GFM2 Module 5: Principles of Spatial Data Quality Individual assignment

This exercise is the individual assignment of Module 5. This is given because you should get credit for the practical component of the module and not be assessed only based on the written test. Please write your answers to the questions in the boxes provided and save the document with the name <surname>_<student_number>_m5.doc. It must be submitted, via Blackboard (Assignment tab) by 28 January 2018 at 8 pm.

The total number of marks for the questions adds up to 100 which will be converted to a mark with one decimal from 1.0 to 10.0 for the individual assignment.

Overview

Geoinformatics covers a range of topics including data archiving and discovery; data acquisition, processing and dissemination; the representation of spatial phenomena; new technology (new sensors and new computing tools); data integration; uncertainty and data quality. We are further interested in the application of robust analytical techniques and the development of new techniques. The correct application of regression is important here.

You are provided with a pre-processed air pollution dataset from one location in the city of Eindhoven, the Netherlands. The pollutant is nitrogen dioxide (NO2), measured in µg m^{-3}. NO2 is measured using two methods: a high quality reference monitor from the Dutch Institute for Public Health and Environment (RIVM) and a low-cost air quality sensor ("Airbox").

The measurements taken by the RIVM instruments are considered to give the highest quality measurements of NO2. Unfortunately, the instruments are expensive, so only a limited number of measurements can be made. Therefore, a network of low-cost air quality sensors (airboxes) has been installed in the city of Eindhoven. We will evaluate the data quality of these low-cost sensors and see if we can improve the measurements by adding other observations, such as weather variables, to the regression model.

The data provided are daily averages for October 2016. A summary is given below. The data are taken from one location, at which both an RIVM monitor and an Airbox measure the concentration of NO2. Each row of the table provides hourly averaged values. The data also include meteorological variables measured at a nearby weather station (KNMI).

```
> head(ap.cal)
                 Time Ref_NO2 Airbox_NO2 Airbox_O3 Mean_temp Wind_speed  Rel_hum
1 2016-10-01 01:00:00   27.17   27.53365 15.846734  13.46002         30 63.25254
2 2016-10-01 02:00:00   24.29   27.09239 17.769346  13.31349         30 65.30465
3 2016-10-01 03:00:00   22.55   23.29825 15.362536  12.93225         20 69.54315
4 2016-10-01 04:00:00   23.68   25.68212 12.065449  12.77198         10 70.96192
5 2016-10-01 05:00:00   28.48   24.49750  9.758195  12.56574         30 72.93417
6 2016-10-01 06:00:00   32.43   26.79655  7.473479  12.42833         10 74.47611
```

The column headings are given as follows:

Date	- indicates the date of the measurements (day-month-year)
Ref_NO2	- NO2 concentrations measured using the RIVM instrument (our reference), in µg m^{-3}
Airbox_NO2	- NO2 concentrations measured using the low-cost Airbox, in µg m^{-3}
Airbox_O3	- Ozone (O3) concentrations measured using the low-cost Airbox, in µg m^{-3}
Mean_temp	- Mean_temp (in °C)
Wind_speed	- wind speed (in 0.1 m s^{-1})
Rel_hum	- Relative humidity (%)

As indicated, the Airbox is known to be less accurate than the RIVM measurements, but is useful as a predictor variable (covariate).

Exercise

The dataset (M5_cal.csv) contains the part of the dataset we will use for this assignment. It contains hourly values for October 2016 on the air pollution and meteorological variables. You should import the tables into R, and convert the time column to a suitable format:

```
ap.cal = read.csv("M5_cal.csv", head = T, sep =";")
ap.cal$Time <- strptime(ap.cal$Time, format="%d-%m-%Y %H:%M")
```

Begin by exploring the variables (summary statistics, histograms, scatterplots, boxplots, QQ plots etc).

1) Explore the dataset. Looking at the units supplied above, do the summary statistics values for each variable fall within reasonable and naturally possible limits?
The Royal Netherlands Meteorological Institute (KNMI) defines wind speeds of ≥ 8 m s^{-1} as stormy. How many stormy hours were there in October 2016?
Explain how you reached your answer. (10 marks)

I checked the summary of the variables. I especially focus on checking the minimun, maximun, median and mean, variables. Yes. They fall within reasonable and naturally possible limits. Because the unit of wind speed in the csv file is 0.1m/s. I calculate it to m/s. Then I do the summary statistics on it. This conversion on the unit make it more easy to check whether the value is reasonable or not.

```
> summary(ap.cal$Time )
                 Min.               1st Qu.                Median
"2016-10-01 01:00:00" "2016-10-08 18:45:00" "2016-10-16 12:30:00"
                 Mean               3rd Qu.                  Max.
"2016-10-16 12:33:42" "2016-10-24 06:15:00" "2016-11-01 00:00:00"
> summary(ap.cal$Ref_NO2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   6.18   19.17   25.54   27.69   34.36   67.89      52
> summary(ap.cal$Airbox_NO2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  13.22   22.99   26.43   28.13   32.45   58.76      65
> summary(ap.cal$Airbox_O3 )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  2.891   6.502  16.446  25.161  39.289 117.929       6
```

```
> summary(ap.cal$Mean_temp )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  4.525   9.395  10.789  10.919  12.347  20.712       6
> summary(ap.cal$Wind_speed )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    20.0    30.0    30.2    40.0   100.0
> summary(ap.cal$Rel_hum )
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  35.84   61.34   70.81   67.49   75.50   82.53       6
```

```
> summary(ap.cal$Wind_speed*0.1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    2.00    3.00    3.02    4.00   10.00
```

R commands:

```
count_stormy_hours=0
for(i in 1:length(ap.cal$Wind_speed)+1)
{
 if (ap.cal$Wind_speed[i]*0.1>=8)
 {
  count_stormy_hours=count_stormy_hours+1
 }
}
count_stormy_hours
```

answer: 4
In October, 4 hours are stormy hours in total.

2) Take a look at the histogram and Q-Q plot of ozone (Airbox_O3). Do you consider the ozone data symmetrically distributed? Do you consider the ozone data normally distributed? Take a look at the boxplot of ozone. Do you find outliers? Now transform the data using a log-transformation:

```
ap.cal$log_O3 <- log(ap.cal$Airbox_O3)
```

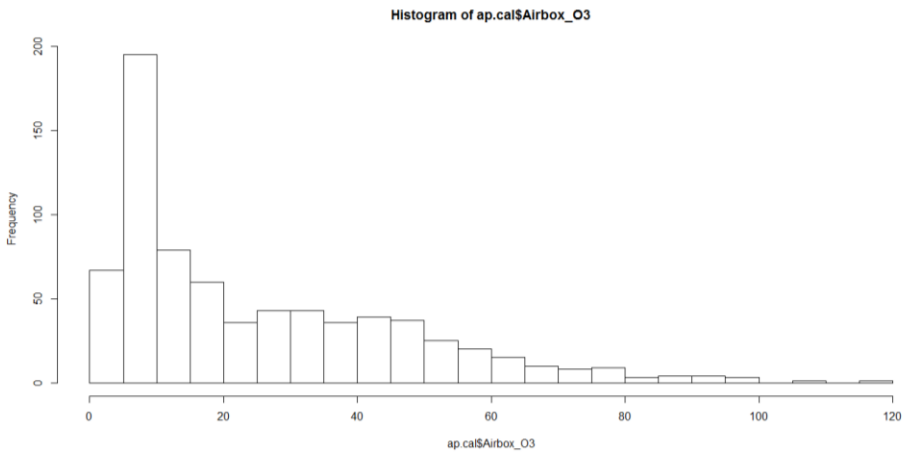
Examine the graphical plots of the transformed data. What changes do you observe? Explain your answers and provide evidence (e.g., graphical plots). (10 marks)

No, ozone data is not symmetrically distributed. Because, in the histogram, it has a peak on the left side of the median. It has a very long tail on its right side. In QQ-plot, a lot of points do not locate on the line. In the low-left corner of the QQ-plot, a lot of points are very far away from the line. This indicates that ozone data has a heavy tail. In boxplot, the location of the median line is not in the middle of the box. Instead it is very close to the bottom of the box. Additionally, the top whisker has much longer length than the low whisker.

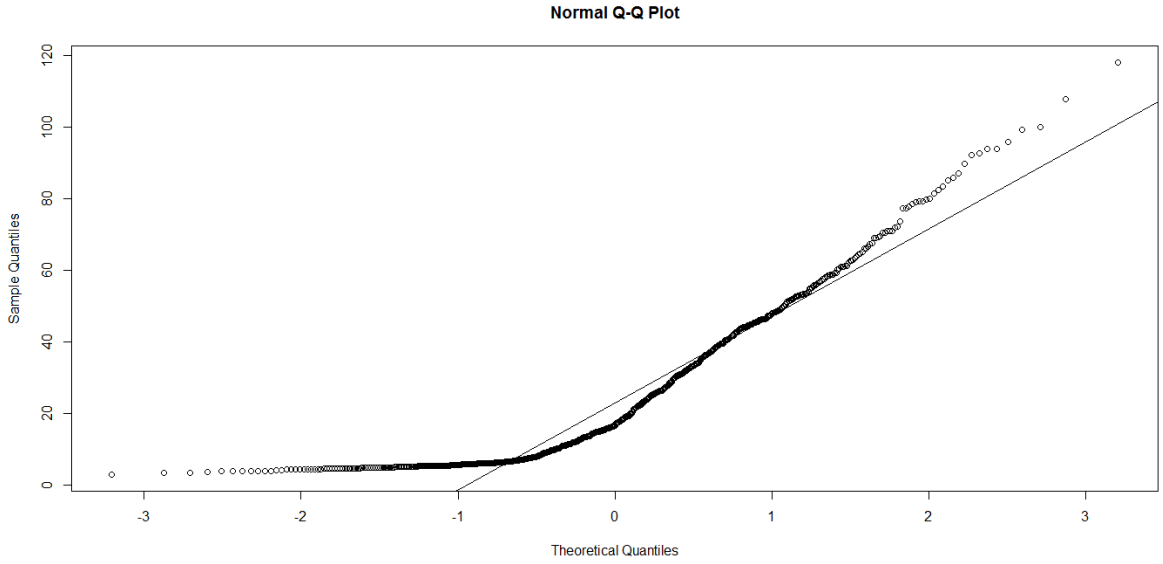
Ozone data is not normally distributed. Symmetricity is one of the requirements of normal distribution. Since ozone data is not symmetrically distributed, it must not be normally distributed. What I have observed in histogram and QQ-plot (described in the above paragraph) has proved this.

Yes, I have found outliers. The small circles that locate above the top whisker are the outliers.

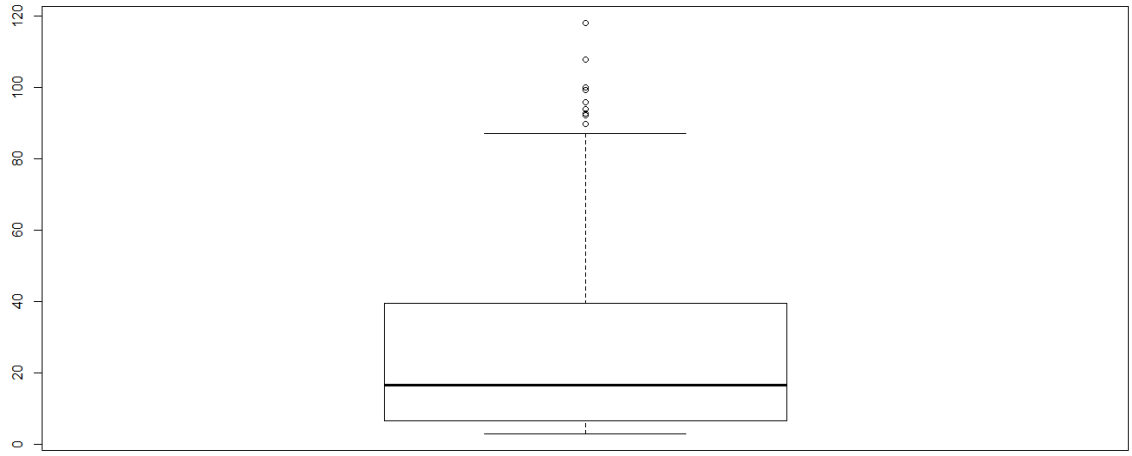
Histogram of ozone:
hist(ap.cal$Airbox_O3)



QQ-plot of ozone:
qqnorm(ap.cal$Airbox_O3)
qqline(ap.cal$Airbox_O3)



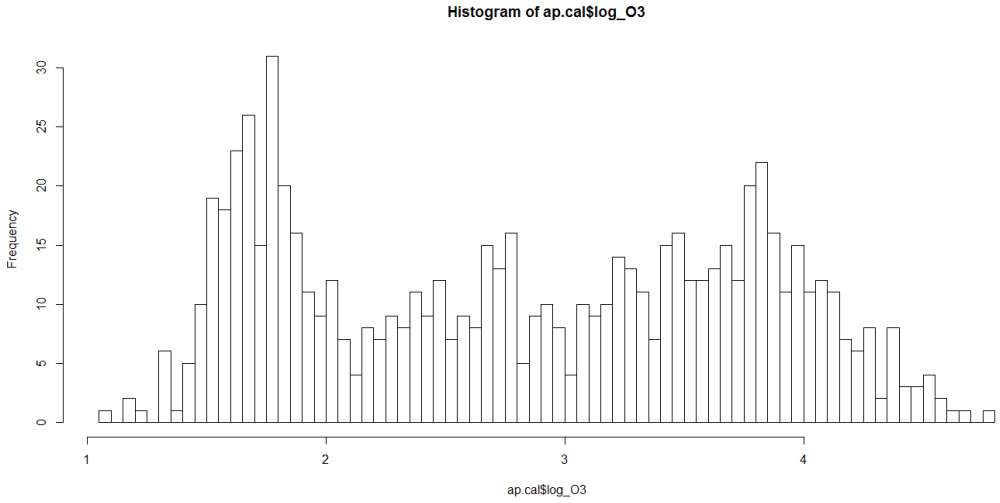
Boxplot of ozone:
boxplot(ap.cal$Airbox_O3)
points(1,mean(ap.cal$Airbox_O3))



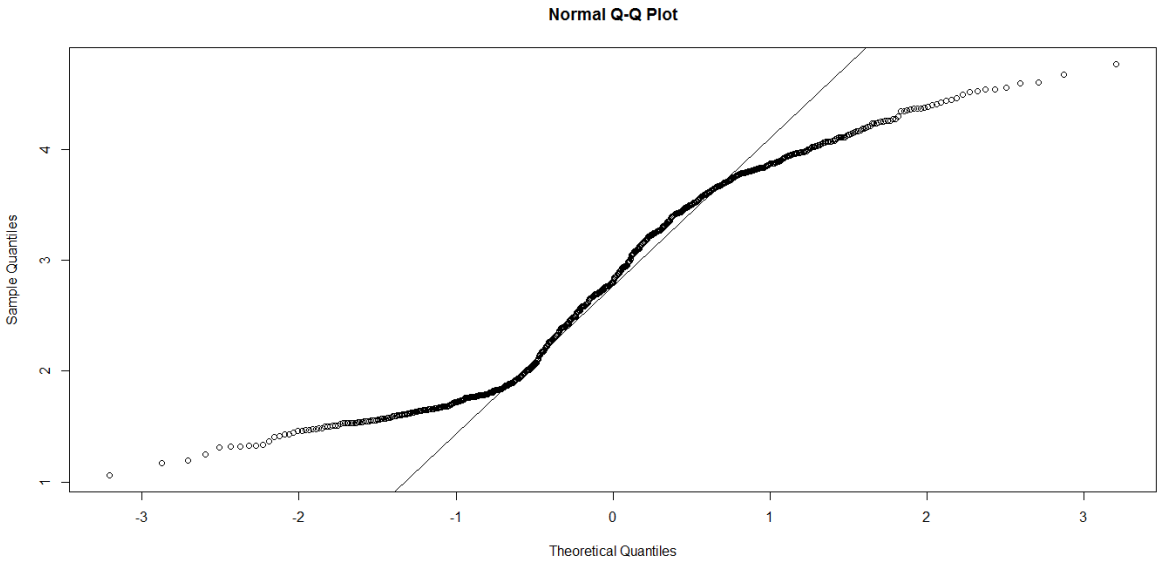
After transformation: log_O3 is roughly symmetrically distributed. Left disribution of the mean is similar, but not exactly the same, as right distribution of the mean. Log_O3 is not normally distributed. In boxplot, log_O3 does'nt have outliers.

Explanation: In histogram, it shows that log_O3 has two peaks. One peak locates on the left of the mean. The other peak locate on the right of the mean. The left peak is a bit higher than the right peak. So we can say that log_O3 is roughly, but not precisely, symmetrically distributed. In QQ-plot, two tails of the log_O3 are very far away from normal distribution. So it is definitely not a normal distribution. In boxplot, median line is located in the middle of the box. The top whisker is slightly longer than the low whisker. It doesn't have outliers.

Histogram of logtransformed-ozone
hist(ap.cal$log_O3)

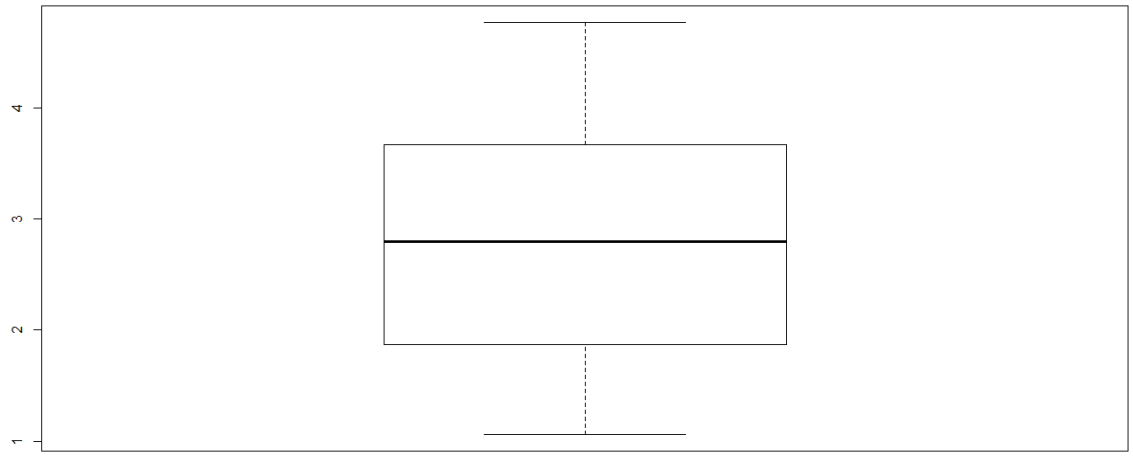


QQ-plot of of logtransformed-ozone
qqnorm(ap.cal$log_O3)
qqline(ap.cal$log_O3)



Boxplot of logtransformed-ozone

```
boxplot(ap.cal$log_O3)
points(1,mean(ap.cal$log_O3))
```



As stated above, we are going to construct regression models using "Ref_NO2" as the response variable, and "Airbox_NO2" as the explanatory variable. Besides that, we might add some extra explanatory variables. The explanatory variables can also be referred to as "covariates" or "predictor variables". First, we would like to examine the relationship between Ref_NO2 and several potential explanatory variables. We can do this using scatterplots. When you make a scatterplot, put the response variable on the y-axis and the explanatory variable on the x-axis.

3) Make scatterplots to investigate the following potential covariates: Airbox_NO2, Airbox_O3, log_O3, Mean_temp, Wind_speed, Rel_hum. Use Ref_NO2 as the response variable in all scatterplots. Examine the scatterplots and comment on the linearity and strength of the relationships.

Are the relations positive or negative? Did the transformation of O3 also improve its linearity? Explain your answers and provide evidence (e.g., graphical plots).
(10 marks)

R command:
X11()
par(mfrow=c(2,3))
Ref_NO2=ap.cal$Ref_NO2
plot(ap.cal$Airbox_NO2, Ref_NO2)
plot(ap.cal$Airbox_O3 ,Ref_NO2)
plot(ap.cal$log_O3 ,Ref_NO2)
plot(ap.cal$Mean_temp ,Ref_NO2)
plot(ap.cal$Wind_speed ,Ref_NO2)
plot(ap.cal$Rel_hum ,Ref_NO2)

Airbox_NO2: Airbox_NO2 is positively linearly related with Ref_NO2. The linearity is strong. Points lie in a straight line.

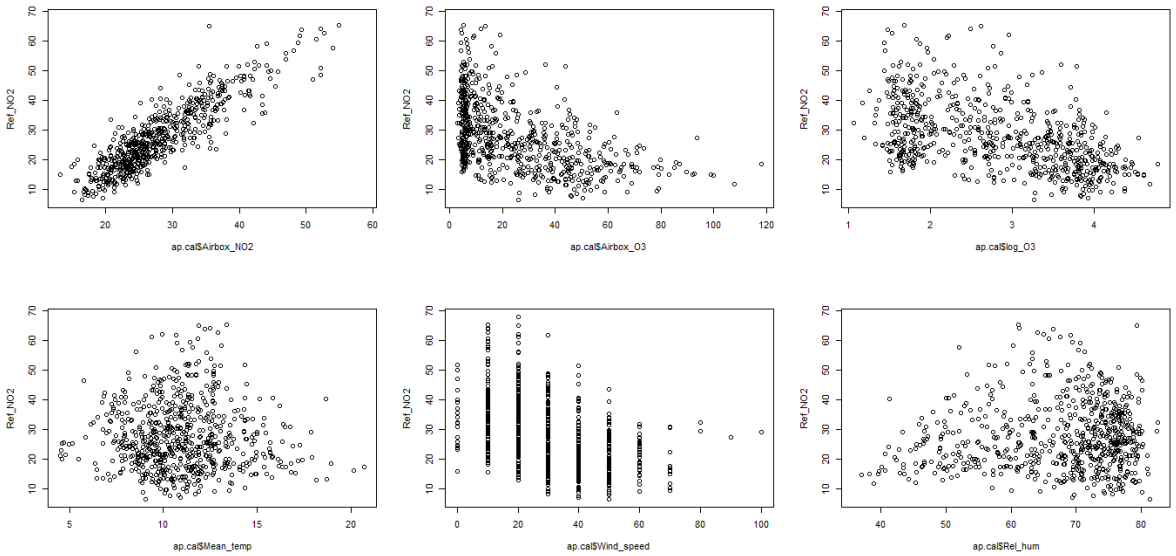
Airbox_O3: Airbox_O3 is negatively linearly related with Ref_NO2. The linearity is weak.

log_O3: log_O3 has a negative association with Ref_NO2. They are negative related. But the linearity is very weak. Log_O3 doesn't have improvement on the linearity, compared with Airbox_O3.

Mean_temp: there is no positive or negative association with Ref_NO2. With the increase of Mean_temp, Ref_NO2 may increase and may decrease.

Wind_speed: Wind_speed is negatively linearly related with Ref_NO2. The linearity is ok, but not strong. On each x value (wind_speed value), the corresponded Ref_NO2 shows a large variability. The mean of Ref_NO2 decreases, when wind_speed increases.

Rel_hum: relative humidity has a kind of positive association with Ref_NO2. But the association is not linear. The linearity is very weak. It is like a cone shape.



4) Build 6 linear models, each using one of the covariates we used in the previous question: Airbox_NO2, Airbox_O3, log_O3, Mean_temp, Wind_speed, Rel_hum. Use Ref_NO2 as the response variable in all linear models. Make a table to state for each covariate the intercept, slope, p-value and R^2-adj. Comment on the strength of the associations and whether the slopes are positive or negative. Does this agree with your findings in Question 3?
(10 marks)

Potential covariates	intercept	slope	p-value	R^2-adj
Airbox_NO2	-11.12190	1.37431	intercept <2e-16 *** slope <2e-16 ***	0.7775
	Linearity is strong. Association is strong. Correlation coefficient is close to 1. Slope is positive. This agrees with my findings in Question 3.			
Airbox_O3	34.15303	-0.26538	intercept <2e-16 *** slope <2e-16 ***	0.2468
	Linearity is weak. Correlation coefficient is very close to zero. Slope is negative. Airbox_O3 is weakly negatively associated with Ref_NO2. This agrees with my findings in Question 3.			
log_O3	45.062	-6.210	intercept <2e-16 *** slope <2e-16 ***	0.2556
	Linearity is weak. Correlation coefficient is close to zero. Slope is negative. Both O3 and log_O3 are very weakly related with Ref_NO2. These information match with my findings in questions 3. In the model output, the slop in log_O3 is steeper than the slop in O3. This is what I missed in question 3. The log-transformation on O3 indeed doesn't have improvement on the linearity. This agrees with my findings in Question 3.			
Mean_temp	29.0365	-0.1359	intercept <2e-16 *** slope 0.439	-0.0005864
	Association is very weak, or even "no association". There is almost no linearity. Slope is negative. But not significantly negative. Because p value of slope is very close to 0.5. This agrees with my findings in Question 3.			
Wind_speed	37.13406	-0.30690	intercept <2e-16 *** slope <2e-16 ***	0.1999
	Association is weak. There is a little linearity. Slope is negative. This agrees with my findings in Question 3.			
Rel_hum	20.98001	0.09717	intercept: 2.02e-12 *** slope: 0.0237 *	0.00601
	Association is weak. There is almost no linearity. Slope is positive. This agrees with my findings in Question 3.			

We would like to examine the correlation between variables before adding them together in a model with more than one covariate. Use the following code to obtain the correlation matrix (rounded to 2 decimal places). We exclude the first column which contains the time.

```
round(cor(ap.cal[,2:8], use="pairwise.complete.obs"),2)
```

5) First, have a look at the first row of the correlation matrix (Ref_NO2 vs. covariates). What is the relation between these correlation coefficients and the R^2 values you obtained in the linear model? Now have a look at the correlations between the covariates. Examine the values. Which covariates are correlated to each other? What influence does this have on including them together in a linear model? (10 marks)

The following table shows the first row in the correlation matrix. I am going to analyse the relation between these correlation coefficients and the R^2 values:

	Ref_NO2	Airbox_NO2	Airbox_O3	Log_O3	Mean_temp	Wind_speed	Rel_hum
Correlation with Ref_NO2	1.00	0.88	-0.50	-0.51	-0.03	-0.45	0.09
R^2 in the linear model		0.78	0.25	0.26	0.001	0.20	0.01

The value of R^2 in the linear model is equal to the square of correlation coefficient (R) with Ref_NO2. R^2 is in range from 0 to 1. R is in range from -1 to 1. When R is close to either -1 or 1, R^2 is close to 1. This means a high correlation. When R is close to zero, R^2 is also close to zero. It means a low or even no correlation.

correlations between the covariates:
In the following table, light-yellow part means the values that are not relevant to this question. Red mark means the value which is bigger or equal than 0.4.

```
> round(cor(ap.cal[,2:8], use="pairwise.complete.obs"),2)
           Ref_NO2 Airbox_NO2 Airbox_O3 Mean_temp Wind_speed Rel_hum log_O3
Ref_NO2
Airbox_NO2                    -0.43      -0.12      -0.40    -0.02  -0.43
Airbox_O3    -0.43                       0.44       0.57    -0.47   0.93
Mean_temp    -0.12     0.44                         0.17    -0.63   0.48
Wind_speed   -0.40     0.57     0.17                       -0.12   0.63
Rel_hum      -0.02    -0.47    -0.63      -0.12              -0.45
log_O3       -0.43     0.93     0.48       0.63    -0.45
```

Which covariates are correlated to each other?
Airbox_NO2 and Airbox_O3.
Airbox_NO2 and Mean_temp.
Airbox_NO2 and Wind_speed
Airbox_O3 and Wind_speed
Airbox_O3 and Rel_hum
Mean_temp and Rel_hum
Log_O3 with all the other covariates (Airbox_NO2, Airbox_O3, Mean_temp, Wind_speed, Rel_hum)

If correlated variables are included in one model, then they will have much bigger influence on the response variable, compared with other independent variables. Small errors in the correlated vairables would lead to a big propagation. Bias on the response variable will be generated.

In Question 4, we made linear models including one covariate. We can also combine covariates in the model. We continue with three covariates: Airbox_NO2, log_O3 and Wind_speed. Use Ref_NO2 as the response variable in all linear models.

6) Try different combinations of two covariates from the three covariates listed above (Airbox_NO2, log_O3 and Wind_speed). Also build a regression model of all three covariates. Comment on the change in R^2-adjusted, the significance of the covariates and the strength and sign of the slopes (positive/negative). Which model do you prefer as the final model? Which covariate would you drop when there is not enough money available to collect data on all covariates? Explain your answer. (20 marks)

```
# R command for question6
Ref_NO2=ap.cal$Ref_NO2
ap.cal$log_O3 <- log(ap.cal$Airbox_O3)
Airbox_NO2=ap.cal$Airbox_NO2 #NO2
Wind_speed= ap.cal$Wind_speed #wind
log_O3=ap.cal$log_O3 #logO3

Model_NO2_logO3.lm=lm(Ref_NO2~Airbox_NO2+log_O3) # 2 explanatory variables.
summary(Model_NO2_logO3.lm)
Model_NO2_wind.lm=lm(Ref_NO2~Airbox_NO2+Wind_speed) # 2 explanatory variables.
summary(Model_NO2_wind.lm)
Model_logO3_wind.lm=lm(Ref_NO2~log_O3+Wind_speed) # 2 explanatory variables.
summary(Model_logO3_wind.lm)
Model_3covariates.lm=lm(Ref_NO2~Airbox_NO2+log_O3+Wind_speed) # 3 explanatory variables.
summary(Model_3covariates.lm)
```

The following table shows my result:
Text in grey background is the information copied from question 4

Linear model with the following covariates	**R^2 adjusted**	**Strength and sign of the slopes**	**Significances of covariates**
2 covariates: Airbox_NO2 and log_O3	0.8023	Slope of Airbox_NO2: 1.25734 positive Slope of log_O3: -2.10471 negative	Airbox_NO2: <2e-16 *** log_O3: <2e-16 ***
1 covariate: Airbox_NO2	0.7775	Slope of Airbox_NO2: 1.37431 positive	Airbox_NO2: <2e-16 ***
1 covariate: log_O3	0.2556	Slope of log_O3: -6.210 negative	log_O3: <2e-16 ***
	Change in R^2 adjusted: Airbox_NO2 has a very big improvement on the linear model. Its contribution is much larger than log_O3. R^2 adj of Airbox_NO2 increases from 0.7775 to 0.8023, which is a small increase. R^2 adj of log_O3 increases from 0.2556 to 0.8023, which is a very big increase. Slope of Airbox_NO2 remains positive and increases a little bit. Slope of log_O3 remains negative. Its strength of linearity become less, changing from very negative (-6.2) to less negative (-2.1). There is no change in the significance of covariates.		

2 covariates: Airbox_NO2 and Wind_speed	0.7977	Slope of Airbox_NO2: 1.28036 positive Slope of Wind_speed: -0.10530 negative	Airbox_NO2: <2e-16 *** Wind_speed: 5.75e-15 ***
1 covariate: Airbox_NO2	0.7775	Slope of Airbox_NO2: 1.37431 positive	Airbox_NO2: <2e-16 ***
1 covariate: Wind_speed	0.1999	Slope of Wind_speed: -0.30690 negative	Wind_speed: <2e-16 ***
	Adjusted R^2 of Airbox_NO2 increases a little bit. Adjusted R^2 of Wind_speed increases significantly. This means that Airbox_NO2 has a very big improvement on the linear model. Its contribution is much larger than Wind_speed. Slope of Airbox_NO2 remains positive and increases by 0.1. Slope of Wind_speed changes from negative (-0.3) to less negative (-0.1). Its strength of linearity become less. There is no change in the significance of Airbox_NO2. The significance of Wind_speed becomes a little bit small, changing from 10^-16 to 10^-15.		
2 covariates: Log_O3 and Wind_speed	0.2797	Slope of log_O3: -4.57827 negative Slope of Wind_speed: -0.13940 negative	log_O3: <2e-16 *** Wind_speed: 1.32e-06 ***
1 covariate: Log_O3	0.2556	Slope of log_O3: -6.210 negative	log_O3: <2e-16 ***
1 covariate: Wind_speed	0.1999	Slope of Wind_speed: -0.30690 negative	Wind_speed: <2e-16 ***
	Change in adjusted R^2: Adjusted r^2 increases for both Log_O3 and Wind_speed. But the amount of increase is different: Difference between Log_O3 and the 2 covariates=0.2797-0.2556=0.0241 Difference between Wind_speed and the 2 covariates=0.2797-0.1999=0.0798 0.0798<0.0241. So Log_O3 has bigger contribution to the linear model, than the wind_speed. Slope of both two covariates remains negative. Both of their strength of linearity become less. Slope value of log_O3 increases from -6.2 to -4.5. Slope value of wind_speed increases from -0.3 to -0.1. In general, both Log_O3 and Wind_speed are negatively related with the response variable. There is no change in the significance of log_O3. The significance of Wind_speed becomes smaller, changing from 10^-16 to 10^-6. But the significance level (***) remains the same.		
All three covariates:	0.8059	Slope of Airbox_NO2: 1.24008 Slope of log_O3: -1.50395 Slope of Wind_speed: -0.05676	Airbox_NO2: <2e-16 *** log_O3: 2.3e-07 *** Wind_speed: 0.000377 ***

1 covariate: Airbox_NO2	0.7775	Slope of Airbox_NO2: 1.37431 positive	Airbox_NO2: <2e-16 ***
1 covariate: Log_O3	0.2556	Slope of log_O3: -6.210 negative	log_O3: <2e-16 ***
1 covariate: Wind_speed	0.1999	Slope of Wind_speed: -0.30690 negative	Wind_speed: <2e-16 ***
	R^2 adjusted (0.8059) is almost the same as R^2 adjusted in the AirboxNO2–logO3 model (0.8023). Slope of Airbox_NO2 remains positive. Its strength of linearity remains almost the same. Slope of log_O3 become less negative. Slope of Wind_speed also become less negative. This means that the strength of linearity of these 2 covariates become less than the single-covariate model. The significance level is all on the same level (***). But the significance level of wind_speed become smaller than the one in the 2-covaraite model.		

Which model do you prefer as the final model? I will choose the linear mode with all three covariates. Because its adjusted R^2 is the largest.

Which covariate would you drop when there is not enough money available to collect data on all covariates? I would drop the Wind_speed. Because, in the first two 2-covariate models, Airbox_NO2 has the biggest contribution to the linear model. Log_O3 and Wind_speed are both less correlated with the response variable. Then, the following evidences help me with my decision:

1) R^2 adjusted of Wind_speed is the smallest.
2) The slope of wind speed is very close to zero. The significance of wind speed keep becoming small when adding other variables to the model.
3) When we put Log_O3 and Wind_speed in one model, we are able to see which one of them makes the least contribution to the model.
   Difference between Wind_speed and the 2 covariates=0.0798
   Difference between Log_O3 and the 2 covariates=0.2797-0.2556=0.0241
   0.0798<0.0241 So wind_speed makes the least improvement to the linear model.
4) R^2 adjusted (0.8059) in 3-covariate model is almost the same as R^2 adjusted in the AirboxNO2–logO3 model (0.8023). This also proves that wind_speed makes the least improvement to the linear model.

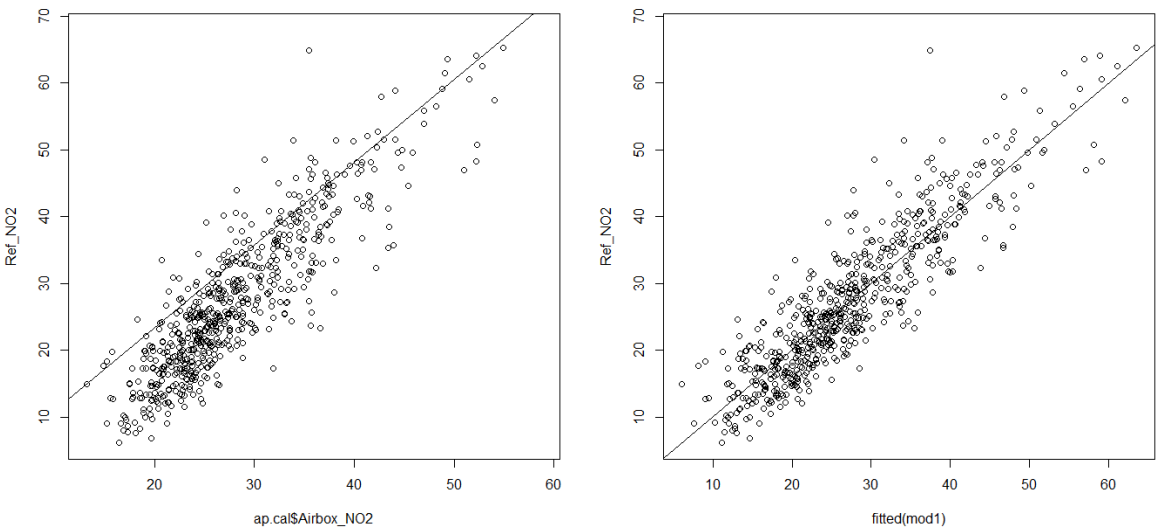
We continue with the model that takes all covariates. Make sure to set na.action to "na.exclude":

```
mod1 <- lm(Ref_NO2~Airbox_NO2+log_O3+Wind_speed, data=ap.cal,
na.action=na.exclude)
```

This will keep the NAs included in the list of fitted values, which will make sure the list of fitted values is of the same length as the list of input values. You need this to make scatterplots.

7) Make a scatterplot of the original data, with Airbox_NO2 on the x-axis and Ref_NO2 on the y-axis, as we did in Question 3. Now also make a scatterplot using `fitted(mod1)` on the x-axis and Ref_NO2 on the y-axis. `fitted(mod1)` includes the "fitted values" of the model. What does this mean? How are they computed? How does the scatterplot change when using the fitted values instead of the Airbox_NO2 values? Explain your answer. (10 marks)

Left image: a scatterplot of observed value, with Airbox_NO2 on the x-axis and Ref_NO2 on the y-axis. The black line is the line of the regression model.
Right image: a scatterplot using `fitted(mod1)` on the x-axis and observed Ref_NO2 on the y-axis. The black line is the line "vertical value=horizontal value"



R command:

```
mod1 <- lm(Ref_NO2~Airbox_NO2+log_O3+Wind_speed, data=ap.cal, na.action=na.exclude)
X11()
par(mfrow=c(1,2))
plot(ap.cal$Airbox_NO2, Ref_NO2)                          #original data
abline(lm(Ref_NO2~Airbox_NO2+log_O3+Wind_speed))   # line of the regression model
plot(fitted(mod1), Ref_NO2)        #fitted value of response variable
abline(a=0, b=1)                          # plot a line: vertical value=horizontal value
```

In the right image (fitted model), the x axis is the "fitted value" and y axis is the observed response value. "Fitted value" means the predicted response value (predicted Ref_NO2) calculated by the regression model. "Fitted value" is calculated by substituting the input values (observed Airbox_NO2, the observed log_O3 and the observed Wind_speed in the csv file) from the data into the model:

Fitted value
= Predicted Ref_NO2
=1.24008 * observed_Airbox_NO2+(-1.50395) * observed_log_O3+ (-0.05676) * observed_Wind_speed

The goal of this scatter plot is to see whether the model is good enough such that the predicted Y is close to the observed Y. We have known that residual= Observed Y – Predicted Y. There is a difference between Observed response value and Predicted response value. I have drawn a line: vertical value=1*horizontal value. This line means "Observed Y= Predicted Y". In this scatter plot, the points are very close to this line. The deviation above the line looks the same as the deviation underneath the line. So this model is a reasonable model. In general, scatter plot helps us visualize how much the difference is between Observed response value and Predicted response value.

The left image shows, with the observed Airbox_NO2, what the corresponded observed response variable (Ref_NO2) is. The straight line is the model line. The plots deviate from the model line. Because Airbox_NO2 is just one covariate among the three. The other two covariates are not shown

in this picture. It is reasonable that this scatterplot doesn't fully follow the line of the regression model.

8) We do not have reference NO2 measurement at all locations in the city, because this instrument is much more expensive than the Airbox measurement instrument. We have an Airbox at a different location in the city, where the NO2 measured according to the airbox is 22.5 µg m^{-3} at a moment in time. At the same moment, log_O3 = 3.95 and Wind_speed = 40. Predict the value of Ref_NO2 and compute the relevant 95% prediction interval. Continue using `mod1`. Explain how you have reached your answer. (10 marks)

R command:
new=data.frame(Airbox_NO2=22.5,log_O3=3.95, Wind_speed = 40)
predict(Model_3covariates.lm, new, interval="prediction", level=0.95)

result: fit=18.23016, lower bound=8.541107, upper bound=27.91921

This means that the predicted Ref_NO2 value is 18.23016 µg m^{-3}. It means that the predicted mean of the response value (Ref_NO2), with the given input (Airbox_NO2=22.5,log_O3=3.95, Wind_speed = 40), is 18.23016 µg m^{-3}.
With 95% prediction interval, the true Ref_NO2 value will locate in the range [8.541107, 27.91921]. It means that, if I randomly pick a value with the given input (Airbox_NO2=22.5,log_O3=3.95, Wind_speed = 40), the response variable (Ref_NO2) will give me a value in the range [8.541107, 27.91921] with 95% probability.

I calculated the predicted value by 2 different ways. The first way is using R command. It has shown above. The second way is a manual calculation:
Predicted Ref_NO2 = 1.24008* Airbox_NO2+(-1.50395)* log_O3+ (-0.05676)* Wind_speed
=1.24008* 22.5 µg m^{-3}+(-1.50395)* 3.95+ -0.05676* 40
=19.6907975 µg m^{-3}<18.23016 µg m^{-3}
18.23016 is more trustworthy than 19.69. The manually-calculated value is slightly larger than the R-calculated value. The reason is probably that two ways use different amount of decimal number, in the process of calculation. So, in conclusion, the predicted response value (Ref_NO2) is 18.23016 µg m^{-3}.

9) Imagine a client would like to use the Airbox measurement instrument for a project in the city of Beijing, China. The airbox measures NO2 and O3. From a local weather station, measurements of wind speed are available. There is no reference NO2 available. Can we use the regression model that we produced to predict the reference NO2 levels in Beijing? Why (not)? Explain your answer. (10 marks)

No, we cannot use this regression model to predict the reference NO2 levels in Beijing. The reasons are given here:
1) Different countries have different weather. The characteristics of wind speed in Eindhoven might be very different from the one in Beijing. Besides wind speed, the other meteorological variables in Eindhoven might also be different from the one in Beijing.
2) Some potential covariates that have been dropped in the Eindhoven case might should not be dropped in the Beijing case. For example, Airbox_O3, Mean_temp and Rel_hum.

3) Furthermore, variables other than meteorological variables (eg. the amount of people, the amount of industrial factories) might also have an impact on the concentration of NO2.
4) This regression model is based on the data only in Eindhoven. This regression model has not been tested in other cities in the Netherland. It is difficult to say whether this model is applicable in whole Netherland. So it is much harder to say whether it is applicable to Beijing.