Nitrogen Oxide Pollution in Switzerland

Klaudia Ludwisiak & Marika P. Paraskevopoulou

MSc. Computational Statistics and Machine Learning

Department of Computer Science, University College London, UK

STATG001: Statistical Models and Data Analysis

November 21, 2016

1 Introduction

A study was designed to examine the influence of three specific measurements on the concentration of nitrogen oxide, NOx, in the ambient air at a certain place in Switzerland close to a motorway. Table 1 contains a brief description of the response and the depended variables, which are repeated measurements over one year. Our main goal through this assignment is to provide efficient statistical data analysis of the above data set, and to conclude by recommending and optimal fitted model.

Variable	Description
nox	NOx concentration in ambient air **(response variable)**
noxem	Sum of NOx emission of cars on this motorway (not known units)
ws	Wind speed
humidity	Absolute humidity in the air

Table 1: Variables description.

2 Explanatory data analysis

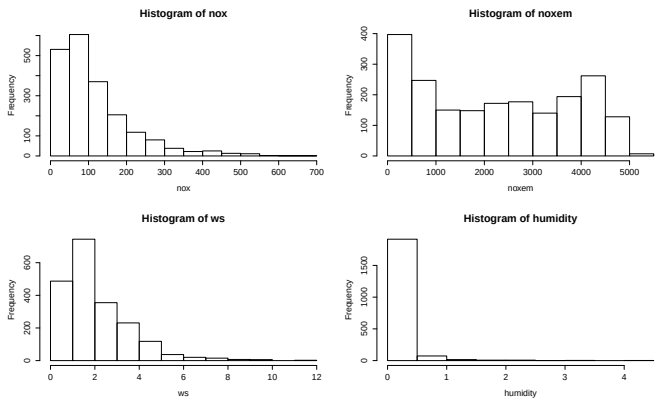


Figure 1: Histograms for all variables.

In this section our goal is to investigate the way that our variables are distributed, to look for asymmetries, possible relationships, dependencies and transformations. We will attempt to address all of those, by producing some measures of the central tendency and dispersion and explanatory plots. To begin with,in Figure 1 we represent the histograms for all of the four variables. It is clear that the response variable **nox** exhibits right asymmetry as does the wind speed (ws), a log transformation will possibly be a good way to correct it. Whereas for the variable noxem, it should be noted that the lack of knowledge about the units can be problematic, as these could include a transformation already. However, further investigation reveals that a possible relationship between this variable and the response is also attained by taking a log-transform.

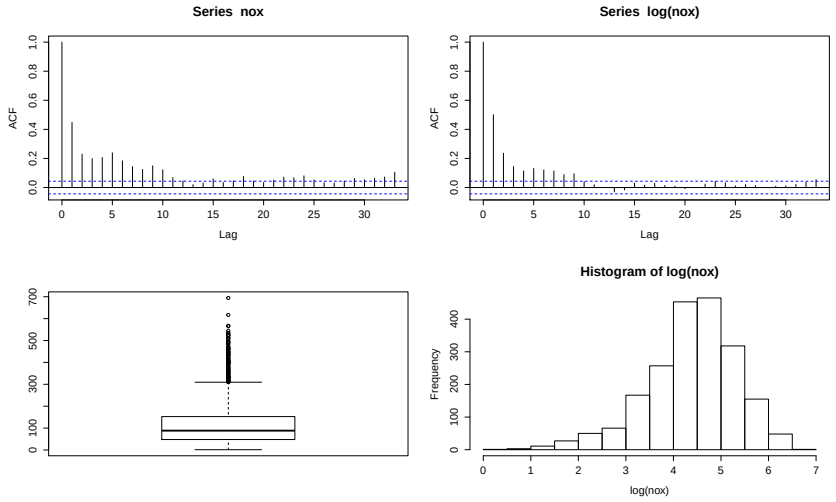


Figure 2: ACF, box-plot for the initial response and ACF and histogram for the log response **nox**.

Secondly, as mention above, the data set consists of repeated measurements over time, so it is logical to expect the existence of dependencies. This is visible from Figure 2, where the ACF plot for the original response indicates dependencies of up to lag 30. After the log transformation, the variable distribution becomes more symmetrical, and the dependence is reduced to lag 9. Still, the fact that there is dependence is problematic, as is the existence of possible extreme values. It should be noted that different other transformations were tested on the data, for instance taking the mean and mean of log, but all indicated that the logarithm is the best option. Finally, the dataset was examined for seasonal dependencies, and it was discovered that **nox** exhibits seasonal variation. However, lack of knowledge about the date that the measurements commenced renders it impossible to we make precise inferences (see R output).

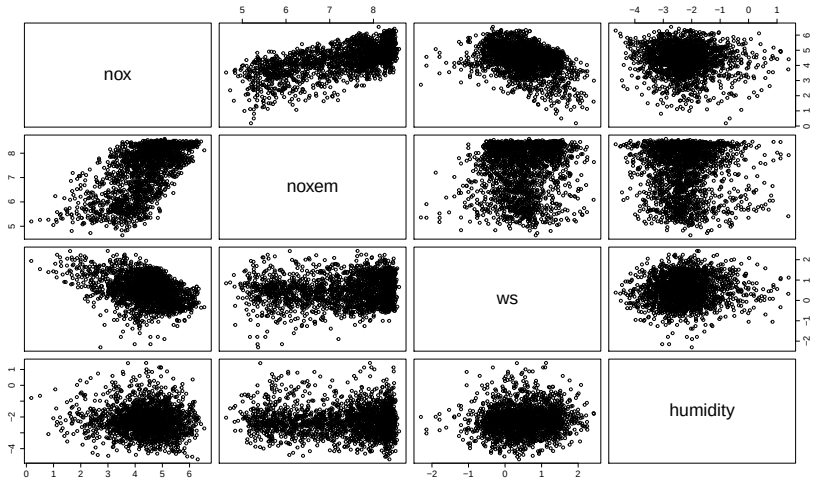


Figure 3: Scatter- plot matrix for the log data.

After having tried different transformations, we conclude that the best way to further investigate the relationship between the response and dependent variables is to log transform all of them. From Figure 3 it can be seen that there is a clear upward trend between nox and noxem. Furthermore, there exists a negative relationship between nox and ws, whereas it can be assumed that humidity and nox are not related. Additionally, there is no multicollinearity between our dependent variables. Finally, in Table 2 we can see the mean and sd of the logged data, the high variability is worth mentioning.

	nox	noxem	ws	humidity
mean (sd)	4.38(0.95)	7.34(1.01)	0.51(0.68)	-2.24(0.87)

Table 2: Mean and standard deviation for the log data.

3 Possible Models

In this section three different models will be fitted; firstly, a full **linear regression model**, where all the dependent variables are included. Assuming normal residuals, the model can be written as;

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i \sim N(0, \sigma^2) \tag{1}$$

Secondly, a nested model is fitted, where the humidity variable is excluded since it is not statically significant. The last fitted model is a **robust regression model**. In all three cases log-transformed data was used.

3.1 Model 1

As mentioned above, the first model is the full model- including all the dependent variables under a log transformation. Table 3 below, summarizes the results of this first fitted model, where from the R-output: $R^2 = 62.74\%$, $R^2_{adj} = 62.69\%$ and residual (ϵ_i) standard error is 0.5776.

	Estimate(S.E)	p-value
intercept	-0.07(0.098)	0.501
noxem	0.65(0.0189)	≤ 0.01
ws	-0.65(0.0128)	≤ 0.01
humidity	-0.02(0.0148)	0.286

Table 3: Summary table for the **Model 1**.

Closer examination of the results in Table 3 reveals that the intercept (β_0 in (1)) and the humidity variable are not statistically significant. Therefore, it is senibel to introduce a simplified model (Model 2) excluding the humidity variable . Below we can see the mathematical expression of the Model 1.

$$\log nox_i = -0.07 - 0.65 \log(ws_i) + 0.65 \log(noxem_i) - 0.02 \log(humidity_i) + \epsilon_i$$

3.2 Model 2

Our second model comprises the log response of all but the humidity variable. Table 4 below provides a summary of the estimators. This time, the residual standard error is 0.5777, whereas $R^2 = 62.72\%$ and $R^2_{adj} = 62.68\%$. It can be seen that the residual standard error, the coefficient of determination and its adjusted version only change slightly between Model 1 and Model 2. This is evidence that humidity does not impact measured NOx levels, which is further supported by the F-test comparing the full model with the nested model. Therefore, we conclude that the humidity variable does not contribute to improved model predictions and will be excluded. Finally, as was the case in Model 1, in the Model 2

	Estimate(S.E)	p-value
intercept	-0.03(0.0939)	0.709
noxem	0.65(0.0189)	≤ 0.01
ws	-0.65(0.0127)	≤ 0.01

Table 4: Summary table for the **Model 2**.

the intercept is not statistically significant, if excluded from the analysis both the adjusted and multiple coefficient determination (see R-output) become very high. However, from the perspective of interpreting the model, the decision of excluding the constant is not very wise as information is lost.

3.3 Model 3

Finally as the last model **Model 3**, robust regression is performed on Model 2. We use the robust regression since from Section 2 we suspect the existence of some extreme- influential points. Although the robust model did not indicate any outliers, the residuals standard error is slightly lower 0.537, $R^2 = 62.08\%$ and $R^2_{adj} = 62.00\%$, we choose to use the simple linear model in **Model 2**.

4 Recommended fitted model

In this section the selected model will be examined more closely, verification of the model's assumptions and an interpretation of results will be provided. According to our discussion in Section 3, we put forward Model 2 to be "best" model, this is given by;

$$\log nox_i = -0.03 + 0.65 \log(noxem_i) - 0.65 \log(ws_i) + \epsilon_i \sim N(0, 0.5777^2) \tag{2}$$

$$nox_i = e^{-0.03+0.65 \log(noxem_i)-0.65 \log(ws_i)+\epsilon_i}, \epsilon_i \sim N(0, 0.5777^2) \tag{3}$$

For the interpretation of the model, we can say that as the NOx emissions of the cars are increasing, the air population is increasing dramatically, whereas, high wind speed can very significantly decrease the total amount of NOx in the air.

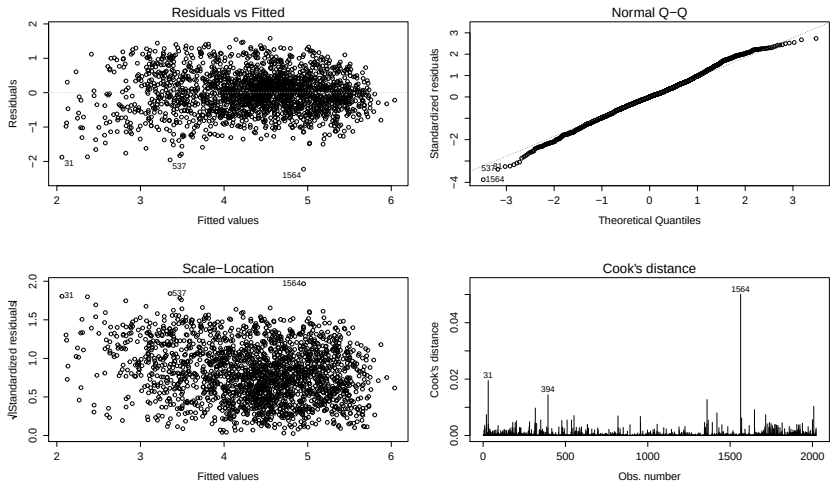


Figure 4: Diagnostic plots for the **Model 2**.

As we can see from Figure 4 and 5, the model's fit looks satisfactory enough since there is no clear trend in the residual vs the fitted values plot, and most of the data points lie between (-2,2). The normal Q-Q plot also reveals a sufficient fit, although some deviation in the upper and lower tails can be observed. The Cook-distances indicate, three extreme values, the detail of which is shown in Table 5. These extreme values are not considered influential since the robust model is Section 3 didn't indicate any extreme value. Further inquiry into the nature of these 'exceptions' would be desirable before any decision to discard them. Finally, Figure 5 depicts the residuals plotted against the explanatory variables, further confirming there is no significant trend and thus indicating a good model fit.

	nox	noxem	ws	humidity
no.31	0.18	5.19	1.94	-0.79
no.394	0.50	5.26	1.54	-0.66
no.1564	2.72	5.36	-2.30	-1.99

Table 5: Outliers (log data) for **Model 2**.

Data points listed in Table 5 have been identified as out-layers from the examination of Cook's distance. Hence, they may be erroneous and require further thought and investigation. Referring back

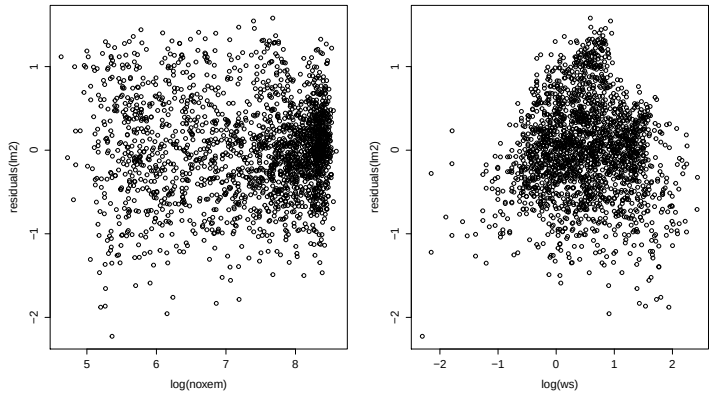


Figure 5: Residuals vs the explanatory variables for **Model 2**.

to Table 2 (Section 2) one can compare the 'suspicious' results with mean values for the variables. NOx levels of data points 31 and 394 are significantly below the mean levels, whereas point 1564 is also on the low side (within the 1st quantile). In all three cases NOx emissions are at similar low levels; below the mean and within the 1st quantile. The wind speed in case of points 31 and 394 is significantly above the mean value, whereas in case of point 1564 it is significantly below the mean and is the lowest value in the entire dataset. Thinking back to the physical meaning of the response variable and the underlying exploratory variables, results for points 31 and 394 are in fact quite logical; at high wind speeds one would expect to observe lower NOx concentrations, especially if emissions are also low. The result 1564 depicts a case with low wind speed and high humidity, intuitively this should give rise to higher NOx concretions as air particles are comparatively still and heavy with water particles. Therefore, physical explanations for the low NOx concentrations at this data points are less obvious and it could be that some measurement error occurred. Never the less, the data points flagged by the Cook distance analysis could be real measurements and are explicable physically. Therefore, they will not be excluded from the model. In order to take the deacons to exclude some or all of these results further investigation with parties involved in collecting the data would be desirable.

5 Concluding remarks

The linear models assumptions are satisfied as there is normality of the residuals, independence between the residuals and the variables as well as a stable variance. It is important however, to bear in mind that there exists a correlation between the residuals themselves(visible in the R-output ACF plot for the residuals). This is hardly surprising since, the response variable exhibits correlation even after transformation. This correlation cannot be ignored as it is part of the data, but it does introduce instability in the predictions. Therefore, extreme care needs to be taken if the model is to be used for prediction. Alternatively, further investigation of the underlying data and refinement of the model could be undertaken.

R code

```
##STATISTICS G001 In- Course Assesment
# Question 1
# Klaudia Ludwisiak & Marika Paraskevopoulou 21.11.2016

################################################################################
#  SETUP

install.packages("robustbase")
library(robustbase)

# load the data and check it
data=read.table("emissionssw.dat",header=TRUE)
head(data)

# set the names of the new variables
nox=data[,1]
noxem=data[,2]
ws=data[,3]
humidity= data[,4]

################################################################################
##  Expanatory data analysis

# Initial observations:
# 1. Measurments taken repeatedly for one year, thus we can expect dependence
# 2. There is dependence up to lag(10), thus up to ten previous results
# have impact on current result

# Plot the data to see their structure
par(mfrow=c(2,2))
hist(ws)
hist(noxem)
hist(nox)
hist(humidity)

# 3. ws and nox exponential relationship thus use log tranform
par(mfrow=c(2,2))
acf(nox)
acf(log(nox)) # notice reduction in Lag upon transformation ???

# 4. Visible left assymetry, again suggestive of transform requirement
hist(nox)

# 5. The logarithm is a reasonable transformation as it yelds a distribution more
# resemblent of the normal distribution, however it is not perfect
# with a longer tail of low log(nox) values
hist(log(nox))

# 6. Try to transform to the "center" the data by taking x_i-mean(x),
# with no visible improvement.
par(mfrow=c(1,2))
hist((nox-mean(nox))) #bad
hist((log(nox)-mean(log(nox)))) #bad

# 7. Further explanatory analysis in the response variable confirm
# the assymetry in the original response.
# Adiitionaly, the presence of possible outlayers, requiring further
```

```
# examination becomes apparent.
par(mfrow=c(1,2))
boxplot(nox)
boxplot(log(nox))

# Explanatory analysis for the full data set:
summary(data)
summary(log(data))

# Matrix plot for looking for multicollinearity:
# 8. It seems that there is no linear relationship between the dependent variables
plot(data)
plot(log(data))

# 9. A possitive trend between nox and noxem exists
# 10. A Negative trend between nox and ws exists
# 11. No visible trend between nox and humidity
# 12. The log transform gives better data, thus it will be used.

################################################################################
##  Linear Regression

# Start from the full model (with all 3 exploratory variables included)
# fit model for data as0is and for the log transformed data:
# lm1i=lm(log(nox)~ws+noxem+humidity)
lm1=lm(nox~ws+noxem+humidity,data=log(data))

# p_value > 5% for the humidity variable, therefore:
# fit the second model without the humidity
# lm2i=lm(log(nox)~noxem+ws)        #Model 2i is acceptable
lm2=lm(nox~noxem+ws,data=log(data)) #Model 2, is better
lm2c=lm(nox~noxem+ws+0,data=log(data)) #Model without intercept
summary(lm2c) #high R^2- for interpertation reason we won't choose it
par(mfrow=c(2,2)) #diagnostics for the model without the constant
plot(lm2c,which=1:4,add.smooth=FALSE,ask=FALSE) #Model 2 qq-plot better
par(mfrow=c(1,2))
plot(log(noxem),residuals(lm2c))
plot(log(ws),residuals(lm2c))

# Initailse single explanatory variabel models, further examinig significance:
lm3=lm(nox~ws,data=log(data))
lm4=lm(nox~noxem,data=log(data))
# observe both noxem and ws are significant

# ANOVA, Perofrm F-test for model selection between following pairs of models:
# model 2 vs  model 1, model 3 and model 4 respectively

anova(lm2,lm1)
# low F value and high P thus, humidity is not significant and can be taken out

anova(lm2,lm3)
anova(lm2,lm4)
# both comparisons yeld high F values and low P values indicating the significance
# of noxem and ws variables. Model 3 has the highest F value and therefore is better

################################################################################
##  Select Model 2 & Further Analysis
```

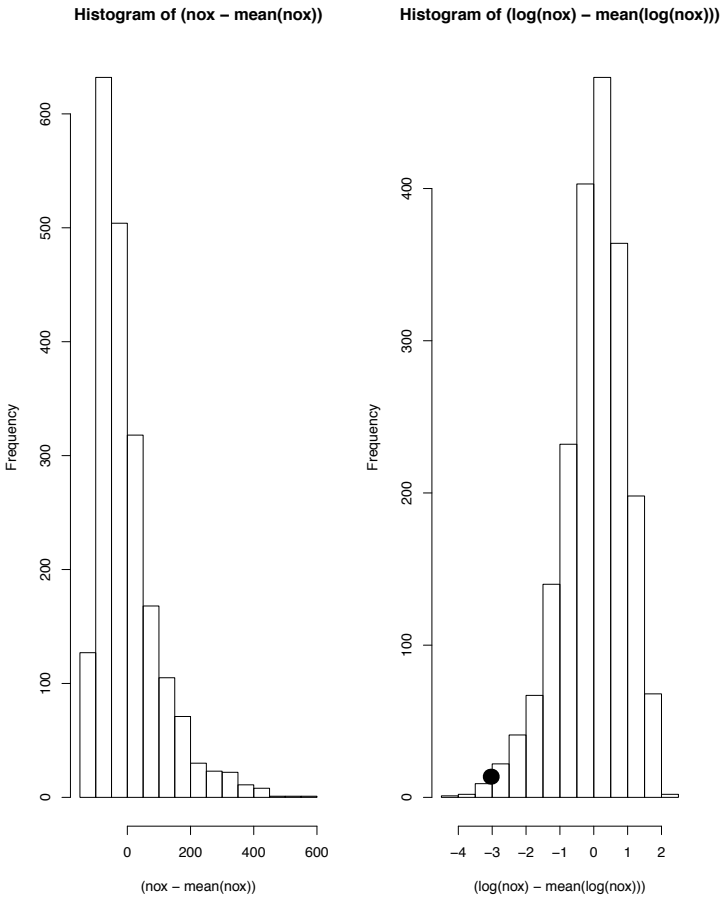
```
# Summarise the results from the linear regression
# Observe that R^2 is sattisfactory but not perfect
summary(lm2)

# Derive the pearson, deviance and standrardized residuals
resdev=rstandard(lm2,type='deviance')
respear=rstandard(lm2,type='pearson')

# Derive the fitted values
fit=lm2$fitted.values

# Plot fitted values vs stand.residuals and personian residuals:
par(mfrow=c(1,2))
plot(fit,lm2$residuals)
plot(fit,respear)
acf(respear) #correlation
# Residuals fitting gives reasonable results; data is scattered wih no visible pattern.

##############################################################################################
##  Plot leverage points (given by the elements of the hat matrix) and the cook distances:

# par(mfrow=c(1,2))
# plot(hatvalues(lm2),main="Leverage Values Plot of the Fitted Model",
ylab="leverage values",type="n")
# text(hatvalues(lm2), labels=1:2022, cex= 0.7, pos=3)
# plot(cooks.distance(lm2),main="Cook's Distances Plot of the Fitted Model",
ylab="cook's distance values",type="n" )
# text(cooks.distance(lm2), labels=1:2022, cex= 0.7, pos=3)
# We expect 5% of the cook distances to be outside of the limit 2p/n
# However we find this assumption is violated.
# Although this method is not valid way to check the cook diastances,
# and is controversional,
# this result is resaon enough to perform robust regression:
d1=cooks.distance(lm2)
5*2022/100
length(d1[d1>4/2022])

##############################################################################################
## Generate the diagnostics plots all in one:

par(mfrow=c(2,2))
plot(lm2,which=1:4,add.smooth=FALSE,ask=FALSE)
par(mfrow=c(1,2))
plot(log(noxem),residuals(lm2))
plot(log(ws),residuals(lm2))
# Observations:
# From the qq plot divergence can be seen at the lower tail,
# supported by original exploratory analysis
# Since we have a large data set, the qq-plot and any normality test became
# very sensitive to extreme values
# Therefore bespite this divergence teh overal result are still a good fit.

##############################################################################################
##  Robust regression in Model 2

lmrob2= lmrob (log(nox) ~ log(noxem) + log(ws))
summary(lmrob2)
# Observations:
```

```
# robust regression does not find extreme values
# The RSS's are close, so are the estimates
# Hence the linear model, is enough.

################################################################################
##  Further analysis for Part1; looking for seasonality

# Use for loops for funding the mean daily concertations
y=nox
ynew1=rep(0,337)
for (i in seq(1, 2022, by= 6)) {
  ynew1[i]=(y[i]+y[i+1]+y[i+2]+y[i+3]+y[i+4]+y[i+5])/6}
ynew2=ynew1[ynew1!=0]
ynew=ynew2[!is.na(ynew2)]

# the acf plot looks better
acf(log(ynew))
hist(log(ynew))

# do the same for the dependent variables
x=noxem
xnew=rep(0,337)

for (i in seq(1, 2022, by= 6)) {
  xnew[i]=(x[i]+x[i+1]+x[i+2]+x[i+3]+x[i+4]+x[i+5])/6}
xnew2=xnew[xnew!=0]
xnew1=xnew2[!is.na(xnew2)]
x=ws
xnew=rep(0,337)
for (i in seq(1, 2022, by= 6)) {
 xnew[i]=(x[i]+x[i+1]+x[i+2]+x[i+3]+x[i+4]+x[i+5])/6}
xnew3=xnew[xnew!=0]
xnew2=xnew3[!is.na(xnew3)]
x=humidity
xnew=rep(0,337)

for (i in seq(1, 2022, by= 6)) {
 xnew[i]=(x[i]+x[i+1]+x[i+2]+x[i+3]+x[i+4]+x[i+5])/6}
xnew4=xnew[xnew!=0]
xnew3=xnew4[!is.na(xnew4)]
lmn1=lm(log(ynew)~xnew1+xnew2)
summary(lmn1)
plot(lmn1,which=1:4,add.smooth=FALSE,ask=FALSE)

#Observe that R^2 the same
acf(residuals(lmn1))

#dependence in residuals
#explore seasonality
#plot the mean daily measurments:
par(mfrow=c(1,1))
plot(ynew,type="l",ylab="NOx",xlab="Days",main="Mean daily NOx on air")
abline(v=100,col=2)
abline(v=250,col=2)
#for the mean daily NOx concentration we can see that
#there is a trend approximate between the days 1-100 very high concentr
#100-250 low and 250-... again high
#we explore it better if we knew the season or the
```

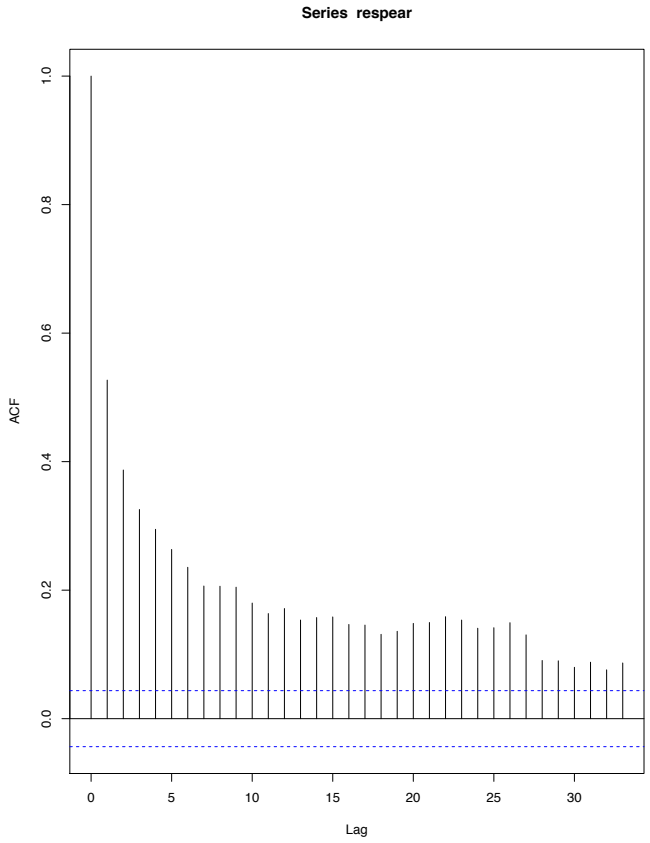
```
#the date the measurments start
#the other looks reasonably ok
plot(xnew1,type="l",ylab="NOx of cars",xlab="Days")
plot(xnew2,type="l",ylab="NOx of cars",xlab="Days")

##############################################################################################
#Further analysis Part2: extreme value theory in the tails of the data set,
#more informations about the data set, i.e: when the measurment start
#LD Analysis, explore better the dependences
#time series anaysis, MA or AR model
```

Appendix - R output

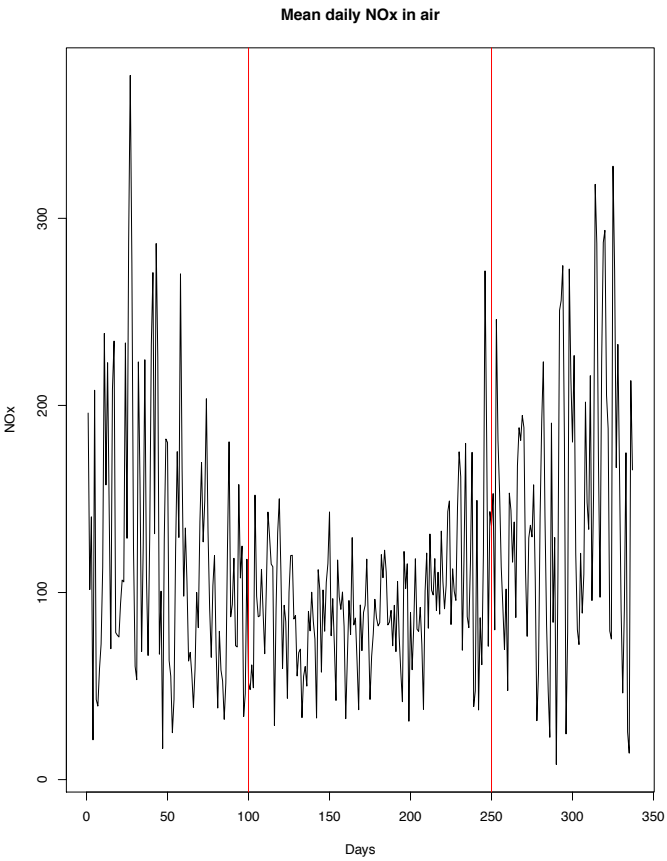
Below is a non exhaustive sample of the most relevant figures. A complete list of figures can be obtained by running the R code.



Exploratory data analysis: testing different transformations for the response variable



Model 2 : ACF plot for the Pearsonian Residuals, it cab be seen that there is lag up to 30.



Exploratory data analysis: Exploring the seasonality in data- two season clearly visible

Report: Summary of Nitrogen Oxide Pollution Study

Introduction

A study was designed to investigate the impact of car nitrogen oxide emissions, wind speed and air humidity on nitrogen oxide (NOx) air pollution levels at a specific site near a motorway in Switzerland. The study was conducted using data collected by the environmental research institute with the aim to make quantitative estimates of the strength of influence between the above mentioned variables. Data was collected 5-6 times a day for a period of one year. A linear regression model was fitted to the data in an attempt to provide a basis for inference.

This report outlines the observations made form the data set and concludes the existence of a relationship between some of the variables in question. The report goes on to suggest methods for improvement and strengthening of the quantitative model.

Observations

Initial data exploration revealed results are summarised in the Table 1 and Figure 1 below.

Variable:	Average value	Minimum value	Maximum value
NOx concentration [ppd]	116.0	1.2	694.5
NOx emissions [units unknown]	2250.7	102.5	5362.1
Wind Speed [m/s]	2.1	0.1	11.4
Absolute air humidity [g/kg]	0.17	0	4.09

Table 1. Summary of variables used in study

Figure 1 depicts all data points for all four variables plotted one agains the other. This results form the basis for the intuition that NOx concentration levels depended on both NOx emissions and wind speeds. Unsurprisingly, in general, as NOx emissions increase so does NOx air pollution. Similarly, high winds yield lower NOx pollution. It can also be noted that NOx emissions, wind speed and humidity do

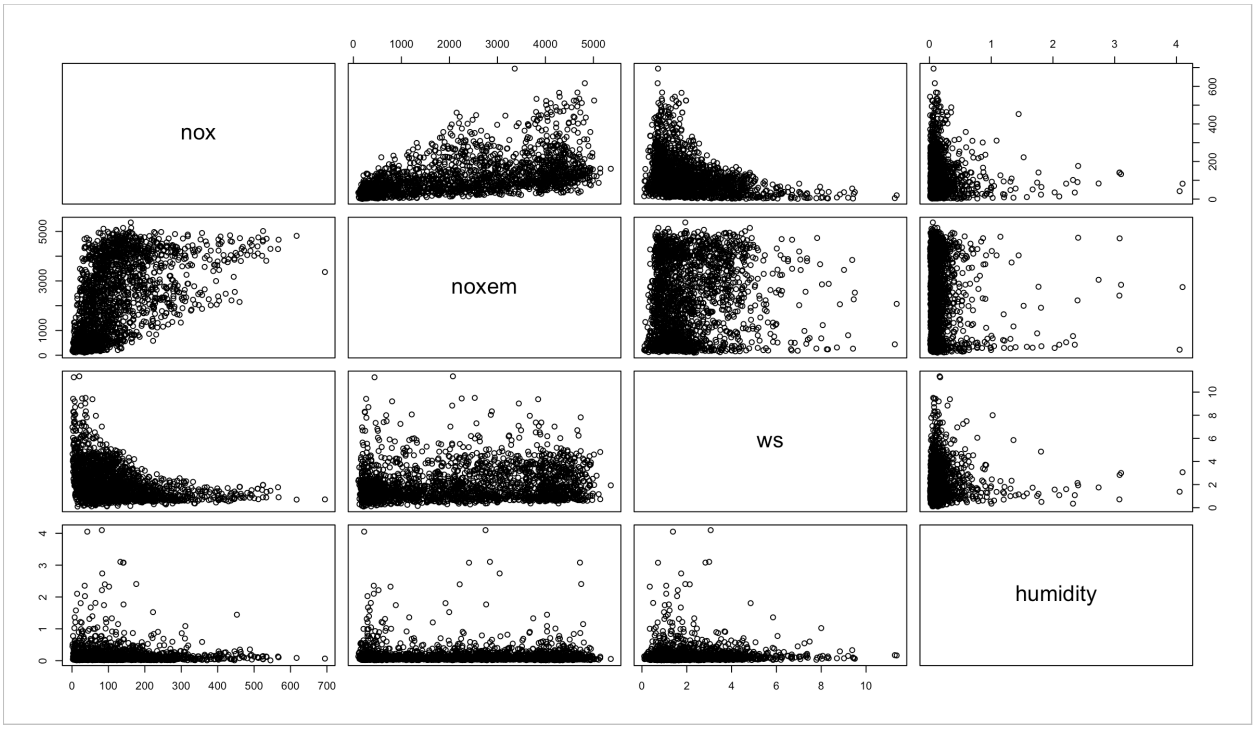


Figure 1. Scatterplot providing an exploratory view of the relationship between variables. Where: nox is the NOx concentration in air, noxem denotes NOx emissions and ws is the wind speed.

not obviously depend on one another and can therefore be called independent. As opposed to the other variables, humidity, with overall low concentrations through the year, does not have a significant impact on NOx air pollution. Humidity remains close to zero in almost all instances of the collected data and NOx concentration values range widely regardless of the observed humidity. Staying true to reality, the statistical model fitted to the data excludes the influence of humidity on NOx levels. Instead, in the formulation of the model emphasis is placed on uncovering a linear relationship between NOx concentrations, NOx car exhaust emissions and wind speeds. The data set is large and robust enough to be give strong and significant evidence for the existence of such a relationship. For completeness the following formula quantitatively describes the above discussed relationship:

$$nox_i = e^{-0.03+0.65\log(noxem_i)-0.65\log(ws_i)+\epsilon_i}$$

where the suffix *i* indicates a particular instance of the variable, *e* is the exponential function and the term *ei* is the error term. This model is advantageous in its relative simplicity and represents the best attainable fit for the data provided. However, it should be acknowledged that the liner model is not perfect and has shortcomings as it does not account for time-dependance. From the data however, it was found that time dependence between NOx concentrations exists, whereby at any given time instance NOx levels depend on previous NOx levels, up to even 30 measurements back. Furthermore, end users of the model should be made aware of the seasonal dependance between NOx levels and the time of year, where independently of wind and emissions, NOx levels are higher in one portion of the year and lower in the other. It could well be that temperature, obviously related to the time of the year, plays a part in the observed NOx concentrations. In light of the above issues the model should be used with caution and should not be relied upon for exact predictions.

Scope for study expansion and model refinement

As explained in the section above the proposed model is the best that could be fitted given limited information. However, the model could be refined if more was known about the existing data set and if new data could be made available. Firstly, it would be beneficial to obtain information about the exact dates and times the data measurements were taken. This would enable further examination of time dependance correlation and seasonal effects. With the idea that different models could be fitted for the distinct seasons, enabling improved prediction. It would also become feasible to further examine time dependance of NOx concentrations, perhaps discovering daily patterns of air pollution and emission levels. Secondly, the formulation and testing of the model could be aided by adding other relevant data variables , for example temperature data. As of now, the model is only useful to make predictions about the local area where NOx measurements were taken. If the model was to be used to make not only local but also more global predictions, then the inclusion of more than one data collection location should be considered. Air pollution can vary depending on local topography, the motorway's layout and numerous other factors. Therefore, analysing data from multiple locations could yield new insights, a more sophisticated and scalable model.

Statistics G001 ICA PART 2

Solution written by Klaudia Ludwisiak

Question:
Show that:
$$\hat{\beta} - \hat{\beta}_{(i)} = C^{-1} x_i \frac{e_i}{1-h_{ii}}$$

Problem definition:

- **Influence vector:** $\hat{\beta} - \hat{\beta}_{(i)}$, where $\hat{\beta}$ is the least squares estimator (LSE) for the full data and $\hat{\beta}_{(i)}$ denotes LSE excluding the ith variable.
- **Design matrices:** X & $X_{(i)}$, where $X_{(i)}$ is the design matrix without the ith variable.
- Defining: $C_{(i)}^{-1} = [X_{(i)}^T X_{(i)}]^{-1}$ and $C^{-1} = [X^T X]^{-1}$ using 'Hint' it follows:

$$[X_{(i)}^T X_{(i)}]^{-1} = [X^T X]^{-1} + \frac{[X^T X]^{-1} x_{(i)} x_{(i)}^T [X^T X]^{-1}}{1 - x_{(i)}^T C^{-1} x_i} \quad (1)$$

- Defining the Hat matrix: $H = X^T [X^T X]^{-1} X$ and it's diagonal element: $h_{ii} = x_i^T [X^T X]^{-1} x_i$, Using fact that H is symmetric and independent, equation (1) can be written as:

$$[X_{(i)}^T X_{(i)}]^{-1} = [X^T X]^{-1} + \frac{[X^T X]^{-1} x_{(i)} x_{(i)}^T [X^T X]^{-1}}{1 - h_{ii}} \quad (2)$$

- Using fact that LSE: $\hat{\beta} = (X^T X)^{-1} X^T y$
And taking out the ith cases for $\hat{\beta}_{(i)} = [X_{(i)}^T X_{(i)}]^{-1} [X^T y - x_{(i)} y_{(i)}]$ (3)

- Substituting EQ.(2) into EQ.(3):

$$\hat{\beta}_{(i)} = \left[[X^T X]^{-1} + \frac{[X^T X]^{-1} x_{(i)} x_{(i)}^T [X^T X]^{-1}}{1 - h_{ii}} \right] [X^T y - x_{(i)} y_{(i)}]$$

$$\hat{\beta}_{(i)} = \underbrace{[X^T X]^{-1} X^T y}_{\hat{\beta}} - \underbrace{[X^T X]^{-1}}_{C^{-1}} x_i y_i + \frac{\overbrace{[X^T X]^{-1}}^{C^{-1}} x_{(i)} x_{(i)}^T [X^T X]^{-1} X^T y}{1 - h_{ii}} - \frac{\overbrace{[X^T X]^{-1}}^{C^{-1}} x_i x_i^T [X^T X]^{-1} x_i y_i}{1 - h_{ii}}$$

so

$$\hat{\beta} - \hat{\beta}_{(i)} = C^{-1} x_i \left[\frac{y_i (1 - h_{ii}) - x_i^T [X^T X]^{-1} X^T y + \overbrace{x_i^T [X^T X]^{-1} x_i}^{h_{ii}} y_i}{1 - h_{ii}} \right]$$

$$\hat{\beta} - \hat{\beta}_{(i)} = C^{-1} x_i \left[\frac{y_i - y_i h_{ii} - x_i^T \beta + h_{ii} y_i}{1 - h_{ii}} \right] = C^{-1} x_i \left[\frac{y_i - x_i^T \hat{\beta}}{1 - h_{ii}} \right]$$

and since $e_i = y_i - x_i^T \hat{\beta}$ it follows that:

$$\hat{\beta} - \hat{\beta}_i = C^{-1} x_i \left[\frac{e_i}{1 - h_{ii}} \right] \blacksquare$$