Assignment 2

Andrei Udriste, Xinyu Hu, Maria Gherghina-Tudor - Group 43

2021/3/8

Exercise 1. ***Moldy bread***

```
library("car")
```

```
## Loading required package: carData
```

```
bread = read.table(file="bread.txt", header=TRUE, sep=" ")
attach(bread)
```

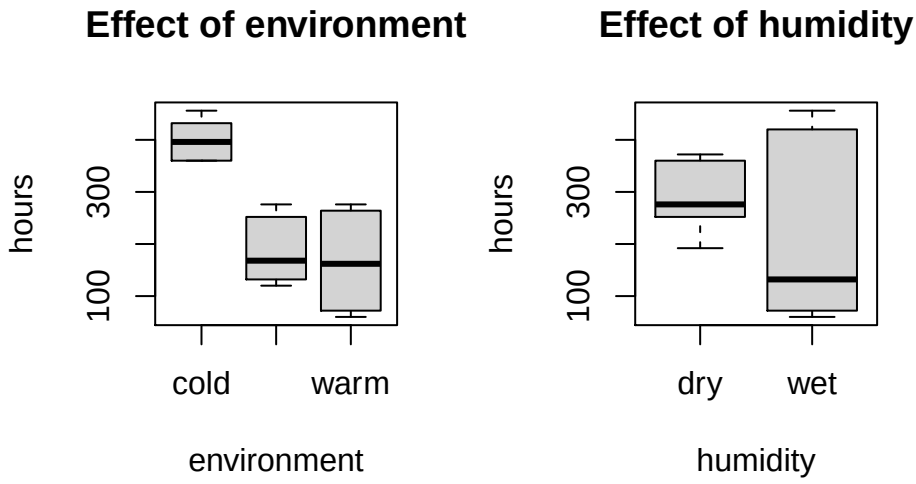
```
env=3; hum=2; N=3
rbind(rep(1:env,each=N*hum),rep(1:hum,N*env),sample(1:(N*env*hum)))
```

a)

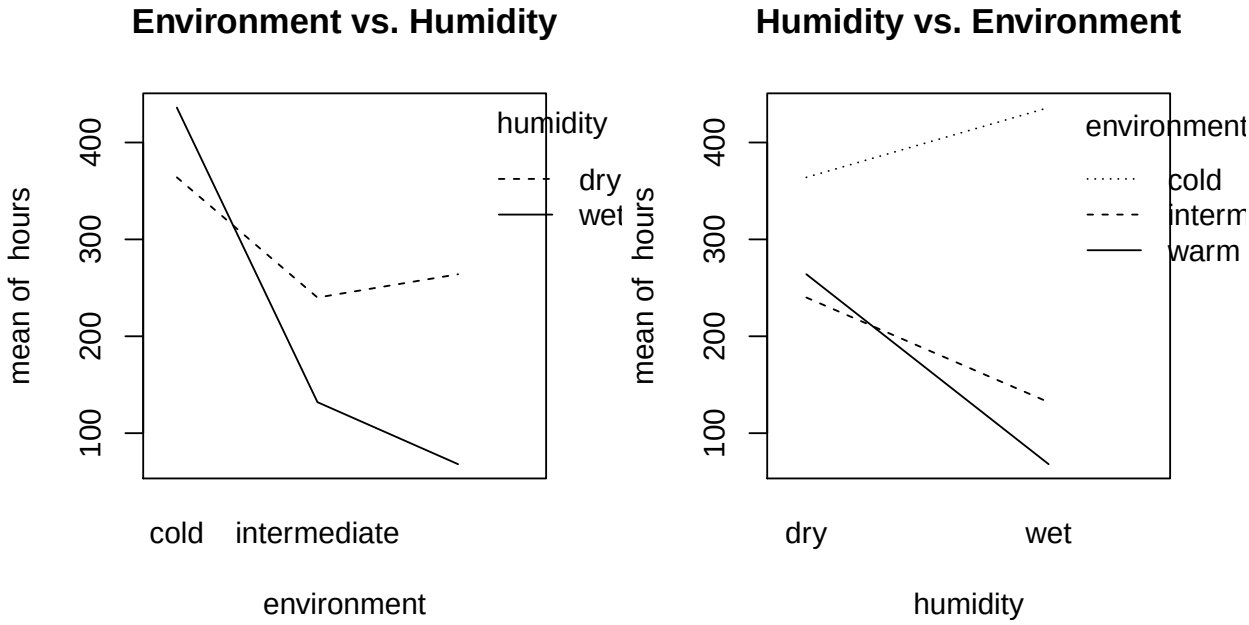
```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## [1,]    1    1    1    1    1    1    2    2    2     2     2     2     3     3
## [2,]    1    2    1    2    1    2    1    2    1     2     1     2     1     2
## [3,]    1    2    4   16   17    9   18   12   15     6     8    11     7     3
##      [,15] [,16] [,17] [,18]
## [1,]     3     3     3     3
## [2,]     1     2     1     2
## [3,]    10    14    13     5
```

Interpreting the table resulting from the randomization process:

- each column can be seen as a different unit (where the ID of the unit is the value in the third row)
- the first row can be seen as the environment that each unit has to be measured in
- the second row can be seen as the humidity group of the unit.



b)




```
aovhoursenv=lm(hours~environment*humidity)
anova_hours = anova(aovhoursenv)
anova_hours
```

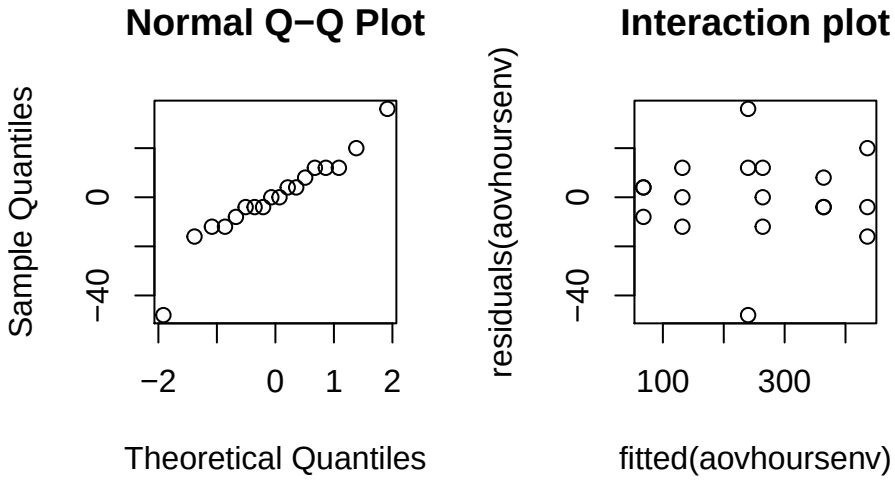
```
## Analysis of Variance Table
##
## Response: hours
##                      Df Sum Sq Mean Sq F value    Pr(>F)
## environment           2 201904  100952 233.685 2.461e-10 ***
## humidity              1  26912   26912  62.296 4.316e-06 ***
## environment:humidity  2  55984   27992  64.796 3.705e-07 ***
## Residuals            12   5184     432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_hours)
```

```
##        Df            Sum Sq          Mean Sq          F value
##  Min.   : 1.00   Min.   :  5184   Min.   :   432   Min.   : 62.30
```

```
##  1st Qu.: 1.75   1st Qu.: 21480   1st Qu.: 20292   1st Qu.: 63.55
##  Median : 2.00   Median : 41448   Median : 27452   Median : 64.80
##  Mean   : 4.25   Mean   : 72496   Mean   : 39072   Mean   :120.26
##  3rd Qu.: 4.50   3rd Qu.: 92464   3rd Qu.: 46232   3rd Qu.:149.24
##  Max.   :12.00   Max.   :201904   Max.   :100952   Max.   :233.69
##                                                    NA's   :1
##      Pr(>F)
##  Min.   :0.0e+00
##  1st Qu.:2.0e-07
##  Median :4.0e-07
##  Mean   :1.6e-06
##  3rd Qu.:2.3e-06
##  Max.   :4.3e-06
##  NA's   :1
```

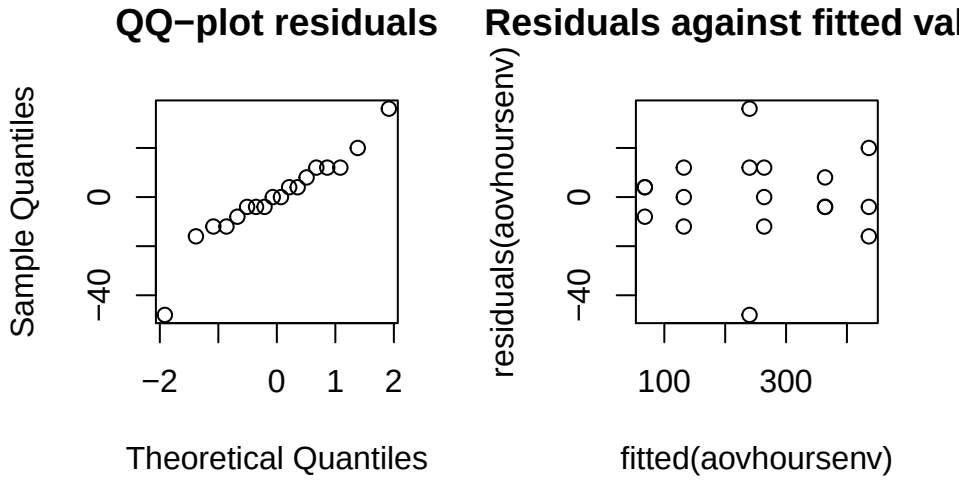
When performing a one way ANOVA test to analysis the effect of the factors on the decay and their interaction, it can be concluded that both factors have a significant effect on the time of the decay: F=131.9, p-value: 4.676e-10



When analysing the graphical representation the interaction (the interaction plot), the interaction of Humidity vs. Environment shows a clear effect: it can be concluded that for both intermediate and warm environments, the mean of the hours of decay is decreasing when humidity is increasing (from a dry to a wet environment). For the cold environment, humidity has an opposite effect - an increase in the duration of the time to decay once humidity increases (from a dry to a wet environment). Furthermore, the spread in the residuals seems to be bigger for middle-fitted values, while a few data points seem extreme, requiring further outliers investigation.

d) Considering the summary of the data discussed in the previous question, it is clear that the interaction of the 2 factors has a significant effect on the time of the decay.

Therefore, when discussing the effects of the main factors, the interaction effect has to also be taken into account, thus the question is not exactly correct.



e)

```
shapiro.test(residuals(aovhoursenv))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(aovhoursenv)
## W = 0.9296, p-value = 0.1911
```

Shapiro-Wilk normality test results lead to a p-value $= 0.1911$, which means the normality assumption is not rejected: there is no reason to suspect that the differences are not resulting from a normal distribution.

Exercise 2. Search engine

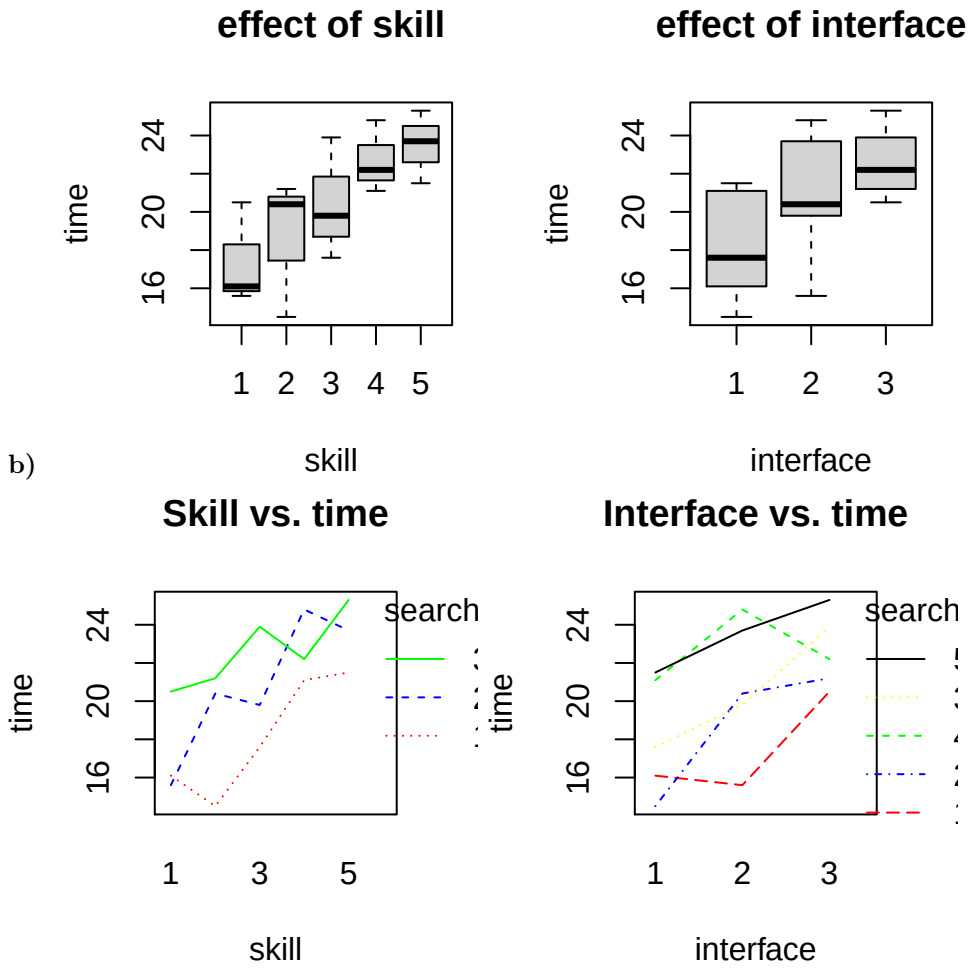
```
I=3;B=5;N=1
for(i in 1:B) print(sample(1:(N*I)))
```

a)

```
## [1] 3 1 2
## [1] 1 2 3
## [1] 2 3 1
## [1] 3 2 1
## [1] 3 2 1
```

To create a randomized block design we can use the function sample from R, which selects a random variable from a given dataset. I represents the number of interfaces, 3. B represents the number of levels of computational skills, 5.

b)



The first step in analyzing the provided data is to create box plots for the skill level of the users and the used interfaces and see what are the differences in those boxplots. If we observe the box plots we can observe that the skill level impacts the search time for the users. The same can be said about the interfaces, where the first interfaces provides a lower search time than the other two interfaces.

```
factor_skills = factor(search_data$skill)
factor_interface = factor(search_data$interface)
search_aov = lm(time~factor_interface + factor_skills, data = search_data)
anova(search_aov)
```

```
## Analysis of Variance Table
##
## Response: time
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## factor_interface  2 50.465 25.2327  7.8237 0.01310 *
## factor_skills     4 80.051 20.0127  6.2052 0.01421 *
## Residuals         8 25.801  3.2252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After performing the ANOVA test the p-value 0.0142 was obtained, this means that the null hypotheses is rejected, so the search time is different for particular interfaces.

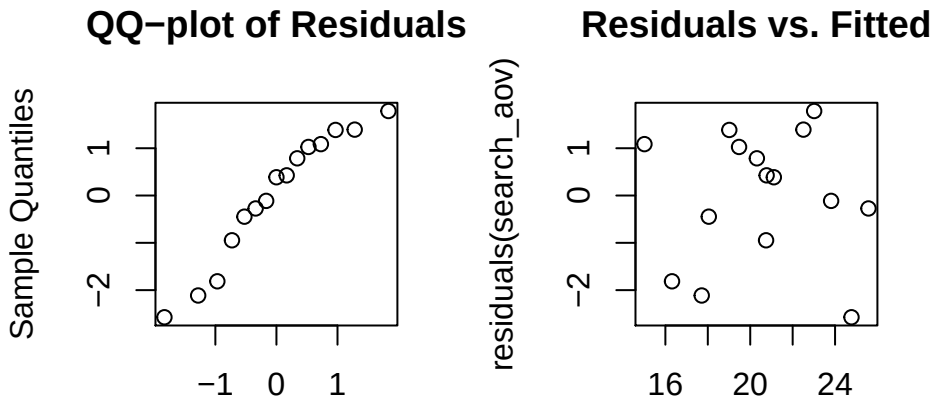
```
summary(search_aov)
```

```
##
## Call:
## lm(formula = time ~ factor_interface + factor_skills, data = search_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5733 -0.6967  0.3867  1.0567  1.7867
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         15.013      1.227  12.238 1.85e-06 ***
## factor_interface2    2.700      1.136   2.377  0.04474 *
## factor_interface3    4.460      1.136   3.927  0.00438 **
## factor_skills2       1.300      1.466   0.887  0.40118
## factor_skills3       3.033      1.466   2.069  0.07238 .
## factor_skills4       5.300      1.466   3.614  0.00684 **
## factor_skills5       6.100      1.466   4.160  0.00316 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.796 on 8 degrees of freedom
## Multiple R-squared:  0.8349, Adjusted R-squared:  0.7111
## F-statistic: 6.745 on 6 and 8 DF,  p-value: 0.008395
```

```
cat("The smallest time is equal with:", mean(search_data$time[search_data$interface == 1]) + mean(searcl
```

```
## The smallest time is equal with: 18.92
```

The interface that has the highest individual and mean search time is number 3, having a standard deviation 4.460 bigger than the normal standard deviation(15.013). The shortest time would be for the first interface(1) and and participant with the highest computational level(1) which is equal with 18.92. If we want to find the time a level 3 user would spend on a level 3 interface we need to compute $mu + alfa_3 + Beta_2$ which is equal with $15.013 + 4.460 + 3.033 = 22.506$.



c)

The two graphs look ok. The QQ-plot look almost normal, the only problem that would make us doubt the normality of the data is the bump present in the upper region, but is not that significant. The fitted plot looks great, the fitted value have a good spread and there doesn't seem to be any pattern.

Another method to check normality would be to use Shapiro-Wilk normality test.

```
shapiro.test(residuals(search_aov))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(search_aov)
## W = 0.93092, p-value = 0.2817
```

After running the test we obtain a p-value = 0.2817, thins means that we do not reject the null hypothesis H0 so the distribution of the data could be normal.

```
friedman.test(search_data$time, search_data$interface, search_data$skill, data=search_data)
```

d)

```
##
##  Friedman rank sum test
##
## data:  search_data$time, search_data$interface and search_data$skill
## Friedman chi-squared = 6.4, df = 2, p-value = 0.04076
```

After computing the Friedman test we obtain a p-value of 0.04076, this means that we will reject the null hypothesis H0, so there is a an effect in the interface that is used.

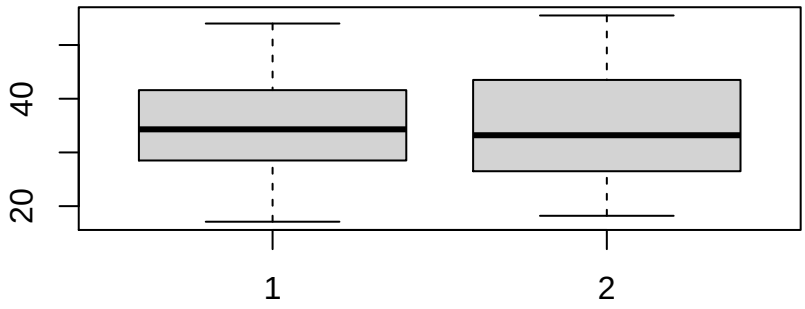
```
search_one_aov = lm(time~interface, data=search_data)
print(anova(search_one_aov), signif.stars=F)
```

e)

```
## Analysis of Variance Table
##
## Response: time
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## interface  1  49.729  49.729  6.0652 0.02852
## Residuals 13 106.588   8.199
```

After performing the one-way-ANOVA test we obtain a p-value of 0.02852, this means that we will reject the null hypothesis, so the interface has an effect on the time a user spends searching. But using a one-way-ANOVA test here is wrong because we only take into consideration the interface and we ignore the skill level of each participant which also has an influence on the time spend searching.

Exercise 3. Feedingstuffs for cows



a)

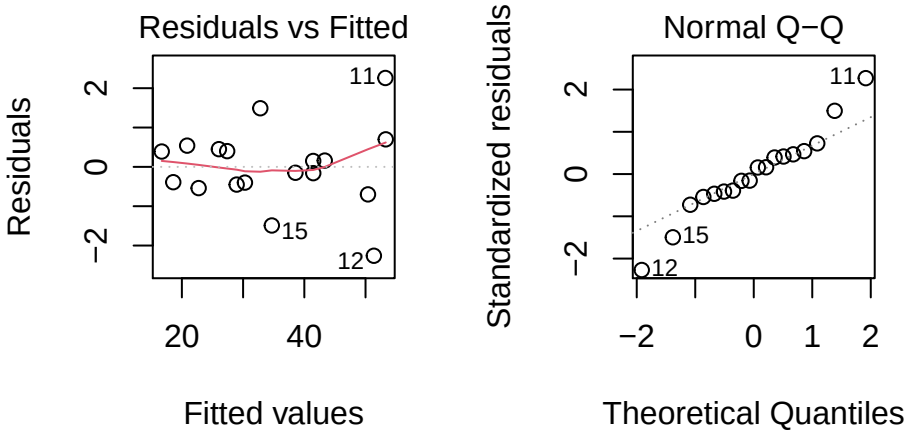
In the boxplot above we can observe the two milk quantities obtained after using each treatment. It can be observed that there is almost no difference between the two milk quantities.

```
cowlm = lm(milk ~ id + per + treatment, data = cow)
anova(cowlm)
```

```
## Analysis of Variance Table
##
## Response: milk
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## id         8 2467.47 308.434 124.4832 7.494e-07 ***
## per        1   24.50  24.500   9.8881   0.01628 *
## treatment  1    1.16   1.156   0.4666   0.51654
## Residuals  7   17.34   2.478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To draw a better conclusion between the two treatments we can use the ANOVA test, which will give us a p-value of 0.51654. This means that we will accept the null hypothesis H0, so there is no difference between the two treatments.

To make sure that the ANOVA test is effective we should verify the QQ-plot and the dispersion of the residual to see if the distribution of the data is normal or not.



The QQ-plot looks normal, and the residuals appear to be evenly distributed so we can conclude that de data has a normal distribution.

```
library(lme4)
```

b)

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

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
cowlmer = lmer(milk ~ per + treatment + (1|id), data = cow, REML = FALSE)
cowlmer1 = lmer(milk ~ per + (1|id), data = cow, REML = FALSE)
anova(cowlmer1, cowlmer)
```

```
## Data: cow
```

```
## Models:
## cowlmer1: milk ~ per + (1 | id)
## cowlmer: milk ~ per + treatment + (1 | id)
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## cowlmer1    4 116.09 119.65 -54.045   108.09
## cowlmer     5 117.51 121.96 -53.755   107.51 0.5807  1      0.446
```

Another way to test the difference between the two treatments is to create two models, one with the treatment and one without the treatment and observe the difference between those two treatments. After computing the difference between the two models we obtain a p-value of 0.446, this means that we accept the null hypothesis H0, so there is no significant difference between the two treatments so we can conclude that the treatment doesn't have a big influence over the model.

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

c)

```
##
##  Paired t-test
##
## data:  milk[treatment == "A"] and milk[treatment == "B"]
## t = 0.22437, df = 8, p-value = 0.8281
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.267910  2.756799
## sample estimates:
## mean of the differences
##               0.2444444
```

The last way we can try to test the difference between the two treatments is to perform a t-test on the two milk samples after using each treatment. After computing the t-test we obtain a p-value of 0.8281 this means that we accept the null hypothesis H0, so there is no significant difference between the two treatments. This means that all three test obtain the same conclusion, that the null hypothesis is accepted, but in the case of the t-test the p-value was considerably bigger than in the case of the other tests. But there is a clear problem with using the t-test in this case, mainly that we only look at the treatment factor and we discard any other factor that might influence the final result.

Exercise 4. Jane Austen

```
austen=read.table("austen.txt",header=TRUE)
austendata=data.frame(austen$Sense, austen$Emma, austen$Sand1)
attach(austendata)
```

a) For a statistical analysis of the word frequencies in relation to Jane Austen's novel Sanditon, the amount of specific words used in her previous works is compared to the ones used by the admirer who tried to replicate her style. For such comparison, it is needed to determine if the frequency counts for specific words are equally distributed for both writers. Therefore, a contingency table test for homogeneity is the most suitable.

b) To analyze Austen's consistency throughout her different novels, a Chi-squared test can be used. This is performed on the columns containing data based on her writings (Sense, Emma, Sand1):

```
consistency=chisq.test(austendata)
consistency
```

```
##
##  Pearson's Chi-squared test
##
## data:  austendata
## X-squared = 12.271, df = 10, p-value = 0.2673
```

Pearson's Chi-squared test: X-squared = 12.271, df = 10, p-value = 0.2673

Since the p-value is high (0.26), the null hypothesis is to be rejected, therefore confirming a relationship between the variables. As expected, it can be concluded that there is a consistency throughout Austen's novels, with no significant difference throughout the word counts of the different books.

The main inconsistencies can be found by investigating the residuals:

```
residuals(consistency)
```

```
##      austen.Sense austen.Emma austen.Sand1
## [1,]  -1.02997736  -0.1290203    1.5937736
## [2,]   0.44728806  -0.1590968   -0.3746273
## [3,]   0.05133600   0.2938669   -0.5036577
## [4,]   0.74817619   0.2865778   -1.4423521
## [5,]  -0.04747379   0.5205063   -0.7035205
## [6,]   1.06544255  -1.5884103    0.8926239
```

```
admirer=chisq.test(austen)
admirer
```

c)

```
##
##  Pearson's Chi-squared test
##
## data:  austen
## X-squared = 45.578, df = 15, p-value = 6.205e-05
```

Pearson's Chi-squared test: X-squared = 45.578, df= 15, p-value = 6.205e-05. Since the p-value is very high (6.205e-05), it can be concluded that the null hypothesis is rejected, thus confirming a similarity between the writers' style. The main differences between the writers can be found once again by inspecting the residuals:

```
residuals(chisq.test(austen))
```

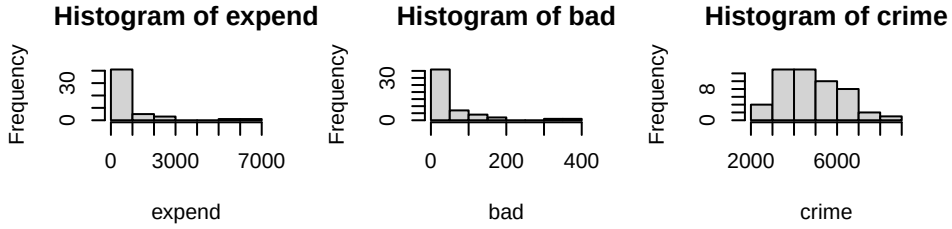
```
##              Sense          Emma      Sand1       Sand2
## a       -1.0149156 -0.1120927868  1.6062866 -0.05889921
## an      -0.5906319 -1.2199545912 -1.0671306  3.72816398
## this     0.1388299  0.3904903154 -0.4436450 -0.32671736
## that     1.5943613  1.1798488360 -0.9099606 -3.04931581
## with    -0.5120944  0.0001916718 -1.0246069  1.74821745
## without  1.3919336 -1.3411962838  1.1365432 -1.06963011
```

The main differences are found for the count of the following words:

- "an" and "with" - both words have been used significantly more by the admirer, compared to Austen
- "that" - has been used significantly less by the admirer, compared to Austen

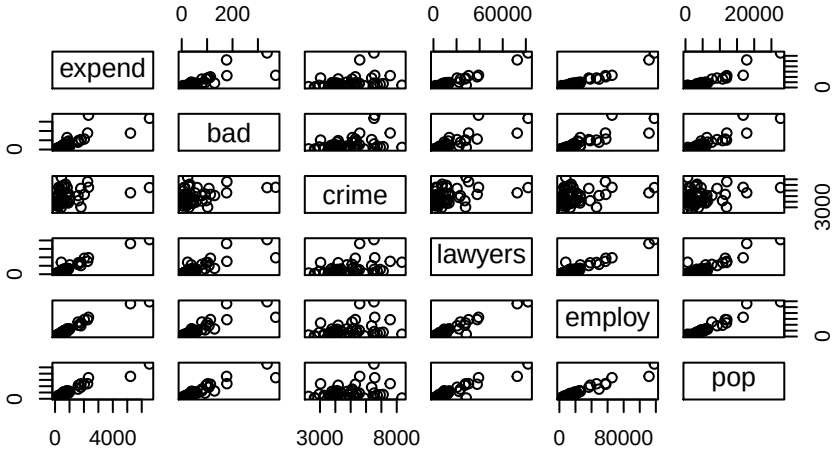
Therefore, it can be concluded that the admirer was quite successful in imitating Austen's writing style, with small differences for a few words.

Exercise 5. Expenditure on criminal activities



a)

In case of finding the potential and influence points, the histograms are shown above. It is clear that the crime factor is normally distributed, and the rest of factors have the similar curve. It shows that there exists collinearity.




```
##         expend  bad crime lawyers employ  pop
## expend    1.00 0.83  0.33    0.97   0.98 0.95
## bad       0.83 1.00  0.37    0.83   0.87 0.92
## crime     0.33 0.37  1.00    0.38   0.31 0.28
## lawyers   0.97 0.83  0.38    1.00   0.97 0.93
## employ    0.98 0.87  0.31    0.97   1.00 0.97
## pop       0.95 0.92  0.28    0.93   0.97 1.00
```

In the graph, (expend, crime), (bad, crime), (crime, lawyers), (crime, employ), (crime, pop) are not linear independently. And we need to see which predictor variables are involved in collinearity.

```
exlm1 = lm(expend~bad+crime+lawyers+employ+pop, data=ex);vif(exlm1)
```

```
##       bad     crime   lawyers    employ       pop
##  8.364321  1.487978 16.967470 33.591361 32.937517
```

```
exlm2 = lm(expend~crime+lawyers+employ+pop, data=ex); vif(exlm2)
```

```
##     crime   lawyers    employ       pop
```

```
##  1.233263 16.372292 33.106158 17.576977
```

```
exlm3 = lm(expend~crime+employ+pop, data=ex); vif(exlm3)
```

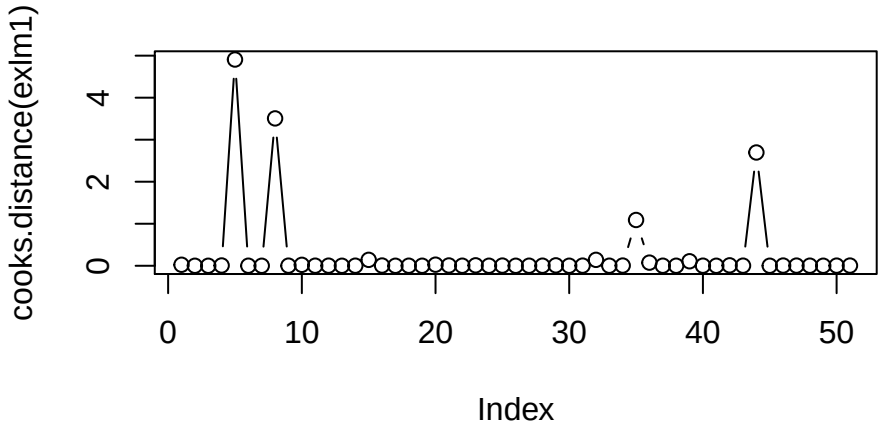
```
##     crime    employ       pop
##  1.121163 17.967808 17.568906
```

```
exlm4 = lm(expend~crime+pop, data=ex); vif(exlm4)
```

```
##   crime     pop
## 1.08213 1.08213
```

```
exlm5 = lm(expend~crime, data=ex);vif(exlm5)
# Error in vif.default(exlm5) : model contains fewer than 2 terms
```

In exlm1, exlm2 and exlm3, all VIF's are large, so there is a collinearity problem, but the exlm4 and exlm5 are OK.




```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 0.02 0.00 0.00 0.01 4.91 0.00 0.00 3.51 0.00 0.02 0.00 0.00 0.00 0.00 0.14 0.01
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
## 0.00 0.00 0.00 0.03 0.01 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.14
##   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
## 0.00 0.00 1.09 0.07 0.00 0.00 0.11 0.00 0.00 0.01 0.00 2.70 0.00 0.00 0.00 0.00
##   49   50   51
## 0.00 0.00 0.00
```

Thus, the potential and influence points are Point(5), Point(8), Point(35) and Point(44).

b) First, we start with step-up method.

```
summary(lm(expend~bad,data=ex))[[8]]
```

```
## [1] 0.6963839
```

```
summary(lm(expend~crime,data=ex))[[8]]
```

```
## [1] 0.1118564
```

```
summary(lm(expend~lawyers,data=ex))[[8]]
```

```
## [1] 0.9372789
```

```
summary(lm(expend~employ,data=ex))[[8]]
```

```
## [1] 0.9539745
```

```
summary(lm(expend~pop,data=ex))[[8]]
```

```
## [1] 0.9073261
```

The employ has highest value: 0.9539745.

```
summary(lm(expend~employ+bad,data=ex))[[8]]
```

```
## [1] 0.955097
```

```
summary(lm(expend~employ+crime,data=ex))[[8]]
```

```
## [1] 0.9550501
```

```
summary(lm(expend~employ+pop,data=ex))[[8]]
```

```
## [1] 0.95431
```

```
summary(lm(expend~employ+lawyers,data=ex))[[8]]
```

```
## [1] 0.9631745
```

The model of expend~employ+lawyers has highest value: 0.9631745.

```
summary(lm(expend~employ+lawyers+bad,data=ex))[[8]]
```

```
## [1] 0.9638741
```

```
summary(lm(expend~employ+lawyers+crime,data=ex))[[8]]
```

```
## [1] 0.9631881
```

```
summary(lm(expend~employ+lawyers+pop,data=ex))[[8]]
```

```
## [1] 0.9637326
```

Since the models did not yield any significant results, the step-up method stopped.

```
summary(lm(expend~bad+crime+lawyers+employ+pop,data=ex))[[8]]
```

```
## [1] 0.9675314
```

```
summary(lm(expend~bad+lawyers+employ+pop,data=ex))[[8]]
```

```
## [1] 0.9665736
```

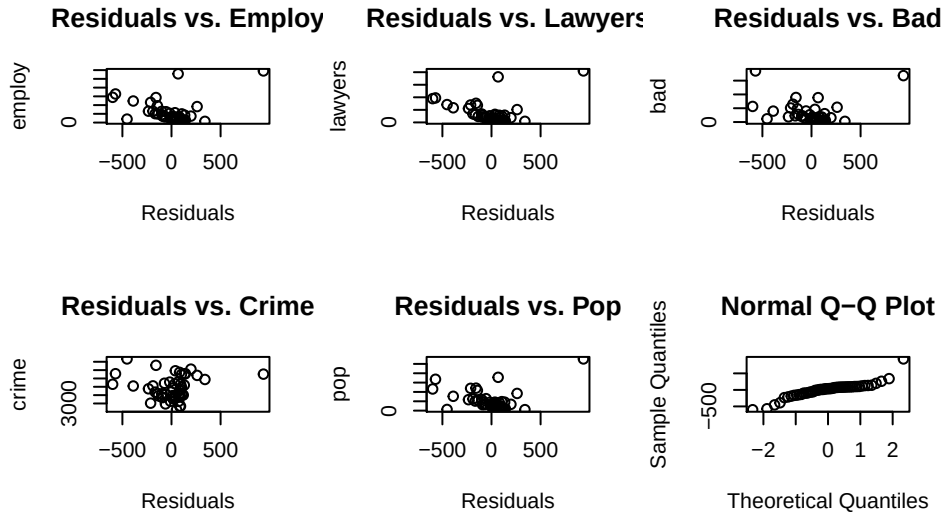
```
summary(lm(expend~bad+lawyers+employ,data=ex))[[8]]
```

```
## [1] 0.9638741
```

```
summary(lm(expend~lawyers+employ,data=ex))[[8]]
```

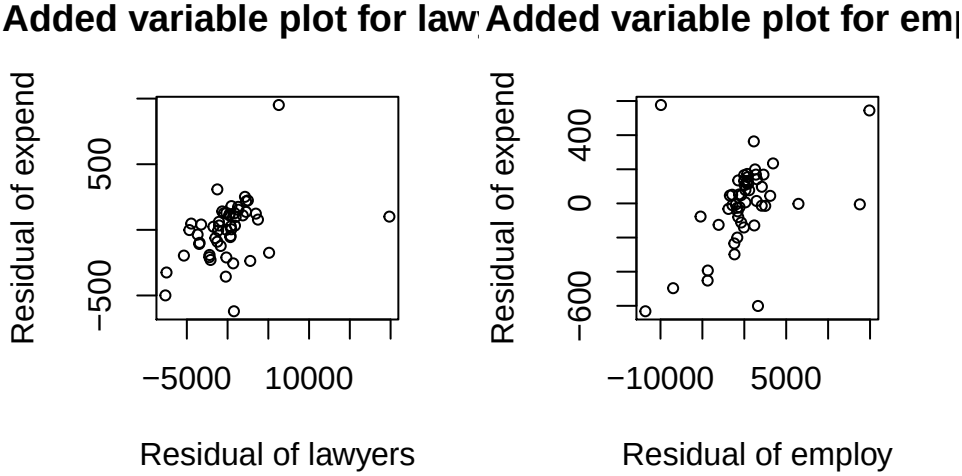
```
## [1] 0.9631745
```

All of these models did not yield any significant results, so the step-down method stopped. Hence, expend~lawyers+employ is the final model for both methods, which $expend = -110.7 + 0.002971 \times employ + 0.02686 \times lawyers$.



c)

From question(a), it already shows the collinearity of dependent and independent variables. The above graphs claims that the spread of residuals against variables did not show such a pattern existing. And the QQ-plot shows the residuals are normally distributed.

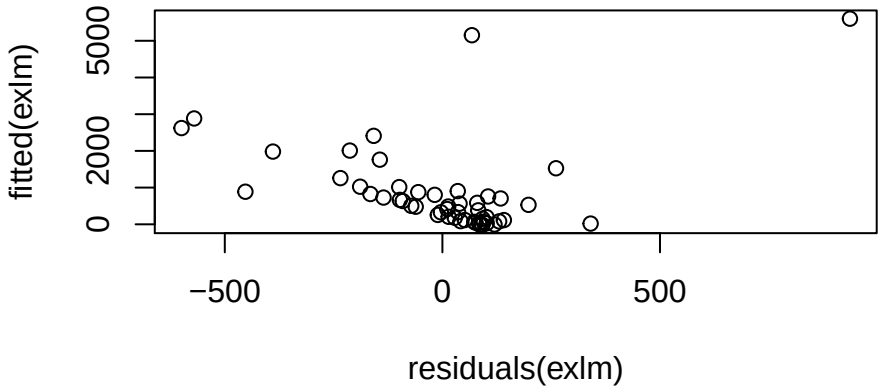


The added variable plots also show that there is no such specific curved pattern visible.

```
shapiro.test(residuals(exlm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(exlm)
## W = 0.8475, p-value = 1.118e-05
```

The Shapiro-Wilk normality test shows the same as the QQ-plot, which means it is still normally distributed since p-value=1.118e-05 < 0.05.



Moreover, there is no patterns or errors are visible in the scatter plot of residuals against Y (and $\hat{Y}$).