EDDA Group 29 Assignment 2

Geoffrey van Driessel (12965065), Yizhen Zhao (2658811) & Sophie Vos (2551583)

3/9/2020

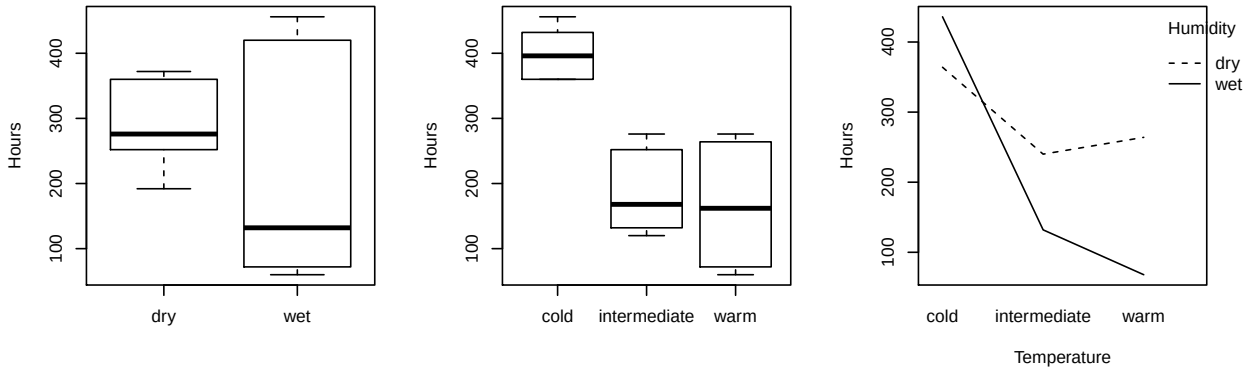
An overview of the R code is shown in the Appendix on page X.

Exercise 1

a) A randomized design with two categorical factors, with (1) the first factor having three categorical levels, (2) the second factor having two levels and (3) having three samples for each unique category, can be produced with the following R code:

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

b) The boxplot and interaction plot below confirms our intuition: (1) a cold environment causes a much slower decay, (2) wet bread has a much wider distribution (variance), (3) on average dry bread decays slower than wet bread, however, (4) wet and cold (frozen) bread has the slowest decay. From the non-parallel lines in the interaction plot and the wide distribution of the wet sample, we conclude that the (wet) humidity amplifies the effect of the temperature and it can thus be explained by the strong interaction between the two factors (opposed to the errors in the measurement).



c) We have three null hyphotheses: (1) H_0 : there is no main effect of first factor (humidity), (2) H_0 : there is no main effect of second factor (environment) and (3) H_0 : there is no interactions between the two factors. From the two-way ANOVA result below, we reject all null hyphotheses. This means that both factors have a main effect on the decay time of bread, and the factors have an interaction effect.

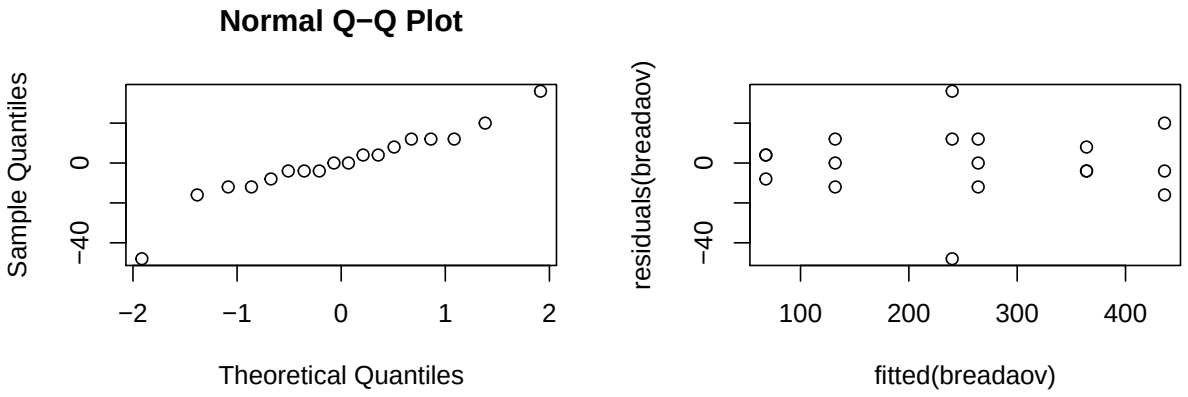
```
## Analysis of Variance Table
##
```

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

d) According to the means of squares, on average the environment has the largest effect on the decay. However, this can not easily be concluded as it is being compared to one base (the first category), instead of a more comprehensive analysis.

```
##                                      Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                               364   12.00000  30.333333 1.032769e-12
## humiditywet                                72   16.97056   4.242641 1.142103e-03
## environmentintermediate                  -124   16.97056  -7.306770 9.389760e-06
## environmentwarm                          -100   16.97056  -5.892557 7.336887e-05
## humiditywet:environmentintermediate      -180   24.00000  -7.500000 7.233671e-06
## humiditywet:environmentwarm              -268   24.00000 -11.166667 1.073751e-07
```

e) The first requirement is that for each unique category, there should be at least 2 samples, which is the case. Then, the most important requirement is that the data among the factors should approximately have equal variances. This has been tested in b) and the conclusion was that they approximately were the same. A different test we can do after the ANOVA test, is check whether the error is normally distributed, which can be expected from a random variable. In the QQ-plot, it can be seen that the residuals are approximately normally distributed. In the fitted residuals plot, it can be seen that the spread is approximately horizontally symmetric among the fitted values, however, there are 2 outliers in the middle.



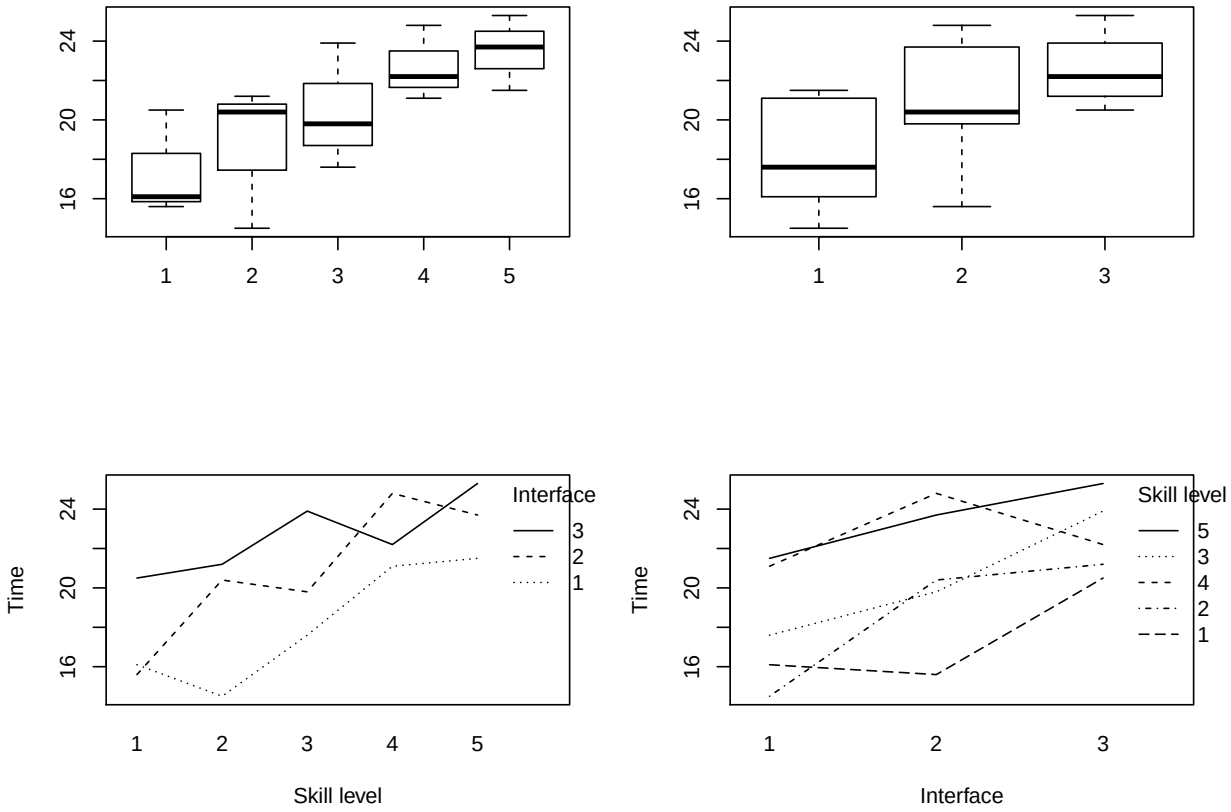
Exercise 2

a) The following code generates a random block design with five blocks, a factor with three levels and one sample per unique category.

```
B=5;
if1 = sample(1:5)
if2 = sample(6:10)
```

```r
if3 = sample(11:15)
for (i in 1:B) print(c(if1[i], if2[i], if3[i]))
```

b) The boxplots below suggest that indeed the skill level and the interfaces matter for the search time. We see that skill level 1 and interface 1 are the fastest. From the interaction plots below, we observe clear interaction effects. Overall, the factors have the same pattern, namely, all lines start in the lower left corner and end towards the upper right corner. However, they are not perfectly parallel, this can be explained by the small sample sizes which cause local irregularities. Thus, we conclude that there is no interaction between the two factors.



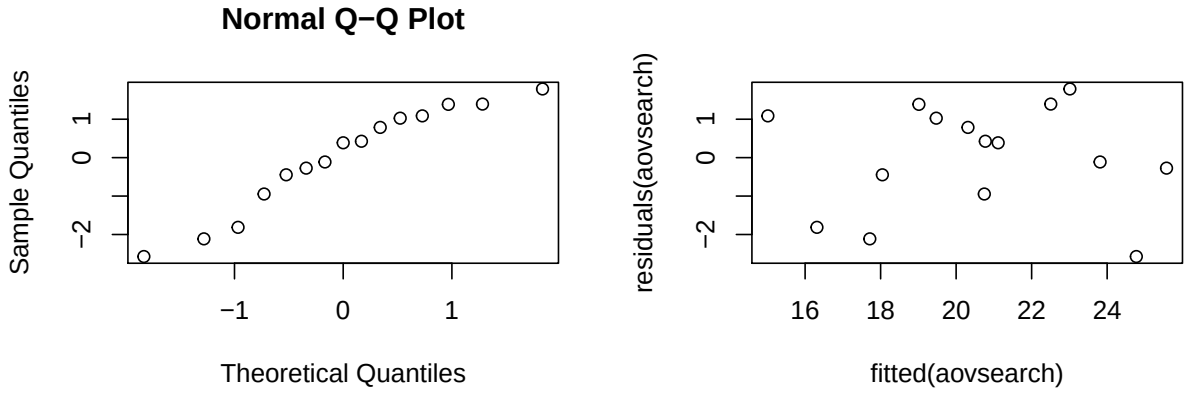
c) H_0 : search time is the same for all interfaces. From the ANOVA results below, it can be concluded that H_0 is rejected. This means that the search time is not the same for all interfaces. Furthermore, we can estimate the time it takes for a user with skill level 3 to find a product using interface 2 by looking at the summary table and adding the coefficients of these two categories to the intercept. In this case, that would be $15.015 + 3.033 + 2.7 = 20.748$.

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value  Pr(>F)
## interface  2 50.465 25.2327  7.8237 0.01310 *
## skill      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

##
## Call:
## lm(formula = time ~ interface + skill, data = search)
##
## 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 ***
## interface2     2.700      1.136   2.377  0.04474 *
## interface3     4.460      1.136   3.927  0.00438 **
## skill2         1.300      1.466   0.887  0.40118
## skill3         3.033      1.466   2.069  0.07238 .
## skill4         5.300      1.466   3.614  0.00684 **
## skill5         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
```

d) The QQ-plot of the residuals below looks normally distributed, which is good. The fitted residuals do not depict any outliers.



e) The result of the Friedman test is the same as the ANOVA test: we reject the H_0 mentioned before, thus, there is a difference in search times.

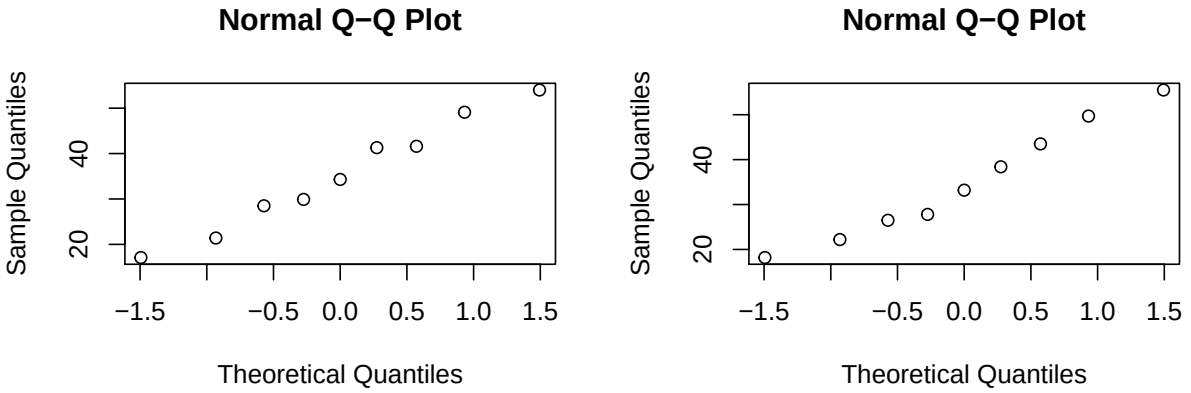
```
##
##  Friedman rank sum test
##
## data:  search$time, search$interface and search$skill
## Friedman chi-squared = 6.4, df = 2, p-value = 0.04076
```

f) The one-way ANOVA returns no significant difference in the search time between the interfaces. This result is not very useful, because (1) we removed a lot of information from the model and (2) the model now assumes that the block is a random selection of all available blocks, which is not the case because the blocks were fixed/predetermined.

```
## Analysis of Variance Table
##
## Response: time
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## interface  2  50.465  25.233  2.8605 0.09642 .
## Residuals 12 105.852   8.821
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 3

a) First, we evaluate normality to determine which test to use. From the QQ-plots below, we conclude that both treatment samples are normally distributed. There are three hypotheses: (1) H_0 (id) : there is no difference in milk production between cows, (2) H_0 (per) : there is no difference in milk production in different periods and (3) H_0 (treatment) : there is no difference in milk production with different treatment. From the ANOVA results below, we can conclude that within-cow variation (see variable "id") the milk production differs. Because the p-value for id is less than the significance level of 0.05, therefore, the first H_0 is rejected. Furthermore, from the summary, we can conclude that most of the cows (except id4) are different from the cow with id1. Afterwards, we could see p-value for per is less than 0.05, so we reject the second H_0 which means that whether a cow is going through the first period or second seems to make a difference. Furthermore, as the p-value of treatment is equal to 0.51654, we do not reject the third H_0. This means that treatment A does not significantly differ from treatment B. This could be seen from the second table (treatment B). The summary below indicates that there is no significant difference in milk production. However, it is important to note that this is not the appropriate way of testing the cross-over design.




```
## 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
```

```
##
## Call:
## lm(formula = milk ~ id + per + treatment, data = cow)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2600 -0.4375  0.0000  0.4375  2.2600
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.3000     1.2444  24.349 5.02e-08 ***
## id2          23.0000     1.5741  14.612 1.68e-06 ***
## id3          11.1500     1.5741   7.084 0.000196 ***
## id4          -1.3500     1.5741  -0.858 0.419480
## id5          -7.0500     1.5741  -4.479 0.002870 **
## id6          23.4500     1.5741  14.898 1.47e-06 ***
## id7          13.5500     1.5741   8.608 5.69e-05 ***
## id8           4.9000     1.5741   3.113 0.017011 *
## id9         -11.2000     1.5741  -7.115 0.000191 ***
## per2         -2.3900     0.7466  -3.201 0.015046 *
## treatmentB   -0.5100     0.7466  -0.683 0.516536
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.574 on 7 degrees of freedom
## Multiple R-squared:  0.9931, Adjusted R-squared:  0.9832
## F-statistic: 100.6 on 10 and 7 DF,  p-value: 1.349e-06
```

b) In this exercise, we model the cows effect as a random effect by using the function lmer.

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

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: milk ~ treatment + order + per + (1 | id)
##    Data: cow
##
##      AIC      BIC   logLik deviance df.resid
##    119.3    124.7    -53.7    107.3       12
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.53112 -0.37104  0.02686  0.26748  1.72489
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 133.145  11.539
##  Residual               1.927   1.388
## Number of obs: 18, groups:  id, 9
##
```

```
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  38.5000     5.8110   6.625
## treatmentB   -0.5100     0.6585  -0.775
## orderBA      -3.4700     7.7685  -0.447
## per2         -2.3900     0.6585  -3.630
##
## Correlation of Fixed Effects:
##            (Intr) trtmnB ordrBA
## treatmentB -0.063
## orderBA    -0.743  0.000
## per2       -0.063  0.111  0.000
```

Based on the p-value below, we do not reject H_0 for treatment. This means that treatment is not important. The result is the same as the result a).

```
## Data: cow
## Models:
## cowlmerTreatment: milk ~ order + per + (1 | id)
## cowlmer: milk ~ treatment + order + per + (1 | id)
##                   Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## cowlmerTreatment  5 117.89 122.34 -53.946   107.89
## cowlmer           6 119.31 124.65 -53.656   107.31 0.5807      1      0.446
```

Based on the p-value below, we do not reject H_0 : there is no difference in milk production if the order differs. Therefore, the order of treatment AB is not important.

```
## Data: cow
## Models:
## cowlmerOrder: milk ~ treatment + per + (1 | id)
## cowlmer: milk ~ treatment + order + per + (1 | id)
##               Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## cowlmerOrder  5 117.51 121.96 -53.755   107.51
## cowlmer       6 119.31 124.65 -53.656   107.31 0.1973      1     0.6569
```

Based on the p-value below, we reject H_0 for per, which means whether a cow is going through the first treatment or second is important.

```
## Data: cow
## Models:
## cowlmerPer: milk ~ treatment + order + (1 | id)
## cowlmer: milk ~ treatment + order + per + (1 | id)
##            Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## cowlmerPer  5 125.43 129.88 -57.714   115.43
## cowlmer     6 119.31 124.65 -53.656   107.31 8.1151      1    0.00439 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c) From the result below we conclude that the p-value is equal to 0.8281 and we do not reject H_0 for treatment. This means that the treatment is not important. From previous analysis, the same conclusion was reached. Given the design, it is inappropriate to use the paired t-test. Since the previous analysis shows that factors such as per have a significant effect on the milk production, it might be unwise to ignore such factors.

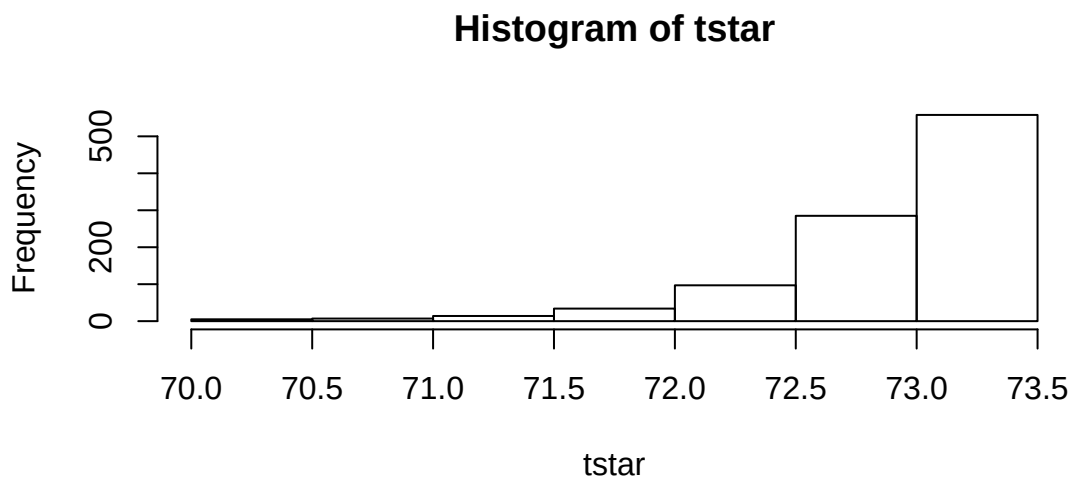
```
##
##  Paired t-test
##
## data:  milk[treatment == "A"] and milk[treatment == "B"]
## t = 0.22437, df = 8, p-value = 0.8281
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.267910  2.756799
## sample estimates:
## mean of the differences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

Exercise 4

a) We observe that the result is the same as the dataset "nauseatable.txt", so the dataframe we created is correct.

b) We observe that the p-value is 0.041, this is less than 0.05. Therefore, H_0 should be rejected which means that the medicine does not work equally well against nausea. By performing the chi-squared test, we get the result 6.62. We could use this to calculate the p-value. By performing the permutation test, we obtain the result 71.82. This test uses different values to calculate the final p-value.

```
## The following objects are masked _by_ .GlobalEnv:
##
##     label_medic, nausea
```



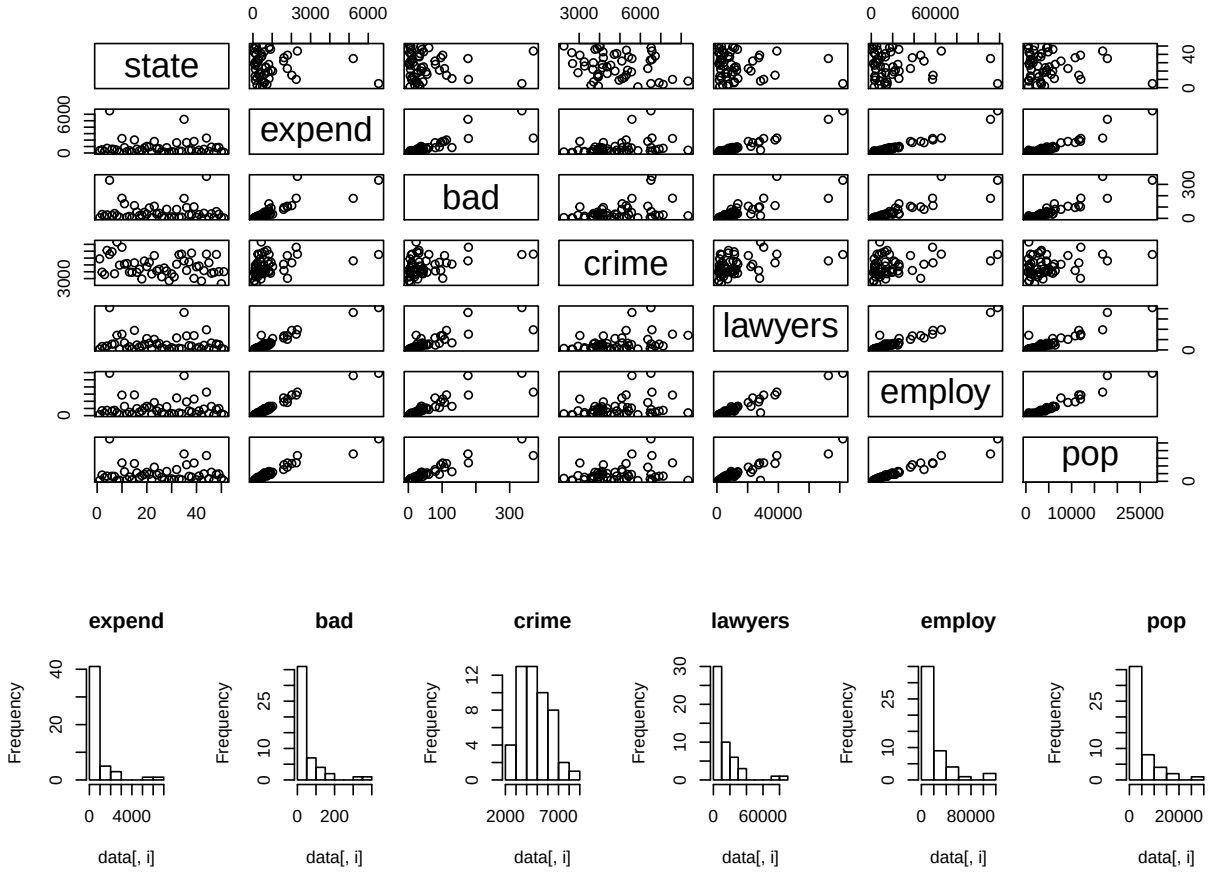
c) We could see from chi-squared test that the p-value is equal to 0.03643, this is smaller than 0.05, therefore, H_0 should be rejected. The p-values of both tests are close to each other. In conclusion, they use different methods but the results are the same.

```
##
##  Pearson's Chi-squared test
##
## data:  xtabs(~newNausea$medicine + newNausea$nausea)
## X-squared = 6.6248, df = 2, p-value = 0.03643
```

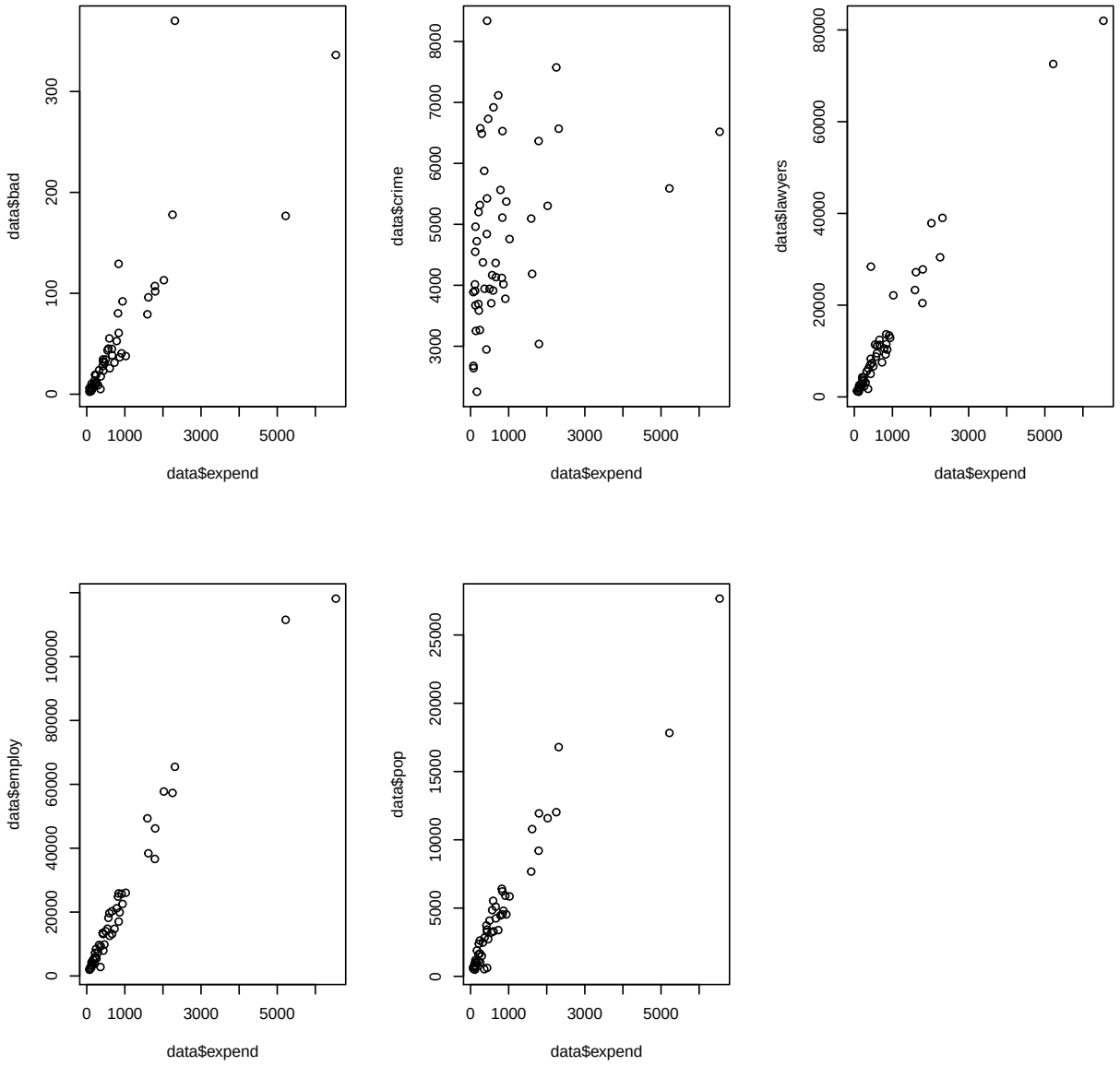
Exercise 5

In our regression analysis, the response variable is "expend" and the explanatory variables are: "bad, crime, lawyers, employ and pop". The purpose is to explain expend by a numerical function of the explanatory variables.

a) First, we make a graphical summary of the data by plotting each variable against the others. Furthermore, we construct a histogram of all the numerical variables. Looking at the plots, we observe that expend, lawyers, employ and pop all approximate a linear relationship with each other. Furthermore, state and crime have nonlinear relationhsips with all the other variables. Lastly, the variable bad can be argued to have a weak linear relationship with the variables expend, lawyers, employ and pop. Looking at the histograms, it is interesting to see that almost all variables (expend, bad, lawyers, employ and pop) follow a similar pattern, namely, the lowest value appears frequently and as the value increases, the frequency decreases steeply. Except for a few outliers of frequently occuring high values. In contrast, the variable crime shows a different pattern. Namely, the values in the middle occur also relatively frequently. But the rule: as the value increases, the frequency decreases, applies as well.



To build an intuition of the linear relationship between the depandable variable (expand) and its explanatory variable, we have plotted the explanatory variables against the dependaple variable below. These are the same plots as from the first plot with all factors plotted against each other, but now we can take a closer look. From the plots we can see a strong linear relationship between the dependable expend and 4 of the five factors: bad, lawyers, employ and pop. However we see very big outliers, which will have to be adjusted for, otherwise they will have a very big weight in the factor coefficient from the linear regression.



A potential point is an outlier in an explanatory variable. The effect can be studied by fitting the model with and without the potential point. If the estimated parameters change drastically when removing the potenetial point, the observation is called an influence point. Using the Cook's formula, the distance of an observation on the predictions can be calculated. Whenever the Cook's distance for an observation approximates or is larger than 1, the observation can be considered to be an influence point. As we have not constructed a model yet, we analyse the potential and influence points of our chosen model in c). Another relevent concept is collinearity. This is the problem of linear relations between explanatory variables. Collinearity can be detected by a straight line in a scatter plot or by calculating the correlation coefficient. Looking at the scatter plots of the data, we suspect collinearity between the variables expend, lawyers, employ and pop. We confirm this by calculating the correlaction coeffiecients of all possible variable combinations. Looking at the output below, we observe that all the combinations of the variables expend, lawyers, employ and pop have a correlation coefficient above 93. Thus, we conclude that these variables have a collinear relation. The variable bad has a weaker collinear relation with the variables expend, lawyers, employ and pop, namely, ranging from 0.83 to 0.93. Lastly, the variable crime has no collinear relation with

any of the other variables. When collinearity is detected among variables, we should avoid having both explanatory variables in the model.

```
##         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
```

b) To fit a linear regression model to the data, first, we start with the step-up method. Using this method, we start by fitting all possible simple linear regression models and calculate the determination coefficient (R^2). The results are shown in the table below. Looking at this table, we observe that employ has the largerst value of R^2 (0.954) and is thus selected. Next, we combine this variable with all the variables that do not have a collinear relation with employ. These are the explanatory variables bad and crime. Note that the variable bad still can be considered to be linearly correlated to employ. Adding the variables bad and crime to the models yields in $R^2 = 0.9551$. This is an improvement compared to the previous model. Therefore, we continue to add the other vairables to the model. For both models, there is just one variable to add. This results in the last possible option: a model of employ, bad and crime combined. This result in $R^2 = 0.9568$, the highest value so far. As there are no more variables to add, the method stops here. The resulting model is: expend = -2.857e^{+02} + 4.979e^{-02} * employ - 1.391e^{+00} * bad + 3.810e^{-02} * crime + error. We have to be carefull as it could be argued that the variables employ and bad are collinear. Therefore, the model that is constructed with the variables employ and crime (expend = -2.484e^{+02} + 4.630e^{-02} * employ + 2.962e^{-02} * crime + error) might be a better model as it contains fewer variables and the value of R^2 is similar.

Explantory Variable(s)	bad	crime	lawyers	employ	pop	bad, employ	crime, employ	bad, crime, employ
Multiple R-squared	0.6964	0.1119	0.9373	0.954	0.9073	0.9551	0.9551	0.9568

Second, we use the step-down method. This method start with fitting all explanatory variables in the so-called full model. In each iteration, one explanatory variable is removed. This time, we try the model with all variables, regardless of collinearity. In round 1, we observe that the variable crime has the highest p-value, $0.25534 > 0.05$, therefore, the variable crime will be removed. In round 2, pop has the highest p-value, $0.06012 > 0.05$, therefore, the variable pop will be removed. In round 3, bad has the highest p-value, $0.34496 > 0.05$, therefore, the variable bad will be removed. In round 4, lawyers has the highest p-value, $0.00113 < 0.05$, therefore, the variable will not be removed and the method stops. This results in the model expend = -1.107e^{+02} + 2.686e^{-02} * lawyers + 2.971e^{-02} * employ + error.

Round 1: expend ~ bad, crime, lawyers, employ, pop

Explantory Variables	bad	crime	lawyers	employ	pop
p-value	0.02719	0.25534	0.00592	0.00354	0.03184

Round 2: expend ~ bad, lawyers, employ, pop

Explantory Variables	bad	lawyers	employ	pop
p-value	0.05402	0.00106	0.00380	0.06012

Round 3: expend ~ bad, lawyers, employ

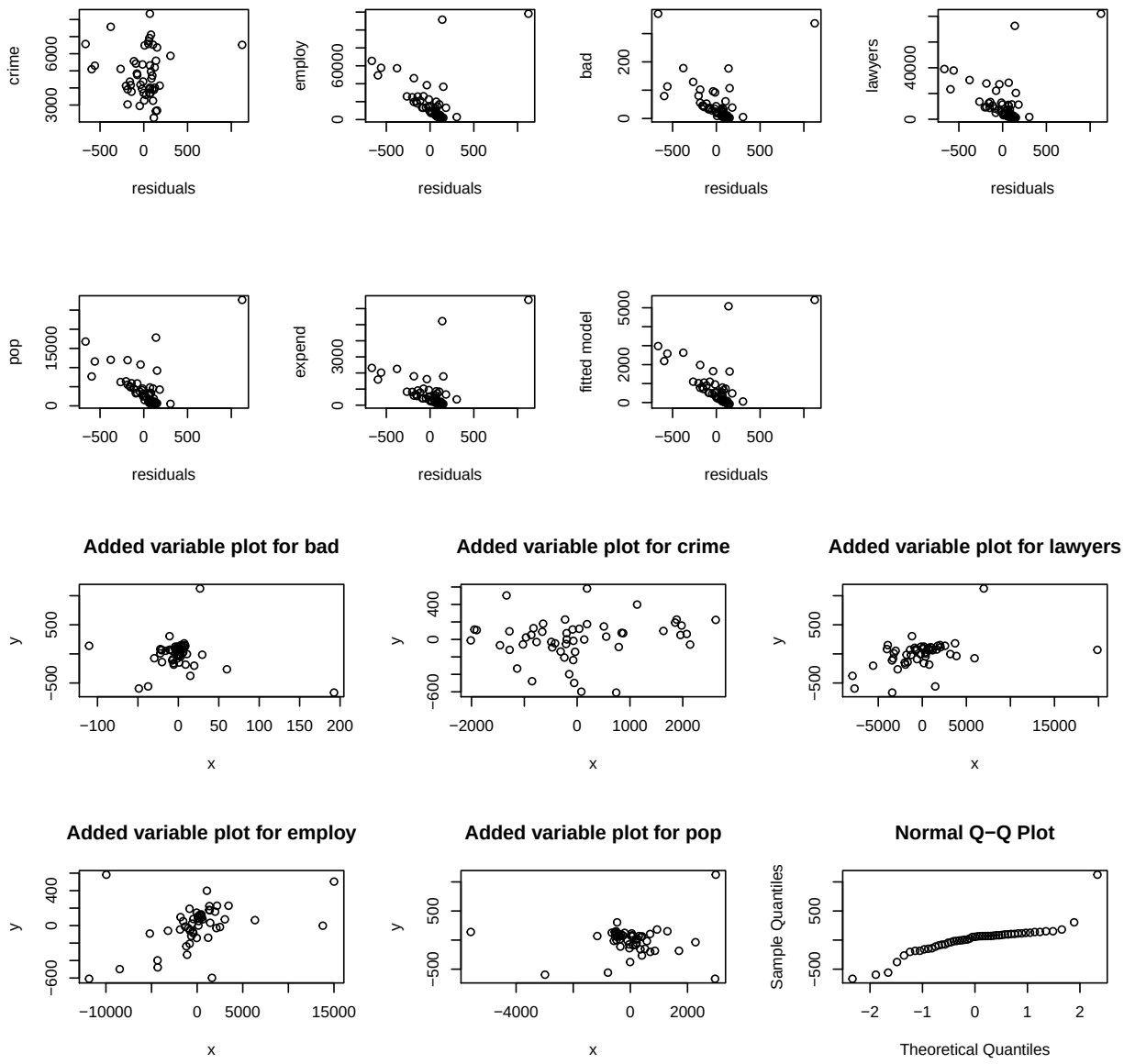
Explantory Variables	bad	lawyers	employ
p-value	0.34496	0.00147	$1.2e^{-06}$

Round 4: expend ~ lawyers, employ

Explantory Variables	lawyers	employ
p-value	0.00113	$4.89e^{-07}$

Using the step-up and step-down method resulted in two different models. The advantage of the model constructed using the step-up method (expend ~ employ bad crime) is that the variables are not collinear. The advantages of the model constructed using the step-down method (expend ~ lawyers employ) are that the value of R^2 (0.9632) is higher compared to the R^2 value of the step-up model (0.9568) and the model contains fewer explanatory variables. However, as the collinearity of variables weight higher compared to number of variables and the difference of R^2 is relatively small, we prefer the model constructed using the step-up model. We even consider removing the variable bad of the step-up model as it can be argued to have a collinear relation with the variable employ.

c) We check the model (expend ~ employ crime) assumptions (linearity of the relation and normality of the errors) using both graphical and numerical tools. First, we look at the scatter plot of the response variable against each explanatory variable separately. This is visualized in a). For each combination with expend, we see two outliers at the right of the scatter plots. Moreover, the relation with the variables bad, lawyers, employ and pop is linear and nonlinear combined with state and crime. Second, we construct the scatter plot of the residuals against each explanatory variable that is in the model (crime and employ) seperately. The plot with crime contains a cluster around the middle line and the plot with employ has a cluster in the bottom left diagnal with two outliers in the upper right corner. This means... Third, we construct the added variable plot. In this plot, the residuals of the explanatory variables are plotted against the residuals of the model without that specific variable. This shows the effect of adding an explanatory variable to the model. Looking at the figure, we observe that the plots for bad, lawyers and pop contain compact clusters. The plots of crime and employ are more spread. This means... Next, we construct the scatter plots of the residuals against each explanatory variable that is not in the model (bad, lawyers and pop) seperatly. The outputs are similar to each other. When looking at the pattern we do not observe a linear relation and therefore should not include more variables. Afterwards, we construct the scatter plot of the reduals against the response variable and the fitted model. We observe little spread as in both plots there is a cluster around zero with only few outliers. This means... Lastly, we check the normality assumption by constructing a qq-plot and conduction Shapiro-Wilk's test. We cannot assume normality as the qq-plot does not approximate a straight line, this means that the model is invalid. Unfortunetely, none of the other models resulted in a normal distribution of the residuals.


crime
employ
bad
lawyers
pop
expend
fitted model
residuals
Added variable plot for bad
Added variable plot for crime
Added variable plot for lawyers
Added variable plot for employ
Added variable plot for pop
Normal Q−Q Plot
x
y
Sample Quantiles
Theoretical Quantiles

Appendix