EDDA - Assignment 2

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

Exercise 1

A)

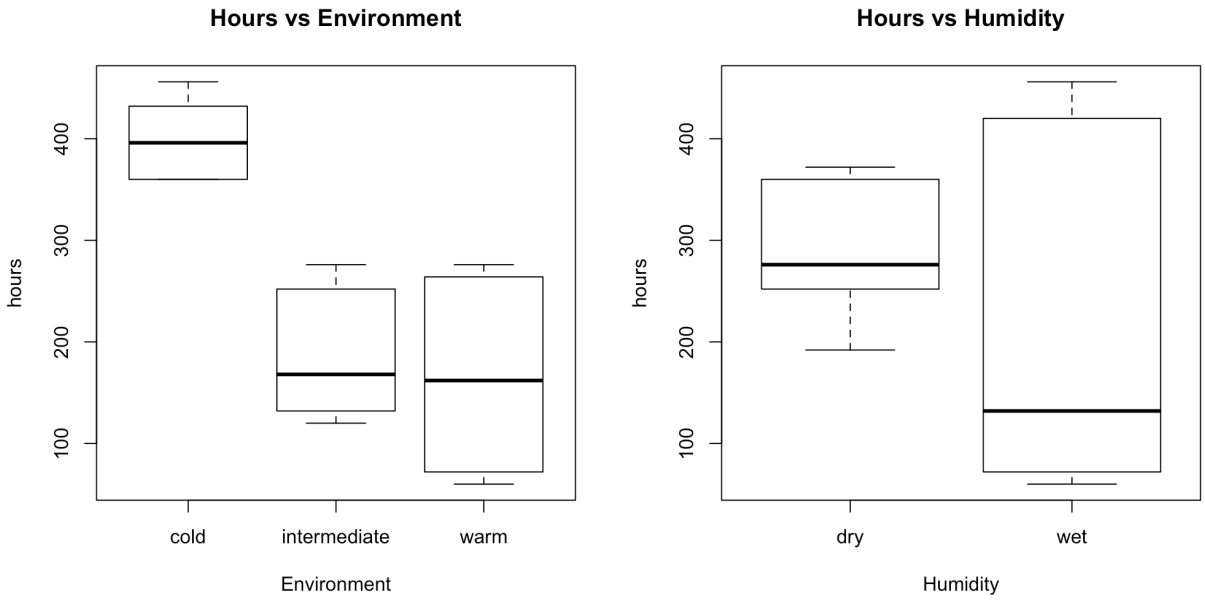
The code below was used to randomize the deviation of 18 slices of loafs into 6 different combinations of conditions. The result of the deviation can be found below the code.

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

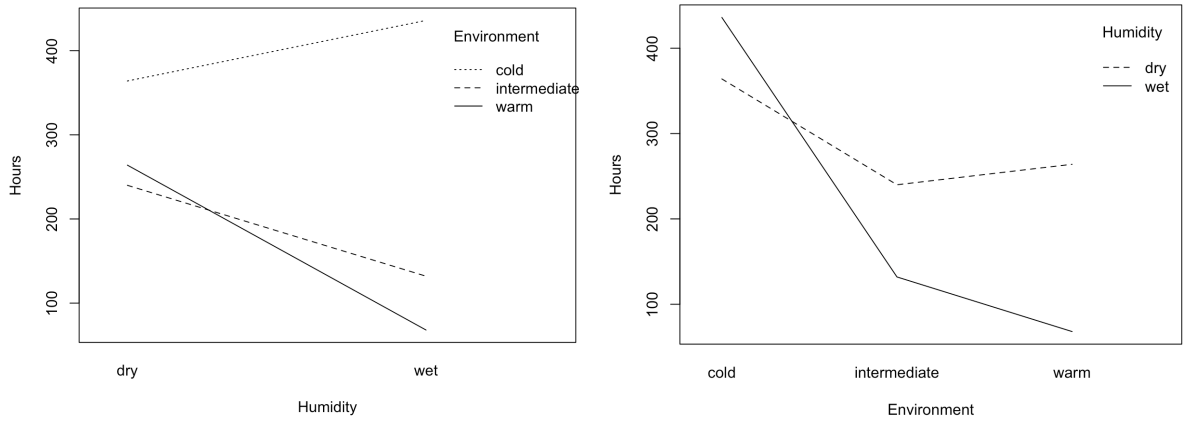
```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18]
[1,]    1    1    1    1    1    1    2    2    2     2     2     2     3     3     3     3     3     3
[2,]    1    2    1    2    1    2    1    2    1     2     1     2     1     2     1   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

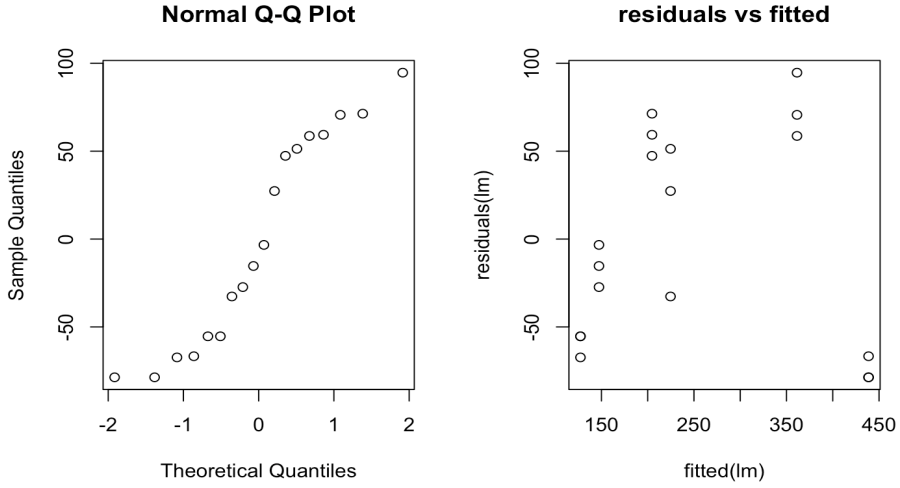
	F	*p*
environment	233.685	2.461e-10
humidity	62.296	4.316e-6
environment:humidity	64.796	3.705e-7

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.

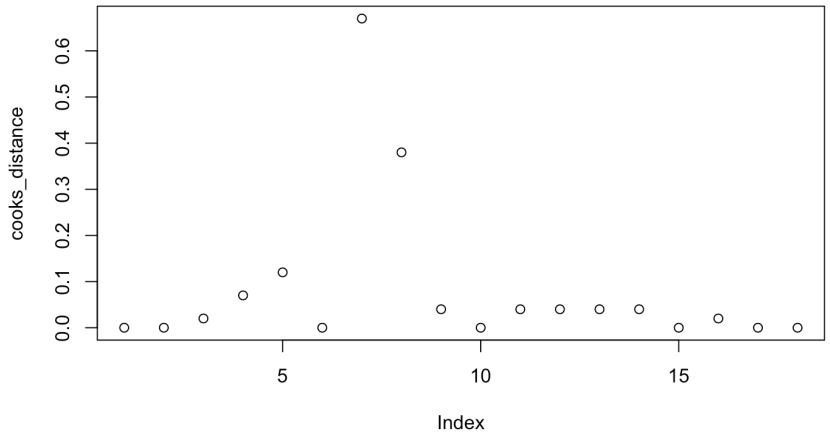


Outliers:

```
> 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)
```

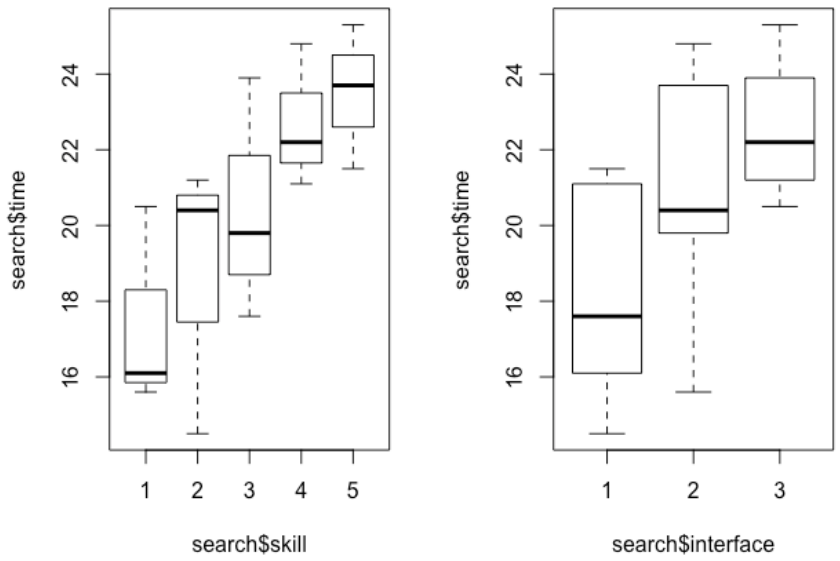
Using the command below we found that the search time of interface 3 is 4.460 seconds longer than the first one.

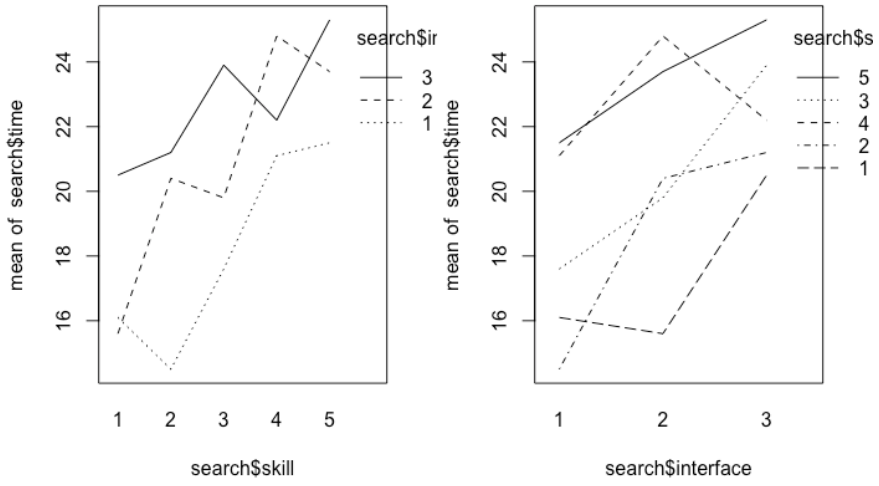
```
> summary(searchaov)
```

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

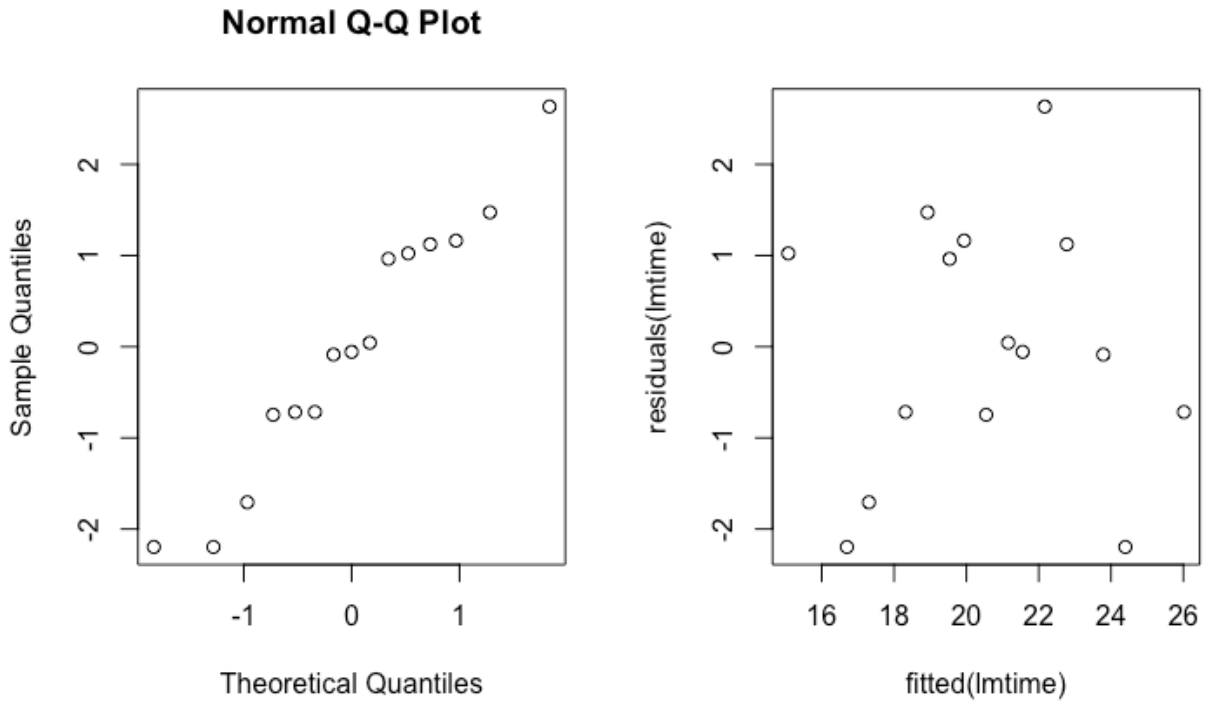
Some graphical summaries are added below.





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)
```

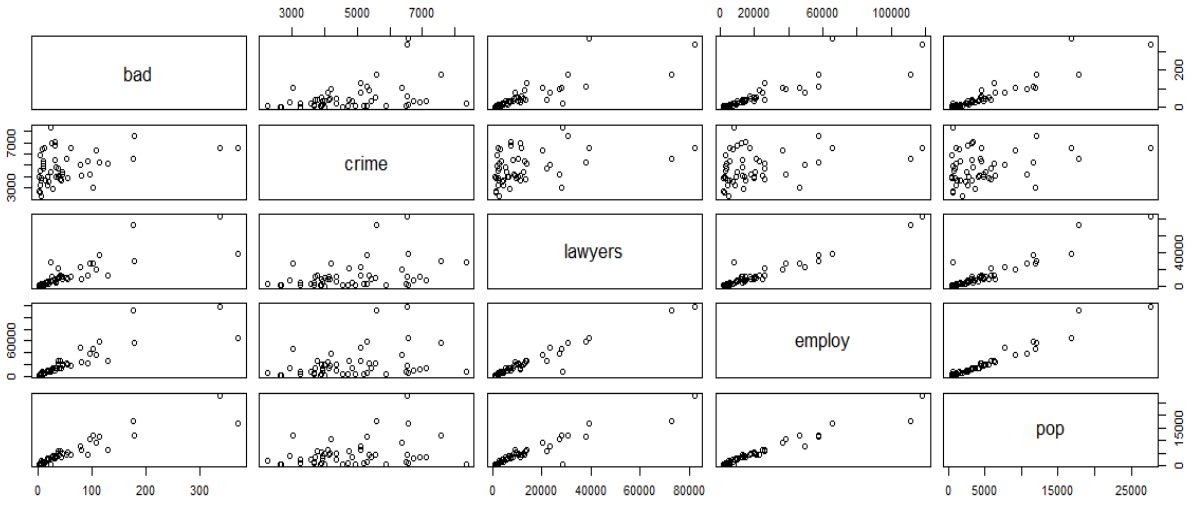
```
              Sense          Emma       Sand1       Sand2
a       -1.0149156 -0.1120927868  1.6062866 -0.05889921
an      -0.5906319 -1.2199545912 -1.0671306  3.72816398
this     0.1388299  0.3904903154 -0.4436450 -0.32671736
that     1.5943613  1.1798488360 -0.9099606 -3.04931581
with    -0.5120944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

Exercise 5

A)

Problem of collinearity

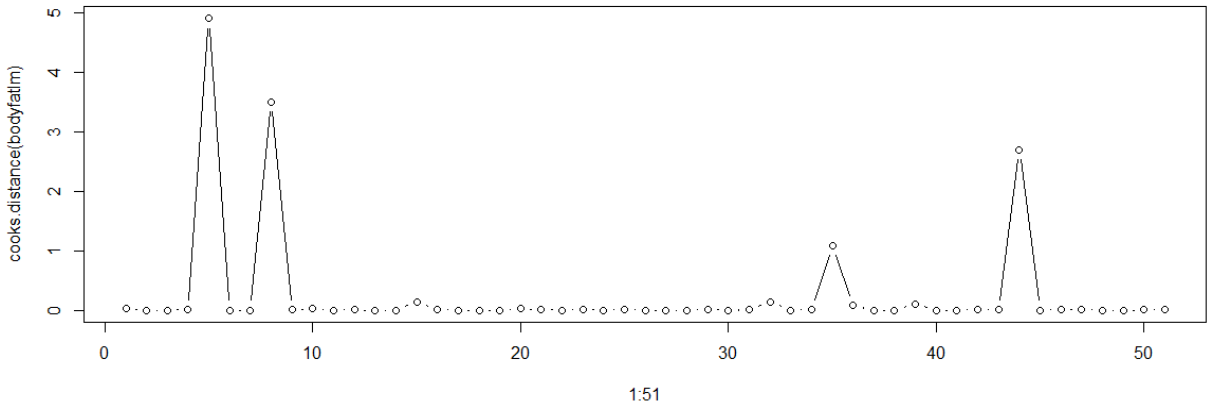
To check collinearity, scatterplots of the different variables were made and set against each other except from the states because that isn't a numeric value. These were created by the first line in the code below. We also leave expend out because this is a response variable and not an explanatory one. 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.

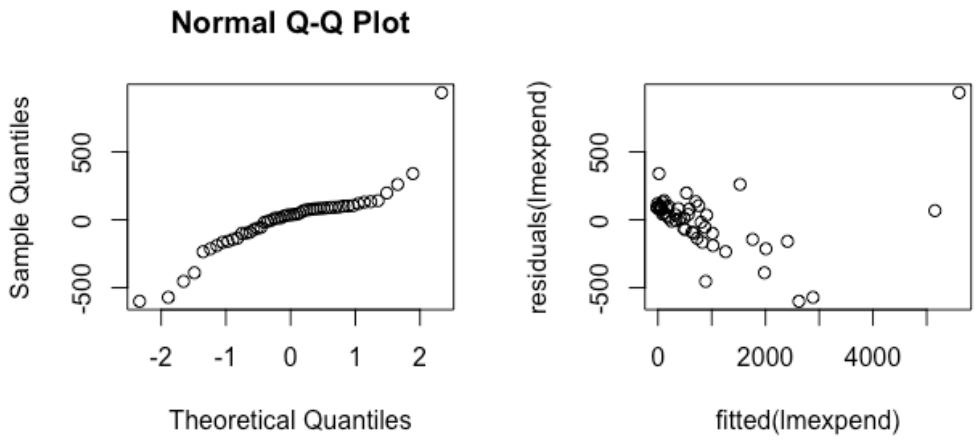
Step-down

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))
```