EDDA - Assignment 2 - Group 77

Dante de Lang, Ignas Krikštaponis and Kamiel Gülpen

Exercise 1

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.

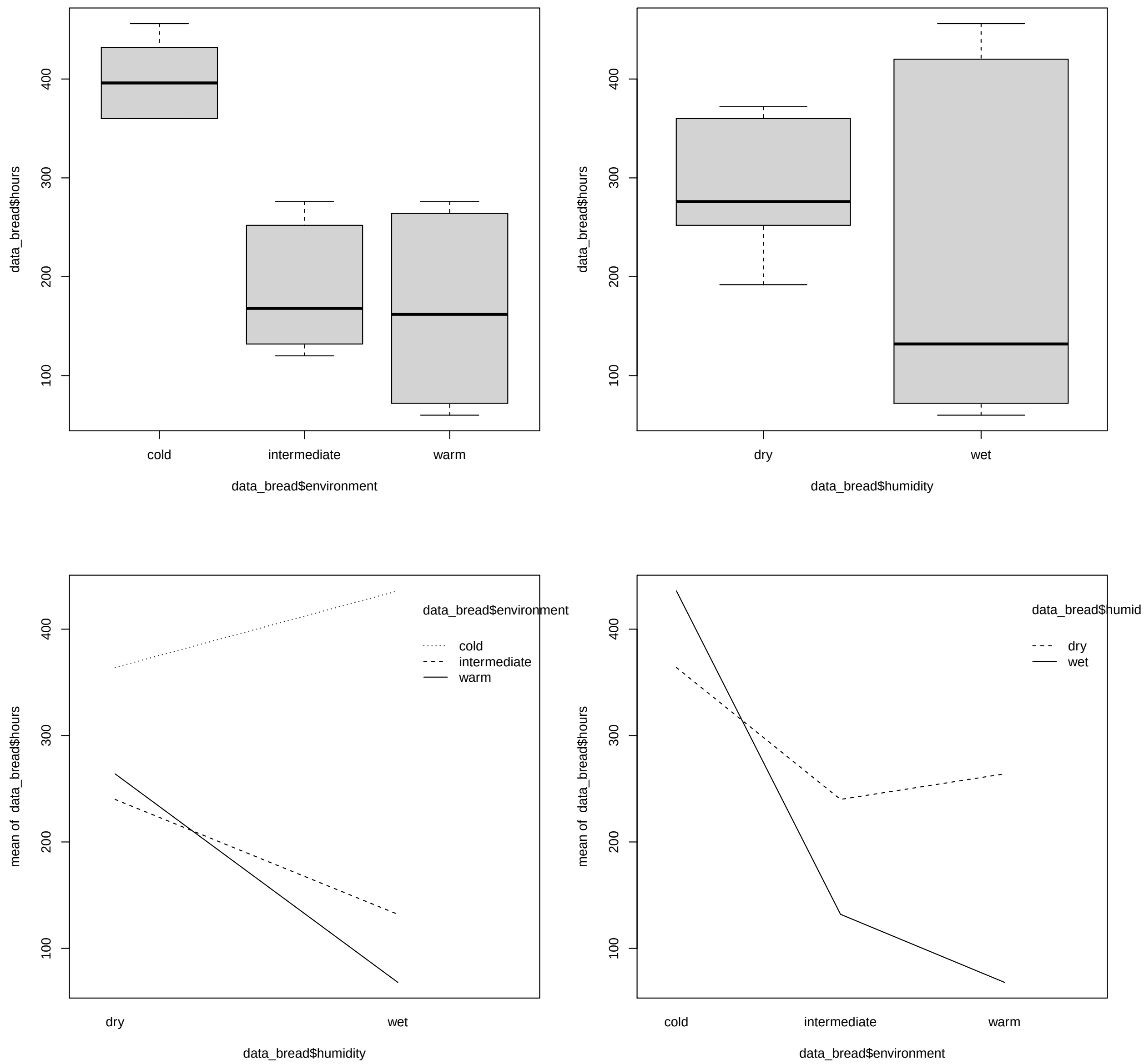
```
data_bread <- read.table(file="data/bread.txt", header=TRUE)
humid <- factor(rep(c("dry","wet"),each = 9))
temp <- factor(rep(c("cold", "intermediate","warm"), times = 6))
knitr::kable(data.frame(humid,temp,slice = sample(1:18)),
             caption = "Randomised combinations")
```

Table 1: Randomised combinations

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

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

```
par(mfrow=c(2,2))
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 in 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}.

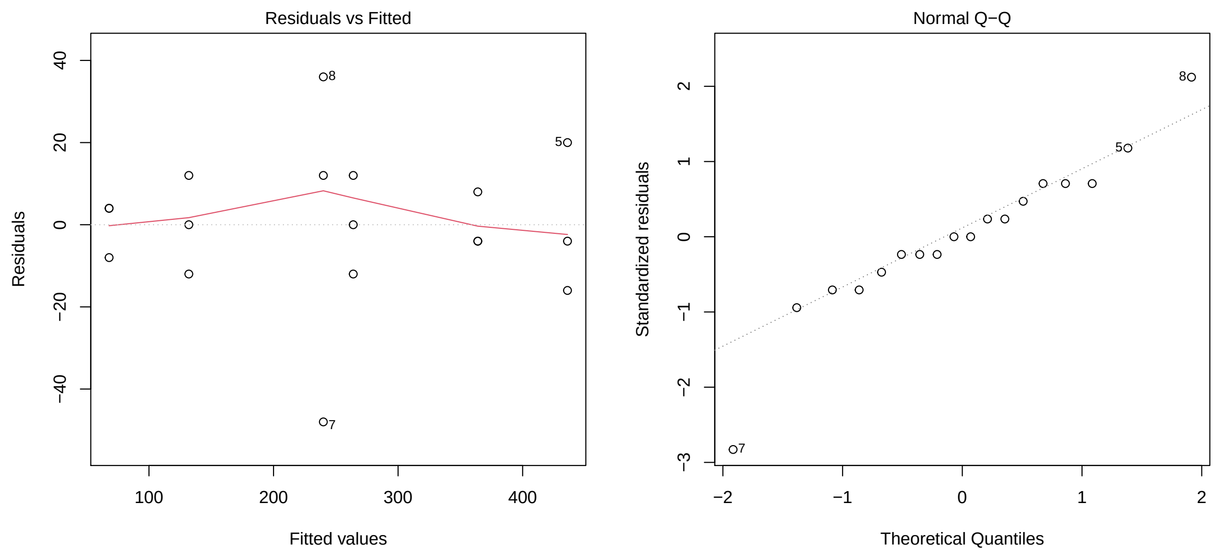
The interaction effect looks at the difference of differences, for example: it looks at the difference in hours for environment = cold and environment = warm for humidity = wet. Then it looks at the difference between environment = cold and environment = warm for humidity = dry. It then looks at the 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?

To answer this question we would need to construct an additive model. However, in *c*) we concluded that there is significant interaction between the two factors, therefore, it is not proper to simply decompose the effects of the two factors with the additive model - the question is not correct.

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 we can conclude that the data is normally distributed - however, there are some outliers at the extremes marked with 5, 7 and 8. We also looked at the Residuals vs Fitted plot, which showed no obvious relationship - which is an acceptable behavior.

Exercise 2

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 <- sample(c(1:15))
knitr::kable(data.frame(student,skill,interface), caption = "Randomised block design")
```

Table 2: Randomised block design

student	skill	interface
7	1	1
3	2	1
10	3	1
6	4	1
12	5	1
8	1	2
15	2	2
9	3	2
5	4	2
13	5	2
4	1	3
11	2	3
2	3	3
14	4	3
1	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) - therefore, the search times are not the same between the interfaces. Furthermore, the summary shows that interface 3 gives the highest α parameter value, making search time the longest for this interface type. For the shortest search time, interface 1 can be combined with skill levels 1, 2 or 3 since all three have the lowest β parameter values with 2 and 3 not being significantly different from skill level 1. If significance is disregarded, combination of interface 1 with skill level 1 would give the shortest search time.

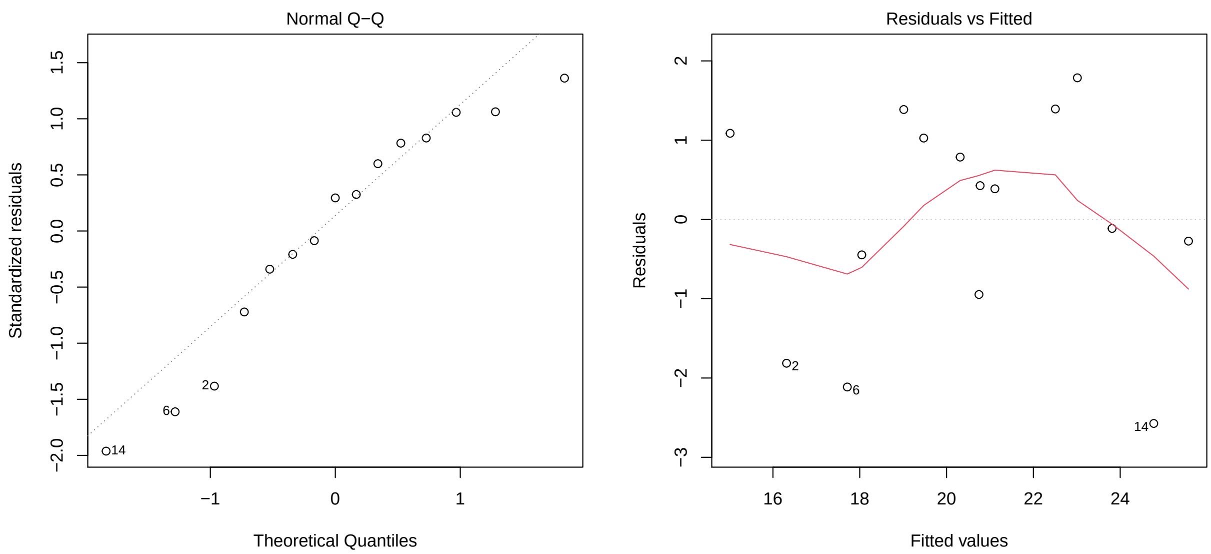
For the estimation of time it takes a typical user of skill level 3 using interface 3 we can calculate Y by summing the estimates and adding the error, giving a time of 23.97 units:

```
# Estimate interface 3 and skill 3:
options(digits=10)
Y = 15.01+4.46+3.03+1.47; Y
```

```
## [1] 23.97
```

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

```r
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. There seems to be some relationship visible, albeit small, in the Residuals vs Fitted plot - it would be a good idea to perform an extra test suitable for non-normal data.

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

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

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

The test shows a p-value < 0.05 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?

```r
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
```

Looking at the p-value (> 0.05) of the one-way ANOVA test, we see that it 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 randomised block design, it is not correct to use this test since it leaves out important effects of the skill level, which we observed to be significant in *b*) and *d*).

Excercise 3

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",]
```

```
## 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
```

```
print("Estimate for Treatment B:"); table
```

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

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

From the results of the 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 $\beta_{treatment_B}$ = -0.51 (meaning that with treatment B we have 0.51 less milk production than 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 effects" introduced by 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 of 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 (different 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 - a 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

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 H_0 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 are 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
```

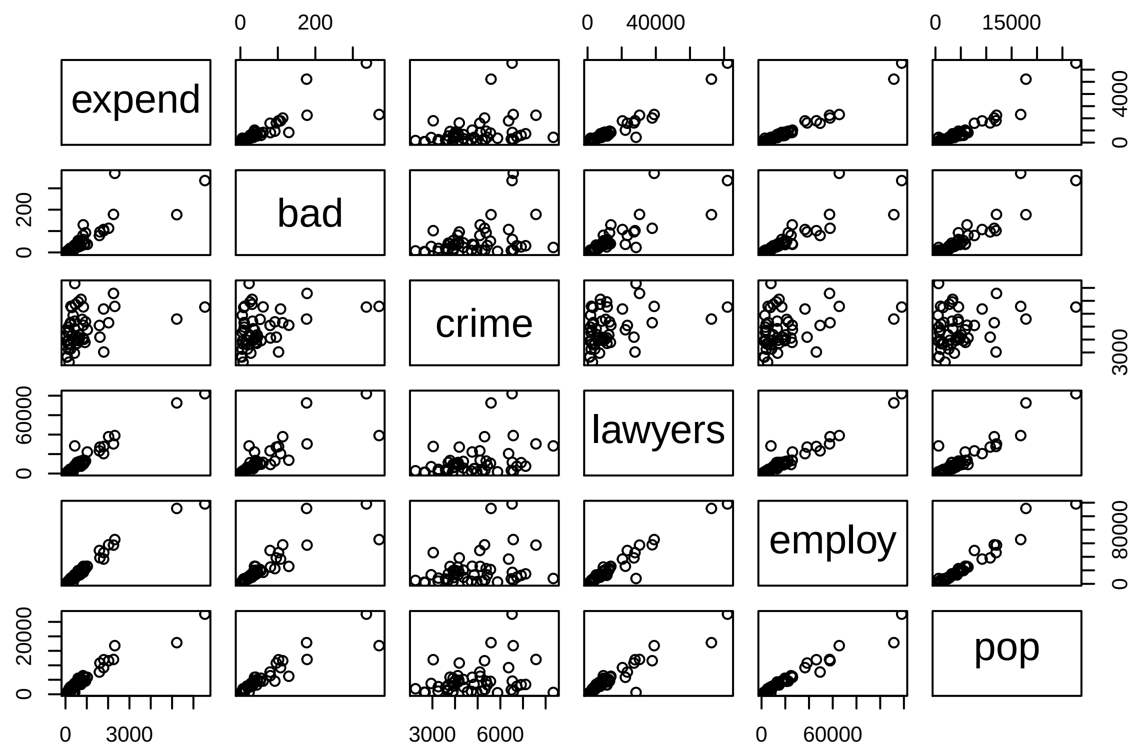
```
##           Sense      Emma  Sand1   Sand2
## a        -1.015 -0.112093  1.606 -0.0589
## an       -0.591 -1.219955 -1.067  3.7282
## this      0.139  0.390490 -0.444 -0.3267
## that      1.594  1.179849 -0.910 -3.0493
## with     -0.512  0.000192 -1.025  1.7482
## without   1.392 -1.341196  1.137 -1.0696
```

The fan is inconsistent as the p-value of the test is below 0.05. Therefore, we have to reject the H_0 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

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

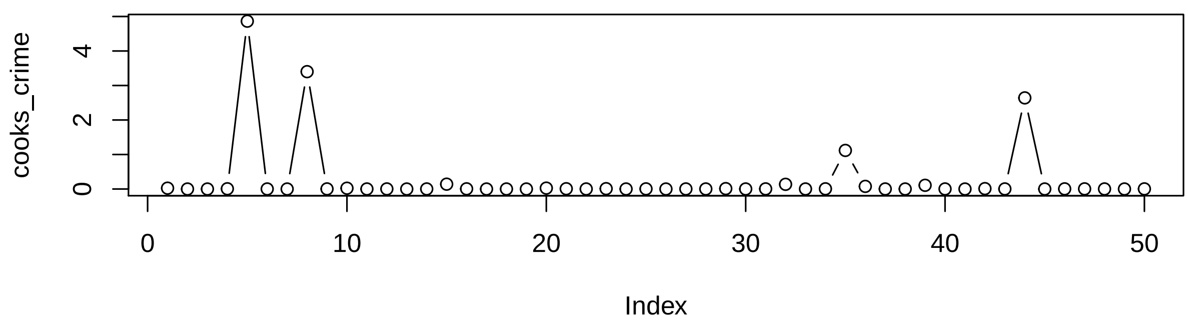
```
data_crime = read.table(file="data/expensescrime.txt",header=TRUE)
regression_data = data_crime[2:7]; pairs(regression_data)
```



From the pairs plot above we can clearly see linear relationships between predictors *bad*, *lawyer*, *employ* and *pop* - this could cause a problem of collinearity. Looking at predictors vs *expend* we can see a strong linear relationship with *bad*, *lawyers*, *employ* and *pop*; no obvious relationship can be observed with *crime*.

Influence points

```r
cooks_crime <- cooks.distance(lm(expend~crime + bad + lawyers + employ + pop,
                                 data = regression_data))
plot(cooks_crime, type="b");
```




```r
# remove influence points from the data
remove <- which(cooks_crime > 1);
regression_data <- regression_data %>% mutate(row_n = row_number()) %>%
  filter(!(row_n %in% remove)) %>% select(-row_n)
```

Model containing all of predictors was chosen to find influence points as we have not yet selected which predictors will be used in the final model. From the figure above we can see that points 5, 8, 35 and 44 have Cook's distance of > 1, therefore they were regarded as influence points and removed from the data.

Collinearity

```r
knitr::kable(round(cor(regression_data),2), caption = "Correlation coefficients")
```

Table 3: Correlation coefficients

	expend	bad	crime	lawyers	employ	pop
expend	1.00	0.90	0.34	0.96	0.98	0.96
bad	0.90	1.00	0.33	0.85	0.90	0.91
crime	0.34	0.33	1.00	0.25	0.25	0.18
lawyers	0.96	0.85	0.25	1.00	0.97	0.95
employ	0.98	0.90	0.25	0.97	1.00	0.97
pop	0.96	0.91	0.18	0.95	0.97	1.00

Correlation table above produced results similar to what was observer with pairs plot before - we see that (excluding variable *crime*), the lowest correlation coefficient is 0.85 (between *bad* and *lawyers*), which is still a high correlation coefficient. This means that all predictors (excluding *crime*) carry similar information.

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:

```r
# level 1
model_bad <- lm(expend~bad, data=regression_data);
model_crime <- lm(expend~crime, data=regression_data)
model_lawyers <- lm(expend~lawyers, data=regression_data);
model_employ <- lm(expend~employ, data=regression_data)
model_pop <- lm(expend~pop, data=regression_data)
level1_models <- list(model_bad, model_crime, model_employ, model_lawyers, model_pop);
variables_1 <- c("bad", "crime", "employ", "lawyears", "pop")

r_squared_values_1 <- c()
for(model_1 in level1_models){
  r_squared <-  summary(model_1)$r.squared
  r_squared_values_1 <- c(r_squared_values_1, r_squared)
}

knitr::kable(data.frame(Predictor = variables_1, r.squared = r_squared_values_1) %>%
               mutate(Selected = if_else(r.squared == max(r.squared),"Yes", "No")),
             caption = "Step-up: Level 1")
```

Table 4: Step-up: Level 1

Predictor	r.squared	Selected
bad	0.817	No
crime	0.117	No
employ	0.955	Yes
lawyears	0.927	No
pop	0.930	No

```r
# level 2
model_bad_2 <- lm(expend~employ+bad, data=regression_data);
model_crime_2 <- lm(expend~employ+crime, data=regression_data)
model_pop_2 <- lm(expend~employ+pop, data=regression_data);
model_lawyers_2 <- lm(expend~employ+lawyers, data=regression_data)
level2_models <- list(model_bad_2, model_crime_2, model_lawyers_2, model_pop_2);
variables_2 <- c("employ+bad", "employ+crime", "employ+lawyears", "employ+pop")

r_squared_values_2 <- c()
significance_2 <- c()
for(model_1 in level2_models){
  is_significant <- sum(c(summary(model_1)$coefficients[,4] < 0.05)) > 2
  r_squared <-  summary(model_1)$r.squared
  r_squared_values_2 <- c(r_squared_values_2, r_squared)
  significance_2 <- c(significance_2, is_significant)
}

knitr::kable(data.frame(Predictor = variables_2, r.squared = r_squared_values_2,
  Significant = significance_2) %>% mutate(Selected = if_else(Significant,"Yes", "No")),
  caption = "Step-up: Level 2")
```

Table 5: Step-up: Level 2

Predictor	r.squared	Significant	Selected
employ+bad	0.957	FALSE	No
employ+crime	0.964	TRUE	Yes
employ+lawyears	0.960	FALSE	No
employ+pop	0.960	FALSE	No

```
# level 3

model_bad_3 <- lm(expend~employ+crime+bad, data=regression_data);
model_pop_3 <- lm(expend~employ+crime+pop, data=regression_data);
model_lawyers_3 <- lm(expend~employ+crime+lawyers, data=regression_data);
level3_models <- list(model_bad_3, model_lawyers_3, model_pop_3);
variables_3 <- c("employ+crime+bad", "employ+crime+lawyers", "employ+crime+pop")

r_squared_values_3 <- c()
significance_3 <- c()
for(model_1 in level3_models){
  is_significant <- sum(c(summary(model_1)$coefficients[,4] < 0.05)) > 3
  r_squared <-  summary(model_1)$r.squared
  r_squared_values_3 <- c(r_squared_values_3, r_squared)
  significance_3 <- c(significance_3, is_significant)
}

knitr::kable(data.frame(Predictor = variables_3, r.squared = r_squared_values_3,
  Significant = significance_3) %>%
    mutate(Selected = if_else(Significant & r.squared == max(r.squared),"Yes", "No")),
  caption = "Step-up: Level 3")
```

Table 6: Step-up: Level 3

Predictor	r.squared	Significant	Selected
employ+crime+bad	0.965	FALSE	No
employ+crime+lawyers	0.970	TRUE	No
employ+crime+pop	0.974	TRUE	Yes

Step up method: First we started by fitting a linear model with one explanatory variable. Out of the 5 available variables *employ* performed the best according to r-squared value, therefore it was selected for further fitting. Next, another layer of explanatory variables was added to the linear model. Here, only adding *crime* resulted in a model that had all significant parameters, therefore it was the model we carried on with. Next, we added a third layer of variables. *pop* seemed to give the best results according to r-squared values (and also had all significant parameters). However, looking back at previous analysis, we know that *pop* and *employ* are strongly correlated and it would not make sense to have them in the same model, therefore we performed VIF analysis:

```
# VIF analysis
vif(model_pop_3)
```

```
## employ  crime    pop
##   17.87   1.14  17.31
```

From the VIF analysis we see that indeed having these two parameters should be avoided as VIF for them is > 5. We removed *pop* from the model even though it had lower VIF - removing *employ* resulted in lower r-squared value. Removal of this variable did not have a high impact on the r-squared value as addition of it only brought marginal improvement. VIF of the model without *pop* (as seen below) is now acceptable:

```
vif(model_crime_2)
```

```
## employ  crime
##   1.07   1.07
```

Therefore, the final model from the step-up method is as follows:

```
summary(model_crime_2)$coefficients
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -165.9828   59.78132   -2.78 8.11e-03
## employ         0.0363    0.00113   31.96 1.33e-31
## crime          0.0432    0.01280    3.38 1.57e-03
```

$expend = -165.9828 + 0.0363 \times employ + 0.0432 \times crime + error$

Step-down method:

```
# step down method - start with all
summary(lm(expend~bad+crime+lawyers+employ+pop, data=regression_data))$coefficients
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -241.6684   60.15675 -4.0173 0.000252
## bad           -0.0321    0.94970 -0.0338 0.973216
## crime          0.0531    0.01214  4.3772 0.000084
## lawyers        0.0118    0.00659  1.7888 0.081216
## employ         0.0162    0.00481  3.3606 0.001720
## pop            0.0625    0.02100  2.9765 0.004930
```

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

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -240.8488   54.37368   -4.43 6.87e-05
## crime          0.0530    0.01106    4.79 2.21e-05
## lawyers        0.0119    0.00615    1.93 6.03e-02
## employ         0.0161    0.00454    3.55 9.79e-04
## pop            0.0622    0.01836    3.39 1.57e-03
```

```
# remove lawyers
summary(lm(expend~crime+employ+pop, data=regression_data))$coefficients
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -252.6057   55.76140   -4.53 4.82e-05
## crime          0.0548    0.01138    4.81 1.95e-05
## employ         0.0207    0.00399    5.19 5.66e-06
## pop            0.0727    0.01810    4.02 2.40e-04
```

Step-down method: We started with a model that used all of the available predictors. From the summary table we see that *bad* had the highest (and insignificant) p-value - therefore, it was removed. Next we see that *lawyers* was the only insignificant parameter and therefore, was removed from the model. The resulting model of $crime + employ + pop$ has all significant parameters and is the same as produced with the step-up method. However, as before, *pop* should be removed as it is colinear with *employ*.

Conclusion Both step-up and step-down methods bring us to the same model of $employ + crime$. However, we can see that both models posses a negative intercept. Conceptually, this does not make sense as the expenditure of the government can not be negative even if all of the predictors are 0. Taking this into consideration, in addition to Occam's razor principle and the fact that introduction of extra predictor (*crime*) to *employ* did not bring substantial improvement to r-squared (only $\approx +0.01$) we choose the following model as the final model:

```
summary(lm(expend~employ, data=regression_data))$coefficients
```

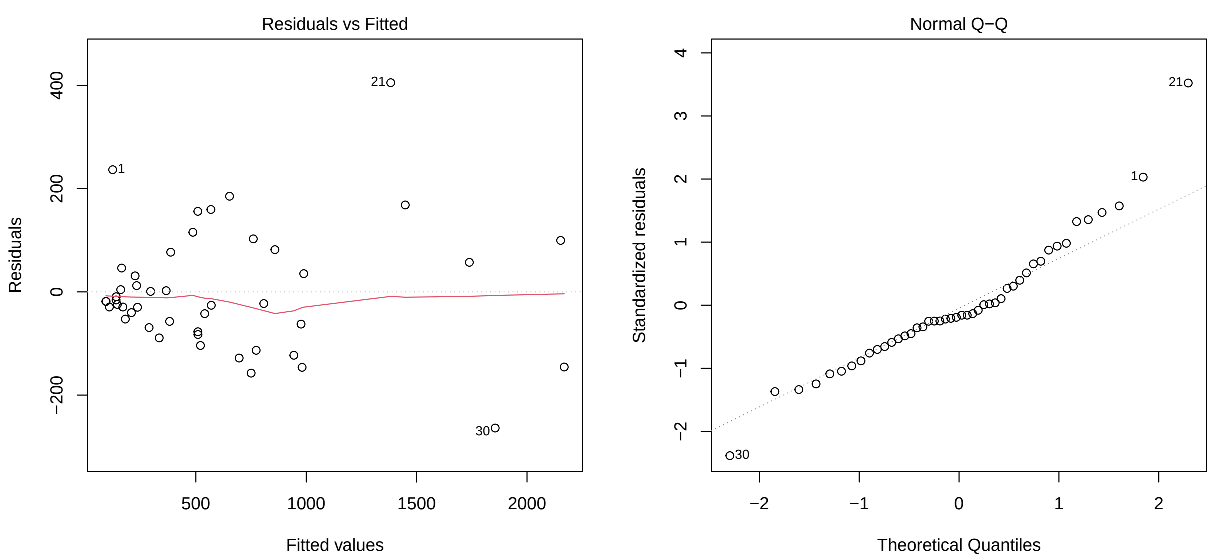
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.1763   26.42996   0.726 4.72e-01
## employ        0.0372    0.00122  30.510 3.07e-31
```

Even though the intercept parameter is insignificant, it is in principle not always a problem and could mean that the model starts at the origin. However, it will not be removed as we do not have enough information to make a concrete decision on this. The final model is as follows:

$expend = 19.1763 + 0.0372 \times employ + error$

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

```
right_plot = lm(expend~employ , data=regression_data)
par(mfrow=c(1,2)); plot(right_plot, 1); plot(right_plot, 2)
```



From the diagnostic plots above:

- Normal qq-plot: residuals follow a straight line pretty well, however there are some outliers at the extremes - removing them could improve the results.
- Residuals vs fitted: as desired, there does not seem to be any trend.