EDDA - Assignment 2 - Group 77

Dante de Lang, Ignas Krikštaponis and Kamiel Gülpen

Exercise 1

Moldy bread If left alone bread will become moldy, rot or decay otherwise. To investigate the influence of temperature and humidity on this process, the time to decay was measured for 18 slices of white bread, which were placed in 3 different environments and humidified or not. The data are given in the filebread.txt, with the first column time to decay in hours, the second column the environment (cold, warm or intermediate temperature) and the third column the humidity.

a) The 18 slices came from a single loaf, but were randomized to the 6 combinations of conditions. Present an R-code for this randomization process.

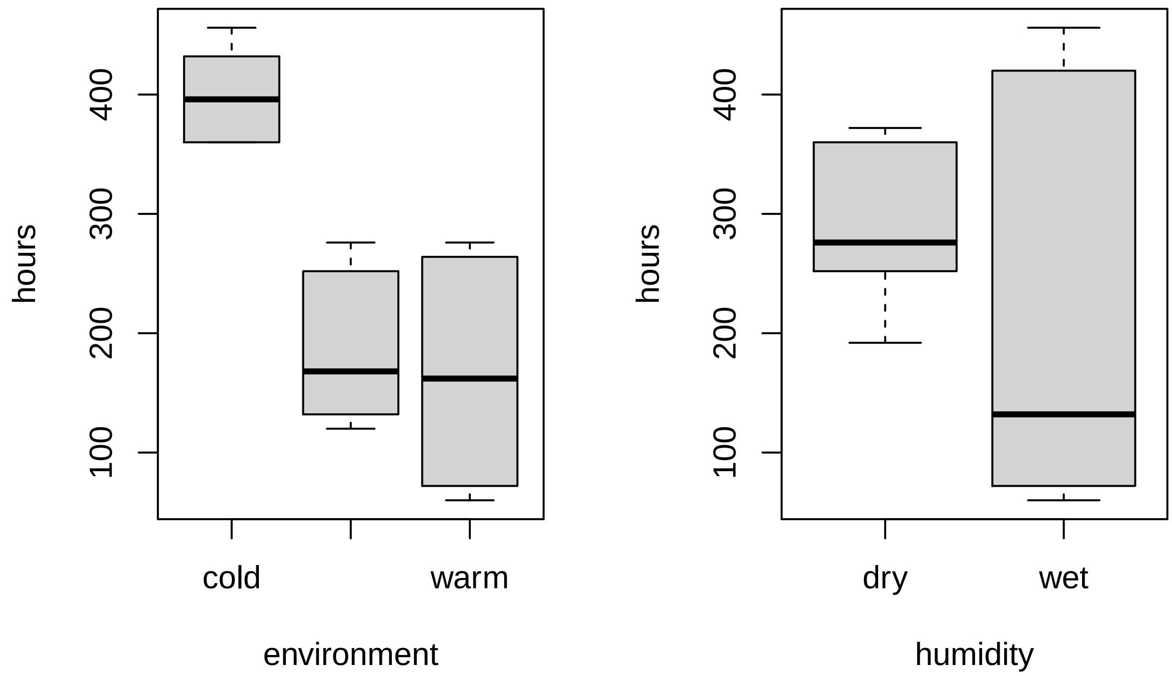
```
data_bread <- read.table(file="data/bread.txt",header=TRUE)
humid <- c("dry","wet")
temp <- c("cold", "intermediate","warm")

combination <- data.frame(cbind(c(humid,humid,humid),c(temp,temp)))
```

b) Make two boxplots of hours versus the two factors and two interaction plots (keeping the two factors fixed in turn).

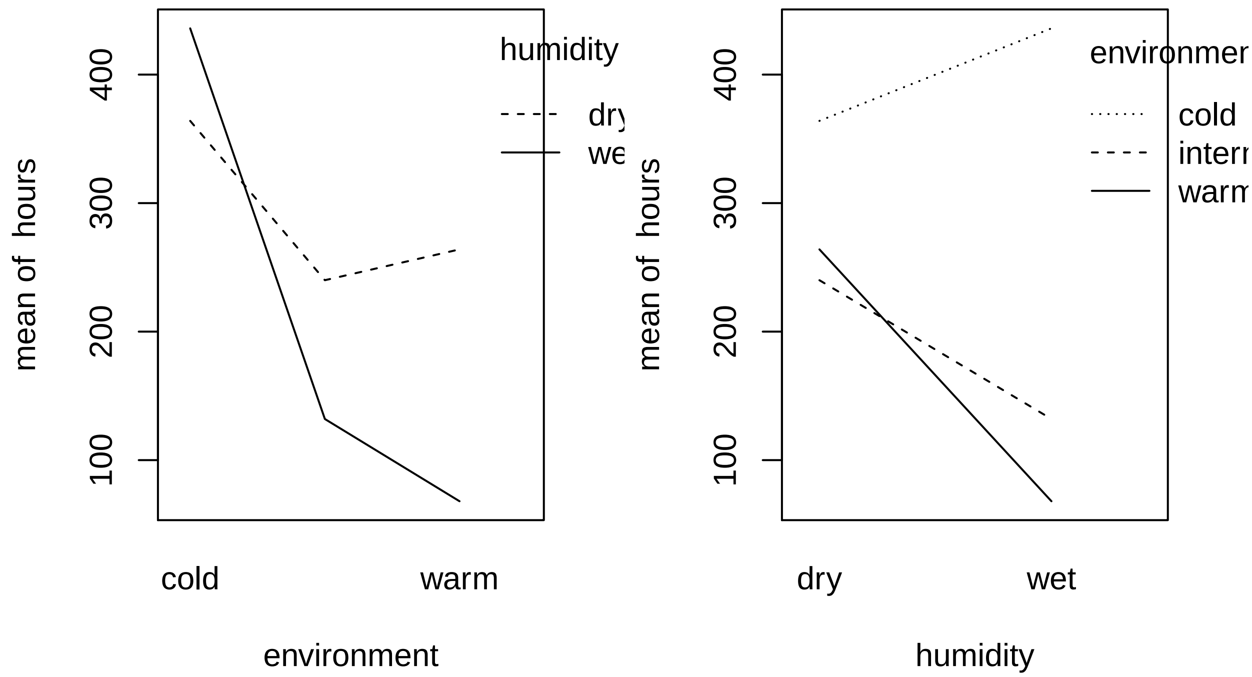
```
data_bread
```

```
##    hours  environment humidity
## 1    360         cold      dry
## 2    360         cold      dry
## 3    372         cold      dry
## 4    420         cold      wet
## 5    456         cold      wet
## 6    432         cold      wet
## 7    192 intermediate      dry
## 8    276 intermediate      dry
## 9    252 intermediate      dry
## 10   132 intermediate      wet
## 11   120 intermediate      wet
## 12   144 intermediate      wet
## 13   252         warm      dry
## 14   276         warm      dry
## 15   264         warm      dry
## 16    60         warm      wet
## 17    72         warm      wet
## 18    72         warm      wet
```

```
attach(data_bread)
par(mfrow=c(1,2))
boxplot(hours~environment)
boxplot(hours~humidity)
```




```
par(mfrow=c(1,2))
interaction.plot(environment,humidity,hours)
interaction.plot(humidity,environment,hours)
```



c) Perform an analysis of variance to test for effect of the factors temperature, humidity, and their interaction. Describe the interaction effect in words.

```
aovbread = lm(hours~environment+humidity)
anova(aovbread)
```

```
## Analysis of Variance Table
##
## Response: hours
##             Df Sum Sq Mean Sq F value  Pr(>F)
## environment  2 201904  100952   23.11 3.7e-05 ***
## humidity     1  26912   26912    6.16   0.026 *
## Residuals   14  61168    4369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aovbread)
```

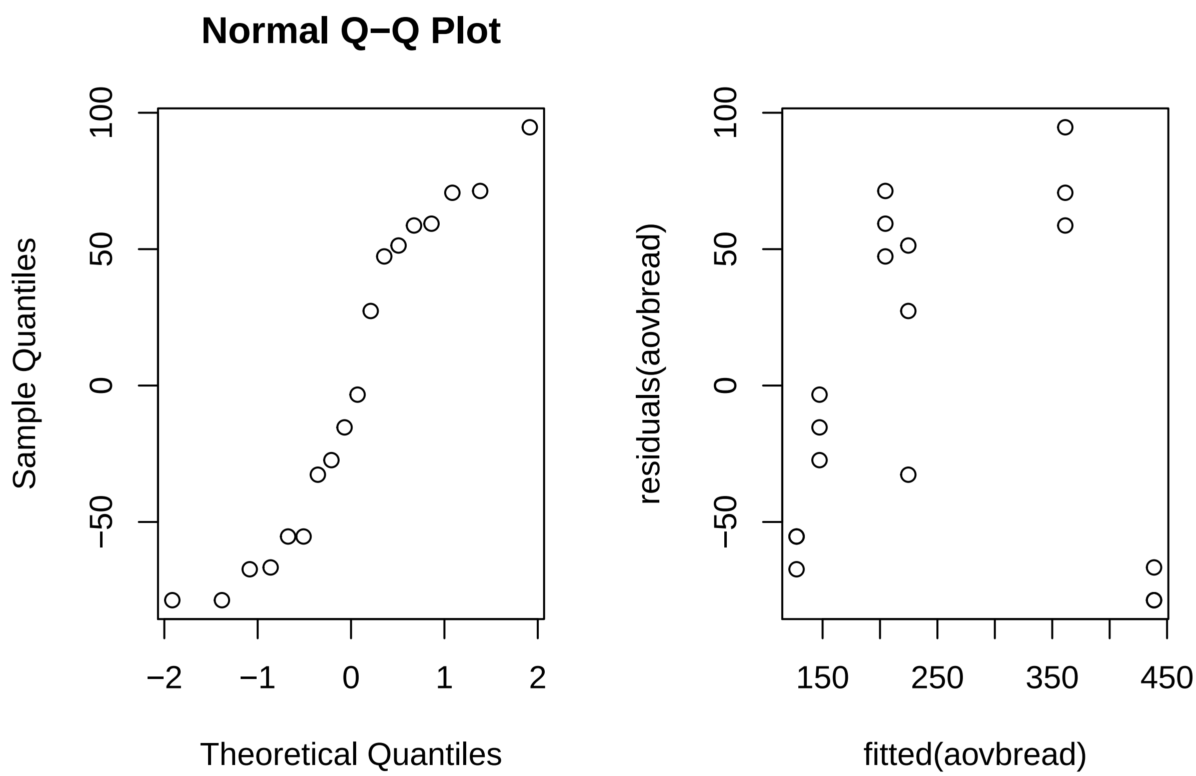
```
##
## Call:
## lm(formula = hours ~ environment + humidity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -78.67 -55.33  -9.33  56.83  94.67
```

```
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                438.7       31.2   14.08  1.2e-09 ***
## environmentintermediate   -214.0       38.2   -5.61  6.5e-05 ***
## environmentwarm           -234.0       38.2   -6.13  2.6e-05 ***
## humiditywet                -77.3       31.2   -2.48    0.026 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.1 on 14 degrees of freedom
## Multiple R-squared:  0.789,  Adjusted R-squared:  0.744
## F-statistic: 17.5 on 3 and 14 DF,  p-value: 5.27e-05
```

d) Which of the two factors has the greatest (numerical) influence on the decay? Is this a good question?

e) Check the model assumptions by using relevant diagnostic tools. Are there any outliers?

```
par(mfrow=c(1,2))
qqnorm(residuals(aovbread))
plot(fitted(aovbread),residuals(aovbread))
```



Exercise 2

A researcher is interested in the time it takes a student to find a certain product on the internet using a search engine. There are three different types of interfaces with the search engine and especially the effect of these interfaces is of importance. There are five different types of students, indicating their level of computer skill (the lower the value of this indicator, the better the computer skill of the corresponding student). Fifteen students are selected; three from each group with a certain level of computer skill. The data is given in the file search.txt. Assume that the experiment was run according to a randomized block design which you make in a). (Beware that the levels of the factors are coded by numbers.)

a) Number the selected students 1 to 15 and show how (by using R) the students could be randomized to the interfaces in a randomized block design.

```
data_search <- read.table(file="data/search.txt",header=TRUE)

interface <- factor(rep(c("1","2","3"),each = 5))
skill <- factor(rep(c("1","2","3","4","5"),times = 3))
students <- c(1:15)
block <- data.frame(students,skill,interface)

block
```

```
##    students skill interface
## 1         1     1         1
## 2         2     2         1
## 3         3     3         1
## 4         4     4         1
## 5         5     5         1
## 6         6     1         2
## 7         7     2         2
## 8         8     3         2
## 9         9     4         2
## 10       10     5         2
## 11       11     1         3
## 12       12     2         3
## 13       13     3         3
## 14       14     4         3
## 15       15     5         3
```

b) Test the null hypothesis that the search time is the same for all interfaces. What type of interface does require the longest search time? For which combination of skill level and type of interface is the search time the shortest? Estimate the time it takes a typical user of skill level 3 to find the product on the website if the website uses interface 3.

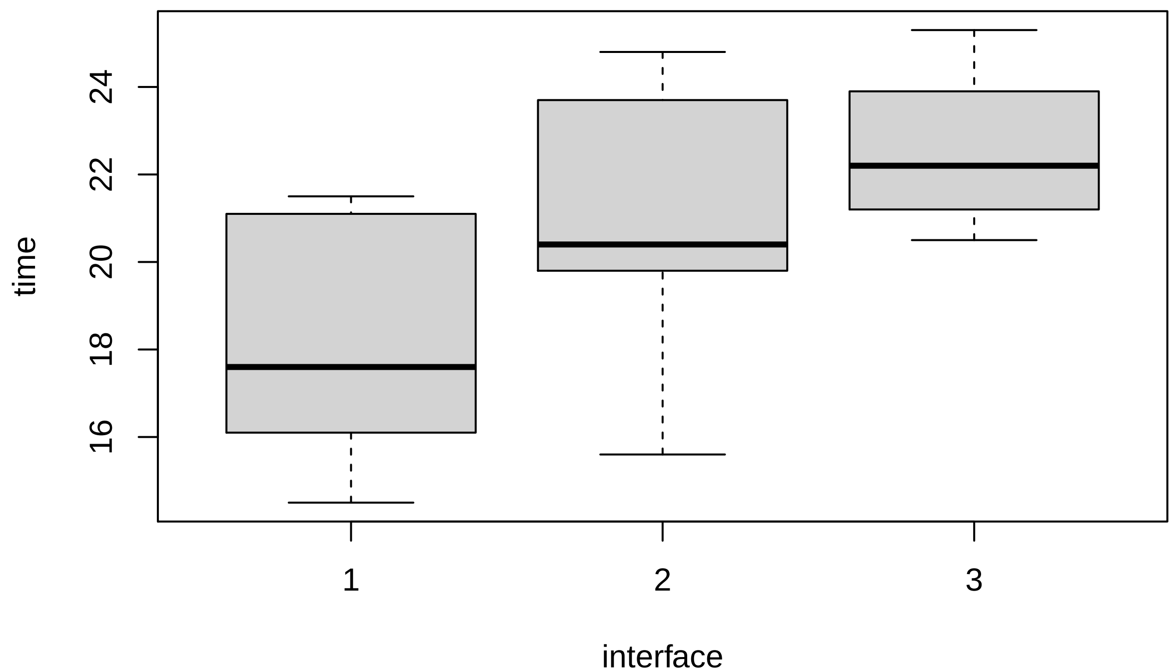
```
attach(data_search)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     interface, skill
```

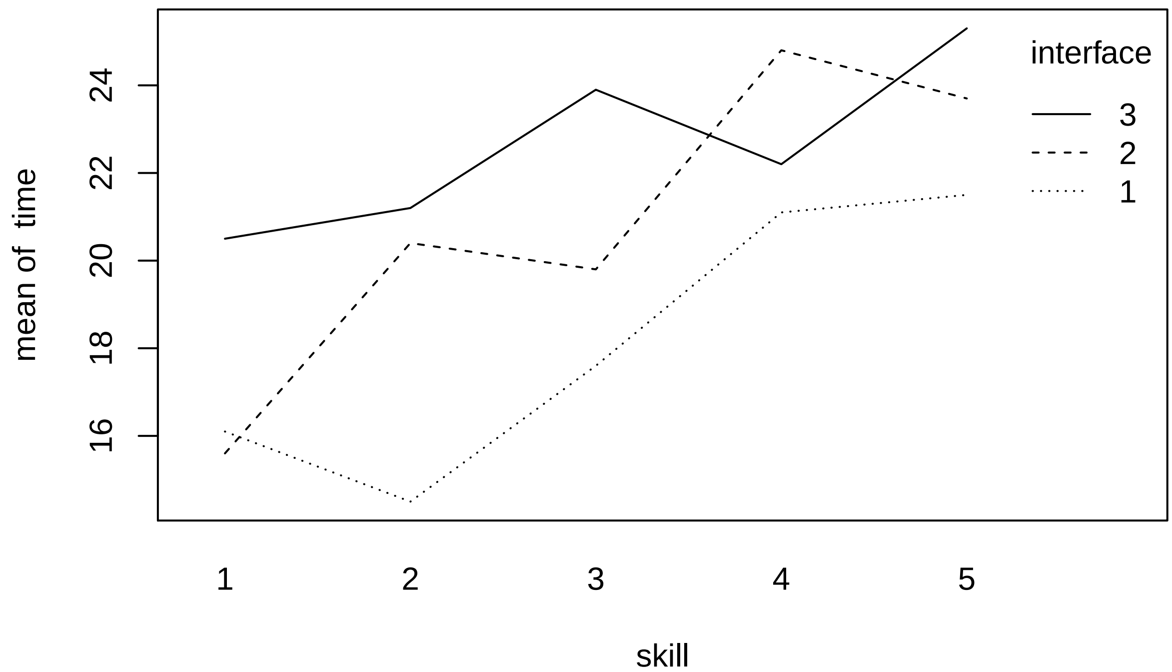
```
aovsearch = lm(time~interface+skill)
anova(aovsearch)
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value Pr(>F)
## interface  2   50.5   25.23    7.82  0.013 *
## skill      4   80.1   20.01    6.21  0.014 *
## Residuals  8   25.8    3.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
boxplot(time~interface) # Interface 3 has the longest search time
```




```
interaction.plot(skill,interface,time) # Skill 2 and interface 1 is the fastest
```




```
summary(aovsearch) # Estimate interface 3 = 4.46, skill 3 = 3.03, so 3-3 gives:
```

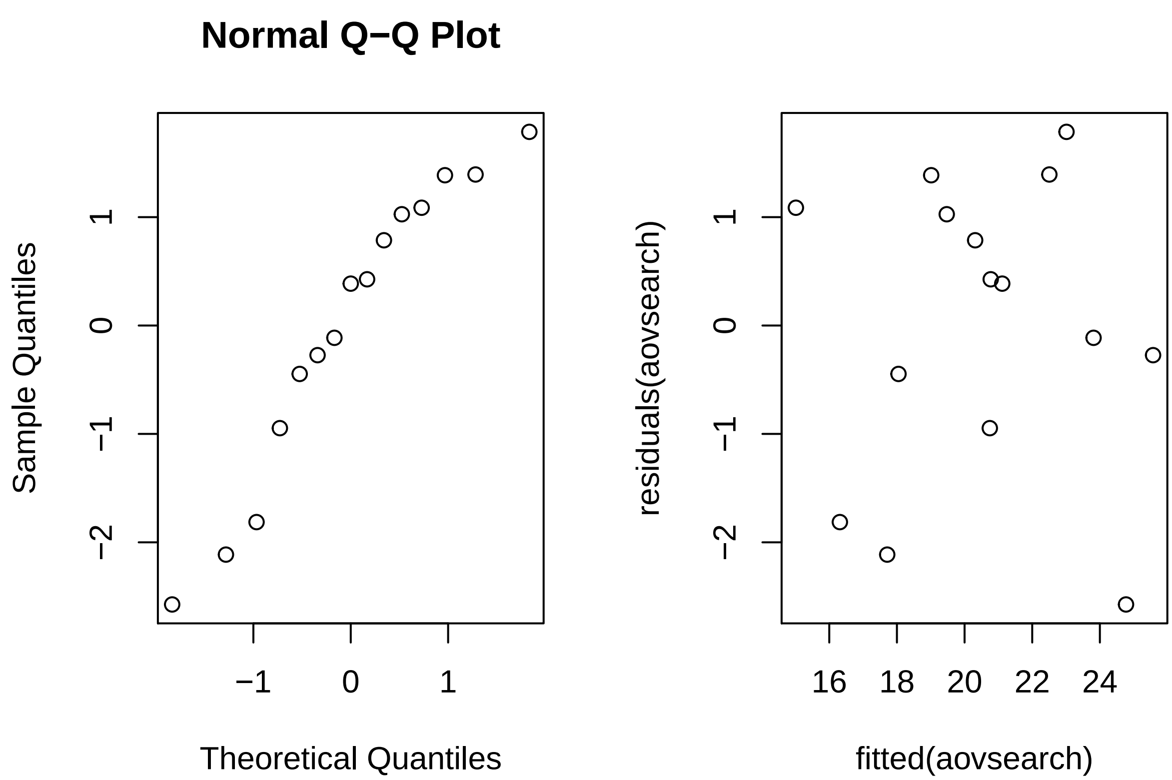
```
##
## Call:
## lm(formula = time ~ interface + skill)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.573 -0.697  0.387  1.057  1.787
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15.01       1.23   12.24  1.8e-06 ***
## interface2      2.70       1.14    2.38   0.0447 *
## interface3      4.46       1.14    3.93   0.0044 **
## skill2          1.30       1.47    0.89   0.4012
## skill3          3.03       1.47    2.07   0.0724 .
## skill4          5.30       1.47    3.61   0.0068 **
## skill5          6.10       1.47    4.16   0.0032 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.8 on 8 degrees of freedom
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.711
## F-statistic: 6.74 on 6 and 8 DF,  p-value: 0.0084
```

```
(4.46+3.03)/2 # 3.75 seconds ????
```

```
## [1] 3.75
```

c) Check the model assumptions by using relevant diagnostic tools.

```
par(mfrow=c(1,2))
qqnorm(residuals(aovsearch))
plot(fitted(aovsearch),residuals(aovsearch))
```



d) Perform the Friedman test tot test whether there is an effect of interface.

```
friedman.test(time,interface,skill)
```

```
##
##  Friedman rank sum test
##
## data:  time, interface and skill
## Friedman chi-squared = 6, df = 2, p-value = 0.04
```

e)Test the null hypothesis that the search time is the same for all interfaces by a one-way ANOVA test, ignoring the variable skill. Is it right/wrong or useful/not useful to perform this test on this dataset?

```
attach(data_search)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     interface, skill

## The following objects are masked from data_search (pos = 3):
##
##     interface, skill, time
```

```
aovsearch = lm(time~interface)
anova(aovsearch)
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value Pr(>F)
## interface  2   50.5   25.23    2.86  0.096 .
## Residuals 12  105.9    8.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 3

In a study on the effect of feedingstuffs on lactation a sample of nine cows were fed with two types of food, and their milk production was measured. All cows were fed both types of food, during two periods, with a neutral period in-between to try and wash out carry-over effects. The order of the types of food was randomized over the cows. The observed data can be found in the file cow.txt, where A and B refer to the types of feedingstuffs.

a) Test whether the type of feedingstuffs influences milk production using an ordinary "fixed effects" model, fitted with lm. Estimate the difference in milk production.

```
data_cow <- read.table(file="data/cow.txt",header=TRUE)
attach(data_cow)

aovcow <- lm(milk~factor(treatment)+factor(order)+id)
anova(aovcow)
```

```
## Analysis of Variance Table
##
## Response: milk
##                   Df Sum Sq Mean Sq F value Pr(>F)
## factor(treatment)  1      0     0.3    0.00   0.97
## factor(order)      1     54    53.5    0.33   0.58
## id                 1    161   161.5    0.98   0.34
## Residuals         14   2295   163.9
```

```
summary(aovcow)
```

```
##
## Call:
## lm(formula = milk ~ factor(treatment) + factor(order) + id)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16.94 -10.72   2.35   9.43  19.72
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          42.972      7.985    5.38  9.7e-05 ***
## factor(treatment)B   -0.244      6.036   -0.04     0.97
## factor(order)BA       6.970     12.147    0.57     0.58
## id                   -2.320      2.338   -0.99     0.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.8 on 14 degrees of freedom
## Multiple R-squared:  0.0857, Adjusted R-squared:  -0.11
## F-statistic: 0.438 on 3 and 14 DF,  p-value: 0.73
```

b) Repeat a) and b) by performing a mixed effects analysis, modelling the cow effect as a random effect (use the function lmer). Compare your results to the results found by using a mixed effects model.

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
attach(data_cow)
```

```
## The following objects are masked from data_cow (pos = 5):
##
##     id, milk, order, per, treatment
```

```
cow_lmer <- lmer(milk~factor(treatment)+factor(order)+(1|id), REML=FALSE)
summary(cow_lmer)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: milk ~ factor(treatment) + factor(order) + (1 | id)
##
##      AIC      BIC   logLik deviance df.resid
##    125.4    129.9    -57.7    115.4       13
##
## Scaled residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1.3726 -0.4871 -0.0255  0.3507  1.6768
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 131.73   11.48
##  Residual               4.75    2.18
## Number of obs: 18, groups:  id, 9
##
## Fixed effects:
##                    Estimate Std. Error t value
## (Intercept)          37.172      5.813    6.39
## factor(treatment)B   -0.244      1.027   -0.24
## factor(order)BA      -3.470      7.768   -0.45
##
## Correlation of Fixed Effects:
##             (Intr) fct()B
## fctr(trtm)B -0.088
## fctr(rdr)BA -0.742  0.000
```

c) Study the commands:

```
attach(data_cow)
```

```
## The following objects are masked from data_cow (pos = 3):
##
##     id, milk, order, per, treatment

## The following objects are masked from data_cow (pos = 6):
##
##     id, milk, order, per, treatment
```

```
t.test(milk[treatment=="A"],milk[treatment=="B"],paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  milk[treatment == "A"] and milk[treatment == "B"]
## t = 0.2, df = 8, p-value = 0.8
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.27  2.76
## sample estimates:
## mean of the differences
##                   0.244
```

Does this produce a valid test for a difference in milk production? Is its conclusion compatible with the one obtained in a)? Why?

Exercise 4

Stochastic models for word counts are used in quantitative studies on literary styles. Statistical analysis of the counts can, for example, be used to solve controversies about true authorships. Another example is the analysis of word frequencies in relation to Jane Austen's novel Sanditon. At the time Austen died, this novel was only partly completed. Austen, however, had made a summary for the remaining part. An admirer of Austen's work finished the novel, imitating Austen's style as much as possible. The file austen.txt contains counts of different words in some of Austen's novels: chapters 1 and 3 of Sense and Sensibility (stored in the Sense column), chapters 1, 2 and 3 of Emma (column Emma), chapters 1 and 6 of Sanditon (both written by Austen herself, column Sand1) and chapters 12 and 24 of Sanditon (both written by the admirer, Sand2).

a) Discuss whether a contingency table test for independence or for homogeneity is most appropriate here.

```
# Not directly? Because the data is not really a contingency table but more just frequency table, there
```

b) Using the given data set, investigate whether Austen herself was consistent in her different novels. Where are the main inconsistencies?

```
data_aus <- read.table(file="data/austen.txt",header=TRUE)
data_aus
```

```
##         Sense Emma Sand1 Sand2
## a         147  186   101    83
## an         25   26    11    29
## this       32   39    15    15
## that       94  105    37    22
## with       59   74    28    43
## without    18   10    10     4
```

```
z <- chisq.test(data_aus[1:3],simulate.p.value=TRUE); z
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  data_aus[1:3]
## X-squared = 12, df = NA, p-value = 0.3
```

```
residuals(z)
```

```
##           Sense   Emma  Sand1
## a       -1.0300 -0.129  1.594
## an       0.4473 -0.159 -0.375
## this     0.0513  0.294 -0.504
## that     0.7482  0.287 -1.442
## with    -0.0475  0.521 -0.704
## without  1.0654 -1.588  0.893
```

```
# Looking at this table we can state that in the Sanditon book she used more "a" and less "that" then i
```

c) Was the admirer successful in imitating Austen's style? Perform a test including all data. If he wasnot successful, where are the differences?

```r
z <- chisq.test(data_aus,simulate.p.value=TRUE); z
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  data_aus
## X-squared = 46, df = NA, p-value = 5e-04
```

```r
residuals(z)
```

```
##          Sense      Emma  Sand1   Sand2
## a       -1.015 -0.112093  1.606 -0.0589
## an      -0.591 -1.219955 -1.067  3.7282
## this     0.139  0.390490 -0.444 -0.3267
## that     1.594  1.179849 -0.910 -3.0493
## with    -0.512  0.000192 -1.025  1.7482
## without  1.392 -1.341196  1.137 -1.0696
```

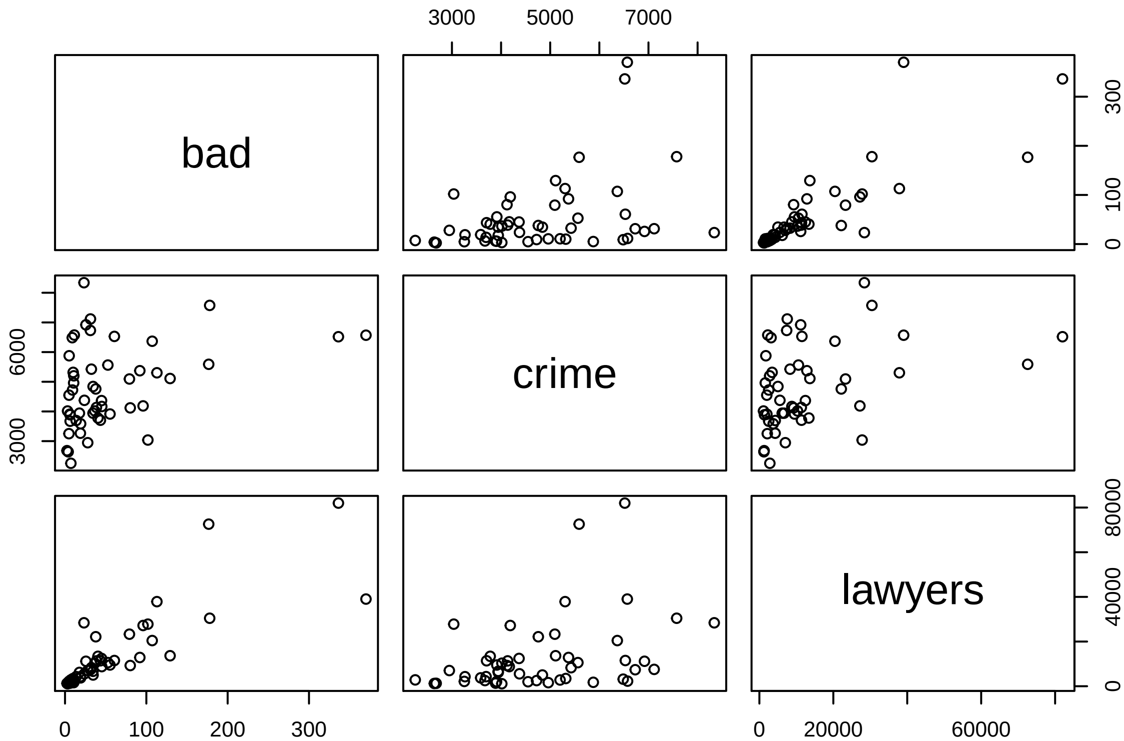
```r
# Looking at this table we can see that the admirer uses considerably more "an" and less "that" then Au
```

Exercise 5

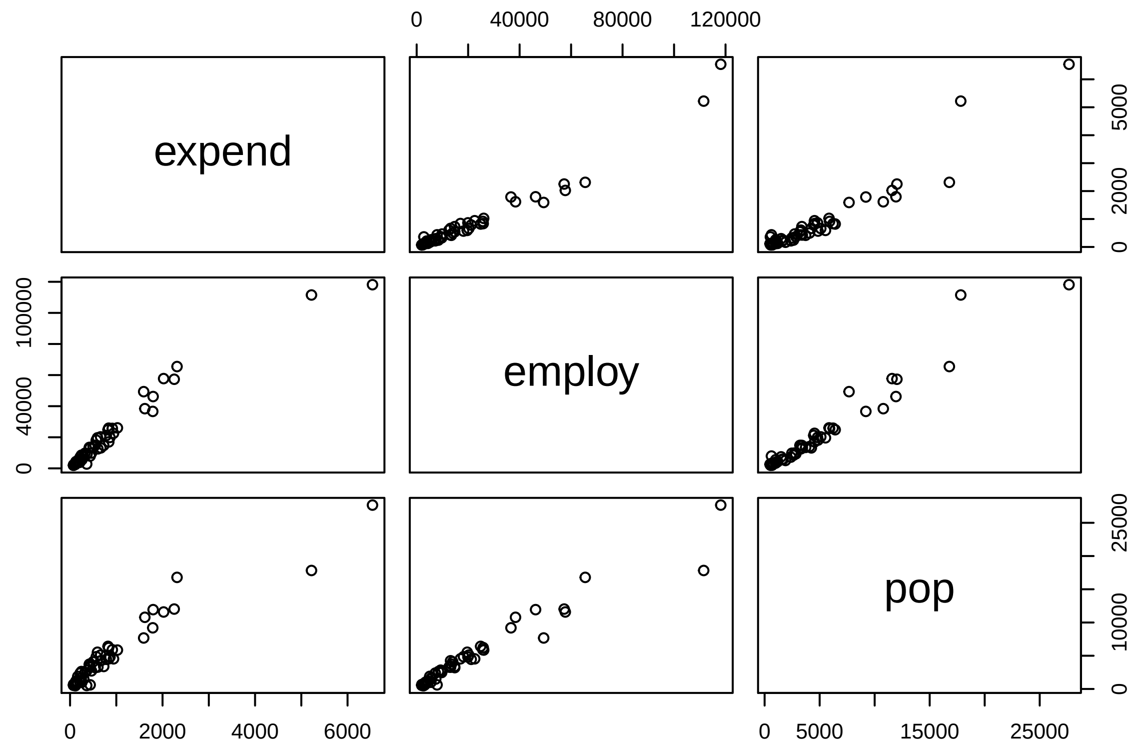
The data in expensescrime.txt were obtained to determine factors related to state expenditures on criminal activities (courts, police, etc.) The variables are: state (indicating the state in the USA), expend (state expenditures on criminal activities in \$1000), bad (crime rate per 100000), crime (number of persons under criminal supervision), lawyers (number of lawyers in the state), employ (number of persons employed in the state) and pop (population of the state in 1000). In the regression analysis, take expend as response variable and bad, crime, lawyers, employ and pop as explanatory variables.

a) Make some graphical summaries of the data. Investigate the problem of potential and influence points, and the problem of collinearity.

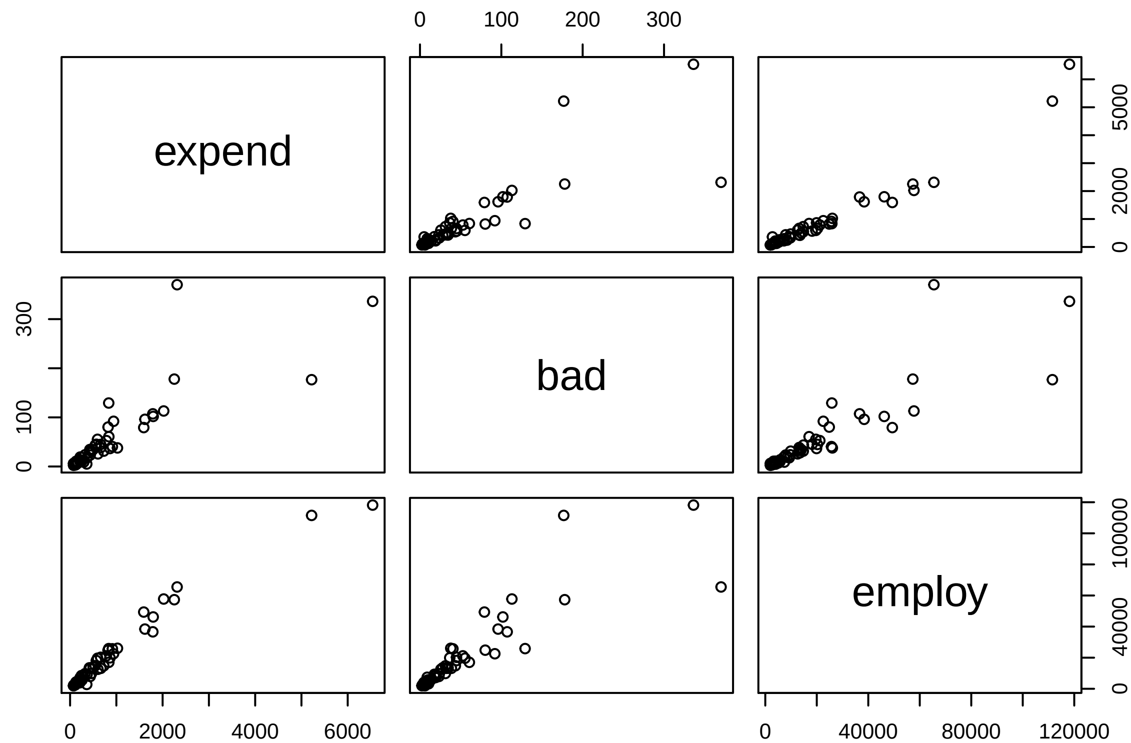
```r
data_crime <- read.table(file="data/expensescrime.txt",header=TRUE)

par(mfrow=c(1,3));plot(data_crime[,c(3,4,5)])
```



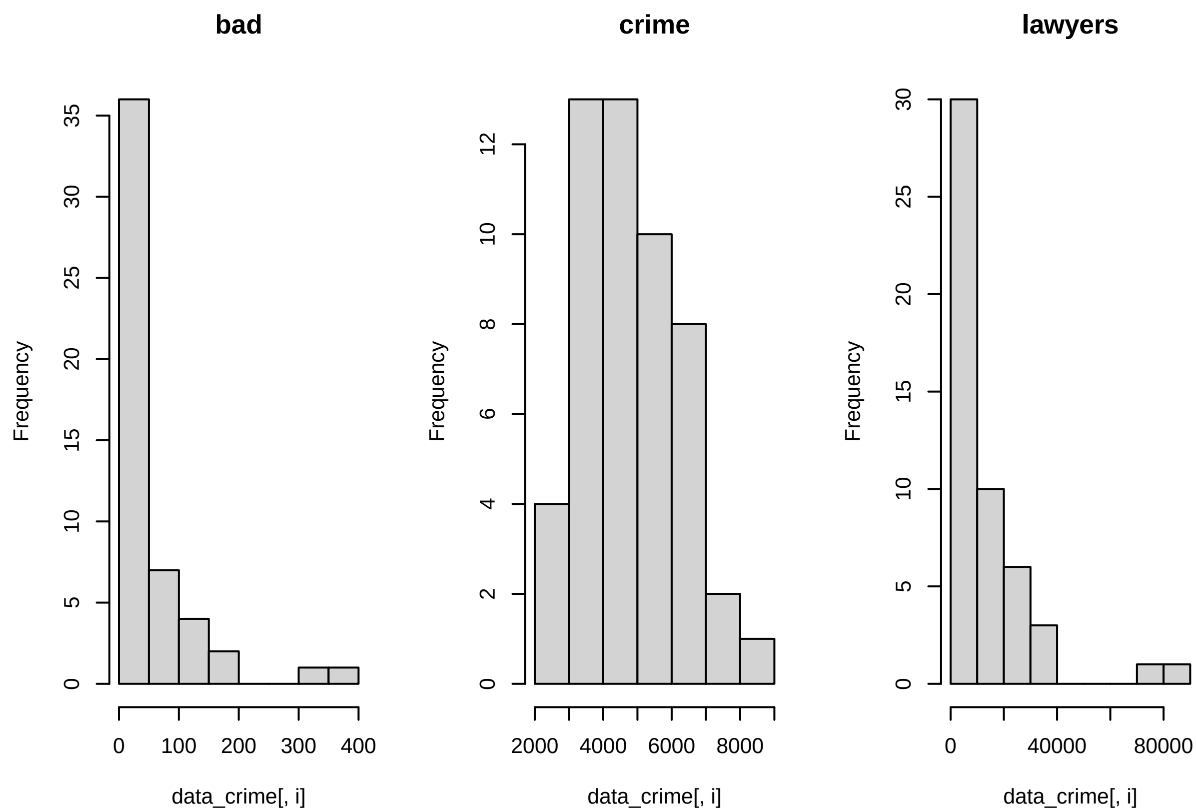

```
par(mfrow=c(1,3));plot(data_crime[,c(2,6,7)])
```




```
par(mfrow=c(1,3));plot(data_crime[,c(2,3,6)])
```




```
par(mfrow=c(1,3));for (i in c(3,4,5)) hist(data_crime[,i],main=names(data_crime)[i])
```




```r
attach(data_crime)
lm_crime <- lm(expend~bad+crime+employ,data=data_crime)
```