Analysis of Longitudinal Data SoSe 2015 **Sheet 1**

Prof. Dr. Sonja Greven and Dipl. Stat. Jona Cederbaum 20.04.2015

This exercise sheet will familiarize you with the structure and characteristics of longitudinal data, as well as with their graphical representations in `R`. The exercises refer to the content of the first and second lecture slides.

Note: Due to the wealth of material, this tutorial is designed in a way, that it is expected that at least an attempt to solve the exercises was made **in advance**.

Exercise 1:

In this exercise, we are working with the data set `rats`. First, please read through the description of the data set (on the homepage).

a) Download the data set from the homepage and import it in `R`. Make sure that NAs are recognized as such (*Note:* Use option `na.strings()`). Convert the variable `GROUP` into a factor variable with corresponding labels for the three treatment groups. Convert the variable `SUBJECT` into a factor variable as well and take a first look at the data.

b) Now use the function `reshape()` to reformat the data set `rats` in a way, that there is one row for each animal in the data set, and name the reformatted data set `rats.wide`. Use `rats.wide` to gain an overview of the failure rates in the different treatment groups.

i) Do the failures (`NAs`) follow a specific mechanism?

ii) For how many animals in each treatment group are all six or only five, four, three, etc. measurements available? Represent the quantities over time graphically. Is it possible to detect a pattern?

c) Using the `groupedData` format from the `R` package `nlme`, many calls for estimates and plots for longitudinal data can be simplified. Therefore, generate a new data set `rats2` in `groupedData` format, based on the original data. Now plot the individual curves for each animal by treatment group.

d) Fit a linear model for all rats (pooled) with `TIME` as covariate and visualize the individual curves of the residuals of each animal (using the package `lattice`). You can use the code available on the homepage for this.

i) What do you notice considering the plot ?

ii) Are the assumptions of the ordinary linear model met?

Exercise 2:

In this exercise, we are working with the data set `cd4` (on the homepage). It includes 2 376 measurements of the number of CD4 cells in the blood of 369 men infected with HIV/sick with AIDS before and after the date on which the presence of HIV antibodies was detected in the blood for the first time (*seroconversion*). The number of CD4 cells serves as a biomarker for the condition of the immune system. The main interest is to determine the progress of the CD4 content over time.

a) Import the data set in `R`, convert the variables `drug` and `ID` into factor variables (with corresponding labels for the variable `drug`) and get a first overview of the data.

b) Which of the covariates in the data set are time constant, which are time varying?

c) Now plot the individual curves for each patient and estimate a smooth mean function using the code available on the homepage.

d) Repeated measurements of a subject are usually correlated, which is reflected in the residuals. Calculate the residuals by subtracting the smooth mean curve from c).

i) For each subject, build the pairwise products of the residuals as an estimator of the covariance using the code available on the homepage.

ii) Plot all pairwise products in a scatter plot against the distance of the corresponding measurements and add a smooth mean curve using the function `loess()`. What can you see from the plot?

e) Consider the following model for the cd4 data

$$\texttt{CD4}_{ij} = \alpha_d z_{ij} + \alpha_{\bar{d}}(1 - z_{ij}) + \beta_d z_{ij} t_{ij} + \beta_{\bar{d}}(1 - z_{ij}) t_{ij} + \varepsilon_{ij}, \tag{1}$$

where $\texttt{CD4}_{ij}$ denotes the jth CD4 measurement of subject i at time t_{ij} and $z_{ij} = 1$ if no drugs were taken and 0 otherwise.

i) Fit a general linear model using the function `gls()` from the package `nlme` assuming independence of all measurements.
Note: Use the specification `correlation=NULL`.

ii) It can be assumed that the residuals of a subject, $\varepsilon_{ij},\ i = 1, \ldots, n_i$, are not independent. For a new estimation assume that the correlations of all deviations of subject i - regardless of the time lag - are equal and compare the coefficients and standard errors of both estimations. What do you notice?
Note: Use `correlation=corCompSymm(form=~ 1|ID)` for the second estimation.

iii) Which other assumptions could reasonably be made for the correlation structure of the residuals?