Analysis of Longitudinal Data SoSe 2016 **Sheet 7**

Prof. Dr. Sonja Greven and Dipl. Stat. Jona Cederbaum 12.07.2016

This exercise sheet will familiarize you with marginal models for non-normal longitudinal data on the one hand – in particular with the estimation using generalized estimating equations (GEEs) – and with missing values on the other hand. The exercises refer to the content of the tenth and eleventh lecture slides.

Exercise 1: GEE

The data set `leprosylong.txt` contains the number of leprosy bacilli before and after treatment with antibiotics. For each patient, the following variables exist: `Drug`, `count` (count of bacilli) and `time`. Drug C corresponds to a treatment with placebo. `time = 0` corresponds to the condition before treatment, `time = 1` to the condition after treatment.

(a) Determine the mean and the variance of the numbers of bacilli before and after treatment for each treatment group and plot the courses separated by groups. For which variables do you expect a significant effect and what would you criticize in the study?

(b) At first, fit a Poisson GLM with time, treatment and their interaction as variables.

- **(i)** Interpret the coefficient estimates.
- **(ii)** Compute an estimate for the dispersion parameter ϕ based on the Pearson residuals.

(c) Which assumptions have to be made for the estimation of a marginal model?

(d) Estimate a marginal model with the same variables as in model (b) using GEE. Assume independent observations at first.
Note: Use the function `gee` included in the package `gee` with the argument `corstr = ‘independence’`.

- **(i)** What lies behind the “robust” and “naive” estimations of the variances of the parameters?
- **(ii)** In which cases is the robust variance estimation appropriate?
- **(iii)** In which case are both variance estimations equal?

(e) Read the `R`-help of the function `?quasipoisson` and estimate a quasi-Poisson model using the function `glm`.
Note: Use `family=quasipoisson(link=‘log’)`.

- **(i)** What do you notice when you compare the results with those from (d)?
- **(ii)** Show that the coefficient estimators of the simple Poisson model correspond to those of the quasi-Poisson model.

(f) Use the Pearson residuals of the quasi-Poisson model to obtain a preliminary estimate of the correlation between two measurements of a patient.

(g) Estimate the model once more under the more realistic assumption that observations on the same patient are not independent. Use an unstructured correlation.

(h) What does the "generalized" in the name of the estimating equations GEE refer to?

Exercise 2: Missing values

This exercise is about classifying missing values in missing mechanisms. This can be very important in order to decide which methods are applicable and which are not. Consider the following LMM

$$Y_{ij} = \beta_{x:t} x_{ij} t_{ij} + b_i + \epsilon_{ij},$$

with $i = 1, \ldots, N = 30$ individuals and two time points $t_{i1} = 0$ and $t_{i2} = 1$ for all $i = 1, \ldots, N$. Let x_{ij} be a binary variable.

Let R_i be an indicator for observing the second measurement of the i-th individual, i.e. $R_i = 1$ if Y_{i2} was observed and $R_i = 0$ if Y_{i2} is missing.

(a) Why is it dropout when missing values for Y_{i2} exist?

In the following, we consider different dropout probabilities.

(b) Consider $P(R_i = 1|Y_{i1}, Y_{i2}, \mathbf{X})$ for $R_i \sim \text{Bernoulli}(\pi_i)$ with

$$\text{logit}(\pi_i) = \gamma_0 + \gamma_x x_{i1} + \gamma_1 Y_{i1} + \gamma_2 Y_{i2}.$$

Specify the restrictions for $(\gamma_0, \gamma_1, \gamma_2)$ if dropout is

(i) MCAR

(ii) MAR

(iii) NMAR

Besides, find examples of all three types of missing values.

(c) Consider instead

$$\text{logit}(\pi_i) = \alpha_0 + \alpha_1 b_i$$

with $\alpha_0 = -0.5$ and $\alpha_1 = 3$. What kind of dropout mechanism do we have here? Give reasons for your answer.