Experimental Design and Data Analysis

Assignment 1

Tommaso Castiglione Ferrari 2673807

Daniyal Selani 2692551

Simone Korteling 2671463

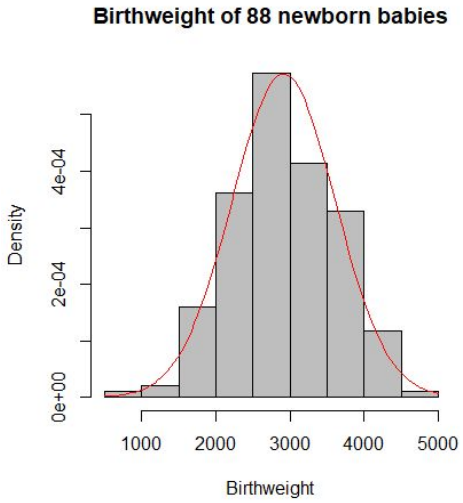
Group 71

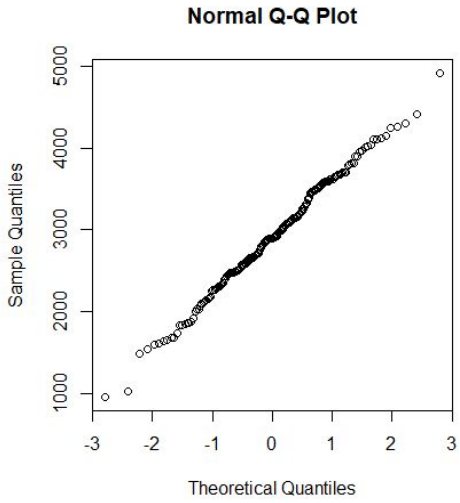
Exercise 1. Birthweight

a) Check normality of the data. Compute a point estimate for μ. Derive, assuming normality (irrespective of your conclusion about normality of the data), a bounded 90% confidence interval for μ.

Normality

First, to check for normality we made an histogram and a QQplot of the data.





Graph 1.1. Histogram of 88 newborn babies; *Graph 1.2 QQplot of 88 newborn babies*

In both the histogram and the QQplot all the points fall approximately along the reference line, therefore we assume normality. Based on the observed data of 188 newborn babies, we computed a point estimate of $\hat{\mu} = 2913{,}293$.

Based only on the observed data, we computed a bounced 90% confidence interval(CI) that contains the true value of the parameter with probability at least $1 - 0.10$. Because the standard deviation is unknown we estimated it by s and the CI is based on a t-distribution.

To compute the boundaries of the CI, we used 1) the estimated μ , 2) the estimated standard deviation and 3) the error. We then computed the lower_bound and upper_bound by subtracting and adding the error from the center of the distribution.

Code:

```
stddev= sd(data, na.rm=TRUE)
n = 188
qt(.95, n-1)

me <- qt(.95, 187)* stddev/sqrt(n-1)

lower_bound = mean(data) - me
upper_bound = mean(data) + me
```

Resulting in a 90% confidence interval of: 2829> $\hat{\mu}$ < 2998

b) An expert claims that the mean birthweight is bigger than 2800, verify this claim by using a t-test. What is the outcome of the test if you take $\alpha = 0.1$? And other values of α?

To test whether the mean birthweight is bigger than 2800 we are testing the following hypothesis: H_0: $\hat{\mu}$ equal/smaller than 2800 vs. H_1: $\hat{\mu}$ bigger than 2800.

Output:

```
One Sample t-test

data:  data
t = 2.2271, df = 187, p-value = 0.01357
alternative hypothesis: true mean is greater than 2800
90 percent confidence interval:
 2847.868      Inf
sample estimates:
mean of x
 2913.293
```

Looking at the results of our t-test with alpha level of $\alpha = 0.1$, our p-value= 0.014 is significant. So we reject our H_0 hypothese, and conclude that the mean birthweights of the population is greater than 2800. Performing a t-test with both an alpha level of $\alpha = 0.05$ and $\alpha = 0.01$ results again in p-value = 0.014. So we reject our H_0 hypothesis, and conclude again that the mean μ of birthweights is greater than 2800.

c) In the R-output of the test from b), also a confidence interval is given, but why is it different from the confidence interval found in a) and why is it one-sided?

The confidence interval is different because in question b the estimated mean population was set to 2800, where in question a we used a mean of μ = 2913. The confidence interval from question b is one-sided because we tested whether the population mean was *greater than* 2800. In this case the alternate hypothesis gives the alternate in only one direction (greater than) of the value of the mean specified in the

null hypothesis. The critical region is only right sided (i.e. a right tailed test) and therefore the confidence interval only contains a *lower* boundary.

Exercise 2. Power function of the t-test

Using the follow code for computing the power of the t-test, we can get the results shown below:

```
get_power=function(n,m,mu,sd,B, sequence){
  power = numeric(length(sequence))

  for (index in 1:length(sequence)){
    nu_val <- sequence[index]
    p <- numeric(B)

    for (b in 1:B) {
      x<-rnorm(n,mu,sd);
      y<-rnorm(m,nu_val,sd)
      p[b] <- t.test(x,y,var.equal=TRUE)[[3]]
    }

    power[index] <- mean(p.value(n, m, mu, nu_val, sd, B)<0.05)
  }
  return(power)
}
```

a) Set n=m=30, mu=180 and sd=5. Calculate now the power of the t-test for every value of nu in the grid seq(175,185,by=0.25). Plot the power as a function of nu.

Graph 2.1

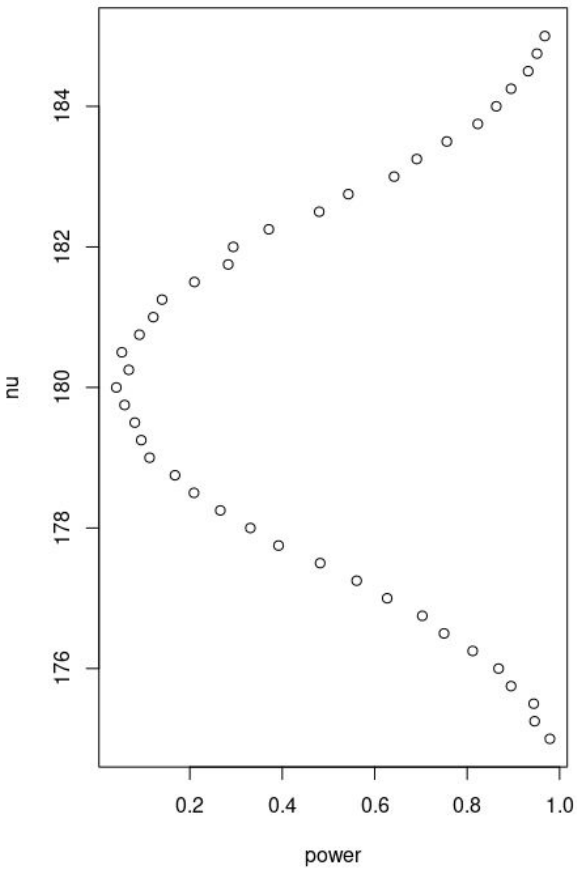
b) Set n=m=100, mu=180 and sd=5. Repeat the preceding exercise. Add the plot to the preceding plot.

Graph 2.2

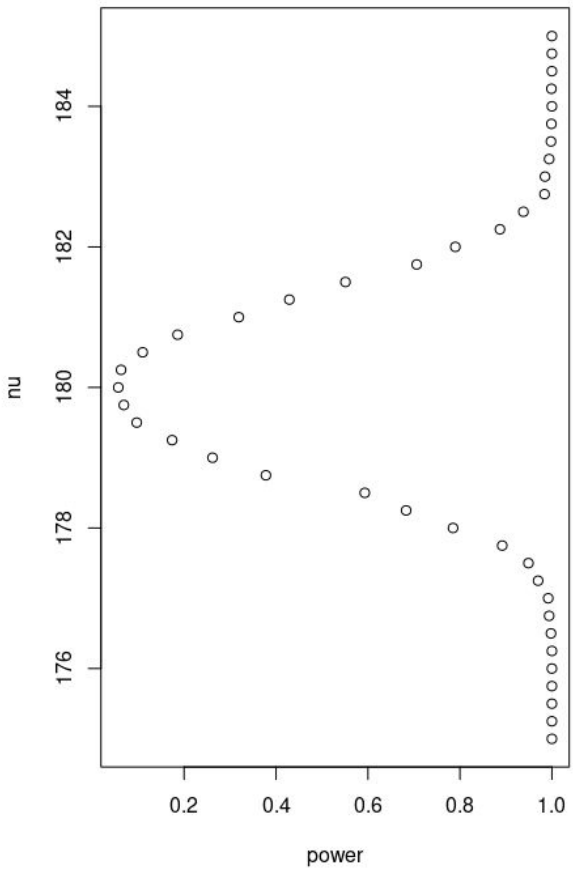
c) Set n=m=30, mu=180 and sd=15. Repeat the preceding exercise.

Graph 2.3

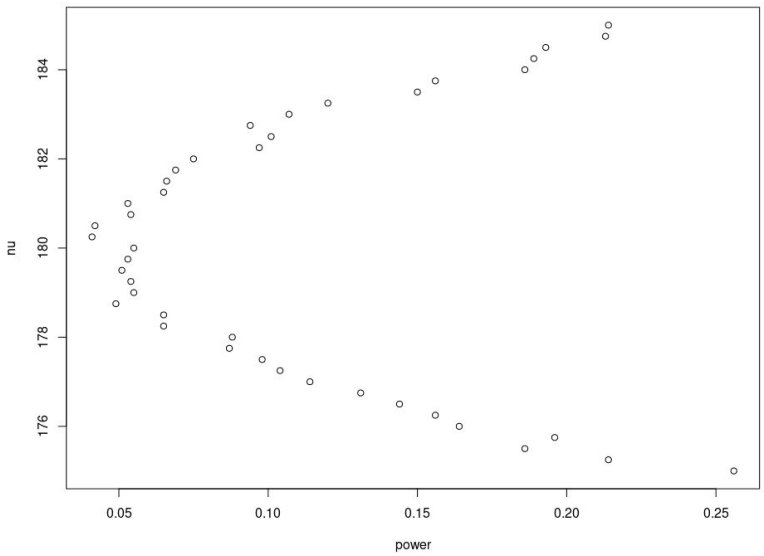
Graph 2.1



Graph 2.2



Graph 2.3



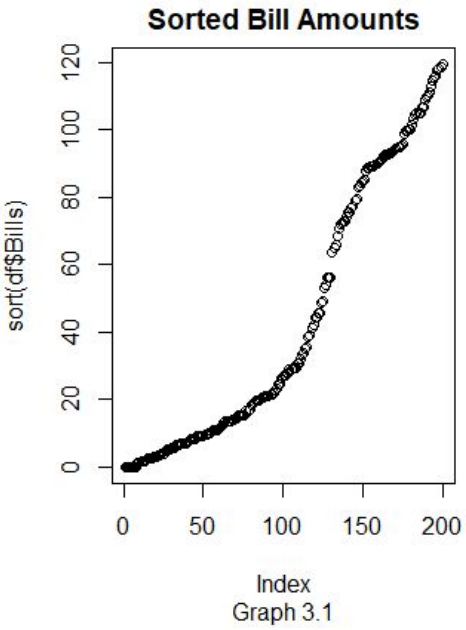
d) Explain your findings.

In the results presented above we can observe two distinct trends in action. The first one correlates the number of samples considered with a more tight and centralized curve - e.g. graph 2.2 -. The second observation considers the different standard deviation considered. In fact, as shown by the graphs, to a smaller standard deviation corresponds a more precise and well defined approximation curve, while when the standard deviation is bigger - e.g. in graph 2.3 - the points are more scattered across when forming the curve.

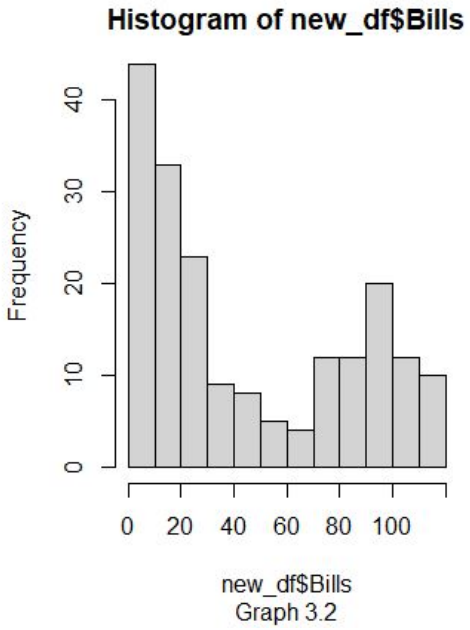
Exercise 3. Telephone Bills

a) Make an appropriate plot of this data set. What marketing advice(s) would you give to the marketing manager? Are there any inconsistencies in the data? If so, try to fix these.

Graph 3.1 shows that there are a few data points in the data that show the monthly fee as 0. As this data represents the first month bills of the new subscribers, we can assume that all of the new subscribers have used their subscription and their bill should not be 0, hence these data points can be removed.



Graph 3.1



Graph 3.2

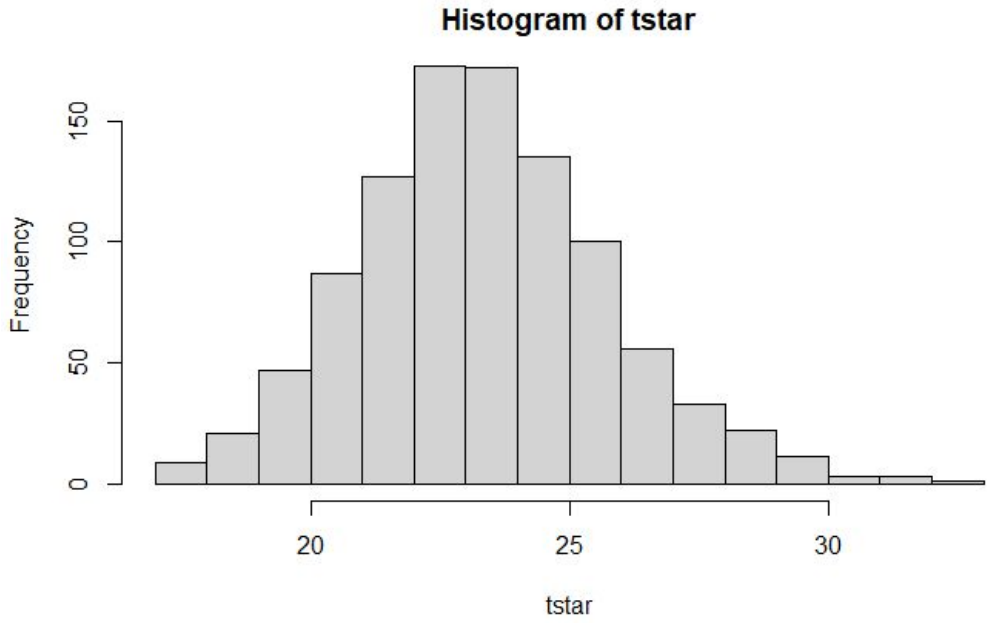
Graph 3.2 shows that most customers rarely use their telephone connections, and some customers use their telephones a lot. However, there are even fewer customers that are somewhere in the middle of the 2 extremes.

b) By using a bootstrap test with the test statistic T = median(X1, . . . , X200), test whether the data telephone.txt stems from the exponential distribution Exp(λ) with some λ from [0.01, 0.1].

We define the null hypothesis H_0 as: the data from telephone.txt stems from the exponential distribution Exp(λ) with some λ.

For the bootstrap test we take the λ parameter for the exponential distribution as 0.03. We set B = 1000 to run our test for the t test statistic t-median.
The median of the original data (t) is 28.905

The resulting distribution from out bootstrap test as outlined can be seen in *graph 3.3*:



We calculate the p-value (using the formula in lecture 2). This results in a p-value of: 0.044. Hence we reject H_0.

Code:

```
B<- 1000
t <- median(new_df$Bills)
tstar <- numeric(B)
n = length(new_df$Bills)
l = 0.03
for (i in 1:B){
  xstar <- rexp(n,l)
  tstar[i] <- median(xstar)}
pr <- sum(tstar>t)/B
pl <- sum(tstar<t)/B
p <- 2*min(pl,pr)
```

c) Construct a 95% bootstrap confidence interval for the population median of the sample.

Constructing a 95% bootstrap confidence interval can be done by sampling the data B number of times, with replacement, for the chosen t statistic, in this case median, and calculating the value using the formula provided in lecture 2. For a 95% confidence interval, with alpha = 0.05

The resulting values are:
[16.505 36.575]

Code:
```
B <- 1000
tstar <- numeric(B)
t<-median(new_df$Bills)
for (i in 1:B){
  xstar <- sample(new_df$Bills, replace=TRUE)
  tstar[i] <-median(xstar)}
tstar25<-quantile(tstar,0.025)
tstar975<-quantile(tstar, 0.975)
ci = c(2*t-tstar975,2*t-tstar25)
```

d) Assuming X1, . . . Xn ~ Exp(λ) and using the central limit theorem for the sample mean, estimate λ and construct again a 95% confidence interval for the population median. Comment on your findings.

Assuming the data is from an exponential distribution, lambda = 1/sample mean
```
lbd <- 1/mean(new_df$Bills)
```

Lambda = 0.02202461

Median = 28.905

```
t <- qt(1-((0.025/2)*(sd(new_df$Bills)/sqrt(length(new_df$Bills)))),df=n-1)
med-t; med+t
```

```
[1] 27.08197
[1] 30.72803
```

For a 95% confidence interval, with alpha = 0.05 the population median is between 27.082 and 30.728. The confidence interval is small, which indicates that we can be confident about our findings.

e) Using an appropriate test, test the null hypothesis that the median bill is bigger or equal to 40 euro against the alternative that the median bill is smaller than 40 euro. Next, design and perform a test to check whether the fraction of the bills less than 10 euro is less than 25%.

The distribution in the data cannot be assumed to be symmetric, hence we will use the binomial test to test the null hypothesis (H_0) that the median is equal to 40. We set p=0.5.

```
> binom.test(40, length(new_df$Bills), 0.5)
```

Output:

```
        Exact binomial test

data:  40 and length(new_df$Bills)
number of successes = 40, number of trials = 192, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1532039 0.2726995
sample estimates:
probability of success
          0.2083333
```

We get a p-value < 2.2e-16, and hence we can reject H_0.

To test the null hypothesis (H_0) that the fraction of the bills are less than $10 is 25%, we can perform the same test but set p=0.25. We do this because if the fraction of bills less than $10 is 25%, then the probability of a bill being less than $10 is 0.25.

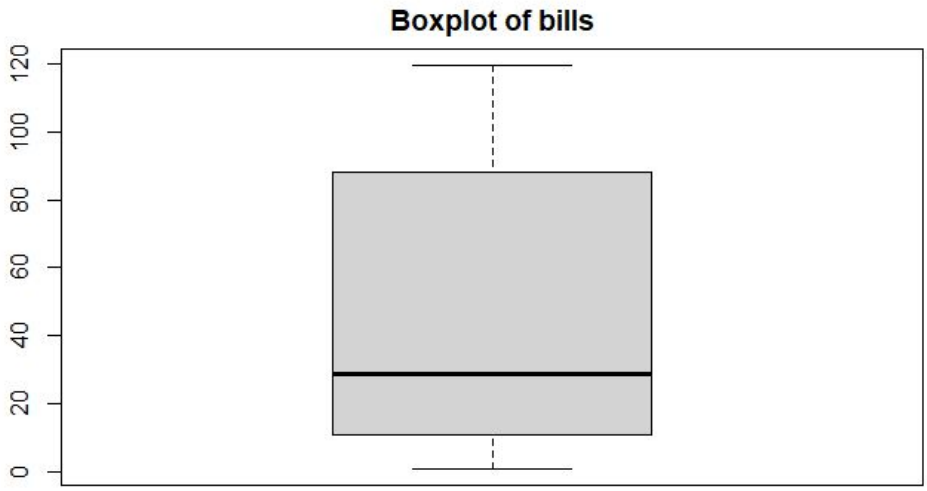
```
>binom.test(10, length(new_df$Bills), 0.25)
```

Output:

```
        Exact binomial test

data:  10 and length(new_df$Bills)
number of successes = 10, number of trials = 192, p-value = 5.661e-13
alternative hypothesis: true probability of success is not equal to 0.25
95 percent confidence interval:
 0.02525533 0.09369536
sample estimates:
probability of success
          0.05208333
```

We get a significant p-value = 5.661e-13, hence we reject H_0 that the fraction of the bills are less than $10 is 25%.

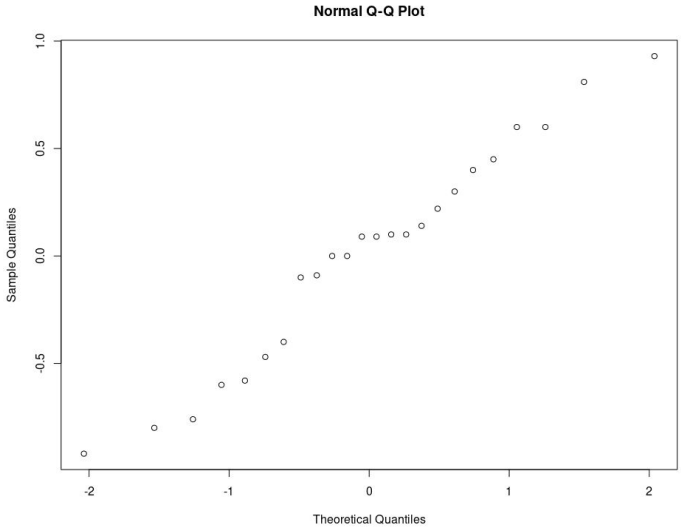
These results can be further corroborated with a box plot:



Exercise 4. Energy drink

a) Disregarding the type of drink, test whether the run times before drink and after are correlated.

Asserting the normality of the difference between the "before" and "after" performances with *qqnorm(data[,1]-data[,2])*



We can use the paired t-test to determine whether the run times before and after the drinks are correlated:

t.test(data[,1],data[,2],paired=TRUE)

This returns a p-value = 0.96, so we can determine that the run times before and after the drink are correlated.

b) Test separately, for both the softdrink and the energy drink conditions, whether there is a difference in speed in the two running tasks.

Using Pearson's product-moment correlation, we can determine if there is a difference in speed in the two running tasks. Considering the two drinks to be

lemo <- data[1:12, 1:2]
energy <- data[13:24, 1:2]

We have for the softdrink

```
cor.test(lemo[, 1], lemo[, 2])
```

That returns a p-value of 0.018, which suggests that there is a difference between before and after the two runs.

Considering the energy drink, we have:

```
cor.test(energy[,1], energy[,2])
```

That gives us a p-value of 0.009. Once again, this p-value allows us to determine that there is a difference between before and after the two runs, considering the energy drink consumption.

c) For each pupil compute the time difference between the two running tasks. Test whether these time differences are effected by the type of drink.

For testing this, we can use a permutation test for two paired samples, by testing the null hypothesis of no difference between the distributions of the two populations - aka the two types of drinks -. So, if the p-value is greater than 0.05, we can assume that there actually is a difference in the running tasks, considering the two types of drinks.

```
lemo_diff = lemo[,1] - lemo[,2]
energy_diff = energy[,1] - energy[,2]
mystat=function(x,y) {mean(x-y)}
tstar=numeric(1000)
for (i in 1:1000)
{
  datastar=t(apply(cbind(lemo_diff,energy_diff),1,sample))
  tstar[i]=mystat(datastar[,1],datastar[,2])
}
myt=mystat(lemo_diff,energy_diff)
pl=sum(tstar<myt)/B
pr=sum(tstar>myt)/B
p=2*min(pl,pr)
```

The obtained p-value is 0.17, so we can safely assume that the type of drink actually affects the time difference between the two running tasks.

d) Can you think of a plausible objection to the design of the experiment in b) if the main aim was to test whether drinking the energy drink speeds up the running? Is there a similar objection to the design of the experiment in c)? Comment on all your findings in this exercise.

For question b, a possible objection is that the second run, even if provided with energy drink or soft drink, is going to be altered by the fact that a physical challenging activity was already performed shortly before. This can alter the second run time obtained, by possibly reducing the speed. This can create a false negative, or at least a reduced effect of the energy drink/softdrink consumption on the performance of the students, due the fact that is the second run, and so having performed an already physically taxating exercise.
This does not apply to question c, because the way the test is being performed, it only considers the difference of the results of the two drinks, normalizing the input to not be strictly dependent on the second run - like point b -.

Exercise 5. Chick weights

a) Test whether the distributions of the chicken weights for meatmeal and sunflower groups are different by performing three tests: the two samples t-test (argue whether the data are paired or not), the Mann-Whitney test and the Kolmogorov-Smirnov test. Comment on your findings.

T-test

There is one numerical outcome per experimental unit (chicken weights) and there are two levels of experimental units (i.e. 'meatmeal' and 'sunflower'). The chickens are randomly assigned to either one of these levels. This means the data is *not paired* and we will use a non-paired t-test to see whether the distributions of chickens weights differ for the two different levels of feed supplements.

The two samples t-test assumes that samples come from a normal population. We test the null hypothesis $H_0 : \mu$ equals v vs. $H_1 : \mu$ not equal v

```
Output:
Welch Two Sample t-test

data:  meatmeal_sunflower$weight by meatmeal_sunflower$feed
t = -2.1564, df = 18.535, p-value = 0.04441
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -102.572435   -1.442716
sample estimates:
mean in group 4 mean in group 6
      276.9091        328.9167
```

The results of the T-test show that the effect of levels on the weight of chickens is significant with $p = 0.04$ ($p < 0.05$), hence H_0 is rejected.

Mann-Whitney test

We test the hypothesis H_0: $F = G$ that the populations are the same. So we test whether the ranks of the distributions are the same between chickens who ate sunflower and meatmeal supplements.

Output:

```
        Wilcoxon rank sum exact test

data:  meatmeal_sunflower$weight by meatmeal_sunflower$feed
W = 36, p-value = 0.06882
alternative hypothesis: true location shift is not equal to 0
```

The results of the Mann-Whitney test show that the levels are not significant with a p-value of 0.069 ($p>0.05$), hence H_0 is not rejected and we conclude that there are no significant differences between the chicken weights for meatmeal and sunflower groups.

Kolmogorov-Smirnov test

We test the null hypothesis H_0: $F = G$ that the populations are the same. We computed the highest vertical difference in summed histograms to check whether the populations come from the same distributions.

Output:

```
Two-sample Kolmogorov-Smirnov test

data:  sunflower and meatmeal
D = 0, p-value = 1
alternative hypothesis: two-sided
```

The result of the Kolmogorov-Smirnov test shows a p-value = 1 ($p>0.05$). Hence, H_0 is not rejected and we conclude that chicken weights for meatmeal and sunflower groups are not significantly different from each other.

b) Conduct a one-way ANOVA to determine whether the type of feed supplement has an effect on the weight of the chicks. Give the estimated chick weights for each of the six feed supplements. What is the best feed supplement?

First we made a linear restriction on the parameters to perform an ANOVA. We made a linear model with 'feed' as a factor with 6 different levels. We then performed an ANOVA test with 'weight' as dependent variable and 'feed' as independent and tested whether the weights of the chickens can be explained by the different food supplements.

Output:

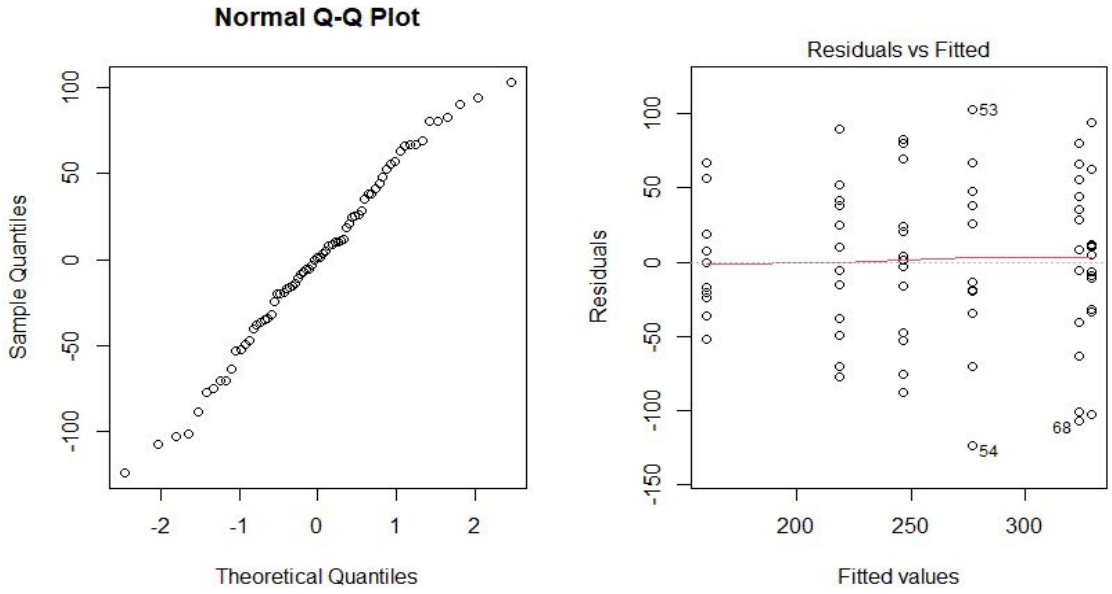
```
Analysis of Variance Table

Response: chick$weight
          Df Sum Sq Mean Sq F value    Pr(>F)
chick$feed  5 231129   46226  15.365 5.936e-10 ***
Residuals  65 195556    3009
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for testing $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ is $p<0.001$, hence H_0 is rejected, i.e. the factor 'feed' is significant.

Furthermore, with the *command fitting(chickmodel)* we obtained the estimated means of the different levels, resulting in the following means:
Casian : $\mu = 323.58$
Horsebean : $\mu = 160.2$
Linseed : $\mu = 218.75$
Meatmeal : $\mu = 276.9$
Soybean : $\mu = 246.43$
Sunflower: $\mu = 328.92$

This would mean that the chickens who ate 'sunflower' have the highest mean weight, and thus we interpret sunflower as the best food supplement.

c) Check the ANOVA 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).



Graph 5.1. Shows normality QQplot; Graph 5.2. Shows relationship between residuals and the means of 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 $p = 0.63$. 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) Does the Kruskal-Wallis test arrive at the same conclusion about the effect of feed supplement as the test in b)? Explain possible differences between conclusions of the Kruskal-Wallis and ANOVA tests.

We performed an Kruskal-Wallis test to check whether this test arrived with the same results as the one-way ANOVA-test. The results are as follows:

```
Kruskal-Wallis rank sum test

data:  chick$weight and chick$feed
Kruskal-Wallis chi-squared = 37.343, df = 5, p-value = 5.113e-07
```

The p-value for testing $H_0 : F_1 = F_2 = F_3 = F_4 = F_5 = F_6$ is 5.113e-07, hence H_0 is rejected. Even though we arrive at the same conclusion (i.e. rejecting H_0), the p-value is different compared to the ANOVA test. The differences could be due to the fact that an ANOVA is a parametric test while kruskal test is a non-parametric approach. Thus, kruskal test does not need any distributional assumption and doesn't rely on the assumptions of normality. Instead the test is based on differences between levels in ranks.

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).

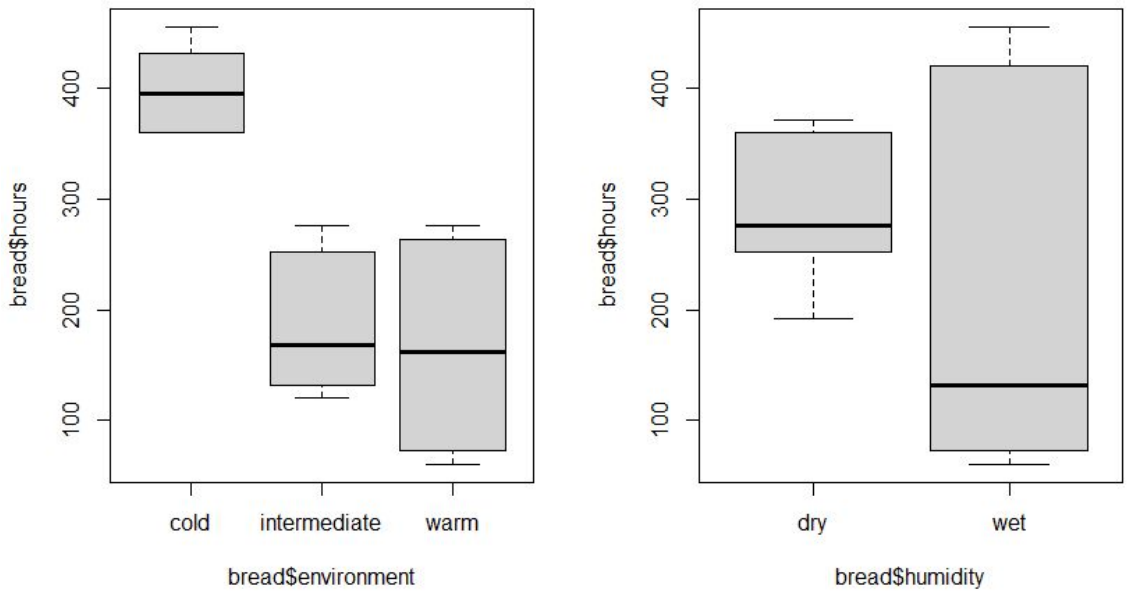


Fig. 1.1 Box-plot showing the main effects of the factors

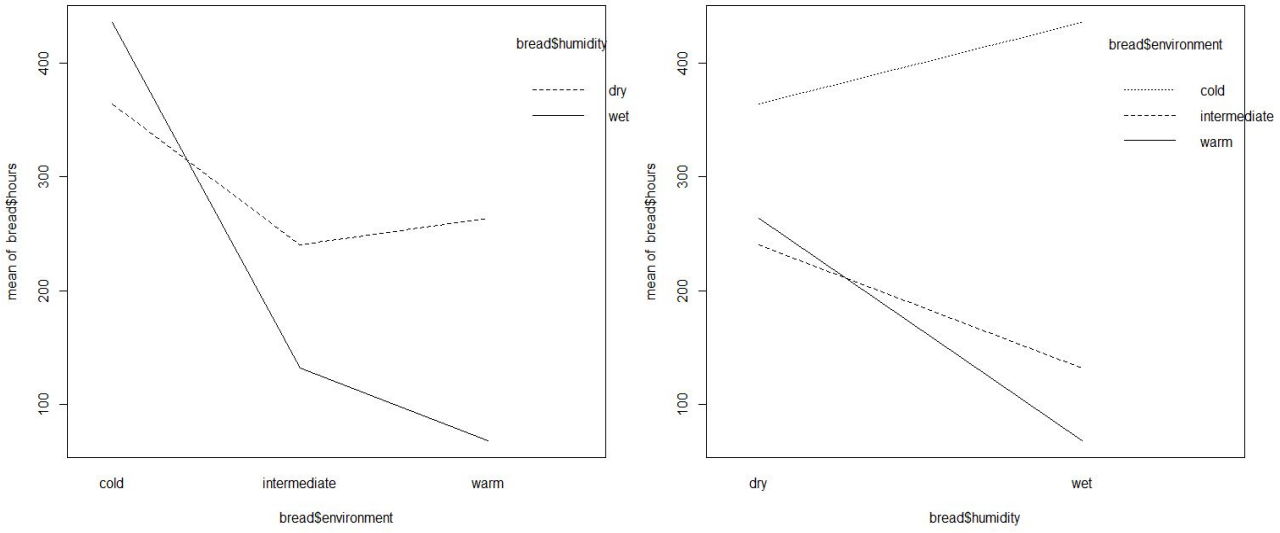


Fig. 1.2 Box-plot showing the interaction of environment and humidity on hours to decay; Interaction of humidity and environment on hours to decay.

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.

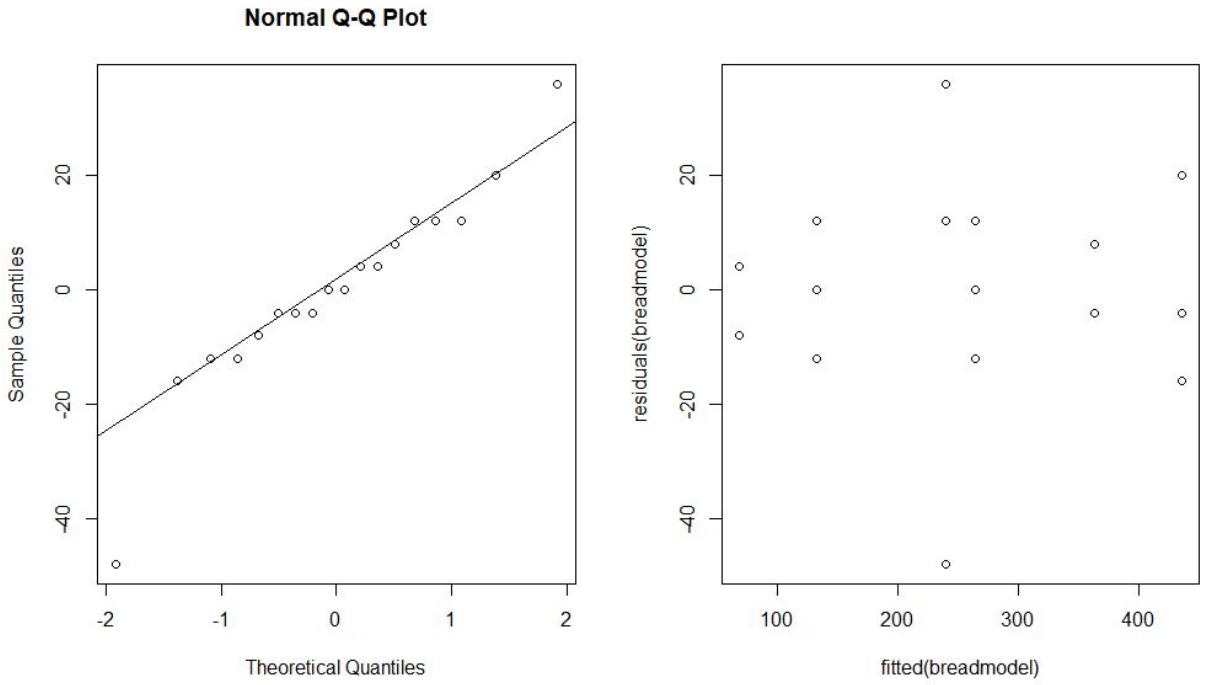


Fig. 1.3 QQplot to check normality; Plotted spread of residuals to check for equal variances.

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.

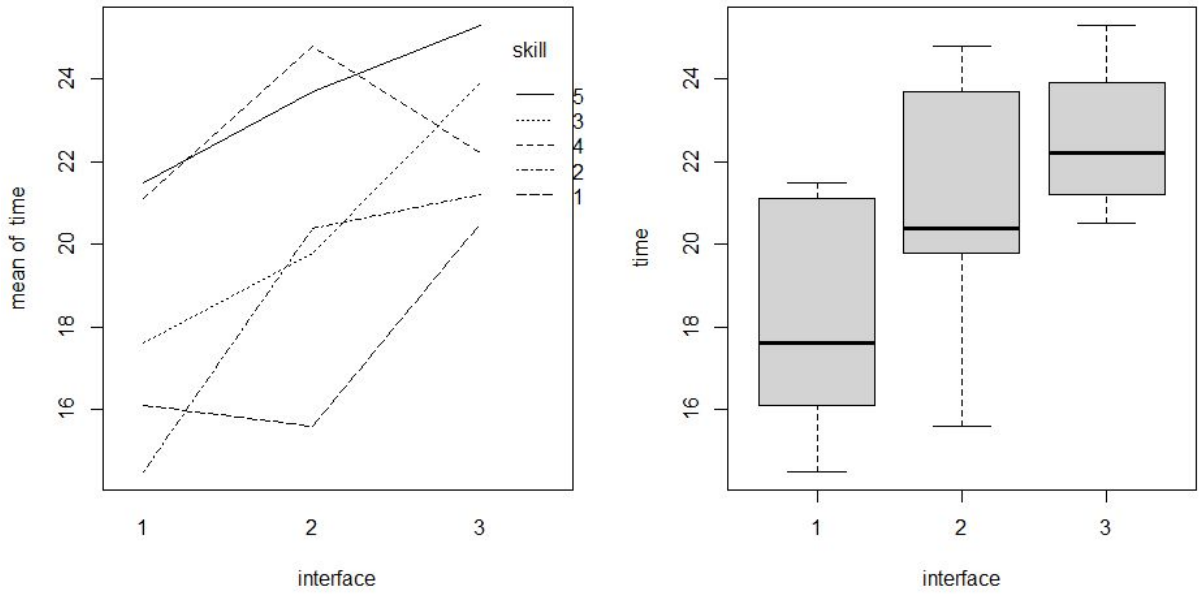


Figure 2.1: interaction plot of interface and skil ; Figure 2.2 Boxplot of interface and search time.

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).

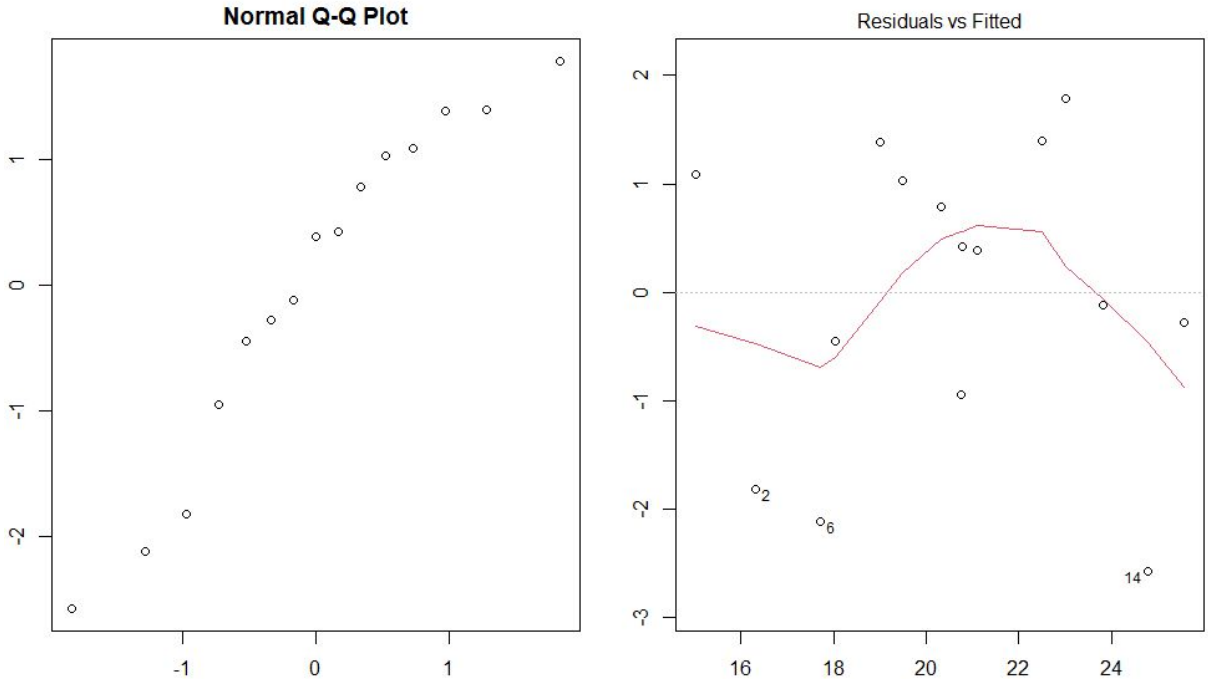


Figure 2.3. Shows normality QQplot; Figure 2.4. Shows relationship between residuals and the means of the different levels.

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.

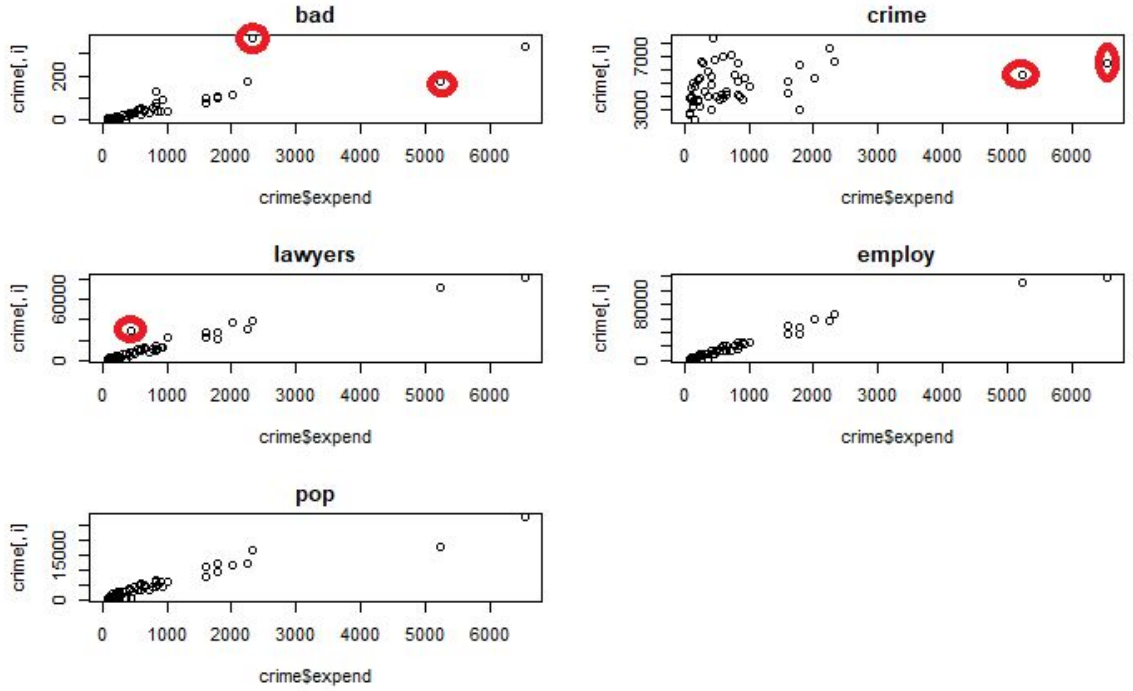


Fig 5.1 Plots of predictor variables against dependent variables

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

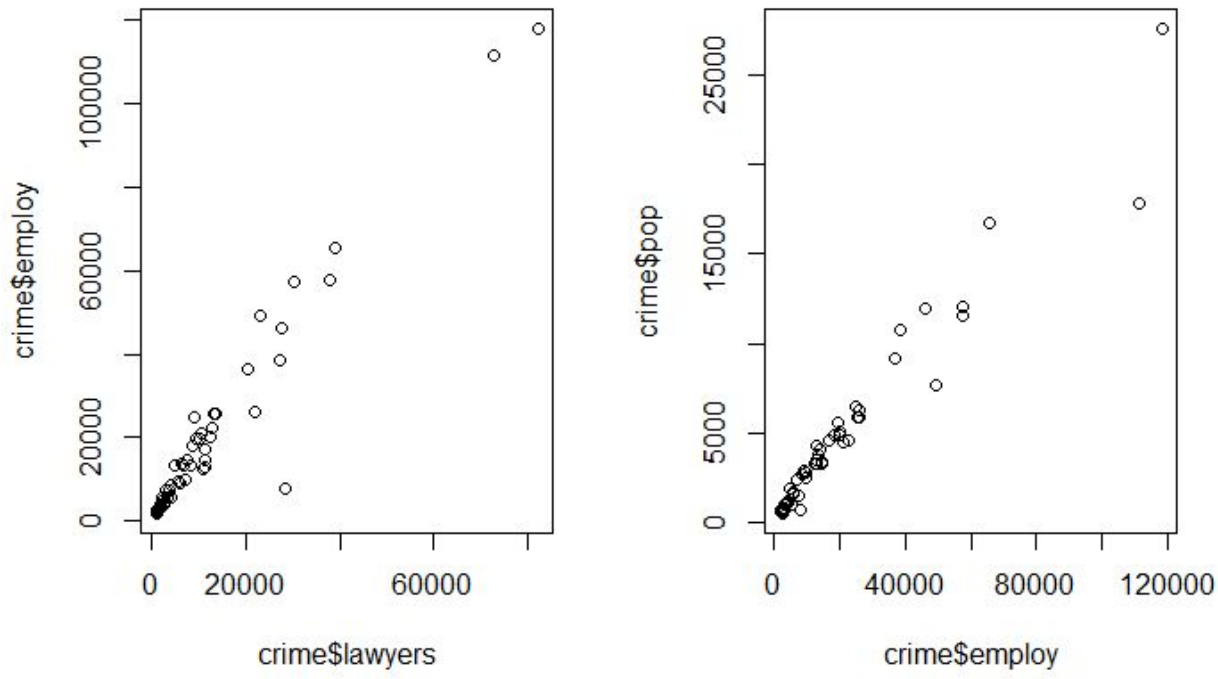


Fig 5.2 Plot of dependent variables against each other

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.

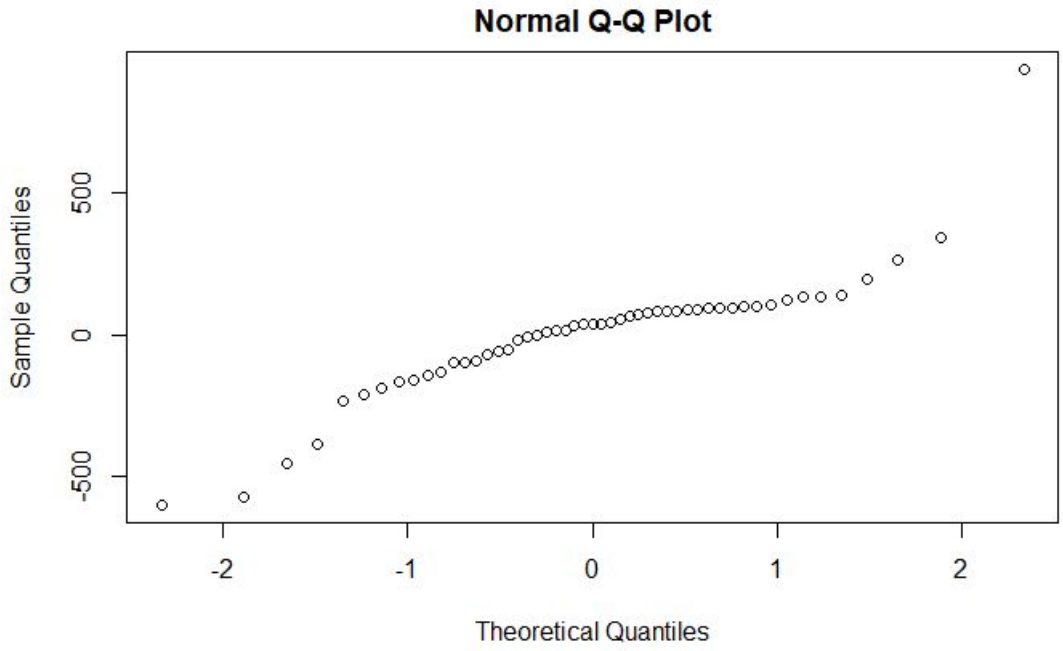


Fig. 5.3. QQ-plot of the residuals of the step up model.

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.

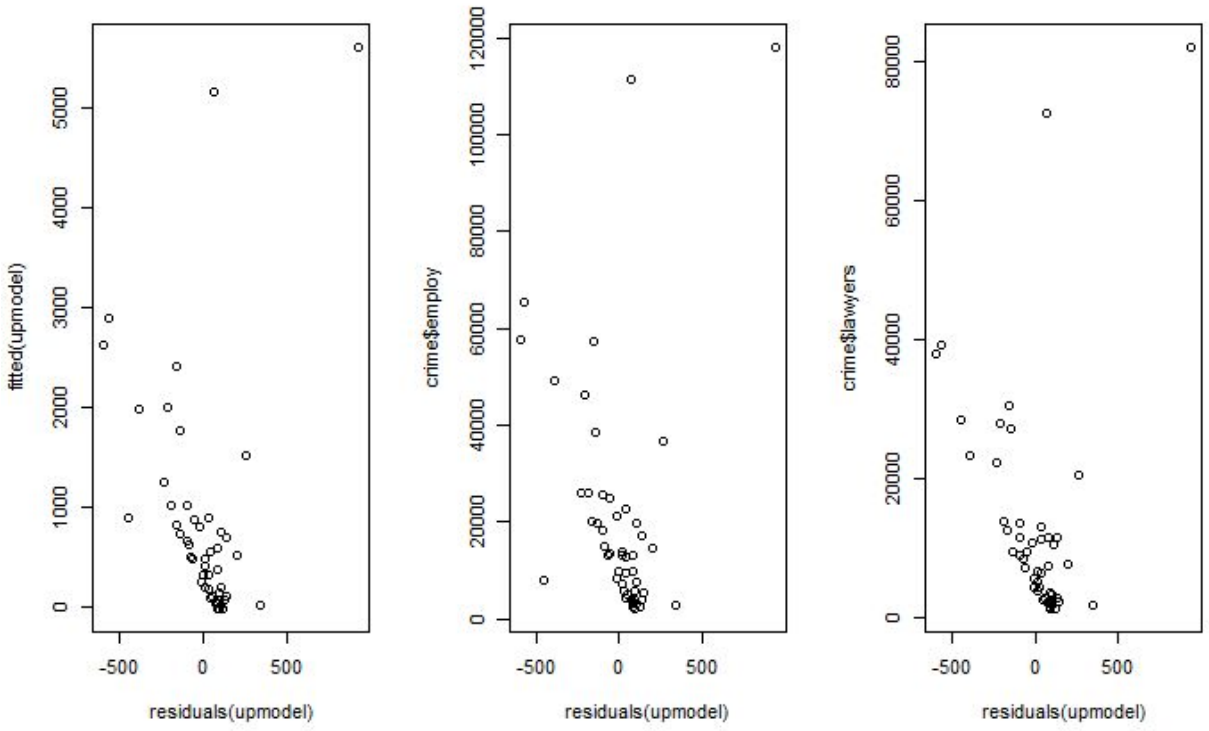


Fig 5.4: Residuals of the step up model, plotted against fitted and independent variables.

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.

Experimental Design and Data Analysis

Assignment 3

Tommaso Castiglione Ferrari 2673807

Daniyal Selani 2692551

Simone Korteling 2671463

Group 71

Exercise 1.

A.

Make an informative plot of the data. Investigate whether sexual activity influences longevity by performing a statistical test, without taking the thorax length into account. What are the estimated longevities for the three conditions? Comment.

```
df$loglongevity = log(df$longevity)
plot(df$longevity~df$thorax,pch=as.character(df$activity))
```

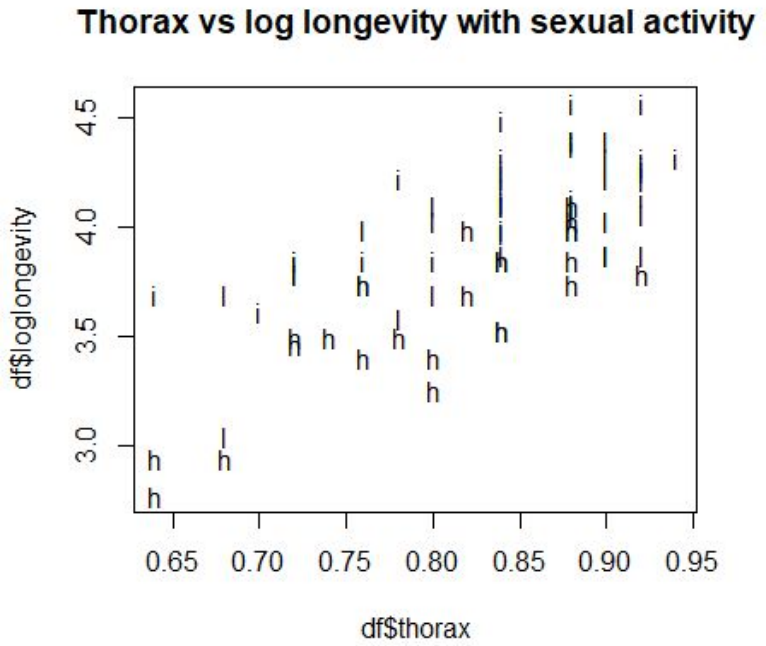


Fig 1.1

```
m1 = lm(loglongevity~activity, data=df)
print(anova(m1))

Analysis of Variance Table

Response: loglongevity
          Df Sum Sq Mean Sq F value    Pr(>F)
activity   2 3.6665  1.8333  19.421 1.798e-07 ***
Residuals 72 6.7966  0.0944
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P value < 0.05, and hence we can reject H0. Sexual activity is significant for loglongevity. However 1-way anova test with a factor is not correct. We must consider the numeric variable, or atleast investigate the interaction between numeric and factor variable.

```
summary(m1)
Call:
lm(formula = loglongevity ~ activity, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95531 -0.13338  0.02552  0.20891  0.49222

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.60212    0.06145  58.621  < 2e-16 ***
activityisolated  0.51722    0.08690   5.952 8.82e-08 ***
activitylow       0.39771    0.08690   4.577 1.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3072 on 72 degrees of freedom
Multiple R-squared:  0.3504,  Adjusted R-squared:  0.3324
F-statistic: 19.42 on 2 and 72 DF,  p-value: 1.798e-07
```

mean loglongevity for high activity = 3.6
isolated activity = 3.6+0.52 = 4.12
low activity = 3.6 + 0.40 = 4.0

B. Investigate whether sexual activity influences longevity by performing a statistical test, now including thorax length as an explanatory variable into the analysis. Does sexual activity increase or decrease longevity? What are the estimated longevities for the three groups, for a fly with average thorax length?

```
m2 = lm(loglongevity~thorax+activity,data=df)
drop1(m2,test='F')
```

```
Model:
loglongevity ~ thorax + activity
         Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>                 2.9180 -235.50
thorax    1    3.8786 6.7966 -174.08  94.374 1.139e-14 ***
activity  2    2.1129 5.0309 -198.64  25.705 4.000e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P value < 0.05. We can reject H0, sexual activity does affect log longevity

```
summary((m2))
```

```
Call:
lm(formula = loglongevity ~ thorax + activity, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4858 -0.1612  0.0104  0.1510  0.3574

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.21893    0.24865   4.902 5.79e-06 ***
thorax            2.97899    0.30665   9.715 1.14e-14 ***
activityisolated  0.40998    0.05839   7.021 1.07e-09 ***
activitylow       0.28570    0.05849   4.885 6.18e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2027 on 71 degrees of freedom
Multiple R-squared:  0.7211,  Adjusted R-squared:  0.7093
F-statistic:  61.2 on 3 and 71 DF,  p-value: < 2.2e-16
```

The estimates from the summary show that high sexual activity results in shorter longevity.

This is corroborated by the loglongevity for the average thorax length for each activity level.

```
avgthorax = mean(df$thorax)
for (i in c('low', 'isolated', 'high')){


ndata = data.frame(thorax=avgthorax, activity=i)
ndata$activity = as.factor(ndata$activity)
print(i)
print(predict(m2, ndata, type='response'))
}
```

```
[1] "low"
      1
3.96091
[1] "isolated"
      1
4.08519
[1] "high"
       1
3.675209
```

C.

How does thorax length influence longevity? Investigate graphically and by using an appropriate test whether this dependence is similar under all three conditions of sexual activity.

```
plot(loglongevity~thorax,pch=unclass(activity), data=df)
for (i in c('high', 'low', 'isolated')) abline(lm(loglongevity~thorax,data=df[df$activity==i,]))
```

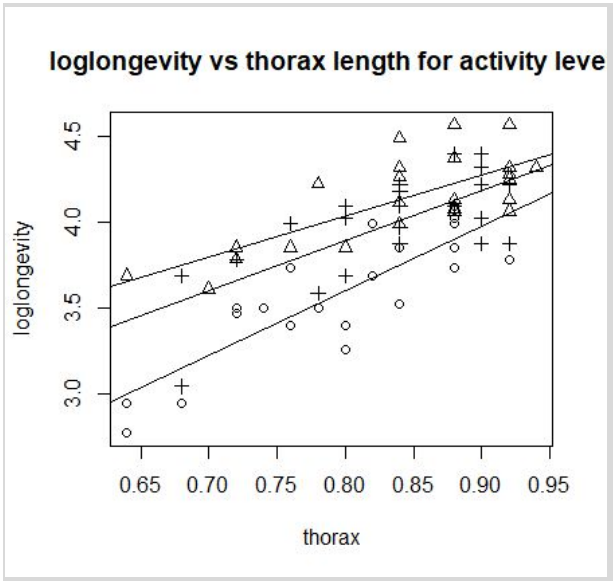


Fig 1.2

```
m3 = lm(loglongevity~activity*thorax,data=df)
anova(m3)
```

Analysis of Variance Table

Response: loglongevity

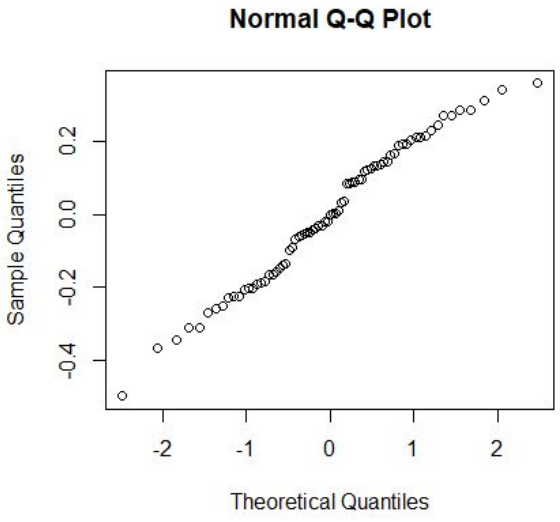
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
activity	2	3.6665	1.8332	45.7687	2.228e-13	***
thorax	1	3.8786	3.8786	96.8327	9.020e-15	***
activity:thorax	2	0.1542	0.0771	1.9251	0.1536	
Residuals	69	2.7638	0.0401			

P value for thorax is < 0.05. Hence we can reject H0, and deduce that thorax is indeed significant for loglongevity.
P value for interaction between activity and thorax is not significant, p > 0.05. H0 cannot be rejected, and there is no significant interaction between the two variables.

D.
Which of the two analyses, without or with thorax length, do you prefer? Is one of the analyses wrong?

As there is no interaction between the predictor variables, analysis with and without thorax is valid. However, because r squared error for model without thorax is considerably lower, we prefer that analysis.

E.
Verify normality and heteroscedasticity by making a normal QQ-plot of the residuals, and a residuals versus fitted plot, for the analysis that includes thorax length.



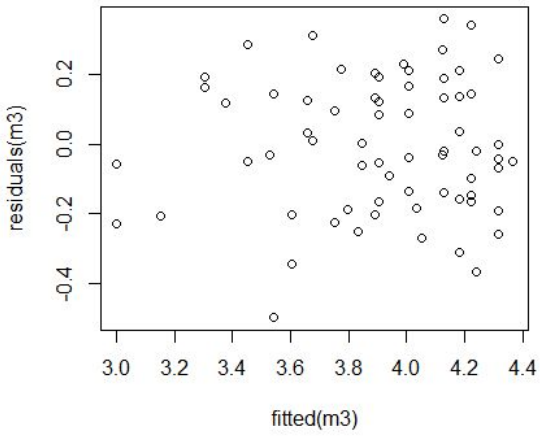


Fig 1.3

Fig 1.4
The qqplot shows residuals are normally distributed and the residuals vs fitted plot is highly heteroscedastic.

F.
Perform the ancova analysis with the number of days as the response, rather than its logarithm. Verify normality and homoscedasticity of the residuals of this analysis. Was it wise to use the logarithm as response?

```
m4 = lm(longevity~thorax+activity,data=df)
drop1(m4, test='F')
```

```
Model:
longevity ~ thorax + activity
         Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>                7673 355.10
thorax    1    7686.8 15360 405.15  71.127 2.624e-12 ***
activity  2    4966.7 12640 388.53  22.979 2.016e-08 ***
---
```

ANCOVA reveals that both thorax and activity are significant ($p < 0.05$). Hence, we can reject H0 for both variables.

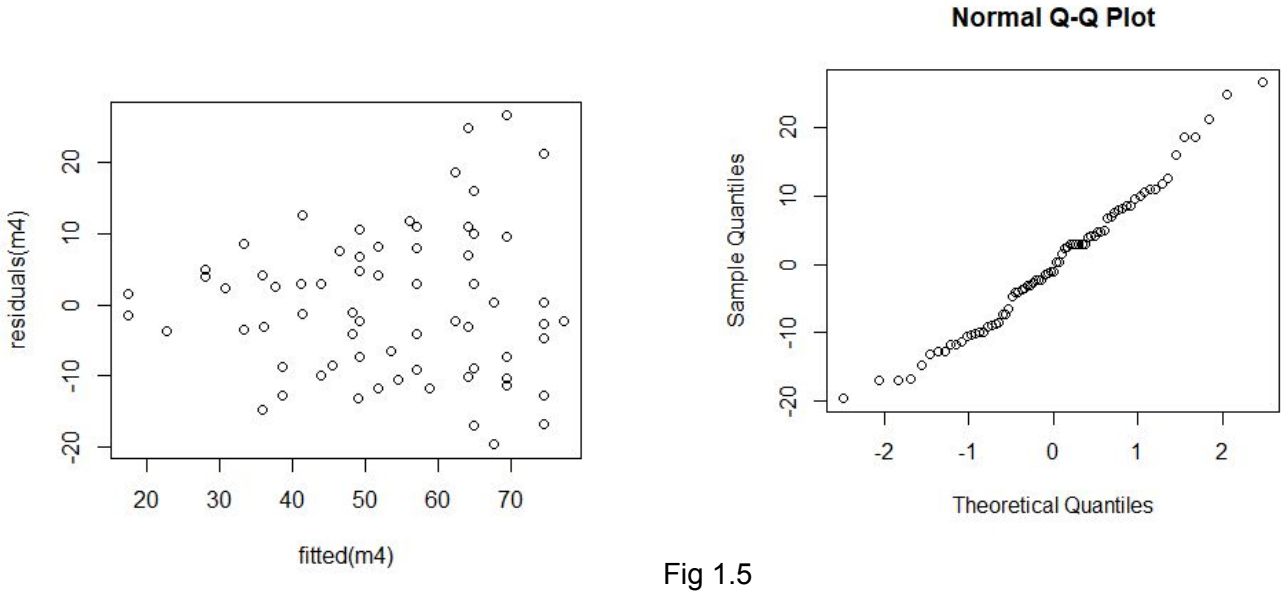


Fig 1.5

Fig 1.6

The qqplot of the residuals reveals a slightly less ideal normal distributions, and the fitted vs residuals plot reveals slightly more homoscedasticity.

Exercise 2. Military coups in Africa

To study the influence of different political and geographical variables on the number of military coups, these data are collected for several Sub Saharan African countries in the file africa.txt.

a) Study the data and give a few (>1) summaries (graphics or tables).

Code:

```
titanic$Age = as.numeric(titanic$Age)
titanic$PClass = as.numeric(titanic$PClass)
hist(titanic[,3],main="Age")
hist(titanic[,2], main="PClass")
```

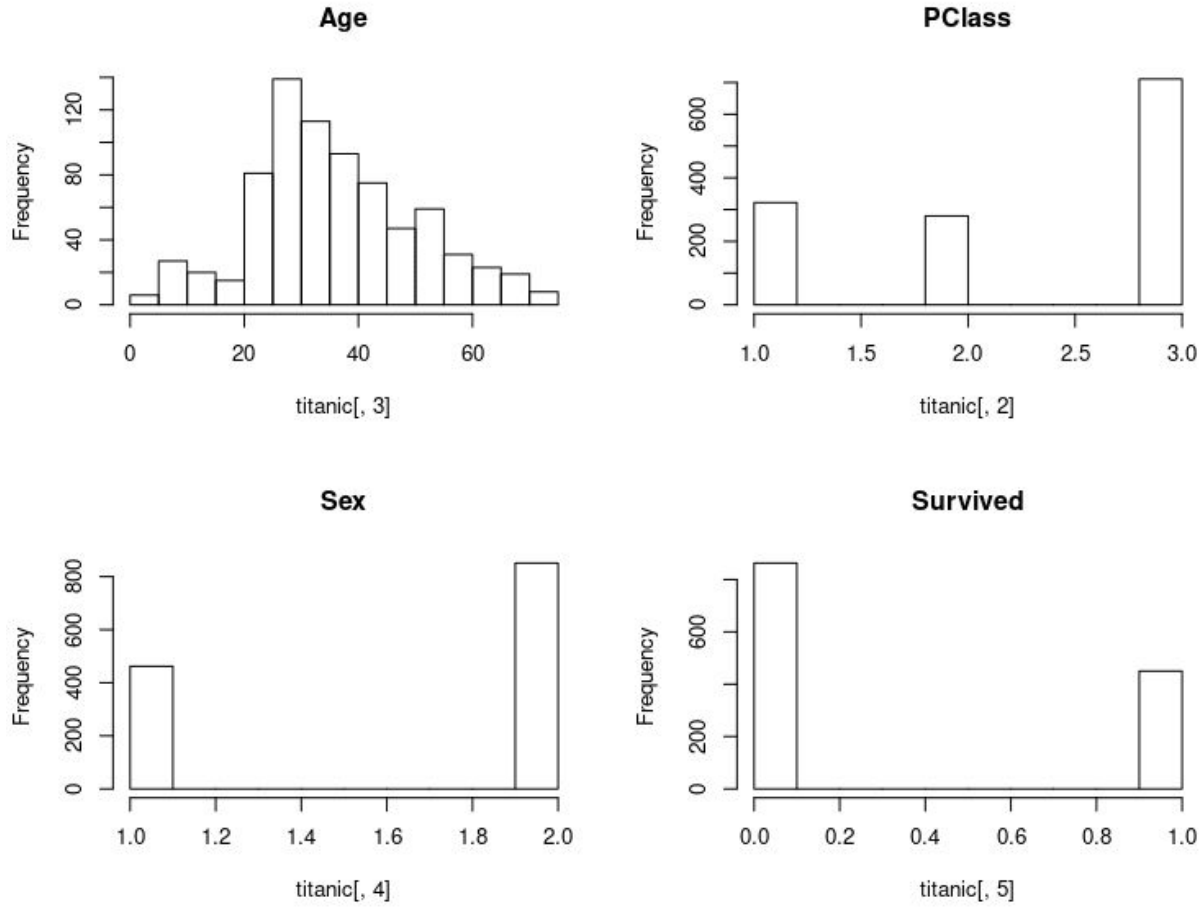


Fig. 2.1

In the images Fig. 2.1 we can observe the number of data available related to the different variables. As we can see, in our data, most of the people that were onboard of the Titanic were male, in their late 20s and from and residing in the 3 passenger class. Unfortunately, the available data shows, as we know, that most of these people did not survive when the Titanic sank.

Code:
```
xtabs(~titanic$PClass+titanic$Sex, data=titanic)
```

	Sex	
PClass	1	2
1st	143	179
2nd	107	173
3rd	212	499

In the table above we can see the combination of people divided by gender and passenger class. As already seen in the histograms above, the vast majority of the population resides in the third class, with an overwhelming difference between males (2) and females (1)

Code:
```
xtabs(~titanic$Survived+titanic$Sex, data=titanic)
```

	Sex	
Survived	1	2
0	154	709
1	308	142

In the table above we can see the combination between who survived and the gender of the passengers, and we can determine that there is a correlation between the gender of each passenger and its probability of surviving the catastrophe.

Code:
```
xtabs(~titanic$Survived+titanic$PClass, data=titanic)
```

	PClass		
Survived	1st	2nd	3rd
0	129	161	573
1	193	119	138

In the table above we can see the combination between who survived and the residing class of the passengers, and we can determine that there is a correlation between the class of each passenger and its probability of surviving the catastrophe.

Code:

```
tot = xtabs(titanic$Survived~titanic$PClass+titanic$Sex, data=titanic)
print(round(tot/xtabs(~titanic$PClass+titanic$Sex), 2))
```

```
          Sex
PClass    1    2
1st       0.94 0.33
2nd       0.88 0.14
3rd       0.38 0.12
```

Finally, in the table above we can see an early probability result of the chance of survival of a passenger, considering its gender and class on the cruise. As already anticipated, we can see a clear increase in the probability of surviving the event when moving to a more prestigious class, and moving from male to female. We can see, in fact, that the highest number of survivors in percentage were women from the 1st class with a 94% survival rate, while men from the 3rd class had an overall survival rate of 12%.

b) Fit a logistic regression model (no interactions yet) to investigate the association between the survival status and the predictors PClass, Age and Sex. Interpret the results in terms of odds, comment.

Considering the variables PClass, Age and Sex as predictors, we obtain the following logistic regression model:

Code:

```
titanic$Sex = as.numeric(titanic$Sex)
titanic$PClass = as.numeric(titanic$PClass)
titanic$Age = as.numeric(titanic$Age)
call <- glm(Survived~PClass+Age+Sex,family=binomial, data=titanic)
summary(call)
```

The estimated odds from this model are the following:

$$\hat{o}_k = Pr(Y_k =1) \setminus Pr(Y_k =0) \approx$$
$$\exp\{7.88 -1.26\ PClass_k - 0.04\ age_k - 2.63\ Sex_k \}$$

Analyzing the odds obtained from this model, we can preliminary determine some interesting features. First of all, we can see that by increasing the passenger class (as a remainder, 1st class is the most prestigious, while the 3rd is the lowest), the probability of surviving sensibly decreases. Moreover, being a male (or Sex = 2) has the most drastic change in the linear predictor, by reducing the chance of surviving of a factor of -2.63. Finally, it seems that the age variable has the minimum effect on the possibility of surviving.

c) Investigate for interaction of predictor Age with factors PClass and Sex. From this and b), choose (and justify) a resulting model. For this model, report the estimate for the probability of survival for each combination of levels of the factors PClass and Sex for a person of age 53.

For investigating the interaction between the variable Age and, separately, the two variables PClass and Sex (i.e. Age:PClass and Age:Sex) we can implement the following logistic regression model:

Code:
```
titanic$Sex = factor(titanic$Sex)
titanic$PClass = factor(titanic$PClass)
call1 <- glm(Survived~Age*(PClass+Sex),family=binomial, data=titanic)
anova(call1_alternativa, test="Chisq")
```

```
          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      755    1025.57
Age            1    2.972       754      1022.60    0.08471 .
PClass         2  113.028       752      909.57   < 2.2e-16 ***
Sex            1  214.986       751      694.59   < 2.2e-16 ***
Age:PClass  2    5.477         749       689.11     0.06467 .
Age:Sex        1   26.979       748      662.13      2.057e-07 ***
```

As we can see from the result obtained by the Anova executed with the Chi-squared test, there is no meaningful interaction between Age and PClass, while it seems that a strong interaction between Age and Sex is present. Considering this, and the importance represented by the variable PClass from the odds analysis in the previous sub-question, we can implement a resulting model on the basis of this information:

Code:
```
titanic$PClass = factor(titanic$PClass)
titanic$Sex = factor(titanic$Sex)
titanic$Age = as.numeric(titanic$Age)
model = glm(Survived~PClass+Age*Sex,data=titanic,family=binomial)
```

Given this model, we can estimate the probability of survival of a 53 years old person for each combination of levels of the factors PClass and Sex:

Code:
```
for (pc in levels(titanic$PClass)){
  for (sex in levels(titanic$Sex)) {
    newdata = data.frame(Sex=sex, PClass=pc, Name="Mr. Mc Buttface", Age=53)
    print(newdata)
    print(predict(model, newdata, type="response"))
    }}
 Sex PClass Name Age
   1      1        test    53   = 0.95 %

  Sex PClass Name Age
   2      1        test    53   = 0.23 %

  Sex PClass Name Age
```

```
 1     2        test    53   = 0.79 %

Sex PClass Name Age
 2     2        test    53   = 0.06 %

 Sex PClass Name Age
 1     3        test    53   = 0.55 %

 Sex PClass Name Age
 2     3        test    53    = 0.02 %
```

Remembering that Sex = 1 is female and Sex = 2 is male, and PClass = 1 is first class (most luxurious) and PClass = 3 is third class (most humble).

d) Propose a method to predict the survival status and a quality measure for your prediction and describe how you would implement that method (you do not need to implement it).

Fitting the observed data $(X_1, Y_1), \ldots, (X_N, Y_N)$ in logistic regression, we have

$$\Pr(Y_k = 1) = 1 / (1 + e^{-x^{(T)}_{(k)} \theta}) \quad \text{for } k = 1, \ldots, N,$$

And we obtain (by the maximum likelihood) an estimate $\hat{\theta}$ of the parameter θ.
For a new predictor vector X_{new}, we can predict its success probability

$$\hat{P}_{new} = 1 / (1 + e^{-x^{(T)}_{(new)} \hat{\theta}})$$

Now we can use $\hat{P}_{new}$ to predict the new label $\hat{Y}_{new}$ as

$$\hat{Y}_{new} = \begin{cases} 1 & \text{if } \hat{P}_{new} \geq p_0 \\ 0 & \text{if } \hat{P}_{new} < p_0 \end{cases}$$

for some threshold $p_0 \in [0, 1]$.

Considering this, because we used the maximum likelihood, we can consider the value of $\hat{P}_{new}$ as a measure of the quality of the prediction. In fact, if $\hat{P}_{new} \gg p_0$ or $\hat{P}_{new} \ll p_0$ we can have a more “safe” prediction of $\hat{Y}_{new}$ - i.e. we can assert with more certainty that $\hat{Y}_{new}$ is either equal to 1 or to 0, namely if $\hat{P}_{new} \gg p_0$ or $\hat{P}_{new} \ll p_0$ -, than if $\hat{P}_{new}$ is close to the value of p_0.

e) Another approach would be to apply a contingency table test to investigate whether factor passenger class (and gender) has an effect on the survival status. Implement the relevant test(s).

As an alternative approach, we can use the contingency table tests Chi-squared and Fisher tests, for studying the dependency of namely the class and the survival rate, and the gender and the survival status.

Code:
```
chisq.test(titanic$PClass, titanic$Survived)
fisher.test(titanic$Sex, titanic$Survived)
```

The data obtained is:

```
data:  titanic$PClass and titanic$Survived
X-squared = 172.3, df = 2, p-value < 2.2e-16
```

And

```
data:  titanic$Sex and titanic$Survived
p-value < 2.2e-16
```

So it looks like there is a strong dependency relationship for both the passenger class and the survival rate, and for the gender and the survival rate - as we already assessed -.

f) Is the second approach in e) wrong? Name both an advantage and a disadvantage of the two approaches,relative to each other.

The second approach is not wrong, but it retrieves different information, and has some dataset sizes limitations.
In fact, the output obtained using the logistic regression model uncovers only the probability of survival.
Conversely, the contingency tests retrieve a different type of information, namely the dependency between different variables.
Moreover, the Fisher test can only be used on tables that are 2x2 or, in other words, one variable against another variable. Therefore, the Fisher test is more precise when applied to small datasets, which is not true for logistic regression. But, the disadvantage of the fisher test with regards to the logistic regression model resides exactly in its size limitations, that the logistic regression does not have.
On the other side, also the Chi-squared test is size bounded, being more accurate in its approximation with a wider dataset. This problem arises because chi-squared follows the chi-squared distribution only approximately. The more observations we have, the better the approximation is going to be. This, once again, is not true for the logistic regression model.
Finally, the latter allows for predictions to be made, because it has an intercept and coefficients.
For this analysis, the logistic regression approach seems to be more accurate and an overall better choice, considering the dataset size and the prediction task of the exercise.

Exercise 3. Military coups in Africa

To study the influence of different political and geographical variables on the number of military coups, these data are collected for several Sub Saharan African countries in the file africa.txt.

a) Perform Poisson regression on the full data set africa, taking miltcoup as response variable, Comment on your findings.

In this research the response variable *Y* is a count, namely = miltcoup (numerical). The explanatory variables are both numerical and factorial. Before we performed the Poisson regression we changed the variables onto the right data type, resulting in the following variables: oligarchy(numerical), pollib (factor, with three levels 0-2), parties (numerical), pctvote (numerical), popn (numerical), size (numerical), numelec (numerical), numregim (numerical).

We performed the Poisson regression with the following code:

Code:

```
glm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote + popn + size + numelec + numregim,
family = poisson, data = africa)
```

Output:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2334274	0.9976112	-0.234	0.81500
oligarchy	0.0725658	0.0353457	2.053	**0.04007** *
pollib1	-1.1032439	0.6558114	-1.682	0.09252 .
pollib2	-1.6903057	0.6766503	-2.498	**0.01249** *
parties	0.0312212	0.0111663	2.796	**0.00517** **
pctvote	0.0154413	0.0101027	1.528	0.12641
popn	0.0109586	0.0071490	1.533	0.12531
size	-0.0002651	0.0002690	-0.985	0.32444
numelec	-0.0296185	0.0696248	-0.425	0.67054
numregim	0.2109432	0.2339330	0.902	0.36720

With an alpha of <0.05 the variables ‘oligarchy’, ‘pollib2’ and ‘parties’ have a significant effect on the number of military coups. The coefficient for oligarchy is 0.073, which indicates that the expected log count for the number of military coups for a one-unit increase in oligarchy is 0.073. The coefficient for pollib2 is -1.69, which indicates that the expected log count for the number of military coups for a one-unit increase in pollib2 is -1.69. The coefficient for parties is 0.031, which indicates that the expected log count for the number of military coups for a one-unit increase in parties is 0.031. The interpreting of the output results in the following model for estimating the number of military coups: -0.23intercept + 0.073oligrchy - 1.69pollib2 + 0.031parties.

b) Use the step down approach (using output of the function summary) to reduce the number of explanatory variables. Compare the resulting model with your findings in a).

By using the step down method we iteratively excluded the variable with the highest p-value, resulting in a final model with only significant variables explaining the outcome. We respectively excluded the following variables to improve the model: numregim, numelec,size, popn, pctvote. The step down procedure results

in a model with the following variables: oligarchy, pollib and parties, as significant explanatory variables for the response variable.

Code
```
summary(glm(miltcoup ~ oligarchy + pollib + parties, family = poisson,  data= africa))
```

Output:
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.825480   0.527632   1.565  0.11770
oligarchy    0.092622   0.021779   4.253 2.11e-05 ***
pollib      -0.574103   0.204383  -2.809  0.00497 **
parties      0.022059   0.008955   2.463  0.01377 *
```

Comparing this with the model of question 3a, we see a difference in the number of variables. Where we used 8 variables to explain the outcome variable in the first model, we now have 3 variables (oligarchy + pollib + parties) that are essential in explaining *Y*. The first model has a degrees of freedom of 27, which is lower than the top-down model with a degrees of freedom of 32. The p-values of the variables in the first model are (except for parties) higher than our final top-down model. So for most of the variables the statistics increase (lower p-values), when we exclude non-significant variables from the model.

c) Predict the number of coups for a hypothetical country for all the three levels of political liberalization and the averages (over all the counties in the data) of all the other (numerical) characteristics. Comment on your findings.

First we changed the data such that we use the average of the numerical explanatory variables. The response variable *Y* is the number of military coups. We then predicted the number of military coups for the three levels of political liberalization.

Code:
```
pred_lib_data = data.frame(oligarchy=avg_oligarchy, pollib=0, parties=avg_parties)  #level 0
output = predict(topdown_model, pred_lib_data, type="response"); output

pred_lib_data = data.frame(oligarchy=avg_oligarchy, pollib=1, parties=avg_parties)  #level 1
output = predict(topdown_model, pred_lib_data, type="response"); output

pred_lib_data = data.frame(oligarchy=avg_oligarchy, pollib=2, parties=avg_parties)  #level 2
output = predict(topdown_model, pred_lib_data, type="response"); output
```

The output shows that for level 0 on ‘political liberalization’ the number of predicted military coups of a hypothetical country is: 3.04. For level 1 on ‘political liberalization’ the number of predicted military coups of a hypothetical country is: 1.71. Finally for level 2 on ‘political liberalization’ the number of predicted military coups of a hypothetical country is: 0.96. We observe that the more civil rights a country has (level 2) the lower the corresponding number of military coups is.