Statistical Models and Data Analysis ICA 2022
(STAT0028)

November 2022

1 Question 1

1.1 Introduction

This report examines a data set consisting of 2022 measurements of Nitrogen Oxide (NOx) pollution content in the ambient air, local car NOx emissions, wind speed and humidity, in Switzerland over the course of a year. In this report we aim to find a suitable fitted statistical model which best describes the data such that accurate quantitative statements on the response variable nox can be made.

1.2 Explanatory Data Analysis

We note that the response variable and all three of the explanatory variables are quantitative and continuous. As a preliminary to regression modelling, pairwise plots between the predictor variables and response can be used to assess linearity, outliers and homoscedasticity.

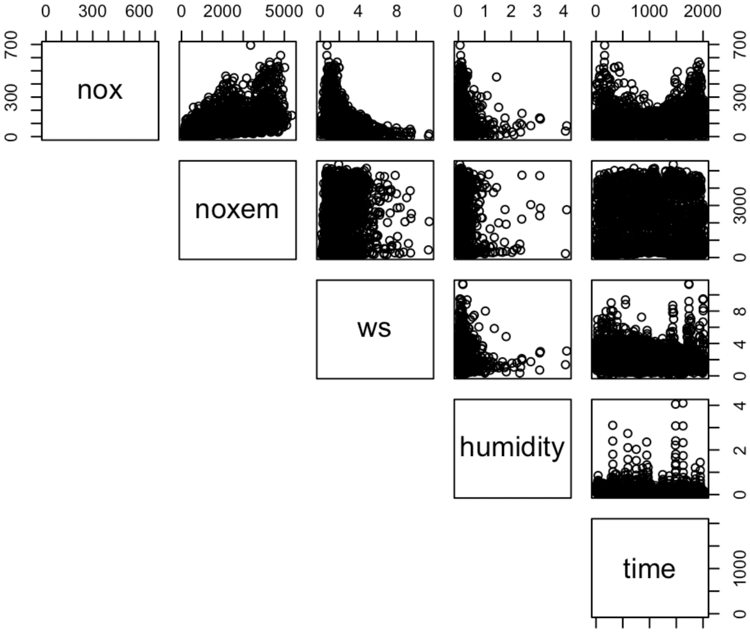


Figure 1: Matrix plot for emissions data.

From Figure 1, we observe some weak linear relationship exists between noxem and nox. Otherwise, we see non-linear relationships between the rest of the predictor variables and the response and also no collinearity between the pairwise predictor variables. Note that we mustn't over interpret these plots at this stage as the single predictors do not always give the full impression. We also included a time variable into plot under the assumption that each observation was made at equally spaced times throughout the year. We noticed some quadratic behaviour between time and the response.

1.3 Model Selection

In this section we fit and assess three different statistical models.

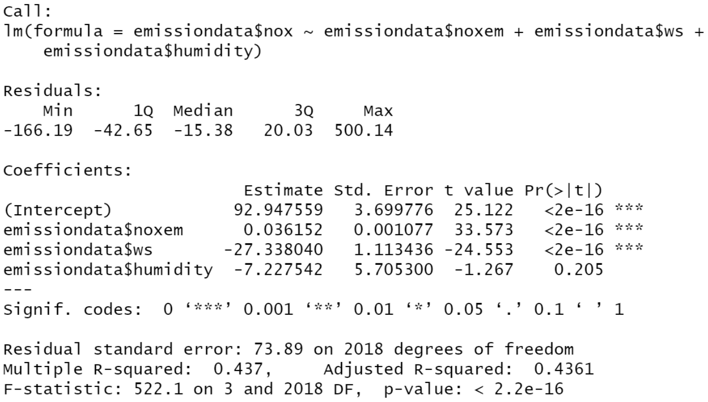
1.3.1 Model 1: Standard Multiple Linear Regression

As a baseline, we decided to use the simplest statistical model including all variables:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

where Y_i denotes, for the ith observation, the response of NOx pollution content in the ambient air, corresponding to the values of x_{i1}, x_{i2}, x_{i3}, of the explanatory variables x_1 (Noxem), x_2 (Wind Speed) and x_3 (Humidity). e_i is the 'error' associated with the ith observation, where $e_i \sim \mathcal{N}\left(0, \sigma^2\right)$ independently, $i = 1, \ldots, 2022$ and β_0, β_1, β_2 and β_3 are unknown parameters to be estimated from the data.

Our estimates of the regression coefficients, $\hat{\beta_0}, \hat{\beta_1}$ and $\hat{\beta_2}$ were computed in R by calling the lm function on the chosen variables. The relevant statistics for the fitting of Model 1 are returned in Figure 2. As we can see, the R^2 value is 0.437 which indicates a moderate explanatory strength. F-test: p-value is very small which indicates dependence of nox on at least one of the explanatory variables. Looking at the p-values in the $Pr(>|t|)$ column, the small p-values in the 2nd and 3rd rows indicate evidence for stronger dependence of nox on noxem and ws instead of humidity. The p-value for humidity is sufficiently large to suggest that humidity could be omitted from the model given that the other two explanatory variables are in the model. This is taking into consideration that we are using a 95% confidence interval.



```
Call:
lm(formula = emissiondata$nox ~ emissiondata$noxem + emissiondata$ws +
    emissiondata$humidity)

Residuals:
    Min      1Q  Median      3Q     Max
-166.19  -42.65  -15.38   20.03  500.14

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           92.947559   3.699776  25.122   <2e-16 ***
emissiondata$noxem     0.036152   0.001077  33.573   <2e-16 ***
emissiondata$ws      -27.338040   1.113436 -24.553   <2e-16 ***
emissiondata$humidity  -7.227542   5.705300  -1.267    0.205
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73.89 on 2018 degrees of freedom
Multiple R-squared:  0.437,     Adjusted R-squared:  0.4361
F-statistic: 522.1 on 3 and 2018 DF,  p-value: < 2.2e-16
```

Figure 2: Model 1 fit.

In Figure 3 we return the summary statistics for residuals. From the Residuals vs Fitted values plot, we do not observe a randomly scattered plot, which reveals issues in the formulated model assumptions (non-linearity, heteroscedasticity). The plot reveals that the mean residual changes with the fitted values and also the spread of the residuals is increasing too (non constant variance). This was expected as our explanatory data analysis revealed no linear relationships between the predictors and the response, indicating a violation of the Linearity assumption of the multiple linear regression model. Similarly, the Normal Probability plot of the residuals (Q-Q plot) should roughly look like a straight line if the errors are distributed normally, however it does not, which indicates deviations from normality (heavy tails), including outliers. The Scale-Location plot should also ideally be a horizontal line with equally (randomly) spread points. Here this is not true also indicating heteroscedasticity. Lastly, the Cook's plot identifies the most influential observations. We notice some points with large cooks distance's which indicates some leverage points and outliers.

Looking at Figure 4, we also see heteroscedasticity along some of the explanatory variables. Taking this first model into account, we will attempt to deal with these problems (heteroscedasticity and non-linearity) in our next model using a transformation of variables by changing everything to logarithms.

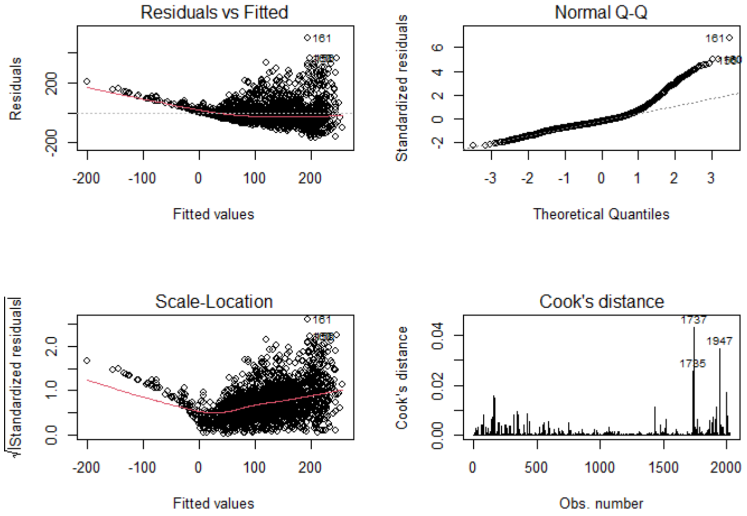


Figure 3: Residual plots for Model 1.

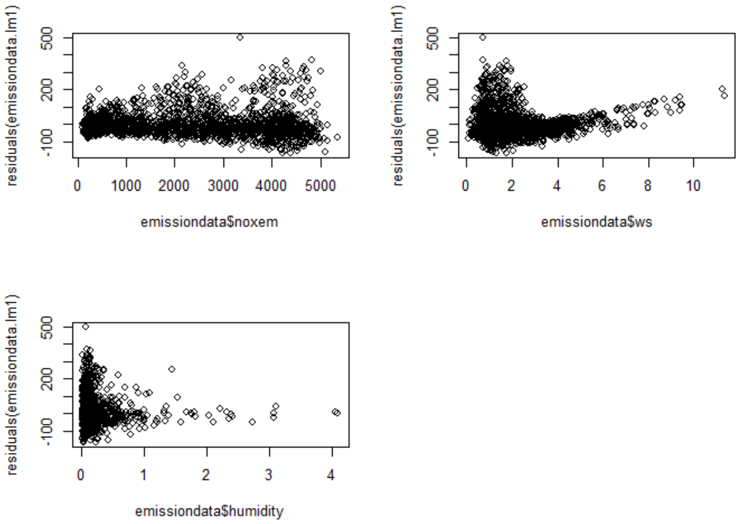


Figure 4: Residual vs explanatory variables for Model 1.

1.3.2 Model 2: Multiple Linear Regression with Transformed Variables

Thus the new model becomes:

$$log(Y_i) = \beta_0 + \beta_1 log(x_{i1}) + \beta_2 log(x_{i2}) + \beta_3 log(x_{i3}) + e_i$$

under the logarithmic transformation of variables and same assumptions as before. This transformation can not only be justified from the observations in Model 1 to tackle its short-comings but also by looking at the pairwise plots of the transformed variables (replaced each variable value by its corresponding log). We immediately notice much stronger linear relationships between the new nox and noxem and between the new nox and ws variables, while the relationship with the humidity remains non-linear.

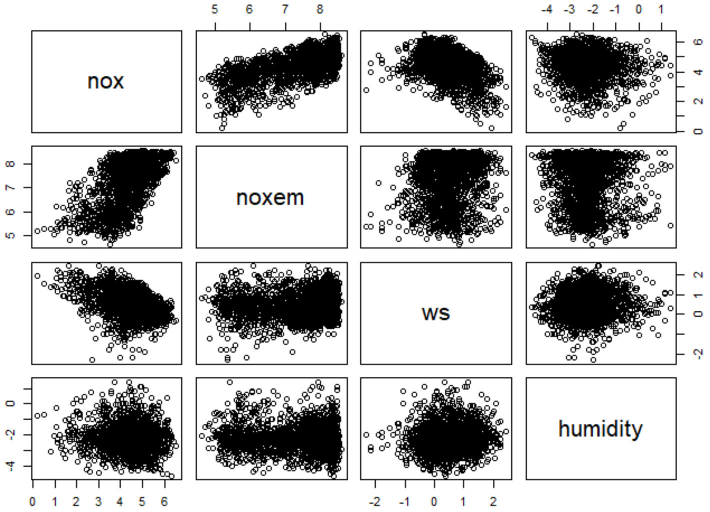
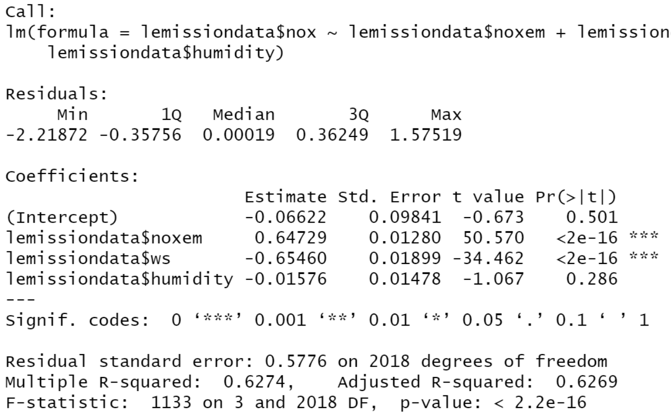


Figure 5: Matrix plot for log of emissions data (transformed variables).

We again estimate our regression coefficients, $\hat{\beta_0}, \hat{\beta_1}$ and $\hat{\beta_2}$ in R by calling the lm function on the chosen variables. The relevant statistics for the fitting of Model 2 are returned in Figure 6. As we can see, the R^2 value has now greatly increased to 0.6274 suggesting a good explanatory strength for the model. Looking at the p-values in the $Pr(>|t|)$ column, the small p-value in the third row once again suggests that humidity could be omitted from the model given that the other two explanatory variables are in the model.



```
Call:
lm(formula = lemissiondata$nox ~ lemissiondata$noxem + lemissiondata$ws +
    lemissiondata$humidity)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.21872 -0.35756  0.00019  0.36249  1.57519 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)            -0.06622    0.09841  -0.673    0.501    
lemissiondata$noxem     0.64729    0.01280  50.570   <2e-16 ***
lemissiondata$ws       -0.65460    0.01899 -34.462   <2e-16 ***
lemissiondata$humidity -0.01576    0.01478  -1.067    0.286    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5776 on 2018 degrees of freedom
Multiple R-squared:  0.6274,	Adjusted R-squared:  0.6269 
F-statistic:  1133 on 3 and 2018 DF,  p-value: < 2.2e-16
```

Figure 6: Model 2 fit.

We look at the summary statistics for residuals in Figure 7. The residuals vs fitted values plot has now drastically improved from Model 1 and the plot is now showing a randomly scattered pattern, furthermore, the Normal Q-Q plot is now looking roughly like a straight line. This is evidence that our new model has much stronger Linearity and Normality modelling assumptions, as well reduced heteroscedasticity in the statistical model.The residual standard error has also improved vastly from 73.89 to 0.5776, supporting that our statistical model is a much better fit to the data.

Figure 8 backs up the claims made from the diagnostic plots and summary statistics above. We observed that the heteroscedasticity along the explanatory variables has reduced, thus we will now concentrate on introducing non-linearity in our next statistical model to try and improve it.

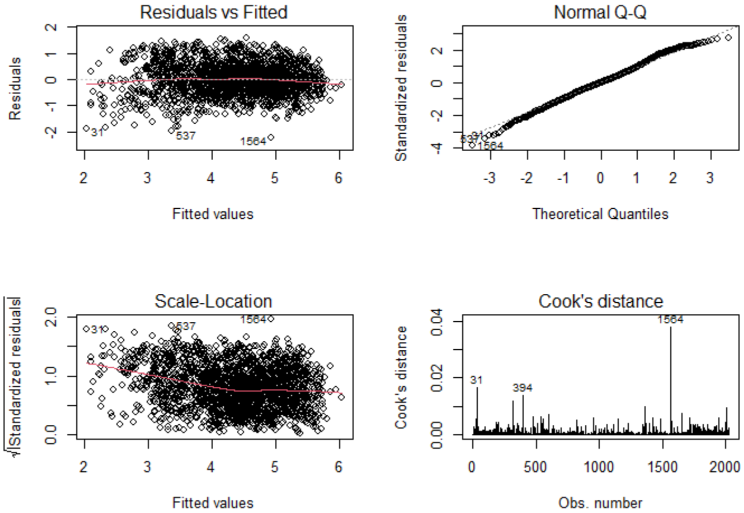


Figure 7: Residual plots for Model 2.

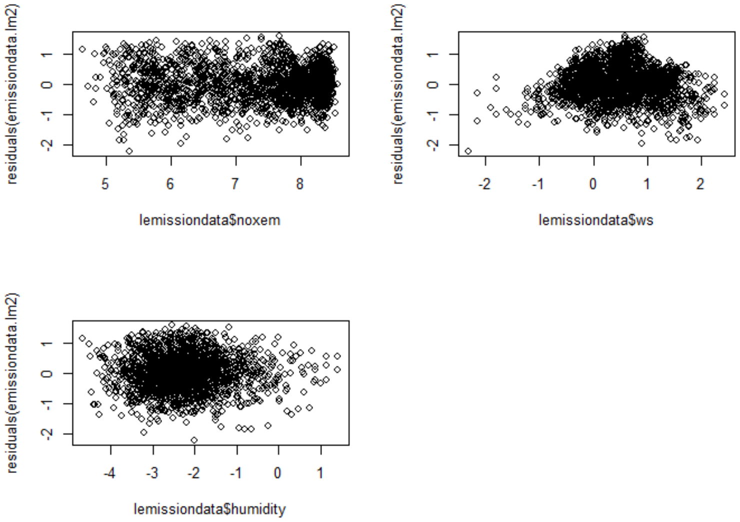


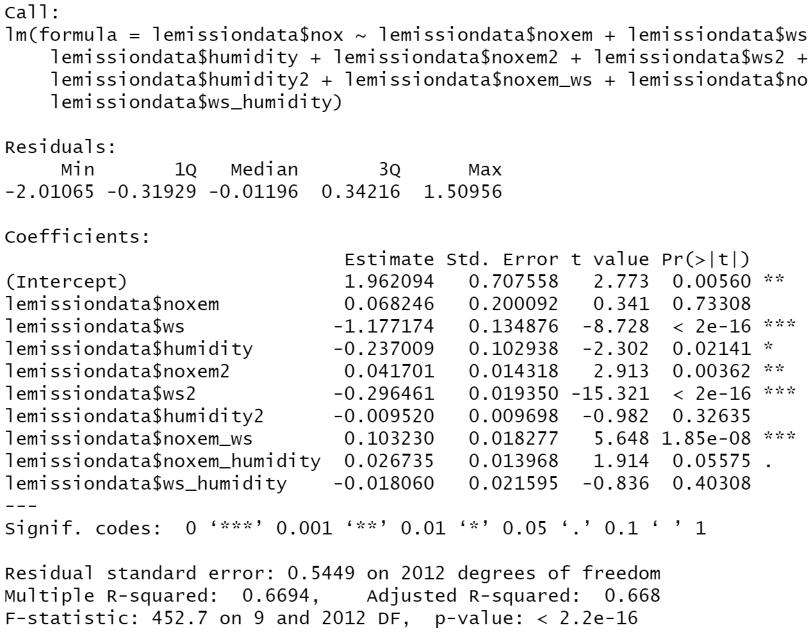
Figure 8: Residual vs explanatory variables for Model 2.

1.3.3 Model 3: Multiple Linear Regression with Transformed Variables and Interactions.

To improve on Model 2, we keep working with the log transformation of the variables and in addition, consider interactions up to degree 2 in our variable space. Thus our model is reformulated as:

$$\begin{gathered} log(Y_i) = \beta_0 + \beta_1 log(x_{i1}) + \beta_2 log(x_{i2}) + \beta_3 log(x_{i3}) \\ +\beta_4 log(x_{i1}) log(x_{i2}) + \beta_5 log(x_{i1}) log(x_{i3}) + \beta_6 log(x_{i2}) log(x_{i3}) \\ +\beta_7 log(x_{i1})^2 + \beta_8 log(x_{i2})^2 + \beta_9 log(x_{i3})^2 + e_i \end{gathered}$$

under the same assumptions as before. We estimate the $\hat{\beta}_i$ in R by calling the lm function on the chosen variables. The relevant statistics for the fitting of Model 3 are returned in Figure 9. As we can see, the R^2 value has now increased to 0.6694, and the residual standard error has decreased to 0.5449, making a small improvement to Model 2.



```
Call:
lm(formula = lemissiondata$nox ~ lemissiondata$noxem + lemissiondata$ws +
    lemissiondata$humidity + lemissiondata$noxem2 + lemissiondata$ws2 +
    lemissiondata$humidity2 + lemissiondata$noxem_ws + lemissiondata$noxem_humidity +
    lemissiondata$ws_humidity)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01065 -0.31929 -0.01196  0.34216  1.50956

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.962094   0.707558   2.773  0.00560 **
lemissiondata$noxem          0.068246   0.200092   0.341  0.73308
lemissiondata$ws            -1.177174   0.134876  -8.728  < 2e-16 ***
lemissiondata$humidity      -0.237009   0.102938  -2.302  0.02141 *
lemissiondata$noxem2         0.041701   0.014318   2.913  0.00362 **
lemissiondata$ws2           -0.296461   0.019350 -15.321  < 2e-16 ***
lemissiondata$humidity2     -0.009520   0.009698  -0.982  0.32635
lemissiondata$noxem_ws       0.103230   0.018277   5.648 1.85e-08 ***
lemissiondata$noxem_humidity  0.026735   0.013968   1.914  0.05575 .
lemissiondata$ws_humidity   -0.018060   0.021595  -0.836  0.40308
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5449 on 2012 degrees of freedom
Multiple R-squared:  0.6694,	Adjusted R-squared:  0.668
F-statistic: 452.7 on 9 and 2012 DF,  p-value: < 2.2e-16
```

Figure 9: Model 3 fit.

We look at the summary statistics for residuals in Figure 10. The residuals vs fitted plot looks randomly scattered and has improved on that of Model 2 in Figure 7 as the mean of the residuals does not vary and is more linear. Furthermore the Normal Q-Q plot is strong and slightly more linear, suggesting we managed to correct and also further improve the non-linearity and heteroscedasticity issues we had with respect to Model 1. The distances in the Cook's plot also look harmless, suggesting there is no need for a robust model.

Lastly, in Figure 11 we look at the plot of the residuals against each explanatory variable in Model 3. These look fairly randomly scattered which suggests independence of the errors from the explanatory variables.

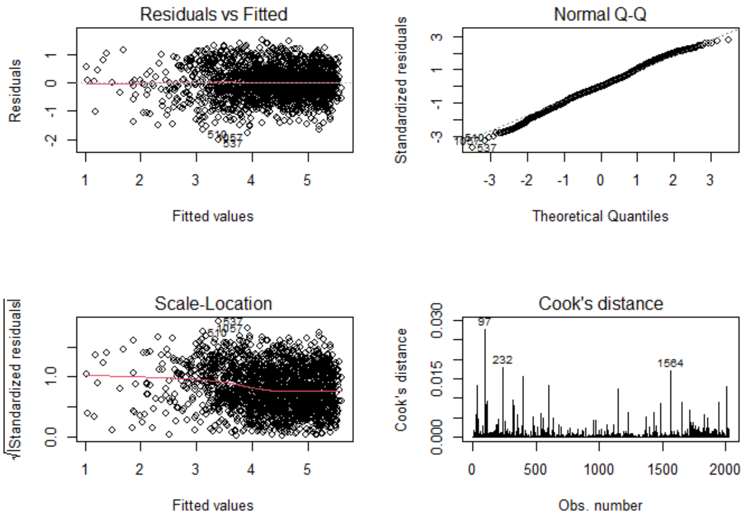


Figure 10: Residual plots for Model 3.

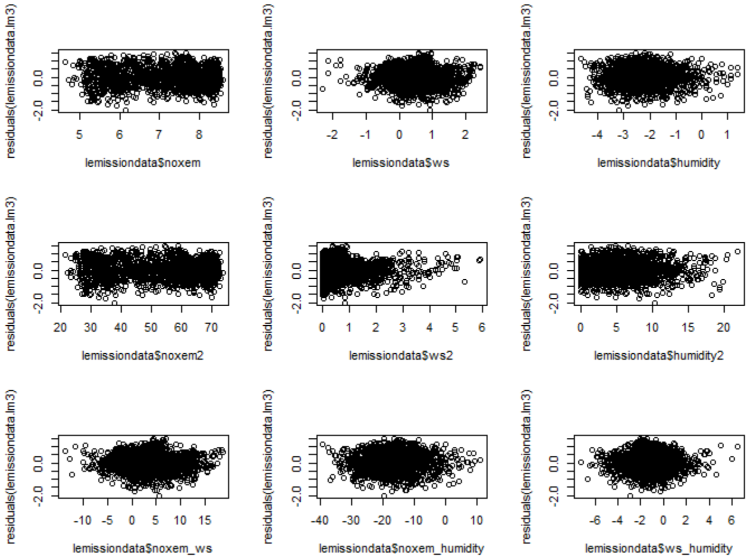


Figure 11: Residual vs explanatory variables for Model 3.

1.3.4 Model Comparison

Table 1 gives a brief overview and summary of our 3 statistical models used in the model selection process above. We conclude that model 3 has the highest R^2, lowest Residual Standard Error, and strongest Linearity, Normality and Homogeneity of variances assumptions and is therefore the recommended fitted model that we have provided.

Model	$\mathbf{R^2}$	**Residual Standard Error**	**Linearity**	**Normality**	**Variances**
Model 1	0.4370	73.89	Violated	Violated	heteroscedasticity
Model 2	0.6274	0.5776	Good	Good	Constant
Model 3	0.6694	0.5449	Best	Best	Constant

Table 1: Comparison of Statistical Models

1.4 Hypothesis Test for two of the same regression coefficient

Sometimes we are interested in testing joint hypotheses which impose restrictions on multiple regression coefficients. For example, in model 3:

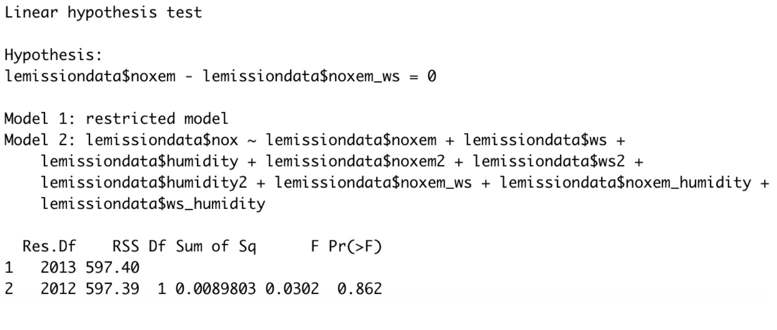
$$log(Y_i) = \beta_0 + \beta_1 log(x_{i1}) + \beta_2 log(x_{i2}) + \beta_3 log(x_{i3})$$
$$+\beta_4 log(x_{i1}) log(x_{i2}) + \beta_5 log(x_{i1}) log(x_{i3}) + \beta_6 log(x_{i2}) log(x_{i3})$$
$$+\beta_7 log(x_{i1})^2 + \beta_8 log(x_{i2})^2 + \beta_9 log(x_{i3})^2 + e_i$$

, we may test the null hypothesis $H_0 : \beta_1 = \beta_4$ against the alternative hypothesis $H_0 : \beta_1 \neq \beta_4$. Note that β_1 is the coefficient of noxem and β_4 is the coefficient of the interactive term between noxem and wind speed.

The procedure is as followed:

1. Estimate the restricted model i.e. imposing our null hypothesis (assumed to be true) onto our model.
2. Compute the residual sum of squares of this restricted model. (RRSS).
3. Estimate the unrestricted model, i.e. imposing our alternative hypothesis, which is our original model.
4. Compute the residual sum of squares residuals of this unrestricted model (URSS).
5. Finally compute the F-test statistic. The F statistic is equal to $\frac{(RRSS-URRS)(N-K)}{q*URSS}$, where $N-K$ is the degrees of freedom and q is the number of restrictions imposed. Additionally, we calculate the p value for the F statistics i.e. $Pr\,(> F)$.

All of this can be done by simply calling the *linearHypothesis* package from the **car** library.



```
Linear hypothesis test

Hypothesis:
lemissiondata$noxem - lemissiondata$noxem_ws = 0

Model 1: restricted model
Model 2: lemissiondata$nox ~ lemissiondata$noxem + lemissiondata$ws +
    lemissiondata$humidity + lemissiondata$noxem2 + lemissiondata$ws2 +
    lemissiondata$humidity2 + lemissiondata$noxem_ws + lemissiondata$noxem_humidity +
    lemissiondata$ws_humidity

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   2013 597.40
2   2012 597.39  1 0.0089803 0.0302  0.862
```

Figure 12: Linear hypothesis test between the $log\,(noxem)$ and $log\,(noxem)\,log\,(ws)$.

From Figure 12, we see that the sum of residual squares has not been affected. Given that $p = Pr\,(> F) = 0.862 > 0.05$, we conclude that a significant difference does not exist and thus are more likely to accept the null hypothesis.

1.5 Report for Institute

From our findings above we have shown that the proposed model with the strongest predictive power includes a logarithmic variable transformation and the addition of some interaction terms to capture non-linearity.

We present to the institute the impacts on the NOx pollution content in the ambient air, by varying each predictor variable by 1% with respect to our final fitted statistical model.

- A 1% increase in noxem resulted in $\sim 0.1798\%$ increase in nox
- A 1% increase in ws resulted in $\sim 0.9324\%$ decrease in nox
- A 1% increase in humidity resulted in $\sim 0.0627\%$ decrease in nox

Note that all these effects are assuming that all other variables are held fixed at their mean values. From these impacts it can be seen that there is a significant influence from both the NOx emissions from cars on the nearby motor way and also the wind speed. In contrast there seems to be only a very weak effect from the humidity, which agrees with previous findings in our model selection process. It should also be noted that the data set was limited to the time span of one year only. This could potentially introduce some doubt about its reliability since their may be some other factors that could have some underlying effects on the data such that having data over more than a single year could reveal something new. For example there could be some randomness involved from measurements taken in just one year (rare weather conditions / unusually high traffic). We suggest to the institute that perhaps in the future, a more carefully designed and controlled experiment can be carried out or more data collected such that a better statistical model can be formulated. For this reason although we noticed some quadratic behaviour between time and the response, we decided not to include the time variable in our model. This could have potentially lead to inaccuracies in using the model for inference in future years. Another thing to note is that due to the nature of our final statistical model, by taking logarithms and interaction terms, we are increasing the dimensionality of the model and thereby complexity. This can be hard to simplify the relationship of the data and also may be overfitting the noise in some sense.

2 Question 2

Given that $\hat{y} = Hy$ where $H = X(X^TX)^{-1}X^T$ (projection matrix), we can show the following properties:

- $H^T = H$ i.e. symmetric
- $H^2 = H$ i.e. idempotent

Proof:

$$\begin{aligned} H^T &= (X(X^TX)^{-1}X^T)^T \\ &= X^T((X^TX)^{-1})^TX \\ &= X(X^TX)^{-1}X^T \\ &= H \end{aligned}$$

$$\begin{aligned} H^2 &= HH^T \\ &= H^TH \\ &= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\ &= X(X^TX)^{-1}X^T \\ &= H \end{aligned}$$

Since $\hat{e} = y - Hy$ (from lectures), we see that:

$$\begin{aligned} \hat{e}^T\hat{e} &= (y - Hy)^T(y - Hy) \\ &= y^Ty - y^THy - y^TH^Ty + y^TH^THy \\ &= y^Ty - y^TH^THy - y^TH^THy + y^TH^THy \\ &= y^Ty - y^TH^THy = y^Ty - (Hy)^THy \\ &= y^Ty - \hat{y}^T\hat{y} \\ \therefore y^Ty &= \hat{y}^T\hat{y} + \hat{e}^T\hat{e} \end{aligned}$$

3 Question 3

3.1 (a)

$$X = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ 1 & x_3 - \bar{x} \\ \vdots & \\ 1 & x_n - \bar{x} \end{pmatrix}$$

$$X^T = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 - \bar{x} & x_2 - \bar{x} & x_3 - \bar{x} & \dots & x_n - \bar{x} \end{pmatrix}$$

$$X^T X = \begin{pmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n (x_i - \bar{x}) \\ \sum_{i=1}^n (x_i - \bar{x}) & \sum_{i=1}^n (x_i - \bar{x})^2 \end{pmatrix} = \begin{pmatrix} n & n\bar{x} - n\bar{x} \\ n\bar{x} - n\bar{x} & S_{xx} \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}$$

where $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$.

3.2 (b)

If $x_1 = x_2 = \dots = x_n$ then $\bar{x} = x_1 = x_2 = \dots = x_n$
$\Rightarrow S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (0)^2 = 0$

$$\therefore X^T X = \begin{pmatrix} n & 0 \\ 0 & 0 \end{pmatrix}$$

which cannot be inverted.

3.3 (c)

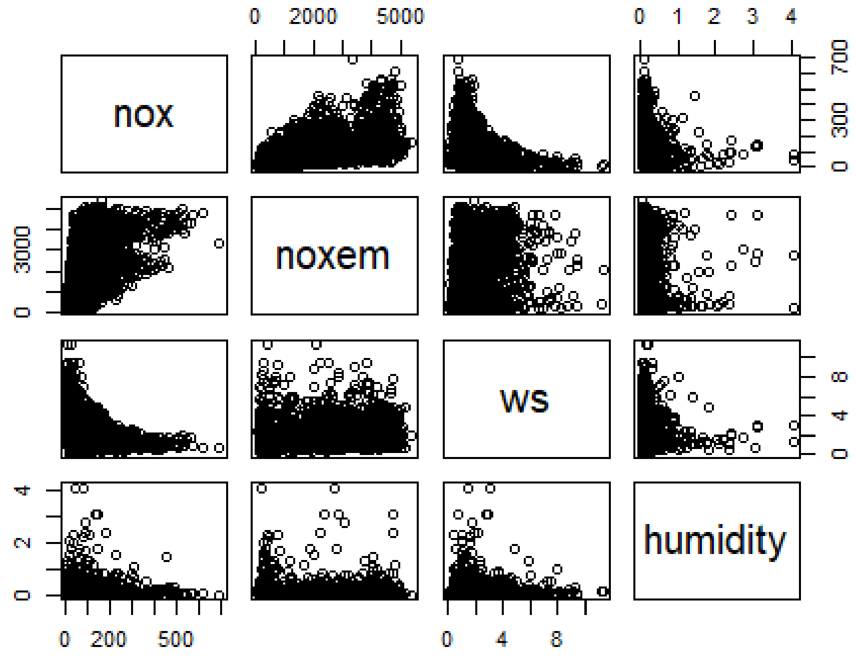
$X^T X \hat{\beta} = X^T y$

$$\Rightarrow \begin{pmatrix} n & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta_0} \\ \hat{\beta_1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

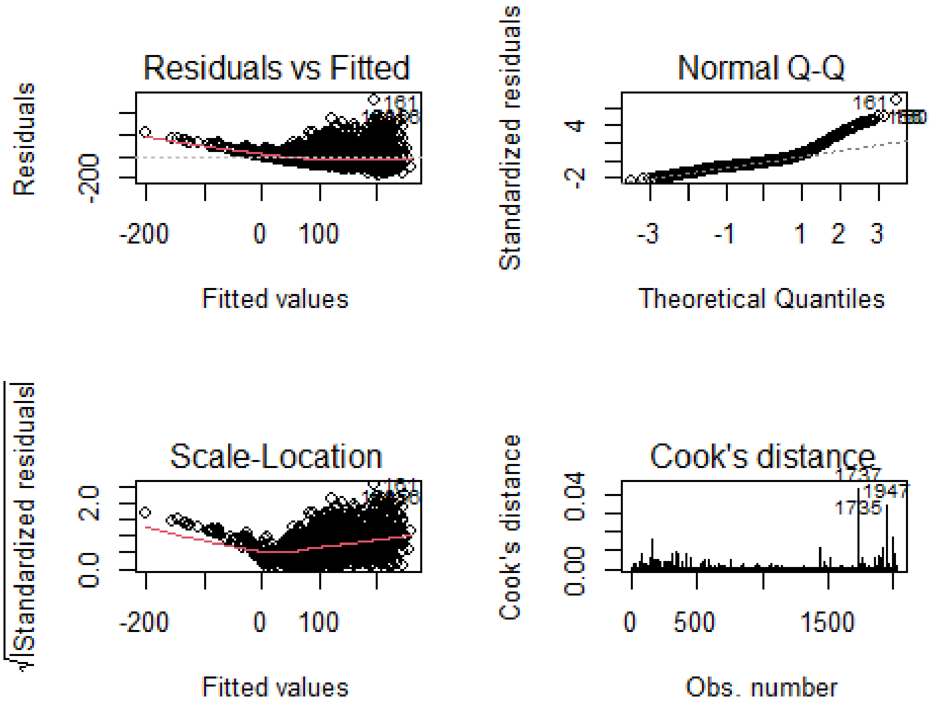
$$\Rightarrow \begin{pmatrix} n\hat{\beta_0} \\ 0 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ 0 \end{pmatrix}$$

Thus $\hat{\beta_1}$ can take any values as the above expression is always true no matter the value of $\hat{\beta_1}$.

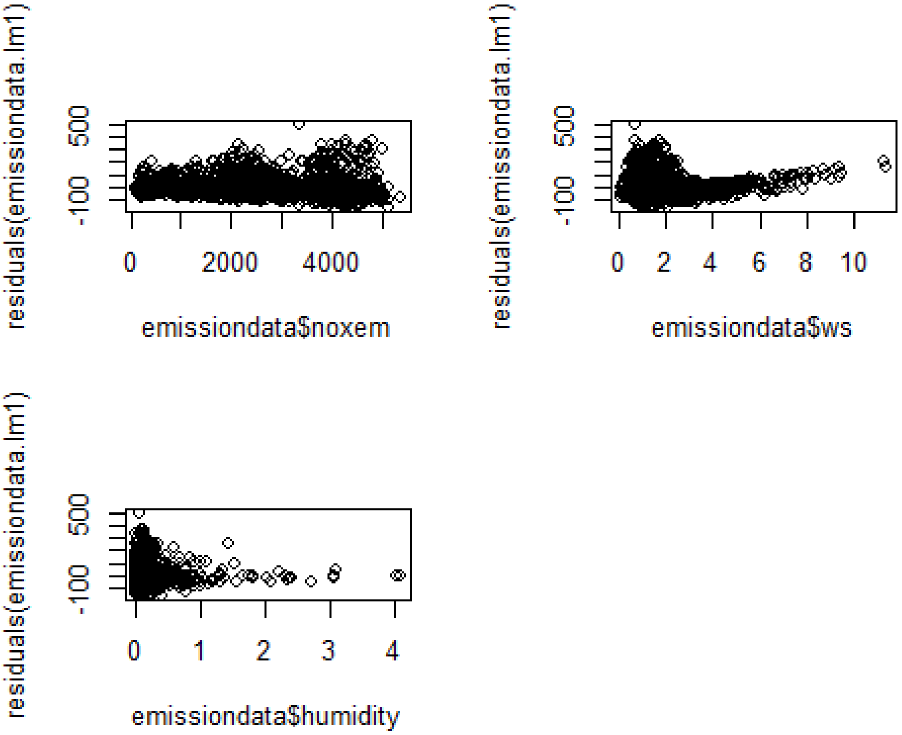
```
#Read data as table
emissiondata <-read.table("emissionssw.dat", header = TRUE)

# Or use the following code to import data
#loading in dataset
#emissiondata <-read.table("emissionssw.dat", header = TRUE)

#plot of data overleaf
pairs(emissiondata)
```

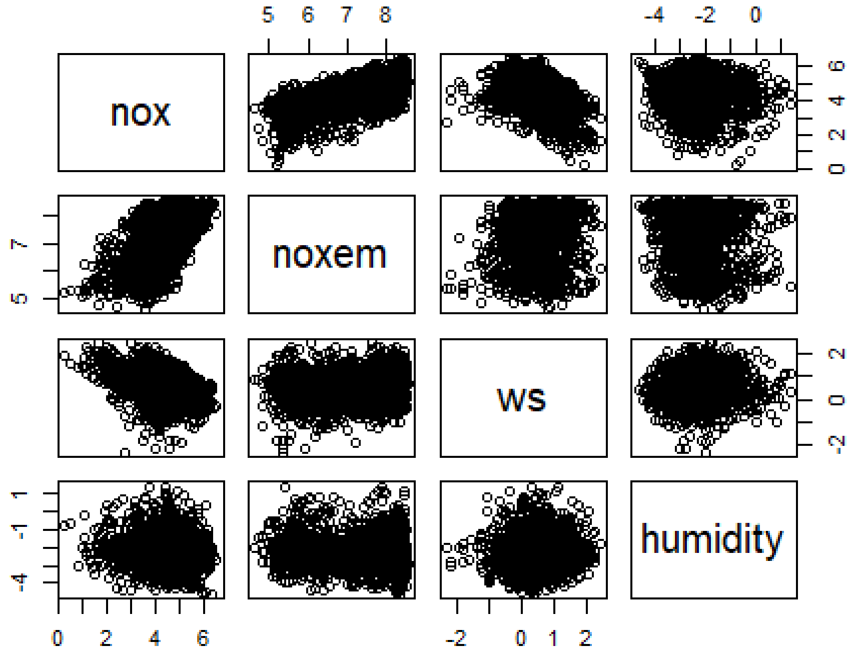


```
#Model 1
emissiondata.lm1 <-lm(emissiondata$nox~ emissiondata$noxem + emissiondata$ws
+ emissiondata$humidity)
summary(emissiondata.lm1)

##
## Call:
## lm(formula = emissiondata$nox ~ emissiondata$noxem + emissiondata$ws +
##     emissiondata$humidity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -166.19  -42.65  -15.38   20.03  500.14
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             92.947559   3.699776   25.122   <2e-16 ***
## emissiondata$noxem      0.036152   0.001077  33.573   <2e-16 ***
## emissiondata$ws       -27.338040   1.113436 -24.553   <2e-16 ***
## emissiondata$humidity  -7.227542   5.705300  -1.267    0.205
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73.89 on 2018 degrees of freedom
## Multiple R-squared:  0.437,  Adjusted R-squared:  0.4361
## F-statistic: 522.1 on 3 and 2018 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(emissiondata.lm1, which=1:4, ask=FALSE)
```



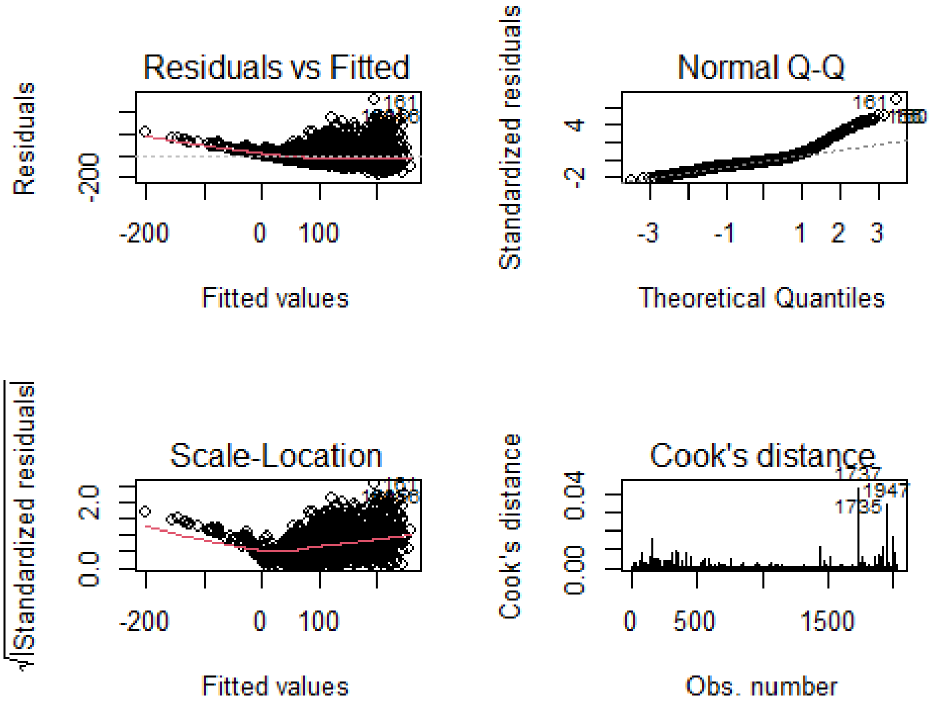
```
plot(emissiondata$noxem, residuals(emissiondata.lm1))
plot(emissiondata$ws, residuals(emissiondata.lm1))
plot(emissiondata$humidity, residuals(emissiondata.lm1))




#logarithmic transformation
lemissiondata <-log(emissiondata)
pairs(lemissiondata)
```



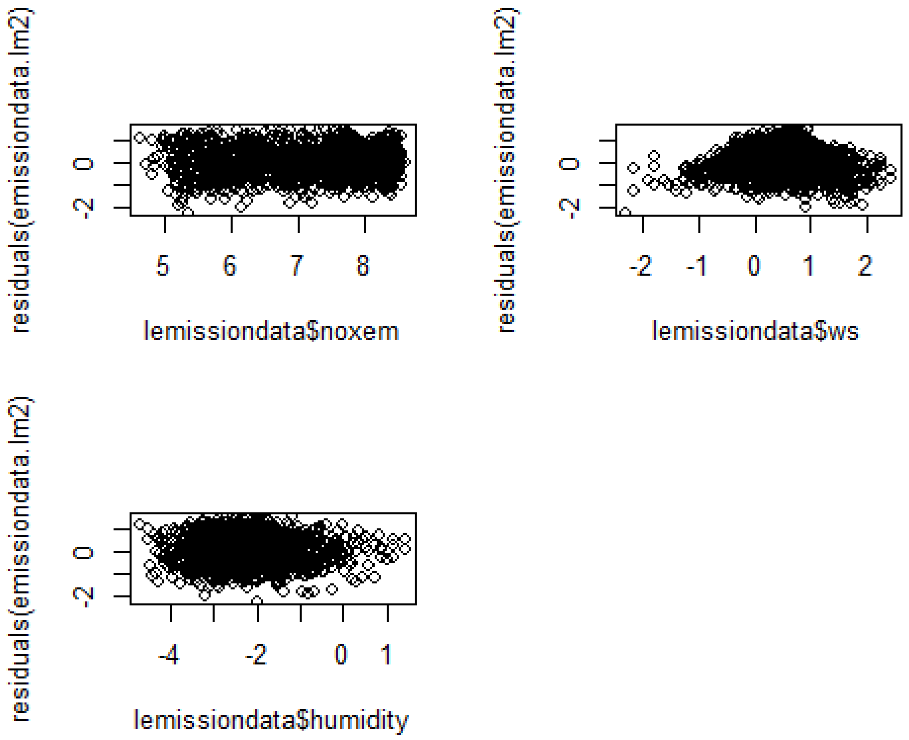


```
par(mfrow=c(1,1))
```

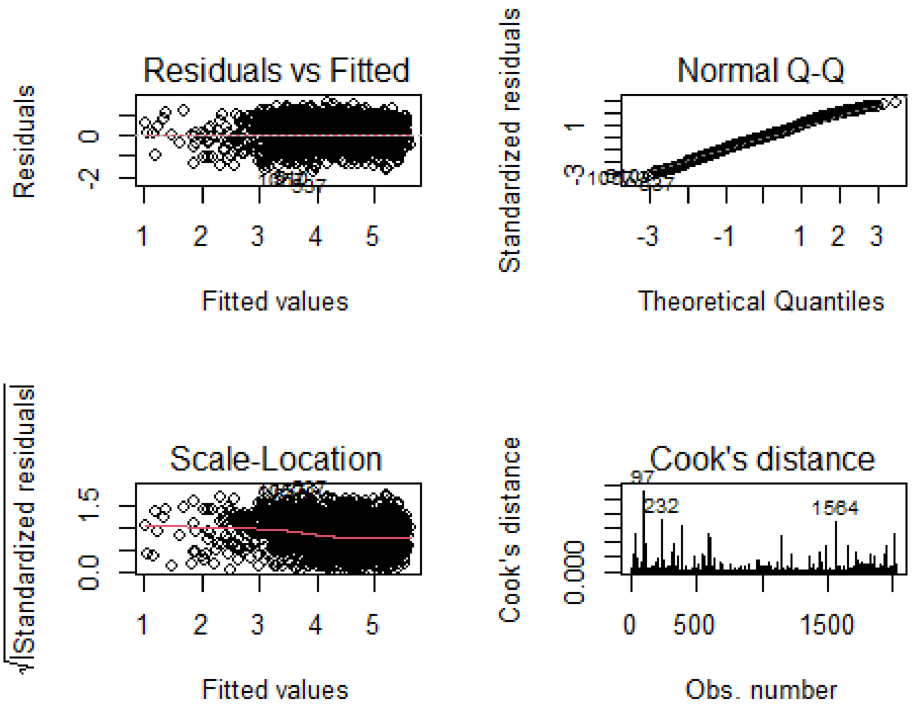
```
#Model 2
emissiondata.lm2 <-lm(lemissiondata$nox~ lemissiondata$noxem +
lemissiondata$ws + lemissiondata$humidity)
summary(emissiondata.lm2)

##
## Call:
## lm(formula = lemissiondata$nox ~ lemissiondata$noxem + lemissiondata$ws +
##     lemissiondata$humidity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.21872 -0.35756  0.00019  0.36249  1.57519
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -0.06622    0.09841  -0.673    0.501
## lemissiondata$noxem     0.64729    0.01280  50.570   <2e-16 ***
## lemissiondata$ws       -0.65460    0.01899 -34.462   <2e-16 ***
## lemissiondata$humidity -0.01576    0.01478  -1.067    0.286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5776 on 2018 degrees of freedom
## Multiple R-squared:  0.6274, Adjusted R-squared:  0.6269
## F-statistic:  1133 on 3 and 2018 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(emissiondata.lm1, which=1:4, ask=FALSE)
```



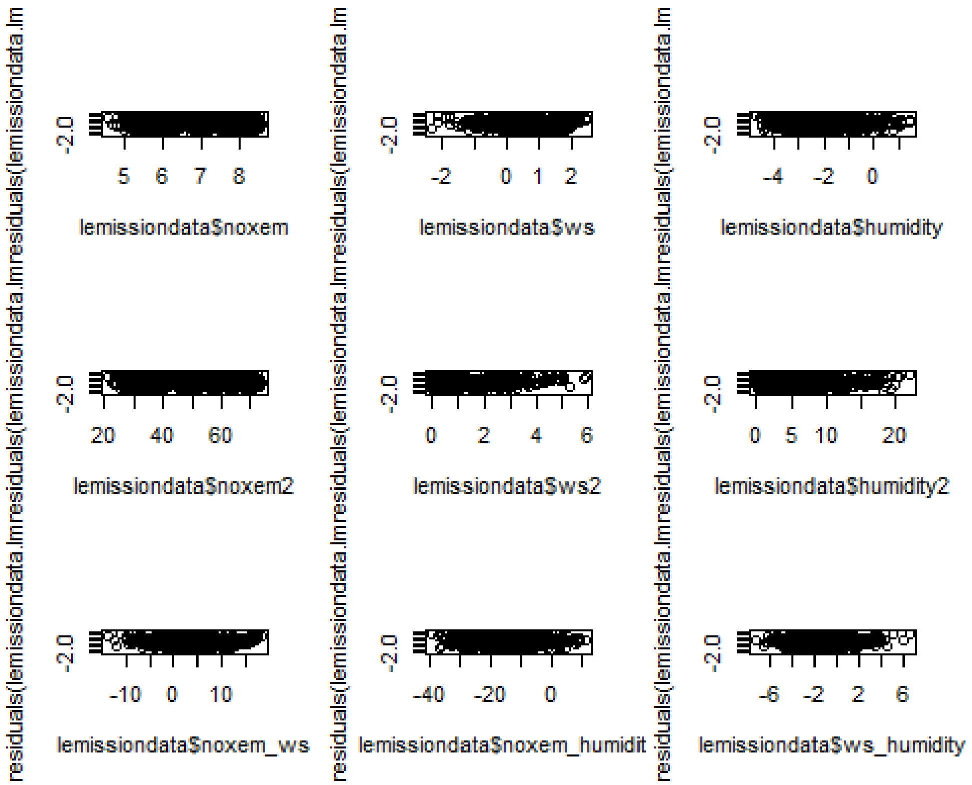
```
plot(lemissiondata$noxem, residuals(emissiondata.lm2))
plot(lemissiondata$ws, residuals(emissiondata.lm2))
plot(lemissiondata$humidity, residuals(emissiondata.lm2))






#Model 3
lemissiondata$noxem2 <- (lemissiondata$noxem)^2
lemissiondata$ws2 <- (lemissiondata$ws)^2
lemissiondata$humidity2 <- (lemissiondata$humidity)^2
lemissiondata$noxem_ws <- lemissiondata$noxem * lemissiondata$ws
lemissiondata$noxem_humidity <- lemissiondata$noxem * lemissiondata$humidity
lemissiondata$ws_humidity <- lemissiondata$ws * lemissiondata$humidity
lemissiondata.lm3 <-lm(lemissiondata$nox~ lemissiondata$noxem +
lemissiondata$ws + lemissiondata$humidity + lemissiondata$noxem2 +
lemissiondata$ws2 + lemissiondata$humidity2 + lemissiondata$noxem_ws +
lemissiondata$noxem_humidity + lemissiondata$ws_humidity)
summary(lemissiondata.lm3)

##
## Call:
## lm(formula = lemissiondata$nox ~ lemissiondata$noxem + lemissiondata$ws +
##     lemissiondata$humidity + lemissiondata$noxem2 + lemissiondata$ws2 +
##     lemissiondata$humidity2 + lemissiondata$noxem_ws +
lemissiondata$noxem_humidity + 
##     lemissiondata$ws_humidity)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.01065 -0.31929 -0.01196  0.34216  1.50956 
## 
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    1.962094   0.707558   2.773  0.00560 ** 
## lemissiondata$noxem            0.068246   0.200092   0.341  0.73308    
## lemissiondata$ws              -1.177174   0.134876  -8.728  < 2e-16 ***
## lemissiondata$humidity        -0.237009   0.102938  -2.302  0.02141 *  
## lemissiondata$noxem2           0.041701   0.014318   2.913  0.00362 ** 
## lemissiondata$ws2             -0.296461   0.019350 -15.321  < 2e-16 ***
## lemissiondata$humidity2       -0.009520   0.009698  -0.982  0.32635    
## lemissiondata$noxem_ws         0.103230   0.018277   5.648 1.85e-08 ***
## lemissiondata$noxem_humidity  0.026735   0.013968   1.914  0.05575 .  
## lemissiondata$ws_humidity     -0.018060   0.021595  -0.836  0.40308    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5449 on 2012 degrees of freedom
## Multiple R-squared:  0.6694, Adjusted R-squared:  0.668 
## F-statistic: 452.7 on 9 and 2012 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
```



```
plot(lemissiondata.lm3, which=1:4, ask=FALSE)
```



```
par(mfrow=c(3,3))
plot(lemissiondata$noxem, residuals(lemissiondata.lm3))
plot(lemissiondata$ws, residuals(lemissiondata.lm3))
plot(lemissiondata$humidity, residuals(lemissiondata.lm3))
plot(lemissiondata$noxem2, residuals(lemissiondata.lm3))
plot(lemissiondata$ws2, residuals(lemissiondata.lm3))
plot(lemissiondata$humidity2, residuals(lemissiondata.lm3))
plot(lemissiondata$noxem_ws, residuals(lemissiondata.lm3))
plot(lemissiondata$noxem_humidity, residuals(lemissiondata.lm3))
plot(lemissiondata$ws_humidity, residuals(lemissiondata.lm3))
```



```
# Hypothesis Testing
library(car)

## Warning: package 'car' was built under R version 4.1.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.1.3

linearHypothesis(lemissiondata.lm3,
"lemissiondata$noxem2=lemissiondata$noxem_ws")

## Linear hypothesis test
##
## Hypothesis:
## lemissiondata$noxem2 - lemissiondata$noxem_ws = 0
##
## Model 1: restricted model
## Model 2: lemissiondata$nox ~ lemissiondata$noxem + lemissiondata$ws +
##     lemissiondata$humidity + lemissiondata$noxem2 + lemissiondata$ws2 +
##     lemissiondata$humidity2 + lemissiondata$noxem_ws +
lemissiondata$noxem_humidity +
##     lemissiondata$ws_humidity
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   2013 599.38
## 2   2012 597.39  1    1.9907 6.7045 0.009686 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```