lab 1

Eric Börjesson, Ludvig Ekman, Wenli Zhang, Tim Grube

17/11/2020

Time

Question a.

Get the descriptive statistics for this data (mean, median, standard deviation, variance). Suggested functions: mean, median, sd, var.

First, we created a vector for the time data from the assignment.

```
d <- c(229, 186, 396,233,238,158,259,317,222,375,156,108,197,227,379,234)
df<-data.frame(Time = d)
kable(df)
```

Time
229
186
396
233
238
158
259
317
222
375
156
108
197
227
379
234

Then we used four R methods for calculating the mean, the median, the standard deviation and the variance of the data set:

```
mean(d)
```

```
## [1] 244.625
```

```
median(d)
```

```
## [1] 231
```

```
sd(d)
```

```
## [1] 83.46726
```

```
var(d)
```

```
## [1] 6966.783
```

Question b.

On the previous question, what is being calculated? The population standard deviation or the sample standard deviation? Why?

The sample standard deviation is being calculated, since the company only provides 16 features that were chosen randomly. The data is a sample and therefore only a subset of the population (all features provided by the company).

Question c.

Set up appropriate hypothesis for investigating the issue. Think about carefully if it is one tailed or two tailed hypothesis.

As we want to find out whether the mean time to develop a feature is higher than 225 h, we have a null hypothesis which states that it is actually below or equal 225h. This enables us to either reject h0 which would be in favor of hA and therefore be an indication that the mean time is higher than 225h. If H0 cannot be rejected, there is not enough evidence that the mean time is actually higher than 225h. Consequently, these are the hypothesis:

- h0: mu(d) <= 225
- hA: mu(d) > 225

These are one tailed hypothesis since we are only interested in testing one end of the distribution, so only one direction.

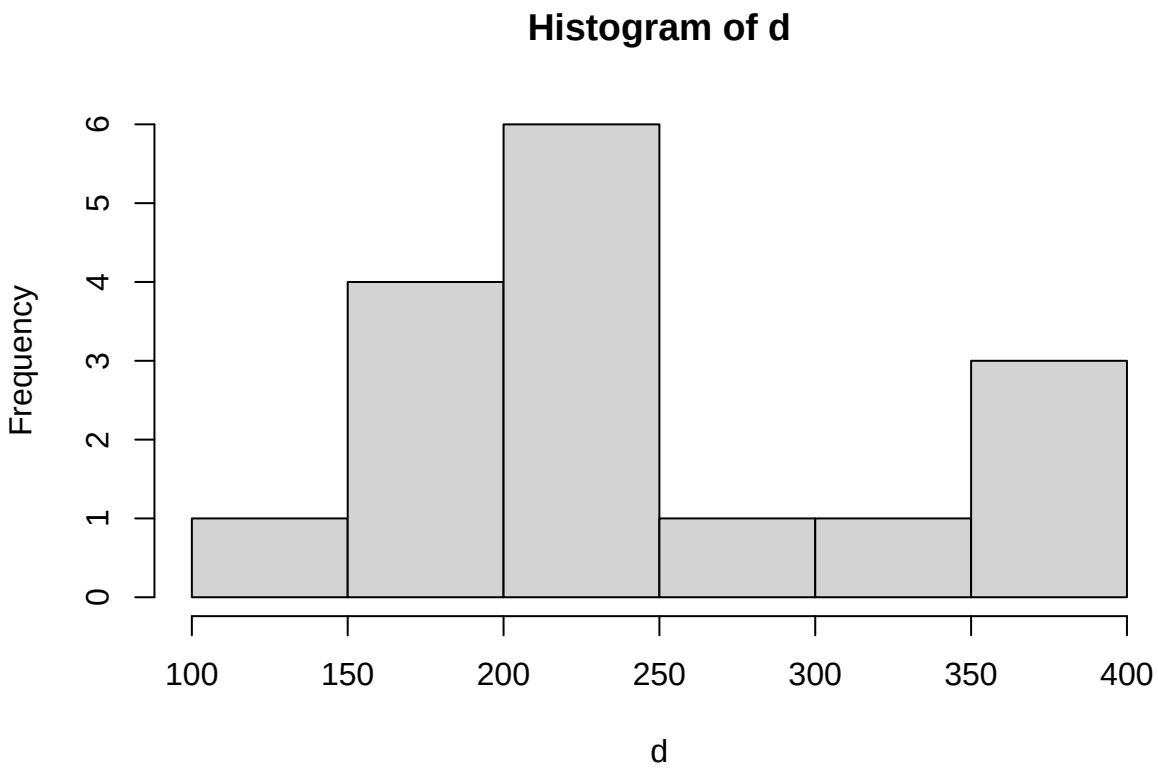
Question d.

Plot a histogram of the data with a bin size of 50, 75 and 100. Does the bin size change your perception of normality of the data?

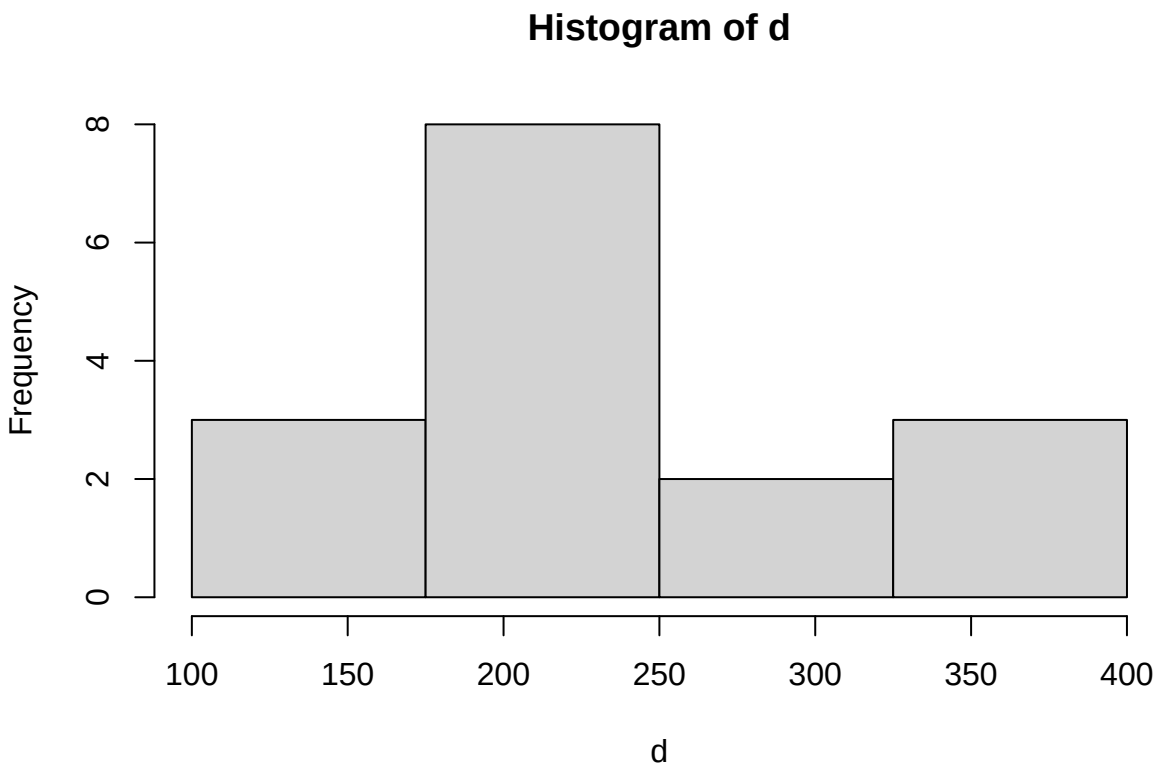
All histograms are plotted with the R function hist. For configuring the bin sizes, we configured the amount of breaks. For example, the first histogram has bin sizes of 50. All times are between 100 and 400 and so we wanted to have 6 bins with a size of 50. *length.out* has to be equal to *"number of bins" + 1*. The choice where the bins start can have an influence on how the histogram looks like and also whether it looks normally distributed. An alternative would be to do not at start at 100 but at the minimum value of the data set.

With a bin size of 100, the data looks normally distributed. When reducing the size to 75, there is a first indication that it could be not normally distributed, as the number of values between 250 and 325 is quite low in comparison to 325 and 400. With a bin size of 50 it is even more obvious that it is probably not normally distributed.

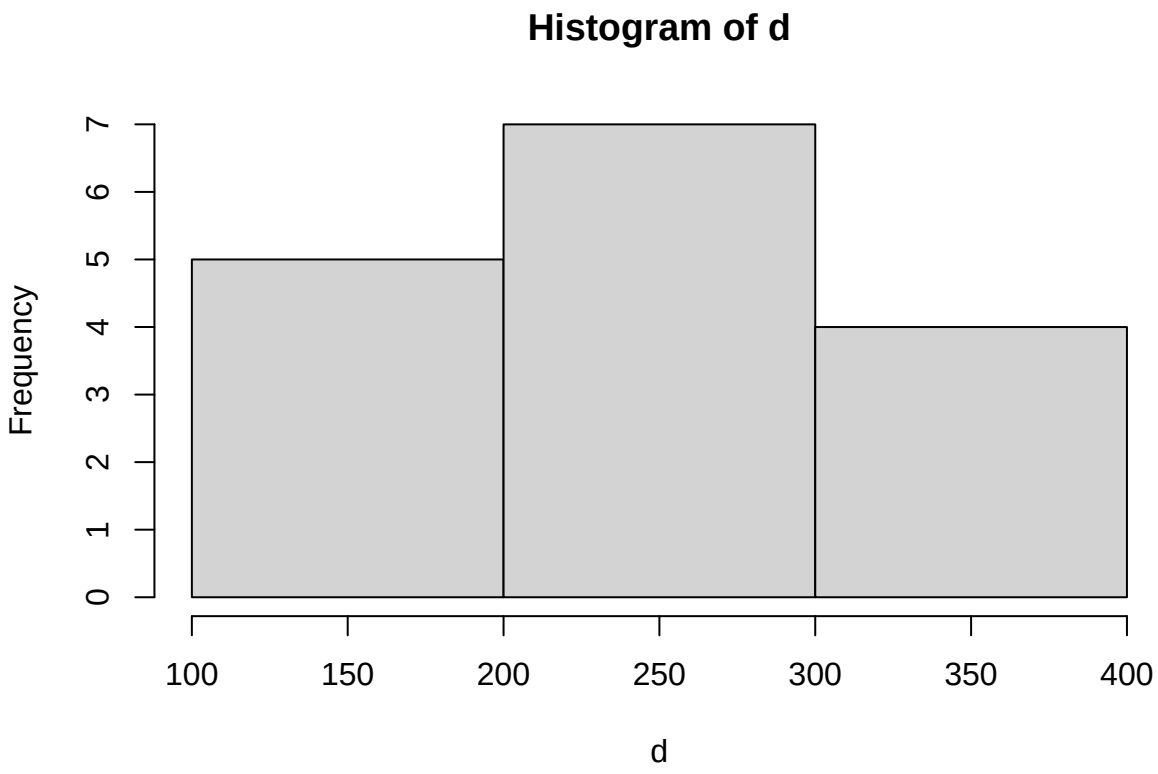
```
hist(d, breaks = seq(100, 400, length.out = 7))
```




```
hist(d, breaks = seq(100, 400, length.out = 5))
```




```
hist(d, breaks = seq(100, 400, length.out = 4))
```



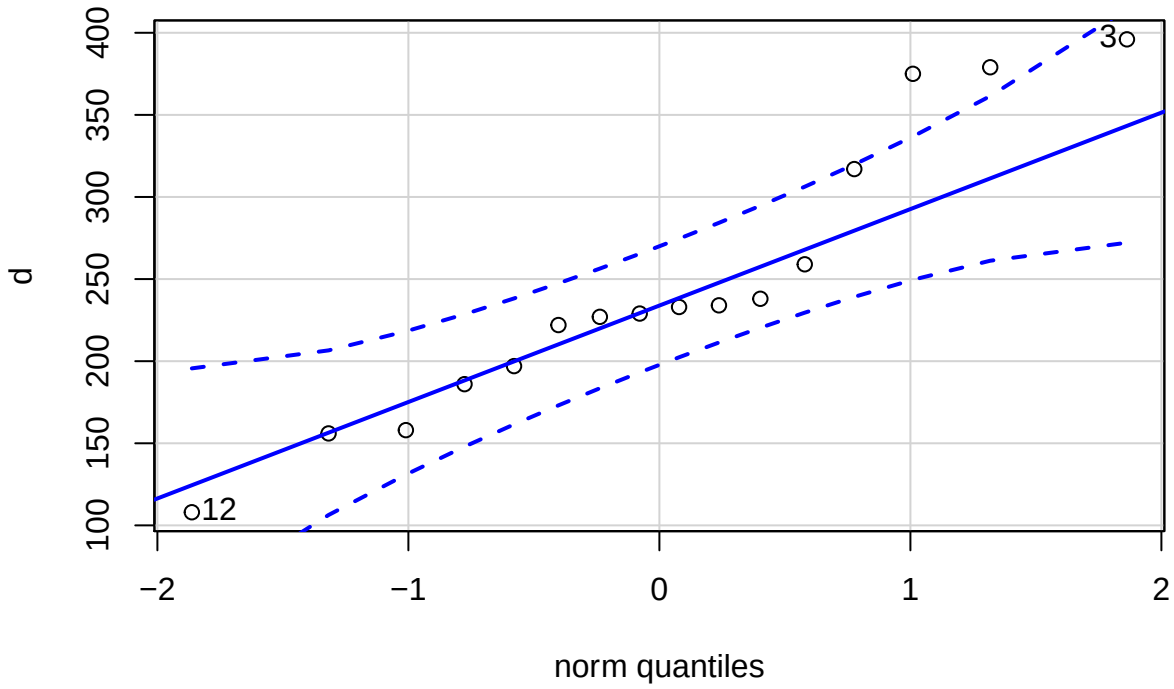
Question e.

Plot the qq-plot for this data. We suggest using the function qqPlot from the car package. This function has the advantage of providing 95% confidence intervals in the qq-plot. However, you can also plot it using ggplot.

The following diagram is a qq-plot of the data set. If all data points would be on the straight blue line, the data would be perfectly normally distributed. The lower values are quite good on that line and easily inside the 95% confidence interval which is shown by the blue dashed lines. However, the higher values tend to be much above the straight line and some of them are even out of the 95% confidence interval.

Besides, the return result of qq-plot is 3 and 12, which means the third data (396) is the largest and the twelfth data (108) is the smallest on the data set.

```
qqPlot(d)
```




```
## [1]  3 12
```

Question f.

Test the normality of the data using the Shapiro-Wilk test. Based on the Shapiro test and the histograms and qqplots. Is the data normally distributed? You can use the function shapiro.test

If the data is normally distributed, the W-value from the Shapiro-Wilk test should be 1. Since it is 0.92, there could be an indication that the data is not normally distributed. Furthermore, the p-value is 0.17 which is higher than 0.05, so the null hypothesis (data is normally distributed) cannot be rejected. However, like explained in the video about the Shapiro-Wilk test, the test has not much power to reject a null hypothesis which is true is in this case (n=16). These are the calculated values from the test:

```
shapiro.test(d)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d
## W = 0.92033, p-value = 0.1708
```

Combining the results from the histograms, the qq-plot and the Shapiro-Wilk test, there are signs that the data in general is normally distributed. However, as seen in the qq-plot and the histogram using a bin size of 50, there are some exceptions for the higher values of the data set.

Question g.

Investigate if the hypothesis is true using a one-sample t-test with a = 5%. Report and discuss the results appropriately. Why using a one-sample t-test and not a z-test? Why a one-sample test and not a two-sample test? Does it follow the assumptions of a one-sample t-test? Don't forget to report the confidence interval (95%) for the mean value.

- *Use the function t.test with the options mu and the conf.level*
- *Note that confidence level is 1 - a and it is already given by the t.test function*

Since the p-value (0.18) is higher than 0.05, the null hypothesis will not be rejected, which means that the mean time to develop a feature is lower or equal to 225. Moreover, the t-value at the 95% border is 1.753. This can be found in a t-table by looking for 15 degrees of freedom (df) and 0.95 (one-tail). The calculated t-value is a lot smaller than this.

These are all results from the execution of the R function t.test:

```
t.test(d, conf.level = 0.95, mu = 225, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  d
## t = 0.94049, df = 15, p-value = 0.1809
## alternative hypothesis: true mean is greater than 225
## 95 percent confidence interval:
##  208.0444      Inf
## sample estimates:
## mean of x
##   244.625
```

The t-test is chosen because of two reasons: The variance of the population is unknown and the sample size is too small (i.e. not greater than 30). This means, that a z-test cannot be used in this case.

There is only one sample and the one-sample t-test compares the mean of one sample with the known mean 225. In contrast, two-sample t-tests compare whether there are statistical differences in the mean of two samples to determine whether they are from the same population.

Since the data is continuous and most of the data is normally distributed, it mostly follows the assumptions of the one-sample t-test.

As it is a one-tailed test and we are only focusing on one direction (mean time is higher than 225), the confidence interval (95%) is 208.04-Inf, so only a lower border is specified.

Performance

Question a.

Get the descriptive statistics by group using the long format data.

```
performance <- read.csv("~/Downloads/performance.csv")
df2 <- tidyr::pivot_longer(performance, cols=everything(), names_to="Group", values_to="Time")

psych::describeBy(df2, group = df2$Group)
```

```
##
##  Descriptive statistics by group
## group: timeOptimized
```

```
##        vars  n mean   sd median trimmed  mad   min   max range  skew kurtosis
## Group*    1 10    1 0.00   1.00    1.00 0.00  1.00  1.00  0.00   NaN      NaN
## Time      2 10   16 0.03  16.01   16.01 0.02 15.96 16.04  0.08 -0.43    -1.19
##          se
## Group* 0.00
## Time   0.01
## ------------------------------------------------------------
## group: timeOriginal
##        vars  n  mean   sd median trimmed  mad   min   max range  skew kurtosis
## Group*    1 10  1.00 0.00   1.00    1.00 0.00  1.00  1.00  0.00   NaN      NaN
## Time      2 10 16.02 0.03  16.02   16.02 0.04 15.96 16.05  0.09 -0.43    -1.28
##          se
## Group* 0.00
## Time   0.01
```

Question b.

The function str() allows you to inspect what type of data composes your data frame. Double check the type of data you have and convert the time to numeric and the group to categorical. Explain what numeric, categorical (factor) and ordinal data are.

Numeric data is measured and expressed as numbers and consists of either discrete or continuous data. Categorical or factor data is composed of a limited number of categories, which in this case are the categories **timeOptimized** and **timeOriginal**. Ordinal data is similar to categorical, but in comparison there exists an intrinsic ordering of the categories. One difference between numeric data and categorical/ordinal is that the latter two can consist of natural language description while numeric data always consists of numbers.

The code below checks the type of the data before and after conversion.

```
str(df2)
```

```
## tibble [20 x 2] (S3: tbl_df/tbl/data.frame)
##  $ Group: chr [1:20] "timeOriginal" "timeOptimized" "timeOriginal" "timeOptimized" ...
##  $ Time : num [1:20] 16 16 16 16 16 ...
```

```
df2$Group <- as.factor(df2$Group)
df2$Time <- as.numeric(df2$Time)
str(df2)
```

```
## tibble [20 x 2] (S3: tbl_df/tbl/data.frame)
##  $ Group: Factor w/ 2 levels "timeOptimized",..: 2 1 2 1 2 1 2 1 2 1 ...
##  $ Time : num [1:20] 16 16 16 16 16 ...
```

Question c.

Get the parameters of this linear model and complete the equation that you wrote above. What is the value of the intercept a in the data? What is the value of b? What is the value that X assumes for selecting the optimized group? What value does it assume for the original (non optimized) group?

The linear model below uses the group as a predictor to estimate the outcome of time. Running a summary on the model gives the following values:

- Intercept a: 16.005
- Coefficient for b: 0.01
- X assumes *0* for the optimized group and *1* for the original group.

```r
model <- lm(Time~Group, data=df2)
summary(model)
```

```
##
## Call:
## lm(formula = Time ~ Group, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.0550 -0.0175  0.0050  0.0175  0.0350
##
## Coefficients:
##                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)       16.005000   0.008851 1808.350   <2e-16 ***
## GrouptimeOriginal  0.010000   0.012517    0.799    0.435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02799 on 18 degrees of freedom
## Multiple R-squared:  0.03425,    Adjusted R-squared:  -0.01941
## F-statistic: 0.6383 on 1 and 18 DF,  p-value: 0.4347
```

Question d.

The linear model provided by R also gives a p-value for the parameters. Is the factor 'Group' statistically significant for this model? How do you interpret this result? Don't forget to analyze the assumptions.

Statistically significant: The t-test from the linear model above shows that the p-value for the Group-coefficient is 0.435, which is relatively high. We have previously used a significance level of 0.05 which is much lower than the p-value for the coefficient. Therefore, there is not enough evidence in the given sample to reject the null hypothesis.

Further, this indicates that there is no association between changes in the independent variable *Group* and the dependent variable *Time*. To conclude, the factor *Group* is not statistically significant which further indicates that it is not possible to see any effects on the population.

Assumptions: We make two assumptions:

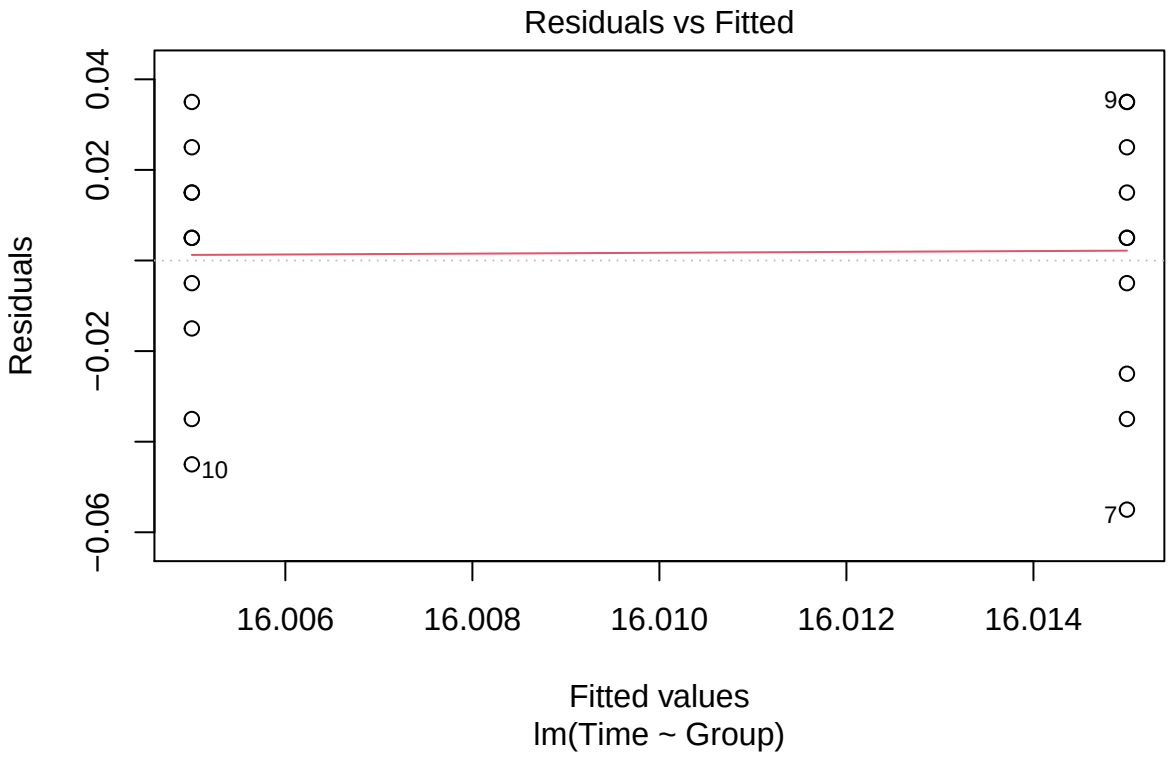
- The data consists of random samples.
- There is a linear regression that includes the following assumptions:
  1. Linearity
  2. Independence of residuals
  3. Homoscedasticity
  4. Normality of residuals

The first assumption is that the data consists of random samples. According to the provided description, each data point has been randomly selected.

Below diagnostics are presented to investigate if any of the four linear model assumptions have been violated.

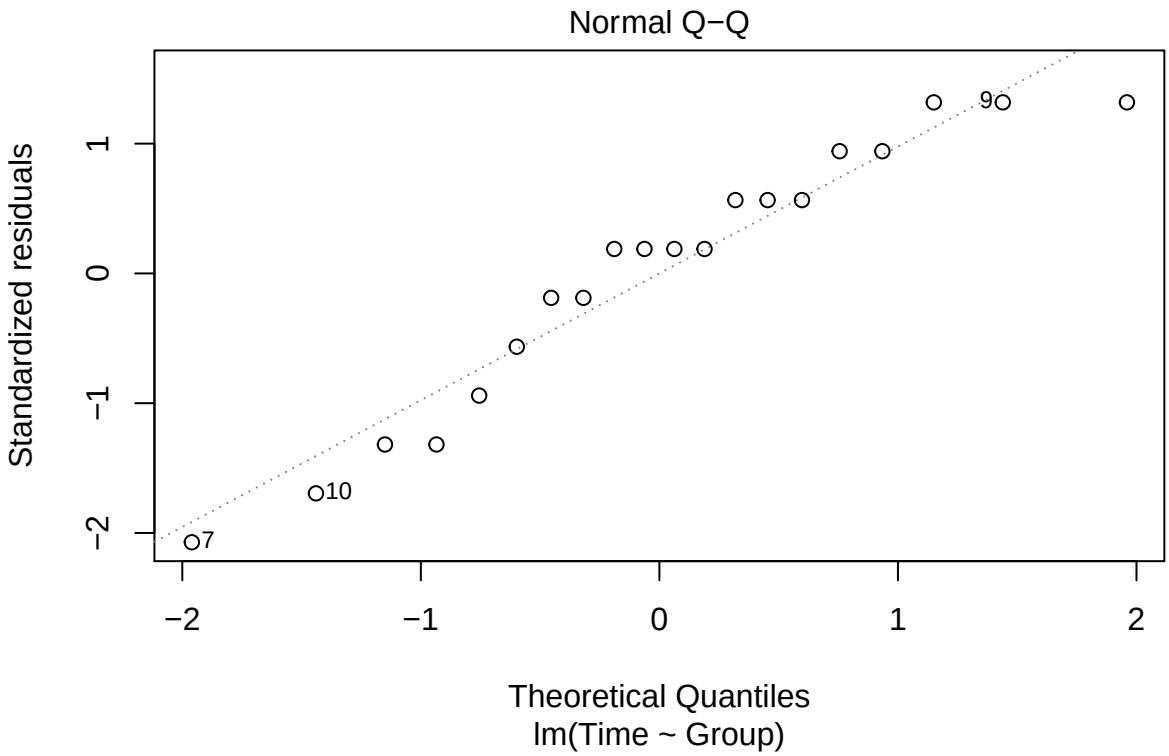
The figure below is used to evaluate the linearity and homoscedasticity of the model. It plots the residuals on the y-axis and the predicted values on the x-axis. In order to meet the two mentioned assumptions, there should not exist any distinct patterns, the residuals should be equally spread around 0 and the red line should be horizontal. A visual inspection of the figure below suggests that these criteria are met.

```r
plot(model, which=1)
```



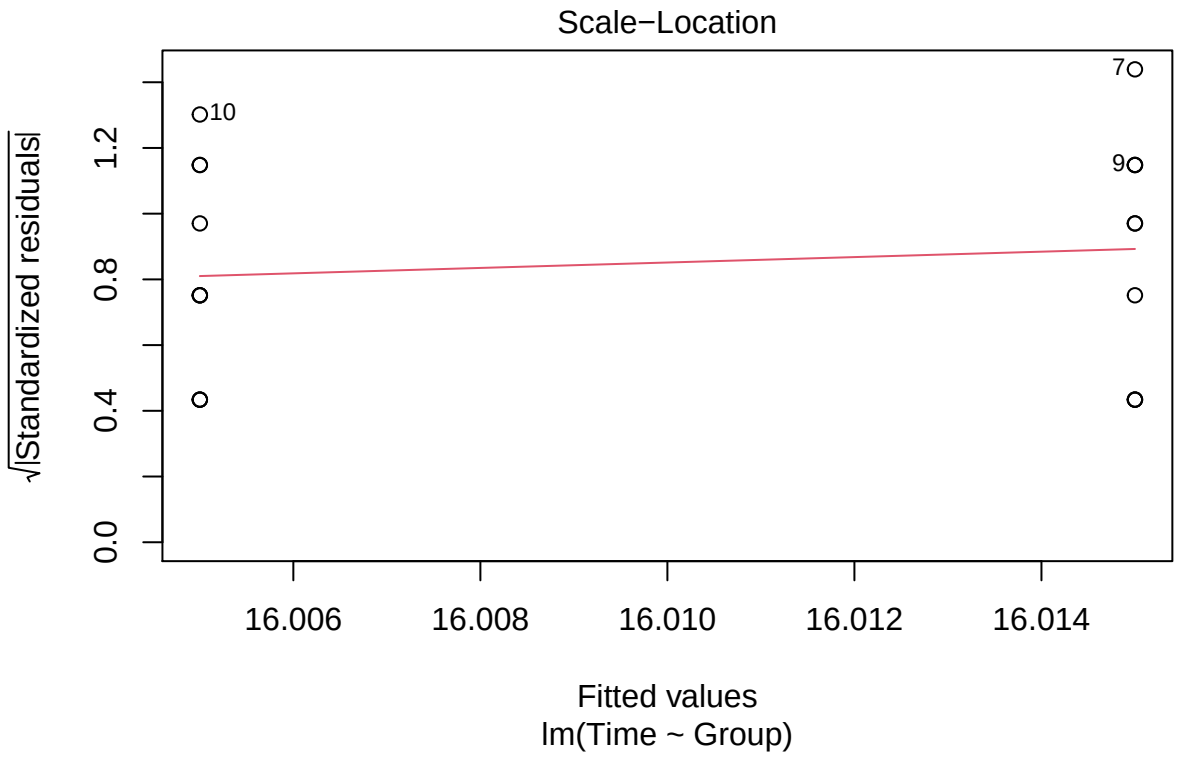
Moving on, the below QQ-plot shows that the residuals are normally distributed since they follow the line. However, there seems to be some deviations in the lower quantiles. Despite this, the assumption of normality of residuals is assumed to be met.

```
plot(model, which=2)
```



The third plot from the diagnostics could be used to evaluate the homoscedasticity and the independence of residuals. A horizontal red line indicates that these assumptions are being met. The plot below indicates that there exists a slight positive effect which could indicate that the residuals are not being randomly spread and that the assumptions are being violated.

```
plot(model, which=3)
```



Contributions

All four of us attended the lab. During the lab we went through all questions, discussed them and wrote down our first thoughts. There were a few aspects that we could not solve immediately and needed to do more detailed research about it, for example task 1g and 2d. Therefore, we decided do split up the group.

Wenli and Tim concentrated on the first task and Ludvig and Eric on the second task. In these smaller groups we had a closer look at the subtasks again, improved the work from the lab and added missing information. After that we collected everything in one document and everyone of the group checked it again, so that everybody knows about all final results.