Prof. Dr. Sonja Greven,
Dipl. Stat. Jona Cederbaum
Department of Statistics

Question	1	2	3	4	5	6	$\sum$
Points							

Exam in Analysis of longitudinal Data

Summer term 2016, 22. July 2016

Name (in block capitals): _____

Matriculation number: _____

Degree Program: _____

I confirm that I have taken note of the information below and that I have checked the exam copy for completeness. I agree with the publication of my exam results (matriculation number, grade) in the Department of Statistics, Ludwigstr. 33. (If not, please cross out last sentence.).)

Signature: _____

Information for the exam (1–8):

1. First check if the exam copy is complete. The instructions should consist of **7 sheets** with **6 questions**.
2. The examination duration is **90 minutes**. It is not possible to submit the exam during the first 30 minutes and during the last 15 minutes.
3. A maximum of **85 points** can be scored. A comprehensible solution is required in order to obtain the full score. The bonus question (Question 6) is optional. The full number of points can also be scored without solving it.
4. The following auxiliary materials are allowed:
   - calculator
   - 2 DIN-A4 sheets of paper with handwritten notes on front and back
   - a dictionary if necessary
5. For your answers, please use the exam sheets only. In case the space is not enough, additional sheets of paper can be requested from the supervisors during the exam. Mark each sheet (once) with your name and matriculation number.
6. Have your photo ID and a valid student card ready.
7. In case of peculation, the examination office will be informed. It is your obligation to rule out any suspicion concerning this matter.
8. Leave the examination room only after you have handed over your exam to the supervisor personally. You are responsible that the supervisor receives the complete exam. Please leave the lecture tract after the exam fast and quiet, so that you do not disturb the participants of other exams.

Good luck!

Question 1	14 Points

The following model stems from a study on the treatment of Adipositas (obesity). The obese patients were sent to two different clinics (`clinic`) in which they were treated for 60 days. To avoid that the patients are too far away from their families, they were assigned to the clinics by residence (city/country side). In both clinics, the state of health of each patient was measured daily (`day`) in form of a score (`score`). One is mainly interested in whether there is a difference in the score development between the two clinics.

```
model1 <- lme(score ~ day * clinic, random = ~ 1 + day | id, data = adipos)
```

a) (1P) What kind of model was estimated here?

b) (2P) Why would it not be useful to apply a fixed effects model here?

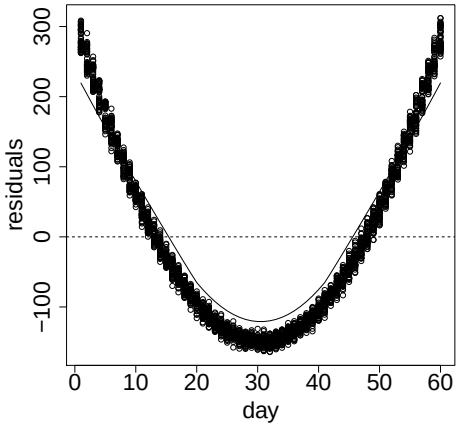
c) (3P) Why can you not necessarily assume that the random effects assumption regarding the covariate `clinic` is fulfilled in the specified model? Justify your answer based on a suitable example.

d) (1P) Which assumption is made in this model for the association between the random intercepts and the random slopes?

e) (5P) Which test could you use to test whether the assumption on the association between random intercepts and random slopes could be simplified? Also specify the null hypothesis H_0 and the distribution of the test statistic under H_0. Name all used components.

f) (2P) What can you conclude from the following plot concerning the model diagnostics? **Briefly** justify your answer.

```
r <- resid(model1, level = 0)
plot(adipos$day, r, xlab = "day", ylab = "residuals")
lines(lowess(adipos$day, r))
abline(h = 0, lty = 2)
```



Solution 1

Question 2 27 Points

In the last years, more and more so-called "super foods", which are said to be very healthy, entered the market. For example, chia seeds are said to counteract thrombosis due to their high content of Omega3 fatty acids. In order to evaluate the actual benefit of the daily consumption of chia seeds, the consumer protection ministry carried out a study in which the subjects were randomly assigned to two groups. Over a two year period, the subjects in group 1 (`group=0`) should not eat chia seeds and the subjects in group 2 (`group=1`) should eat chia seeds every day. Starting one month after the beginning of the study, monthly (`month`) information for all subjects was collected on whether a thrombosis (`thrombosis`) occurred or not (yes/no). The model that was used yields among others the following output:

```
model3 <- glmmPQL(fixed = thrombosis ~ month + group,
          random = ~ 1 | id, family = binomial, data = dat)

summary(model3)
Linear mixed-effects model fit by maximum likelihood
 Data: dat
  AIC BIC logLik
   NA  NA     NA

Random effects:
 Formula: ~1 | id
        (Intercept)  Residual
StdDev:   0.8639935 0.9511475

Variance function:
 Structure: fixed weights
 Formula: ~invwt
Fixed effects: thrombosis ~ month + group
                 Value  Std.Error   DF    t-value p-value
(Intercept)  0.0332044 0.09474085 4599   0.350476  0.7260
month        0.0447339 0.05913039 4599   0.756531  0.4494
group       -1.9518854 0.14313599  198 -13.636580  0.0000

Number of Observations: 4800
Number of Groups: 200
```

a) (13P) Write down the model for this concrete application. Specify and name **all** components and write down **all** assumptions. *Note:* Unless specified differently, the natural link function was used.

b) (6P) Interpret the estimate of covariate `group`.

c) (2P) For which reason does the estimation criterion in the estimation of generalized linear mixed models have to be approximated?

d) (2P) For which reason does the model output not supply an AIC, BIC and a log-likelihood value?

e) (3P) The ministry is mainly interested in the expected change of the odds to get a thrombosis in the population if everyone were to consume chia seeds on a daily basis. The two statisticians who evaluate the study come to different conclusions: The first statistician suggests that the average log odds to get a thrombosis for a fixed month fall by approx. 1.952 when chia seeds are consumed every day. The second statistician argues that the effect is smaller and the log odds fall by less than 1.952 (ceteris paribus).
Which of the two is correct? Justify your answer.

f) (1P) Which model could alternatively have been used to estimate the odds for thrombosis in the population if everyone were to consume chia seeds on a daily basis?

Solution 2

Question 3	8 Points

a) (1P) Why are the random effects in the linear mixed model not predicted on the basis of the marginal log-likelihood?

b) (1P) **Briefly** describe how the marginal model for Gaussian data can be derived from the conditional model?

c) (2P) How does the spread of the predictions of the random effects $\boldsymbol{b}_i$ change compared to the spread which is obtained if the $\boldsymbol{b}_i$ are instead estimated as fixed effects (in words)? What is this effect called?

d) (4P) Now consider the compact matrix representation of the linear mixed model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon}.$$

What does the covariate matrix, $\boldsymbol{Z}$, of the random effects look like for all individuals together and what is its dimension? Name all components.

Solution 3

Question 4 16 Points

In a growth study, the height (`height`) of boys and girls is measured monthly by the parents of the children for three years starting with age 10. The parents should actually measure the height every month (`month`). However, some parents are not very consequent and do not take regular measurements. For this reason, the measurements in the last year of the study were taken for all children together in the study center.

a) (2P) Are the data balanced? **Briefly** justify your answer.

b) (3P) In order to get a first impression of the data, you want to depict the mean curve over all children. Propose an explorative method to proceed in this data situation. Why should you be careful to only use it as an explorative tool?

The following model is used for the analysis of the data:

```
model2 <- lme(height ~ sex * month, random = ~1|child, data = growth),
```

where `sex` is a dummy variable for the sex of the children and `child` is a factor-coded ID-variable.

c) (3P) Name three properties of the marginal correlation between two measurements on the same child in `model2`.

d) (1P) How large is the correlation between two measurements on the same child conditional on the random intercept in `model2`?

e) (2P) Do we have dropout here for the missing values? **Briefly** justify your answer.

f) (3P) Which missing mechanism can be assumed here? Justify your answer.

g) (1P) Which estimation method was used to estimate the variance parameters in your model?

h) (1P) **Briefly** explain why the estimates of the fixed effects can differ depending on whether the variance parameters were estimated using maximum likelihood or restricted maximum likelihood.

Solution 4

Question 5	15 Points

In the following, you see the beginning of five different statements (A-E). Complete each sentence with one of the three given options. For each sentence, exactly one option is correct. Mark the matching option on your exam copy (no justification necessary). For each correct answer, you obtain 3 points, for each incorrect answer 1.5 minus point. Maximal 15 and minimal 0 points can be obtained in this question.

A. The procedure for maximum likelihood estimation in the linear mixed model is as follows:

(1) The fixed effects and the variance parameters are estimated simultaneously by maximizing the residual maximum likelihood.

(2) First, the fixed effects are estimated under a working independence assumption and then plugged into the estimates of the variance parameters. As the fixed effects are estimated first, the loss of degrees of freedom leads to a biased maximum likelihood estimator for the variance parameters.

(3) The variance parameters are estimated by maximizing the profile-likelihood and the estimates are then plugged in to obtain the estimates for the fixed effects.

B. In how far does the linear mixed model (LMM) represent a very special particular case of the generalized linear mixed model (GLMM)?

(1) In contrast to the the GLMM, for the LMM there is an analytic formula for the estimators of the fixed effects and of the variance components.

(2) In the LMM, additional assumptions on the error term can be made and one does not always assume conditional independence as in the GLMM.

(3) The conditional and the marginal perspective mutually imply each other in the LMM.

C. Missing values frequently occur in the collection of longitudinal data, which should be accounted for in the choice of the analysis methods. When there are missing values,

(1) only subjects for which there are no missing values should enter the analysis.

(2) the estimators based on generalized estimation equations (GEE) are only consistent when the probability of missingness does not depend on the response values.

(3) linear mixed models are in general no valid method for the analysis.

D. A major advantage of generalized estimation equations over maximum likelihood methods is that even for misspecification of the correlation structure,

(1) the estimators of the fixed effects and the sandwich estimator of the covariance of the fixed effects estimator are consistent estimators.

(2) the estimators of the fixed effects and of the random effects are consistent estimators.

(3) the estimators of the fixed effects are consistent and efficient estimators.

E. At the end of each month, a job center collects the number of applications sent out this month by long-term unemployed people. Instead of writing applications, some unemployed people went on holiday without permission. They prefer to keep to themselves that they did not send out applications and thus do not fill out the questionnaire of the job center. Which missing mechanism do we have here?

(1) MCAR

(2) MAR

(3) NMAR

Bonus question (Question 6) (5P)

The best linear unbiased predictor (BLUP) can be derived in several ways. Name the two ways which were discussed in the lecture and outline **one** of the derivations. Name all used components.

Solution 6