EXAM COVER SHEET

Module:	**D2, Design and Analysis of Experiments**
Date of exam:	11 June 2019, 1.00 – 3.00pm
Duration:	120 min
Type of exam:	Open book, open internet, no communication with other human beings (i.e. printed material allowed, laptop allowed, wlan allowed, not allowed to communicate with other people)
Module coordinator:	Christoph Kopp (BFH)

Name of the student (readable!): ……
Please write your name also on <u>every</u> page (at the top of the sheet)!

School: ❒ ZHAW ❒ BFH ❒ FHNW ❒ HES-SO

Venue of exam: ……

Exam

Version A

Briefing

- Next to each problem, the number of points is indicated in parentheses, e. g. (3).
- The level of significance is 5% unless otherwise mentioned. Accordingly, confidence bounds and intervals have a confidence level of 95% by default.
- Include a short reasoning (e. g. *"I used a marginal F test and obtained a p value of . . . "*) for your results. This permits giving points for your approach even in the final answer should not be not correct.
- Give numeric results (such as p values) to at least three digits.
- All answers must be copied on this exam, file submissions are not accepted.

Best of luck!

For correction, please do not write into.

Prob. 1	Prob. 2

Problem 1 (16)

The federer.tobacco data frame from the agridat library contains data on the total height (in cm, the sum of 20 tobacco plants per treatment and block) which had been exposed to seven different doses of radiation (dose, in roentgen). The seedlings were transplanted in a randomized complete block design with eight blocks. It is the main aim to model the effect of the radiation on the height, while accounting for potential block effects.

Do turn dose, block and row into factors now, e. g. as follows:

```
> federer.tobacco$dose <- factor(federer.tobacco$dose)
```

a.) Give the R code to fit a suitable parametric model to this data set to answer the research question. (1)

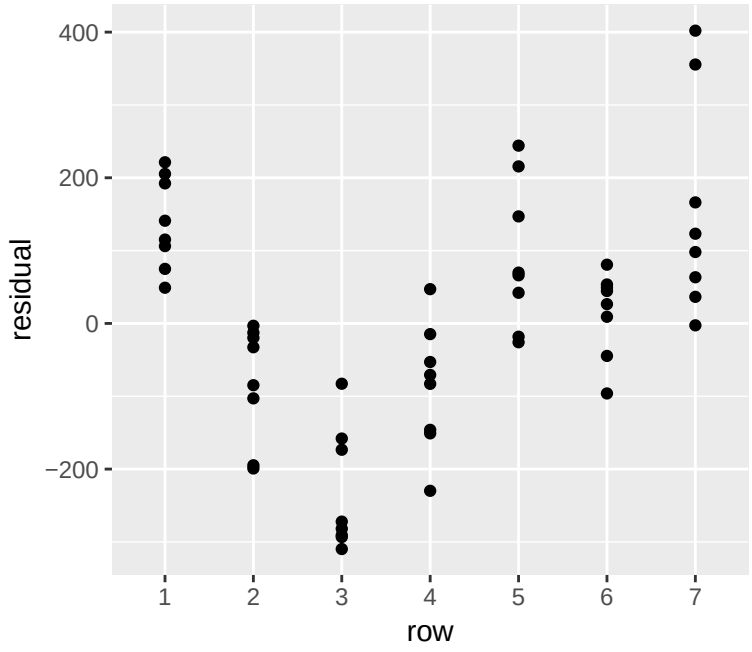
```
> fed.lme <- lme(height ~ dose , random = ~1 | block, data = federer.tobacco)
```

b.) Is the dose effect significant according to the model from a.)? Name the tools used, give the R code to perform the tests, report the p value and use it to answer the question.(2)

```
> anova(fed.lme, type = "marginal")

#             numDF denDF   F-value p-value
# (Intercept)     1    42 265.29233  <.0001
# dose            6    42   1.51004  0.1985
```

A marginal F test showed no significant dose effect ($p = 0.199$).

c.) The layout of the experiment was in eight rows and seven columns. Shortly after the plants were transplanted to the field it became apparent that an environmental gradient existed. The following residual plot for the model in a.) was produced:



Use this residual plot to explain that there is a gradient and discuss what it looks like.(2)

The residuals are observed values − fitted values. Accordingly, the rows 2, 3, and 4 (with negative residuals) seem to produce systematically lower values than the rows 1,5, and 7.

d.) To take care of the mentioned gradient, one can add the fixed effect of the row to the model. Give the R code to fit this model. (1)

```
> fed.grad <- lme(height ~ dose + row, random = ~1 | block,
+                 data = federer.tobacco)
```

e.) According to the model in d.), is the radiation effect significant? And the row effect? Use the same tools as in b.), give your R code, p values and your decision. (1)

```
> anova(fed.grad, type = "marginal")

#             numDF denDF  F-value p-value
# (Intercept)     1    36 487.4003  <.0001
# dose            6    36   2.7141  0.0281
# row             6    36  22.7816  <.0001
```

According to the marginal F tests, the row ($p < 0.001$) and the radiation dose ($p = 0.028$) both have a significant effect on the average total height.

f.) According to the model in d.), by how many cm does the total plant height (of the 20 plants) decrease on average for a "typical" block if the dose was 5'000 roentgen, compared to 0 roentgen? Explain your approach. (2)

```
> fixef(fed.grad)

# (Intercept)     dose250     dose500    dose1000    dose1500
# 1185.530716   23.070073   21.832451    2.666409  -32.124437
#    dose2500    dose5000        row2        row3        row4
#   -6.580593 -128.810507 -243.365185 -400.603993 -251.300986
#        row5        row6        row7
#  -47.647904 -129.270664  -19.254680
```

The notion of "typical" block means that we should ignore the random effects. According to the fixed effects estimates, the average decrease of the total height is equal to 128.8 cm (in row 1, but see the next question).

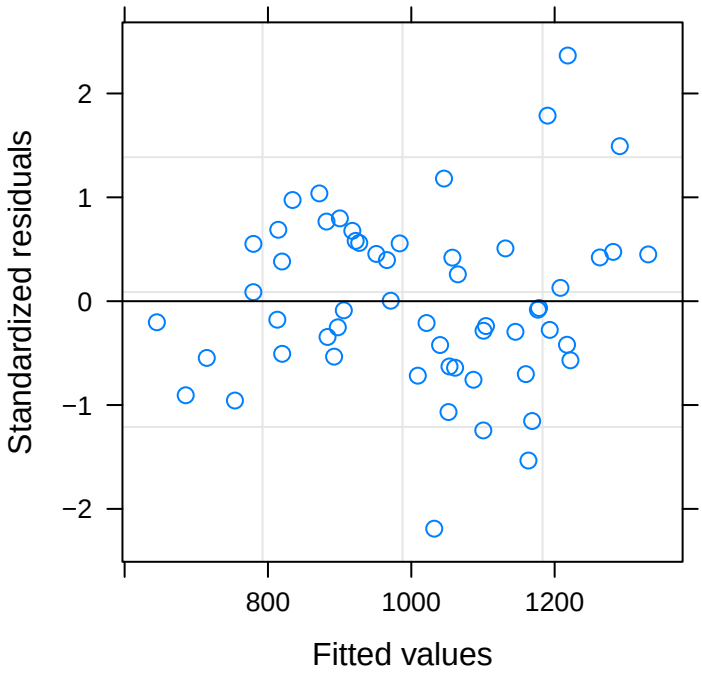
g.) Is the answer to f.) different for different rows or not? Why or why not? (1)

Since the model has no dose × row interaction (the row effect is additive), the dose effect does not depend on the row.

h.) Assess the model assumptions: explain what you assess, with which method, give your R code, discuss the results (give p values where you calculate them) and report your conclusions regarding the assumptions. (2)

We assess residual normality and equal variance:

```
> library(car)
> #qqPlot(resid(fed.grad))
> #shapiro.test(resid(fed.grad))
> plot(fed.grad)
```



Residual normality is unproblematic according to the normal QQ plot and the Shapiro-Wilk test ($p = 0.726$). The residual plot suggests problems with the equal variance assumption: the variance of the residuals seems to increase with the fitted values.

i.) You should have found a problem with the assumptions. What could you do about it? Give a precise answer for full points, not only a rough idea. (2)

It is possible to model heteroskedastic error terms explicitly with lme. We saw how to do this according to the levels of a factor in the lecture notes, so we could try this here as follows:

```
> fed.het <- lme(height ~ dose + row, random = ~1 | block,
+                weights = varIdent(form = ~ dose),
+                data = federer.tobacco)
> plot(fed.het)
```

This approach does not help here, because the variance is not well explained by the dose or the row (as you can see by plotting residuals vs. these variables). An alternative would be to use the fitted values (essentially, the combination of dose and row effect) fo model the heteroskedasticity, but we did not discuss this in the module.

j.) We want to compare each of the dose levels only to its adjacent levels (0 to 250, 250 to 500, ...). Show how to test which of these differences are significant, and give the significant results only, with p values. (2)

```
> library(multcomp)
> summary(glht(fed.grad, mcp(dose = "Sequen")))


#
#   Simultaneous Tests for General Linear Hypotheses
#
# Multiple Comparisons of Means: Sequen Contrasts
#
#
# Fit: lme.formula(fixed = height ~ dose + row, data = federer.tobacco,
#     random = ~1 | block)
#
# Linear Hypotheses:
#                   Estimate Std. Error z value Pr(>|z|)
# 250 - 0 == 0       23.070     44.140   0.523   0.9913
# 500 - 250 == 0     -1.238     47.047  -0.026   1.0000
# 1000 - 500 == 0   -19.166     47.811  -0.401   0.9978
# 1500 - 1000 == 0  -34.791     47.044  -0.740   0.9560
# 2500 - 1500 == 0   25.544     46.761   0.546   0.9892
# 5000 - 2500 == 0 -122.230     44.833  -2.726   0.0357 *
# ---
# Signif. codes:
# 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# (Adjusted p values reported -- single-step method)
```

This is implemented in multcomp's glht. The only significant difference is found between the 2500 and the 5000 roentgen dose (adjusted $p = 0.0356$).

Method Description in glht method
Dunnett: Compare all levels to a reference level
GrandMean: Compare all levels to the overall mean
Sequen:Test a specific sequence
Tukey: Tukey's HSD

Problem 2 (14)

The data set gregory.cotton from the agridat package contains data from an experiment to study the effects of the nitrogen level (levels: N0: None, N1: 600 rotls/feddan [a local unit in Sudan]), the sowing date (levels: D1 up to D4, see help for exact dates), and two other factors (water and spacing) on the yield (in a local unit) of cotton plants. Data were recorded for two years (Y1, Y2). For this problem, we focus only on the effect of the sowing date, the nitrogen and the year on the yield.

a.) According to the sample data from year Y1: which of the four dates should we choose to get the highest average yield if we use the nitrogen fertilizer, and how high is the average yield for this sowing date in year Y1 for the fertilized plots? (1)

```
> gc <- gregory.cotton
> with(gc[gc$year == "Y1", ], tapply(yield, list(date, nitrogen), mean))

#          N0       N1
# D1 1.317778 2.773333
# D2 1.888889 3.074444
# D3 1.797778 2.664444
# D4 1.427778 1.735556
```

We should choose date D2 if we use the nitrogen fertilizer (N1), the average yield is 3.074.

b.) We begin with a model that has no year effect. Give the R code to model the effect of the sowing date, the nitrogen fertilizer and their interaction on the yield. (1)

```
> gc.lm.base <- lm(yield ~ date * nitrogen, data = gc)
```

c.) Does the effect of the nitrogen fertilizer on the yield depend on the date according to the model from b.)? Give your code, and answer the question, argue with p values. (2)

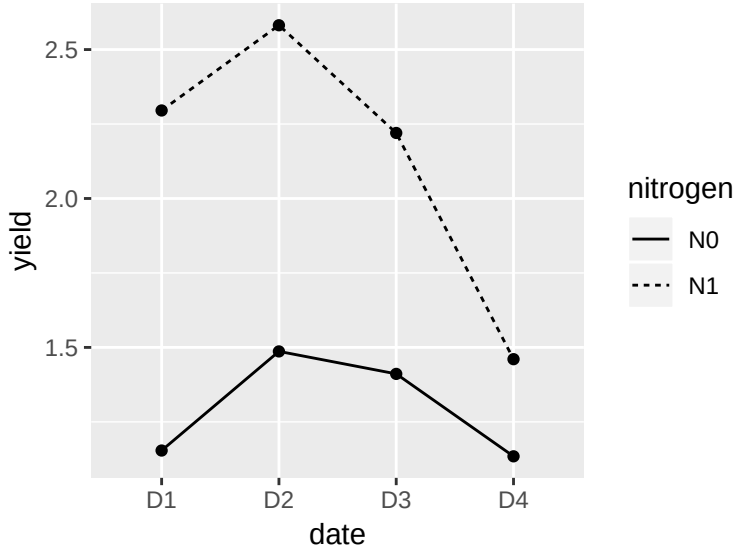
```
> library(car)
> Anova(gc.lm.base, type = 2)

# Anova Table (Type II tests)
#
# Response: yield
#               Sum Sq  Df  F value    Pr(>F)
# date          10.308   3  14.4077 3.324e-08 ***
# nitrogen      25.578   1 107.2497 < 2.2e-16 ***
# date:nitrogen  3.789   3   5.2963  0.001748 **
# Residuals     32.435 136
# ---
# Signif. codes:
# 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the interaction effect of nitrogen and the date is significant ($p = 0.0017$), the nitrogen effect depends on the date according to the model from b.).

d.) How many interaction terms are involved in the interaction effect of this model? (1)

The date has 4 levels and the nitrogen has 2 levels, so their interaction effect involves $(4-1)\cdot(2-1)=3$ terms (as the Anova output above confirms).

e.) According to the interaction plot below, it seems that only a part of the interaction terms may be needed from a statistical point of view. Which one looks the most needed, and why? (2)



The lines for the two nitrogen levels are more or less parallel for dates D1, D2 and D3, but the lines are comparatively much closer for date D4. Therefore, we expect that the interaction of date D4 with nitrogen N1 is the most important term.

f.) Use significance tests for the interaction terms to check your answer to problem e.). Give your code and argue with p values to show that your answer was correct. (2)

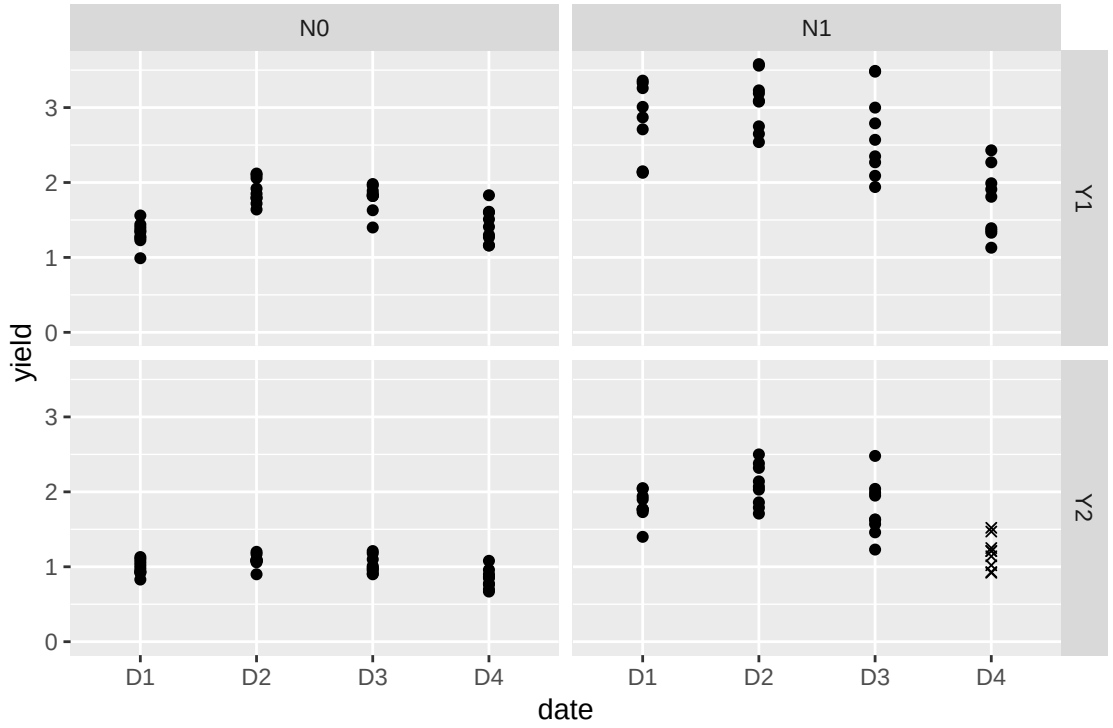
```
> coef(summary(gc.lm.base))

#                   Estimate Std. Error t value  Pr(>|t|)
# (Intercept)        1.15389     0.1151 10.0245 4.757e-18
# dateD2             0.33278     0.1628  2.0443 4.286e-02
# dateD3             0.25722     0.1628  1.5801 1.164e-01
# dateD4            -0.01944     0.1628 -0.1194 9.051e-01
# nitrogenN1         1.14167     0.1628  7.0133 9.935e-11
# dateD2:nitrogenN1 -0.04667     0.2302 -0.2027 8.397e-01
# dateD3:nitrogenN1 -0.33278     0.2302 -1.4455 1.506e-01
# dateD4:nitrogenN1 -0.81556     0.2302 -3.5426 5.434e-04
```

The p values of the model's coefficient estimates confirm that only the dateD4:nitrogenN1 interaction term is significant ($p = 0.0005$).

g.) We now start to investigate the effect of the year. Add the missing yield values for date D4 in year Y2 with nitrogen level N1 to the plot below. (1)

We look up and plot the values

```
> sort(gc$yield[gc$date == "D4" & gc$year == "Y2" & gc$nitrogen == "N1"])

# [1] 0.92 0.93 1.02 1.14 1.21 1.21 1.25 1.47 1.52
```



h.) Add the year to the model. Keep in mind that the year effect could also depend on the levels of the nitrogen and the date. Give your R code to fit a suitable model. (1)

We fit a model that contains the interaction of the year with the date and the nitrogen as well as the three-way interaction:

```
> lm.full <- lm(yield ~ date * nitrogen * year, data = gc)
```

i.) Refer to your model from h.). Which (if any) of the interaction effects with the year is significant? Give your R code and argue with p values. (1)

```
> Anova(lm.full, type = 2)

# Anova Table (Type II tests)
#
# Response: yield
#                      Sum Sq  Df  F value    Pr(>F)
# date                10.3084   3  38.5714 < 2.2e-16 ***
# nitrogen            25.5783   1 287.1234 < 2.2e-16 ***
# year                19.3967   1 217.7330 < 2.2e-16 ***
# date:nitrogen        3.7894   3  14.1790 4.913e-08 ***
# date:year            0.6419   3   2.4018   0.07072 .
# nitrogen:year        0.4433   1   4.9765   0.02743 *
# date:nitrogen:year   0.5503   3   2.0590   0.10892
# Residuals           11.4028 128
# ---
# Signif. codes:
# 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Only the nitrogen:year interaction is significant, this means that the nitrogen effect depends on the year according to the interaction model. (This remains true if we remove the non-significant three-way interaction effect first.)

j.) According to your model in h.) (without removing any nonsignificant effects), how high is the average yield for a plot with N1 nitrogen level, at date D4 in year Y2? Give your code and your answer. (2)

```
> df.pred <- data.frame(date = "D4", nitrogen = "N1", year = "Y2")
> predict(lm.full, df.pred)

#        1
# 1.185556
```

The fitted average yield is 1.1856 (in the local unit). This seems compatible with the data we plotted above. To do this manually, you would simply add the correct coefficients from the interaction model (also the non-significant ones).