7. Model building and model choice

Sonja Greven

Summer Term 2016

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

Residual diagnostics 1

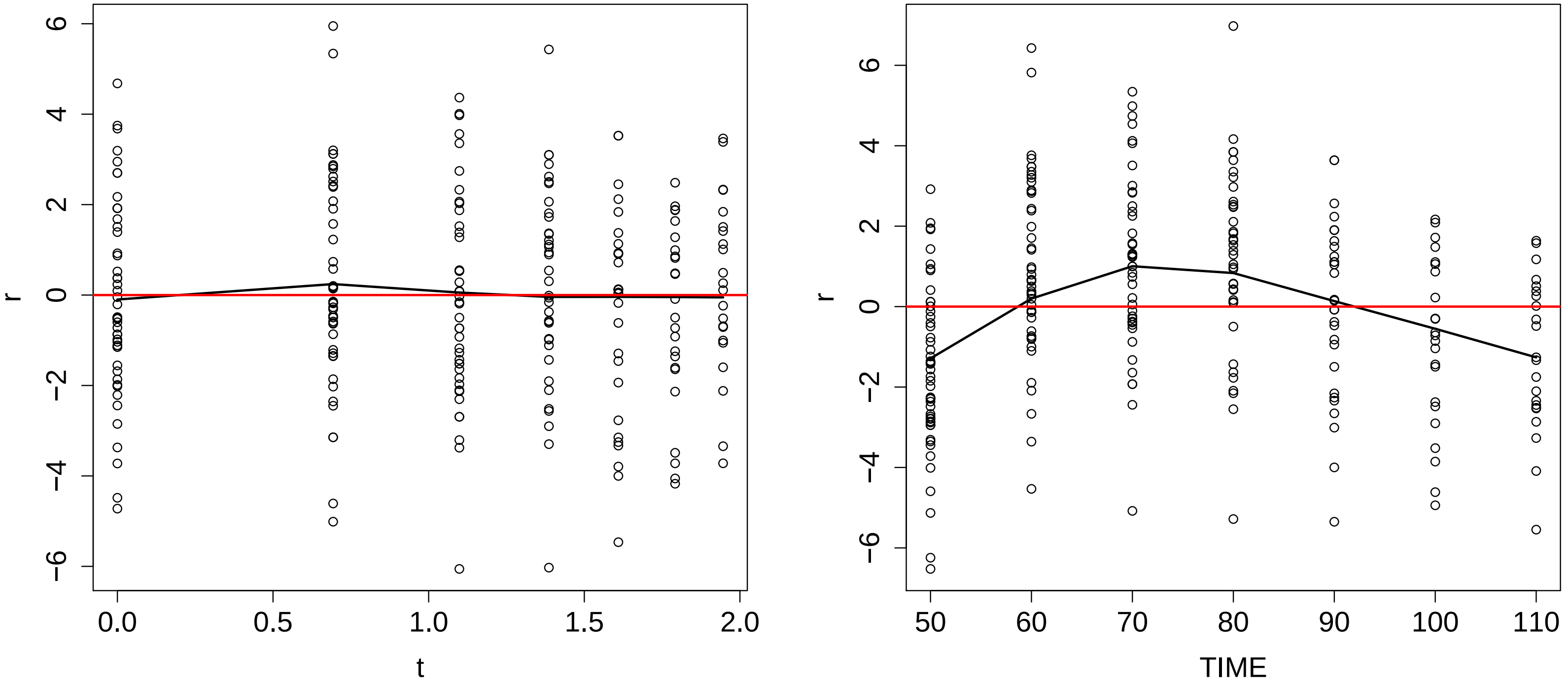
Plotting the residuals $r_{ij} = y_{ij} - \mathbf{x}_{ij}^T \widehat{\boldsymbol{\beta}}$ against covariates can help in diagnosing a misspecified mean structure, e.g. an omitted variable or a missing quadratic term. There should be no systematic trend!

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.

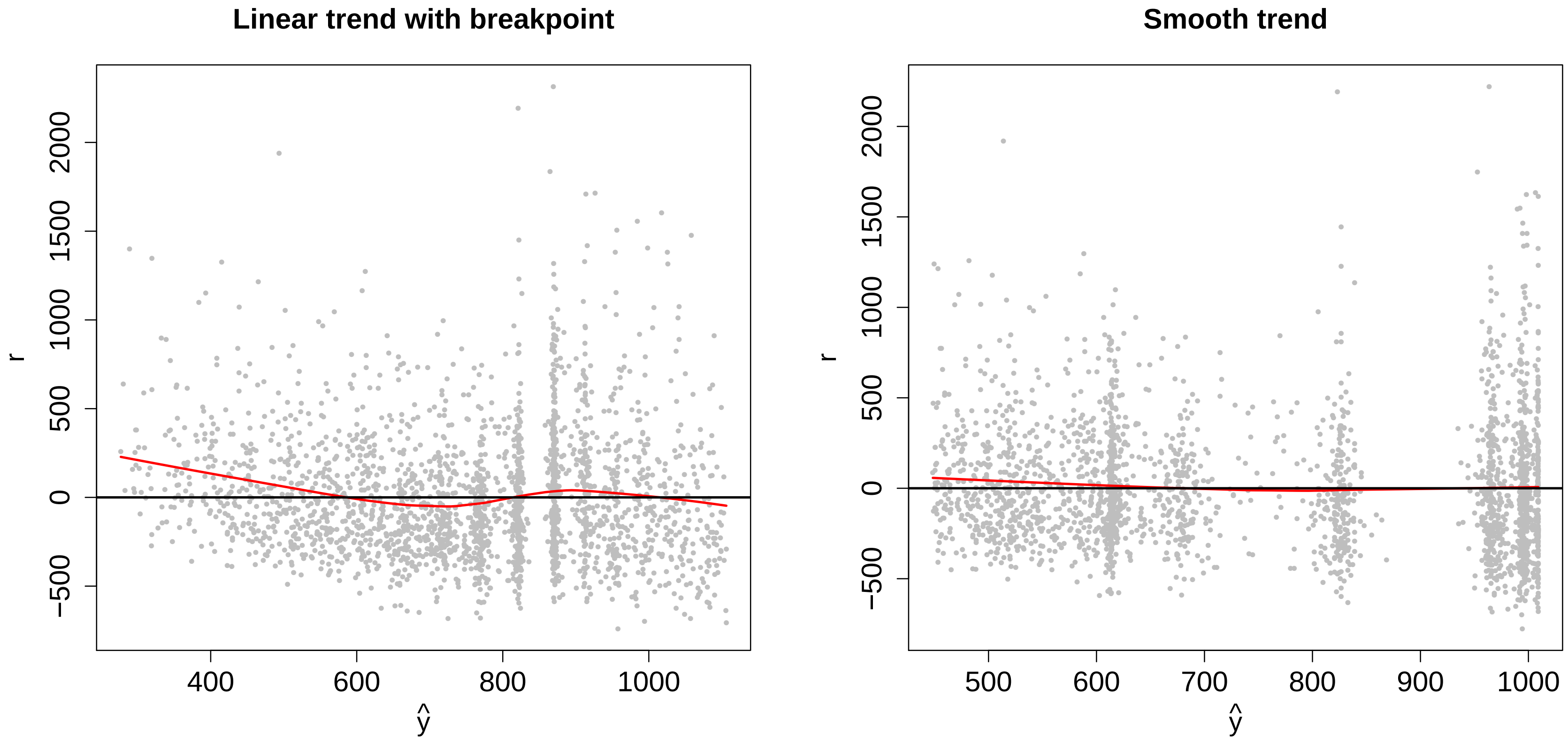
CD4 example, random intercept, linear time trend with breakpoint in 0:

```
> 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

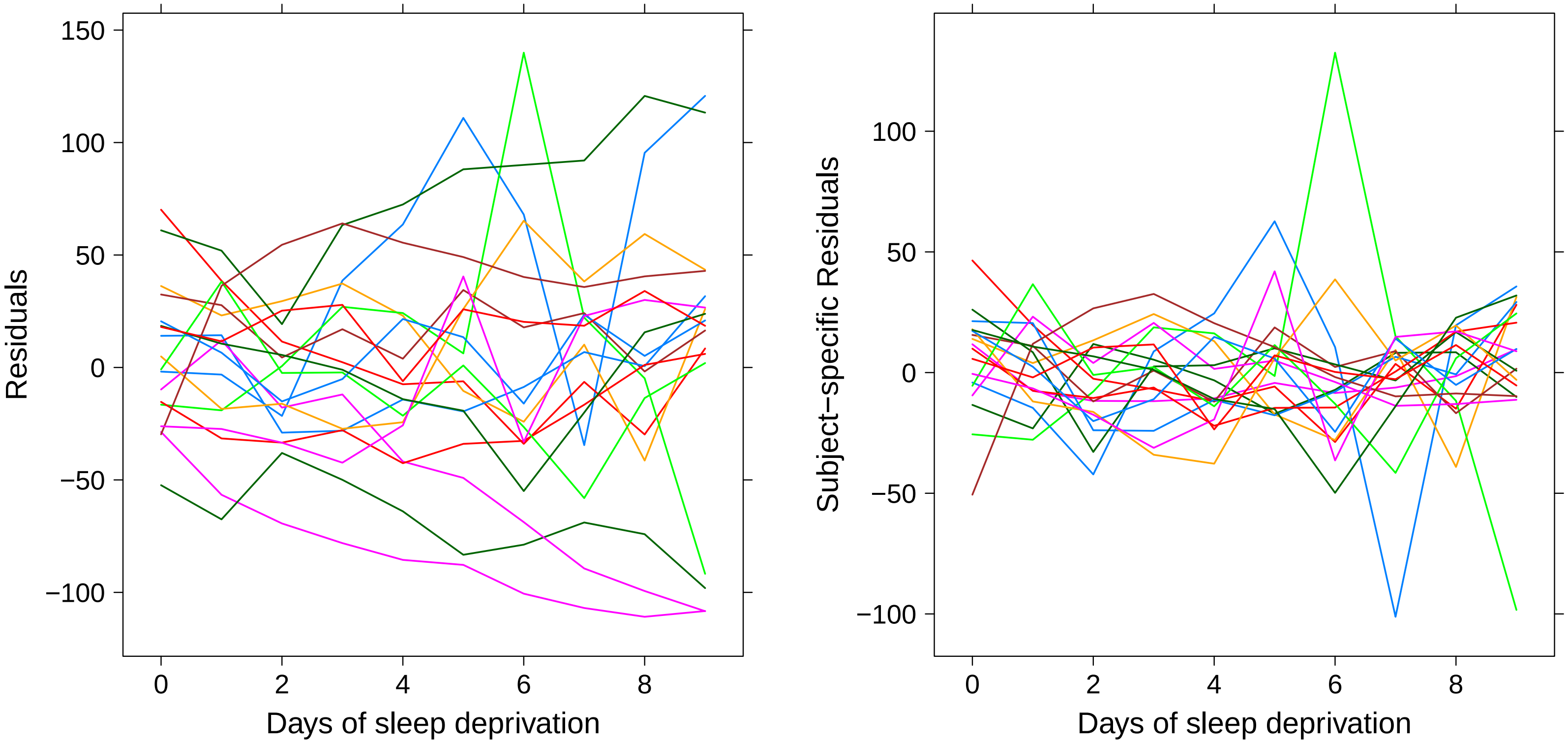
Plotting the residuals $r_{ij} = y_{ij} - \mathbf{x}_{ij}^T\widehat{\boldsymbol{\beta}}$ against covariates, e.g. time, can also indicate a missing random slope.

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 **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

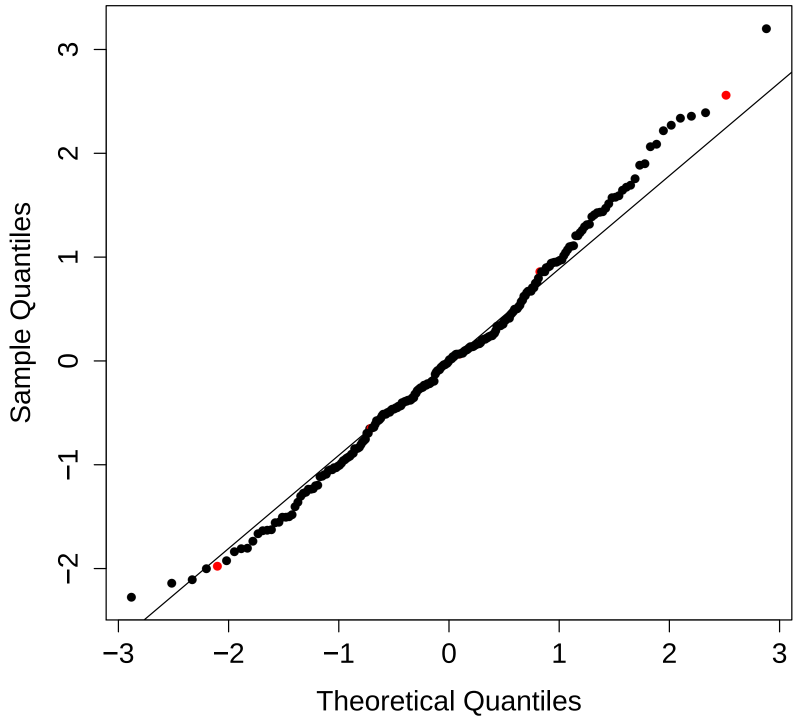
```
> library(RLRsim) # useful to extract lme model components
> r.star <- function(m){ # takes an lme object
+   design <- extract.lmeDesign(m)
+   Z <- design$Z
+   D <- design$Vr * design$sigmasq
+   R <- design$sigmasq * diag(nrow(Z))
+   V <- Z %*% D %*% t(Z) + R
+   L <- t(chol(V))
+   r.star 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

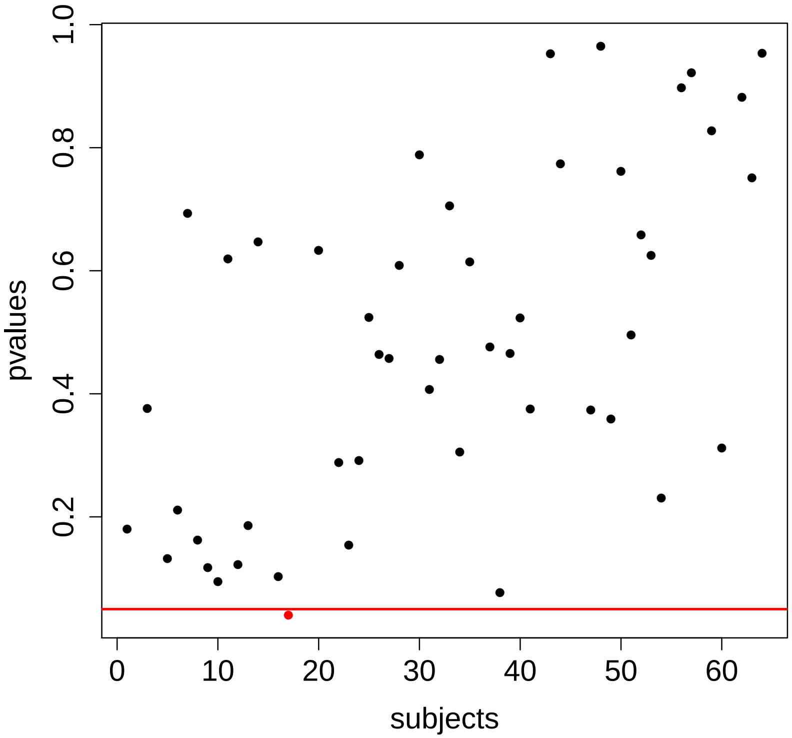
Example rat data

Random intercept model:

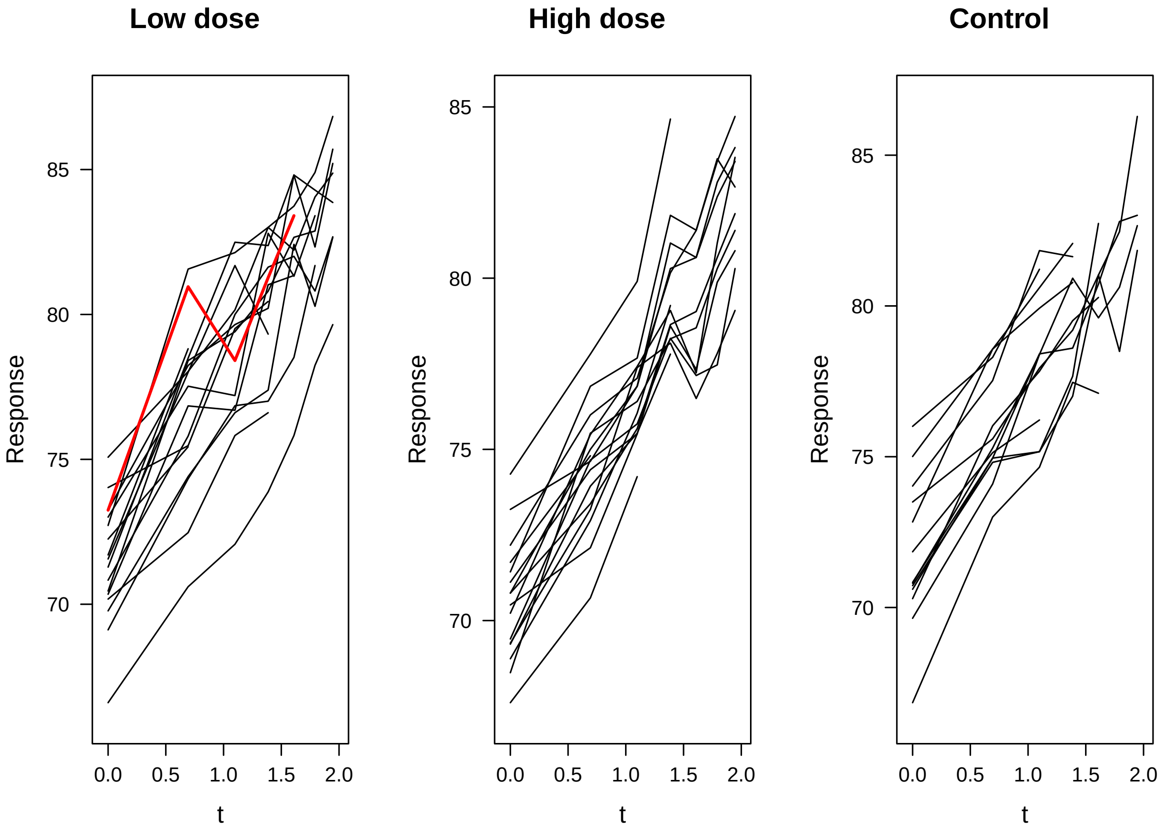
```
### Transformed residuals ###
> lme1 <- lme(RESPONSE ~ group * t - group,
              random = ~ 1 | SUBJECT, data = rats)
> r.star1 <- r.star(lme1) # transformed model residuals
### QQ-Plot ###
> qqnorm(r.star1)
> qqline(r.star1)
### Outlier Diagnostics ###
> subjects <- unique(sort(rats$SUBJECT))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

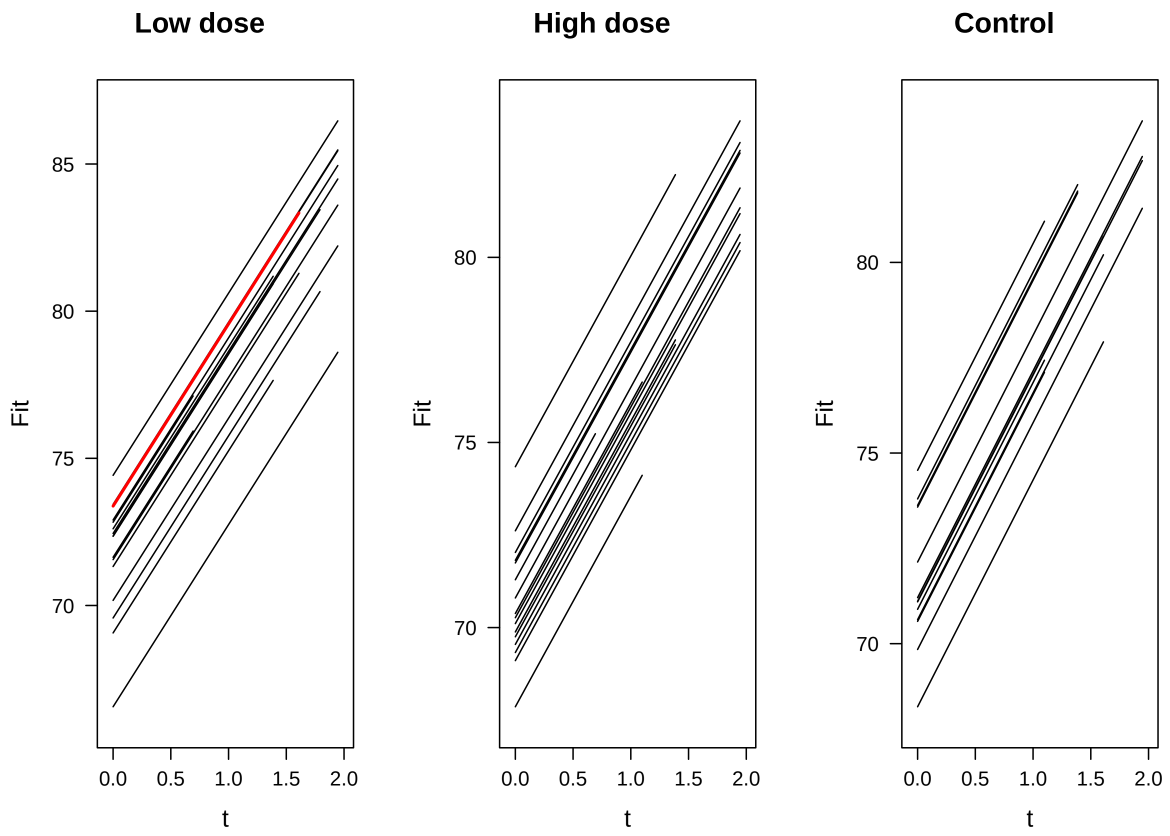




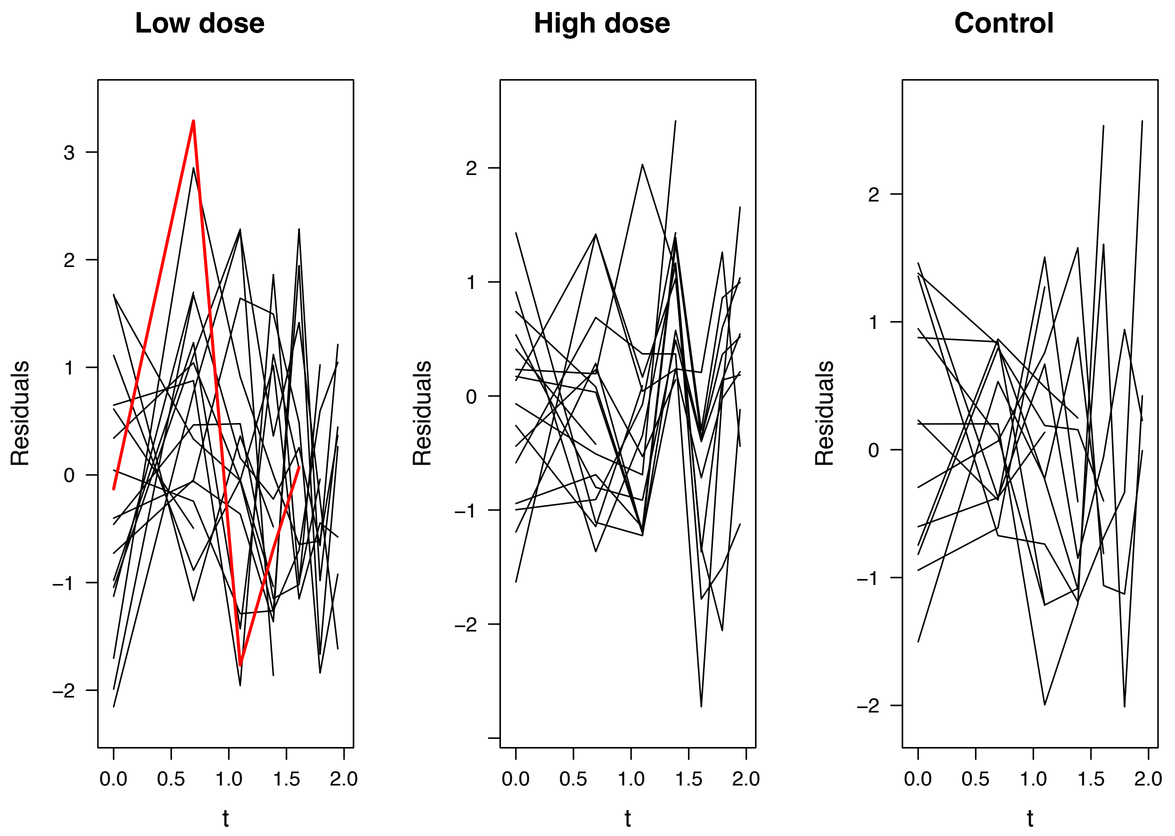
Example rat data - Data



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.

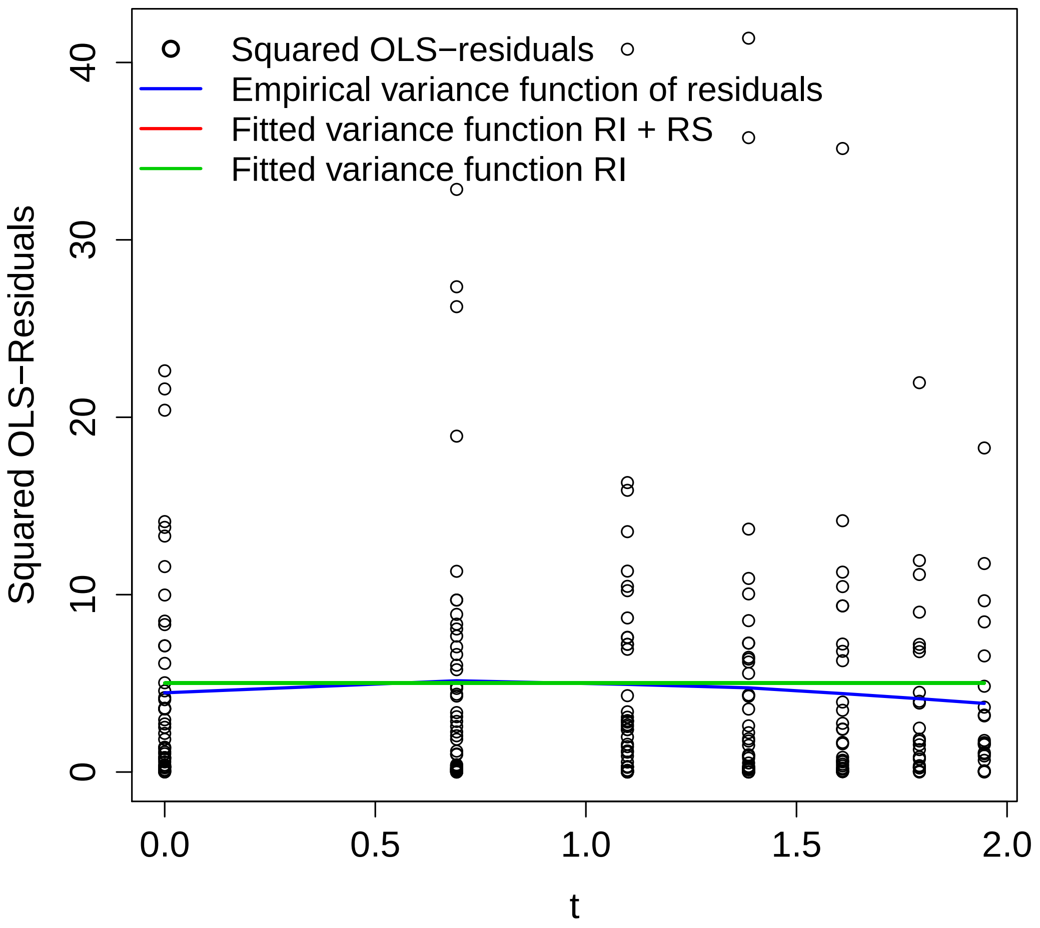
An informal check is to plot the squared OLS residuals

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

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.

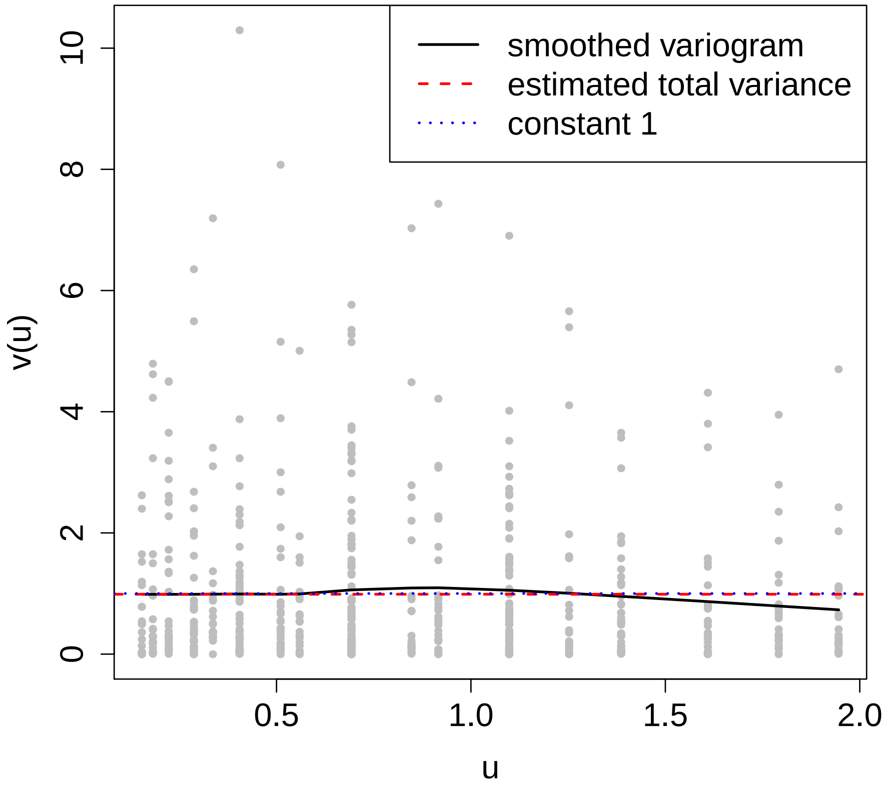
As the transformed residuals are approximately uncorrelated with mean zero and variance one, we have

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

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

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),

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

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?