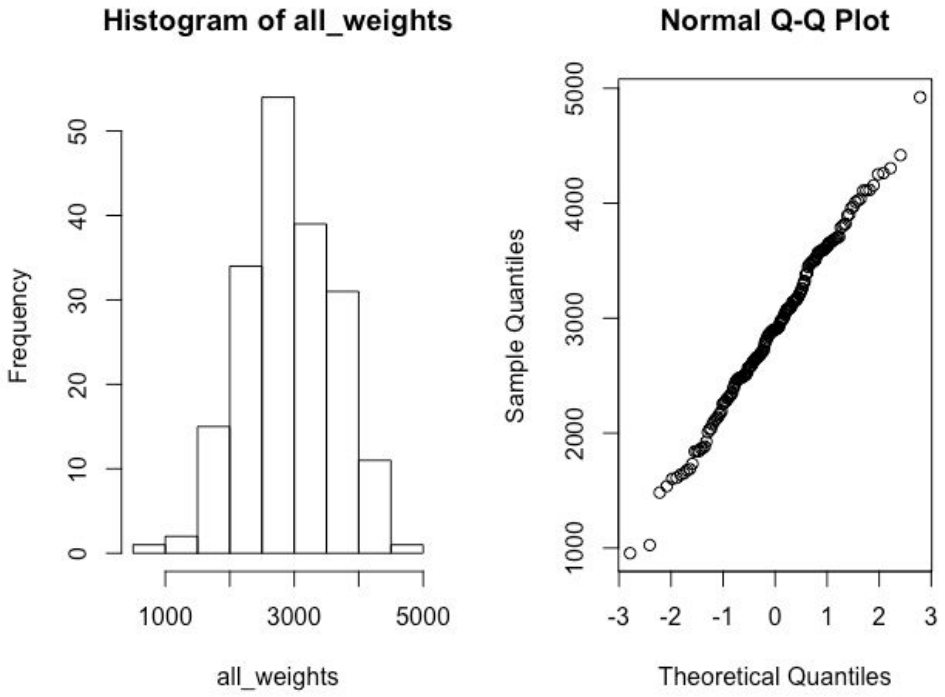
EDDA - Assignment 1

Imme Glaudé 2682588, Hidde van Oijen 2590893, Tiddo Loos 25674974

Exercise 1

a) Considering the histogram and QQ-plot of the data below, we assume the data to be normally distributed. We estimated a mean of μ = 2913.293 (90% CI 2829.202 - 2997.384).




```
> t.test(all_weights,conf.level = 0.90)

       One Sample t-test

…

90 percent confidence interval:
 2829.202 2997.384
sample estimates:
mean of x
 2913.293
```

b) Taking α = 0.1 leads to significant test outcome, meaning the population mean to be >2800. Any α > 0.01357 would also be insignificant, as p = 0.01357.

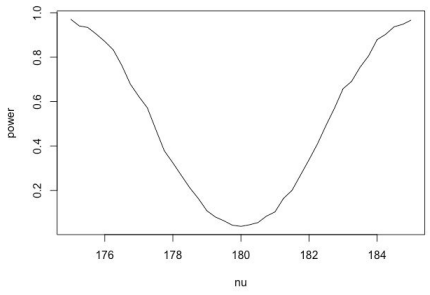
```
> t.test(all_weights,mu=2800,alternative = 'greater')

       One Sample t-test

data:  all_weights
t = 2.2271, df = 187, p-value = 0.01357
alternative hypothesis: true mean is greater than 2800
95 percent confidence interval:
 2829.202      Inf
sample estimates:
mean of x
 2913.293
```

c) It's different from question A because it's a 95% CI by default. It's one-sided, since the alternative hypothesis is 'greater than' in question B.

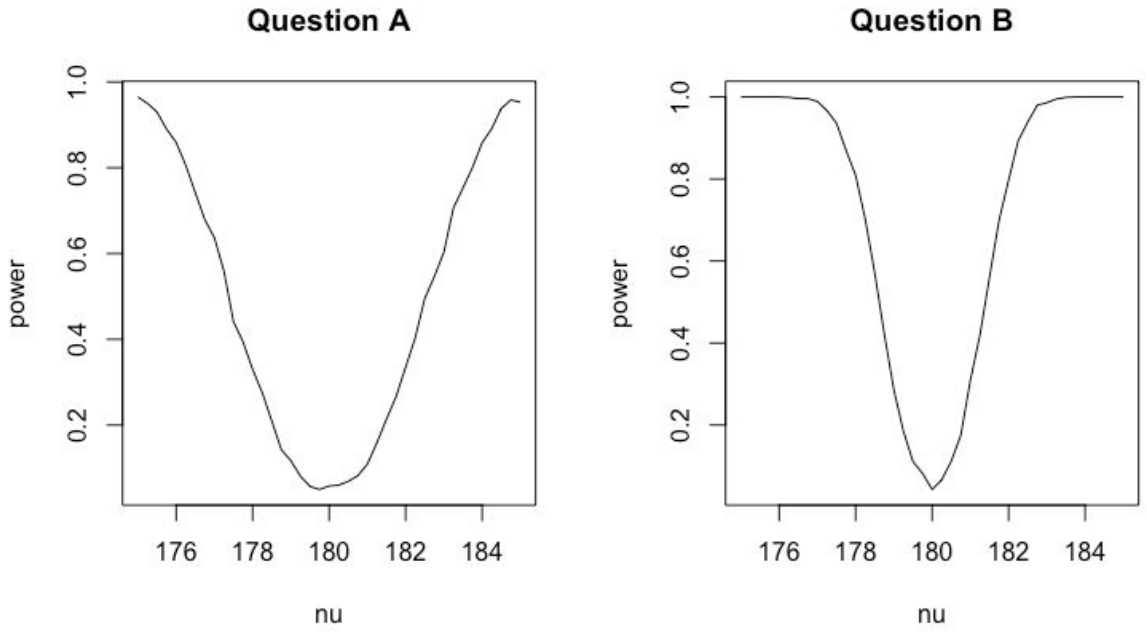
Exercise 2

a)

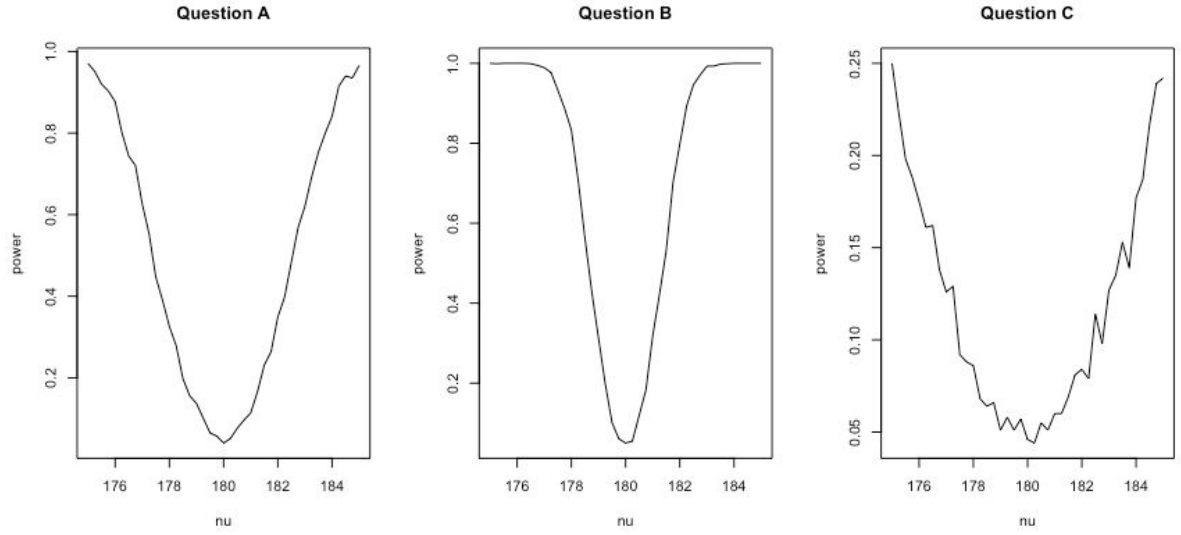
```
> n=m=30; mu=180;nu=175; sd=5;B=1000; nus=seq(175,185,by=0.25);
> p=numeric(B); C=length(nus); powers=numeric(C);

> for (c in 1:C){
+   nu = nus[c]
+   for (b in 1:B) {
+     x=rnorm(n,mu,sd); y=rnorm(m,nu,sd);
+     p[b]=t.test(x,y,var.equal=TRUE)[[3]]}
+   powers[c]=mean(p<0.05)}

> plot(nus,powers,type="l",xlab="nu",ylab="power")
```



b)



c)



d) A higher sample size increases the probability of rejecting H0 (power) at a certain point, as seen in A vs B. Question C suggests that a higher sd will cause a decrease in power.
So:
- Sample size increases: Power increases.
- SD increase: Power decreases.

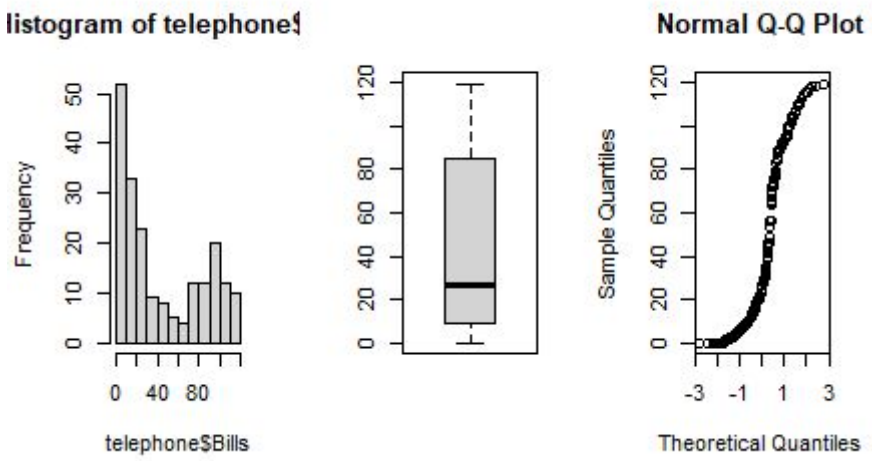
Exercise 3

a)

Plotted data

Through the commands made below we can view the data in three different plots which give us extra information about the distribution of data. The different plots show us that the data is not normally distributed. This can be seen best in the QQ-plot where the line is not a straight line but more of a curve. If we look at the histogram we can see that the data has a skewness to the left. This makes it so that we can say that the data is binomially distributed. The plots and functions to make the plots can be found below.

```
> hist(telephone$Bills)
> boxplot(telephone$Bills)
> qqnorm(telephone$Bills)
```



Marketing advice

Our advice to the marketing manager is to set a minimum price on the telephone bill. This because of the fact that some clients have really low bills that are close to €0. On these bills the company could lose money, because the fixed costs of having that client are higher than the yield on that client. An idea to solve this is to have a minimum telephone bill to compensate for those fixed costs.

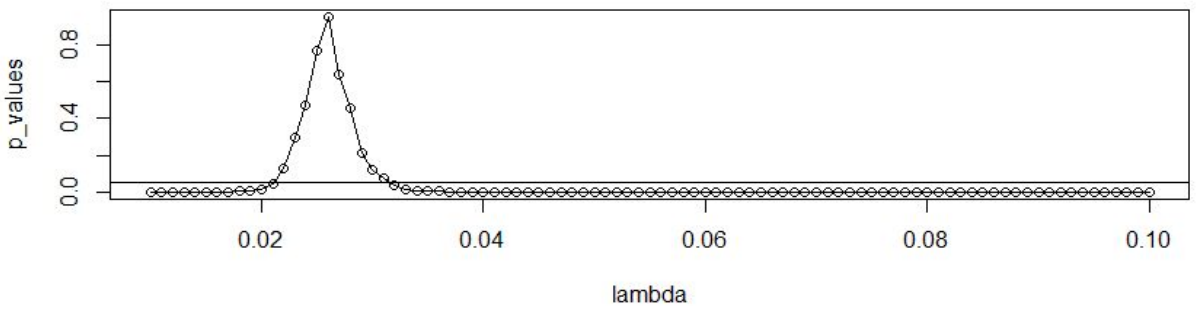
Inconsistencies

We could not find any inconsistencies in the data that could influence the different tests. Some could argue that the €0 bill clients could give a wrong view of the clients because those clients should be dropped because they do not contribute to the companies revenue. But we think those clients are still clients and should not be left out until dropped.

b) When looking at the p-values for each lambda, it can be seen that the p-value is higher than 0.05 with a lambda between [0.022,0.031]. This means that an exponential distribution with a lambda between [0.022,0.031] will not significantly differ for the telephone data. This makes it possible to conclude that the telephone

data is distributed exponentially. The code to construct the bootstrap test and the plot of p-values against the different values of lambda can be found below.

```
> t = median(telephone$bills)
> B = 1000
> tstar = numeric(B)
> n = 200
> p_values = list()
> lambda = seq(0.01, 0.1, by=0.001
> for (j in lambda){
+   for (i in 1:B){
+     xstar = rexp(n, j)
+     tstar[i] = median(xstar)}
+   pl = sum(tstar<t)/B; pr = sum(tsar>t)/B; p =2*min(pl, pr)
+   p_values <- c(p_values, p)
+   if (p>0.05){
+     print(j)}}
```



c) The code below shows how the 95% bootstrap interval is constructed. The 95% bootstrap interval: [20.12, 40.94] with a median of 26.905.

```
> B = 1000
> Tstar = numeric(B)
> for (i in 1:B) {
+   Xstar = sample(telephone$Bills, replace = TRUE)
+   Tstar[i] = median(Xstar)}
> T1 = median(telephone$Bills)
> Tstar25 = quantile(Tstar, 0.025)
> Tstar975 = quantile(Tstar, 0.975)
> c(2*T1-Tstar975, 2*Tstar25)
```

d) With the central limit theorem and with the mean we can estimate lambda, because the *mean = 1 / lambda*. Next to that we know that the *exponential distribution = ln(2)/ lambda*. By combining these formulas we estimate the *median = ln(2) * lambda*. Using the formula *ln(2) * upper/lower* bound of the mean we can estimate the *upper/lower* bound of the median. With the input below a 95% confidence interval is calculated by making use of the central limit theorem. This confidence interval acquired [26.450, 33.975] has a median of 30.212.

Comparing the 95% confidence interval [26.450,33.975] acquired by using the central limit theorem to the 95% confidence interval [20.12, 40.94] acquired with the bootstrap interval test, we can see that the central limit theorem interval is way more precise, assuming our assumptions are correct.

```
> median_limit = log(2)* mean(telephone$Bills)
> bound = (1.97 * sd(telephone$Bills)/sqrt(200))
> median_upper = (mean(telephone$Bills)+bound)*log(2)
> median_lower = (mean(telephone$Bills)+bound)*log(2)
```

e) To conduct these experiments we used a binomial test to test if the fraction is bigger or smaller than a certain amount of euro's. The two experiment designs with their conclusions are presented below.

In the first test we took a look if the median was equal or bigger than €40. The test that was executed can be found below. Looking at the test results, we found a p-value of 0.001 which is smaller than 0.05 which means that H0 does not hold. Thus, the median bill is not equal to €40 or higher, but the median is smaller than €40.

```
> sum(telephone$Bills>40)
[1] 83
> binom.test(83, 200, p = 0.5, alternative = c("less")
```

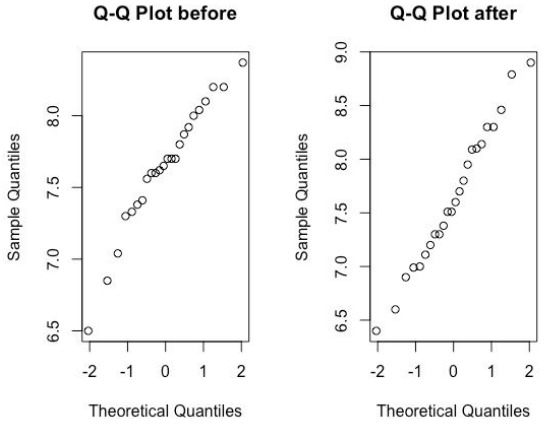
In the second test we looked if the amount of €10 bills is less than 25%. Therefore we constructed a test that can be found below. Here we found a p-value of 0.398, which is bigger than 0.05, which means that H0 holds. This means that 25% of the bills are lower than €10.

```
> sum(telephone$Bills<10)
[1] 52
> binom.test(52, 200, p = 0.25, alternative = c("less")
```

Exercise 4

a) Looking at the QQ-plots below, we assume the samples are normally distributed. So we use Pearson's correlation coefficient. Looking at the outcome below, there seems to be a correlation between the 'before' and 'after' sample (*r = 0.639, p-value = 0.001*).

```
> qqnorm(run$before,main="Q-Q Plot before")
> qqnorm(run$after,main="Q-Q Plot after")
```




```
> cor.test(run$before,run$after)
...
t = 3.8944, df = 22, p-value = 0.00078
r = 0.638803
```

b) A paired t-test on softdrink/lemo group on the ‘before’ and ‘after’ measures resulted in a mean of difference of *-0.145*, p-value = 0.437. Therefore, we can not reject H0.

```
> t.test(run[1:12,]$before,run[1:12,]$after,paired=TRUE)
...
t = -0.80596, df = 11, p-value = 0.4373
mean of the differences
                -0.145
```

A paired t-test on energy drinking group on the ‘before’ and ‘after’ measures resulted in a mean of difference of 0.1541667, p-value = 0.1264. Therefore, we can not reject H0.

```
> t.test(run[13:24,]$before,run[13:24,]$after,paired=TRUE)
...
t = 1.6538, df = 11, p-value = 0.1264
mean of the differences
              0.1541667
```

c) The time difference does not seem to be significantly affected by the type of drink *(p-value = 0.1586)*.

```
> differences = numeric(24)
> for (i in 1:24){ differences[i] = run$before[i]-run$after[i]}

> t.test(differences[1:12],differences[13:24])
...
```

```
t = -1.4764, df = 16.509, p-value = 0.1586
 mean of x  mean of y
-0.1450000  0.1541667
```

d) If the main aim was to test whether drinking the energy drink speeds up the running, it should be taken into account that a person might run slower the second time, regardless of the drink. A control group, in which no drink was given, should have been added to the design. In question C, a similar objection arises.

Exercise 5

a) In this exercise we test if the distributions of weights between the meatmeal and sunflower feed groups differ. We conduct this test with three methods:

1) First, we conducted the unpaired, two-sided t-test as there are no repeated measurements and the differences between the groups is investigated. The test indicated that the distribution between the groups differ (*t=2.18, df=21, p=0.04*)

```
> t.test(sunf_w, meat_w, alternative = c("two.sided"), paired = FALSE, var.equal
= TRUE)
```

2) Secondly, we conducted the Mann-Whitney test *(W=96, p=0.069)*. Based on the results of the Mann-Whitney test it cannot be concluded that the distributions of weights differ between the sunflower and meatmeal feed groups, as *p>0.05*.

```
> wilcox.test(sunf_w, meat_w, alternative = c("two.sided"), paired = FALSE)
```

3) At last, the Kolomogorov-Smirnov test is used to test whether the distributions of the two groups differ. The test showed that there is no significance *(D=0.47, p=0.11)* and therefore the test concludes that the two distributions of the sunflower and meatmeal feed groups do not differ.

```
> ks.test(sunf_w, meat_w, alternative = c("two.sided"), paired = FALSE)
```

The Mann-Whitney and the Kolomogorov-Smirnov test are nonparametric tests. This means that the tests compare two unpaired groups in contrast to the parametric t-test. Both sunflower and meatmeal feed groups have a sample size n<20 and therefore checking the normality of the distributions with qq-plots is not reliable. Therefore, the t-test is in this particular case not a reliable test. Thus, we conclude that there is no significant difference in the distributions between the two groups, based on the results of the Mann-Whitney and the Kolomogorov-Smirnov test.

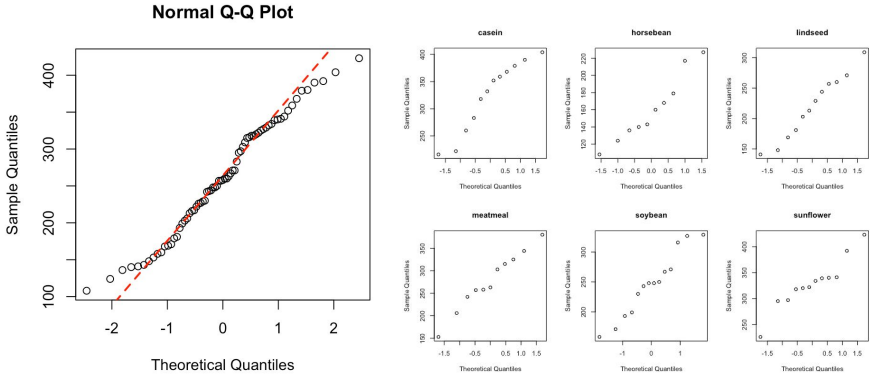
b) In this exercise we investigated whether the feed supplements have an effect on the chicken weights. The ANOVA test indicates that the supplements do have an effect on the chicken weights *(F-statistic: 15.36 on 5 and 65DF, p<0.001).* To estimate the chicken weights we command a summary of the ANOVA test to calculate the mean estimated weights. From the table it can be concluded that sunflower is the best feed supplement.

```
> chick_w_anov=lm(yield~feed,data=chick_frame)
> anova(chick_w_anov)
> summary(chick_w_anov)
```

```
> summary(chick_w_anov)

Call:
lm(formula = yield ~ feed, data = chick_frame)

Residuals:
     Min       1Q   Median       3Q      Max
-123.909  -34.413    1.571   38.170  103.091

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   323.583     15.834  20.436  < 2e-16 ***
feedhorsebean -163.383     23.485  -6.957 2.07e-09 ***
feedlinseed   -104.833     22.393  -4.682 1.49e-05 ***
feedmeatmeal   -46.674     22.896  -2.039 0.045567 *
feedsoybean    -77.155     21.578  -3.576 0.000665 ***
feedsunflower    5.333     22.393   0.238 0.812495
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.85 on 65 degrees of freedom
Multiple R-squared:  0.5417,    Adjusted R-squared:  0.5064
F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

Supplement	Estimated mean weight
casein	323.58
horsebean	160.20
linseed	218.75
meatmeal	276.90
soybean	246.43
sunflower	328.92

c) In this exercise we evaluate the ANOVA assumptions with diagnostic tools.



after plotting the data points in qq-plots and after conducting a shapiro-wilk test *(p=0.21)* we can conclude that the data is normally distributed as *p>0.05*.

```
> shapiro.test(data$weight)
```

As we now know that the data is normally distributed we can check the homogeneity of variance with Bartlett's test *(K-squared =2.51, df = 5 and p = 0.77)*.

```
> bartlett.test(data$weight~group, data)
```

From Bartlett's test we can assume that the variance of the chicken weights are statistically significantly different for the 6 treatment groups as *p>0.05.*

d) At last we conducted Kruskal-Wallis to test whether the supplements have an effect on the chicken weights. From the test we concluded that there is a significant difference between the groups of chickens and their presented feed supplement *(Chi-squared = 37.34, df = 5, p<0.001)*. This, conclusion is consistent with the conclusion drawn from the ANOVA.

```
> attach(chick_frame); kruskal.test(yield,feed)
```

Like the ANOVA test, the Kruskal-Wallis test has a *p<0.001*. This means that the conclusions correspond. We proved the ANOVA assumptions and the ANOVA test is known to be more precise Than the Kruskal-Wallis test. The Kruskal-Wallis test is also usable for non normal distributed data but is less precise. However, it is sufficient to conclude whether some supplements are better than others.

EDDA - Assignment 2

Imme Glaudé 2682588, Hidde van Oijen 2590893, Tiddo Loos 25674974

Exercise 1

A)

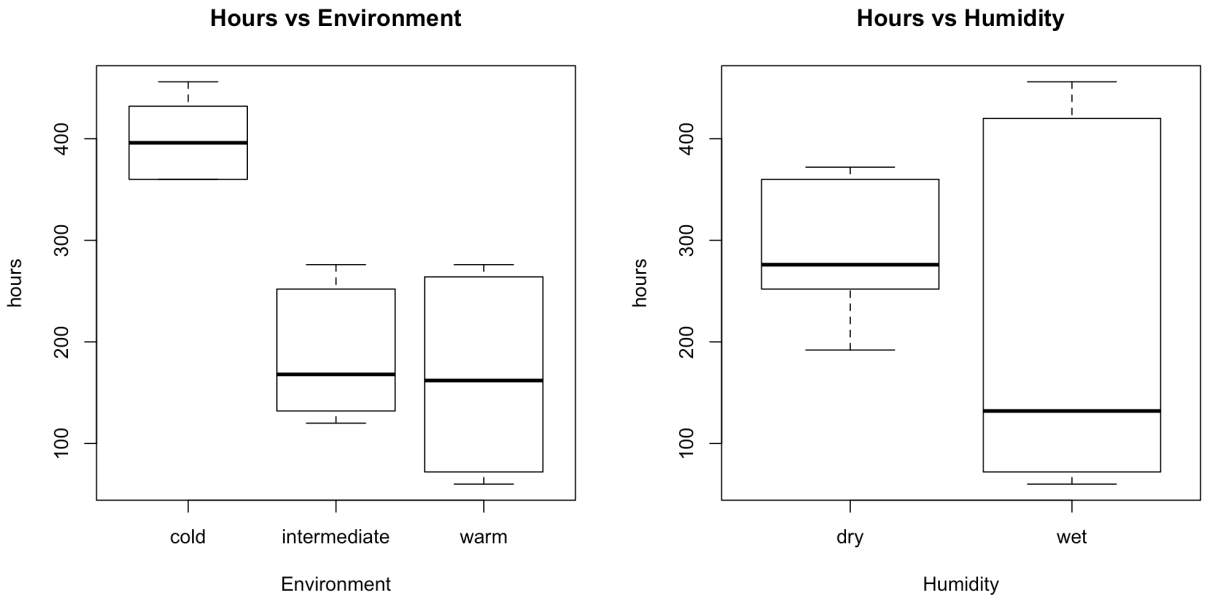
The code below was used to randomize the deviation of 18 slices of loafs into 6 different combinations of conditions. The result of the deviation can be found below the code.

```
> I=3; J=2; N=3
> rbind(rep(1:I,each=N*J),rep(1:J,N*I),sample(1:(N*I*J)))
```

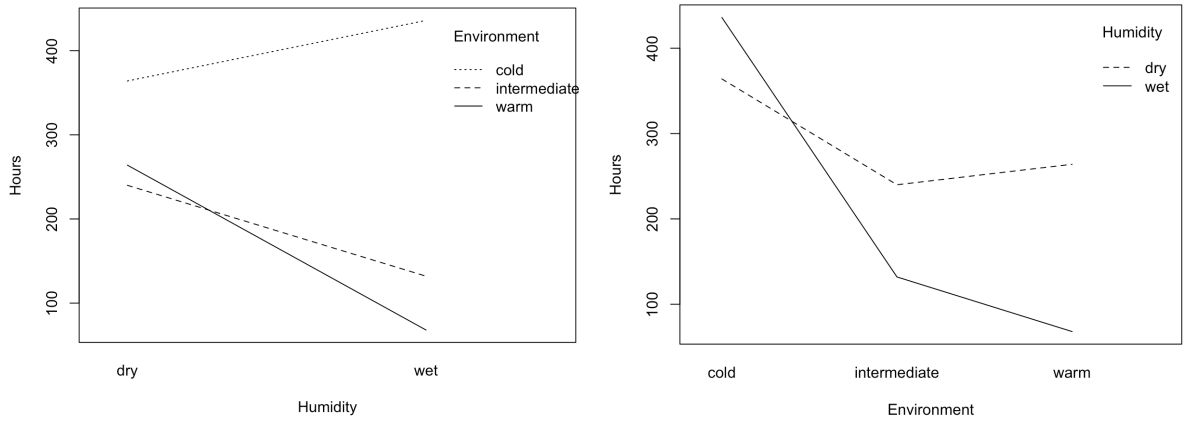
```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18]
[1,]    1    1    1    1    1    1    2    2    2     2     2     2     3     3     3     3     3     3
[2,]    1    2    1    2    1    2    1    2    1     2     1     2     1     2     1   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

B)

```
> plot(as.factor(data$environment ,data$hours, main="Hours vs Environment",
xlab="Environment", ylab="hours")
> plot(as.factor(data$humidity),data$hours, main="Hours vs Humidity",
xlab="Humidity", ylab="hours")
```



```
> interaction.plot(as.factor(data$environment), as.factor(data$humidity),
data$hours, xlab="Environment", ylab="Hours", trace.label = "Humidity")
> interaction.plot(as.factor(data$humidity), as.factor(data$environment),
data$hours, xlab="Humidity", ylab="Hours", trace.label = "Environment")
```



C)

For this sub-question we conducted an ANOVA-test to find evidence for interaction between environment and humidity.

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

The p-value for testing H0 : $\alpha_i = 0$ for all i is < 2.461e-10; for H0: $\beta_j=0$ for all j is 4.316e-6; for H0: $\gamma_{i,j} = 0$ for all (i,j) is 3.705e-7. So, there is evidence for interaction (both factors seem also to have a main effect).

Interaction effect:

An interaction effect means that there are different independent variables that influence the relationship between other independent variables and the dependent variable. In this case it means that the influence of the temperature on the decay time is affected by the humidity. For the other way around it is the same: the influence of humidity on the decay time is affected by the temperature.

D)

Looking at the F-values the environment has the biggest influence on the decay with a F-value of 233.685. But next to that the humidity and the interaction effect also have a significant effect. Because the interaction effect is significant the two factors cannot be looked at independently.

This makes the question which effect has the greatest influence not a good question to ask because the interaction effect has a significant influence.

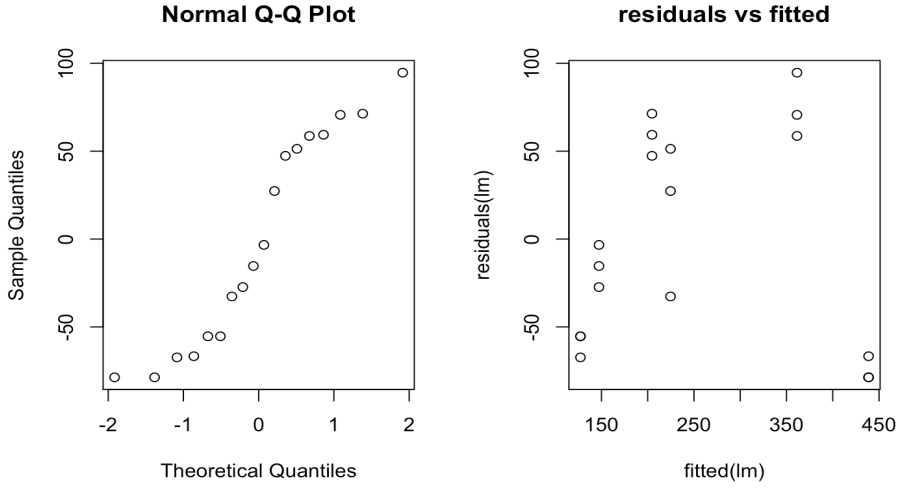
	F	*p*
environment	233.685	2.461e-10
humidity	62.296	4.316e-6
environment:humidity	64.796	3.705e-7

E)

Normality:

```
> lm=lm(hours~environment+humidity, data=data)
> qqnorm(residuals(lm))
> plot(fitted(lm),residuals(lm), main="residuals vs fitted")
> shapiro.test(residuals(lm))
```

Looking at the QQ-plot, we can say that the residuals are approximately normally distributed. Also, the sharpirotest of the residuals concludes that the data is normally distributed with a p-value of 0.0590. Looking at the residuals vs. fitted data we cannot say that the residuals change systematically.

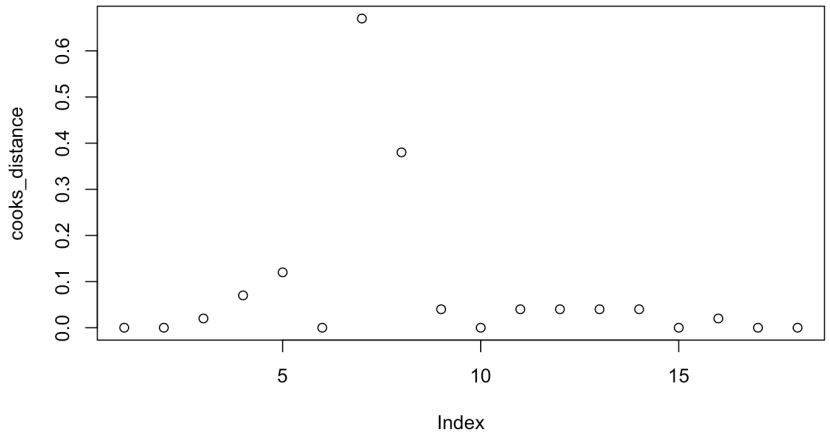


Outliers:

```
> cooks_distance = round(cooks.distance(bread_aov),2)
> plot(cooks_distance)
```

Looking at the QQ-plot there are a couple of suspicious outliers. But looking at the cook's distance from all of the data points we see that the cook's distances are smaller than 1. This

means that they are not influence points and that the points can be maintained in the database.



Exercise 2

A)

The code below assigns a student to one of the three interfaces within their skill levels

```
B=5; I=3; N=1
> for (i in 1:B) {
+     cat( i, '-', sample(1:(N*I)) + 3*(i-1), '\n')}
```

```
1 - 2 3 1
2 - 4 5 6
3 - 8 7 9
4 - 12 10 11
5 - 13 14 15
```

B)

The test of H0, that the search time for every interface is the same, was conducted with a two-way ANOVA. The two-way ANOVA was used to compute the mean difference between groups that were split on the two factors. For the interface we got a p value of 0.013 which means that H0 can be rejected. This means that the search times are not the same for every interface.

```
> searchaov = lm(search$time~search$skill+search$interface, data=search)
> anova(searchaov)
```

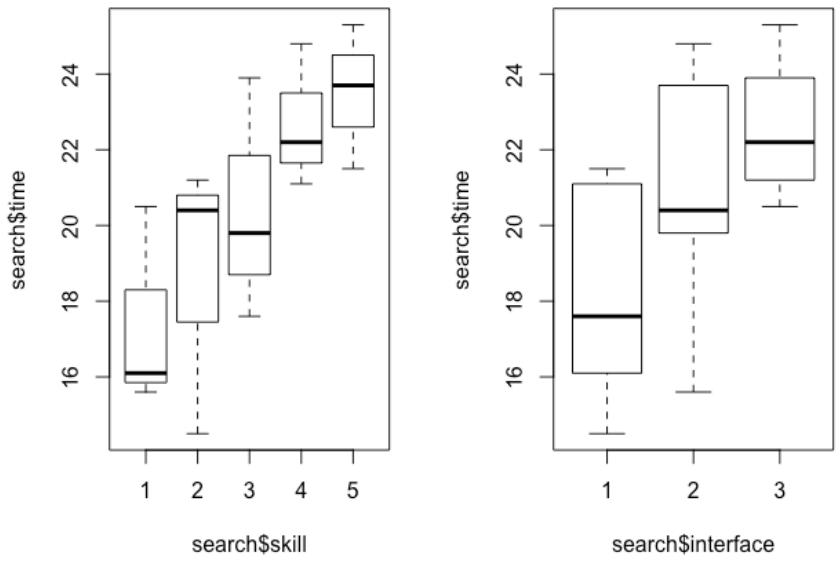
Using the command below we found that the search time of interface 3 is 4.460 seconds longer than the first one.

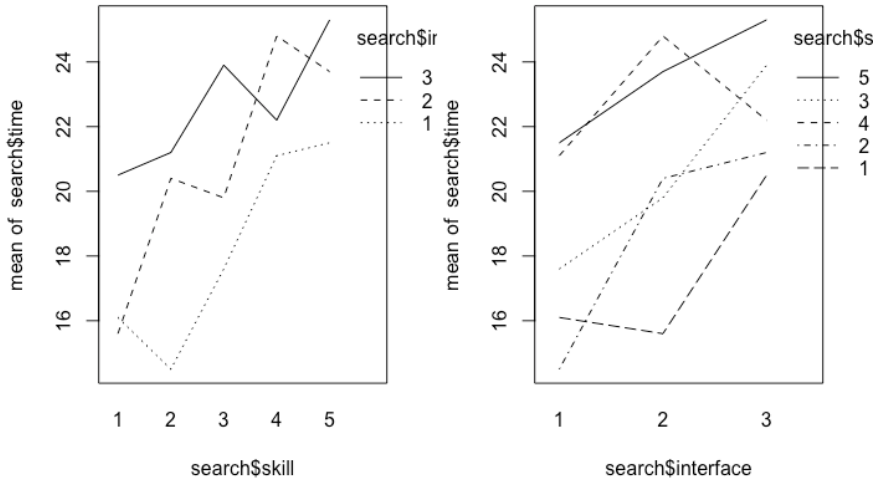
```
> summary(searchaov)
```

Again using the command below, we see that the intercept has the lowest amount of time needed to do a search. This means that skill level 1 and interface 1 together will give the shortest amount of search time of 15.013.

At last, with the use of the summary given by the command below, we estimate that a typical user with skill level 3 and interface 3 needs 15.013 + 3.033 + 4.460 = 22.506 search time

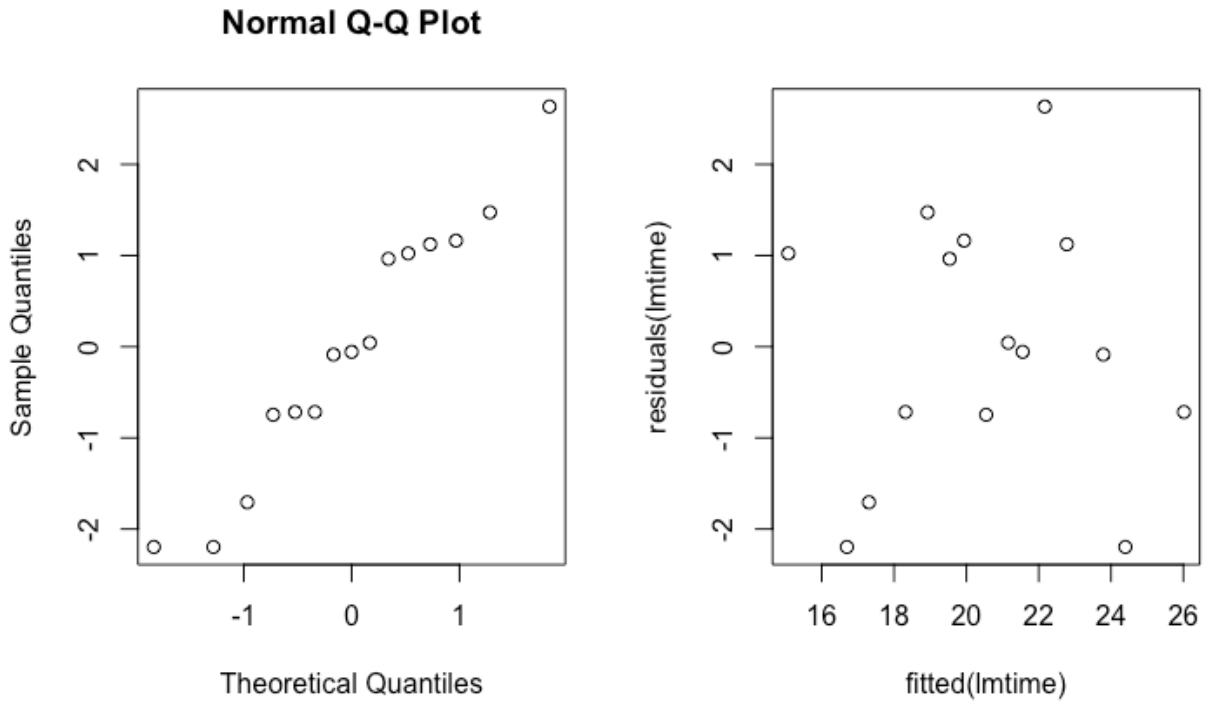
Some graphical summaries are added below.





C)

The QQ-plot doesn't seem very convincing for a normal distribution. However, a p value of 0.6023 was found using a Shapiro-Wilk test, indicating that the residuals are normally distributed. Looking at the residuals vs. fitted data we cannot say that the residuals change systematically.




```
> lmtime=lm(search$time~search$skill+search$interface)
> qqnorm(residuals(lmtime))
> plot(fitted(lmtime),residuals(lmtime))
> shapiro.test(residuals(lmtime))
```

D)

Friedman's test p-value = 0.04076, so we reject H0. There does seem to be a significant effect of the interface.

```
> friedman.test(search$time,search$interface,search$skill)
```

E)

One-way Anova test resulted in a p-value of 0.09, indicating an insignificant effect of interface. However, using this Anova would not be appropriate, since we are not taking the factor skill in consideration, which seems to be influential.

```
> anova(lm(search$time~search$interface))
```

Exercise 3

A)

For the first test we used an ANOVA test if there is a difference between milk production when using different kinds of feedingstuffs. There are 2 different kinds of feedingstuffs, namely treatment A and B. Therefore our H0: a0 = a1 = 0, where a0 is feedingstuff A and a1 is feedingstuff B. Using the commands below we acquired the difference between the various feedingstuffs has a p-value of 0.517. Therefore we cannot say that there is a difference between milk production between the different feedingstuffs.

```
> cowaov = lm(cow$milk~cow$order+cow$id+cow$per+cow$treatment, data=cow)
> summary(cowaov)
```

B)

To analyse the mixed effects with the cows as random effects the commands below were used. Here we found that the variance of the random effect of the cows is 133.145.

```
> cowlmer=lmer(milk~treatment+order+per+(1|id),data=cow,REML=FALSE)
> summary(cowlmer)
```

Only through using this command we couldn't extract the p-value for the difference between the amounts of milk when using different kinds of feedingstuff. Therefore we needed to analyse it with the code below. Through comparing the model without treatment with the bigger model, we get a p-value of 0.446. This means that models don't differ significantly. This means that treatment is not significant. This is the same as the outcome as the ANOVA test in question A.

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

C)

When using the command below, a T test will be executed comparing the two different treatments of feedingstuffs. The p-value extracted out of this test is 0.828 and says the difference between the treatments is not significant. This is the same as in the analysis of question a and b. But looking at question A, we found that the order and period are significant factors with 0.0150 and 0.00235. This means that the t test procedure doesn't give valid results. So, given that the result of C is in line with A and B, this is not a valid result and it's dangerous to use it.

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

Exercise 4

A)

For this test we want to check if the different columns are significantly different or that they do not differ. If the columns are significantly different, we can say that the works are written significantly differently. Therefore we need to check if the distribution over the different columns is equal. A test for independence is inappropriate here, because we don't want to check if a row and column variable are independent. This means that we need to test if the distributions are homogeneous over the columns.

B)

To check whether austen was consistent in her work we executed a chi squared test. Here we checked if the first 3 parts of the book were the same. This means we formulated H0 as Sense = Emma = Sand1 = 0. From the test executed below we found that the p-value is 0.2673. This means that we don't reject H0 and austen is consistent in her writing. A couple of inconsistencies can still be found in the writing. So is the "a" usage in Sand1 higher than in the other parts. Next to that the "that" usage in Sand1 is lower than in the other parts. At last, the "without" usage in Emma is lower than in the other parts

```
> austen_2=select(austen,-Sand2)
> z=chisq.test(austen_2); z
> residuals(z) # = (z$observed-z$expected)/sqrt(z$expected)
```

	Sense	Emma	Sand1
a	-1.02997736	-0.1290203	1.5937736
an	0.44728806	-0.1590968	-0.3746273
this	0.05133600	0.2938669	-0.5036577
that	0.74817619	0.2865778	-1.4423521
with	-0.04747379	0.5205063	-0.7035205
without	1.06544255	-1.5884103	0.8926239

C)

To check whether Austen and the admirer was successful in imitating her work, we executed a chi squared test. Here we checked if the four different parts of the book were the same. This means we formulated H0 as Sense = Emma = Sand1 = Sand2 = 0. From the test executed below we found that the p-value is 6.205e-05. This means that we need to reject H0 and conclude that the parts are significantly different from each other. This means that the imitator did not succeed in imitating austen. This could be explained because,The "an" usage of Sand2 was higher, the "that" usage was lower and the "with" usage was higher.

```
> z=chisq.test(austen); z
> residuals(z) # = (z$observed-z$expected)/sqrt(z$expected)
```

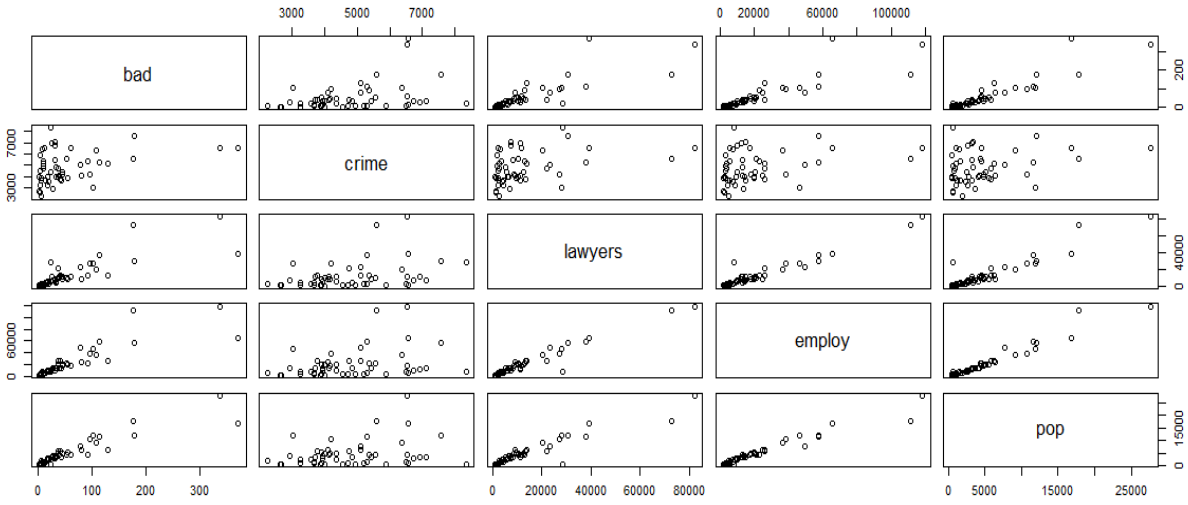
```
              Sense          Emma       Sand1       Sand2
a       -1.0149156 -0.1120927868  1.6062866 -0.05889921
an      -0.5906319 -1.2199545912 -1.0671306  3.72816398
this     0.1388299  0.3904903154 -0.4436450 -0.32671736
that     1.5943613  1.1798488360 -0.9099606 -3.04931581
with    -0.5120944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

Exercise 5

A)

Problem of collinearity

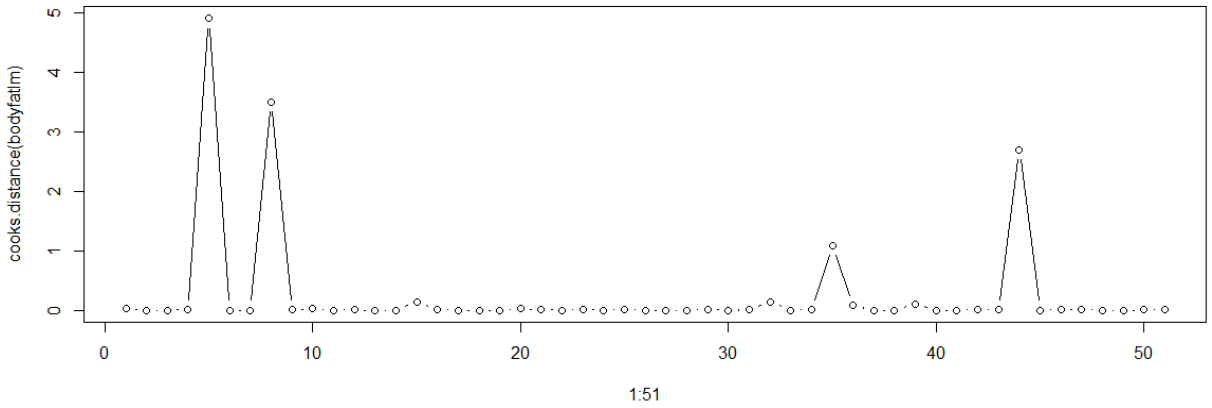
To check collinearity, scatterplots of the different variables were made and set against each other except from the states because that isn't a numeric value. These were created by the first line in the code below. We also leave expend out because this is a response variable and not an explanatory one. When looking at the scatterplots only the crime data does not seem to be collinear with the other data. This can be said because the scatterplots with crime don't show a straight line of a trend. Also with the code below the VIF-values are calculated. Here see that all of the variables except for crime have a VIF-value above 5. This means that there is a collinearity problem in between the variables except for the crime one.




```
> plot(expensescrime[,c(3:7)])
> install.packages("car")
> library(car)
> crimelm = lm(expend~bad+crime+lawyers+employ+pop, data=expensescrime)
> vif(crimelm)
```

Influence points

We used the Cook's distance to look if there were any influence points. This was done with the code and the graph below. Here we see that there are 4 points around or higher than 1 or 4 cook's distance. This means that they could be seen as influence points, but we left them in the data set, because we couldn't say they were errors.




```
> plot(1:6, cooks.distance(crimelm),type="b")
```

B)

Step-up

The resulting model acquired with the step-up method consists of employ + lawyers. This model is acquired with the use of the code below.

First step

```
> summary(lm(expend~bad, data=expensescrime))
> summary(lm(expend~crime, data=expensescrime))
> summary(lm(expend~lawyers, data=expensescrime))
> summary(lm(expend~employ, data=expensescrime))
> summary(lm(expend~pop, data=expensescrime))
```

Variable	R^2	P-values significant
bad	0.696	Yes
crime	0.112	yes
lawyers	0.937	yes
employ	**0.954**	yes

pop	0.907	yes

Take employ for model

Second step

```
> summary(lm(expend~employ + bad, data=expensescrime))
> summary(lm(expend~employ + crime, data=expensescrime))
> summary(lm(expend~employ + lawyers, data=expensescrime))
> summary(lm(expend~employ + pop, data=expensescrime))
```

Variable	R^2	P-values significant
Employ + bad	0.955	No
Employ + crime	0.955	No
Employ + lawyers	**0.963**	Yes
Employ + pop	0.954	No

Take lawyers for model

Third step

```
> summary(lm(expend~employ + lawyers + bad, data=expensescrime))
> summary(lm(expend~employ + lawyers + crime, data=expensescrime))
> summary(lm(expend~employ + lawyers + pop, data=expensescrime))
```

Variable	R^2	P-values significant
Employ + lawyers + bad	0.964	No
Employ + lawyers + crime	0.963	No
Employ + lawyers + pop	0.964	No

No p-values are significant. This means that no third variable needs to be selected. The Employ + lawyers variables make the best model.

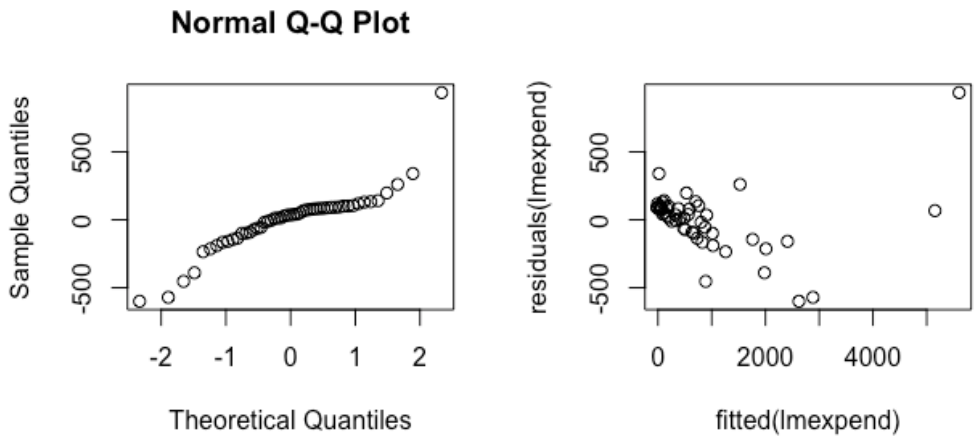
Step-down

The resulting model acquired with the step down method consists of lawyers + employ. This model is acquired with use of the code below.

```
> summary(lm(expend~bad+crime+lawyers+employ+pop, data=expensescrime))
-crime
> summary(lm(expend~bad+lawyers+employ+pop, data=expensescrime)) -pop
> summary(lm(expend~bad+lawyers+employ, data=expensescrime)) -bad
> summary(lm(expend~lawyers+employ, data=expensescrime)) resulting model
```

Which model is better?
Both models from stepping-up and stepping-down are the same, so no choice has to be made. But the models both have employ and lawyers and these variables are collinear. So, both in the same model are not good. Better is to take the first variable chosen by the step-up model. So the best choice is to only choose for the employ variable in the model.

C)
The residuals plot does not seem to be randomly distributed and the QQ-plot does not seem to fit a normal distribution. Also, performing a Shapiro-Wilk test p-value of 1.118e-05, meaning the residuals are significantly different from the normal distribution. With these findings, we could already conclude that these findings do not fit the model assumptions. Next to that, as said earlier, the employ and lawyer variable are collinear. This means that using them in the same model is not a good thing to do. Since the different model assumptions are not met, we can say that this is not a good model to use.




```
> par(mfrow=c(1,2))
> lmexpend=lm(expend~lawyers+employ, data=expensescrime)
> qqnorm(residuals(lmexpend))
> plot(fitted(lmexpend),residuals(lmexpend))
> shapiro.test(residuals(lmexpend))
```