Analysis of Longitudinal Data SoSe 2015 **Sheet 5**

Prof. Dr. Sonja Greven and Dipl. Stat. Jona Cederbaum 15.06.2015

On this exercise sheet we repeat and deepen model diagnostic and model choice in linear mixed models (lecture slides 7) and start considering non-normal longitudinal data and their modeling capabilities (lecture slides 8).

Exercise 1: Model diagnostic

In this exercise, we focus on the diagnosis of estimated linear mixed models to verify the underlying assumptions and specifications.

(a) Download the data set `vitamin` from the homepage and read through the description. Take a first look at the data.

(b) Estimate a linear mixed model (m_RIRS) with random intercepts and random slopes for each child as well as fixed effects for `time` and for the interaction of `group` and `time`. Assume that there is no serial correlation.
Note: In order to estimate no main effect for `group`, use `group*time - group` in the `formula`.

(c) Why do we not need to consider the main effect for `group` here?

(d) What will you get if you call `predict(m_RIRS)` and what does `predict(m_RIRS,level=0)` give you?

(e) You now want to evaluate the model fit. Therefore, it is common to plot the residuals against the covariates. Which two model weaknesses can be found by this?

(f) Now look at the population-specific residuals $r_{ij} = y_{ij} - x_{ij}^\top \hat{\beta}$ and plot the residuals against the covariate `time`. Interpret the plot.

(g) Which other plot could you look at to check for misspecifications of the mean?

(h) Why is the consideration of a quantile-quantile plot for the residuals r_{ij} inappropriate?

(i) Which alternatives to the residuals r_{ij} could be considered?

Exercise 2: Model choice

In this exercise, we will compare linear mixed models and select the one which is more appropriate.

(a) In exercise 1, we have seen that the mean was not well specified, it may therefore make sense to use the transformed variable `log(time)` instead of `time`. In the following, estimate the model `m_RIRSlog` which is identical to the model fitted in exercise 1 apart from the transformation of `time`. Use **ML** (instead of REML) for the estimation and estimate model `m_RIRS` once more **using ML** as well.
Note: Keep in mind that the random slopes have to be adjusted by using the transformation as well.

(b) How can the models `m_RIRS` and `m_RIRSlog` be compared, i.e. how can you choose which of the two models is more appropriate? To which decision do you come regarding your model selection?

(c) What difficulty would arise if the above models, which we want to compare, were estimated by REML?

(d) What assumption is made when considering the marginal and the conditional AIC and what can be concluded from this for their use?

(e) In the following, consider the model `m_RIRSlog` once more but without random slopes. What would happen if you used the marginal AIC for a comparison of the model with random slopes and the model without random slopes?

(f) Instead, use the conditional AIC (included in the `R` package `cAIC4`) to compare the two models. To which decision do you come?
Note: In order to use the function `cAIC{cAIC4}`, estimate the models using the function `lmer` included in the package `lme4` (cf. lecture slides). Consider the help `?lme4` to understand the main differences compared with the already known function `lme`.

Exercise 3: Non-normal longitudinal data

In the following, we are leaving the normal distribution assumption of $\mathbf{Y}_i$ and allow that the distribution of the data belongs to the exponential family. This is analogous to the transition from linear models to generalized linear models (GLM).

(a) As a first step, consider a couple of own examples of **non-normal longitudinal** data (not the ones from the lecture!).

(b) In the data set `epil` in the `R` package `MASS`, the numbers of seizures over time are available for 59 epileptics. The probands were randomly assigned to a treatment group receiving the drug Progabide and a placebo group.

(i) Which model would you choose if you were interested in individual predictions for the probands?

(ii) Which model would you choose if you were interested in the population effect of the drug Progabide?

(iii) Let $\hat{\beta}$ be the estimated effect of Progabide derived from the model fitted in (i). What do you have to consider regarding the interpretation of $\hat{\beta}$? Does the interpretation correspond to the one of the effect of Progabide derived from the model in (ii)?