EDDA - Assignment 2 - Group 77

Dante de Lang, Ignas Krikštaponis and Kamiel Gülpen

Exercise 1

If left alone bread will become moldy, rot or decay otherwise. To investigate the influence of temperature and humidity on this process, the time to decay was measured for 18 slices of white bread, which were placed in 3 different environments and humidified or not. The data are given in the file bread.txt, with the first column time to decay in hours, the second column the environment (cold, warm or intermediate temperature) and the third column the humidity.

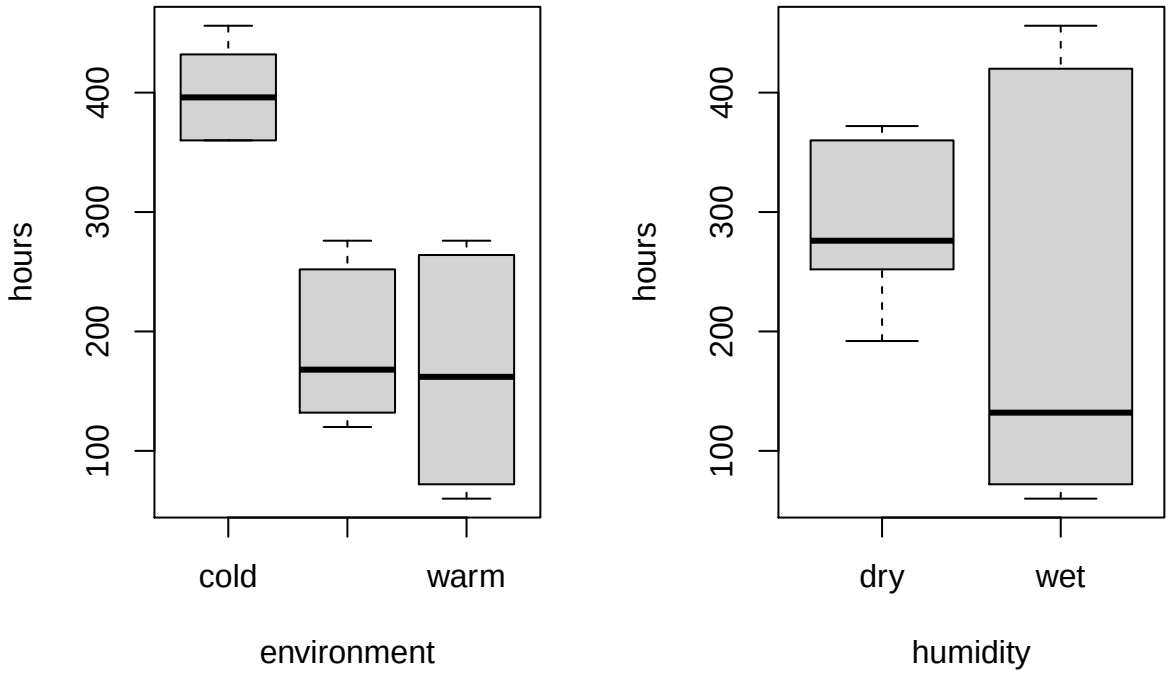
a) The 18 slices came from a single loaf, but were randomized to the 6 combinations of conditions. Present an R-code for this randomization process.

```
# randomization code
breads <- 18; environments <- 3; humidities <- 2;
as_tibble(
      cbind(bread = 1:breads,
      environment = sample(rep(c("cold", "warm", "intermediate"),breads/environments)),
      humidity = rep(c("dry", "wet"), breads/humidities))
)
```

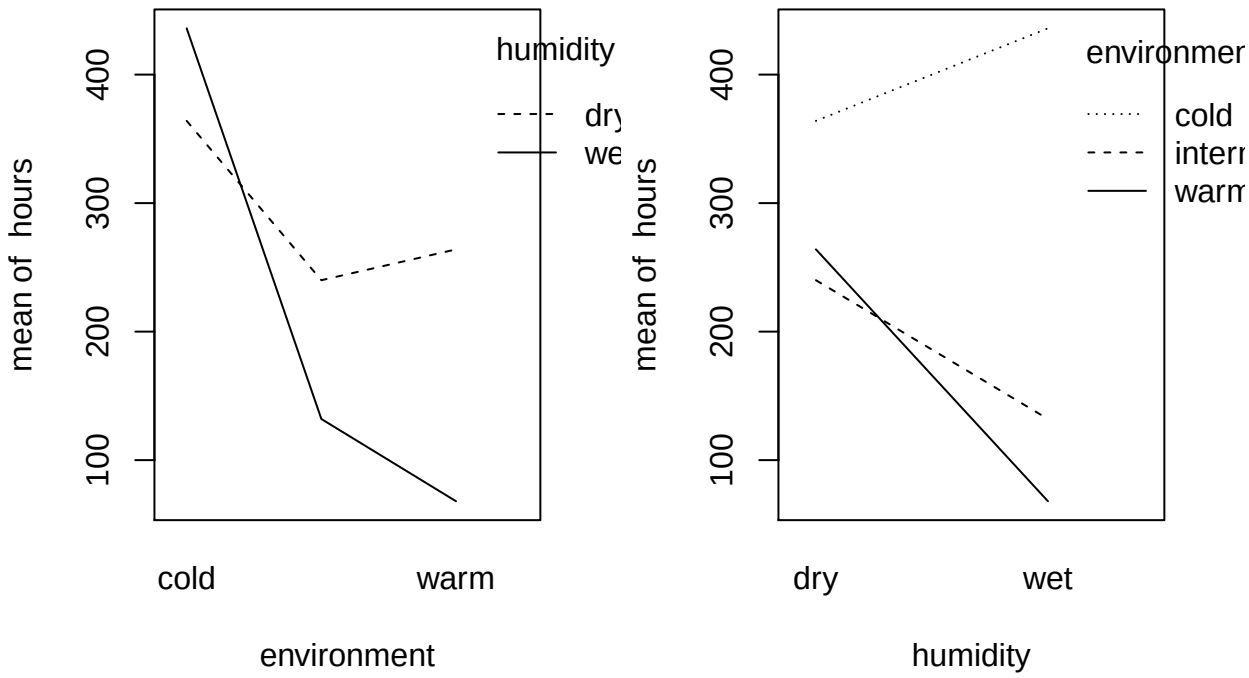
```
## # A tibble: 18 x 3
##    bread environment  humidity
##    <chr> <chr>        <chr>
##  1 1     warm         dry
##  2 2     intermediate wet
##  3 3     warm         dry
##  4 4     warm         wet
##  5 5     intermediate dry
##  6 6     intermediate wet
##  7 7     cold         dry
##  8 8     cold         wet
##  9 9     cold         dry
## 10 10    intermediate wet
## 11 11    warm         dry
## 12 12    cold         wet
## 13 13    warm         dry
## 14 14    intermediate wet
## 15 15    cold         dry
## 16 16    intermediate wet
## 17 17    cold         dry
## 18 18    warm         wet
```

b) Make two boxplots of hours versus the two factors and two interaction plots (keeping the two factors fixed in turn).

```r
# read data
data <- read.table(file="data/bread.txt",header=TRUE)
data$environment <- as.factor(data$environment); data$humidity <- as.factor(data$humidity)
attach(data)
# boxplots
par(mfrow=c(1,2))
boxplot(hours~environment)
boxplot(hours~humidity)
```




```r
# interaction plots
interaction.plot(environment,humidity,hours)
interaction.plot(humidity,environment,hours)
```



c) Perform an analysis of variance to test for effect of the factors temperature, humidity, and their interaction. Describe the interaction effect in words.

```
aov <- lm(hours ~ environment * humidity, data = data)
anova(aov)
```

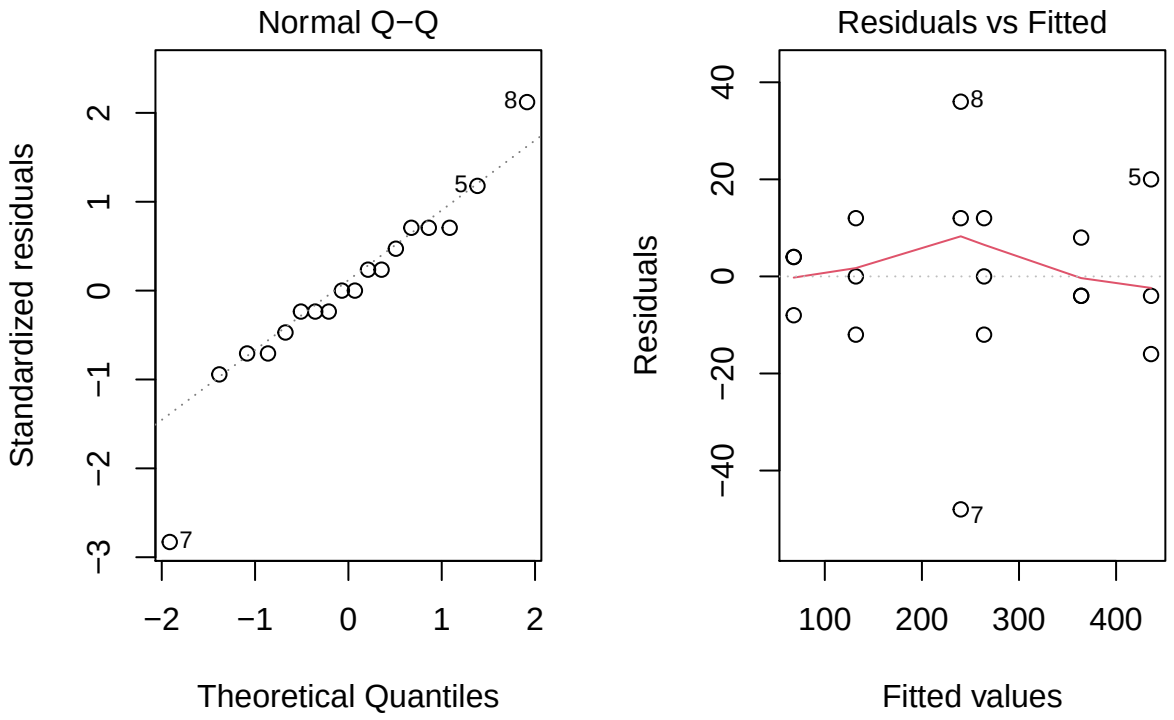
```
## Analysis of Variance Table
##
## Response: hours
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## environment           2 201904  100952   233.7 2.5e-10 ***
## humidity              1  26912   26912    62.3 4.3e-06 ***
## environment:humidity  2  55984   27992    64.8 3.7e-07 ***
## Residuals            12   5184     432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above p-values we can conclude that: * that the levels of environment are associated with significant different decay hours. * that the levels of humidity are associated with significant different decay hours. * the relationships between decay hours and environment depends on the humidity level.

d) Which of the two factors has the greatest (numerical) influence on the decay? Is this a good question?

e) Check the model assumptions by using relevant diagnostic tools. Are there any outliers?

```
par(mfrow=c(1,2))
plot(aov, 2);plot(aov, 1)
```



The above normality and residuals diagnostics signal that there are outliers. By removing them we would be able to satisfy the normality assumptions required for the two-way ANOVA test.

Exercise 2

A researcher is interested in the time it takes a student to find a certain product on the internet using a search engine. There are three different types of interfaces with the search engine and especially the effect of these interfaces is of importance. There are five different types of students, indicating their level of computer skill (the lower the value of this indicator, the better the computer skill of the corresponding student). Fifteen students are selected; three from each group with a certain level of computer skill. The data is given in the file search.txt. Assume that the experiment was run according to a randomized block design which you make in a). (Beware that the levels of the factors are coded by numbers.)

a Number the selected students 1 to 15 and show how (by using R) the students could be randomized to the interfaces in a randomized block design.

b

```
# read data
data <- read.table(file="data/search.txt",header=TRUE)
data$skill <- as.factor(data$skill); data$interface <- as.factor(data$interface)
# perform ANOVA
```

```r
aov <- lm(time ~ interface + skill, data = data)
anova(aov)
```

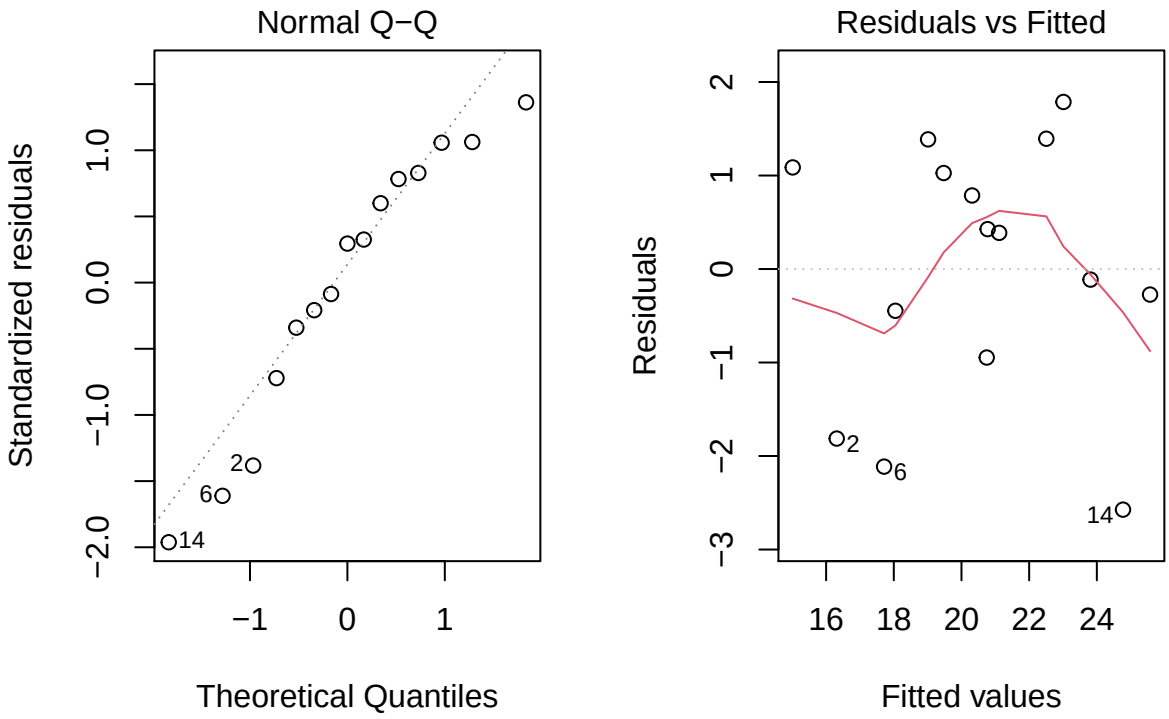
```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value Pr(>F)
## interface  2   50.5   25.23    7.82  0.013 *
## skill      4   80.1   20.01    6.21  0.014 *
## Residuals  8   25.8    3.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# summary table
summary(aov)
```

```
##
## Call:
## lm(formula = time ~ interface + skill, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.573 -0.697  0.387  1.057  1.787
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15.01       1.23   12.24  1.8e-06 ***
## interface2      2.70       1.14    2.38   0.0447 *
## interface3      4.46       1.14    3.93   0.0044 **
## skill2          1.30       1.47    0.89   0.4012
## skill3          3.03       1.47    2.07   0.0724 .
## skill4          5.30       1.47    3.61   0.0068 **
## skill5          6.10       1.47    4.16   0.0032 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.8 on 8 degrees of freedom
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.711
## F-statistic: 6.74 on 6 and 8 DF,  p-value: 0.0084
```

c) Check the model assumptions by using relevant diagnostic tools.

```r
par(mfrow=c(1,2))
plot(aov, 2);plot(aov, 1)
```



There are outliers therefore it would be best to use a different test or remove them.

d) Perform the Friedman test to test whether there is an effect of interface.

```
attach(data)
friedman.test(time, interface, skill)
```

```
##
##  Friedman rank sum test
##
## data:  time, interface and skill
## Friedman chi-squared = 6, df = 2, p-value = 0.04
```

According to the Friedman test there is an effect of the interface.

e) Test the null hypothesis that the search time is the same for all interfaces by a one-way ANOVA test, ignoring the variable skill. Is it right/wrong or useful/not useful to perform this test on this dataset?

```
one_aov <- lm(time ~ interface, data = data)
anova(one_aov)
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value Pr(>F)
## interface  2   50.5   25.23    2.86  0.096 .
```

```
## Residuals 12  105.9    8.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the test above, we could conclude that the interface has no siginificant affect on the time taken to find a certain product. However as we have multiple factors here, one-way ANOVA is not appropriate here as it will ignore the effects the ignored factor could have.

Excercise 3

In a study on the effect of feedingstuffs on lactation a sample of nine cows were fed with two types of food, and their milk production was measured. All cows were fed both types of food, during two periods, with a neutral period in-between to try and wash out carry-over effects. The order of the types of food was randomized over the cows. The observed data can be found in the file cow.txt, where A and B refer to the types of feedingstuffs.

a) Test whether the type of feedingstuffs influences milk production using an ordinary "fixed effects" model, fitted with lm. Estimate the difference in milk production.

```r
# read data
data <- read.table(file="data/cow.txt",header=TRUE)
data$treatment <- as.factor(data$treatment); data$order <- as.factor(data$order)
data$id <- as.factor(data$id); data$per <- as.factor(data$per)

# perform fixed effects model analysis
fixed_aov <- lm(milk ~ order + id + per + treatment, data = data)
summary(fixed_aov)
```

```
##
## Call:
## lm(formula = milk ~ order + id + per + treatment, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.260 -0.438  0.000  0.438  2.260
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   30.300      1.244   24.35  5.0e-08 ***
## orderBA      -11.200      1.574   -7.12  0.00019 ***
## id2           23.000      1.574   14.61  1.7e-06 ***
## id3           11.150      1.574    7.08  0.00020 ***
## id4           -1.350      1.574   -0.86  0.41948
## id5            4.150      1.574    2.64  0.03360 *
## id6           34.650      1.574   22.01  1.0e-07 ***
## id7           24.750      1.574   15.72  1.0e-06 ***
## id8           16.100      1.574   10.23  1.8e-05 ***
## id9               NA         NA      NA       NA
## per2          -2.390      0.747   -3.20  0.01505 *
## treatmentB    -0.510      0.747   -0.68  0.51654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.57 on 7 degrees of freedom
## Multiple R-squared:  0.993,  Adjusted R-squared:  0.983
## F-statistic:  101 on 10 and 7 DF,  p-value: 1.35e-06
```

b)

```
attach(data)
mixed_avo <- lmer(milk ~ treatment + order + per + (1|id),REML=FALSE)
mixed_avo_1 <- lmer(milk ~ order + per + (1|id),REML=FALSE)
anova(mixed_avo_1, mixed_avo)
```

```
## Data: NULL
## Models:
## mixed_avo_1: milk ~ order + per + (1 | id)
## mixed_avo: milk ~ treatment + order + per + (1 | id)
##             npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
## mixed_avo_1    5 118 122  -53.9      108
## mixed_avo      6 119 125  -53.7      107  0.58  1       0.45
```

c)

```
attach(data)
```

```
## The following objects are masked from data (pos = 3):
##
##     id, milk, order, per, treatment
```

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

```
##
##  Paired t-test
##
## data:  milk[treatment == "A"] and milk[treatment == "B"]
## t = 0.2, df = 8, p-value = 0.8
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.27  2.76
## sample estimates:
## mean of the differences
##                   0.244
```

Excercise 4

Stochastic models for word counts are used in quantitative studies on literary styles. Statistical analysis of the counts can, for example, be used to solve controversies about true authorships. Another example is the analysis of word frequencies in relation to Jane Austen's novel Sanditon. At the time Austen died, this novel was only partly completed. Austen, however, had made a summary for the remaining part. An admirer of Austen's work finished the novel, imitating Austen's style as much as possible. The file austen.txt contains counts of different words in some of Austen's novels: chapters 1 and 3 of Sense and Sensibility (stored in the

Sense column), chapters 1, 2 and 3 of Emma (column Emma), chapters 1 and 6 of Sanditon (both written by Austen herself, column Sand1) and chapters 12 and 24 of Sanditon (both written by the admirer, Sand2).

a) Discuss whether a contingency table test for independence or for homogeneity is most appropriate here.

```
data <- read.table(file="data/austen.txt",header=TRUE)
```

Homogeneity is the most appropretiate test here since we want to see if all chapters are homogeneous and if they are not, then the admirer had a different writting stile.

b) Using the given data set, investigate whether Austen herself was consistent in her different novels. Where are the main inconsistencies?

```
only_austen <- data %>%
  select(Sense, Emma, Sand1)
test <- chisq.test(only_austen); test
```

```
##
##  Pearson's Chi-squared test
##
## data:  only_austen
## X-squared = 12, df = 10, p-value = 0.3
```

```
test$residuals
```

```
##           Sense   Emma  Sand1
## a       -1.0300 -0.129  1.594
## an       0.4473 -0.159 -0.375
## this     0.0513  0.294 -0.504
## that     0.7482  0.287 -1.442
## with    -0.0475  0.521 -0.704
## without  1.0654 -1.588  0.893
```

From the p-value it seems that there are no significant difference between word counts in Austen novels' chapters.

c) Was the admirer successful in imitating Austen's style? Perform a test including all data. If he was not successful, where are the differences?

```
test <- chisq.test(data); test
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 46, df = 15, p-value = 6e-05
```

```
test$residuals
```

```
##          Sense      Emma  Sand1   Sand2
## a       -1.015 -0.112093  1.606 -0.0589
## an      -0.591 -1.219955 -1.067  3.7282
```

```
## this     0.139  0.390490 -0.444 -0.3267
## that     1.594  1.179849 -0.910 -3.0493
## with    -0.512  0.000192 -1.025  1.7482
## without  1.392 -1.341196  1.137 -1.0696
```

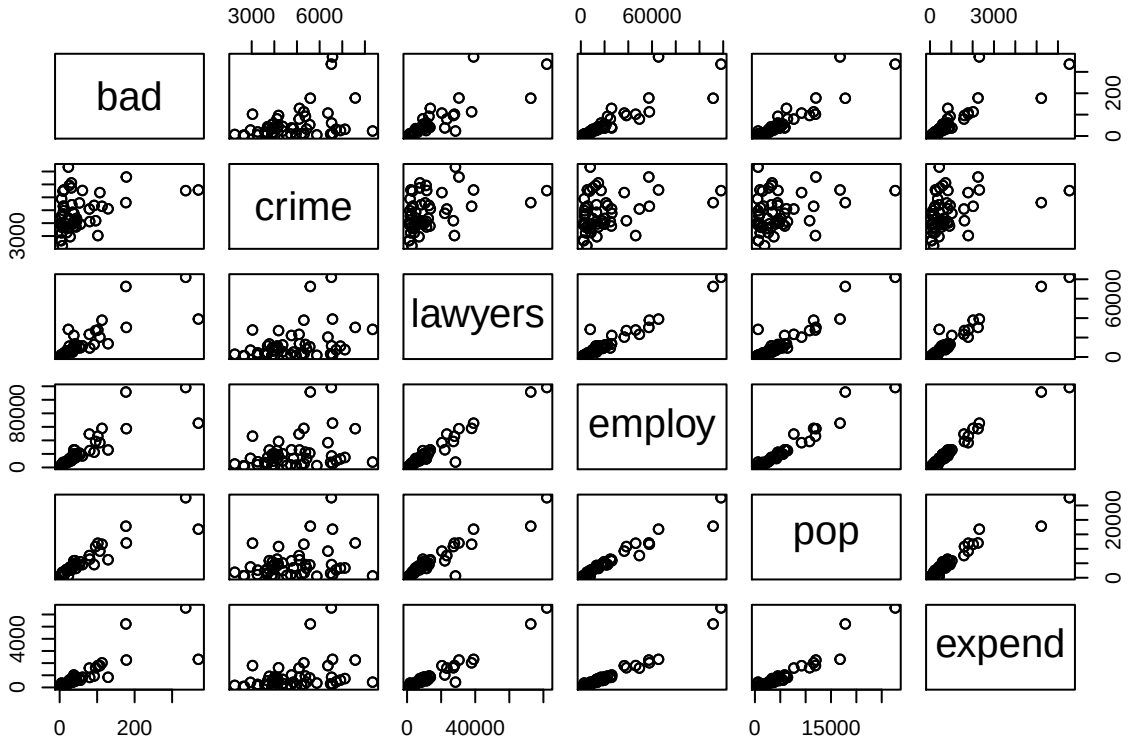
From the p-value above we can conclude that there are significant word count differences between chapters, therefore the admirer fail to mimic Austen's style. The main differences lay in word an and that.

Excercise 5

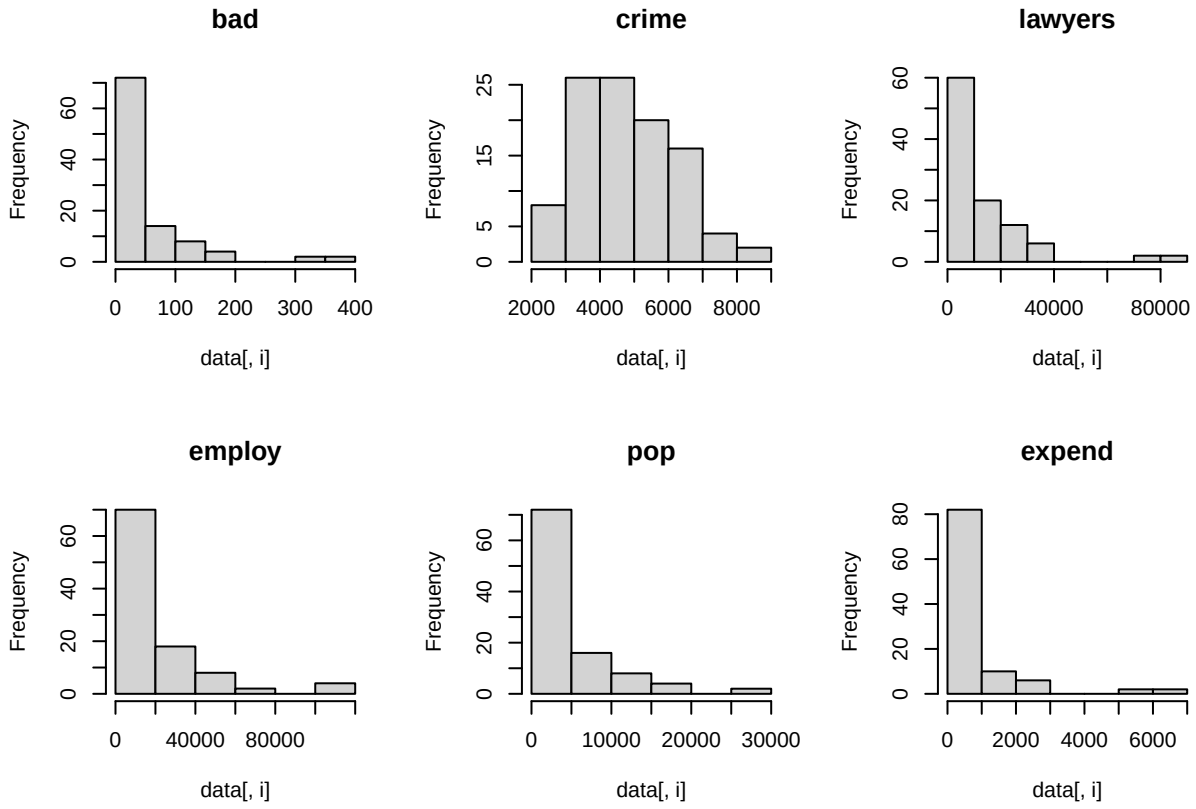
a) The data in expensescrime.txt were obtained to determine factors related to state expenditures on criminal activities (courts, police, etc.) The variables are: state (indicating the state in the USA), expend (state expenditures on criminal activities in $1000), bad (crime rate per 100000), crime (number of persons under criminal supervision), lawyers (number of lawyers in the state), employ (number of persons employed in the state) and pop (population of the state in 1000). In the regression analysis, take expend as response variable and bad, crime, lawyers, employ and pop as explanatory variables.

a) Make some graphical summaries of the data. Investigate the problem of potential and in uence points, and the problem of collinearity.

```
data <- read.table(file="data/expensescrime.txt",header=TRUE)
plot(data[,c(3, 4, 5, 6, 7, 2)])
```



```r
par(mfrow=c(2,3))
for (i in c(3, 4, 5, 6, 7, 2)) hist(data[,i],main=names(data)[i])
```



b) Fit a linear regression model to the data. Use both the step-up and the step-down method to find the best model. If step-up and step-down yield two different models, choose one and motivate your choice.

```r
# step down model
down_lm_1  <- lm(expend ~ bad + crime + lawyers + employ + pop ,data = data); summary(down_lm_1)
```

```
##
## Call:
## lm(formula = expend ~ bad + crime + lawyers + employ + pop, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -638.4  -95.3   22.2  115.5  805.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.99e+02   9.59e+01   -3.12   0.0024 **
## bad         -2.83e+00   8.49e-01   -3.33   0.0012 **
## crime        3.24e-02   1.93e-02    1.68   0.0957 .
## lawyers      2.32e-02   5.51e-03    4.22  5.5e-05 ***
## employ       2.30e-02   5.11e-03    4.50  1.9e-05 ***
## pop          7.79e-02   2.41e-02    3.24   0.0017 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 218 on 96 degrees of freedom
## Multiple R-squared:  0.968,  Adjusted R-squared:  0.966
## F-statistic:  572 on 5 and 96 DF,  p-value: <2e-16
```

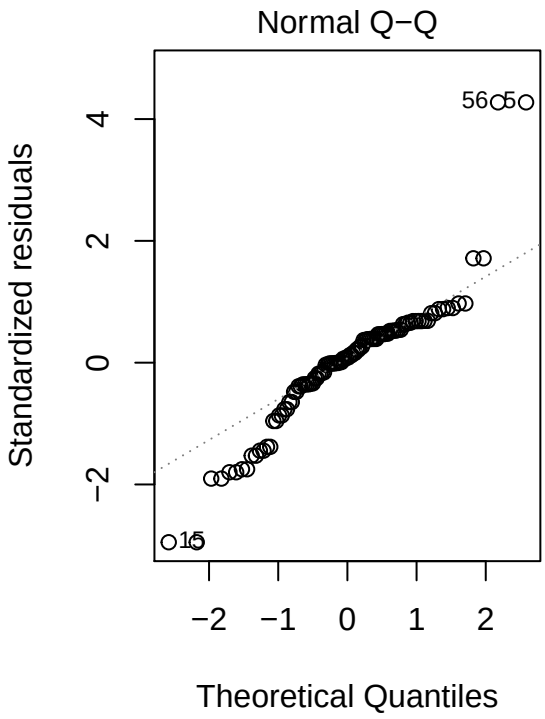
```
down_lm_2  <- lm(expend ~ bad + lawyers + employ + pop ,data = data); summary(down_lm_2)
```

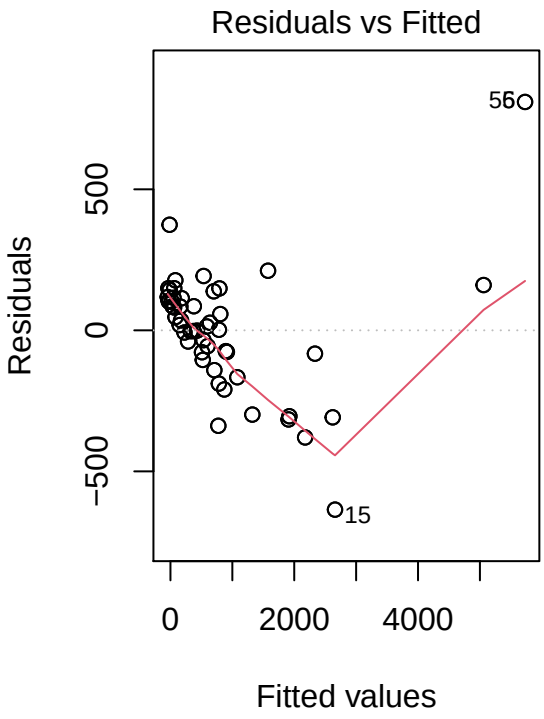
```
##
## Call:
## lm(formula = expend ~ bad + lawyers + employ + pop, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -635.6  -81.4   18.8  114.9  809.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.46e+02   3.13e+01   -4.68  9.2e-06 ***
## bad         -2.24e+00   7.80e-01   -2.87   0.0050 **
## lawyers      2.65e-02   5.21e-03    5.08  1.9e-06 ***
## employ       2.28e-02   5.16e-03    4.43  2.5e-05 ***
## pop          6.37e-02   2.27e-02    2.80   0.0062 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 220 on 97 degrees of freedom
## Multiple R-squared:  0.967,  Adjusted R-squared:  0.965
## F-statistic:  701 on 4 and 97 DF,  p-value: <2e-16
```

```
# step up model
```

c) Check the model assumptions by using relevant diagnostic tools.

```
par(mfrow=c(1,2))
plot(down_lm_2, 2);plot(down_lm_2, 1)
```


Normal Q−Q
Standardized residuals
Theoretical Quantiles
−2
−1
0
1
2
4
56
5
15


Residuals vs Fitted
Residuals
Fitted values
−500
0
500
2000
4000
56
15