Statistics Seminar 2

Group 1-2

Task 1

Four different lipid-lowering drugs (A, B, C and D) were tested in a group of 32 patients. Their cholesterol and triglycerid concentrations were measured, and the data are found in choldata.csv.

Analyse the data with one or more suitable statistical techniques and discuss what can be concluded about potential differences between the drugs.

(1.1) Set null hypothesis

There is no significant difference between the 4 drugs

(1.2) Check correlation between cholesterol and triglyceride

First of all, it is necessary to make sure if the responses depend on each other. If they are related, each figure should be considered as a parameter when analysing data. Therefore, a linear regression analysis and a correlation test is performed.

(A) Linear regression analysis with fixed effect model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.3385      0.4490    9.662    1.01e-10 ***
table1$triglyc   -0.2663      0.2668   -0.998       0.326
Residual standard error: 1.057 on 30 degrees of freedom
Multiple R-squared:  0.03214,      Adjusted R-squared:  -0.0001205
F-statistic: 0.9963 on 1 and 30 DF,  p-value: 0.3262
```

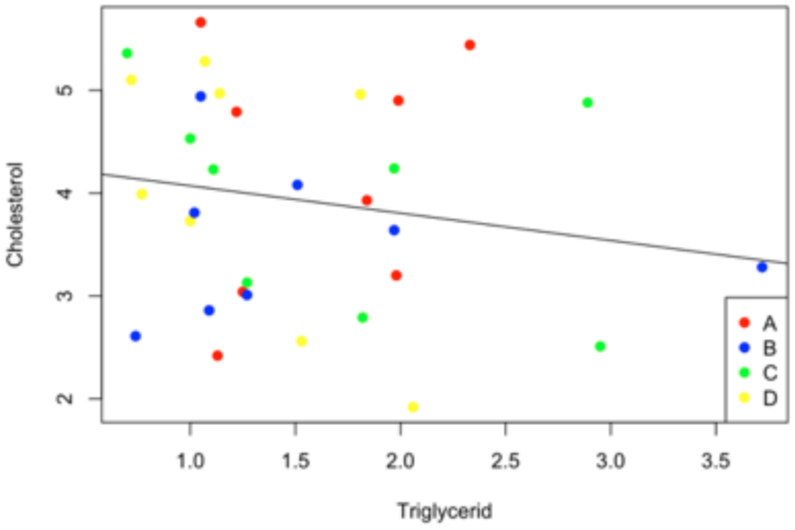


Figure 1-1. *Scatter plot total Cholesterol by Triglycerid*

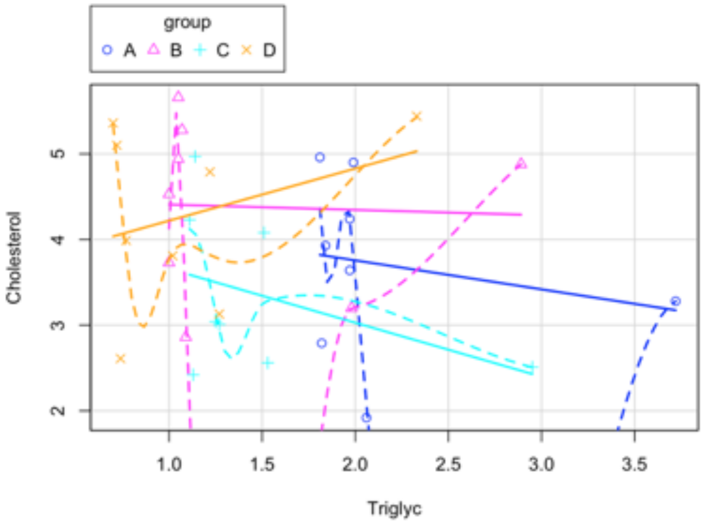


***Figure 1-2**. Scatter plot by groups*

Group	A	B	C	D
p-value	0.6188	0.9198	0.3179	0.4466

***Table 1-1.** The results of linear regression of Cholesterol by Triglycerid*

Figure 1-1 indicates that the data were distributed nonlinearly. Figure 1-2 includes the relationship between cholesterol and trigycerid by each group. If they have a linear relationship, the dash line should be close to the line, but in this graph, the dash line and the line look very different. And Table1-1 shows that the linear regression test results for the cholesterol and triglyceride of the group are respectively shown, and all values are over 5%, indicating that they are not linear.

(B) Correlation test

```
Pearson's product-moment correlation

data:  table1$triglyc and table1$cholesterol
t = -0.99813, df = 30, p-value = 0.3262
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4969104  0.1807107
sample estimates:
    cor
-0.1792802
```

According to the above two test results, there is no significant relationship between both variables. Because both p-values are over 5%, it cannot be rejected the null hypothesis of correlation test that there is no relationship between both of them.

(1.3) Check normality to choose proper analysing model

We will anlayse two responses data respectively, because they are not related.

(A) Cholesterol

If you look at Figure 1-3 and Figure 1-4, the data appear to be normally distributed for each medicine. The Shapiro test was operated to confirm the results clearly. According to Table 1-1, since all p-values are over 5%, we can assure the responses of cholesterol were normally distributed.

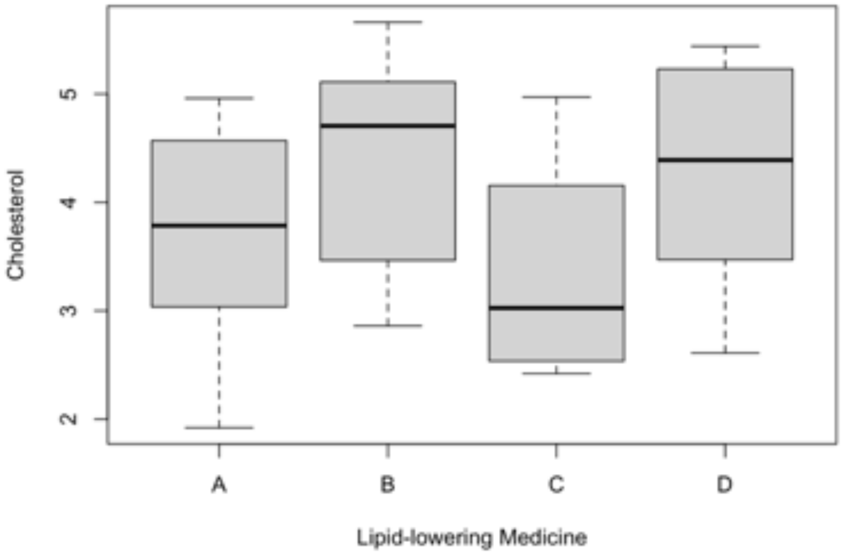
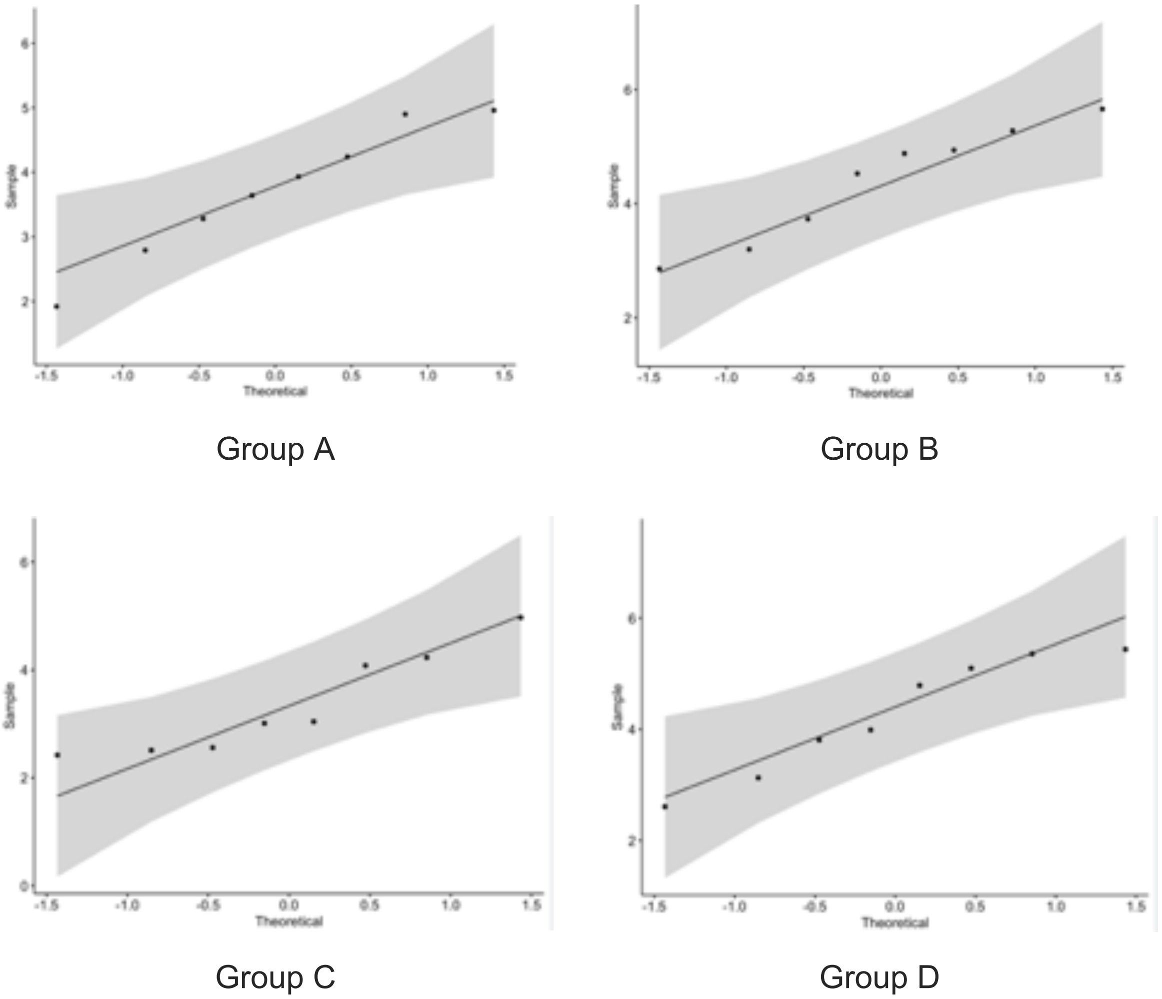


***Figure 1-3**. Boxplot for Cholesterol by 4 different medicine*

***Table 1-1**. The results of Shapiro test for Cholesterol*

Group	A	B	C	D
p-value	0.8059	0.5383	0.1682	0.4272



Group A

Group B

Group C

Group D

Figure 1-4. *Q-Q plot for Cholesterol by 4 group*

(B) Triglycerid

If you look at Figure 1-6 and Figure 1-7, the data appears to be not normally distributed for each medicine. The Shapiro test was operated to confirm the results clearly.

According to Table 1-2, since all p-values are under 5%, we can deny the null hypothesis of the Shapiro test and assure the responses of cholesterol are not normally distributed.

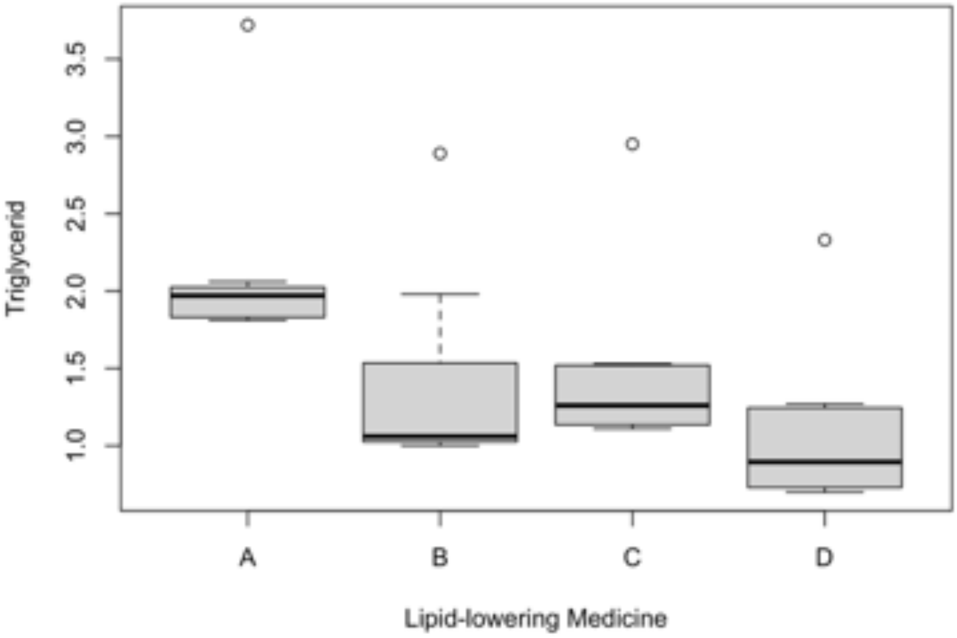
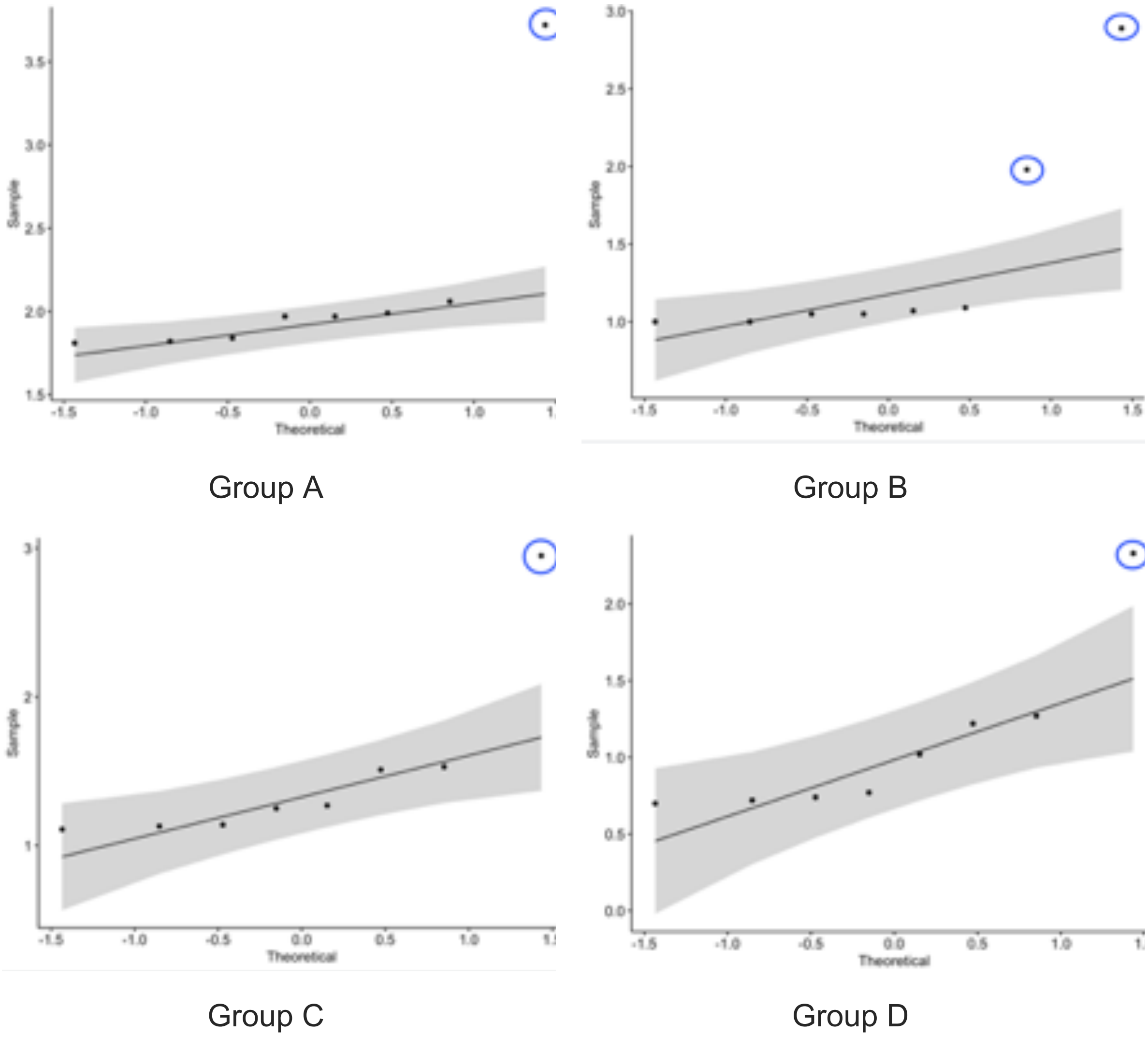


Figure 1-5. *Boxplot for Triglycerid by 4 different medicines*

Table 1-2*. The results of Shapiro test for Triglycerid*

Group	A	B	C	D
p-value	3.941e-05	0.0004943	0.0005561	0.008625



Group A

Group B

Group C

Group D

Figure 1-7. *Q-Q plot for Triglycerid by 4 groups*

(1.4) Choose proper test model for each response

(A) Cholesterol

Since cholesterol data is normally distributed and there are over 2 groups, we chose test models below:

(A.1) Linear regression
(A.2) ANOVA test

(B) Triglyceride

Triglyceride data is not normally distributed and we have over 2 groups:

(B.1) Fisher’s exact test
(B.2) Kruskal-wallis test

(1.5) Compare results

(A) Cholesterol

(A.1) Linear regression

- p-value : **0.1631**
- Coefficients

```
Estimate        Std.        Error    t value    Pr(>|t|)
(Intercept)     3.7075      0.3592   10.321 4.77e-11 ***
groupB           0.6775     0.5080    1.334        0.193
groupC         -0.3550      0.5080   -0.699        0.490
groupD           0.5713     0.5080    1.124        0.270
```

- R-squared : 0.1645

(B.2) ANOVA

- p-value : **0.163**

(B) Triglycerid

(B.1) Fisher's exact test

- p-value : **2.2e-16**

(B.2) Kruskal-Wallis rank sum test

- p-value : **0.006645**

(1.6) Conclusion

According to test results above, in case of Cholesterol, there is no significant difference between 4 different drugs, because all p-value over 5%. In the result of linear regression, any estimated coefficient has significant meaning. And also ANOVA results indicate the same result.

But in case of Triglyceride, there is a significant difference between 4 different drugs. While it cannot obtain the coefficient values from Fisher's exact test and Kruskal-Wallis rank sum test, it can be determined whether there is a difference.

Task 2

The dataset NHANES2 from the National Health and Nutrition Examination Survey with health related data from more than 10,000 individuals. In this task we will focus on the relationship between Total cholesterol (measured in different units than in Task 1) and weight. Try and use the data (nhanes2.csv) to study how Total cholesterol may be influenced by weight, and how that relationship is affected by sex, race and age.

Figure 2-1 shows the relationship between weight and total cholesterol. It is difficult to make any visual sense of this plot.

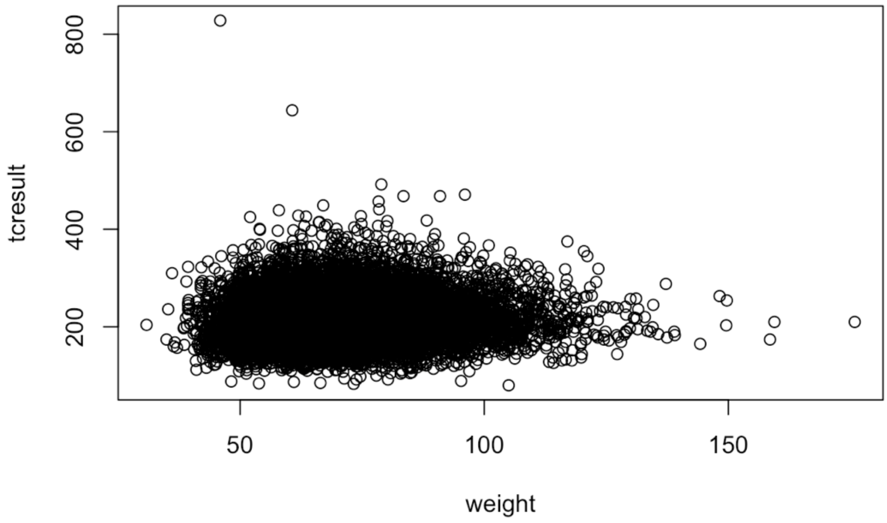


***Figure 2-1**: Relationship between weight and tcresult*

Because the sample size is so large (10 000) and is>30 we can assume normality through the central limit theorem and use parametric tests to test for correlation between the variables. The shapiro test was unsuccessful because of the large amount of sample data. Pearson's correlation test shows a p-value of 2.773e-09 which indicates that we can assume linear relationship. The relationship coefficient is 0.05838569 which indicates a slight increase in cholesterol with increasing weight.

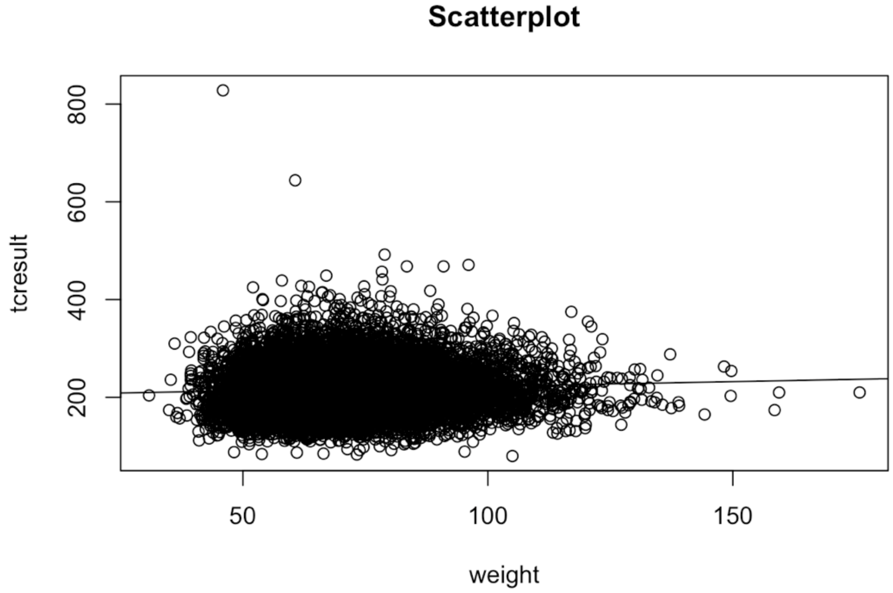


***Figure 2-2**: Linear regression model for weight and tcresult*

Next, we create a linear model of the data for cholesterol and weight. Since the data for cholesterol and weight both are continuous, a linear regression model is appropriate (Figure 2-2).

```
model1=lm(tcresult ~ weight, data=raw)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-143.89  -34.39   -5.05   29.14  615.21 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 204.16942    2.32023   88.00  < 2e-16 ***
weight        0.18777    0.03156    5.95 2.77e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.31 on 10349 degrees of freedom
Multiple R-squared:  0.003409,	Adjusted R-squared:  0.003313 
F-statistic:  35.4 on 1 and 10349 DF,  p-value: 2.773e-09
```

Figure 2-3: *Results model1*

The summary of the linear model shows a statistically significant relationship. Regression diagnostics that confirm linearity are shown Figure 2-4 below.

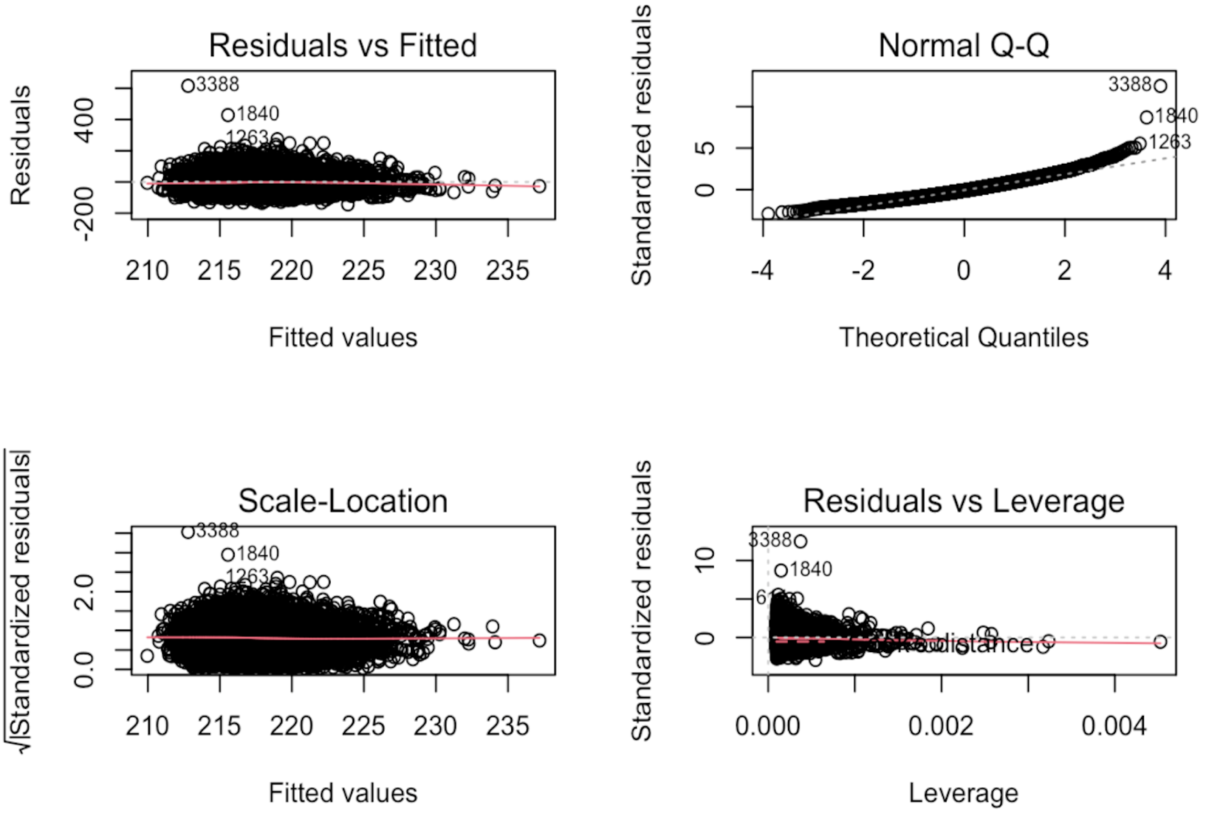
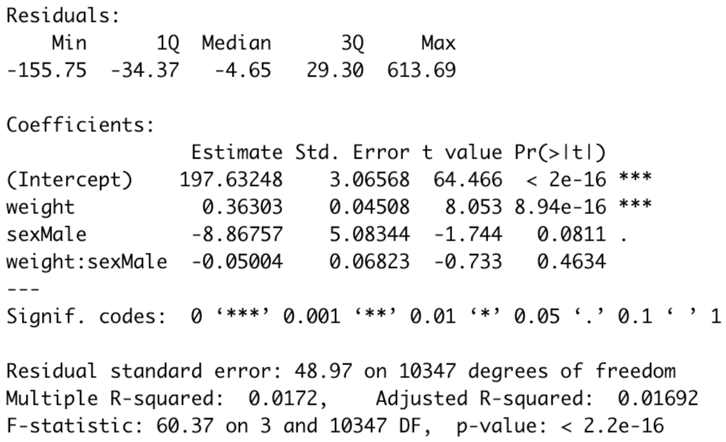


Figure 2-4: *Regression diagnostics of linear regression model for weight and tcresult*

Interaction cholesterol, weight and sex

We want to test if there is a correlation between the total cholesterol and weight depending on sex. The following model is created:

model2=lm(tcresult ~ weight+sex+weight:sex, data=raw)



```
Residuals:
    Min      1Q  Median      3Q     Max 
-155.75  -34.37   -4.65   29.30  613.69 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    197.63248    3.06568  64.466  < 2e-16 ***
weight           0.36303    0.04508   8.053 8.94e-16 ***
sexMale         -8.86757    5.08344  -1.744   0.0811 .  
weight:sexMale  -0.05004    0.06823  -0.733   0.4634    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.97 on 10347 degrees of freedom
Multiple R-squared:  0.0172,	Adjusted R-squared:  0.01692 
F-statistic: 60.37 on 3 and 10347 DF,  p-value: < 2.2e-16
```

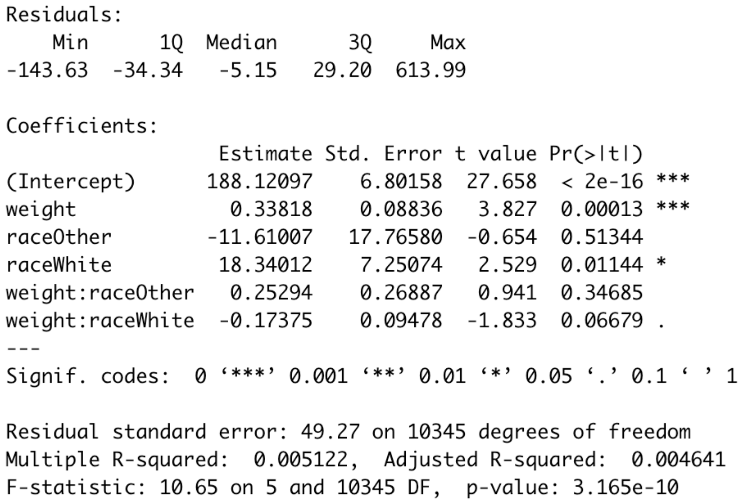
***Figure 2-5**: Results model2*

The summary shows no significant difference depending on the sex as none of the p-values are below alpha=0.05.

Interaction cholesterol, weight and race

We want to test if there is a correlation between the total cholesterol and weight depending on race. The following model is created:

model3=lm(tcresult ~ weight+race+weight:race, data=raw)



```
Residuals:
    Min      1Q  Median      3Q     Max 
-143.63  -34.34   -5.15   29.20  613.99 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      188.12097    6.80158  27.658  < 2e-16 ***
weight             0.33818    0.08836   3.827  0.00013 ***
raceOther        -11.61007   17.76580  -0.654  0.51344    
raceWhite         18.34012    7.25074   2.529  0.01144 *  
weight:raceOther   0.25294    0.26887   0.941  0.34685    
weight:raceWhite  -0.17375    0.09478  -1.833  0.06679 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.27 on 10345 degrees of freedom
Multiple R-squared:  0.005122,	Adjusted R-squared:  0.004641 
F-statistic: 10.65 on 5 and 10345 DF,  p-value: 3.165e-10
```

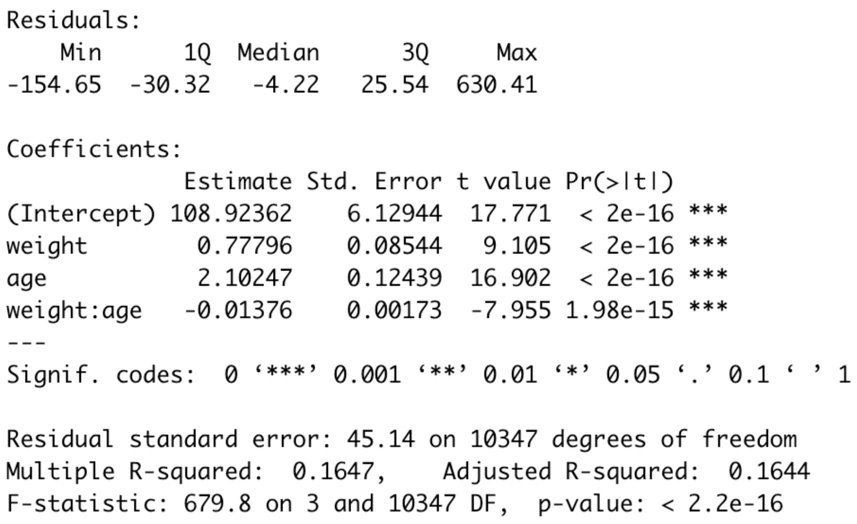
***Figure 2-6**: Results model3*

The summary of the model reveals a small significant difference of the relationship tcresult-weight and between the race black and race white(as the p-value is smaller than alpha=0.05), but not a significant difference with the relationship between cholesterol and weight.

Interaction cholesterol, weight and age

We want to test if there is a correlation between the total cholesterol and weight depending on age. The following model is created:

model4=lm(tcresult ~ weight+age+weight:age, data=raw)



```
Residuals:
    Min      1Q  Median      3Q     Max 
-154.65  -30.32   -4.22   25.54  630.41 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 108.92362    6.12944  17.771  < 2e-16 ***
weight        0.77796    0.08544   9.105  < 2e-16 ***
age           2.10247    0.12439  16.902  < 2e-16 ***
weight:age   -0.01376    0.00173  -7.955 1.98e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.14 on 10347 degrees of freedom
Multiple R-squared:  0.1647,	Adjusted R-squared:  0.1644 
F-statistic: 679.8 on 3 and 10347 DF,  p-value: < 2.2e-16
```

***Figure 2-7**: Results model4*

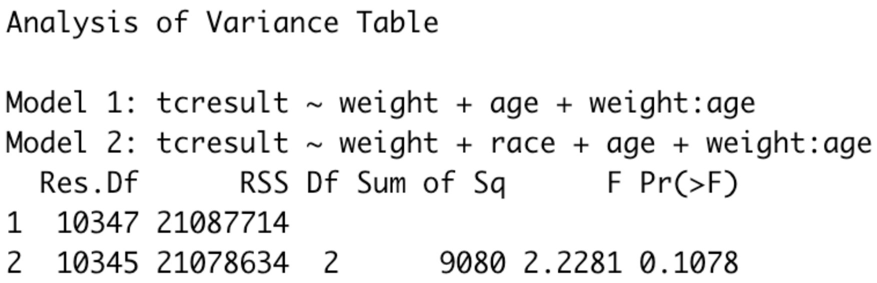
From the summary, we can see that age is a contributing factor, and also the interaction of weight and age. The estimate is negative, indicating that the effect that being overweight has on the total cholesterol levels decrease with age.

Testing the models

Since race showed indications of significance, we test if race should be included in the model. A fifth model was created to compare to model 4.

model5=lm(tcresult ~ weight+race+age+weight:age, data=raw)

The null hypothesis in the f-test is that the models are not significantly different, with an alternative hypothesis that the full model is significantly better. From an f- test (Figure 2-7) the result is that we cannot assume that the full model, including the variable race, is significantly better than the reduced one because the p-value is not lower than alpha=0.05. The race variable is therefore excluded.

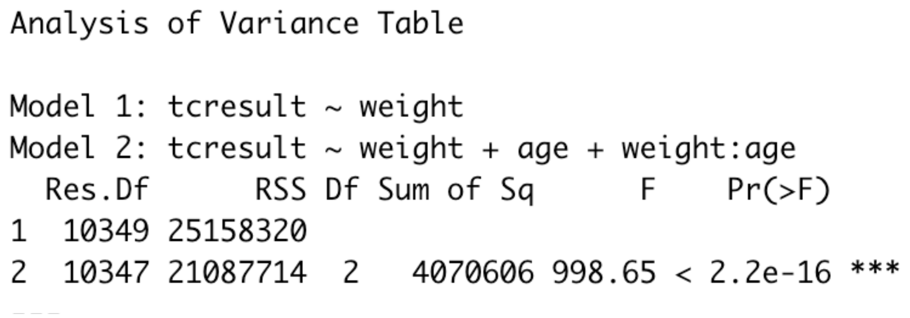


```
Analysis of Variance Table

Model 1: tcresult ~ weight + age + weight:age
Model 2: tcresult ~ weight + race + age + weight:age
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1  10347 21087714                            
2  10345 21078634  2      9080 2.2281 0.1078
```

***Figure 2-8**: Results f-test model4 and model5*

When testing model1 with just the relationship of total cholesterol and weight compared to model4 with included weight and age interaction we get a significantly low p-value which indicates that the full model including the interaction term weight:age is significantly better than the one with only cholesterol and weight.



```
Analysis of Variance Table

Model 1: tcresult ~ weight
Model 2: tcresult ~ weight + age + weight:age
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1  10349 25158320
2  10347 21087714  2   4070606 998.65 < 2.2e-16 ***
```

Figure 2-9: *Results f-test model1 and model4*

The proposed model is therefore model4= lm(tcresult ~ weight+age+weight:age) as it also produces better regression diagnostic plots than model1 (Figure 2-10).

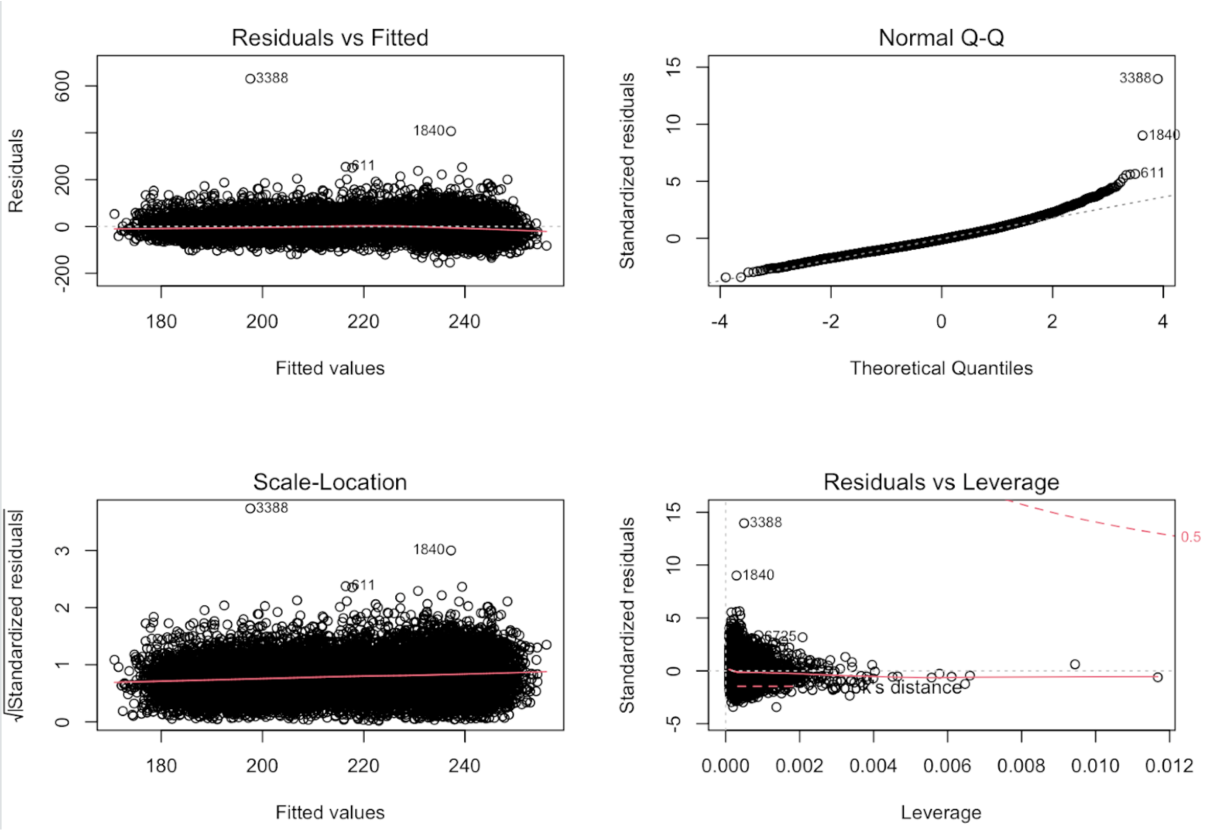


Figure 2-10: *Regression diagnostics model4*

Task 3

Data on the number of Covid-19 cases and fatalities in Sweden and UK until Sept 4 from the European Centre for Disease control ([https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide (Länkar till en externa sida.)]) are found in the file [coviddata.xls.]

(3.1) Data analysis

The given data reports the pandemic development in Sweden and in the UK. It shows both the cases and deaths reported on each day since the 31st of December 2019. The raw data is shown in Figure 3-1.

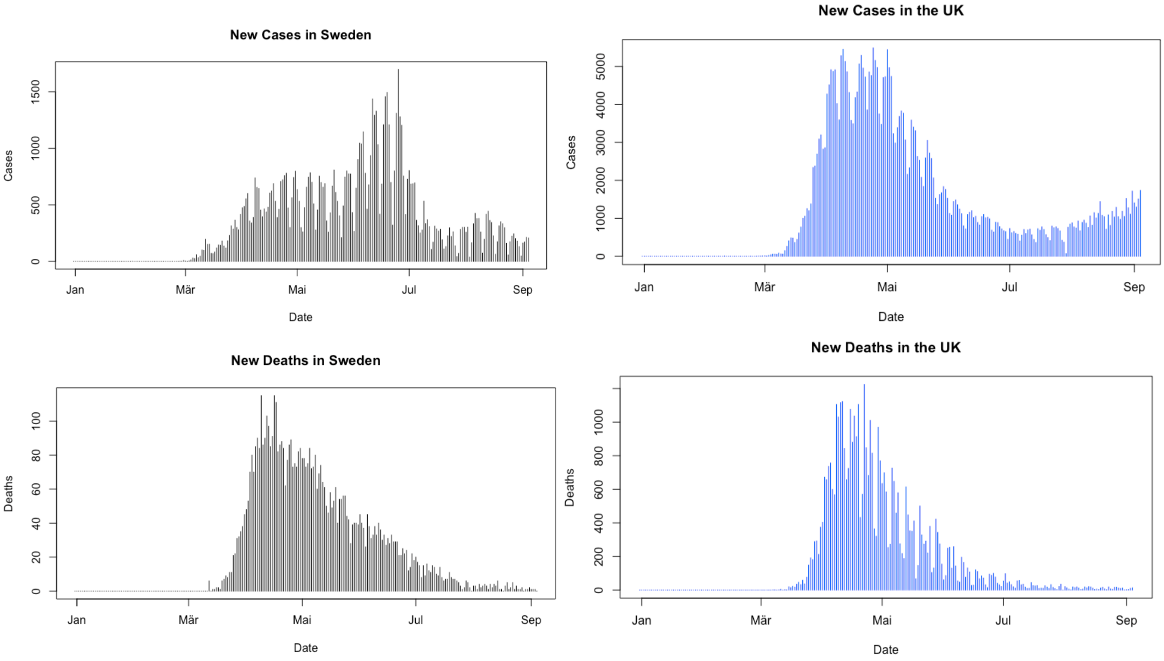


***Figure 3-1**: Raw data*

As can be seen in Figure 3-1, the reported data is very unregular since not all cases and deaths are reported during the weekends. For that reason, instead of the place data set *cases,* a moving average value *av_cases*$_i$ is used, calcualted on each day *i* as:

$$av\_cases_i = \frac{\sum_{n=i-13}^{i} cases(n)}{14}$$

Same applies for the moving average of the deaths *av_deaths*$_i$.Futhermore, the data is scaled to 100,000 inhabitants to make better comparison between both countries. The scalted average data is shown in Figure 3-2.

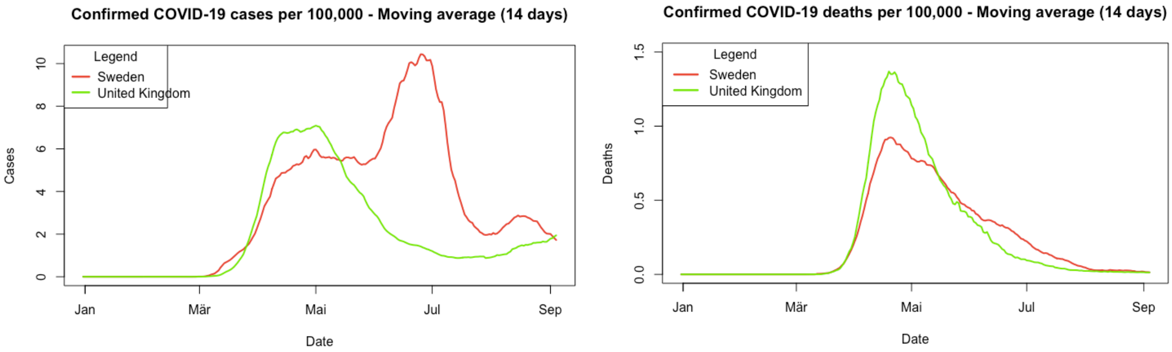


Figure 3-2: *Scaled average data*

(3.2) Try and construct a statistical model describing the pandemic development in Sweden and UK.

Reported Cases

While the curve of the deaths is very regular, the curves for the number of cases is very irregular. It has to be noted that the number of reported cases is affected enormously by the number of tests. For this reason, the number of reported cases can be much smaller than the number of infected people. For this reason, it is not useful to create a model with the number of reported cases.

Reported Deaths

On the other hand, the number of reported deaths can be considered as more reliable. It can be seen that after a first increase the number of deaths has slowly fallen again. Finding a model describing this data is possible, for example a logit() function of the cumulative deaths. However, it is not useful to use such a model for predicting a second wave of the pandemic.

Cases vs. Deaths

Another option is to investigate the correlation of reported cases and deaths. This is done separatly for both countries.

(1) UK

A linear regression is used to see if there is a linear correlation between both variables (cases and deaths). The modeled line as well as the confidence and prediction intervals are shown in Figure 3-3. The R^2 for the model is 0.926, so the linear regression con be considered as good. This means, there is a linear relationship between cases and deaths.

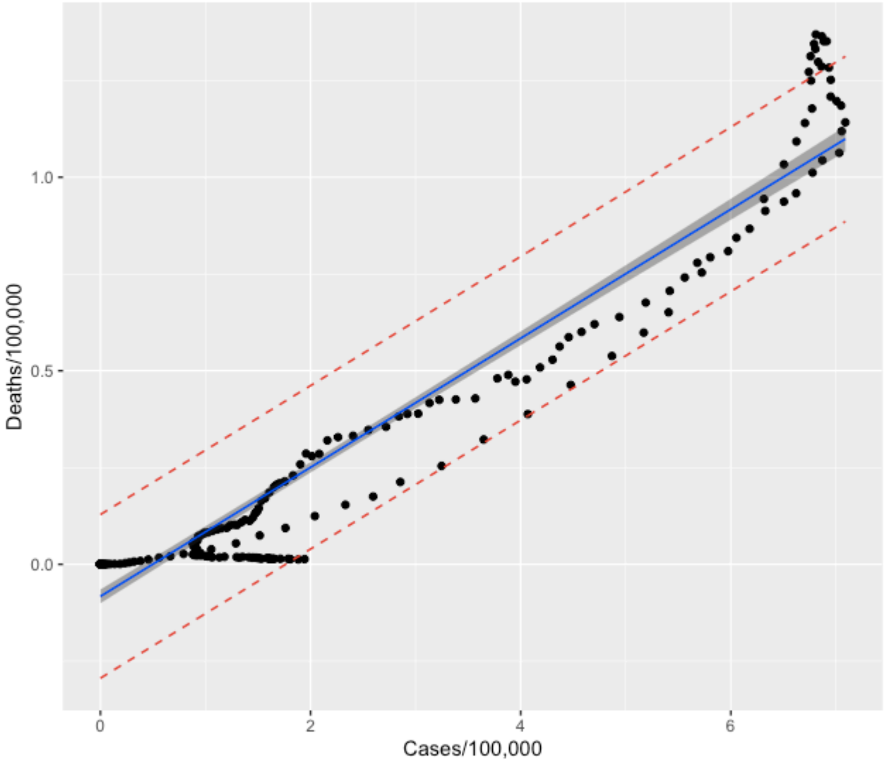


Figure 3-3: *Linear regression of COVID-19 cases and deaths in the UK; Blue line: linear regression; grey area: 95%-confidence interval; Red lines: 95%-prediction interval*

(2) Sweden

The same linear model can used to investigate the cases-deaths-relationship in Sweden. However, applied on all data the model is not useful (Figure 3-4). The R^2 is 0.3962, which confirms the observation of the plot.

There can be several different explanations, why the relationship between cases and deaths is so irregular in Sweden. One possible explanation is the different test rate in different phases of the COVID-19 pandemic. For that reason, the relationship is studied in the first 150 days of the pandemic (Figure 3-5a) and in the last 70 days (Figure 3-5b). As can be seen in the figures, there is a linear relationship between cases and deaths in different phases of the corona. The R^2 is 0.9387 for the model for the first 150 days and 0.8543 for the last 70 days.

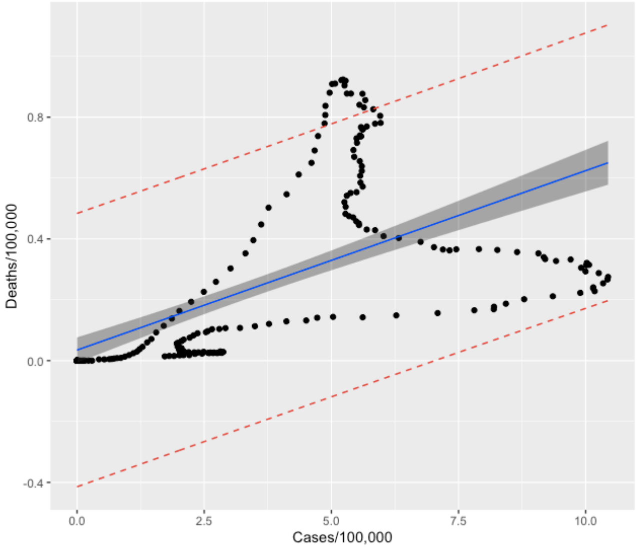
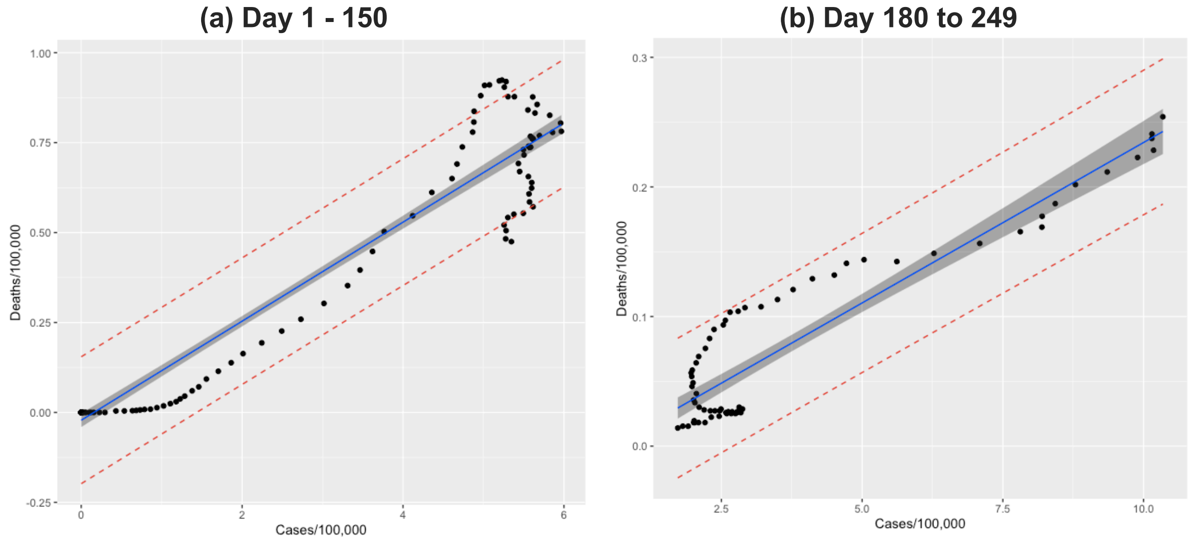


***Figure 3-4:** Linear regression of COVID-19 cases and deaths in Sweden; Blue line: linear regression; grey area: 95%-confidence interval; Red lines: 95%-prediction interval*


(a) Day 1 - 150

(b) Day 180 to 249

***Figure 3-5:** Linear regression of COVID-19 cases and deaths in Sweden for different phases of the pandemic; Blue line: linear regression; grey area: 95%-confidence interval; Red lines: 95%-prediction interval*

(3.3) Discuss the possibility of using this model for making comparisons between the two countries of the future course of the disease.

Figure 3-6 shows the comparison of the four linear models. It can be seen, that the fatality rate (based on the reported cases) in the UK is much higher, even higher than in the first phase of the pandemic in Sweden. In Sweden, the fatality rate decreased significantly.

However, if we assume that the corona virus hasn't mutated much over the time and between both counties, the real fatality rate must be the same. The model therefore provides more information whether really all cases have been reported, which was most likely not the case in the first phase of the pandemic due to law test capacities in both countries.

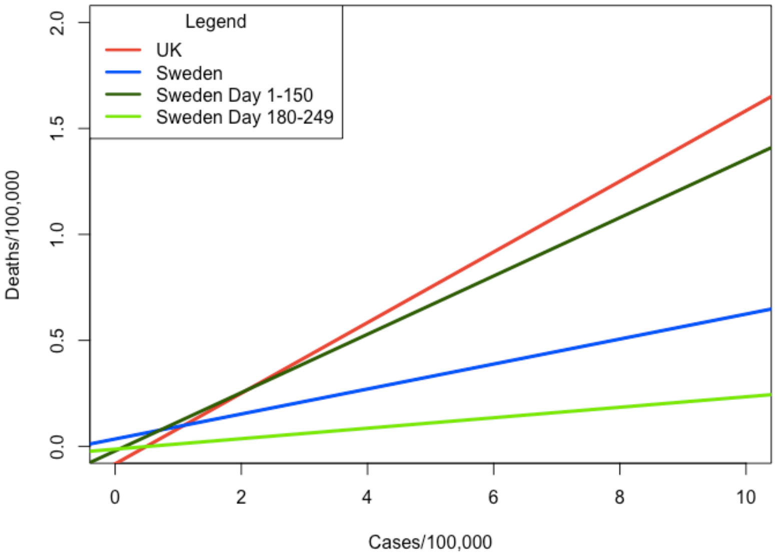


***Figure 3-6:** Comparison of all four linear regression models*

The model can be used to make a comparison between both countries. However, there are several factors which are not considered by just comparing the reported data. This is for example, the test rate. Without knowing the difference between the reported case number and the real case number, it is difficult to really compare the data.

Moreover, the model can't be used to predict the development of COVID-19 in the future. As the example of Sweden has shown, the relationship of cases and deaths is influenced by a lot of background factors and therefore changes of the time and in different phases of the pandemic.

Furthermore, it is very difficult to predict the future of a pandemic in general.

Task 4

In X-ray imaging, the choice of kV to the X-ray tube is critical for obtaining good image quality. In order to maximise the image quality, a medical engineer obtained images of a phantom at a number of different kV settings, and then measured the contrast-to-noise ratio (CNR) between the phantom and the background, a central parameter for quantitative studies of image quality. The data are given in kvdata.csv.
Study the relationship between kV and CNR, and try to use this description to find an optimal value of kV. If possible, try and obtain both a point estimate of the optimal kV and an interval estimate or some uncertainty measure of the point estimate.

The data given was that of kV settings and the corresponding Contrast-to-Noise (CNR) ratios. What we are looking for is an estimate of which setting that produces the best CNR value. The higher the CNR value, the better the result. Thus, we are looking for the kV value which gives the highest CNR ratio.

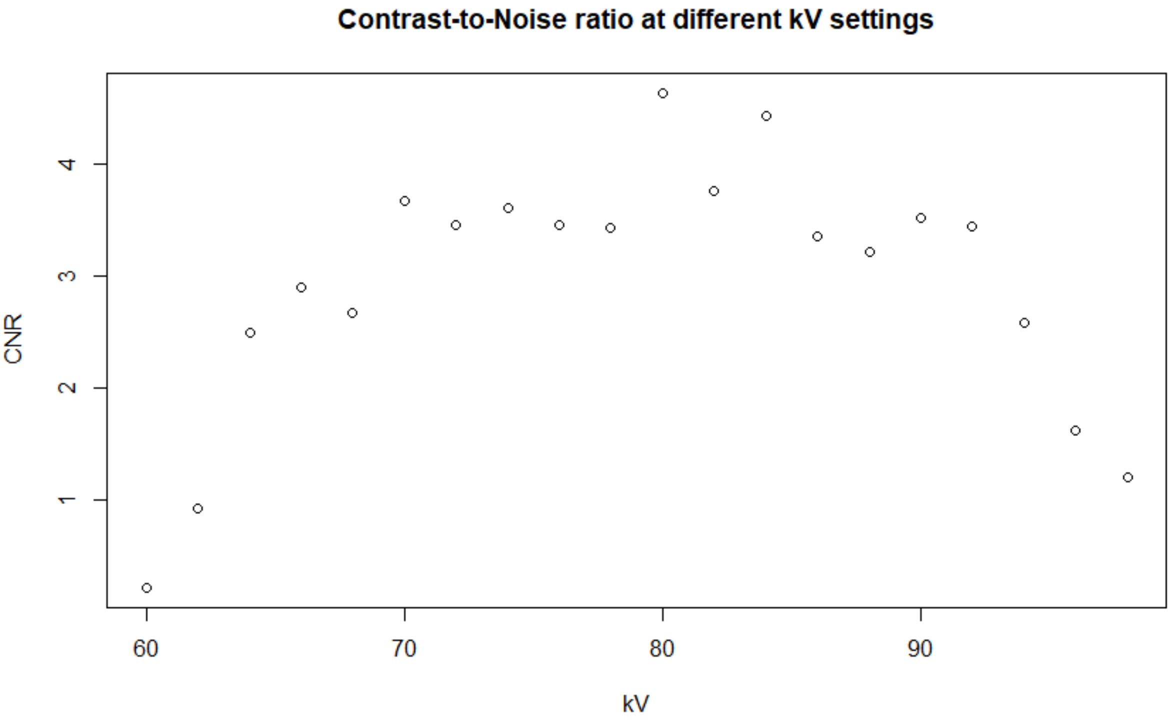


Figure 4-1: *Contrast-to-Noise ratio at different kV settings*

Reading the data using R we get the plot which can be seen above. Just by looking at the values we can see the kV setting around 70-90 kV produce high results.

The data seems to fit to a second-order polynomial. Thus, we moved on by using a polynomial fit to model the data. Below, one can observe the polynomial fit to the data plotted (Figure 4-2). The R-squared of the model is 0.8679.

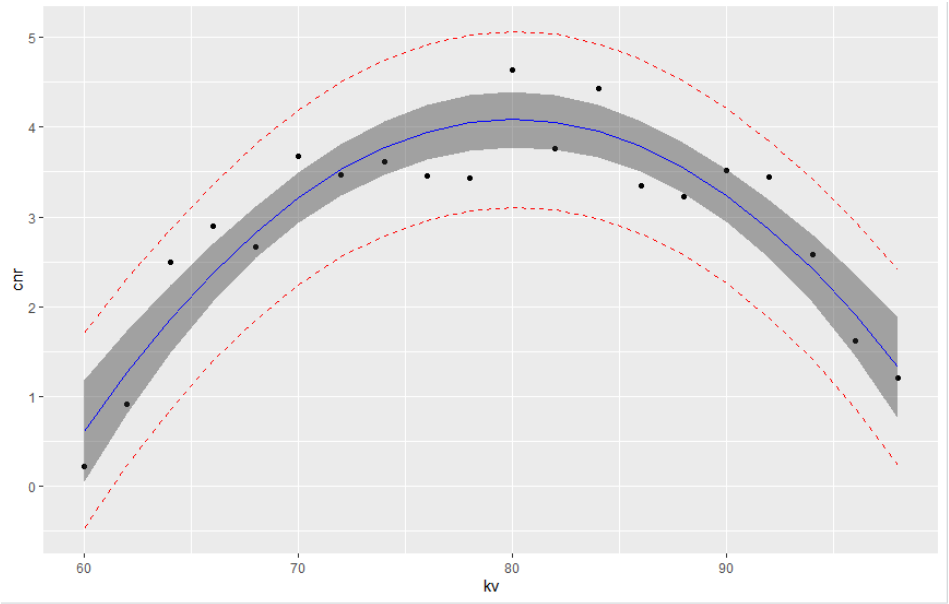


Figure 4-2: *Second-order polynomial fit for contrast-to-Noise ratio and kV settings*

In the plot above one can observe a blue line which is the second-order polynomial fit. The original data can be observed as black dots in the plot. The grey area is the 95 % confidence interval of the polynomial model as a whole, whilst the area between the lower and upper red dotted line is the 95 % prediction interval of certain individual values. Based on the coefficients of the mean curve

cnr = **-0.008604***(kV^2)+**1.378***kV**-51.095**

we could calculate an estimation of the optimal kV value:

cnr' = -0.008604*2*kV+1.378
=> cnr'= 0
=> kV = 80.079

Also, the same cnr value can be produced at the upper limit of the confidence interval between 75 and 87 kV.