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 *Correlation in longitudinal data: basics and notation*

Let $\boldsymbol{\xi} \in \mathbb{R}^n$ be a random variable with $\mathbb{E}(\boldsymbol{\xi}) = \mathbf{0}$ and $\mathbf{V} = \text{Cov}(\boldsymbol{\xi})$ and let $\mathbf{A}$ be an $m \times n$ -matrix.

a) Show that $\text{Cov}(\mathbf{A}\boldsymbol{\xi}) = \mathbf{A}\mathbf{V}\mathbf{A}^\mathsf{T}$.

Consider a random sample of length n for one individual $i = 1$ with response $\mathbf{Y}_i = (Y_1, ..., Y_n)^\mathsf{T}$ and co-variate design matrix

$$\mathbf{X}_i = \begin{pmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in} \end{pmatrix}$$

with $x_{ij} \in \mathbb{R},\ j = 1, ..., n$. We consider the linear model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\xi}_i$$

with coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^2$ and $\boldsymbol{\xi}_i = (\xi_{i1}, ..., \xi_{in})^\mathsf{T}$ a random vector of length n with co-variance matrix V_i.
Note: For simplicity we assume data was only sampled for a single individual. Nevertheless, the index 'i' for the individuals is introduced in preparation for the lecture.

b) Compute $\text{Cov}(\hat{\boldsymbol{\beta}})$ for the simple least squares estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$.

c) Let $\xi_{i1}, ..., \xi_{i(n-1)}$ be i.i.d. with a standard normal distribution and $\xi_{in} = \frac{1}{n-1}\sum_{j=1}^{n-1} \xi_{ij} \prod_{g=1}^{n-1} \text{sgn}(\xi_{ig})$. Show that $\text{Cov}(\xi_{ij}, \xi_{in}) = 0$ for all $j = 1, ..., n-1$, if $n \geq 4$. What can you follow for the relationship between correlation and stochastic dependency?

d) Let $\boldsymbol{\xi}_i$ follow a multivariate normal distribution and let $\text{Cov}(\boldsymbol{\xi}_i)$ be a diagonal matrix. Show that $\xi_{i1}, ..., \xi_{in}$ are independent.

e) Now, let $\boldsymbol{\xi}_i = \mathbf{X}_i\mathbf{b}_i$ for a random vector $\mathbf{b}_i \in \mathbb{R}^2$ with $\mathbb{E}(\mathbf{b}_i) = \mathbf{0}$ and $\text{Cov}(\mathbf{b}_i) = \mathbf{D}$. Draw a sketch of a possible realization of the model for $n = 5$ and determine $\text{Cov}(\hat{\boldsymbol{\beta}})$. How many measurements will it take until the estimate does (ideally) not change anymore? How does the co-variance reflect this?

Exercise 2 *Visualizing longitudinal data*

In this problem we discuss the data set `rats`. Please read the data description (on the homepage).

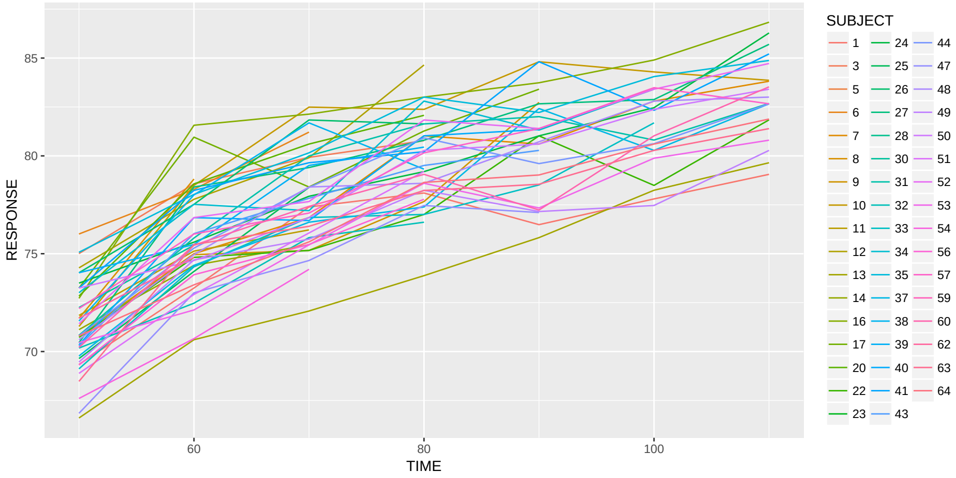
a) i) Download the data set `rats` from the homepage and import it with `read.csv2` in `R`. Convert the variables `GROUP` and `SUBJECT` into factor variables with suitable labels.

ii) Convert the data set to 'long format'. I.e. transform it such that there is one row per measurement in the data set and all measurement times are contained in a column `TIME` and all response values in a column `RESPONSE`. Call this new data set `rats.long`. *Note: You might use the function* `melt` *in the* `R` *package* `reshape2` *for this.*

b) In the following, the data set has been visualized using the `R` package `ggplot2`. Each plot has its particular purpose. However, they are still problematic. Identify the problems and collect ideas for enhancement. Implement one idea in `R`.

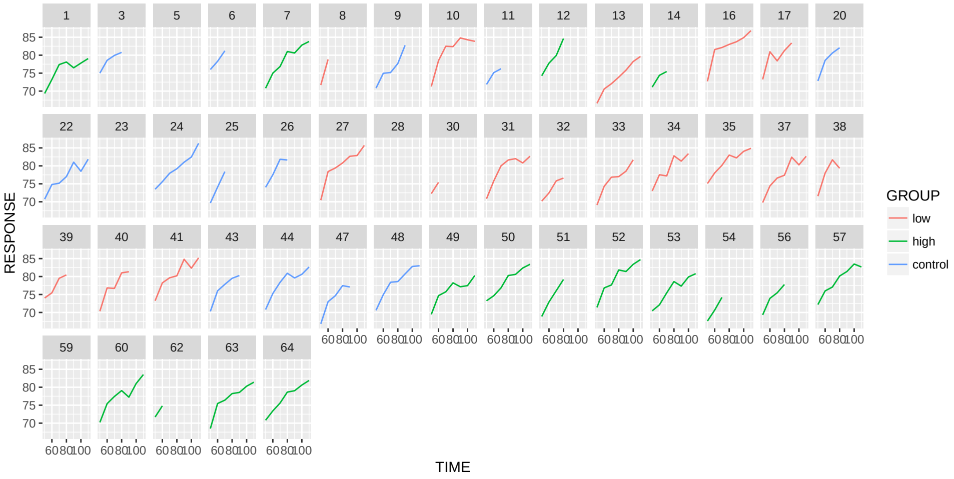
i) Purpose: You want refer to this plot for explaining the data structure (e.g. to a colleague).

```r
library(ggplot2)
ggplot(rats.long, aes(x = TIME, y = RESPONSE, col = SUBJECT)) +
  geom_line()
```



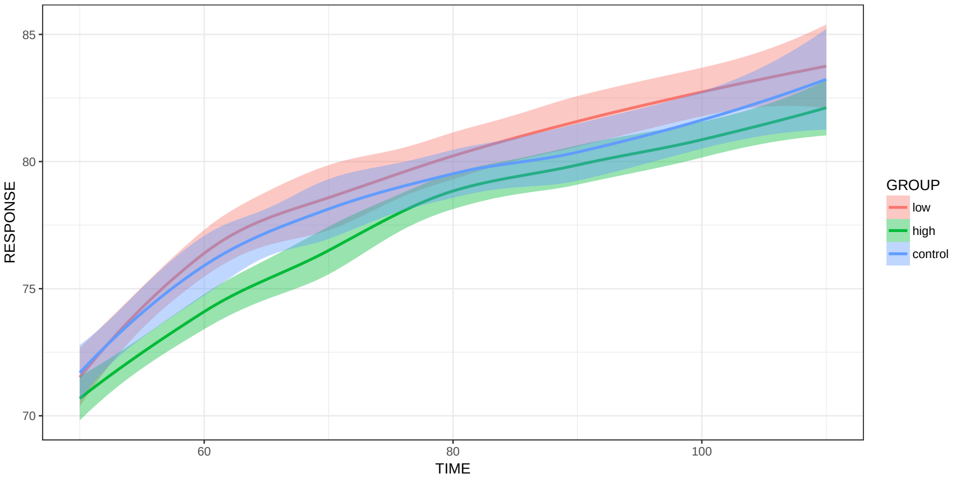
ii) Purpose: You want to discuss patterns in the data with a biologist, who carried out the experiment and might remember details concerning single animals.

```
ggplot(rats.long, aes(x = TIME, y = RESPONSE, col = GROUP)) +
      geom_line() + facet_wrap(~SUBJECT, ncol = 15)
```



iii) Purpose: You want to depict group differences using smooth mean curves.

```
ggplot(rats.long, aes(x = TIME, y = RESPONSE, col = GROUP, fill = GROUP)) +
  stat_smooth() + theme_bw()
```



c) Create a Lasagna Plot for the data. Which advantages and disadvantages do you see comparing Lasagna Plots to usual plots like those above?