Assignment 2 EDDA

Daan Moll (VU: 2559766 / UvA: 11929332) Ramon Bussing (VU: 2687261 / UvA: 10719482) Karim Semin (VU: ksn222 / UvA: 11285990)

Group number: 49

For all exercises, an significance level of .05 is used, unless stated otherwise.

### Exercise 1: Moldy bread

**1a)** Firstly the bread.txt file is loaded in.

bread = read.table("bread.txt",header=TRUE)

The randomization process of the data is done as followed:

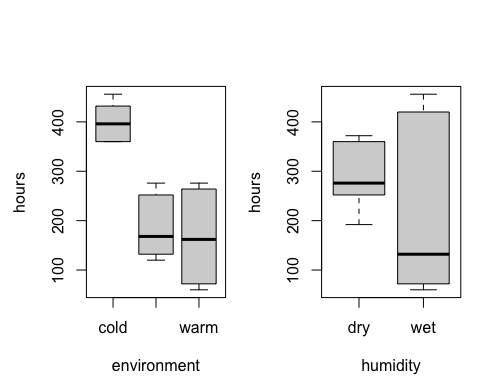
I=3; J=2; N=3  
rbind(rep(1:I,each=N\*J),rep(1:J,N\*I),sample(1:(N\*I\*J)))

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

Here, you can see that all the participants (in this case, slices of bread) are assigned a random condition. In the table above, the first row contains the environment (1 = cold, 2 = intermediate, 3 = warm), the second row contains the humidity (1 = dry, 2 = wet) and the third row contains the participant ID’s (the specific slice of bread). The last row is thus randomized across the different conditions.

**1b)** Next, we create two boxplots of the hours to decay in relation to the two factors.

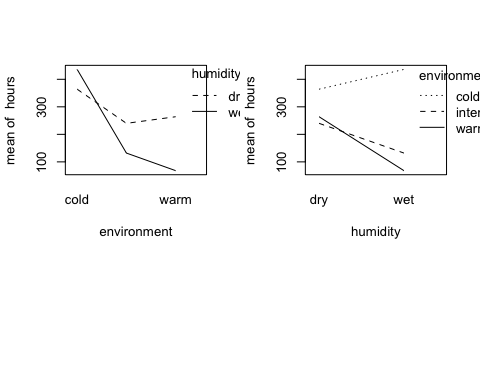
attach(bread)  
par(mfrow=c(1,2), mar=c(4,4,4,1), oma=c(0.5,0.5,0.5,0))  
boxplot(hours~environment)  
boxplot(hours~humidity)

 In the left boxplot we observe the distribution of the dependent variable (hours to decay) for the three environment conditions. Here, it can be observed that the median for the intermediate and warm environment is similar. The median for the bread in the cold environment however is higher than the other two conditions. This probably implies that placing bread in a cold environment will take longer to become moldy.

In the right plot, the distribution of data is plotted for the humidity condition. Here, it can be seen that the median for the hours to decay is lower for wet humidity than for dry humidity. Moreover, a larger spread in data is observed for the wet environment than for the dry environment.

It should be noted that these boxplots only give us an idea of the main effects of the factors can be observed in the boxplot. Therefore interaction plots for both factors are added below.

layout(matrix(c(1,2,3,3), ncol=2, byrow=TRUE), heights=c(2, 1))  
interaction.plot(environment, humidity, hours)  
interaction.plot(humidity, environment, hours)

 In the plots it can be noted that the lines are not parallel. In the left interaction plotm, we note that in the wet condition, the temperature looks to have a larger effect on the hours of decay than in the dry condition. In the right interaction plot, we find a higher hours to decay in the wet condition for the cold environment than in the dry condition. For the intermediate and warm environment, a vice versa effect is found; the mean hours to decay is lower in the wet condition than in the dry condition. Thus, an interaction effect is found. However, noise might also be the cause of dissimilarities in the line slopes. To further investigate the significance of interaction between factors, a two way ANOVA is appropriate.

**1c)** A two way ANOVA is performed to test the effect of the factors and their interaction.

We define the alternative hypothesis as followed: - There is an effect of environment’s temperature on the decay of bread () - There is an effect of humidity on the decay of bread () - There is an interaction effect between the environments temperature and humidity on the decay of bread ()

To research these hypotheses, we need to determine the , and .

bread$environment=as.factor(bread$environment)  
bread$humidity=as.factor(bread$humidity)  
  
breadov=lm(hours~environment\*humidity)  
anova(breadov)

## 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

The p-value for for both factors and is <0.001, which is lower than our significance value (.05). From this, we can conclude that the variation between different environments is larger than within environments. The same can be concluded for humidity; the variation between a dry and wet environment is higher than within those two environments. Therefore, a main effect is found for both variables and we may accept that there is a significant relationship between a bread’s environment and its decay, as well as for ; there is a significant relationship between the humidity and decay of bread (within our confidence interval).

The p-value for the interaction effect between the factors too is <0.001. From this, we may accept the alternative hypothesis that there is a interaction effect between the environment’s temperature and humidity on the decay of the bread. This confirms our findings in the interaction plot in 1c), where we observed that the lines are not parallel to each other.

**1d)** To answer the question of which of the two factors has the greatest (numerical) influence on the decay, the additive model is generated:

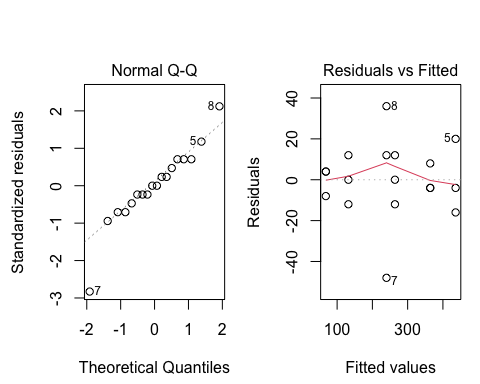
breadov1=lm(hours~environment+humidity)  
anova(breadov1)

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

Based on the additive ANOVA model of the two factors, it can be noted that the (23.11) is higher than (6.16). From this, we may conclude that more of the variance is explained by the environment’s temperature than by humidity within their respective groups. Therefore, environment might have a greater influence on the decay than humidity. However, since an interaction effect was found between the two factors, this conclusion cannot be made. Since an interaction effect was found, the relationship between the environment/humidity and hours to decay may be influenced by a third factor. Therefore “which of the two factors has the greatest (numerical) influence on the decay” is not a good question in this case.

**1e)** Lastly, we check the model assumptions using relevant diagnostic tools. These assumptions consist of normality and equal variances of the residuals.

par(mfrow=c(1,2), mar=c(4,4,4,1), oma=c(0.4,0.4,0.4,0))  
plot(breadov,2)  
plot(breadov,1)

 In the Q-Q plot, normality looks dubious. Moreover, the spread of variance looks slightly unstable for some fitted values, which implies that there may be a relationship between residuals and fitted values. Therefore, homogeneity of variance cannot be assumed. This is caused by some of the data points (marked with 5, 7 and 8), which look to be extreme. These datapoints also have an affect on the normality of the data. We may consider these outliers and trim the data of these values.

### Exercise 2: Search engine

students = read.table("search.txt",header=TRUE)

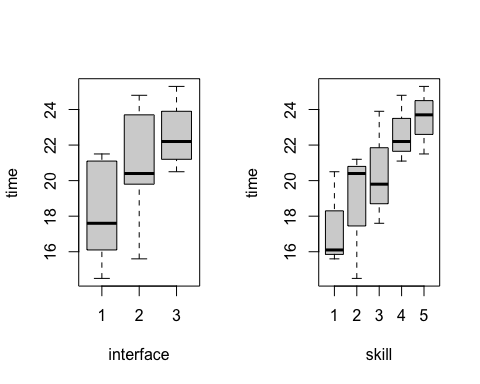
**2a)** Firstly, we randomize the students to the interfaces in a randomized block design.

## [1] 3 2 1  
## [1] 2 3 1  
## [1] 1 2 3  
## [1] 2 1 3  
## [1] 1 3 2

The 5 different blocks (rows) show how the experiment could be conducted by using the randomized block design. Each block is one of the 5 skill levels and it tells us how the 3 students of the skill level should be divided over the 3 interfaces. For example, for block (skill level) 1 student 2 is assigned to interface 3, student 3 is assigned to interface 1 and student 1 to interface 2.

**2b)** Then we test the null hypothesis that the search time is the same for all interfaces.

students$interface = as.factor(students$interface)  
students$skill = as.factor(students$skill)  
attach(students)  
  
par(mfrow=c(1,2))  
boxplot(time~interface, fig.height = 3, fig.width = 3, fig.align = "center")  
boxplot(time~skill, fig.height = 3, fig.width = 3, fig.align = "center")



studentslm = lm(time~skill+interface)  
anova(studentslm)

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

In the boxplot, it can be observed that the medians for all interfaces and skill levels differ from each other. The significance of these difference can be found in the output of the two-way ANOVA. This gives us a p-value for interface of .013 this is below the alpha significance level of .05. This means that we reject the null hypothesis that the search time is the same for all interfaces.

Next, we examine which type of interface requires the longest search time.

inter1 = mean(students$time[interface==1])  
inter2 = mean(students$time[interface==2])  
inter3 = mean(students$time[interface==3])  
which.max(c(inter1, inter2, inter3))

## [1] 3

We find that interface 3 has the highest mean search time.

Moreover, we explore which combination of skill and interface has the shortest search time.

students[which.min(students$time),]

## time skill interface  
## 2 14.5 2 1

The combination of skill and interface that leads to the shortest search time is skill level 2 and interface 1.

The time it takes a typical user of skill level 3 to find the product on the website if the website uses interface 3 is then estimated.

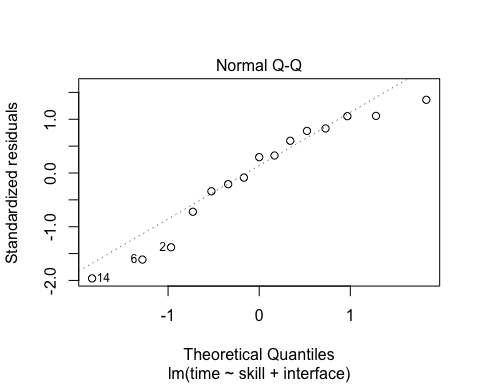
summary(studentslm)

##   
## Call:  
## lm(formula = time ~ skill + interface)  
##   
## 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 \*\*\*  
## 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 \*\*   
## interface2 2.70 1.14 2.38 0.0447 \*   
## interface3 4.46 1.14 3.93 0.0044 \*\*   
## ---  
## 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

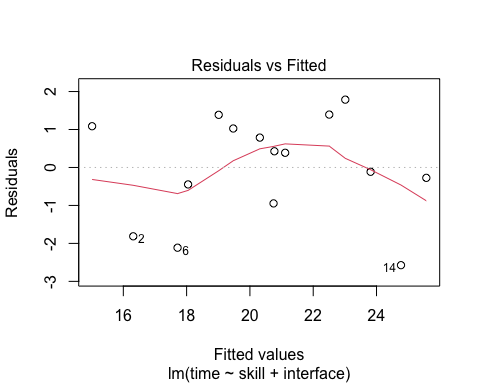
From the summary, we can calculate the answer to our question. This results in the following formula: time = 15.01 + 4.46 + 3.03 This results in the estimation that it takes a user of skill level 3, 22.5 to find a product on the website when using interface 3.

**2c)** Then, we check the model assumptions by using relevant diagnostic tools.

plot(studentslm,2)

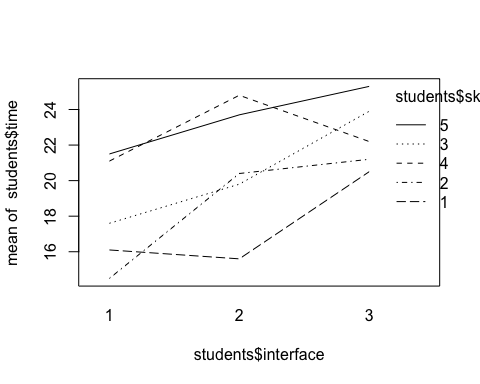


plot(studentslm,1)

 The Q-Q plot seems to deviate from normal. Furthermore, the spread of variance looks slightly unstable for some fitted values, which implies that there may be a relationship between residuals and fitted values. Therefore, homogeneity of variance is questionable.

**2d)** Then, we perform the Friedman test to test whether there is an effect of the interface on search time.

interaction.plot(students$interface, students$skill, students$time)



friedman.test(time,interface,skill,data=students)

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

The Friedman test gives us a p-value of .04. This is smaller than the significance level alpha (.05), therefore we reject H0, implying that the interface has an effect on the search time.

**2e)** Lastly, we test the null hypothesis that the search time is the same for all interfaces with a one-way ANOVA test, ignoring the variable skill.

studentsaov = lm(time ~ interface, data=students)  
  
anova(studentsaov)

## 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

# summary(studentsaov)

The p-value > .05, therefore we would not reject , meaning that we don’t find a significant difference between interfaces. It may be incorrect to use a one-way ANOVA test, because the skill variable may not be spread equally (uniformly) across the population of people who would use the interfaces, meaning that it may not represent the society. Therefore, if only interface would be tested, a random (and preferable larger) sample should be taken from the population.

### Exercise 3. Feedingstuffs for cows

**3a)** We first test whether the type of feedingstuffs influences milk production using an ordinary “fixed effects” model, fitted with lm. To do this, we first read in the data and define the different factors.

Then, an ANOVA is performed for the different factors.

anova(lm(milk~id+per+treatment))

## 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

From this model we can’t conclude that feedingstuffs A and B have an effect on milk production with the fixed effect model since the corresponding ANOVA p-value is 0.517.

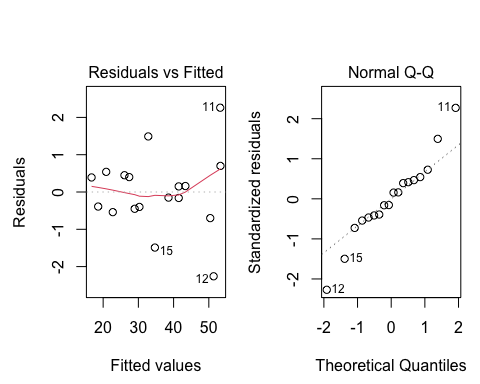
summary((lm(milk~id+per+treatment)))

##   
## Call:  
## lm(formula = milk ~ id + per + treatment)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.260 -0.438 0.000 0.438 2.260   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 30.300 1.244 24.35 5.0e-08 \*\*\*  
## 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 -7.050 1.574 -4.48 0.00287 \*\*   
## id6 23.450 1.574 14.90 1.5e-06 \*\*\*  
## id7 13.550 1.574 8.61 5.7e-05 \*\*\*  
## id8 4.900 1.574 3.11 0.01701 \*   
## id9 -11.200 1.574 -7.12 0.00019 \*\*\*  
## 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

In the given data Feedingstuff B is estimated to give 0.510 times less milk than feedingstuff A (that has an estimate of 30.30 milk production).

Check assumptions:

par(mfrow=c(1,2), mar=c(4,4,4,1), oma=c(0.5,0.5,0.5,0))  
plot(lm(milk~id+per+treatment), 1)  
plot(lm(milk~id+per+treatment), 2)

 This plot does show a relationship between the fitted values and the residuals as the data does not follow the dotted line, allowing us to conclude that the assumption for equality of variances is not met. In addition to this, the assumption for normality is also not met as the data in the normal Q-Q plot does not follow the dotted line.

**3b)** We repeat a) by performing a mixed effects analysis, modeling the cow effect as a random effect (use the function lmer). We compare our results to the results found by using a fixed effects model.

cowlmer = lmer(milk~order+per+treatment+(1|id))  
cowlmer\_reduced = lmer(milk~order+per+(1|id), REML=FALSE)  
(anova(cowlmer\_reduced, cowlmer))

## refitting model(s) with ML (instead of REML)

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

In the mixed effects model the p value of 0.446 is lower than the fixed effects model but still not significant. This could have something to do with an interaction effect.

**3c)** We study the commands: - attach(cow) - t.test(milk[treatment==“A”],milk[treatment==“B”],paired=TRUE) 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)

##   
## 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

These commands do not produce a valid test for a difference in milk production because the assumptions for the t-test are not checked in these two commands. In addition to this, due to the possibility of interaction effects the paired t-test should not be used but rather the mixed effects ANOVA, which takes these into account.

cowlmer = lmer(milk~order+treatment+per+(1|id))  
cowlmer\_reduced = lmer(milk~order+treatment+(1|id), REML=FALSE)  
(anova(cowlmer\_reduced, cowlmer))

## refitting model(s) with ML (instead of REML)

## Data: NULL  
## Models:  
## cowlmer\_reduced: milk ~ order + treatment + (1 | id)  
## cowlmer: milk ~ order + treatment + per + (1 | id)  
## npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)   
## cowlmer\_reduced 5 125 130 -57.7 115   
## cowlmer 6 119 125 -53.7 107 8.12 1 0.0044 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

When we check for the period effect we indeed find that the period has significant effect with a p=0.00439. The paired t-test is not valid for our data because the two orders (8 times AB and 10 times BA) are not equally distributed over the cows and the period of measuring has a significant effect as shown before, meaning that the paired t-test not just includes the difference in milk production but also the significant period effect.

Although the p-values for the fixed model and the paired t-test are both non-significant, this conclusion is not compatible with the one obtained in a) because the paired t-test is the equivalent (gives the same p-value of 0.83) to the anova as shown below:

anova(lm(milk~treatment+id))

## Analysis of Variance Table  
##   
## Response: milk  
## Df Sum Sq Mean Sq F value Pr(>F)   
## treatment 1 0 0.3 0.05 0.83   
## id 8 2467 308.4 57.74 2.8e-06 \*\*\*  
## Residuals 8 43 5.3   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

which is not equal to the fixed model anova used in 3a that also includes the period.

### Exercise 4: Jane Austen

**4a)** Discuss whether a contingency table test for independence or for homogeneity is most appropriate here.

In order to analyze the word count of the chapters of the finished novel, we look to see whether the word count in the chapter written by her admirer are consistent with the word count written by Jane Austen’s herself. We therefore test the null hypothesis that the proportion of word counts are consistent between Austin’s original work and the finished work by her admirer. Because of this, a contingency table for homogeneity is appropriate. With a contigency table for homogeinity we can check whether the distribution of word counts are equal between the original work of Austen (Sense, Emma, Sand1) and her admirer (Sand2).

**4b)** Using the given data set, investigate whether Austen herself was consistent in her different novels. Where are the main inconsistencies?

To investigate whether Austen’s work was consistent in her different novels, we regard the Sense, Emma and Sand1 column of the dataset. We then calculate the sum of each row and column and the total word count. With these sums of the rows and columns and the total word count, we can calculate the expected value for each cell. These computations are done by the build in chisq.test function of R. The result of the Chi-squared test can be observed below.

wordcounts = read.table("austen.txt", header=TRUE)  
austen = subset(wordcounts, select = c('Sense', 'Emma', 'Sand1'))  
z=chisq.test(austen)  
z

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

The p-value is .3. Based on this, we will not reject the null hypothesis that the proportion of word counts differs significantly between Austen’s book chapters. This is to be expected, as these chapters are all written by the same author. To consider inconsistencies in the rows and columns, we observe the residuals of the cells (the square root contribution of each cell to the chi-square statistic.) These values give an indication of how much cell values differ from the expected values under .

z$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 table, inconsistencies with regards to the expected wordcounts can be observed. The main inconsistencies are listed below: - The word ‘a’ was used more often in Sanditon than in Sense and Sensibility book. - The word ‘without’ was used much less in Emma as opposed to the other two books. - The word ‘that’ was used less in Sanditon than in her other two books.

**4c)** Was the admirer successful in imitating Austen’s style? Perform a test including all data. If he was not successful, where are the differences?

Next, we check whether the admirer successful in imitating Austen’s style. We do this by adding back the column with the admirer’s word counts and once again performing a chi-square test.

z1=chisq.test(wordcounts)  
z1

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

The p-value is now smaller than .001. Based on this we will reject the null hypothesis. This implies that after adding back the Sand2 column, we can assume that the proportion of word counts are not equal between cells. We can investigate where these differences lay if we once again investigate the contribution of each to the chi-square test statistic with the residuals.

z1$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

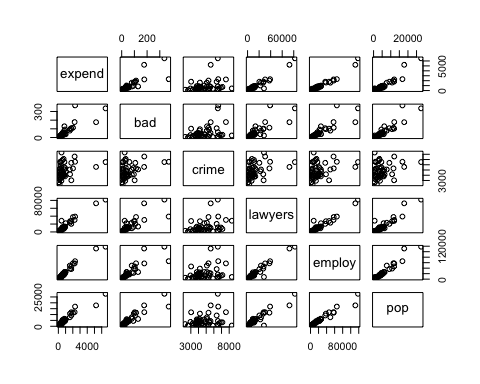
From the table we can gather that these differences lay; - The admirer uses the word ‘an’ much more often than Austen relatively. - The admirer uses the word ‘that’ far less often than Austen relatively. - The admirer uses the word ‘with’ more often than Austen relatively. (In comparison to the expected word count under .)

### Exercise 5:

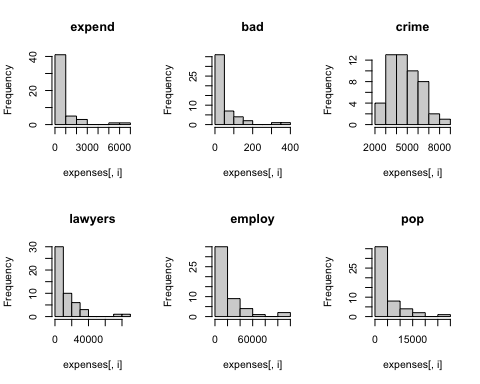
The data in expensescrime.txt were obtained to determine factors related to state expenditures on criminal activities (courts, police, etc.) The variables are: state (indicating the state in the USA), expend (state expenditures on criminal activities in $1000), bad (crime rate per 100000), crime (number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

**5a)** Make some graphical summaries of the data. Investigate the problem of potential and influence points, and the problem of collinearity.

plot(expenses[,c(2,3, 4, 5, 6, 7)])

 When looking at these scatter plots some potential points are clearly visible. An example is the data point most right in the expand-bad scatter plot.

par(mfrow=c(2,3))  
for (i in 2:7) {  
 hist(expenses[,i], main=names(expenses)[i])  
}



problem of influence points:

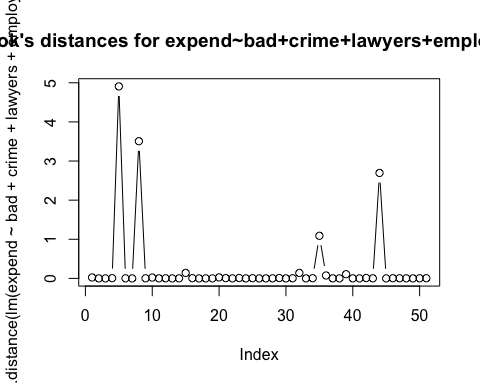
print("Cook's distances for lm(expend~bad+crime+lawyers+employ+pop):")

## [1] "Cook's distances for lm(expend~bad+crime+lawyers+employ+pop):"

round(cooks.distance(lm(expend~bad+crime+lawyers+employ+pop)), 2)

## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16   
## 0.02 0.00 0.00 0.01 4.91 0.00 0.00 3.51 0.00 0.02 0.00 0.00 0.00 0.00 0.14 0.01   
## 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32   
## 0.00 0.00 0.00 0.03 0.01 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.14   
## 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48   
## 0.00 0.00 1.09 0.07 0.00 0.00 0.11 0.00 0.00 0.01 0.00 2.70 0.00 0.00 0.00 0.00   
## 49 50 51   
## 0.00 0.00 0.00

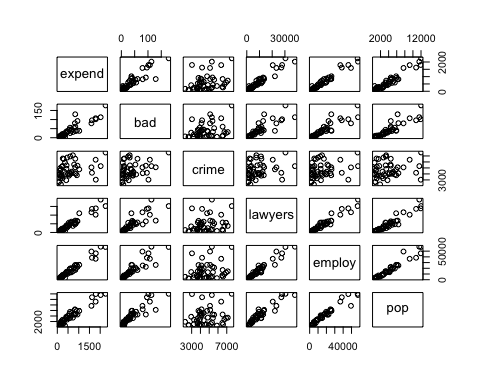
plot(cooks.distance(lm(expend~bad+crime+lawyers+employ+pop)),type="b", main="Cook's distances for expend~bad+crime+lawyers+employ+pop")

 The following influence points are identified using the thumb’s rule that identifies an influence point when Cook’s distances is almost 1 or higher: \* data point 5 with cook’s distance: 4.91 \* data point 8 with cook’s distance: 3.51 \* data point 35 with cook’s distance: 1.09 \* data point 44 with cook’s distance: 2.70

These data points are removed from the data:

expenses = expenses[-c(5, 8, 35, 44), ]  
attach(expenses)

plot(expenses[,c(2,3, 4, 5, 6, 7)])

 The removal of the influence points is clearly visible in the plot above compared to the same plot before the removal. For example, when looking at the expend-bad scatter plot, the outlying data points are removed, resulting in a better visible relation between the expand and bad variables.

Problem of collinearity:

round(cor(expenses[,c(2, 3, 4, 5, 6, 7)]), 2)

## expend bad crime lawyers employ pop  
## expend 1.00 0.91 0.35 0.96 0.98 0.96  
## bad 0.91 1.00 0.33 0.86 0.91 0.91  
## crime 0.35 0.33 1.00 0.25 0.26 0.19  
## lawyers 0.96 0.86 0.25 1.00 0.97 0.95  
## employ 0.98 0.91 0.26 0.97 1.00 0.97  
## pop 0.96 0.91 0.19 0.95 0.97 1.00

From the scatter plot above and correlation values we can see that the following explanatory variables are likely to be collinear since those combinations have data points that show a linear relation and have a high correlation coefficient: pop-employ pop-lawyers pop-bad employ-bad employ-lawyers

**5b)** step-up method

factors = colnames(expenses)[3:length(colnames(expenses))]  
  
# 1st factor  
R2 = numeric(length(factors))  
for (i in 1:length(factors)) {  
 R2[i] = summary(lm(paste('expend', '~', factors[i]), data=expenses))[[8]]  
}  
  
print("R^2 values for addition of first variable:"); factors; R2

## [1] "R^2 values for addition of first variable:"

## [1] "bad" "crime" "lawyers" "employ" "pop"

## [1] 0.820 0.120 0.928 0.956 0.931

factor\_one = factors[which.max(R2)]  
print(paste0("Factor with highest R^2 value: ", factor\_one))

## [1] "Factor with highest R^2 value: employ"

# remove chosen first factor from list  
factors = factors[-which.max(R2)]  
  
# 2nd factor  
R2 = numeric(length(factors))  
for (i in 1:length(factors)) {  
 R2[i] = summary(lm(paste('expend', '~', factor\_one, '+', factors[i]), data=expenses))[[8]]  
}  
  
print("R^2 values for addition of second variable:"); factors; R2

## [1] "R^2 values for addition of second variable:"

## [1] "bad" "crime" "lawyers" "pop"

## [1] 0.958 0.965 0.961 0.961

factor\_two = factors[which.max(R2)]  
print(paste0("Factor with highest R^2 value: ", factor\_two))

## [1] "Factor with highest R^2 value: crime"

summary(lm(expend~employ), data=expenses)

##   
## Call:  
## lm(formula = expend ~ employ)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -263.8 -65.8 -18.5 51.7 405.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 19.20280 25.61534 0.75 0.46   
## employ 0.03722 0.00119 31.15 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 118 on 45 degrees of freedom  
## Multiple R-squared: 0.956, Adjusted R-squared: 0.955   
## F-statistic: 970 on 1 and 45 DF, p-value: <2e-16

The step up method resulted in a model which used only the variable employ and a R^2 value of 0.956. The addition of a second parameter, as shown above, would only increase the R^2 value to a maximum of 0.965. However, this small increase is not worth the addition of another variable and thus the final model used one variable. This resulted in: expend = 23.242 + 0.037 \* employ + error with an R^2 value of 0.956

step-down method

## $coefficients  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -236.1257 59.06144 -3.9980 2.60e-04  
## bad -0.0166 0.94232 -0.0176 9.86e-01  
## crime 0.0526 0.01202 4.3760 8.12e-05  
## lawyers 0.0119 0.00655 1.8109 7.75e-02  
## employ 0.0163 0.00477 3.4275 1.40e-03  
## pop 0.0610 0.02072 2.9466 5.28e-03

## [1] "Crime has highest p-value of 9.86e-1, thus remove crime as variable"

## $coefficients  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 4.6535 25.68761 0.181 0.85712  
## bad 1.5694 1.04095 1.508 0.13913  
## lawyers 0.0176 0.00767 2.293 0.02695  
## employ 0.0178 0.00569 3.134 0.00314  
## pop 0.0255 0.02281 1.118 0.26990

## [1] "Pop has highest p-value of 0.270, thus remove pop as variable"

##   
## Call:  
## lm(formula = expend ~ bad + lawyers + employ, data = expenses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -249.2 -55.5 -10.8 40.0 387.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 16.58769 23.43268 0.71 0.48283   
## bad 2.00134 0.96940 2.06 0.04503 \*   
## lawyers 0.02068 0.00718 2.88 0.00620 \*\*   
## employ 0.02038 0.00523 3.89 0.00034 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 108 on 43 degrees of freedom  
## Multiple R-squared: 0.965, Adjusted R-squared: 0.962   
## F-statistic: 391 on 3 and 43 DF, p-value: <2e-16

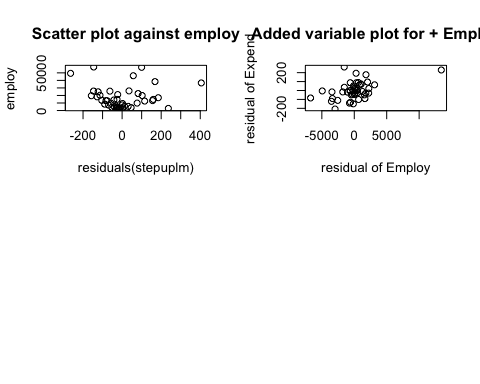
## [1] "Bad has highest p-value of 0.045, which is below 0.05"

All of the three left over variables, bad, lawyers and employ are smaller than 0.05, thus final model keeps these three factors. This results in: expend = 16.5877 + 2.001 \* bad + 0.020 \* lawyers + 0.020 \* employ + error with a R^2 value of 0.965. The step-up method resulted in the model that only uses one variable, employ. This resulted in the following model: expend = 23.242 + 0.037 \* employ + error with an R^2 value of 0.956.

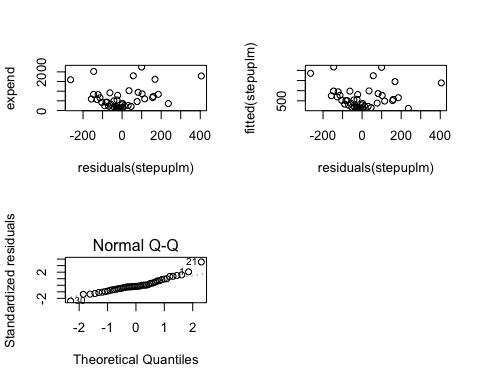
We prefer the model of the step-up method, because it uses one variables instead of three for a slightly lower R^2 value.

**5c)** The model assumptions will be checked for the model that resulted from the step up method.

attach(expenses)  
stepuplm = lm(expend~employ)  
  
# 2  
par(mfrow=c(2,2))  
plot(residuals(stepuplm), employ, main="Scatter plot against employ")  
  
# 3  
x = residuals(lm(employ~bad+crime+lawyers+pop))  
y = residuals(lm(expend~bad+crime+lawyers+pop))  
plot(x,y,main="Added variable plot for + Employ", xlab="residual of Employ", ylab="residual of Expend")  
  
# 4  
# par(mfrow=c(2,2))  
# plot(residuals(stepuplm), bad, main="Scatter plot against bad")  
# plot(residuals(stepuplm), crime, main="Scatter plot against crime")  
# plot(residuals(stepuplm), lawyers, main="Scatter plot against lawyers")  
# plot(residuals(stepuplm), pop, main="Scatter plot against pop")  
  
# 5  
par(mfrow=c(2,2))



plot(residuals(stepuplm), expend)  
plot(residuals(stepuplm),fitted(stepuplm))  
  
# 6  
plot(stepuplm, 2)

 The plot of the residuals against each variable not in the model did not show any linear model and this is why they are not shown. The qq-plot shows a non-linear line. Another step could be to try and transform the data. However, this did not result in any improvements and was thus not included.