Compulsory assignment number 1 spring 2020 STK4900/9900: Statistical methods and applications

Victoria Jensen

Matematisk institutt

UNIVERSITETET I OSLO

Mars 2020

Exercise 1: Air Pollution Study

This is the study how air pollution at a measuring station at Alnabru in Oslo is related to explanatory variables such as traffic volume and meteorological conditions in the same place. There are 500 observations in the data set collected by the Norwegian Public Roads Administration (Vegvesenet). Air pollution is measured by the concentration of NO2 particles.

log.no2	The logarithm of the concentration of NO2
log.cars	The logarithm of the number of cars per hour
temp	Temperature 2 meters above the ground (degrees C)
wind.speed	Wind speed (meters/second)
hour.of.day	Hour of the day the measurements were collected (1-24)

a) Report the main features of the variables log.no2 and log.cars by numerical summaries and plots. Make a scatterplot with log.cars on the x-axis and log.no2 on the y-axis. What do you see?

```
rm(list=ls(all=TRUE))

# Obligatory exercise 1

# Reading datafiles

no2data <-
read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v20/mandatory/no2.txt",sep="
\t",header=TRUE)

# 1a)

# Catogorizing the data

names(no2data)=c("log.no2","log.cars","temp","wind.speed","hour.of.day")

# Summary of the pollution levels and number of cars

summary(no2data)
```

A summary from R shows that the value for the log measuring of cars is much higher than log for concentration of NO2. The more traffic - the more air pollution. Cars produce a certain amount of concentration of NO2 and this concentration can be less than one unit of

concentration of NO2. So, it seems to be true. However, we do not have information about how much concentration is used for NO2. The log measurement shows that parts per million have being used.

```
      log.no2         log.cars
 Min.   :1.224   Min.   :4.127
 1st Qu.:3.214   1st Qu.:6.176
 Median :3.848   Median :7.425
 Mean   :3.698   Mean   :6.973
 3rd Qu.:4.217   3rd Qu.:7.793
 Max.   :6.395   Max.   :8.349
```

```
> summary(no2data$log.cars)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.127   6.176   7.425   6.973   7.793   8.349
```

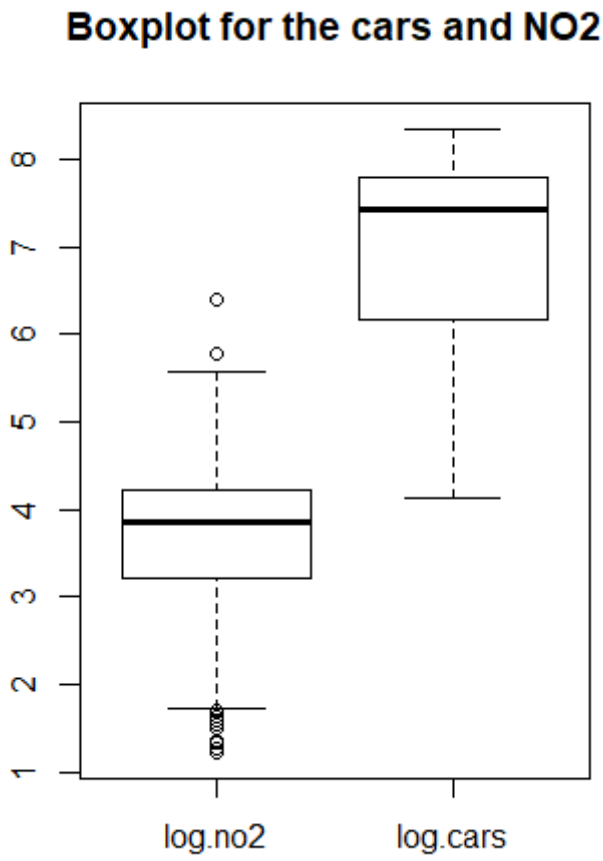
The figure 1 shows the boxplot for the cars and NO2. The numerical values can also be shown via boxplots. A boxplot is a standardized way of displaying the distribution of data based on a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell a lot about outliers and their values. It can also tell if the data is symmetrical, how tightly the data is grouped, and if and how the data is skewed. As a result, boxplot gives better visualization of the summary results and gives a good indication of how the values in the data are spread out. The line inside of the boxplot is the measure for the median of the data. The 1st and 3nd quartiles are represented as the upper and lower ends of the boxes respectively.

The boxplot shows the differences across the units of cars and NO2 concentration and multiple outliers for variable NO2. The data for log cars is positively skewed. The data for NO2 is symmetrical. In addition, we see that dataset for log cars has a wider distribution compared to the data for NO2. This can be the main reason why the mean for logcars (6.973) is much higher than the mean for NO2 (3.698). The median line of the boxplot with the data for NO2 lies outside of the boxplot with the data for car. This indicates that there is likely to be a difference between the two groups.

By comparing of the interquartile ranges or the boxes lengths we can examine how the data is dispersed between each sample. Thus, the data for cars are more dispersed compared to the data for NO2.

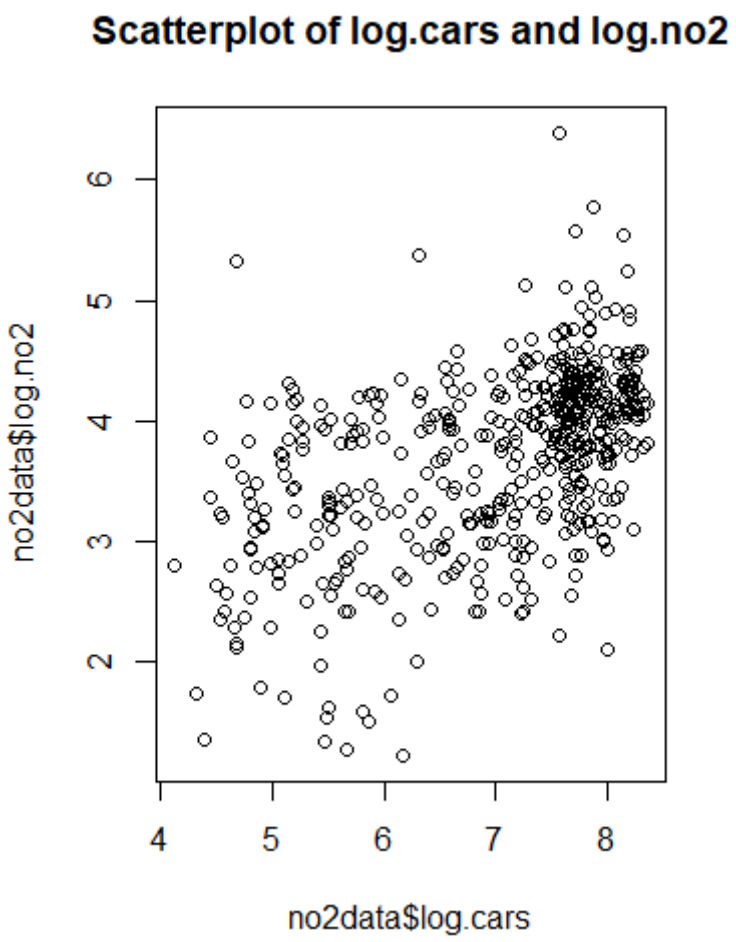
```
# Boxplot of log.no2 and log.cars
boxplot(no2data$log.no2, no2data$log.cars, names = c("log.no2", "log.cars"), main = "Boxplot of car")


# Scatterplot of log.cars and log.no2
plot(no2data$log.cars, no2data$log.no2, main = "Scatterplot of log.cars and log.no2")
```

Figure 1 Boxplot of cars



The scatterplot at the figure 2 confirms the information that we have seen by looking at the boxplot. The data are right skewed for the cars. The scatterplot shows the increase in air pollution (NO2) with the increase the number of cars. We also see that the scatterpoints are concentrated according to both variables median, as seen in table 1. The shape of the scatter distribution also suggests a correlation between the two variables.

Figure 2. Scatterplot. Air pollution and number of cars per hour.



b) Fit a simple linear model where the log concentration of NO2 is explained by the amount of traffic, measured by log (number of cars per hour). Give an interpretation of the estimated coefficients and construct a plot with the observations and the fitted line. Explain what the R2 measure tells you.

```
# 1b)
# Linear fit
fit=lm(no2data$log.no2~no2data$log.cars)
summary(fit)
```

Now we will fit a simple linear model where the log concentration of NO2 is explained by the amount of traffic, measured by log (number of cars per hour). The model for linear regression has the following view:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$

where $y_i = y_1,\ldots,y_n$ is the response variable

x1,...,xn are the explained variable

εi - random noise, β0 is the intercept, β1 determines the slope of the regression

Here is the output from R.

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.18822 -0.40071  0.06428  0.40362  2.48472

Coefficients:Estimate       Std. Error t value     Pr(>|t|)
(Intercept)       1.23310    0.18755   6.575      1.23e-10 ***
no2data$log.cars  0.35353    0.02657  13.303      < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6454 on 498 degrees of freedom
Multiple R-squared:  0.2622,       Adjusted R-squared:  0.2607
F-statistic:   177 on 1 and 498 DF,  p-value: < 2.2e-16
```

R-squared (R2) is a statistical measure that indicate the proportion of the variance for a dependent variable explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, R2 give a measurement how the fitted line fits the data. The R2 for our model is 0,2622 that indicate that only 26 % of the data variance for air pollution (logNO2) is predicted by the amount of cars (log.cars). The measurement for R2 confirms the information from the boxplot and scatterplot. The estimated coefficient shows that without influence of any cars, the air pollution or the concentration of the NO2 is 1,23. However for each additional unit of cars the concentration of NO2 increases with 0,354. The p value is very low (< 2e-16 ***) that indicates that the coefficient is statistically significant for our model. Then the null-hypothesis can be rejected that means that the amount of cars affects concentration of NO2.

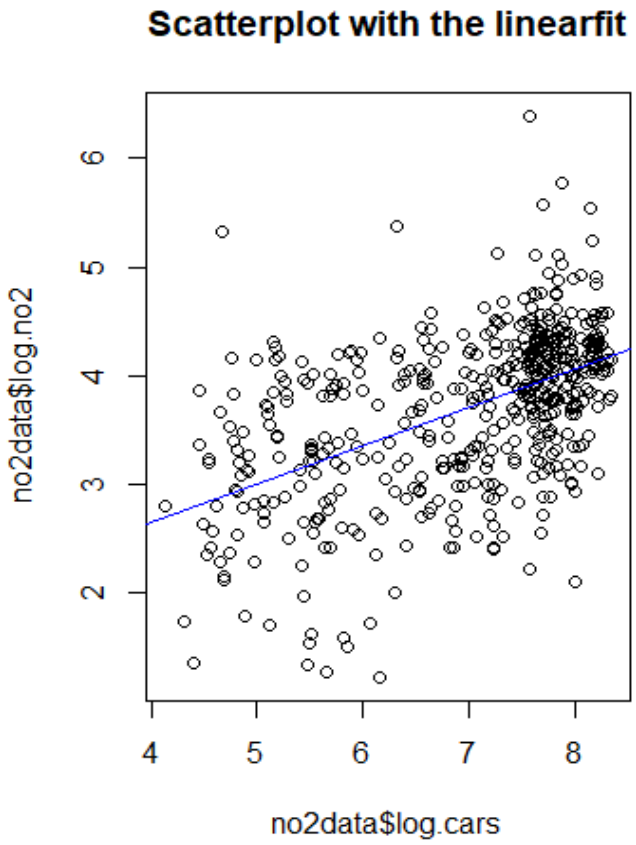
```
# Correlation between no2data$log.cars,no2data$log.no2
cor(no2data$log.cars,no2data$log.no2)**2
```

Calculated R2 = 0.2621956 shows that there is a correlation between variables responsible for air pollution and number of cars.

```
# Scatterplot with the linearfit
plot(no2data$log.cars, no2data$log.no2, main = "Scatterplot with the linearfit")
abline(fit, col = "blue")
```

Figure 3 shows the constructed plot with the observations and the fitted line.

Figure 3. Scatterplot with the linear fit.



c) Check various residual plots to judge if the model assumptions for the model in b) are reasonable.

The linear regression model has 4 assumptions.

1. Linearity:

$\eta i = \beta 0 + \beta 1 x1i + \beta 2 x2i + ... + \beta pxpi$. Where ηi is the systematic part of the simple linear regression model $yi = \eta i + \varepsilon i$,

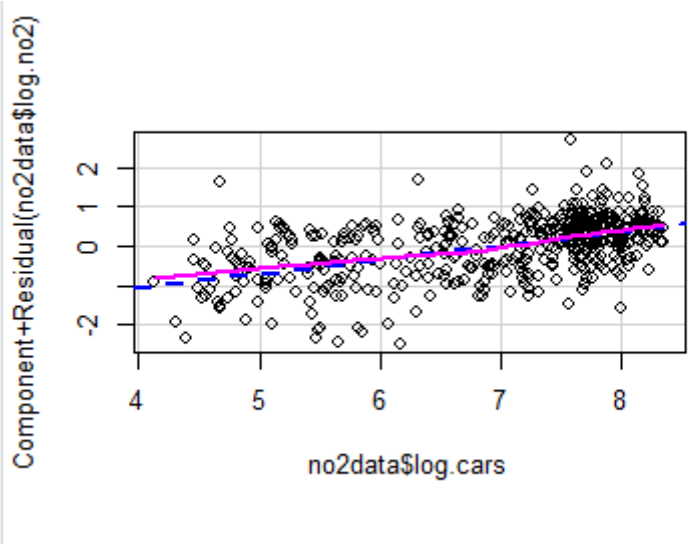
2. Homoscedasticity or constant variance of the error term: Var(εi) = σ2 ε ∀ i.

3. Error should be normally distributed: εi ~ N0,σ2

4. Errors should not correlated with each other : Cov(εi,εj) = 0 ∀ i 6= j.

```
# 1c)
# Linearity check
library(car)
crPlots(fit, terms=~no2data$log.cars)
```

Figure 4. Linearity check.



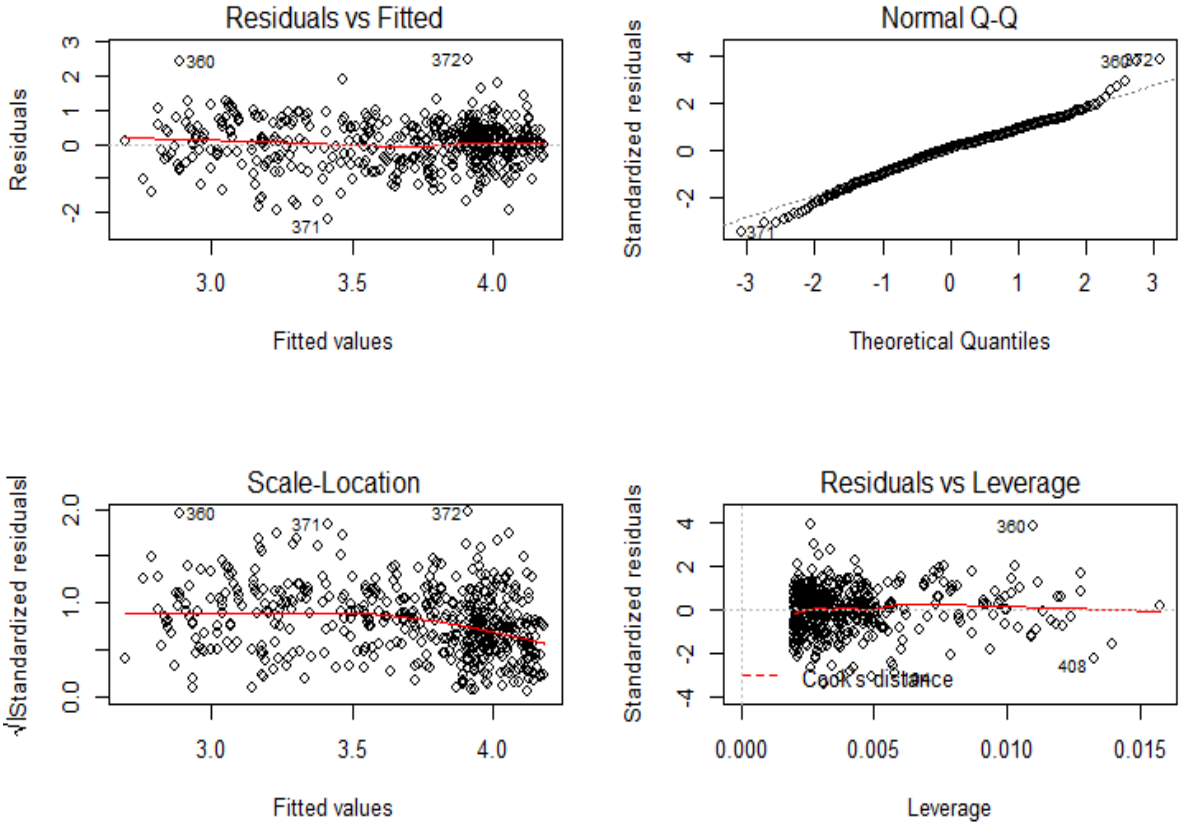
The plot shows that the the lineary distributed by the linear function.

```
# Checking for constant variance
par(mfrow = c(2,2))
plot(fit)
```

The residuals vs fitted plot at top left shows that there are some outliers (values are 360, 371 and 372) that somewhat fits with the plot in figure 1. The QQ plot indicates that the data is not skewed, From the Q-Q plot we see that the sample quantiles mostly follow a straight line, meaning that they are mostly normal. The scale-location plot shows that we have

homoscedasticity. The residuals vs leverage plot at lower right shows that there are not any influential outliers outside of the Cook distance.

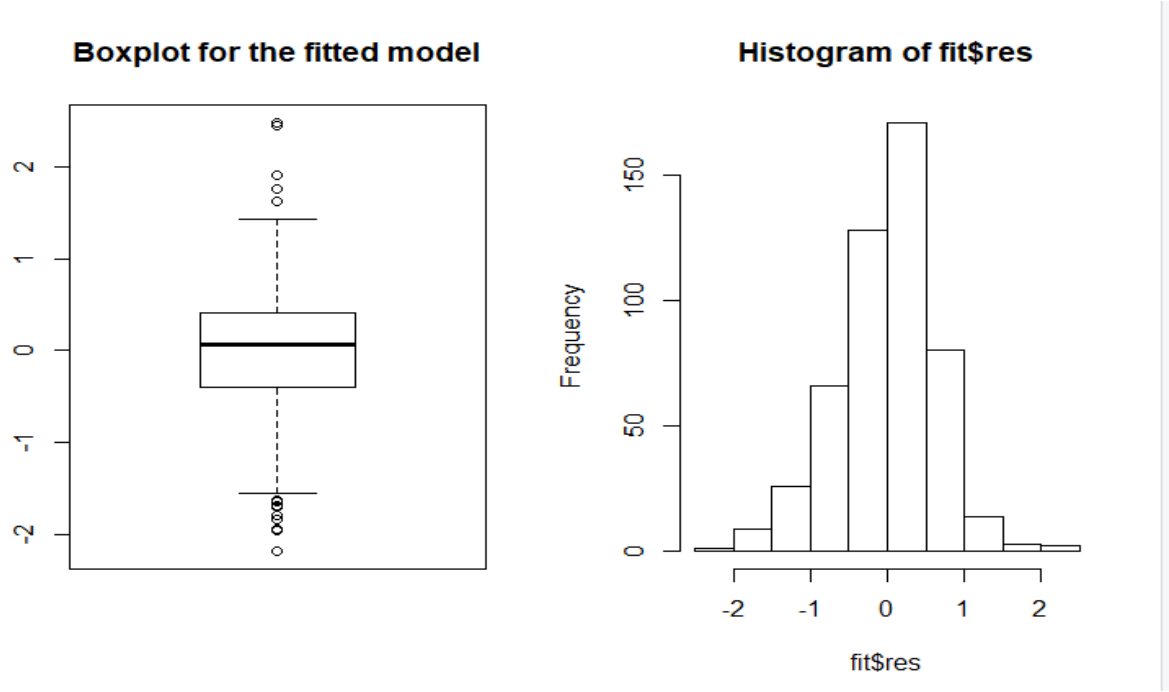
Figure 5. Comparison of the various residual plots to judge if the model assumptions for the model in b) are reasonable.




```
# Checking for normality

par(mfrow = c(1,2))

hist(fit$res)

boxplot(fit$res, main = "Boxplot for the fitted model")
```

Histogram for fitted residuals and boxplot for the fitted model also shows the normal distribution. For the fitted model we can see that there are some outliers, but the model is not skewed, and the data is symmetrically distributed.

Figure 6. Checking homoscedasticity of errors and normality

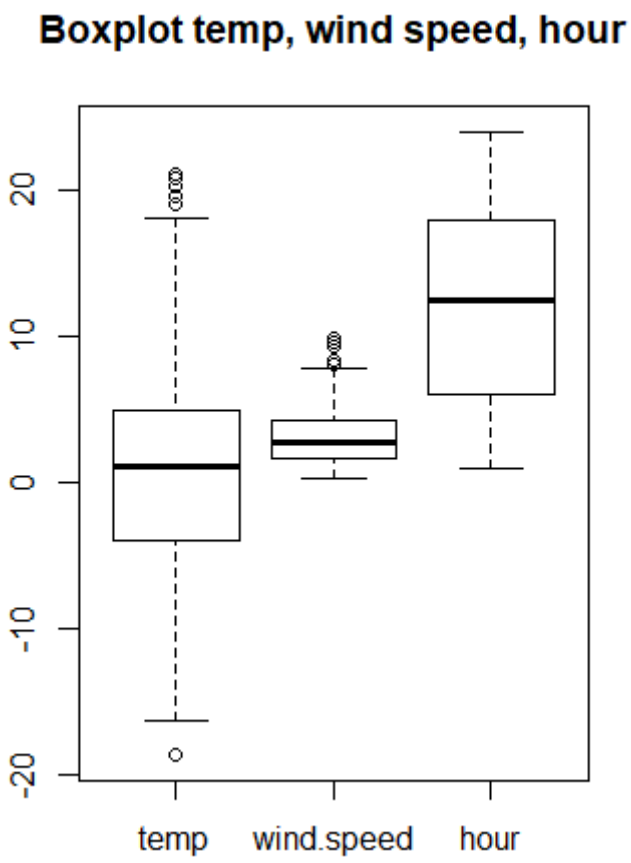


d) Then use multiple regression to study the simultaneous effect of the various covariates on the log concentration of NO2. Use an appropriate measure in order to find the 'best' model for prediction of NO2 concentration at Alnabru. You need not include interactions but check if you should transform some of the variables in addition to the number of cars, which is already log-transformed.

```
# Fitting the model of the multiple linear regression

Multireg = lm(no2data$log.no2~no2data$log.cars+no2data$temp+no2data$wind.speed + no2data$hour.of.day)

summary(Multireg)
```

The summary indicates that the p-values are quite low (>0.05), but I would like to check if I need to log transform the data before I decide which variables I need for the final model. The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics. The boxplot constructed for the temp, wind.speed and hours of day variables shows that only the data responsible for the wind.speed are skewed. So, I will log-transform only the wind.speed variable.

Figure 7. Boxplot for variables temp, wind speed and hour of day.



But first let look at the output from the fitting the model of the multiple linear regression without log transformation of the temp and wind.speed.

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.24876 -0.32070  0.03084  0.33860  1.96057

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.152131   0.175045   6.582 1.19e-10 ***
no2data$log.cars    0.456974   0.028411  16.084  < 2e-16 ***
no2data$temp       -0.026855   0.003905  -6.877 1.85e-11 ***
no2data$wind.speed -0.149334   0.014076 -10.609  < 2e-16 ***
no2data$hour.of.day -0.013025   0.004452  -2.926   0.0036 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5508 on 495 degrees of freedom
Multiple R-squared:  0.4658,   Adjusted R-squared:  0.4615
F-statistic: 107.9 on 4 and 495 DF,  p-value: < 2.2e-16
```

Multiple regression model is used to study how various predictors affects the response variable. First it needs to check if there is a correlation between predictors to exclude highly correlated

ones from the multiple regression model to avoid increase of the standard errors. The results shows that covariates log.cars and hour.per.day are highly correlated with the coefficient correlation of 0.5768567. A multiple linear regression model includes all predictors results. From the output we observe that the R2 has increased significantly and equal 0.4658. So, I can try to run the model without the hour.per.day variable.

Then I will try to run the model without hours of day variable and log transformed wind.speed variable.

```
newwind.speed=log(no2data$wind.speed)

# Fitting of the multiple linear regression model without hours.per.day as it is closely correlated to log.cars

fit.multiregfinal = lm(no2data$log.no2~no2data$log.cars+newwind.speed+no2data$temp)

summary(fit.multiregfinal)
```

The R2 from the multiregfinal model is equal 0,475 that is higher than the previous model that indicates that 47,5 % of the data variance for air pollution (logNO2) is predicted by the amount of cars (log.cars), wind speed and the temperature. Thus, the model is good enough without the variable responsible for the hours of day.

```
Call:
lm(formula = no2data$log.no2 ~ no2data$log.cars + newwind.speed +
    no2data$temp)

Residuals:
     Min       1Q   Median       3Q      Max
-2.07759 -0.33892  0.05458  0.36666  1.83266

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.229009   0.162586   7.559 1.98e-13 ***
no2data$log.cars  0.411979   0.022995  17.916  < 2e-16 ***
newwind.speed    -0.414496   0.036572 -11.334  < 2e-16 ***
no2data$temp     -0.026304   0.003861  -6.813 2.79e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5455 on 496 degrees of freedom
Multiple R-squared:  0.475,    Adjusted R-squared:  0.4718
F-statistic: 149.6 on 3 and 496 DF,  p-value: < 2.2e-16
```

We also observe that the effect of variable log.cars has gone down slightly from 0.456974 to 0.411979. The effect of variable wind.speed has increased from -0.149334 to -0.414496, and effect of temperature has gone also slightly down from -0.026855 to -0.026304. It indicates that

the effect from the variable “hour of the day” has been transferred in the new model into the effect of wind.speed. The correlation coefficients show this magnitude.

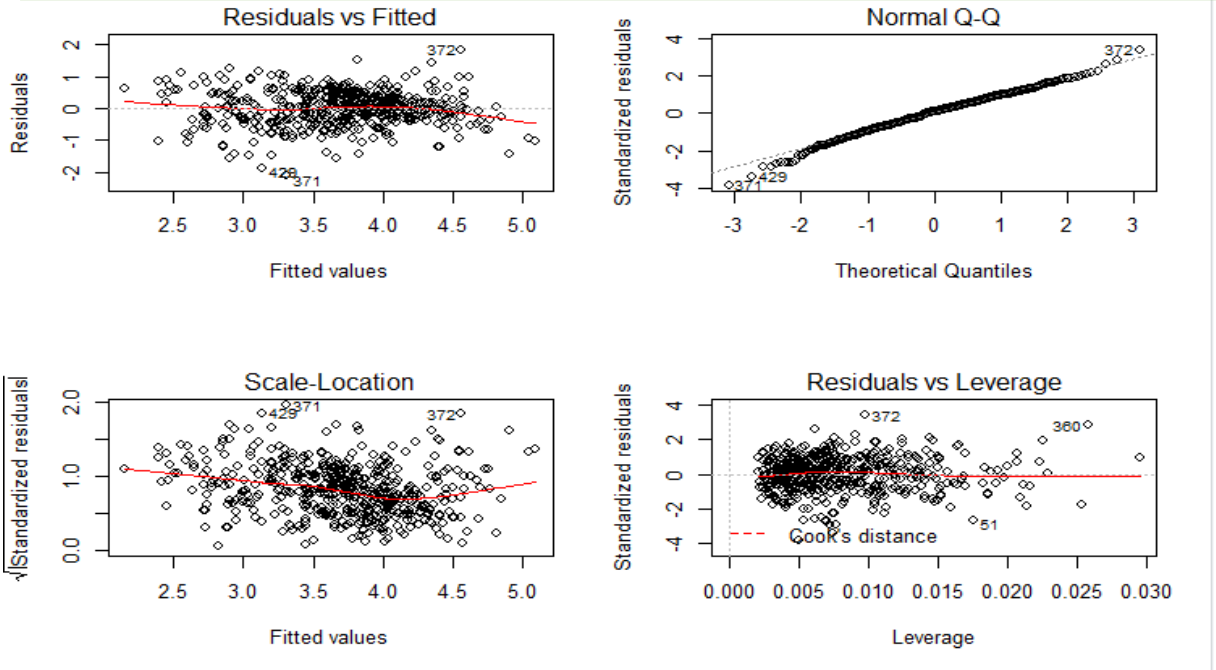
```
> cor(newwind.speed,no2data$hour.of.day)
[1] 0.01550807
> cor(no2data$log.no2,no2data$hour.of.day)
[1] 0.2462013
> cor(no2data$temp,no2data$hour.of.day)
[1] 0.079485
```

e) For the model you have chosen in d), write an interpretation of the model coefficients and check if the model assumptions seem reasonable through various plots. Remember to comment all plots you include in your report.

The final model is the one without the predictor hour.of.day. A CPR plot of the model is seen in figure 8 where we observe a good linearity in all the covariates.

Figure 8. Comparison of the various residual plots from the multiregfinal model without hours of day variable

```
#1e)Summary of the final model
crPlots(fit.multiregfinal, terms=~no2data$log.cars+no2data$temp+newwind.speed)
par(mfrow = c(2, 2))
plot(fit.multiregfinal)
```



```
Call:
lm(formula = no2data$log.no2 ~ no2data$log.cars + newwind.speed +
    no2data$temp)

Residuals:
     Min       1Q   Median       3Q      Max
-2.07759 -0.33892  0.05458  0.36666  1.83266

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.229009   0.162586   7.559 1.98e-13 ***
no2data$log.cars  0.411979   0.022995  17.916  < 2e-16 ***
newwind.speed    -0.414496   0.036572 -11.334  < 2e-16 ***
no2data$temp     -0.026304   0.003861  -6.813 2.79e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5455 on 496 degrees of freedom
Multiple R-squared:  0.475,    Adjusted R-squared:  0.4718
F-statistic: 149.6 on 3 and 496 DF,  p-value: < 2.2e-16
```

By interpreting the coefficients of the final model we see that for each unit of log.cars added the log concentration of NO2 goes up by 0.4119 units.
For each added units of temperature the log concentration of NO2 goes down by 0.026304 units. For each additional unit of log wind.speed, the log concentration of NO2 goes down with 0.414496. All of these have a p-value of less than 0.05, that means that the null-hypothesis is rejected. Both variables newwind.speed or the log variables of wind.speed, and temperature has negative effects that indicates that they are in good help to lower the air pollution. Thus, we observe that the magnitude of the coefficients for log.cars and log(wind.speed) are very similar, meaning that they play an equally important role in observed air pollution. The effect of the coefficients is shown in the figure 8.

The plots shown in the figure 8, indicates that the model is good enough, but the Scale-Location plot has somewhat less horizontal line compared to the Scale-Location plot in figure 5. The residual vs fitted plot shows outliers marked 371,372, similar that is shown in figure 5. The QQ plot shows that the data is normally distributed. The Residuals vs Leverage plot shows no outliers beyond Cooks distance. So, the model seems to be good enough for further analysis.

Exercise 2

In the table below there are measurements of the blood pressure of random samples of 12 men in each of three age groups.

30–45 years	46 – 59 years	60 –75 years
128, 104, 132, 112	120, 136, 174, 166	214, 146, 138, 148
136, 124, 112, 118	138, 124, 160, 157	156, 110, 188, 158
116, 108, 160, 116	108, 110, 154, 122	182, 148, 138, 136

The data are available in the file blood.txt on the course web page. Blood pressure is in column 1 and age group in column 2. Age group is coded with values 1, 2, and 3 and you must remember to specify that they should be considered as categorical (factors).

a) Describe the data using boxplots and numerical summaries. From these, does it seem that blood pressure is varying across age groups?

```
# Reading in data
blood <-"http://www.uio.no/studier/emner/matnat/math/STK4900/v20/mandatory/blood.txt"
blood <-read.table(blood, header = TRUE)
blood$age <-factor(blood$age)
```

We are summarising the data with hjelp of a little macros as the data for the age group has a f actor format.

```
blood %>%
group_by(age) %>%
summarize(q1 = quantile(Bloodpr, 0.25), q3 = quantile(Bloodpr, 0.75), Mean = mean(Bloodpr), Sd = sd(Bloodpr), min= min(Bloodpr), max=max(Bloodpr))
```

A tibble shows that there is a small difference in blood pressure between groups and shows that the blood pressure increased with the age. The youngest age group 1 from 30 to 45 years old has the lowest mean and standard deviation. The next group with the age from 46 to 59 has the higher observed blood pressure and standard deviation and the oldest group from 60 to 75 years old has the highest mean and the highest standard deviation. This makes sense as far as the blood pressure tends to be increased with the age.

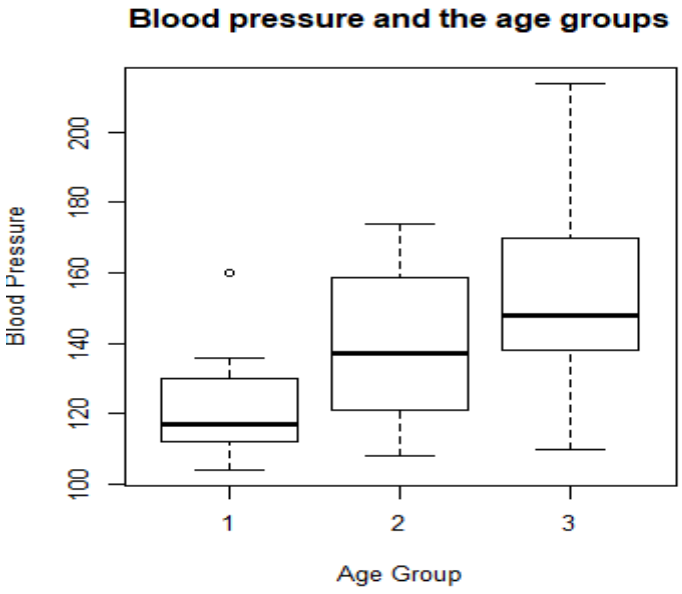
A tibble 1. Summary statistics of blood pressure grouped by age

```
# A tibble: 3 x 7
    age    q1    q3  Mean    Sd   min   max
  <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1     1  112   129   122.  15.3   104   160
2     2  122.  158.  139.  22.6   108   174
3     3  138   164   155.  27.7   110   214
```

```
Bloodpr=blood$Bloodpr

boxplot(blood$Bloodpr ~ blood$age, xlab = "Age Group", ylab = "Blood Pressure")
```

The boxplot gives us the visualisation of the numerical results from the summary. The younger group (1) has lower blood pressure than the other groups, but it doesn't look like there is a significant difference between the two older groups. The group 1 has the most concentrated blood pressure out of all groups, showing that younger people have a relatively stable blood pressure except for the one observation with the value 160 in blood pressure that is probably is the outlier at the figure 9. We also observe that group 3 has the widest range of blood pressure.

Figure 9: Boxplot. The distribution of the blood pressure among three age groups.



b) Use one-way ANOVA to answer the question above. Specify assumptions and the hypotheses you are testing. Write a summary of your findings.

```
#2b) One way anova and summary

aov.blood = aov(Bloodpr~blood$age, data = blood)

summary(aov.blood)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
blood$age    2   6535    3268   6.469 0.00426 **
Residuals   33  16670     505
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

One-way analysis of variance (ANOVA) is used to check if blood pressure varies across age and to determine whether there are any statistically significant differences between the set of independent groups. In general, when performing ANOVA, it needs to make the following assumptions:

- The data has to be normally distributed, i.e. N(μk,σ2).
- The variance among the group needs to be almost similar (assumption for homoscedasticity and homogeneity of the variance).
- The observations must be independent.

H0 : μ1 = μ2 = μ3, where μk is the mean blood pressure of group k. We are interested in testing this null hypothesis to see if μ1, μ2 6 and μ3 are not equal, meaning that the age influences the blood pressure or there is not a differences between blood pressure and age. The output from above shows the main results from ANOVA.

The ANOVA shows that there is a difference between at least one of the groups since the p-value is less than 0.05, but it doesn’t say which group is different. The ANOVA results in a F value of 6.469 is of significant magnitude, suggesting that there is indeed some correlation between age and blood pressure. This means that we can reject the null hypotheses H0.

c) Formulate this problem using a regression model with age group as categorical predictor variable. Use treatment-contrast and the youngest group as reference. Run the analysis, interpret the results and write a conclusion. Compare with the analysis in b).

```
# Doing the linear regression and summary of the fit by the traditional way (without a loop)

fit.blood = lm(Bloodpr~blood$age, data = blood)

summary(fit.blood)

# There is another way to make it possible via loop. The cole is here:

#https://stats.idre.ucla.edu/r/modules/coding-for-categorical-variables-in-regression-models/?fbclid=IwAR1q2l3fR8CHB7DUDhST6wyaOHemFpvckUzZXfYOZg5qvAmj49WddRBxFQ4

blood <- within(blood, {
  Agegroup <- C(age, treatment)
  print(attributes(Agegroup))
})

summary(lm(Bloodpr ~ Agegroup, data=blood))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-45.167 -15.583  -5.167  14.104  58.833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  122.167      6.488  18.829  < 2e-16 ***
blood$age2    16.917      9.176   1.844  0.07423 .
blood$age3    33.000      9.176   3.596  0.00104 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.48 on 33 degrees of freedom
Multiple R-squared:  0.2816,   Adjusted R-squared:  0.2381
F-statistic: 6.469 on 2 and 33 DF,  p-value: 0.004263
```

The same results can be received after running of the following code:

```
# There is another way to make it possible via loop. The cole is here:

#https://stats.idre.ucla.edu/r/modules/coding-for-categorical-variables-in-regression-models/?fbclid=IwAR1q2l3fR8CHB7DUDhST6wyaOHemFpvckUzZXfYOZg5qvAmj49WddRBxFQ4

blood <- within(blood, {

  Agegroup <- C(age, treatment)

  print(attributes(Agegroup))

})

summary(lm(Bloodpr ~ Agegroup, data=blood))
```

In the analysis the youngest group 1 is used as a control or a reference group. Out from the output we can see that the age group 2 is not significantly different from age group 1, but age group 3 is, on the other hand, significantly different from group 1, given the p-values with a threshold at $p>0.05$. This could suggest that blood pressure increasingly affected by age in a non-linear way. Thus, we could find out if the gender influences the blood pressure.

The R2 measurement is quite low, 0.2816, meaning that the data is not well fitted to the regression line. However, this does not necessarily pose a problem and we can still get the meaningful analyse.

This analysis fits well with the analysis in 2b) Both analyses agrees that there is a difference between two or more of our groups. In 2b) we have concluded that the difference exists, and in 2c) we can see that this difference exists between age group 1 and age group 3.

Appendix

The code for exercise 1 and 2

```
rm(list=ls(all=TRUE))


# Obligatory exercise 1


# Reading datafiles

no2data <-
read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v20/mandatory/no2.txt",sep="\t",header=TRUE)


# 1a)

# Catogorizing the data

names(no2data)=c("log.no2","log.cars","temp","wind.speed","hour.of.day")


# Summary of the pollution levels and number of cars

summary(no2data)

summary(no2data$log.cars)


# Boxplot of log.no2 and log.cars

boxplot(no2data$log.no2, no2data$log.cars, names = c("log.no2", "log.cars"), main =
"Boxplot for the cars and NO2")


# Scatterplot of log.cars and log.no2

plot(no2data$log.cars, no2data$log.no2, main = "Scatterplot of log.cars and log.no2")


# 1b)
```

```
# Linear fit
fit=lm(no2data$log.no2~no2data$log.cars)
summary(fit)

# Scatterplot with the linearfit
plot(no2data$log.cars, no2data$log.no2, main = "Scatterplot with the linearfit")
abline(fit, col = "blue")



# Correlation between no2data$log.cars,no2data$log.no2
cor(no2data$log.cars,no2data$log.no2)**2

# 1c)
# Linearity check
library(car)
crPlots(fit, terms=~no2data$log.cars)

# Checking for constant variance
par(mfrow = c(2,2))
plot(fit)

# Checking for normality
par(mfrow = c(1,2))
hist(fit$res)
boxplot(fit$res, main = "Boxplot for the fitted model")
qqnorm(fit$res); qqline(fit$res)

# 1d)

# Correlation check
```

```
cor(no2data$log.cars,no2data$hour.of.day)


# Fitting the model of the multiple linear regression
Multireg = lm(no2data$log.no2~no2data$log.cars+no2data$temp+no2data$wind.speed +
no2data$hour.of.day)
summary(Multireg)


boxplot(no2data$temp, no2data$wind.speed, no2data$hour.of.day , names = c("temp",
"wind.speed", "hour"), main = "Boxplot temp, wind speed, hour")


newwind.speed=log(no2data$wind.speed)


# Fitting of the multiple linear regression model without hours.per.day as it is closely
correlated to log.cars
fit.multiregfinal = lm(no2data$log.no2~no2data$log.cars+newwind.speed+no2data$temp)
summary(fit.multiregfinal)


cor(newwind.speed,no2data$hour.of.day)
cor(no2data$log.no2,no2data$hour.of.day)
cor(no2data$temp,no2data$hour.of.day)


#1e)Summary of the final model


crPlots(fit.multiregfinal, terms=~no2data$log.cars+no2data$temp+newwind.speed)


par(mfrow = c(2, 2))
plot(fit.multiregfinal)


# 2a)
```

```
# Reading in data
blood <-"http://www.uio.no/studier/emner/matnat/math/STK4900/v20/mandatory/blood.txt"
blood <-read.table(blood, header = TRUE)
blood$age <- factor(blood$age)

blood %>%
group_by(age) %>%
summarize(q1 = quantile(Bloodpr, 0.25), q3 = quantile(Bloodpr, 0.75), Mean =
mean(Bloodpr), Sd = sd(Bloodpr), min= min(Bloodpr), max=max(Bloodpr))

Bloodpr=blood$Bloodpr
boxplot(blood$Bloodpr ~ blood$age, xlab = "Age Group", ylab = "Blood Pressure", main =
"Blood pressure and the age groups")

#2b) One way anova and summary
aov.blood = aov(Bloodpr~blood$age, data = blood)
summary(aov.blood)

#2c) Formulate this problem using a regression model with age group as categorical
#predictor variable. Use treatment-contrast and the youngest group as reference.
#Run the analysis, interpret the results and write a conclusion.
#Compare with the analysis in b).

# Doing the linear regression and summary of the fit by the traditional way (without a loop)
fit.blood = lm(Bloodpr~blood$age, data = blood)
summary(fit.blood)
# There is another way to make it possible via loop. The cole is here:
#https://stats.idre.ucla.edu/r/modules/coding-for-categorical-variables-in-regression-
models/?fbclid=IwAR1q2l3fR8CHB7DUDhST6wyaOHemFpvckUzZXfYOZg5qvAmj49Wd
dRBxFQ4
```

```
blood <- within(blood, {
Agegroup <- C(age, treatment)
print(attributes(Agegroup))
})
summary(lm(Bloodpr ~ Agegroup, data=blood))
```