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 file bread.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.

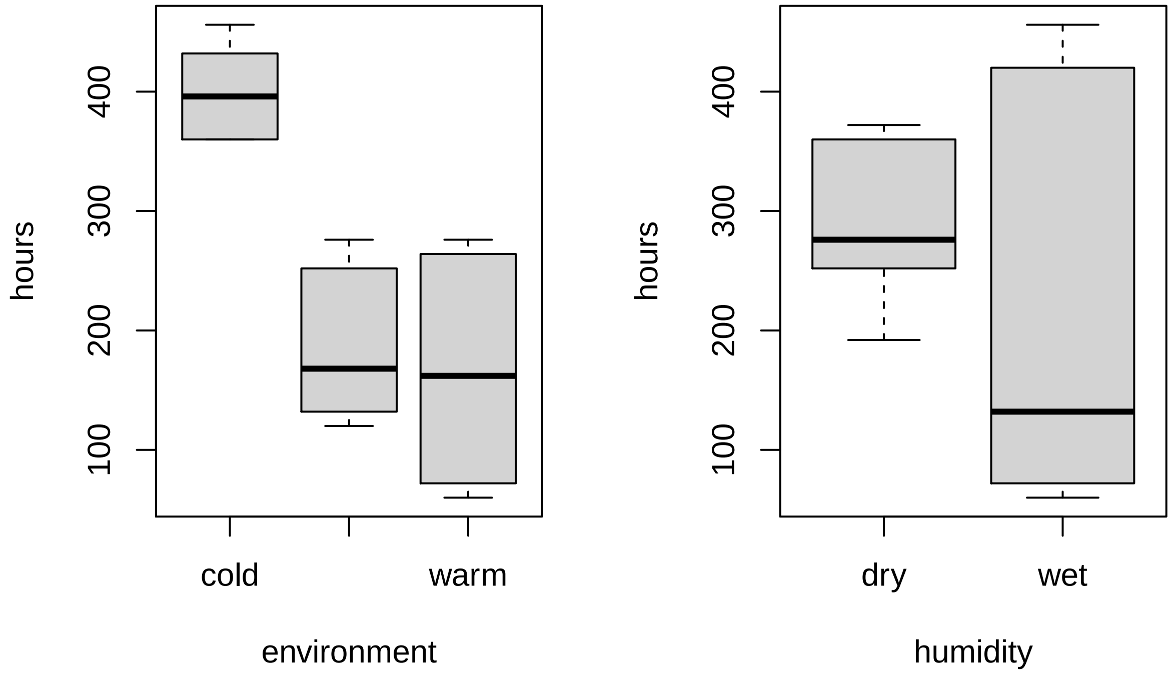
a)The 18 slices came from a single loaf, but were randomized to the 6 combinations of conditions. Present an R-code for this randomization process.

```
data <- read.table(file="data/bread.txt",header=TRUE)
I=3; J=2; N=3

env = rep(c("cold","intermediate","warm"), each=N*J)
hum = rep(c("dry", "wet"),each = N*I)

cbind(env, hum, sample(1:(N*I*J)))
```

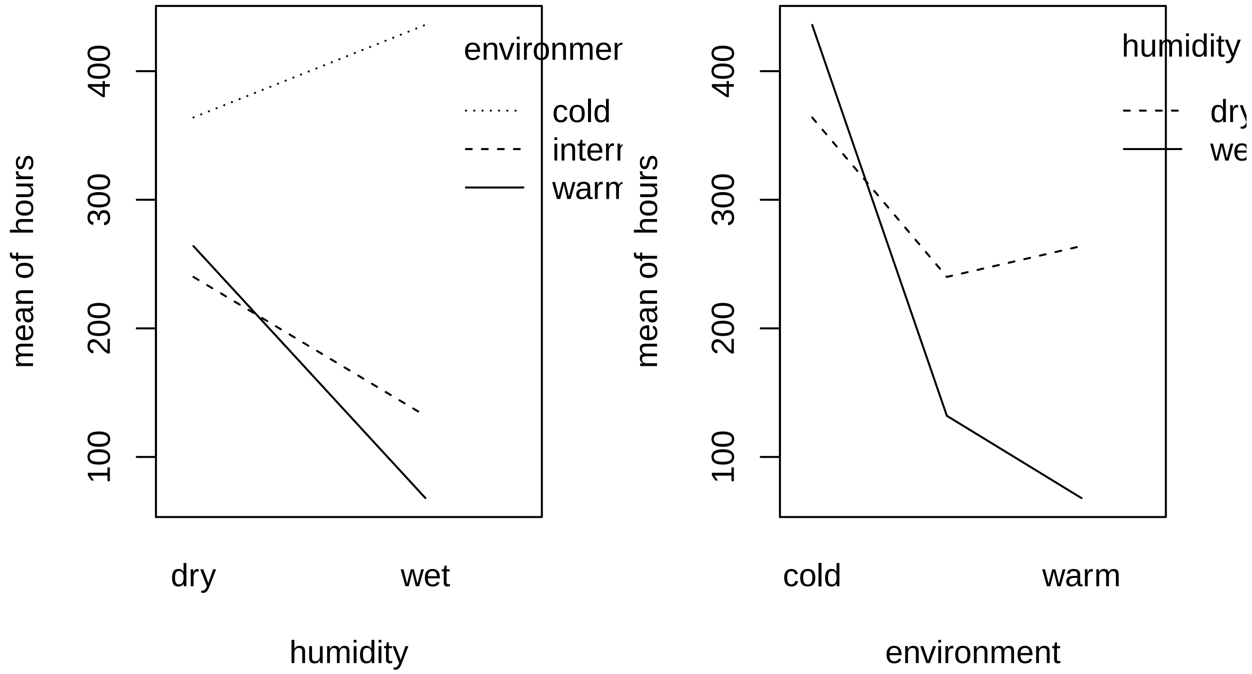
```
##       env            hum
##  [1,] "cold"         "dry" "11"
##  [2,] "cold"         "dry" "12"
##  [3,] "cold"         "dry" "15"
##  [4,] "cold"         "dry" "8"
##  [5,] "cold"         "dry" "9"
##  [6,] "cold"         "dry" "17"
##  [7,] "intermediate" "dry" "1"
##  [8,] "intermediate" "dry" "6"
##  [9,] "intermediate" "dry" "16"
## [10,] "intermediate" "wet" "13"
## [11,] "intermediate" "wet" "5"
## [12,] "intermediate" "wet" "10"
## [13,] "warm"         "wet" "7"
## [14,] "warm"         "wet" "4"
## [15,] "warm"         "wet" "3"
## [16,] "warm"         "wet" "2"
## [17,] "warm"         "wet" "18"
## [18,] "warm"         "wet" "14"
```

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

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



```
interaction.plot(humidity,environment,hours)
interaction.plot(environment,humidity,hours)
```



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

```
data$environment=as.factor(data$environment)
data$humidity=as.factor(data$humidity)
dataaov=lm(hours~humidity*environment)
anova(dataaov)
```

```
## Analysis of Variance Table
##
## Response: hours
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## humidity              1  26912   26912    62.3 4.3e-06 ***
## environment           2 201904  100952   233.7 2.5e-10 ***
## humidity:environment  2  55984   27992    64.8 3.7e-07 ***
## Residuals            12   5184     432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(dataaov)
```

```
##
## Call:
## lm(formula = hours ~ humidity * environment)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
##    -48     -7      0     11     36
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              364         12   30.33  1.0e-12 ***
## humiditywet                               72         17    4.24   0.0011 **
## environmentintermediate                 -124         17   -7.31  9.4e-06 ***
## environmentwarm                         -100         17   -5.89  7.3e-05 ***
## humiditywet:environmentintermediate     -180         24   -7.50  7.2e-06 ***
## humiditywet:environmentwarm             -268         24  -11.17  1.1e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.8 on 12 degrees of freedom
## Multiple R-squared:  0.982,  Adjusted R-squared:  0.975
## F-statistic:  132 on 5 and 12 DF,  p-value: 4.68e-10
```

```
# Without interaction

data$humidity=as.factor(data$humidity)
data$environment=as.factor(data$environment)
dataaov=lm(hours~humidity+environment,data=data)
anova(dataaov)
```

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

When looking at the two-way anova model we see that it consists of the following terms: Y_{ijk} = μ_{ij} + e_{ijk} = $\mu + alpha_i$ + β_j + γ_{ij} + e_{eijk} We decompose the formula it this way such that μ is the overall mean, α_i and β_j are the main effect of level i and j of the first factor and second factor respectively and γ_{ij} the interaction effect.

In order to test the effect of the temperature,humidity, and the interaction we set up 3 hypotheses which are: H_{AB}: $\gamma_{ij} = 0$ for every (i, j) (no interactions between factor A and B)

H_A: $\alpha_i = 0$ for every i (no main effect of factor A)

H_B:$\beta_j = 0$ for every j (no main effect of factor B)

We use the test statistics F_{AB} for H_{AB}, F_A for H_A and F_B for H_B where F is the F-distribution.

To see if the Hypotheses can be rejected we want to look at the probability that P(F>f_{AB}), P(F>f_A) and P(F>f_B), the bigger the F value the lower the probability that the Hypothesis lays under a F-distribution and therefore the Hypothesis can be rejected.

We see that the humidity has a p-value of 4.3e-06, environment a p-value of 2.5e-10 and the interaction between the two (humidity:environment) shows a p-value of 3.7e-07. This means that humidity, environment and the interaction effect between humidity and environment have a significant influence on the hours, which means we can reject H_A, H_B and H_{AB}.

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

When we want to know which factor has the greatest influence we want to use the additive model as used above. This shows a p-value of 0.026 for humidity and a p-value of 3.7e-05 for environemnt. This means that the environment has the greatest influence. It is not

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

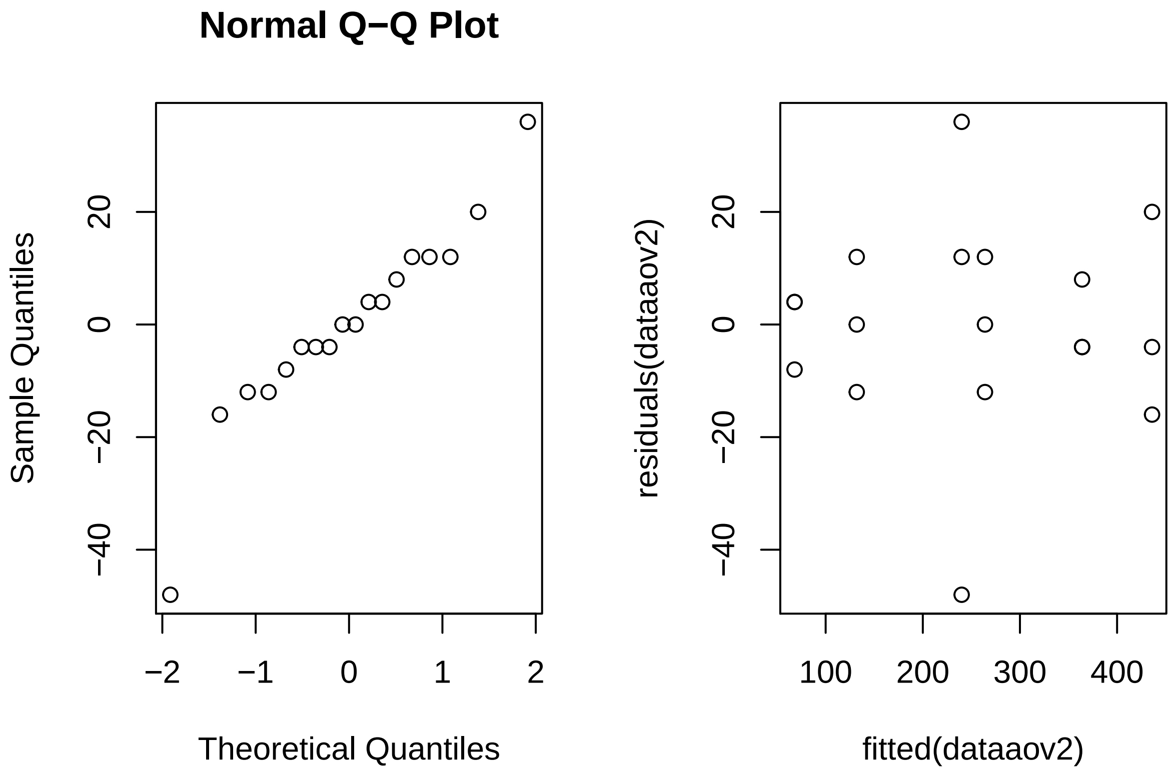
```
par(mfrow=c(1,2))
contrasts(data$humidity)=contr.sum; contrasts(data$environment)=contr.sum
dataaov2=lm(hours~humidity*environment,data=data);
summary(dataaov2)
```

```
##
## Call:
## lm(formula = hours ~ humidity * environment, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##    -48     -7      0     11     36
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)              250.67       4.90   51.17  2.0e-15 ***
## humidity1                 38.67       4.90    7.89  4.3e-06 ***
## environment1             149.33       6.93   21.55  5.8e-11 ***
## environment2             -64.67       6.93   -9.33  7.5e-07 ***
## humidity1:environment1   -74.67       6.93  -10.78  1.6e-07 ***
## humidity1:environment2    15.33       6.93    2.21    0.047 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 20.8 on 12 degrees of freedom
## Multiple R-squared:  0.982,  Adjusted R-squared:  0.975
## F-statistic:  132 on 5 and 12 DF,  p-value: 4.68e-10
```

```
qqnorm(residuals(dataaov2))
shapiro.test(residuals(dataaov2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(dataaov2)
## W = 0.9, p-value = 0.2
```

```
plot(fitted(dataaov2),residuals(dataaov2))
```



The qqplot shows a somewhat linear line which means that based on the qqplot we can state that the data is normally distributed. Furthermore we used a Shapiro-Wilks test to see if the test can back this assumption. The Shapiro-Wils test showed a p-value of 0.2 which means that the residual data is normally distributed. There is also looked at the spread of the residuals, which showed that there are two outliers around 240. # 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.

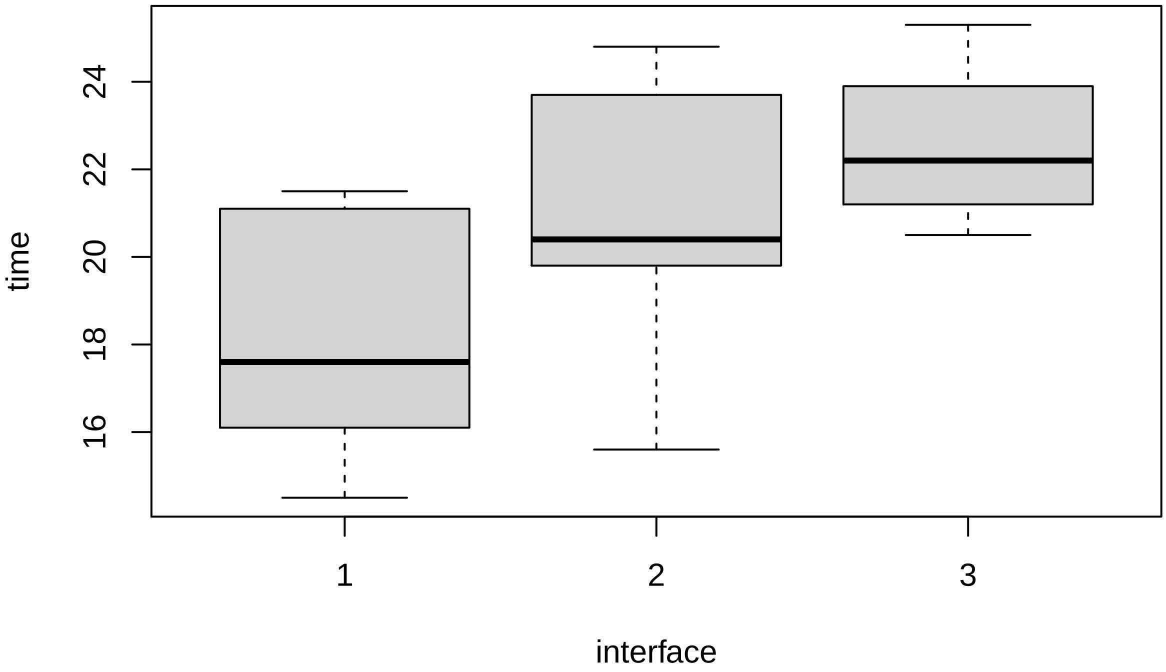
```
attach(data_search)
```

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

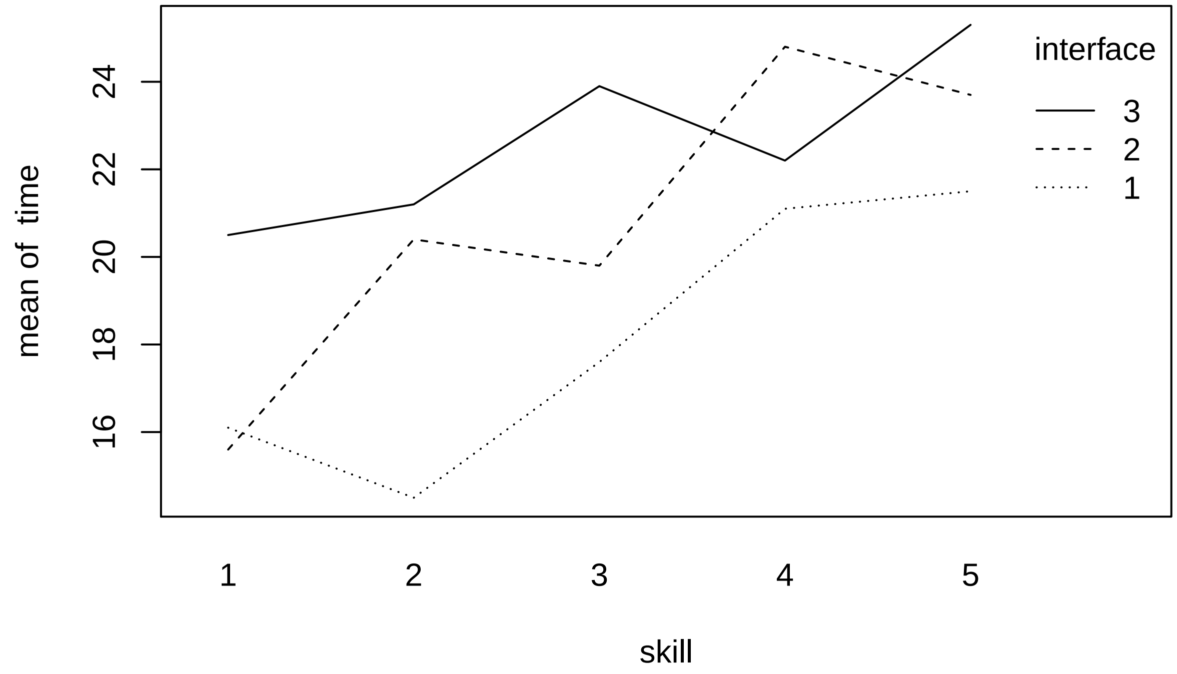
```
aovsearch = lm(time~interface+skill)
anova(aovsearch)
```

```
## 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:
```

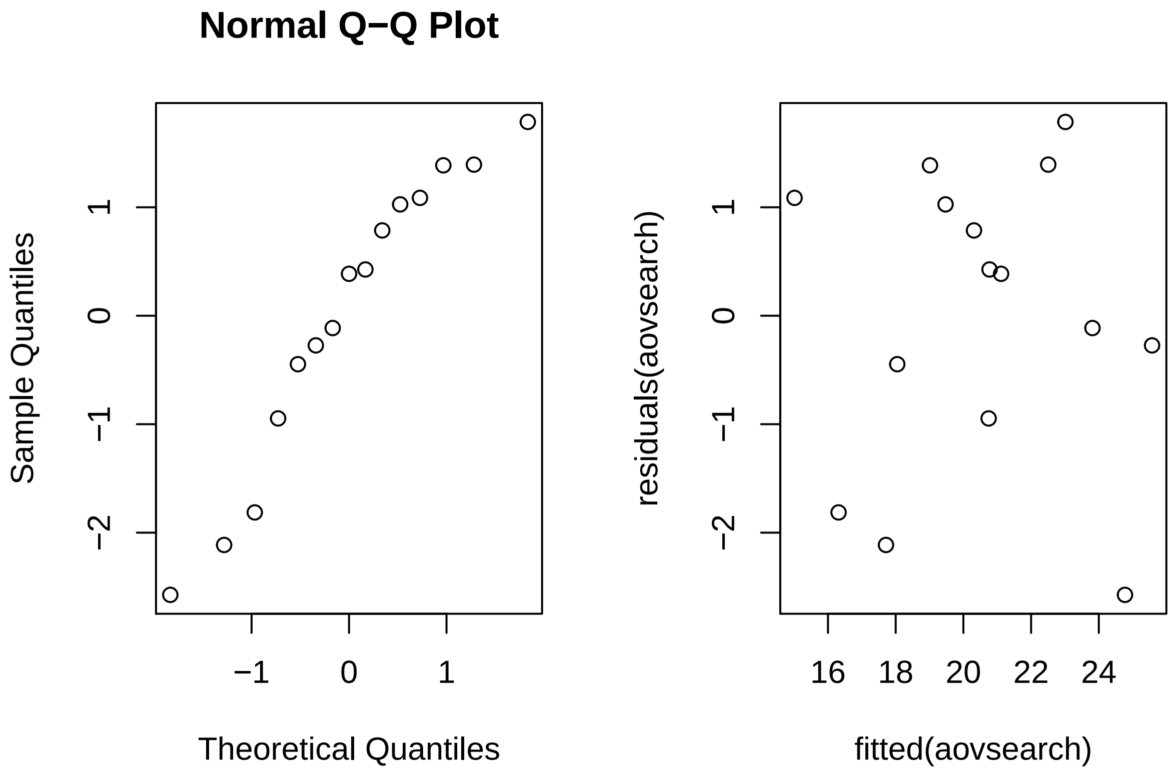
```
##
## Call:
## lm(formula = time ~ interface + skill)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.573 -0.697  0.387  1.057  1.787
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15.01       1.23   12.24  1.8e-06 ***
## interface2      2.70       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 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 author ships. 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.

The contingency table test for homogeneity is appropriate because we want to know if the fan writer imitates Austen in a good way. This means that we want to test whether or not the different columns of data in the table come from the same population (writer) or not. The H0 of the contingency table test for homogeneity states that the distribution of the words is the same for the stories.

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

```
data=read.table(file="data/austen.txt",header=TRUE)
austen = data[,1:3]
z = chisq.test(austen)
z
```

```
##
##  Pearson's Chi-squared test
##
## data:  austen
## X-squared = 12, df = 10, 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
```

She is not inconsistent as the p-value is above 0.05. This means that we cannot reject the H0. She does however have some main inconsistency, which where the words “a”, “that” and “without”. As can be seen in the residual tanle above.

```
z = chisq.test(data)
z
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 46, df = 15, p-value = 6e-05
```

```
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
```

The fan is inconsistent as the p-value of the test is below 0.05. Therefore we have to reject the H0 and accept that the distribution of the words in the stories are not the same. Because Austen herself did not have this inconsistency we can say that the inconsistency is caused by the fan writer. The main inconsistencies were for the words “that” and “an”. As can be seen in the residual tanle above.