransformed time $TIME_j$.

```
> lmet <- lme(RESPONSE ~ group * t - group,
              random = ~ 1 | SUBJECT, data = rats, method = "ML")
> lmeTIME <- lme(RESPONSE ~ group * TIME - group,
              random = ~ 1 | SUBJECT, data = rats, method = "ML")
> anova(lmet, lmeTIME)
        Model df       AIC      BIC    logLik
lmet        1  6  931.9924  953.169 -459.9962
lmeTIME     2  6 1074.0125 1095.189 -531.0063
```

Interpretation?

Example sleep deprivation study

For the sleep deprivation data, compare a model with a random intercept with a model with random intercept and slope.

```
> library(lme4)
> library(cAIC4)
> M1 <- lmer(Reaction ~ Days + (1 | Subject), sleepstudy)
> M2 <- lmer(Reaction ~ Days + (1 + Days | Subject), sleepstudy)
> cAIC(M1)$caic
[1] 1767.118
> cAIC(M2)$caic
[1] 1711.618
```

Interpretation?

Overview Chapter 7 - Model building and model choice

7.1 Model diagnostics

7.2 Model choice

7.3 Assumptions and confounding

Excursus: Linear regression

In linear regression

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \tag{7.1}$$

we have the assumption

$$\mathrm{E}[\epsilon_i | \mathbf{x}_i] = 0 \tag{7.2}$$

s.t. $\mathrm{E}[Y_i | \mathbf{x}_i] = \mathbf{x}_i^T \boldsymbol{\beta}$. If this is not fulfilled, the estimator of $\boldsymbol{\beta}$ will be **biased**.

A common reason for violation of (7.2), i.e. **endogeneity**, is that an important **confounder** z_i was omitted from (7.1). E.g.

- Y_i the number of children born in a village
- x_i the number of stork nests in the same village

Omitted variable: z_i the number of roofs in the village.

Assumptions in LMMs

$$Y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + \epsilon_{ij}$$

In LMMs, we have two random variables in the model and two assumptions, which need to be fulfilled for $\boldsymbol{\beta}$ to be unbiasedly estimated.

- A similar **assumption on the residuals** as in the linear model: $\mathrm{E}(\epsilon_{ij}|\boldsymbol{x}_{ij'}, \boldsymbol{b}_i) = 0$ for all i, j, j', with $\boldsymbol{x}_{ij}$ the jth row of $\boldsymbol{X}_i$.

- In addition, the **random effects assumption** $\mathrm{E}(\boldsymbol{b}_i|\boldsymbol{x}_{ij}) = \mathbf{0}$ for all i, j.

If both are fulfilled, $\mathrm{E}(\mathbf{Y}_i|\mathbf{X}_i, \boldsymbol{b}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$ and $\mathrm{E}(\mathbf{Y}_i|\mathbf{X}_i) = \mathbf{X}_i\boldsymbol{\beta}$.

Sources of Variation and Confounding

Similarly to Chapter 1 and for a single covariate, think of decomposing

$$Y_{ij} = \beta_0 + \beta_B \bar{x}_i + \beta_W (x_{ij} - \bar{x}_i) + b_i + \epsilon_{ij},$$

$\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$. β_B and β_W may be differently affected by **confounding**.

- β_B is estimated from **between-subject** information. Does $\bar{Y}_i$ go up/down on average if $\bar{x}_i$ increases? $\mathrm{E}(b_i|\boldsymbol{x}_{ij}) \neq \mathbf{0}$ can occur due to confounding on the level of subjects, and $\widehat{\beta}_B$ will then be biased.

- β_W is estimated from **within-subject** information. Does Y_{ij} go up/down on average if x_{ij} increases relative to $\bar{x}_i$? $\mathrm{E}(\epsilon_{ij}|\boldsymbol{x}_{ij'}, b_i) \neq 0$ can occur due to confounding within subjects, and $\widehat{\beta}_W$ will then be biased.

We often do not decompose x_{ij}, and our estimate $\widehat{\beta}$ will then be a weighted average of $\widehat{\beta}_B$ and $\widehat{\beta}_W$. It is thus important to consider whether either source of information might be confounded (and if $\beta_B = \beta_W$ can be assumed).

Example of potential confounding within subjects:

- Y_{ij}: mortality count on day j in city i
- x_{ij}: PM_{10} level on day j in city i (time-varying)
- z_{ij}: temperature on day j in city i (time-varying) - both mortality counts (flue etc.) and PM_{10} levels are higher in the winter, see Ch. 1.1.

This leads to a violation of the assumption on the residuals.

Confounding between subjects - Example 1

Consider a **randomized trial** where at the beginning of the trial, subjects are randomized to treatment groups ($x = 1$ or $x = 0$). In this case, x_i is independent of b_i by design and the treatment effect can be unbiasedly estimated (e.g. rat data, TLC trial).

While violations of the second assumption cannot happen for x if we randomize with respect to x, they can occur in **observational studies** where we cannot control the x variables and confounding is possible.

Confounding between subjects - Example 2

Consider the model

$$Y_{ij} = \beta_0 + \beta_1 PM10_{ij} + b_i + \epsilon_{ij}.$$

- $PM10_{ij}$ is the personal exposure to PM_{10}, an air pollutant, and
- Y_{ij} is the FEV1 value, a measure of lung function, for subject i at time t_{ij}

and interest lies in the association between PM_{10} and FEV1. What happens if poorer people are

a) less healthy and thus have lower FEV1 values and

b) tend to live closer to big roads and are exposed to higher PM_{10} levels?

Then b_i, capturing the average FEV1 level $\bar{Y}_i$ of subject i, will be lower for poorer subjects, and due to the correlation between poverty and PM_{10}, also be lower for higher PM_{10} exposure.

As $E[b_i|PM10_{ij}]$ decreases with $PM10_{ij}$, the estimate of β_1 will be confounded, i.e. the estimate is too strongly negative.

The bias can be avoided if socio-economic information is included in the model, provided the effect is modeled well.

Confounding between subjects - Example 3

Consider the model

$$Y_{ij} = f(age_{ij}) + b_i + \epsilon_{ij}.$$

- Y_{ij} is the life satisfaction of subject i at time point t_{ij} in a panel study,
- age_{ij} is the age of subject i at time point t_{ij},
- b_i represents the individual tendency to be satisfied with life

and interest lies in the trend of life satisfaction with age.

What if happy people live longer? Then, $E[b_i|age_{ij}]$ is higher for older age_{ij} and the estimated trend will correspond to the **trend among survivors**.

Fixed vs. random effects b_i

Some people recommend replacing the random effects $\boldsymbol{b}_i$ by fixed effects $\boldsymbol{b}_i$ (fixed effects model) if there are doubts about the random effects assumption. Then, **only intra-individual variability** (within subject information) contributes to the estimates for $\boldsymbol{\beta}$ and each subject serves as its own control, cf. slide 14 in Ch. 1.1. Some pros and cons (see e.g. Townsend et al, 2013 for a full discussion):

\+ In the fixed effects model, estimators for $\boldsymbol{\beta}$ are unbiased if the assumption on the residuals holds.

– When the random effects assumption is satisfied, random effects models are more efficient. ("Bias-variance-tradeoff")

- Some n_i may be too small to estimate all random effects as fixed effects.

- Fixed effects models are more susceptible to violations of the first assumption, which can be more severe than the violation of the random effects assumption.

- Fixed effects models cannot estimate effects of time-constant variables (e.g. gender, treatment). Effects of time-varying covariates are less precisely estimated than in random effects models due to the additional degrees of freedom used. (E.g. for the effect of place of residency, only people who move during the study contribute to the estimate.)

Thus, to decide between a fixed and random effects model, one needs to weigh the plausibility of the two assumptions, whether the effects of interest can be estimated, and the tradeoff between how much bias can be reduced and how much efficiency is lost with the fixed effects model.

Fixed vs. random effects

The **Hausmann test** compares random and fixed effects model estimates of $\boldsymbol{\beta}$. A significant result is fairly reliable evidence for a bias in the effect estimates of the random effects model. (Non-significance unfortunately not necessarily indicates unbiasedness.)

An alternative is to decompose the information as on slide 42 (**hybrid model**). Estimates and standard errors for $\boldsymbol{\beta}_W$ are comparable to the fixed effects model, but time-constant variables can also be included in the model. A test for $\boldsymbol{\beta}_W = \boldsymbol{\beta}_B$ provides a Hausmann-like test.

Fixed vs. random effects - Example

Consider this example from Townsend et al, 2013 on US education policy:

$$NAEP_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \delta Standard_{i,j-1} + b_i + \epsilon_{ij}$$

- $NAEP_{ij}$ the average National Assessment of Educational Progress (NAEP) grade 4 mathematics score in state i in year j
- $\boldsymbol{x}_{ij}$ control variables measuring race composition, poverty etc. in state i and year j as well as year indicators
- $Standard_{i,j-1}$ the state performance standard for its state grade 4 mathematics test (with time lag), the policy variable of interest

Data for the 50 states are available for only 3 years, with some data missing.

Random effects assumption - Example 4

- The random effects model yields $\hat{\delta} = 0.058\ (0.024)\ (p < 0.05)$.
- The fixed effects model and the hybrid model both yield estimates $0.032\ (0.023)$ for δ respectively δ_W, a smaller and non-significant value. The estimate for δ_B is $0.164\ (0.041)$.
- The Hausmann test and the test for $\delta_W = \delta_B$ are both significant. Thus, the fixed effects model may be accounting for heterogeneity between states that can bias the δ estimate in the random effects model.