STK4900 Obligatory 1

Jonas Thoen Faber

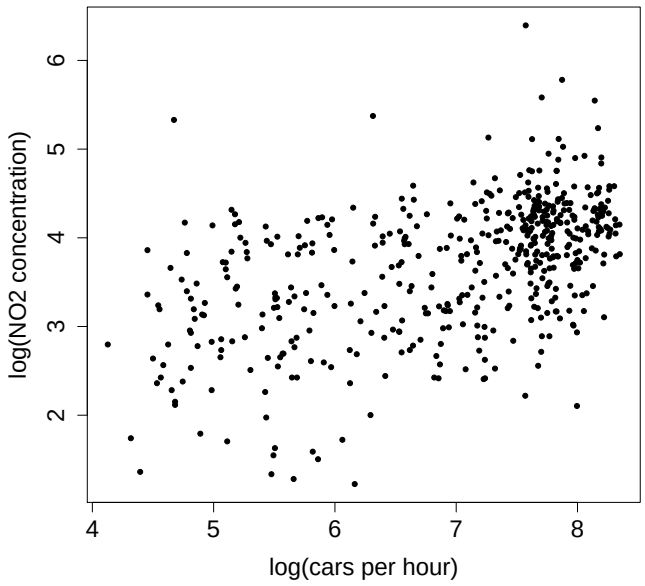
March 4, 2021

All coding is done using jupyter lab and can be found at https://github.com/JonasTFab/STK4900_Statistical_Methods_and_Applications/tree/main/obligatory_1. A compressed version of the code are located in the appendix.

Problem 1

a)

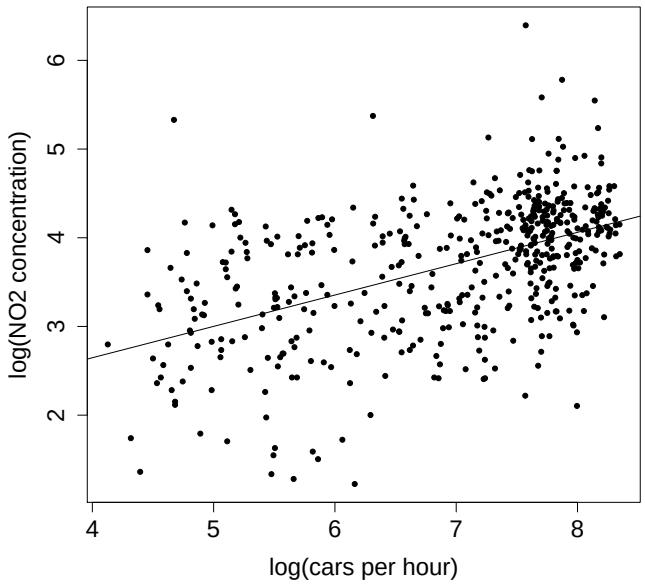
First we import the data from https://www.uio.no/studier/emner/matnat/math/STK4900/data/no2.txt using the R functionality `read.table` and then plot the data as a scatterplot. We plot **log.no2** as a function of **log.cars** and get the following result:



We observe from the plot that the logarithm of the pollution seems to increase somewhat linearly with the logarithm of number of cars and is reasonably expected to be true. However, the deviation is quite large for relatively many datapoints.

b)

Next we introduce the `lm()` function provided that simply fit a linear model based on some data. By use of `summary()` on this function gives additional informative results. We want to estimate the pollution as function of cars per hour (both on a logarithmic scale). The functionality mentioned resulted with a intercept of 1.2331 and a slope of 0.3535. A brief interpretation is that the intercept presents the amount of pollution at Alnabru if there were zero cars per hour. So if it were no cars, the logarithmic pollution would be 1.2331. The slope presents the increase of pollution for every unit increase in cars per hour. For example, if the amount of logarithmic number of cars per hour were 5, then the logarithmic pollution would be $1.2331 + 5 \times 0.3535 = 3.0006$.
The linear model is shown with the datapoints in the plot below:

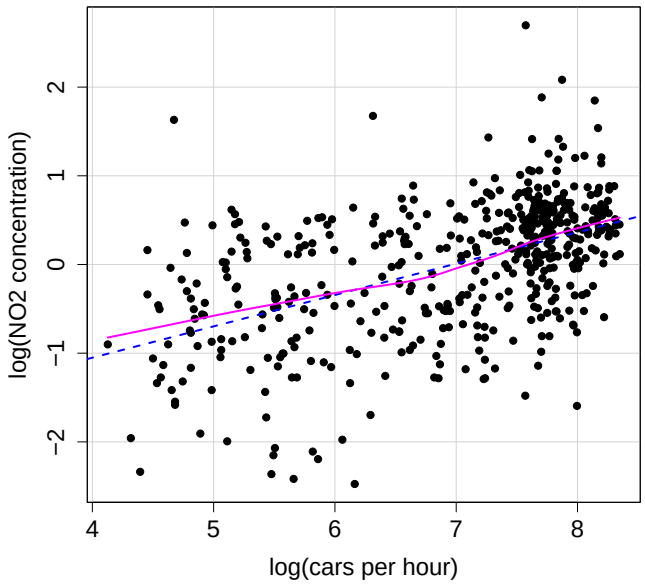


R^2 is a measure that correlates two variables. In this case, we want to find the correlation between the log of NO_2 pollution and the log of cars per hour. This value is printed when using the `summary()` function. We get an correlation value of $R^2 = 0.2621$ telling us that there is an approximate 26 % correlation between the pollution and the number of cars per hour. This result can imply

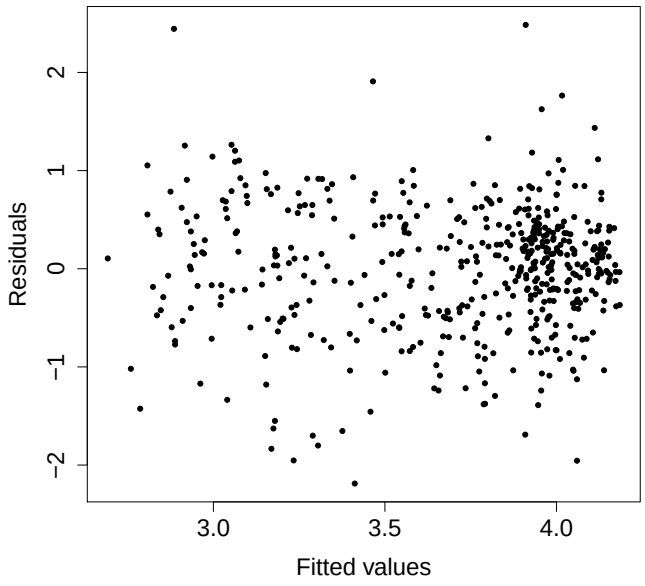
two things. Unwanted noise added to the measured data or that there are other factors contributing to the NO_2 pollution. Both of them seems equally valid.

c)

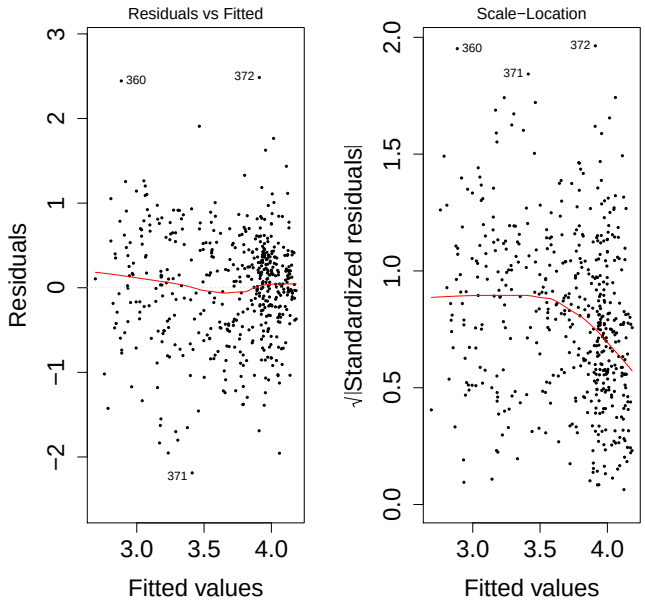
We begin with a CPR plot, or in other words, check of linearity. This is made with `crPlots` provided by the `car` library in R. We get the following plot:



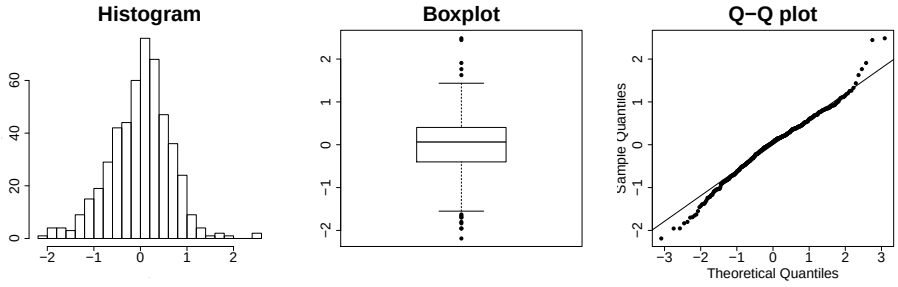
Note that we are still considering the log of NO_2 concentration as a function of the log of number of cars per hour. Looking at the CPR plot, we no that linearity are close to perfect but a second order polynomial might give a slightly better fit.
Next we want to check if the variance of the residuals are constant or not. We use the same fitted model with the log of number of cars per hour as the predictor and plot the residuals as function of fitted values as shown below:



The variance seems to be quite constant even for the part that is less dense. It is hard to judge "by eye" so we will include two additional plots. The first is to fit a line to see if we observe a pattern in the residuals. The other plot is to observe potential increase in variance. The plots are shown below:



We note from the plot to the left that the residuals seems to be somewhat constant with slight deviations. The variance is constant at approximately 0.9 for small amount of cars per hour and it decreases greatly at higher amount of cars per hour as seen from the right plot. The decrease of variance makes us believe that the fitted model is better (i.e. greater accuracy) for higher amount of cars per hour.
Next we wish to check the distribution of the residuals, that is if the residuals is normal distributed or not. This is done by a histogram plot, box plot and a Q-Q plot. We do so by using the three functionalities in R, that is `hist`, `boxplot` and `qqnorm`. Additional line is added to the Q-Q plot by using `qqline`.



The histogram plot immediately makes us believe that the residuals are very likely to be normal distributed. There is approximately an equal distance be-

tween the first and third quantile from the median in the boxplot. The three plots above indicates that there exists outliers in the data we consider in this exercise.
Overall, a regression analysis with a single predictor gives a quite reasonable representation of the data based on the plot presented above.

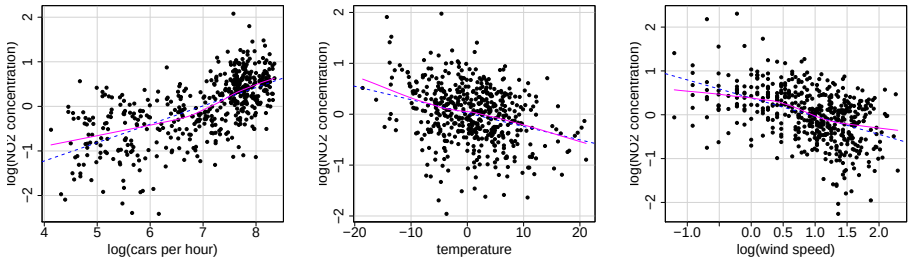
d)

Now we wish to study the simultaneous effect of the covariates between all variables which can be done by using the built-in function `cov` provided. `cov` is used on the extracted data so that we get the covariates between all variables as shown below:

A matrix: 5 × 5 of type dbl

	log.no2	log.cars	temp	wind.speed	hour.of.day
log.no2	1.0000000	0.51205042	-0.1681592	-0.328834938	0.246201259
log.cars	0.5120504	1.00000000	0.2018317	0.097530682	0.576856715
temp	-0.1681592	0.20183175	1.0000000	0.165990712	0.079484997
wind.speed	-0.3288349	0.09753068	0.1659907	1.000000000	-0.002789753
hour.of.day	0.2462013	0.57685671	0.0794850	-0.002789753	1.000000000

We observe that the correlation between `log.cars` and `hour.of.day` is by far the greatest with an value of 0.5769. Hence, `hour.of.day` may be removed when fitting a multi-regression model. Trying to fit a linear model with/without `hour.of.day` increase the R^2 value to 0.4658/0.4566 which is approximately twice the value as before. Removing this variable decreases the standard error of `log.cars` by 0.0051 (0.0284 $\rightarrow$ 0.0233).
Next we try to take the log of all variables except for the `temp` as it contains negative values. Doing so increases the R^2 value to 0.4807/0.4750 with/without `log(hour.of.day)`. The standard error of `log.cars` is worse overall if we include `log(hour.of.day)` as a variable. Excluding this variable reduce the standard error of `log.cars` down to 0.0230 which is slightly better than before.
Creating a CPR plot will visualise the linearity of the different variables:



Here we see that the different variables fit a linear model quite well based on the residual plot.

e)

Next we investigate the coefficients from the 'best' fitted model from d). These can be found in the figure below:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.229009   0.162586   7.559 1.98e-13 ***
log.cars        0.411979   0.022995  17.916  < 2e-16 ***
temp           -0.026304   0.003861  -6.813 2.79e-11 ***
log(wind.speed) -0.414496   0.036572 -11.334  < 2e-16 ***
```

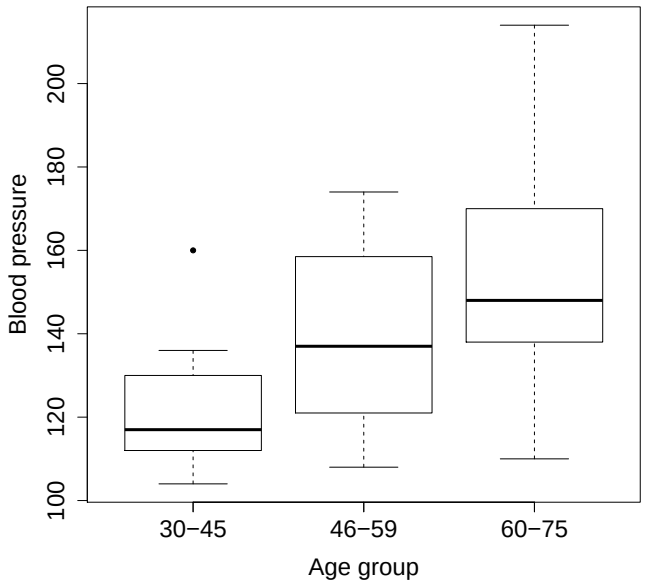
We observe various slopes of the pollution as a function of the three variables (`log.cars`, `temp` and `log(wind.speed)`). This tells us that if we keep all of the predictors constant except one and increase this by one unit, the amount of pollution should increase according to the estimate shown above. This is indeed not certain as we have a standard error, hence a standard deviation of the data.

We note that that temperature and wind speed predictors has a negative sign meaning that the response (pollution) decreases as those increase. The opposite happens for the number of cars which all makes reasonably sense. These conclusion coincides with what we observe from the CPR plot in d). Also, the `temp` predictor from the table above is shown to have a relatively small standard error. The small standard error imply a small standard deviation ($\text{SE} = \sigma/\sqrt{(n)}$) which tells us that the measured temperature are accurate in a greater degree than the other predictors. Investigating the CPR plot in d) once more, we see that the temperature plot represents a better linearity than the two other predictor plot which may be an result of the small standard error.

Problem 2

a)

We start by extracting the data from https://www.uio.no/studier/emner/matnat/math/STK4900/data/blood.txt in R and then categorize each of the three age groups. That is group 1 corresponds to the aging group 30–45 years, group 2 from 46-59 years and finally group 3 from 60–75 years. We then create a boxplot which can be shown below:



We observe that the variety of blood pressure are more spread out with respect to aging group. Also, the first quantile, median and the third quantile is increasing with aging group which support the hypothesis that blood pressure increase with age. There exists one outlier in the 30-45 years group which might be caused by other unknown factors.

b)

Next we want to perform a analysis of variance (ANOVA) to check if the blood pressure indeed is dependent on the age group. More specific, we will perform a one-way ANOVA as we have only one variable in this dataset (age group). Using ANOVA, we assume that obsrvations are independent of each other and that they are normally distributed, that is $N(\mu_k, \sigma^2)$ where $k \in [1, 2, 3]$. We want to test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ and reject it if we get the alternative

hypothesis $H_A : \mu_1 \neq \mu_2 \neq \mu_3$. The null hypothesis can be rejected for large values of the test statistic that is F distributed. We may find this value by using the provided ANOVA functionality `aov()` in R on the data. First we need to categorise into aging groups which can be done with the `factor()` function in R. Then, using ANOVA, we get the following results:

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

We note the F value of 6.469 which is greater than than 3.32 (95th percentile from F table), hence we may reject the null hypothesis. Since the null hypothesis is rejected, we know that $\mu_1 \neq \mu_2 \neq \mu_3$ and therefore the age group should have an impact on the blood pressure.

c)

Now we want to reformulate the problem by solving it with a regression model. Age group should be the predictor while blood pressure is the response. The youngest age group is set as reference, hence a treatment-contrast is to be analysed and is a standard option in R (other use sum-contrast as standard). We now fit a linear model using `lm()` on the data and get the results shown below:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  122.167      6.488  18.829  < 2e-16 ***
age2          16.917      9.176   1.844  0.07423 .
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
```

We first note the difference in the predictor from `age2` and `age3`. The reason for these two significant different value of the covariates, can be explained by the distance between each age group. We remember that age group 1 was our reference point and therefore `age2` is just a single step in the age axis while `age3` is two steps in the age axis. If we then divide the predictor of `age3` by 2, we note that this slope is slightly smaller than the slope for the `age2` predictor. We also note that SE are exactly the same.
Next we investigate the P-value, which are approximately 7.4 % and 0.1 % for age group 2 and 3 respectively. We note that the first P-value is greater than a chosen threshold of 5 %, meaning the predictor is not significant. The predictor of age group 3 are then, because of its P-value, found to be significant.

One could suppose that the blood pressure is not linearly dependent with age but rather quadratic instead. If we take a brief look at the box plot from 2a), we note the variance of each age group to be quite large relative to the various box plots. Hence, an assumption of a linear or quadratic dependency is equally plausible.
Lastly, we observe the multiple R^2 value of 28.2 %. The R^2 value is a measure of the correlation between predictor and response. The correlation value tells us that the blood pressure can be described by the age group to some degree. Noise will affect the measurement and other unknown factors might also describe the blood pressure of men.

1 Appendix

Code

```r
# 1a)
data = read.table("https://www.uio.no/studier/emner/matnat/math/STK4900/data/no2.txt",
     header=T)

plot(log.no2~log.cars, data=data, pch=19,
     xlab="log(cars per hour)", ylab="log(NO2 concentration)",
     cex.lab=1.5, cex.axis=1.5, cex=0.7)




# 1b)
plot(log.no2~log.cars, data=data, pch=19,
     xlab="log(cars per hour)", ylab="log(NO2 concentration)",
     cex.lab=1.5, cex.axis=1.5, cex=0.7)

fit = lm(log.no2~log.cars, data=data)
abline(fit)
summary(fit)




# 1c)
# Check of linearity
library(car)

crPlots(fit, terms=~log.cars, pch=19,
     xlab="log(cars per hour)", ylab="log(NO2 concentration)",
     cex.lab=1.5, cex.axis=1.5, cex=0.7)


# Check of constant variance
plot(fit$fit, fit$res, pch=19,
     xlab="Fitted values", ylab="Residuals",
     cex.lab=1.5, cex.axis=1.5, cex=0.7)

par(mfrow=c(1,2))
plot(fit,1, pch=19, cex.lab=1.5, cex.axis=1.5, cex=0.3)
plot(fit,3, pch=19, cex.lab=1.5, cex.axis=1.5, cex=0.3)


# Check of normality
hist(fit$res, pch=19, main="Histogram", breaks=20,
     cex.lab=0.1, cex.axis=2, cex.main=3)

boxplot(fit$res, pch=19, main="Boxplot",
     cex.lab=0.1, cex.axis=2, cex.main=3)

qqnorm(fit$res, pch=19, main="Q-Q plot",
     cex.lab=2, cex.axis=2, cex.main=3)
qqline(fit$res)




# 1d)
# fitting model with and without hour.of.day
fit.multiple = lm(log.no2 ~ log.cars + temp + wind.speed + hour.of.day, data=data)
summary(fit.multiple)

print("-------------------------------------------------------------------------")
```

```r

fit.multiple = lm(log.no2 ~ log.cars + temp + wind.speed, data=data)
summary(fit.multiple)


# fitting model with the log of variables and also with and without hour.of.day
fit.multiple.log = lm(log.no2 ~ log.cars + temp + log(wind.speed) + log(hour.of.day), data=
      data)
summary(fit.multiple.log)

print("-------------------------------------------------------------------------")

fit.multiple.log = lm(log.no2 ~ log.cars + temp + log(wind.speed), data=data)
summary(fit.multiple.log)


# various plots using the log of variables without 'hour.of.day'
crPlots(fit.multiple.log, terms=~log.cars, pch=19,
     xlab="log(cars per hour)", ylab="log(NO2 concentration)",
     cex.lab=2, cex.axis=2, cex.main=3)

crPlots(fit.multiple.log, terms=~temp, pch=19,
     xlab="temperature", ylab="log(NO2 concentration)",
     cex.lab=2, cex.axis=2, cex.main=3)

crPlots(fit.multiple.log, terms=~log(wind.speed), pch=19,
     xlab="log(wind speed)", ylab="log(NO2 concentration)",
     cex.lab=2, cex.axis=2, cex.main=3)







# 2a)
data = read.table("https://www.uio.no/studier/emner/matnat/math/STK4900/data/blood.txt",
     header=T)

boxplot(data$Bloodpr ~ data$age, pch=19,
     xlab="Age group", ylab="Blood pressure",
     cex.lab=1.5, cex.axis=0.01)
axis(1, at=1:3, labels=c("30-45", "46-59", "60-75"), cex.axis=1.5)
axis(2, cex.axis=1.5)




# 2b)
data$age = factor(data$age)
aov.data = aov(Bloodpr~age, data=data)
summary(aov.data)




# 2c)
fit = lm(Bloodpr ~ age, data=data)
summary(fit)
```