EDDA assignment 2

Mick IJzer, Tirza IJpma, Maud van den Berg

March 10, 2020

Exercise 1

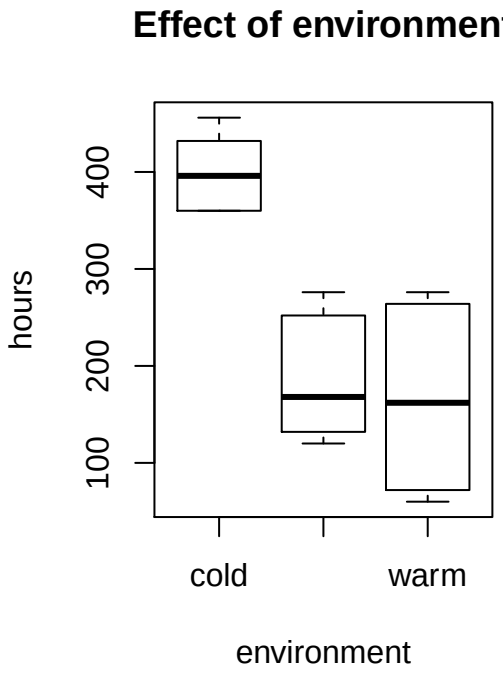
a)
To randomize 18 slices of loaf into 6 different combinations (3 temperatures, 2 humidities) the following code in R can be used:

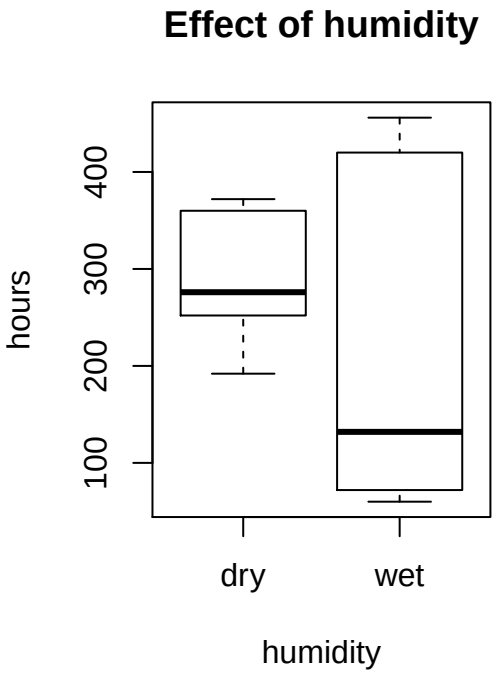
```
temp=3; hum=2; N=3
rbind(rep(1:temp,each=N*hum),rep(1:hum,N*temp),sample(1:(N*temp*hum)))
```

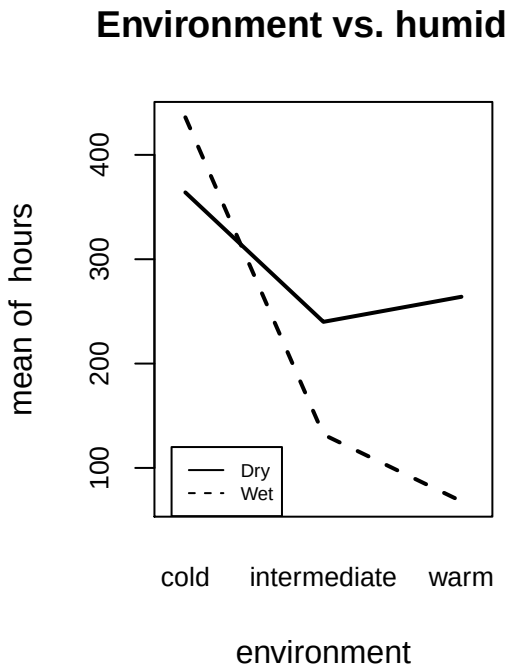
```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## [1,]    1    1    1    1    1    1    2    2    2     2     2     2     3     3
## [2,]    1    2    1    2    1    2    1    2    1     2     1     2     1     2
## [3,]    9   14   13   18    3    8   11    4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

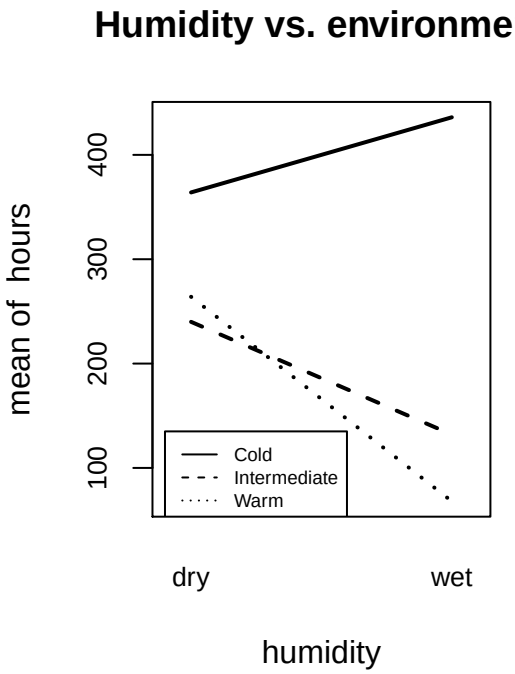
To interpret this table: each column can be seen as a different unit(where the id of the unit is the value in the third row), the first row can be seen as which temperature the unit has to be measured in and the second row can be seen as the humidity group of the unit.

b)









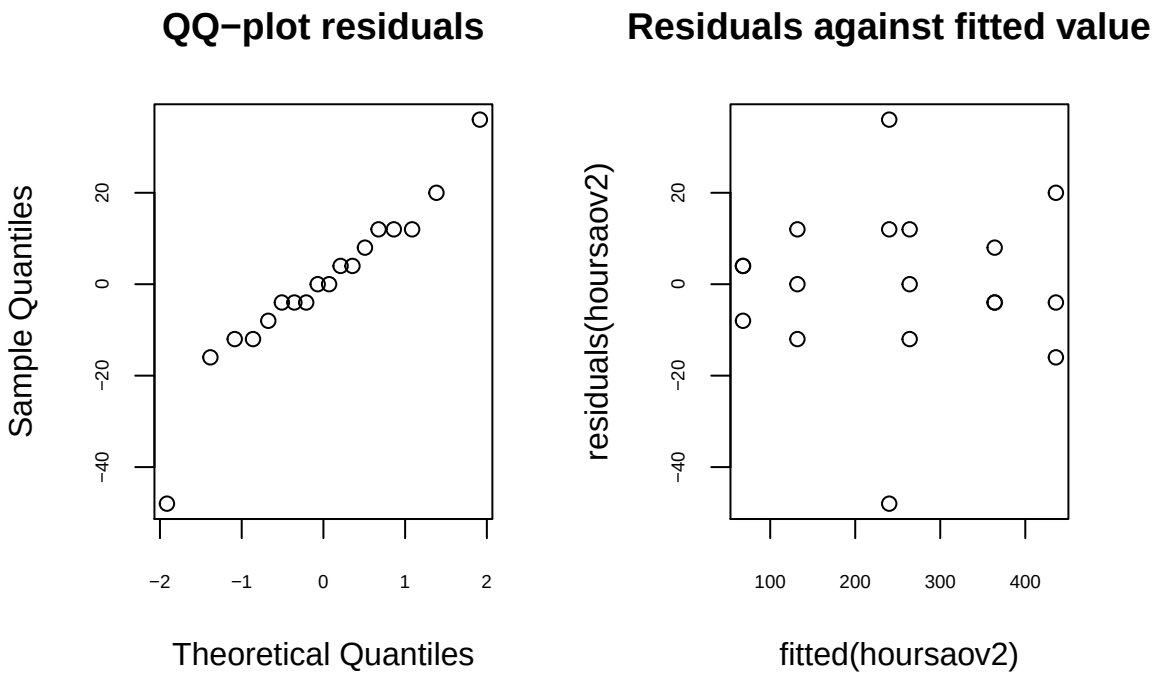
c)
After performing a one way ANOVA test, the following results were obtained: the factor environment has a significant main effect on the time to decay, $F = 233.69$, $p < .000$. Also the type of humidity has a significant

main effect on the time to decay, F = 62.30, p < .000. There is a significant effect for the interaction of the factors on the time to decay as well, F = 64.80, p < .000. To interpret these results, the assumptions of the ANOVA are considered to be satisfied. In question e the ANOVA assumptions are checked.

When inspecting the graphical representation of the interaction effect, the interaction of humidity vs. environment (right graph) shows the differences clearly. It can be seen that for both the intermediate and warm environment the mean of hours decreases when going from a dry to a wet humidity. For a cold environment this effect has the opposite direction, namely an increase in mean of hours when the humidity changes from dry to wet.

d)
When interpreting the results of the ANOVA, it can be seen that environment probably(?) has the most influence on the hours to decay. This can be known through the bigger F-value of the ANOVA as you can see in question c. Which is explained by a higher value for the explained mean of squares for environment (100952) than humidity (26912). This is not a good question, because there is a significant effect for the interaction between the two factors. So the effect on one factor depends on the value of the other factor. If there was no significant interaction, this would have been a good question.

e)
To check the normality and the assumption of equal variances of the ANOVA, a QQ-plot of the residuals is computed. In addition also the fitted values are plotted:



The residuals (the data corrected for the different population means) seem to be normally distributed. Performing the Shapiro-Wilk test confirms this (W = 0.930, p-value = 0.191). Also in the fitted data it can be seen that the spread in the residuals doesn't change systematically when the number of hours increases. Since the 18 bread units are independently measured, it can be stated that the assumptions for the ANOVA are satisfied.

There seem to be two outliers with the predicted value of approximately 240 hours. The actual value is very different from the predicted value. This is probably caused by the large variations in hours to decay in the intermediate-dry condition.

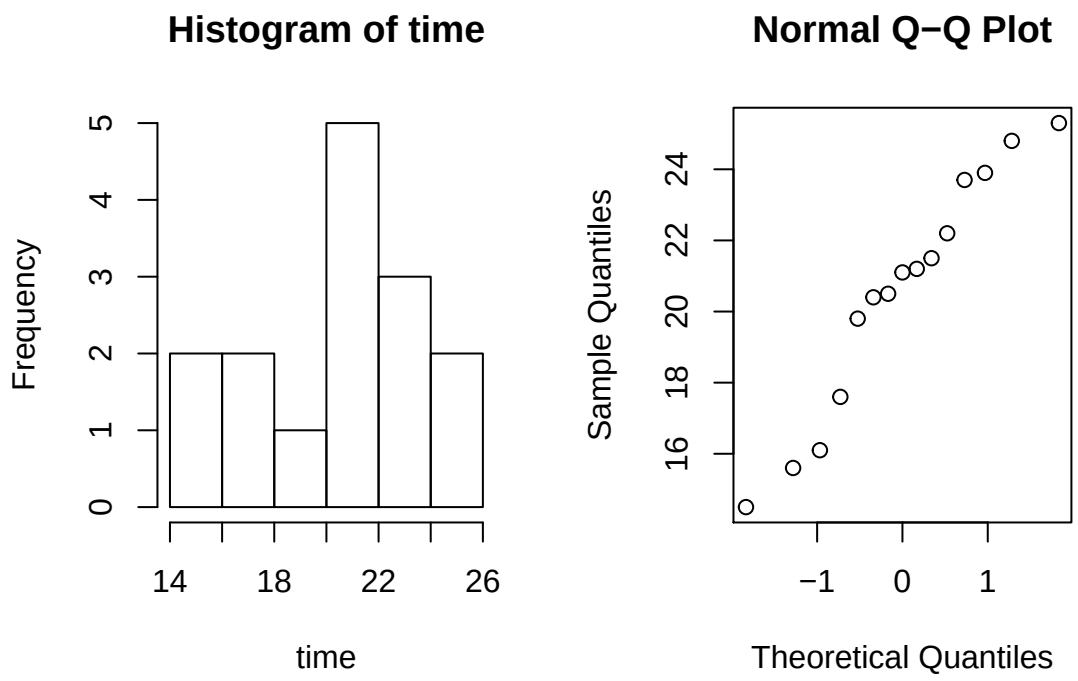
Exercise 2

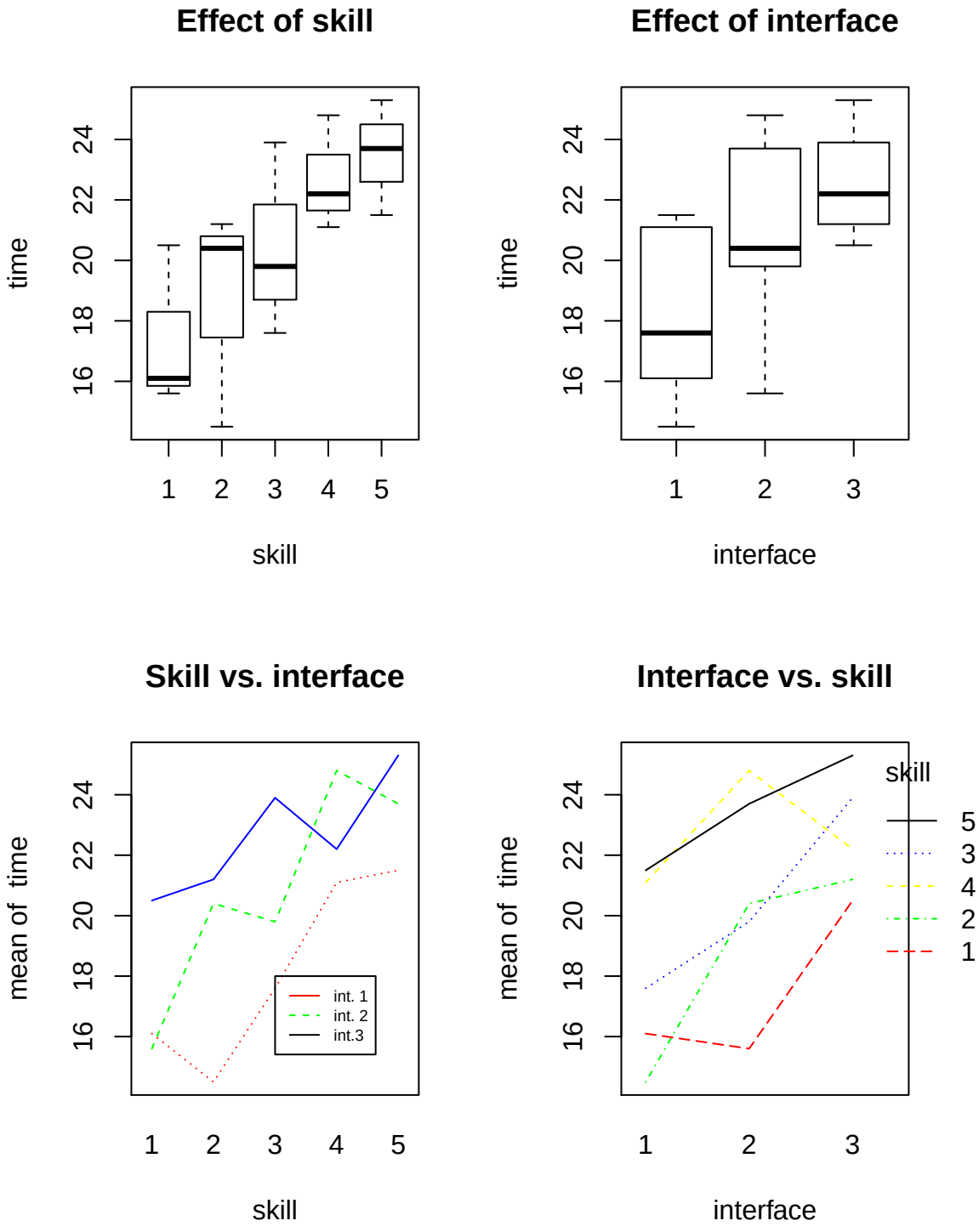
a)
A randomizied block design was used to distribute the 15 students over the 3 possible interfaces. Because both the student ID and the skill level are fixed, only the interface had to be evenly distributed over the students. The code below shows that step. There are three students in each skill-level category, and since there are 3 possible interfaces, every one of those three students should be assigned a different interface. In the output the randomized assignment of students to each interface is shown. The first row indicates the interface, the second row shows the skill level of the students. And the third row represents the ids of the students. Each column corresponds to a student.

```
rbind(c(replicate(5, sample(1:3))),rep(1:5,each=3), rep(1:15))
```

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

b)





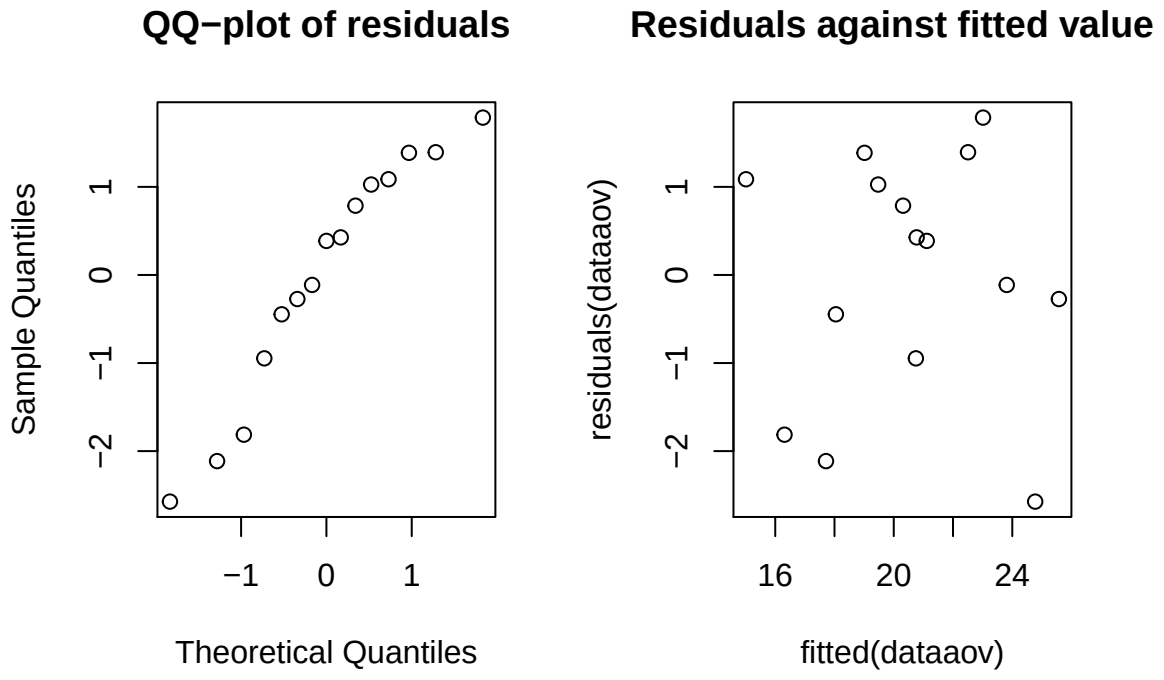
First the distribution of the dependent variable is shown in a histogram and QQ-plot. The data looks normally distributed. Next two boxplots are created to visualize the effect that the independent variables skill and interface have on the time required to finish the task. Overall the students with a lower skill level (better skills) seem to be faster than students with a high skill level. Furthermore, the use of interface 1 seems to result in lower times. Lastly two interaction plots are created. These show that there are no

obvious interaction effects. However, the previous statements are made based on the visualization of the data. No test have been used to underline these statements. Besides, any possible interaction effects cannot be reliably tested, because there is only one observation for each combination of factors.

c)
To test whether interface has a main effect on the time it takes to complete the task, an ANOVA was carried out with both interface and skill as independent variables. No interaction effect was taken into account. Results show that both interface, $F(2,8)=7.82$, $p = 0.013$, and skill, $F(4,8)=6.21$, $p = 0.014$, have a significant main effect. Therefore the search time is not equal for all of the interfaces. To compute the estimated time it takes for a user with skill level 3 who uses interface 2 the summary of the ANOVA was used. This results in an estimated time of (0.5467 + 0.3133 + -0.1133) 20.7467.

```
## 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
```

d)
The model assumptions were checked by looking at both the normality of the residuals and a plot of residuals values against the fitted values. The residuals seem to be normally distributed. This was confirmed by a Shapiro-Wilk normality test, $W=0.93$, $p = 0.282$. The plot of the residuals values against the fitted values shows that there is no systematic error.



e)
A non-parametric Friedman test was carried out to test whether there is an effect of interface. Results show that interface has a significant influence on the time it takes to complete the task, $X^2(2)=6.4$, $p = 0.041$.

f)
When carrying out an analysis of variance with only the interface as independent variable, the results indicate that there is no significant effect of interface, F(2,12)=2.86, p = 0.096 on the time it takes to complete the task. It could make sense to carry out this test, however in the current context it is wrong. This is because earlier tests, as well as the visualization of the data indicate a significant effect of skill. Therefore, it is unwise to remove this factor from the analysis. If there was no effect of skill it would not matter if it was used as a factor in the analysis.

Exercise 3

a)
First, it is investigated whether the dependent variable milk production is normally distributed. According to the QQ-plot, histogram and boxplot it is assumed that the data is normally dsitributed. In addition, a Shapiro-Wilk-test is carried out (W = 0.954, p = 0.495), which supports this assumption. To test whether the type of feedingstuff has an effect on the milk production, an ordinary fixed effects model is carried out, with the treatment, order, period and id as independent variables. From this ANOVA it is concluded that there is a difference in milk production between the two treatments of 0.51 liter milk. However, this result is not significant (F = 0.109, p = 0.751), which indicates that there is no effect of type feedingstuff on the milk production.

```
## Analysis of Variance Table
##
## Response: milk
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## treatment  1    0.27    0.27   0.1085  0.751470
## order      1   53.52   53.52  21.5986  0.002349 **
## per        1   25.39   25.39  10.2462  0.015046 *
## id         7 2413.96  344.85 139.1810 5.632e-07 ***
## Residuals  7   17.34    2.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b)
A mixed effects model is carried out, to investigate if the type of feedingstuff has an effect on the milk production. Now, the effect of cow (id) is an 'random effect'. To get a p-value, another ANOVA is carried out, with this model and a model with the factor treatment left out. This resulted in p = 0.446. So, also in in this model is no effect of type feedingstuff on the milk production.

```
## Data: cow
## Models:
## cowaov2: milk ~ order + per + (1 | id)
## cowaov1: milk ~ treatment + order + per + (1 | id)
##         Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## cowaov2  5 117.89 122.34 -53.946   107.89
## cowaov1  6 119.31 124.65 -53.656   107.31 0.5807      1      0.446
```

c)
The t-test resulted in p = 0.828. This shows that treatment does not have a significant effect on the milk production. This conclusion is compatible with the conclusion from question a. However, this t-test is not a valid test, because the independent variables order and period have a significant effect on milk production.

Exercise 4

a)
The dataframe is created by first creating a column with the number of patients with nausea (180) and without (124). Next the column for medicin is created. This is done by repeating the type of medicin a specified number of times. The number of repetitions is equal to the number of patients that took a certain medicin and did not have nausea.

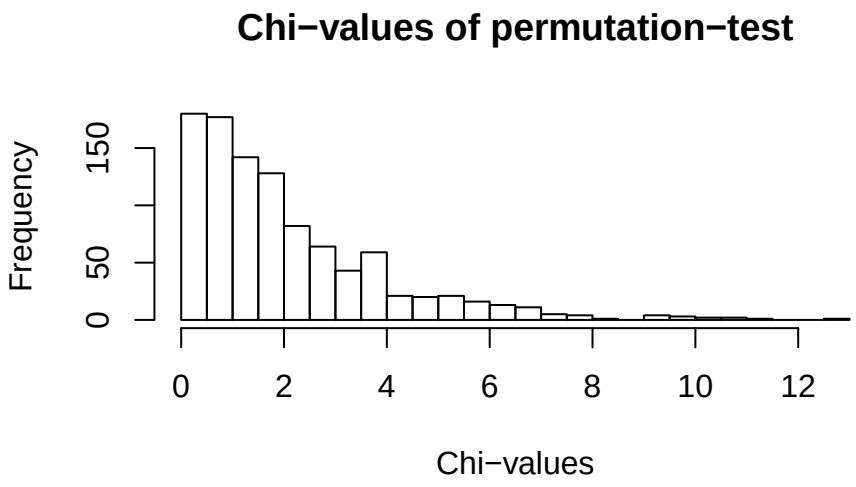
```
nauseadf = data.frame(naus=c(rep('no', each=180), rep('yes', each=124)),
           medicin=c(rep('chlor', each=100), rep('pent100', each=32), rep('pent150', each=48),
                     rep('chlor', each=52), rep('pent100', 35), rep('pent150', 37)))
```

```
##          naus
## medicin    no yes
##   chlor   100  52
##   pent100  32  35
##   pent150  48  37
```

b)
To perform a permutation test, the labels were shuffled 1.000 times. After every shuffle the chi-value was computed and saved. If the null-hypothesis is true, then the chi-value of the original dataset would be a probable outcome. Therefore the chi-value from the original dataset was compared to the chi-values of the 1.000 permutations. The original chi-value was 6.62. Only 3.6% of the permuted chi-values were larger than 6.62. Therefore the null-hypothesis can be rejected, and the conclusion is that medicin and nausea are not independent. So difference medicins don't work equally well against nausea.

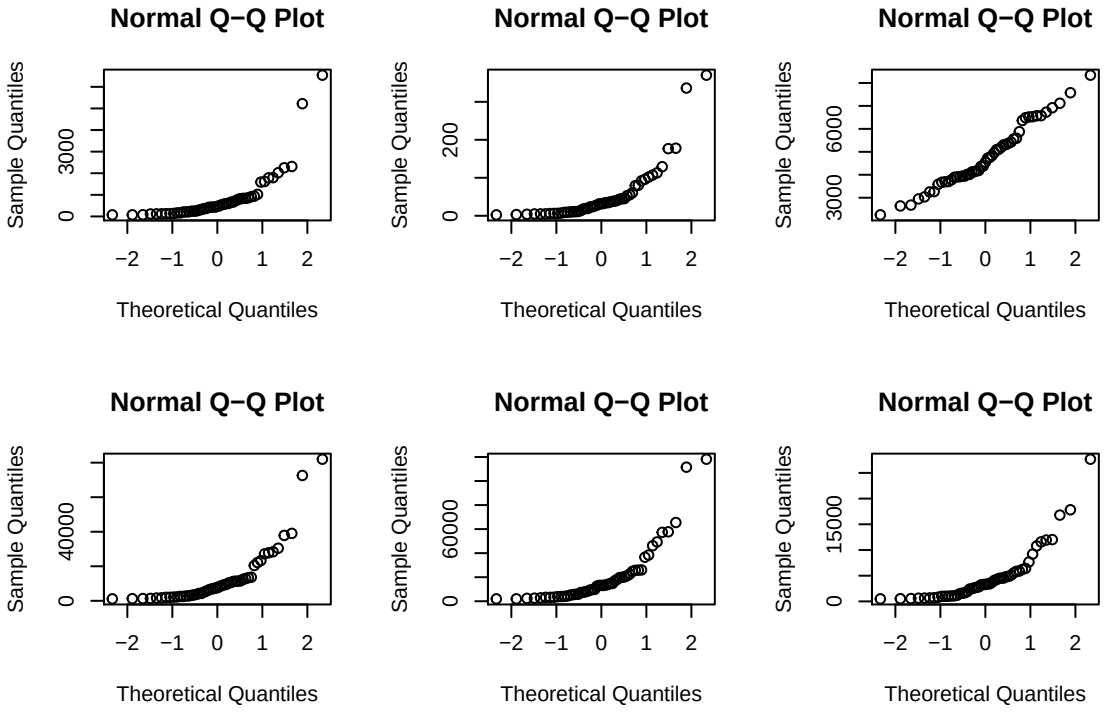
```
B = numeric(1000)
for (i in 1:length(B)){
  nauseadf2 = transform(nauseadf, medicin=sample(medicin))
  B[i] = chisq.test(xtabs(~nauseadf2$medicin+nauseadf2$naus))[[1]]
}
chi_contingency = chisq.test(xtabs(~medicin+naus))[[1]]
p_permutation = mean(B>chi_contingency)
```

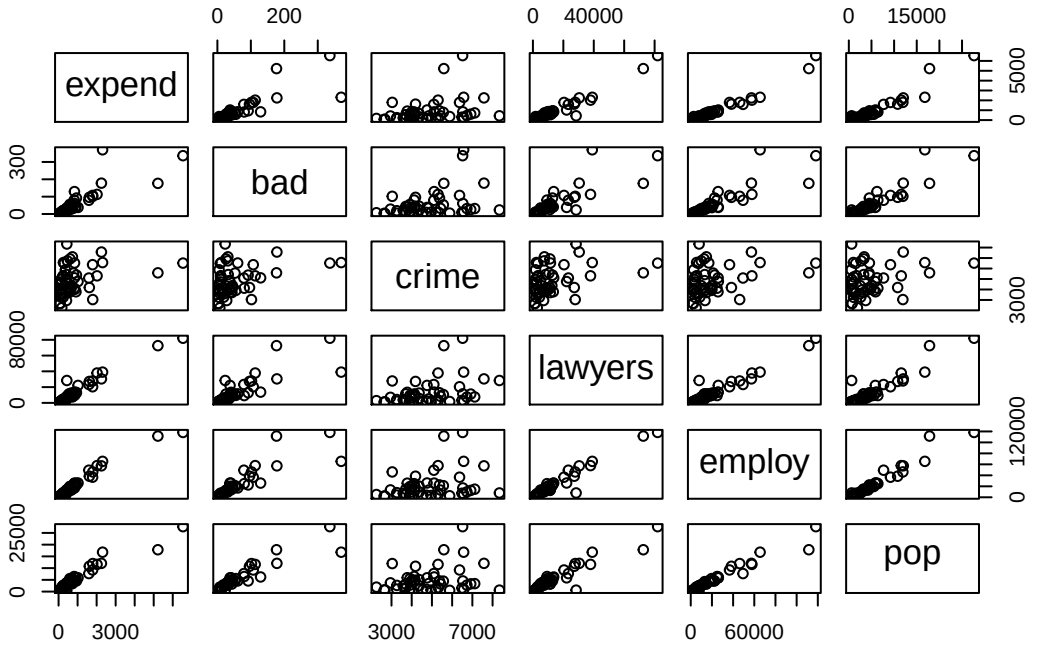
c)
When performing a chi-square test for contingency tables, the outcome is similar. Medicin and reported nausea seem to be related, X^2(2)=6.62, p = 0.036. This makes sense, because the resulting chi-values of the permutations is more or less equal to the true chi-distribution with 2 degrees of freedom. This is illustrated by the histogram. With the permutation test, chi-values are computed under the assumption that the null-hypothesis is true. This results in the actual chi-distribution. Therefore, the p-values from permutation and contingency tables are very similar.



Exercise 5

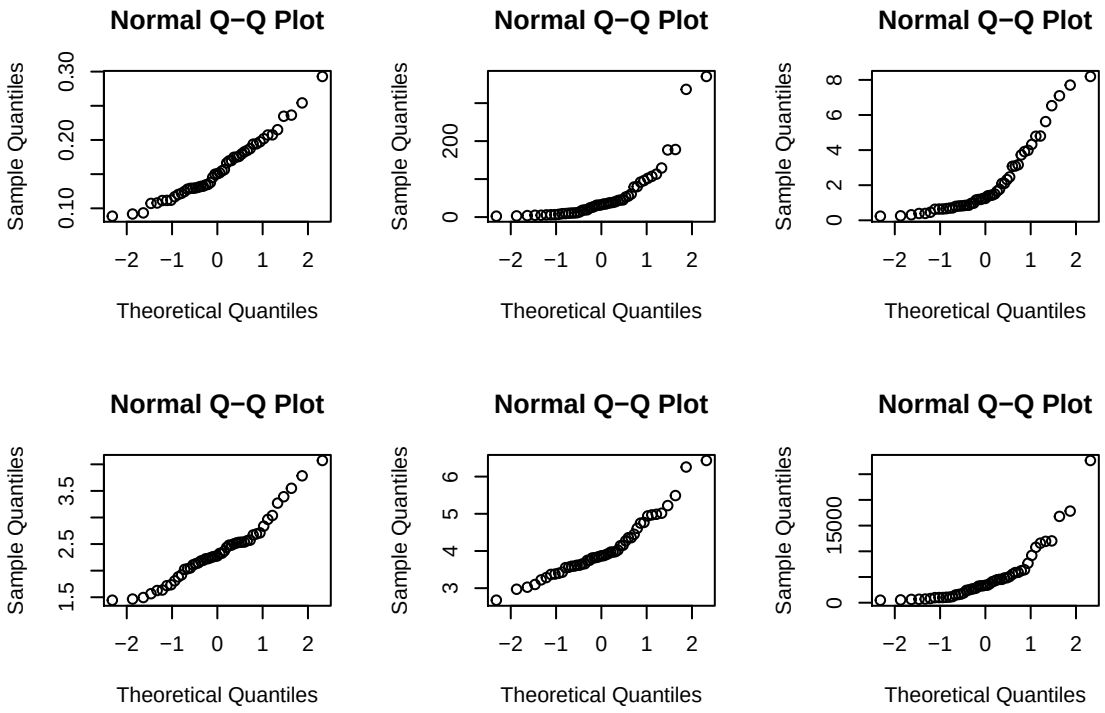
a)
To inspect the distribution of the data and to find outliers, a QQ-plot and histogram (not shown) were computed for all variables. In addition a pairwise plot of all variables is set up and the correlation between all variables is computed, this to check for the collinearity problem.

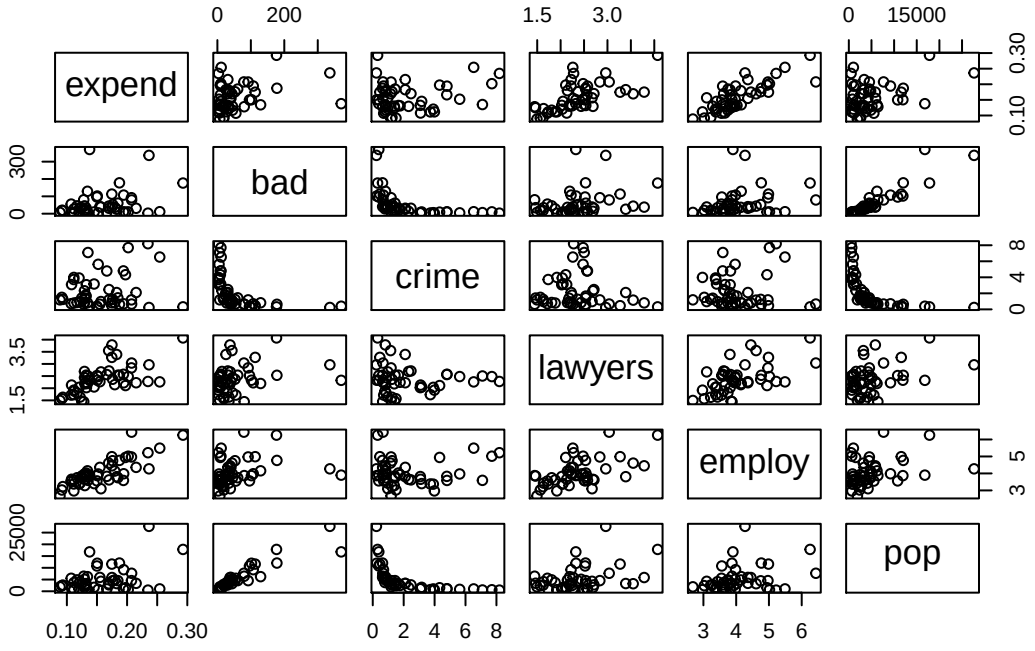





```
##          expend  bad crime lawyers employ  pop
## expend     1.00 0.83  0.33    0.97   0.98 0.95
## bad        0.83 1.00  0.37    0.83   0.87 0.92
## crime      0.33 0.37  1.00    0.38   0.31 0.28
## lawyers    0.97 0.83  0.38    1.00   0.97 0.93
## employ     0.98 0.87  0.31    0.97   1.00 0.97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

As you can see in the QQ-plots and histograms, only the factor crime seems to be normally distributed. There also seems to be an outlier in the data. Many variables seem to be correlated with each other as well, which will result in a problem of collinearity. To fix these problems, the data was further inspected and the outlier (row 8, the data from state DC) was removed from the dataset and the variables expend, crime, lawyers and employ were divided by the population in order to normalize for the population size. The same graphs and correlationvalues were computed with the new data. Results can be seen below. It can be seen that the variables expend, lawyers and employ are distributed more normally. Furthermore, the correlations between variables are lower, which indicated less multicollinearity. This is not the case for variables bad and pop, so these variables will probably not be implemented in the same model.


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−2 −1 0 1 2
0.10 0.20 0.30
0 200
0 2 4 6 8
1.5 2.5 3.5
3 4 5 6
0 15000


expend
bad
crime
lawyers
employ
pop
0 200
1.5 3.0
0 15000
0 300
0.10 0.20 0.30
0 2 4 6 8
3 4 5 6
0 25000

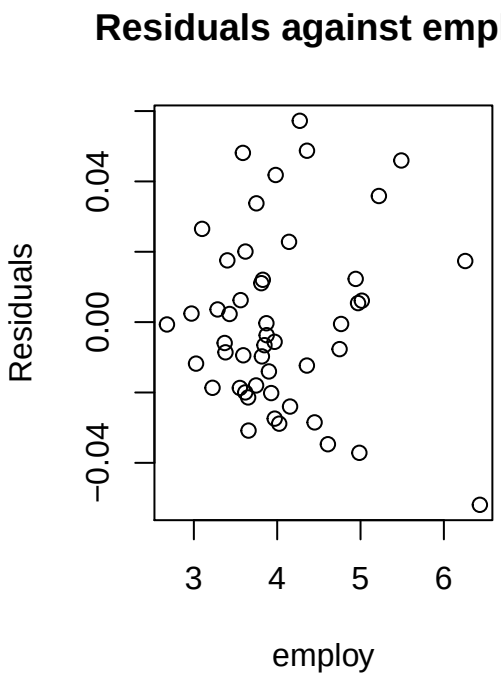
```
##         expend   bad crime lawyers employ   pop
## expend    1.00  0.30  0.18    0.64   0.80  0.37
## bad       0.30  1.00 -0.48    0.29   0.26  0.92
## crime     0.18 -0.48  1.00   -0.09   0.08 -0.57
## lawyers   0.64  0.29 -0.09    1.00   0.57  0.37
## employ    0.80  0.26  0.08    0.57   1.00  0.31
## pop       0.37  0.92 -0.57    0.37   0.31  1.00
```

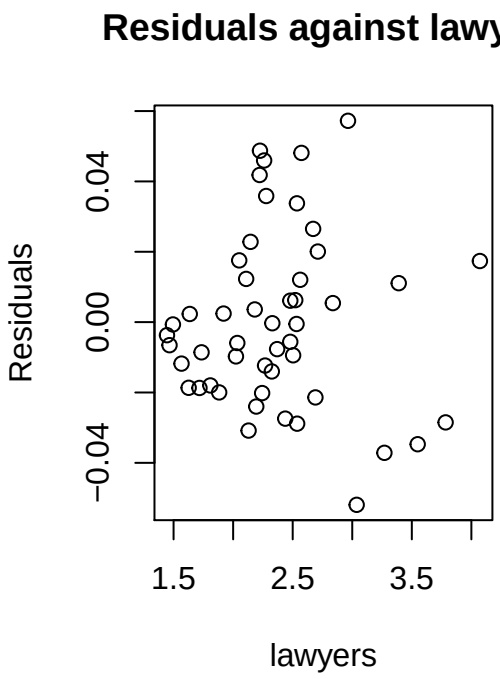
b)
First, the step-up method is used to find the optimal model. R2-values are computed for all independent variables seperately. The variable employ was first added to the model, because it had the highest R2-value amongst the independent variables. This process was repeated multiple times untill there were no signficiant predictors left. The final model included the independent variables employ and lawyers. The R2 for the step-up model was 0.691. Next, the step-down method is used to find an optimal model. All predictors were added to the model, after which the most insignificant predictor is removed. This was done multiple times untill only significant predictors were left in the model. The final model of the step-down method included employ, lawyers, crime and pop. The R2 for this model was 0.766.

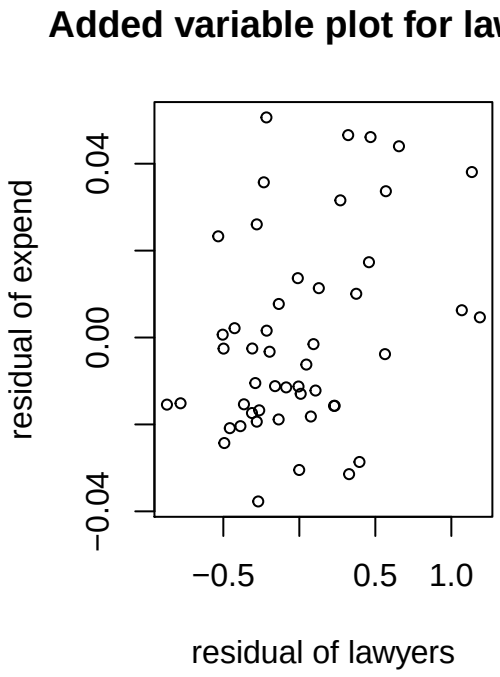
The step-up and step-down method yielded two different models. The step-up model had 2 explanatory variables and 69.1% explained variance. The step-down model had 4 explanatory variables and 76.6% explained variance. The increase in R2 in the step-down model is not big enough to compensate for the extra variables. Therefore, the smallest model is preferred due to simplicity, therefore the step-up model is chosen. The final model with the effects of the predictors is shown below.

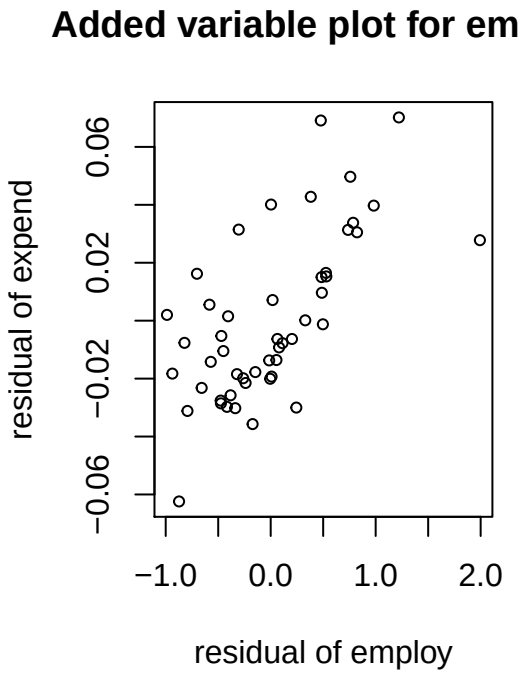
```
##
## Call:
## lm(formula = expend ~ employ + lawyers, data = crimetable2)
##
## Coefficients:
## (Intercept)       employ      lawyers
##    -0.04070      0.03642      0.02168
```

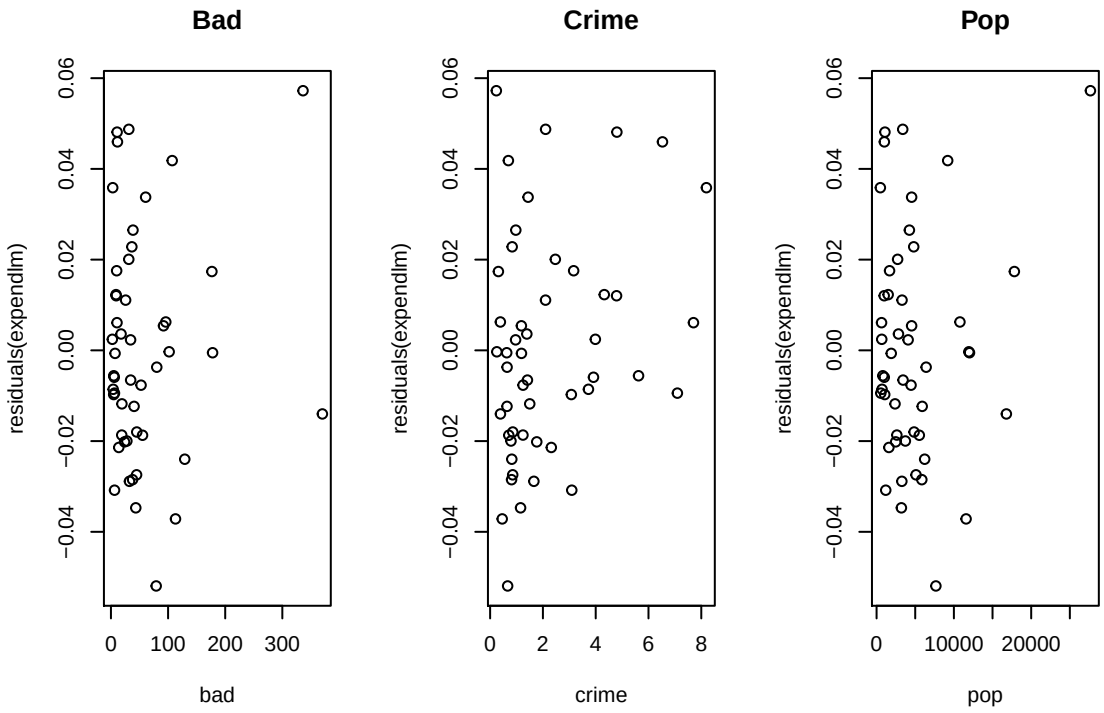
c)
The model assumptions were checked. (1) The plots of the dependent variables against the indpenedent variables are shown in a. No indications of nonlinear relationships between expend and the independent variables are found. (2) The residuals of the model are plotted against the two factors of the model. The residuals do not show a pattern or systematic error. (3) Next partial regression plots are produced where for each predictor a model is created without that predictor. The residuals of that model are plotted against the predictor that was omitted. The slopes in these plots reflect the regression coefficient of the variables that were omitted. (4) The residuals of the model are plotted against the factors that were not included in the model. No linear relations between the residuals and the factors are visible in the plots. Lastly, (5) the QQ-plot of the residuals and (6) a plot of the residuals versus the fitted values are computed. The QQ-plot shows that the residuals are normally distributed, which is confirmed by a Shapiro-Wilk test, W = 0.97, p = .287. No systematic errors or patterns are visible in plot of the residual versus fitted values.

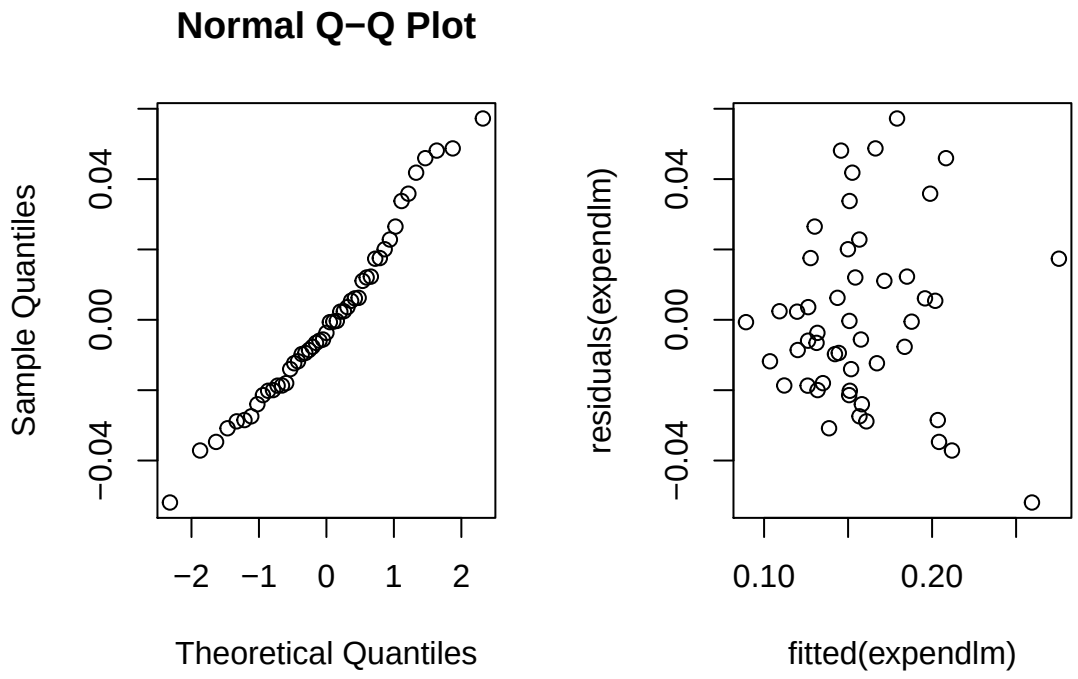

Residuals against employ
Residuals
0.04
0.00
−0.04
3
4
5
6
employ


Residuals against lawyer
Residuals
0.04
0.00
−0.04
1.5
2.5
3.5
lawyers


Added variable plot for lawyer
residual of expend
0.04
0.00
−0.04
−0.5
0.5
1.0
residual of lawyers


Added variable plot for emplo
residual of expend
0.06
0.02
−0.02
−0.06
−1.0
0.0
1.0
2.0
residual of employ


Bad
Crime
Pop
residuals(expendlm)
bad
crime
pop


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
residuals(expendlm)
fitted(expendlm)