# Experimental Design and Data Analysis

**Assignment 2**

Tommaso Castiglione Ferrari 2673807

Daniyal Selani 2692551

Simone Korteling 2671463

Group 71

---

## Exercise 1. Moldy bread

*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.*

In this research there is one numerical outcome: 'time to decay in hours' and there are two factors: 'environment' and 'humidify'. The factor environment consists of three levels ('cold', 'intermediate' & 'warm'). The factor 'humidified' consists of two levels ('dry' & 'wet'). The following code in R randomized factor A and factor B such that we have 18 observations with 6 different conditions:

Code:

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

Output:

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

So now we have for every condition three observations. For example for unit 3 we assign level 1 on factor A and
meaning this unit is assigned the condition of: 'cold' & 'dry'.

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

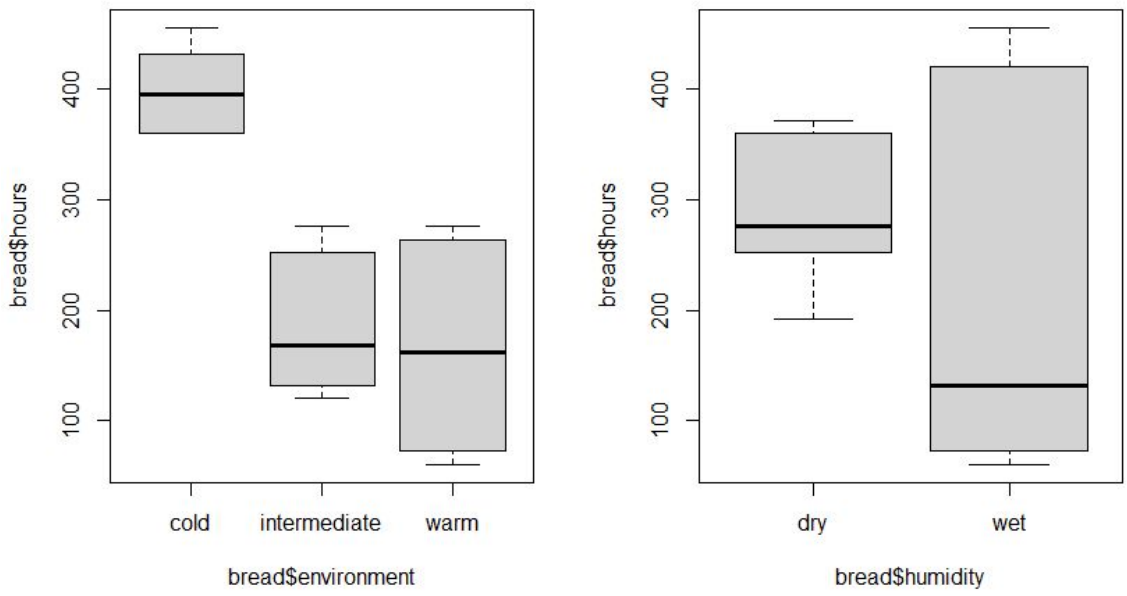


*Fig. 1.1 Box-plot showing the main effects of the factors*

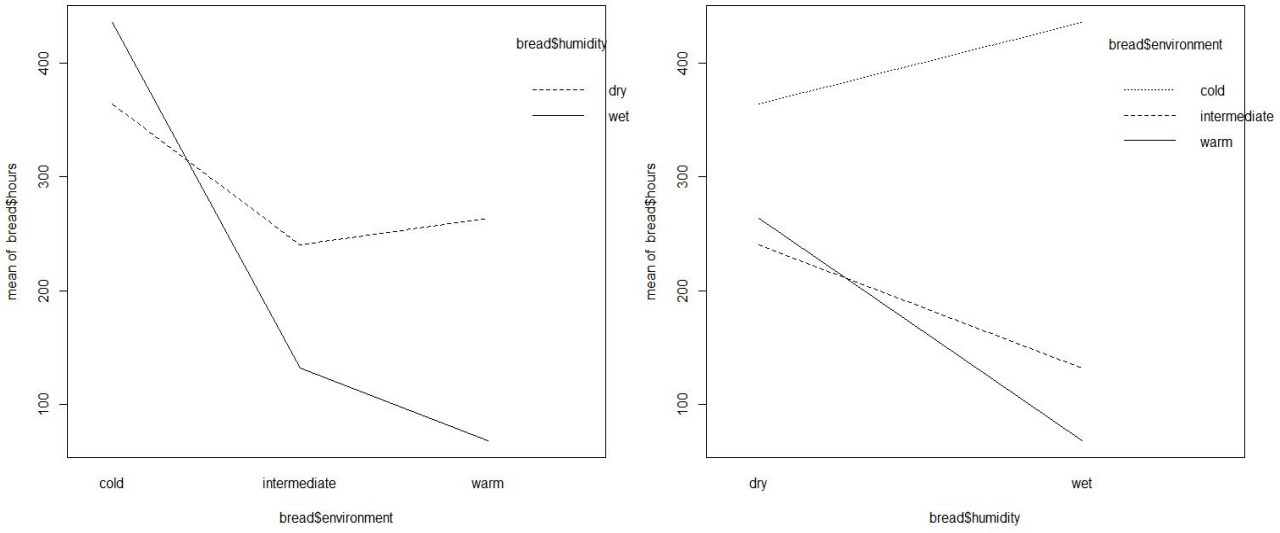


*Fig. 1.2 Box-plot showing the interaction of environment and humidity on hours to decay; Interaction of humidity and environment on hours to decay.*

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

By performing a two-way ANOVA we will test the following three null hypothesis:

- No interaction between environment and humidity on time to decay
- No main effect of environment on time to decay
- No main effect of humidity on time to decay

Code:

```
anova(lm(bread$hours ~ bread$environment * bread$humidity))
```

Output:

Analysis of Variance Table

Response: bread$hours

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| bread$humidity | 1 | 26912 | 26912 | 62.296 | 4.316e-06 *** |
| bread$environment | 2 | 201904 | 100952 | 233.685 | 2.461e-10 *** |
| bread$humidity:bread$environment | 2 | 55984 | 27992 | 64.796 | 3.705e-07 *** |
| Residuals | 12 | 5184 | 432 | | |

The two-way ANOVA test shows that both factor A (humidity) and factor B (temperature) have significant effect on the dependent variable 'time to decay', with p-value < 0.05. Hence we reject H0 and conclude both humidity and temperature have a main effect on the time to decay.
The results also show a significant interaction effect of humidity and temperature on time to decay with p<0.05. This means that the main effect of factor A is bigger depending on the condition of factor B (and vice versa). So the effect of humidity on the time of the bread to decay depends on the temperature. The other way around, the effect of temperature on the time of the bread to decay depends on whether it is humidified or not.

*d) Which of the two factors has the greatest influence on the decay? Is this a good question?*
To see which of the two factors has the greatest influence we performed an two separate ANOVA-test and com the r-squared (i.e. explained variance).

Code:

```
breadmodel1 <- lm(bread$hours ~ bread$environment, data = bread)
breadmodel2 <- lm(bread$hours ~ bread$humidity, data = bread)
summary(breadmodel1); summary (breadmodel2)
```

The results of both tests show that the model with only factor 'environment' explaining the hours to decay has r-squared of : 0.70. The model with 'humidity' as factor explaining the hours to decay has r-squared of : 0.09. We conclude that the factor 'environment' has the greatest influence on decay.
However, we think this is not a good question, because as we saw earlier that the interaction between the two factors are significant. This means the main effects of the two factors depend on each other. So when we m model without this interaction effect, the testing of the main effects of both factors are less meaningful.

If the interaction effect was not significant, we would have considered testing the main effects individually.

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

To check for the assumption of normality and the assumption of equal variances we made a QQplot and also plotted the residuals.

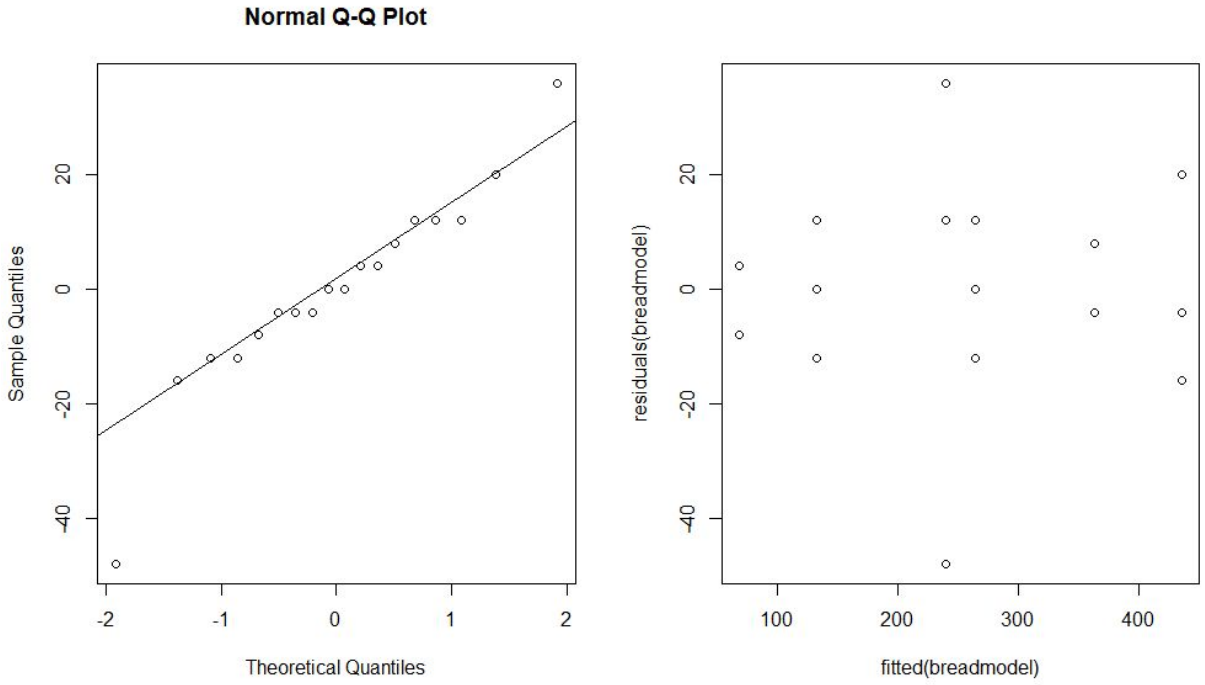


*Fig. 1.3 QQplot to check normality; Plotted spread of residuals to check for equal variances.*

Looking at the QQplot (left graph) we can see that the data points near the tails don't fall exactly along the straight line, but for the most part this sample data appears to be normally distributed. Looking at the residuals plot (right graph) the spread in residuals looks a bit too symmetric. Therefore
we reject the assumption of 'equal variances'. However, as we see some extreme values in the graph we suppose this symmetric pattern can be due to outliers in the data. In order to check for the assumption of equal variances again, these outliers could be excluded.

## Exercise 2. Search engine

*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.*

In this research there are two factors: 'interface' and 'skill'. The factor of interest is 'interface' and consists of 3 levels. The other factor 'skill' is the block factor, which has 5 levels. To see the effect of interface on the dependent variable 'search time', a randomized block design is used.

This means that every treatment (i.e. level of 'interface') is random and equally divided into five blocks of level of 'skill'. This results in the following randomized block design:

Code:
```
search <- tibble::rowid_to_column(search, "student_number")
I = 3 ; B = 5 ; N= 1
for (i in 1:B) print (sample(1:(N*I)))
```

Output:
```
[1] 3 1 2
[1] 1 3 2
[1] 2 3 1
[1] 3 2 1
[1] 3 1 2
```

The numbers are the different interfaces (1,2 or 3). The blocks represent the level of skill (1,2,3,4 or 5). So this block design can be interpreted as block 1 student 3 is assigned to treatment 1, student 1 to treatment 2 and student 2 to treatment 3. For block 2 student 1 is assigned to treatment 1, student 3 to treatment 2 and student 2 to treatment 3 etc.

*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.*

**Main effect**
To test the effect of the factor 'interface' on the search time we use an ANOVA which will test the following hypothesis:
$H_0 : a_1 = a_2 = a_3$

Code:
```
main = lm(time ~ interface+skill, data = search);
anova(main)
summary(main)
```

Output:
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

The ANOVA test shows that there is a significant effect of ‘interface’ on ‘search time’, with $p < 0.05$. hence we reject $H_0$. This means that the search time is not the same for all the interfaces.

**Longest search time**
Furthermore with the summary command we observe that ‘interface level 3’ requires the longest search time, namely a search time of: 19,47.

**Interaction effect**
To test whether there is an interaction effect between skill level and type of interface, we used an ANOVA-test with the following code:

Code:
```
anova(lm(time~interface * skill))
```

We found no interaction effects between skills and interface on search time, meaning that the effect interface has on search time doesn’t depend on skills. To further investigate the interaction effects we also plotted the interaction between type of interface and level of skill on the search time (left graph). As you can see the lines are kina parallel, also suggesting that there is no interaction effect.

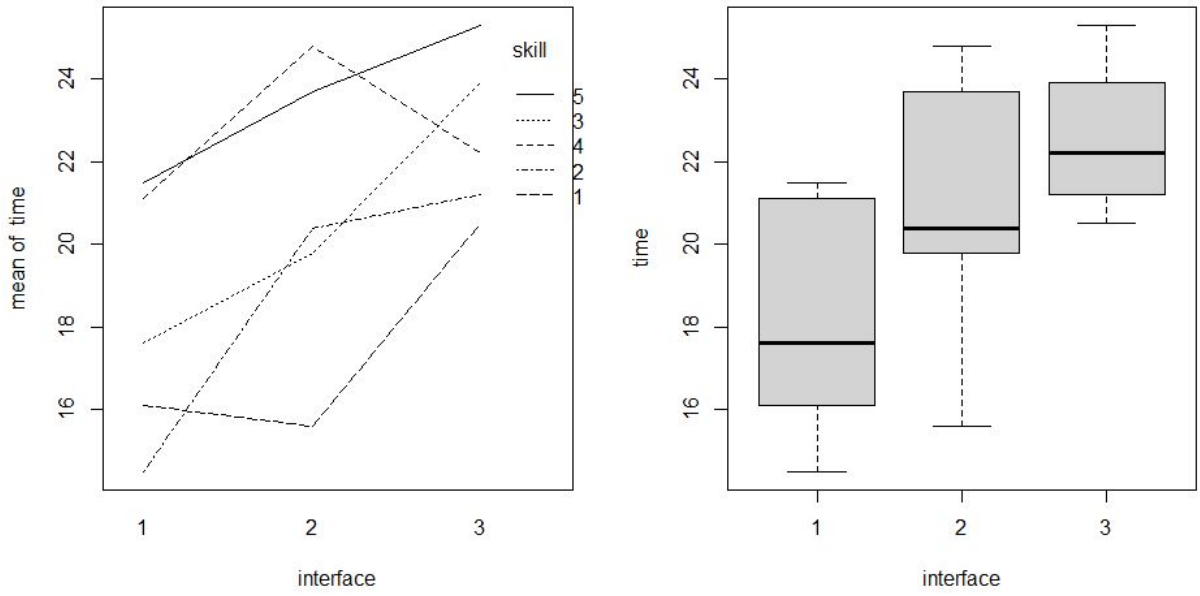


*Figure 2.1: interaction plot of interface and skil ; Figure 2.2 Boxplot of interface and search time.*

**Skill level 3 & interface 3**
After we performed an ANOVA-test we investigated with the command *summary* that the estimated time it takes a typical user of skill level 3 to find the product on the website if the website uses interface 3 is 18**.**

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

By computing the estimators of the residuals we can check for the assumption of ' normality of the population'. The QQplot for the residuals looks as follows (see left graph). We also checked the assumption 'homogeneity of variances' by the residuals versus fits plot (see right graph).

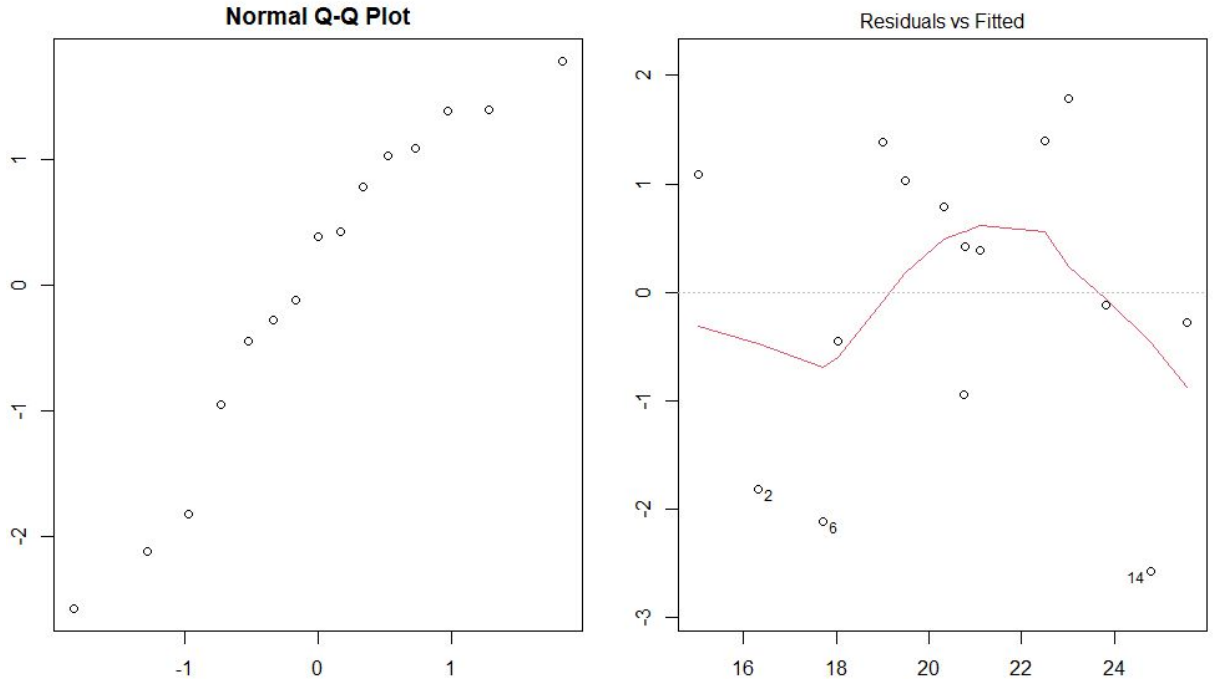


*Figure 2.3. Shows normality QQplot; Figure 2.4. Shows relationship between residuals and the means of the different levels.*

In the left QQplot above, as all the points fall approximately along the reference line, we can assume normality. This conclusion is supported by the Shapiro-Wilk test with a non significant *p = 0.28.*

In the right plot above, there is no evident relationship between residuals and the mean of the different levels of the factor feed, which is good. So, we can also assume the homogeneity of variances.

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

We performed the non-parametric Friendman test to test the null hypothesis of: no treatment effect taking the blocks into account (by using its ranks).

Code:
```
friedman.test(time, interface, skill, data = search)
```

Output:
```
        Friedman rank sum test

data:  time, interface and skill
Friedman chi-squared = 6.4, df = 2, p-value = 0.041
```

We tested the relevance of factor interface taking into account the blocking factor skill. The p-value for testing ($H_0$ : no treatment effect) is 0.041 (*p<0.05*), hence we reject $H_0$ and conclude there is a treatment effect (i.e. of interface) on search time.

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

Output:
```
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
```

We tested the null hypothesis that search time is the same for all interfaces with a one-way ANOVA test, the p-value is p= 0.096 (*p>0.05*), hence we don't reject $H_0$ and conclude that there is no treatment effect.
By performing this one-way ANOVA we excluded the variable level of skill, even though this variable is known to be of influence. We consider this as wrong because when level of skill influences the search time, we should use this factor to create homogeneous groups such that we can detect the effect of interface easier and is not blurred by variation coming from the factor skill.

## Exercise 3. Feedings for cows

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

For studying the feedingstuffs influence of the milk production of various cows, each one tested with both types of feedings with a recovering time in between, we decided to fit the data with lm and passed it through the Anova test.

```
> cowlm=lm(milk~treatment+per+id,data=cow)
> anova(cowlm)
```

The results obtained, with a high p-value of 0.93, we can assume that there is no significant difference between the two types of feedings.

*b) Estimate the difference in milk production.*

Considering the code below, the estimated difference in milk production is such that feeding A results in -0.51 less milk produced by the tested cows.

```
> summary(cowlm)
```

*c) Repeat a) and b) 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 fixed effects model.*

For the mixed effects analysis, we employed the lmer to model-fit the data.

```
> cowlmer=lmer(milk~treatment+order+per+(1|id),REML=FALSE)
> summary(cowlmer)
> cowlmer1=lmer(milk~order+per+(1|id),data=cow,REML=FALSE)
> anova(cowlmer1,cowlmer)
```

We can observe, through the first line, that the estimated difference between the two feeding types is still such that feeding A results in -0.51 less milk produced, remaining consistent with the result obtained above.
For obtaining the p-value, however, not offered by the lmer function, we can refit the model without the treatment type - our focal interest point -, applying the two models to Anova with two arguments to test the fit of the reduced model without the type of feeding inside the full model. The result obtained is 0.446, which shows that there is no significant effect derived from the treatment type.

*d) 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?*

The command shown above represents the paired t-test. This type of test is suitable for this experiment, as the in-between resting period for each cow between the two different types of feeding allows a complete recovery washing out possible residual effects. This frees the two trials from any possible inter-relation and inter-contamination between them.
The result obtained is 0.828, which, even if slightly lower than the one obtained above, shows no significant difference between the two types of feedings.

## Exercise 4. Jane Austen

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

The admirer's purpose is to imitate Austan's style as much as possible, meaning that he wants to use the words 'a', 'an', 'this', 'that', 'with' & 'without' as often as Austen did. Looking at the data this means that the distributions over the *rows* need to be as homogenous as possible. Therefore we consider the most appropriate test to use here is the test for homogeneity (over rows). The null hypothesis in this case would be: the distributions over *row factors* are equal.

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

To investigate whether Austen herself was consistent with the use of specific words, we will look at the square root contributions of each cell to the chi-squared statistic. First we excluded the column 'Sand2' from the matrix, since this isn't relevant for calculating the (in)consistencies between Austen's novels.

Code:

```
df = subset(austen, select = -c(Sand2) )
z= chisq.test(df); z
residuals(z)
```

Output:

```
Pearson's Chi-squared test

data:  df
X-squared = 12.3, df = 10, p-value = 0.27
```

| | Sense | Emma | Sand1 |
|---|---|---|---|
| a | -1.029977 | -0.12902 | 1.59377 |
| an | 0.447288 | -0.15910 | -0.37463 |
| this | 0.051336 | 0.29387 | -0.50366 |
| that | 0.748176 | 0.28658 | -1.44235 |

```
with       -0.047474  0.52051  -0.70352
without  1.065443  -1.58841  0.89262
```

The results from the Chi-squared test show a p-value of 0.27, hence we don't reject the null hypothesis. This means that there are no significant differences between the three novels and we conclude that Austen is consistent in use of words in her different novels.
Furthermore the output shows which values deviate most from the expected under $H_0$. The higher the value (positive or negative) the more the value deviates from $H_0$. The main inconsistencies however are with the word 'a', which is relatively more used in 'Sand1' than 'Sense' and 'Emma'. Another main inconsistency is with the word 'without', which is relatively less used in 'Emma' than in 'Sense' and 'Sand1'.

*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?*

To investigate whether the admirer was successful in imitating Austen's style we will test the null hypothesis $H_0$: the distributions over row factors are equal

Output:

```
Pearson's Chi-squared test

data:  austen
X-squared = 45.6, df = 15, p-value = 6.2e-05

          Sense       Emma    Sand1     Sand2
a       -1.01492 -0.11209279  1.60629 -0.058899
an       -0.59063 -1.21995459 -1.06713  3.728164
this      0.13883  0.39049032 -0.44364 -0.326717
that      1.59436  1.17984884 -0.90996 -3.049316
with     -0.51209  0.00019167 -1.02461  1.748217
without  1.39193 -1.34119628  1.13654 -1.069630
```

The results of the chi-squared test show a p-value of $p<0.05$, hence we reject our H0. This means that there are significant differences between the row factors. If we investigate the residuals we observe that there are high values (both positive and negative) in column 'Sand2', meaning that the counts of the rows in this column are not in proportion to the rows of the other columns (i.e. the other novels). Therefore we conclude the admirer as not successful in imitating Austen's style. The main differences are with the word 'an', which is relatively more used than in the other novels. Also the word 'that', which is relatively less used compared with the other novels.

## Exercise 5. Expenditure on criminal activities

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

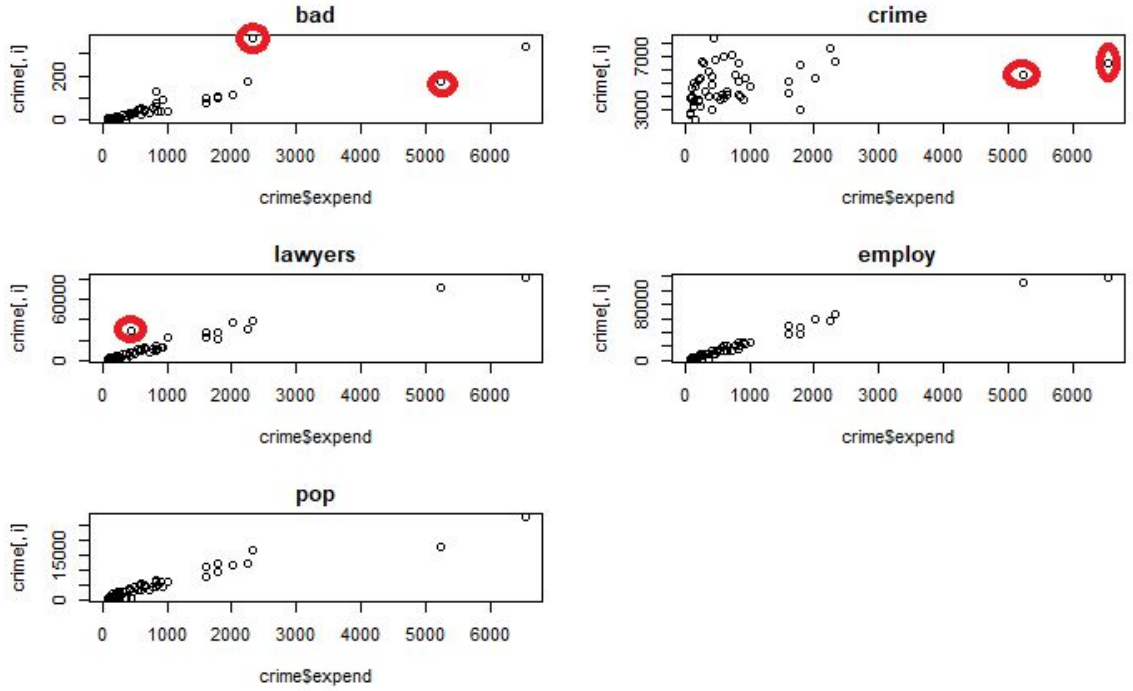


*Fig 5.1 Plots of predictor variables against dependent variables*

Figure 5.1 shows outlier values highlighted in red circles. A deeper investigation could be done into these outliers, however, even through investigation using graphs, we can see that there do exist outliers. Outlier values can tremendously skew the output of a linear regression model

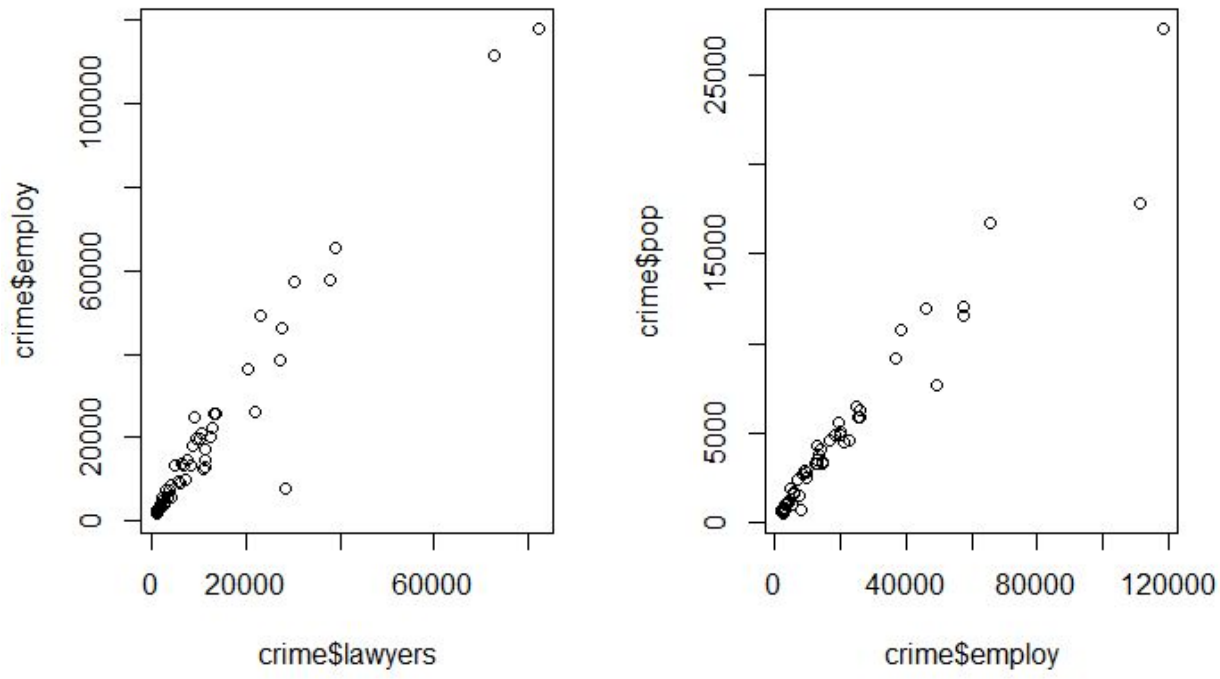


*Fig 5.2 Plot of dependent variables against each other*

To check collinearity we plot 2 independent variables against each other to see if any linearity exists between the 2 plotted variables. If a linear relationship does exist, it means that either variable can be used to predict the other independent variable, and hence they have a dependency. We can then eliminate one of these variable as we dont introduce any loss of information, and at the same time reduce a possible source of noise in the data

*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.*

The step up procedure results in a model with employ and lawyer as the chosen variables:

```
Call:
lm(formula = expend ~ employ + lawyers, data = crime)

Residuals:
   Min      1Q  Median      3Q     Max
-599.47  -94.43   36.01   91.98  936.55

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.107e+02  4.257e+01  -2.600  0.01236 *
employ       2.971e-02  5.114e-03   5.810 4.89e-07 ***
```

```
lawyers      2.686e-02  7.757e-03   3.463  0.00113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 232.6 on 48 degrees of freedom
Multiple R-squared:  0.9632,  Adjusted R-squared:  0.9616
F-statistic: 627.7 on 2 and 48 DF,  p-value: < 2.2e-16
```

The step-down procedure resulted in a model with bad, lawyers, employ and pop as the chosen variables:

```
Call:
lm(formula = expend ~ bad + lawyers + employ + pop, data = crime)

Residuals:
   Min      1Q  Median      3Q     Max
-635.62  -80.18   18.77  114.54  809.66

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.464e+02  4.541e+01  -3.224  0.00232 **
bad         -2.241e+00  1.133e+00  -1.977  0.05402 .
lawyers      2.646e-02  7.571e-03   3.495  0.00106 **
employ       2.283e-02  7.487e-03   3.049  0.00380 **
pop          6.368e-02  3.304e-02   1.927  0.06012 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 226.4 on 46 degrees of freedom
Multiple R-squared:  0.9666,  Adjusted R-squared:  0.9637
F-statistic: 332.5 on 4 and 46 DF,  p-value: < 2.2e-16
```

Both models have very similar R2 values, however, the step up model has half the variables as the step down model, hence we choose the step up model. The step down model has a slightly higher R2 value, and an argument could be made for choosing it as well.

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

To check model assumptions, we use graphical methods to diagnose our model. The residuals (errors) of the model should be normally distributed, as it is assumed that they belong to a normal distribution.

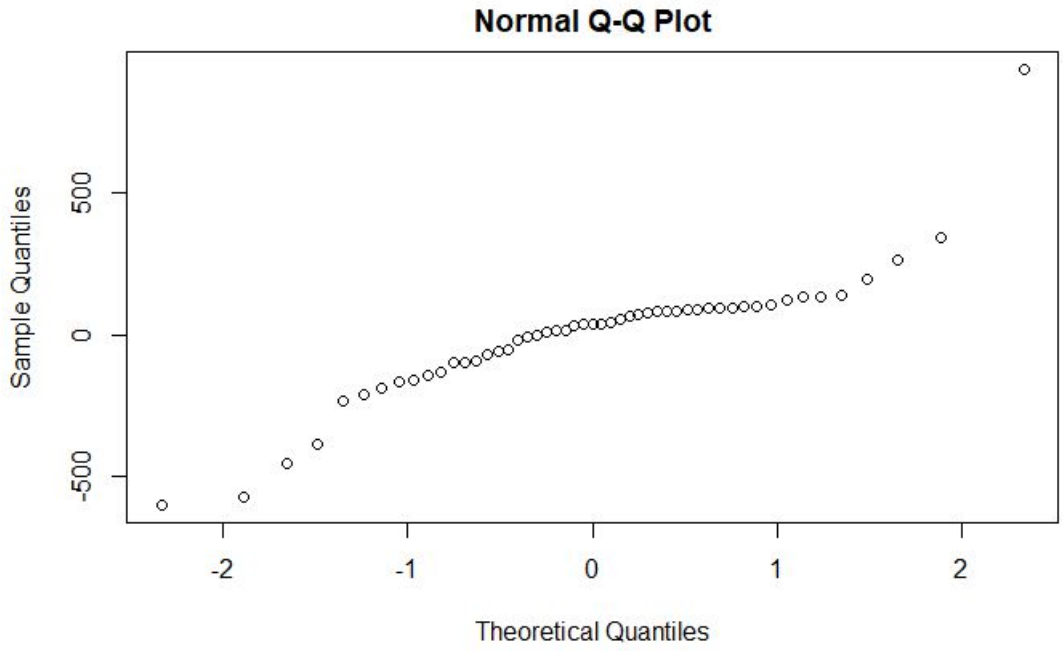


*Fig. 5.3. QQ-plot of the residuals of the step up model.*

The QQ-plot reveals that the errors are not normally distributed.

The errors should also be randomly distributed when plotted against either the fitted values by the model, or any of the dependent variables.

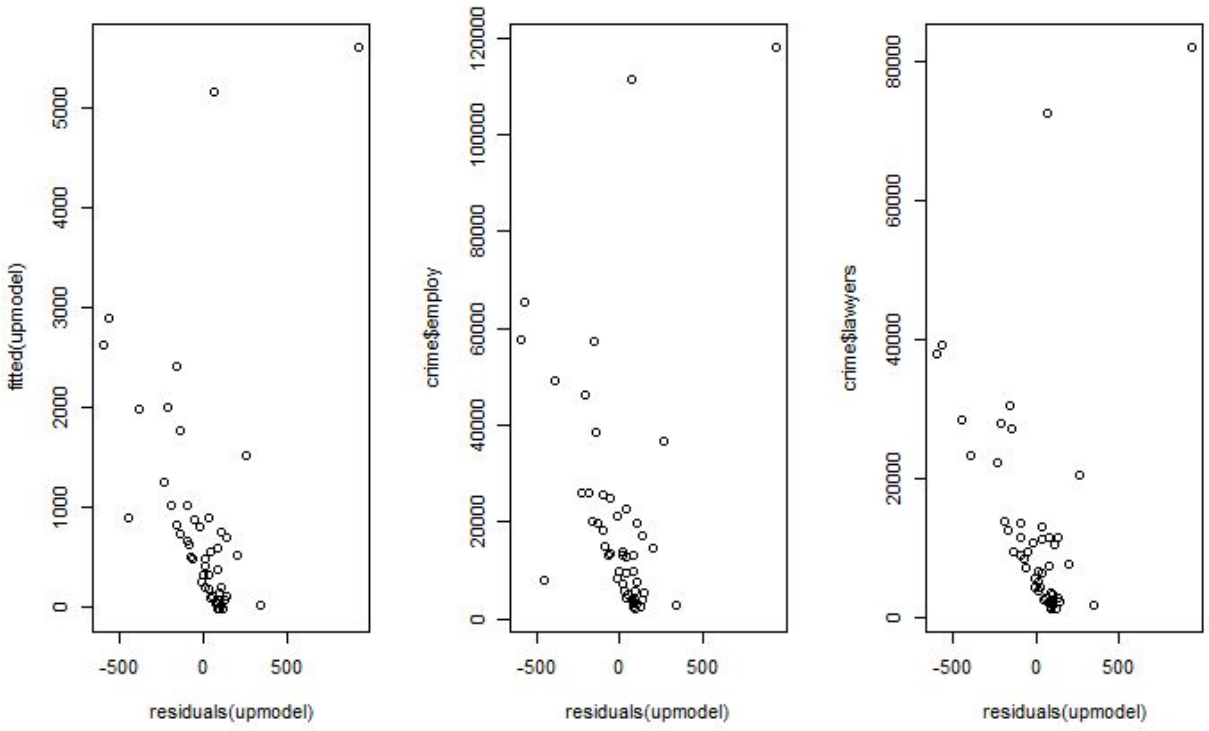


*Fig 5.4: Residuals of the step up model, plotted against fitted and independent variables.*

However fig 5.4 shows that there is a pattern within these plots, and they are far from random. This indicates that the model still has not optimally detected the latent trends within the data.

Even though the R2 value was high, the graphs reveal that the model is not ideal.