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.

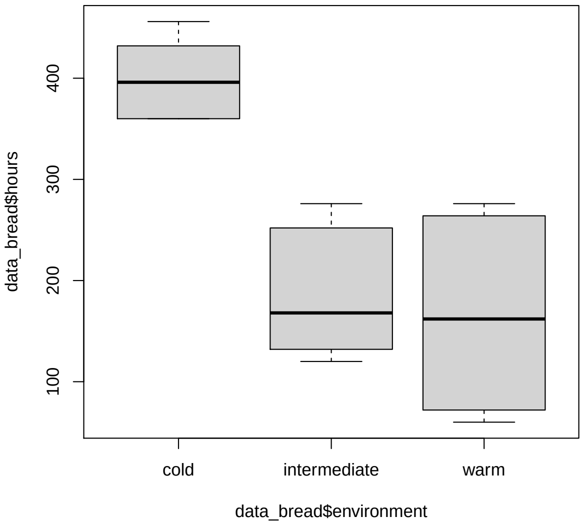
a)The 18 slices came from a single loaf, but were randomized to the 6 combinations of conditions. Present an R-code for this randomization process.

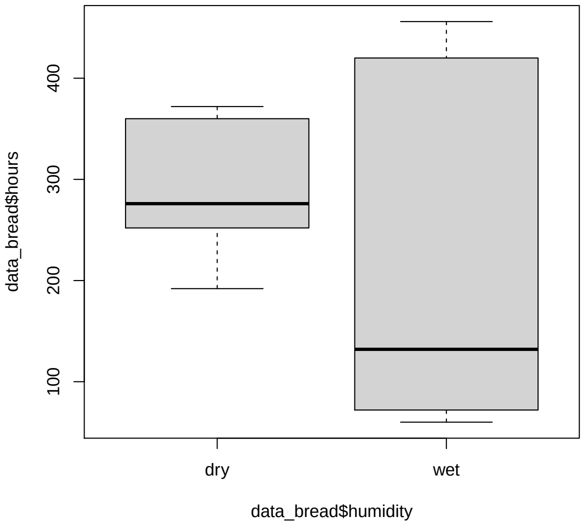
```
data_bread <- read.table(file="data/bread.txt",header=TRUE)
humid <- factor(rep(c("dry","wet"),each = 9))
temp <- factor(rep(c("cold", "intermediate","warm"),times = 6))
knitr::kable(data.frame(humid,temp,slice = sample(1:18)))
```

humid	temp	slice
dry	cold	18
dry	intermediate	16
dry	warm	8
dry	cold	7
dry	intermediate	15
dry	warm	11
dry	cold	6
dry	intermediate	4
dry	warm	3
wet	cold	5
wet	intermediate	10
wet	warm	17
wet	cold	12
wet	intermediate	9
wet	warm	1
wet	cold	13
wet	intermediate	2
wet	warm	14

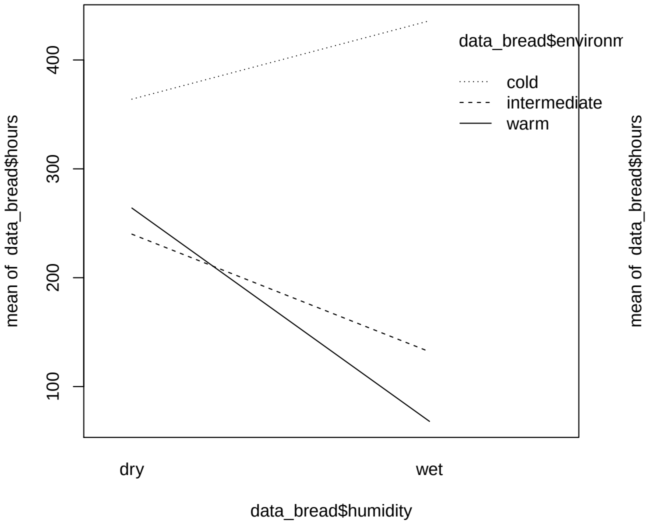
b)Make two boxplots of hours versus the two factors and two interaction plots (keeping the two factors fixed in turn).

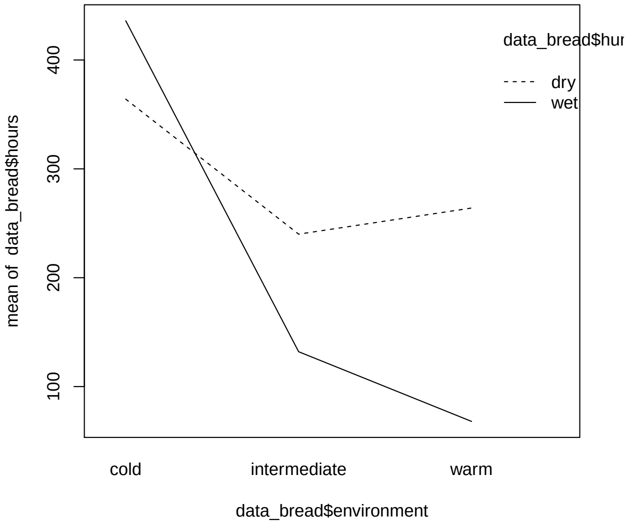
```
par(mfrow=c(1,2))
boxplot(data_bread$hours~data_bread$environment)
boxplot(data_bread$hours~data_bread$humidity)
```





```
interaction.plot(data_bread$humidity,data_bread$environment,data_bread$hours)
interaction.plot(data_bread$environment,data_bread$humidity,data_bread$hours)
```





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

```
attach(data_bread)
environment=as.factor(environment)
humidity=as.factor(humidity)
```

```
dataaov=lm(hours~humidity*environment,data=data_bread)
anova(dataaov)
```

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

```
summary(dataaov)$coefficients
```

```
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              364         12   30.33 1.03e-12
## humiditywet                               72         17    4.24 1.14e-03
## environmentintermediate                 -124         17   -7.31 9.39e-06
## environmentwarm                         -100         17   -5.89 7.34e-05
## humiditywet:environmentintermediate     -180         24   -7.50 7.23e-06
## humiditywet:environmentwarm             -268         24  -11.17 1.07e-07
```

When looking at the two-way anova model we see that it consists of the following terms: $Y_{ijk} = \mu_{ij} + e_{ijk}$ $= \mu + alpha_i + \beta_j + \gamma_{ij} + e_{eijk}$ We decompose the formula it this way such that μ is the overall mean, α_i and β_j are the main effect of level i and j of the first factor and second factor respectively and γ_{ij} the interaction effect.

In order to test the effect of the temperature,humidity, and the interaction we set up 3 hypotheses which are: H_{AB}: $\gamma_{ij} = 0$ for every (i, j) (no interactions between factor A and B)

H_A: $\alpha_i = 0$ for every i (no main effect of factor A)

H_B:$\beta_j = 0$ for every j (no main effect of factor B)

We use the test statistics F_{AB} for H_{AB}, F_A for H_A and F_B for H_B where F is the F-distribution.

To see if the Hypotheses can be rejected we want to look at the probability that P(F>f_{AB}), P(F>f_A) and P(F>f_B), the bigger the F value the lower the probability that the Hypothesis lays under a F-distribution and therefore the Hypothesis can be rejected.

We see that the humidity has a p-value of 4.3e-06, environment a p-value of 2.5e-10 and the interaction between the two (humidity:environment) shows a p-value of 3.7e-07. This means that humidity, environment and the interaction effect between humidity and environment have a significant influence on the hours, which means we can reject H_A, H_B and H_{AB}.

The interaction effect looks at the difference of differences, for example: it looks at the difference in hours for environment = cold and environment = warm for humidity = wet. Then it looks at the difference between environment = cold and environment = warm for humidity = dry. It then looks at the difference between those differences and when this difference is high it shows that there is indeed interaction.

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

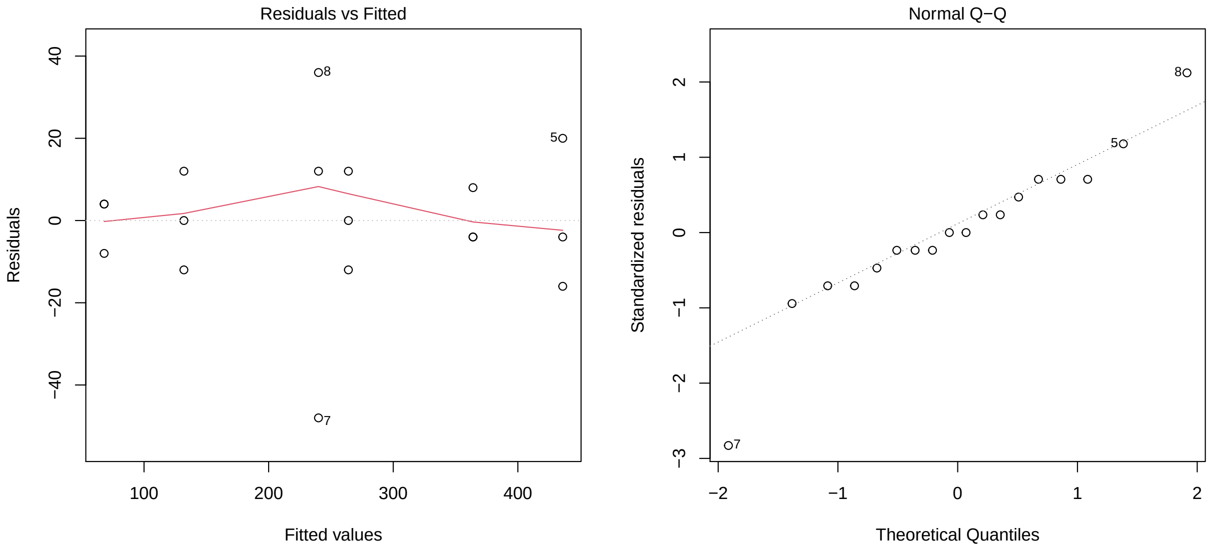
```
# Without interaction
humidity=as.factor(humidity)
environment=as.factor(environment)
dataaov=lm(hours~humidity+environment)
anova(dataaov)
```

```
## Analysis of Variance Table
##
## Response: hours
##             Df Sum Sq Mean Sq F value  Pr(>F)
## humidity     1  26912   26912    6.16   0.026 *
## environment  2 201904  100952   23.11 3.7e-05 ***
## Residuals   14  61168    4369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When we want to know which factor has the greatest influence we want to use the additive model as used above. This shows a p-value of 0.026 for humidity and a p-value of 3.7e-05 for environemnt. This means that the environment has the greatest influence.

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

```
par(mfrow=c(1,2))
dataaov2=lm(hours~humidity*environment,data=data_bread);
plot(dataaov2, 1)
plot(dataaov2, 2)
```



The qqplot shows a somewhat linear line which means that based on the qqplot we can state that the data is normally distributed. We also looked at the spread of the residuals, which showed that there are three outliers which are number 5, 7 and 8 which can be observed in both plot.

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.

```
interface <- factor(rep(c(1,2,3),each = 5))
skill <- factor(rep(c(1,2,3,4,5),times = 3))
student <- c(1:15) # shouldn't we also sample the students here? or then sample those 15 combinations?
knitr::kable(data.frame(student,skill,interface))
```

student	skill	interface
1	1	1
2	2	1
3	3	1
4	4	1
5	5	1
6	1	2
7	2	2
8	3	2
9	4	2
10	5	2
11	1	3
12	2	3
13	3	3
14	4	3
15	5	3

b) Test the null hypothesis that the search time is the same for all interfaces. What type of interface does require the longest search time? For which combination of skill level and type of interface is the search time the shortest? Estimate the time it takes a typical user of skill level 3 to find the product on the website if the website uses interface 3.

```
data_search <- read.table(file="data/search.txt",header=TRUE)
data_search$skill <- as.factor(data_search$skill)
data_search$interface <- as.factor(data_search$interface)
# perform ANOVA
aovsearch = lm(time~interface+skill, data= data_search);  anova(aovsearch);
```

```
## 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
```

```
summary(aovsearch)$coefficients
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15.01       1.23  12.238 1.85e-06
## interface2      2.70       1.14   2.377 4.47e-02
## interface3      4.46       1.14   3.927 4.38e-03
## skill2          1.30       1.47   0.887 4.01e-01
## skill3          3.03       1.47   2.069 7.24e-02
## skill4          5.30       1.47   3.614 6.84e-03
## skill5          6.10       1.47   4.160 3.16e-03
```

Looking at the additive ANOVA test we can conclude that there is a significant main effect of the interface (p-value < 0.05). Furthermore, the summary shows that interface 3 gives the highest α parameter value, making the time it takes for this interface the longest. For the shortest search time, interface 1 can be combined with skill levels 1,2 or 3 since all three have the lowest α parameter values without 2 and 3 being significant different from skill level 0.

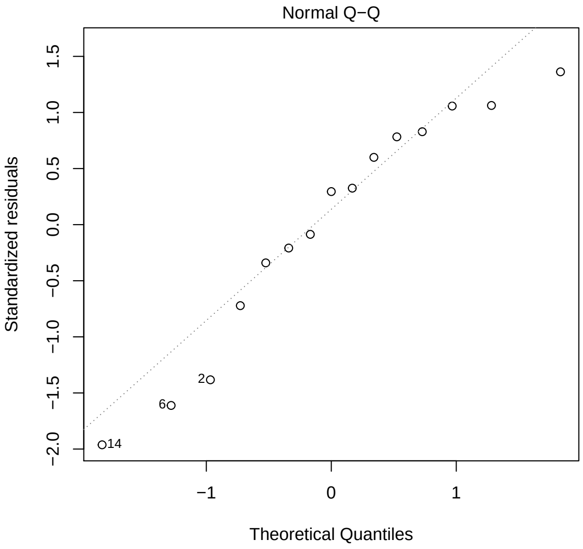
For the estimation of time it takes a typical user of skill level 3 using interface 3 we can calculate Y by summing the estimates and adding the error, giving a time of 24.3 units:

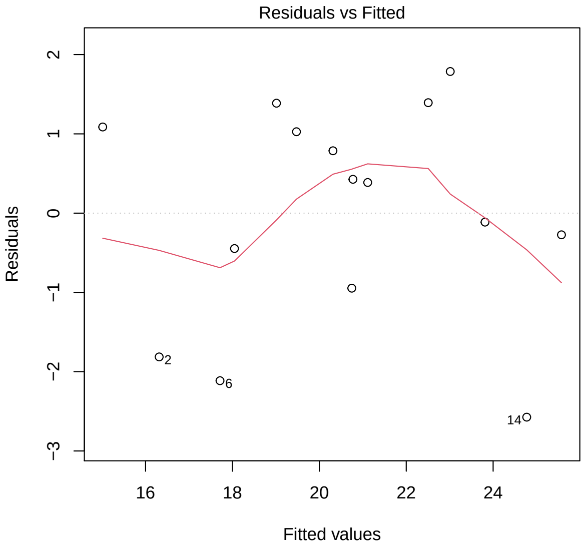
```
# Estimate interface 3 and skill 3:
Y = 15.01+4.46+3.03+1.8; Y
```

```
## [1] 24.3
```

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

```
par(mfrow=c(1,2))
plot(aovsearch,2)
plot(aovsearch,1)
```





As shown in the above QQ-plot and the residuals-fitted plot there are some outliers (points 2, 6, 14) that raise doubt about the normality of the data, however there does not seem to be a clear trend in residuals vs fitter plot.

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

```
friedman.test(data_search$time, data_search$interface, data_search$skill)$p.value
```

```
## [1] 0.0408
```

The test shows a p-value o 0.04 which is significant. This means that the H_0 can be rejected and we can state that there is a significant 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?

```
anova(lm(data_search$time~data_search$interface))
```

```
## Analysis of Variance Table
##
## Response: data_search$time
##                       Df Sum Sq Mean Sq F value Pr(>F)
## data_search$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
```

is it not useless also to ignore skill since the time is clearly also depended on this variable, you can not simply ignore such a variable right?

Looking at the p-value of the one-way ANOVA test, we see that is not significant. We could therefore conclude that the interfaces does not have a significant effect on the search time. However, since the data originates from a random block design, it is not correct to use this test since it leaves out important interactions. Ignas: maybe last sentence as: However, since the data originates from a random block design, it is not correct to use this test since it leaves out important effects skill, which we observed to be significant in b).

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.

```
# 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 ~ id + per + treatment, data = data)
anova(fixed_aov); table <- summary(fixed_aov)$coefficients["treatmentB",]; print("Estimate for Treatmen
```

```
## Analysis of Variance Table
##
## Response: milk
##           Df Sum Sq Mean Sq F value  Pr(>F)
## id         8   2467   308.4  124.48 7.5e-07 ***
## per        1     25    24.5    9.89   0.016 *
## treatment  1      1     1.2    0.47   0.517
## Residuals  7     17     2.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] "Estimate for Treatment B:"

##   Estimate Std. Error    t value   Pr(>|t|)
##     -0.510      0.747     -0.683      0.517
```

From the results of fixed effects model above we see that the p-value for treatment is > 0.05, therefore we can conclude that there is no significant effect of the treatment. From the estimate of TreatmentB we see that $\beta_{treatment_B}$ = -0.051 (meaning that with treatment B we have -0.51 less milk production that with treatment A), however p-value > 0.05 and, therefore, the difference is insignificant. However, this model is not correct as it does not take into consideration the "random effect" introduced by the the order of the types of food randomization over the cows.

b) Repeat a) by performing a mixed effects analysis, modelling the cow effect as a random effect (use the function lmer). Compare your results to the results found by using a mixed effects model.

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

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: milk ~ treatment + order + per + (1 | id)
##
##      AIC      BIC   logLik deviance df.resid
##    119.3    124.7    -53.7    107.3       12
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5311 -0.3710  0.0269  0.2675  1.7249
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 133.15   11.54
##  Residual               1.93    1.39
## Number of obs: 18, groups:  id, 9
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   38.500      5.811    6.63
## treatmentB    -0.510      0.658   -0.77
## orderBA       -3.470      7.768   -0.45
## per2          -2.390      0.658   -3.63
##
## Correlation of Fixed Effects:
##            (Intr) trtmnB ordrBA
## treatmentB -0.063
## orderBA    -0.743  0.000
## per2       -0.063  0.111  0.000
```

```
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
```

The code above implemented the correct model that modeled the cows as "random effects". The fixed effects estimates in the summary table are the same as with the model in a). Furthermore, the code above performed an ANOVA test between the random effect model with and without treatment in it. The p-value for treatment is lower with this model than in a), however it is still > 0.05 - meaning, that there is no significant difference between the models with and without treatment, therefore there is no significant effect of the treatment.

c Study the commands below. Does this produce a valid test for a difference in milk production? Is its conclusion compatible with the one obtained in a)? Why?

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

```
## [1] 0.828
```

The code above performed a paired t-test (same treatment was done on the same cow) to test whether the means of the two populations are significantly different. Here, we see that p-value is > 0.05, which brings us to the same conclusion as with a) and b). However, this is not correct as we can see from the previous analysis that the order had a significant effect on the experimental outcomes - factor this t-test omits. This t-test is the same as performing a two-way ANOVA of treament + id. We can see that the p-value is the same:

```
paste("p-value with two-way ANOVA:",
      round(anova(lm(milk ~ treatment + id, data = data))["treatment", ]$`Pr(>F)`, 3))
```

```
## [1] "p-value with two-way ANOVA: 0.828"
```

Exercise 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 author ships. 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.

The contingency table test for homogeneity is appropriate because we want to know if the fan writer imitates Austen in a good way. This means that we want to test whether or not the different columns of data in the table come from the same population (writer) or not, which would be the case it the fan imitated Austen correctly. The H0 of the contingency table test for homogeneity states that the distribution of the words is the same for the stories.

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

```
data <- read.table(file="data/austen.txt",header=TRUE)
austen <- data[,1:3] # filter data to only have data from Austen
z = chisq.test(austen); z; residuals(z)
```

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

```
##            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
```

She is not inconsistent as the p-value is above 0.05. This means that we cannot reject the H_0. She does however, have some main inconsistency, which where the words “a”, “that” and “without”. As can be seen in the residual table above.

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?

```
z = chisq.test(data); z; residuals(z)
```

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

```
##          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
```

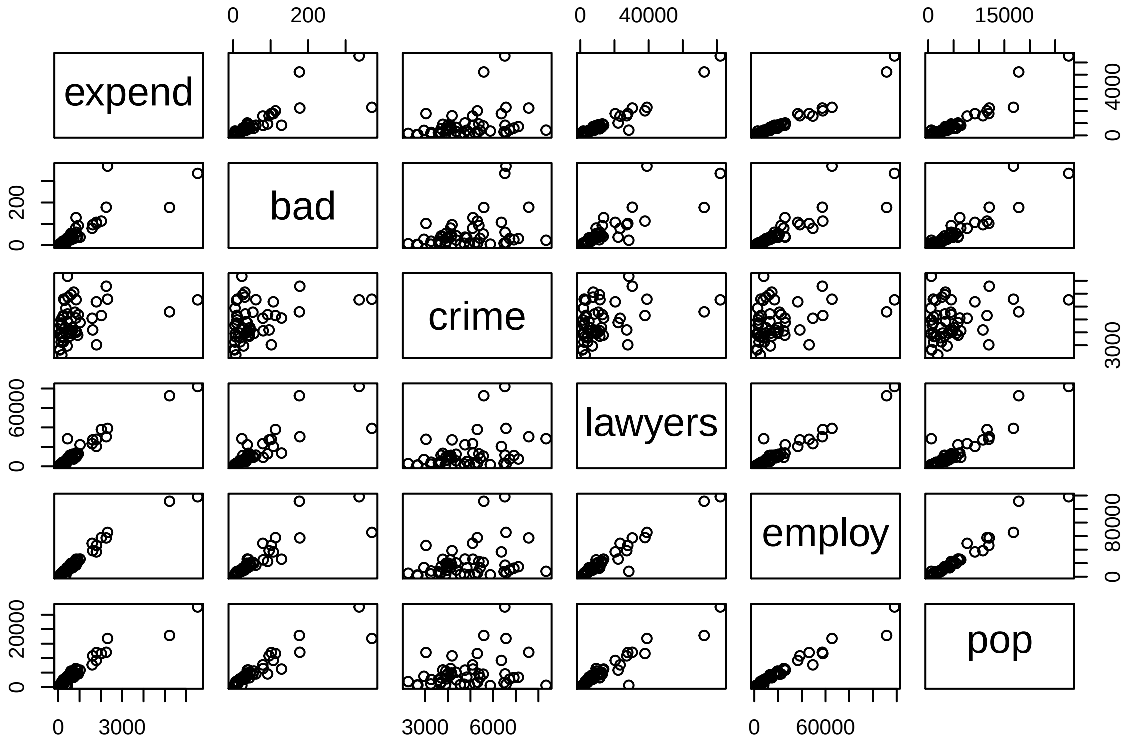
The fan is inconsistent as the p-value of the test is below 0.05. Therefore we have to reject the H0 and accept that the distribution of the words in the stories are not the same. Because Austen herself did not have this inconsistency we can say that the inconsistency is caused by the fan writer. The main inconsistencies were for the words “that” and “an”. As can be seen in the residual table above.

Exercise 5

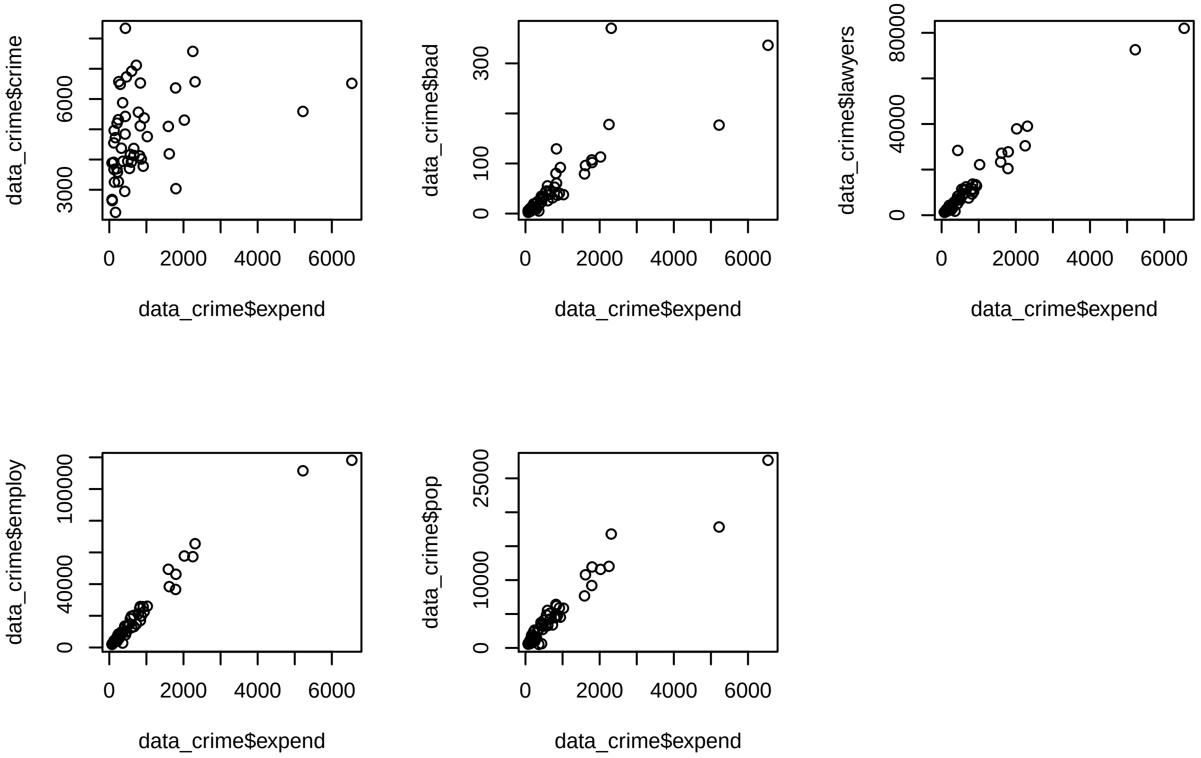
The data in expenses crime.txt were obtained to determine factors related to state expenditures on criminal activities (courts, police, etc.) The variables are: state (indicating the state in the USA), expend (state expenditures on criminal activities in $1000), bad(crime rate per 100000),crime (number 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 influence points,and the problem of collinearity.

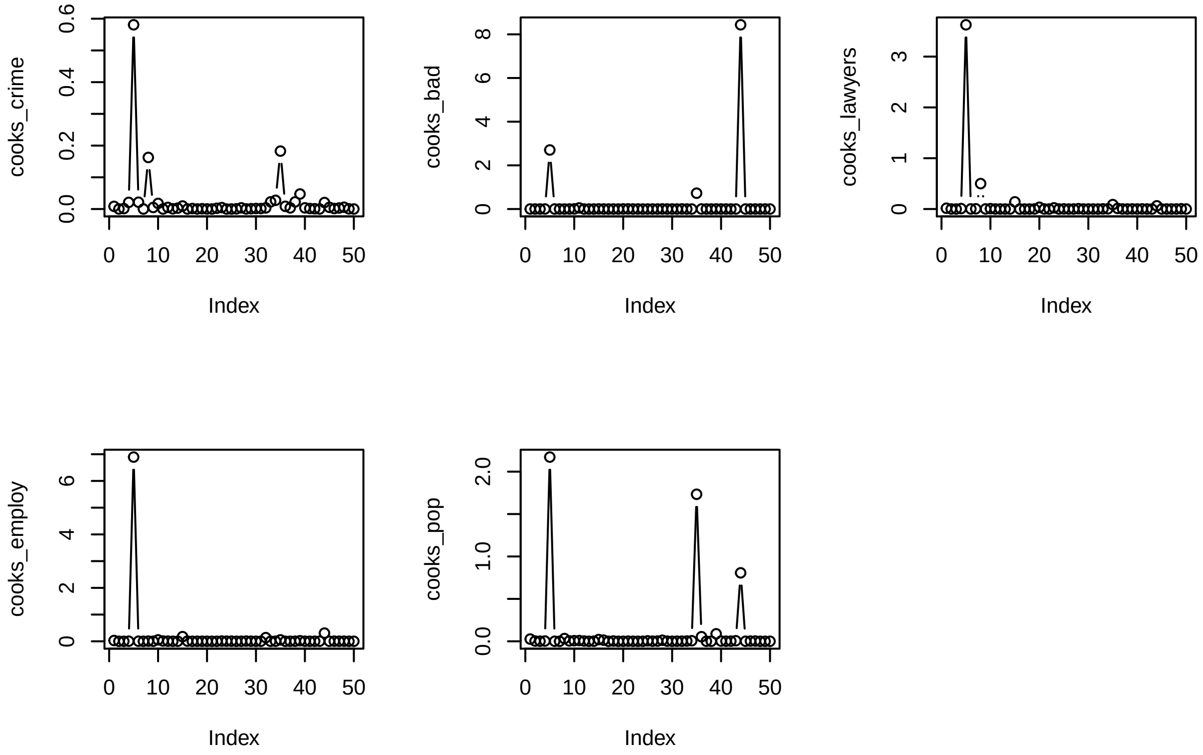
```
data_crime = read.table(file="data/expensescrime.txt",header=TRUE)
regression_data = data_crime[2:7]
pairs(regression_data)
```



```
par(mfrow=c(2,3))
plot(data_crime$expend,data_crime$crime)
plot(data_crime$expend,data_crime$bad)
plot(data_crime$expend,data_crime$lawyers)
plot(data_crime$expend,data_crime$employ)
plot(data_crime$expend,data_crime$pop)
```



```
par(mfrow=c(2,3))
cooks_crime = cooks.distance(lm(expend~crime, data = regression_data))
plot(cooks_crime, type="b")
cooks_bad = cooks.distance(lm(expend~bad, data = regression_data))
plot(cooks_bad, type="b")
cooks_lawyers = cooks.distance(lm(expend~lawyers, data = regression_data))
plot(cooks_lawyers, type="b")
cooks_employ = cooks.distance(lm(expend~employ, data = regression_data))
plot(cooks_employ, type="b")
cooks_pop = cooks.distance(lm(expend~pop, data = regression_data))
plot(cooks_pop, type="b")
```



Looking at the above plots we can see that for the variables bad, lawyers, employ and pop there exist potential and influence points. This is shown through the peaks in the cooks distance plots with peaks going above 1. Based on this rule of thumb the influence points were removed from the data.

```
# Collinearity
round(cor(regression_data),2)
```

```
##         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    0.37   0.30 0.27
## lawyers   0.97 0.83  0.37    1.00   0.97 0.93
## employ    0.98 0.87  0.30    0.97   1.00 0.97
## pop       0.95 0.92  0.27    0.93   0.97 1.00
```

Looking at the collinearity table we see that employee and and lawyers are strongly correlated(0.97). Furthermore, we also see that employee and crime rate per 100000 are strongly correlated(0.87). Also for lawyers and crime rate per 100000 it shows that they are strongly correlated(0.83). Lastly we also see a correlation between pop and bad and pop and lawyers and pop and employ.

```
regressionlm=lm(expend~bad+crime+lawyers+employ, data=regression_data)
car::vif(regressionlm)
```

```
## Registered S3 methods overwritten by ’car’:
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4
```

```
##     bad   crime lawyers  employ
##    4.42    1.30   16.58   20.87
```

```
# We see a value above 5 for lawyers and employees which means we need to take one out

regressionlm=lm(expend~bad+crime+lawyers, data=regression_data)
car::vif(regressionlm)
```

```
##     bad   crime lawyers
##    3.26    1.17    3.27
```

```
# Now it looks good
```

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.

```
# Step-up method

print("1st level:")
```

```
## [1] "1st level:"
```

```
paste("Bad:",round(summary(lm(expend~bad, data=regression_data))$r.squared, 3))
```

```
## [1] "Bad: 0.694"
```

```
paste("Crime: ", round(summary(lm(expend~crime, data=regression_data))$r.squared, 3))
```

```
## [1] "Crime:  0.109"
```

```
paste("Lawyers: ", round(summary(lm(expend~lawyers, data=regression_data))$r.squared, 3)) #0.9369
```

```
## [1] "Lawyers:  0.937"
```

```
paste("Employ: ", round(summary(lm(expend~employ, data=regression_data))$r.squared, 3), "- selected")#0
```

```
## [1] "Employ:  0.954 - selected"
```

```
paste("Pop: ", round(summary(lm(expend~pop, data=regression_data))$r.squared, 3)) # 0.907
```

```
## [1] "Pop:  0.907"
```

```r
print("2nd level: employ +")
```

```
## [1] "2nd level: employ +"
```

```r
paste("Bad:",round(summary(lm(expend~employ+bad, data=regression_data))$r.squared, 3))
```

```
## [1] "Bad: 0.955"
```

```r
paste("Crime: ",round(summary(lm(expend~employ+crime, data=regression_data))$r.squared, 3))
```

```
## [1] "Crime:  0.955"
```

```r
paste("Pop: ",round(summary(lm(expend~employ+pop, data=regression_data))$r.squared, 3))
```

```
## [1] "Pop:  0.954"
```

```r
paste("Lawyers: ",round(summary(lm(expend~employ+lawyers, data=regression_data))$r.squared, 3), "- sele
```

```
## [1] "Lawyers:  0.963 - selected"
```

```r
# expend = -1.146e+02 + 2.690e-02*lawyers + 2.976e-02*employ  + error
# Step-down

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

```
##
## Call:
## lm(formula = expend ~ bad + crime + lawyers + employ + pop, data = regression_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -638.7  -92.6   23.1  117.7  792.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.14e+02   1.42e+02   -2.21   0.0322 *
## bad         -2.90e+00   1.25e+00   -2.32   0.0251 *
## crime        3.42e-02   2.84e-02    1.21   0.2345
## lawyers      2.31e-02   8.08e-03    2.86   0.0064 **
## employ       2.27e-02   7.50e-03    3.03   0.0041 **
## pop          8.06e-02   3.55e-02    2.27   0.0281 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 227 on 44 degrees of freedom
## Multiple R-squared:  0.968,  Adjusted R-squared:  0.964
## F-statistic:  264 on 5 and 44 DF,  p-value: <2e-16
```

```
summary(lm(expend~lawyers+employ+bad + pop, data=regression_data))
```

```
##
## Call:
## lm(formula = expend ~ lawyers + employ + bad + pop, data = regression_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -635.8  -79.6   19.7  116.5  799.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.52e+02   4.65e+01   -3.27   0.0020 **
## lawyers      2.65e-02   7.62e-03    3.48   0.0011 **
## employ       2.26e-02   7.54e-03    3.00   0.0044 **
## bad         -2.27e+00   1.14e+00   -1.99   0.0529 .
## pop          6.54e-02   3.33e-02    1.96   0.0560 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 228 on 45 degrees of freedom
## Multiple R-squared:  0.967,  Adjusted R-squared:  0.964
## F-statistic:  326 on 4 and 45 DF,  p-value: <2e-16
```

```
summary(lm(expend~lawyers+employ + bad, data=regression_data))
```

```
##
## Call:
## lm(formula = expend ~ lawyers + employ + bad, data = regression_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -631.8  -94.9   32.2   92.4  958.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.14e+02   4.36e+01   -2.62   0.0118 *
## lawyers      2.63e-02   7.85e-03    3.36   0.0016 **
## employ       3.23e-02   5.85e-03    5.53  1.5e-06 ***
## bad         -8.55e-01   9.12e-01   -0.94   0.3530
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 235 on 46 degrees of freedom
## Multiple R-squared:  0.964,  Adjusted R-squared:  0.961
## F-statistic:  408 on 3 and 46 DF,  p-value: <2e-16
```

```
summary(lm(expend~lawyers+employ , data=regression_data))
```

```
##
## Call:
```

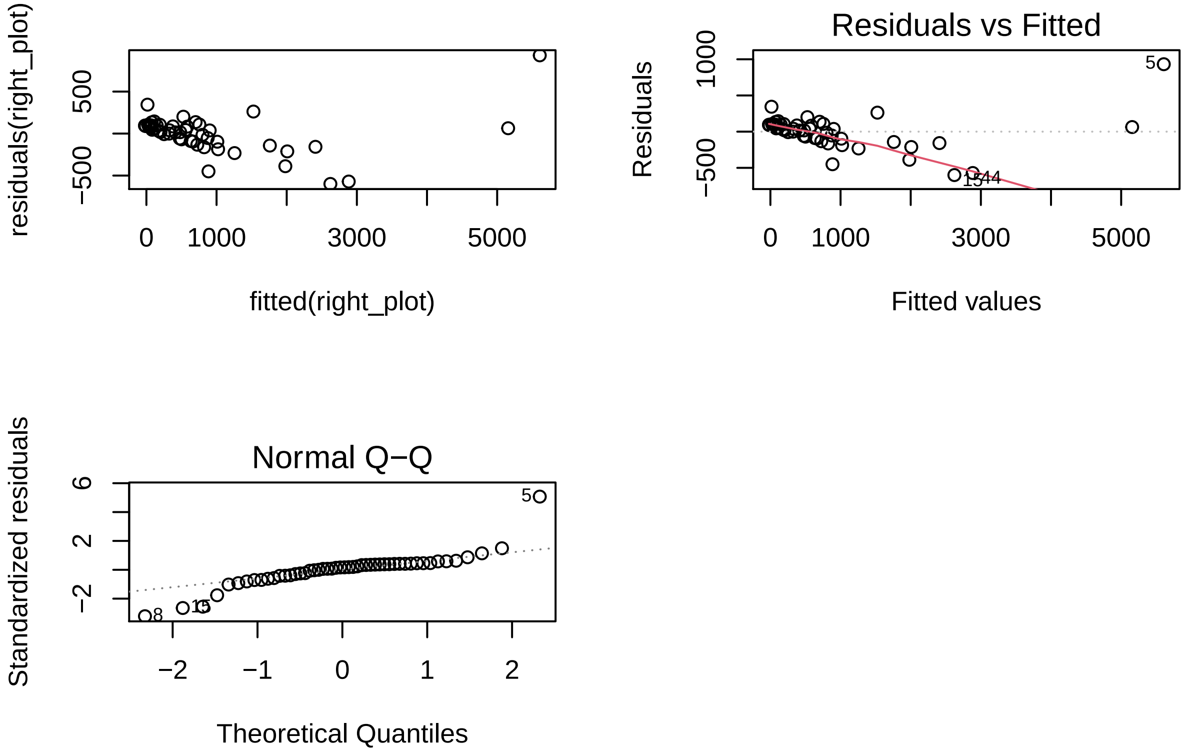
```
## lm(formula = expend ~ lawyers + employ, data = regression_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -599.8  -93.4   38.4   94.8  931.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.15e+02   4.36e+01   -2.63   0.0115 *
## lawyers      2.69e-02   7.82e-03    3.44   0.0012 **
## employ       2.98e-02   5.15e-03    5.77  5.9e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 234 on 47 degrees of freedom
## Multiple R-squared:  0.963,  Adjusted R-squared:  0.962
## F-statistic:  613 on 2 and 47 DF,  p-value: <2e-16
```

Step up method: First we started by fitting a linear model with one explanatory variable. Out of the 5 available variables the employ performed the best according to r-squared value, therefore it was selected for further fitting. Next, another layer of explanatory variables were added to the linear model. Here, only adding Lawyers resulted in a model that had all significant parameters, therefore it was the final model that we chose.

Step down method:

c) Check the model assumptions (of the resulting model from b)) by using relevant diagnostic tools.

```
right_plot = lm(expend~lawyers+employ , data=regression_data)
par(mfrow=c(2,2))
plot(fitted(right_plot), residuals(right_plot))
plot(right_plot, 1)
plot(right_plot, 2)
```


residuals(right_plot)
500
−500
0
1000
3000
5000
fitted(right_plot)
Residuals vs Fitted
Residuals
1000
−500
5
15
44
0
1000
3000
5000
Fitted values
Normal Q−Q
Standardized residuals
6
2
−2
5
8
15
−2
−1
0
1
2
Theoretical Quantiles