1

1.1 Introduction and data exploration

The data is composed of 2022 measurements of NOx pollution content in the ambient air at a certain place in Switzerland close to a motorway and some related variables. We aim to fit and recommend one model in order to explain and make quantitative statements about the relationship between nox and the other three regressors.
Plotting the pairwise relationship of all variables reveals that there is no linear relationship between the variables except for a slight one between nox and noxem.

1.2 Model selection

We are going to fit 3 models and explain our choices in term of selection. The best model will preserve the linearity assumption, have Gaussian residuals, a constant variance, low residual standard error and high R squared.

1.2.1 The full model

We start with a simple additive model including all variables (noxem, ws, humidity) and the response variable (nox). That is $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i$, and $\varepsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, 2020$ independently distributed. Where $x_{i,1}$ = noxem, $x_{i,2}$ = ws and $x_{i,3}$ = humidity.
We observe that residual's normality is violated as **Table 1** shows an asymmetry in quantiles and the dots from the Q-Q plot (**Figure 1**) do not lie on the line. We also observe in the Residuals vs Fitted plot a clear heteroscedasticity as residuals are not randomly spread around a horizontal line. In details, it shows a funnel shape of the residuals. This indicates a potential non-constancy of variance. Also, the Scale-Location, in which are showed the standardized residuals, brings to light the potential inadequacy in the linearity of this model. Plus, the residual standard error is high and the R squared is low; indeed, only 43% of the variations in the response variable are explained by the fitted model. The outliers appearing are not significant and will not be influential as their cook's distance is less than 0.05.

Model	:	lm	(nox	∼	noxem + ws + humidity)
Residuals :	Min:-166.19	1Q:-42.65	Median:-15.38	3Q:20.03	Max:500.14
Residual standard error :	73.89	R^2:	0.437	p-value :	$< 2.2e^{-16}$

Table 1: Main results for the full model [corresponds to R-output 2]

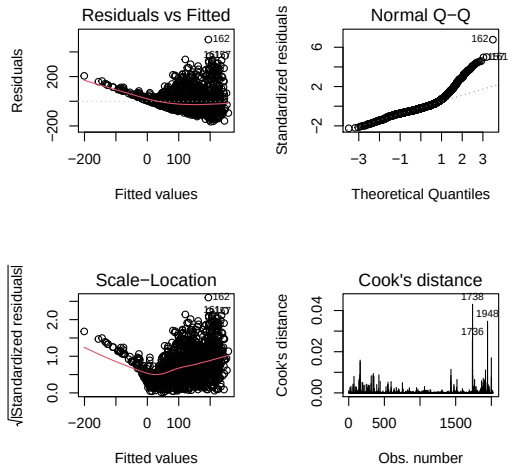


Figure 1: Residual plots for the full model [Corresponds to R-output 3]

1.2.2 The BoxCox model

In order to tackle the heteroscedasticity and the non-normality of the residuals we now perform a box-cox transformation to find the best λ such that the linear model would better fit y^λ. We thus maximize the log-likelihood of the data with respect to λ, for the values between -3 and 3. The closest interpretable value of λ with high log-likelihood is 0, and, therefore, a normal linear model would best fit log(y) under the same assumption of the previous one. That is, we fit : $log(Y_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i$. For this model, the **Table 2** shows a symmetry in the residuals, thus we can assume they are gaussian; the residual standard error is quite low and the R^2 increased compared to the full model. The plots from (**Figure 2**) confirm that this model is clearly better than the full model. Nonetheless, we are going to try to fit a third model, indeed, the symmetry in the residuals and the linearity assumption are not perfect.

Model	:	lm	(log(nox)	∼	noxem + ws + humidity)
Residuals :	Min:-2.19965	1Q:-0.35439	Median:0.00825	3Q:0.36562	Max:1.70010
Residual standard error :	0.5883	R^2 :	0.6135	p-value :	$< 2.2e^{-16}$

Table 2: Main results for the boxcox model [corresponds to R-output 5]

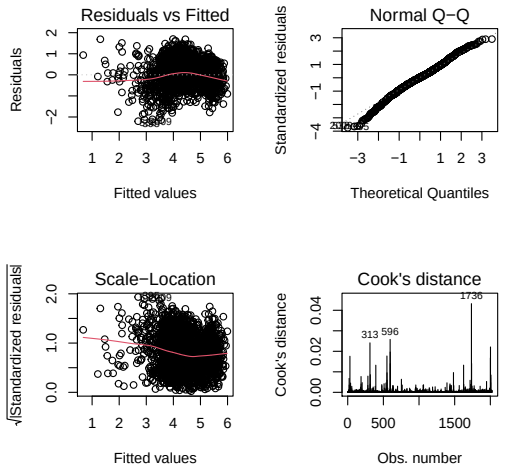


Figure 2: Residual plots for the boxcox model [corresponds to the R-output 6]

1.2.3 The final model

In order to further improve the model, we use the fact that there is no clear relationship between LogHumidity and LogNox and no linear relationship between Logws and LogNox. That is, we decide to drop LogHumidity and add a quadratic term to capture the non-linearity among Logws and LogNox. Thus, we fit the following model : $log(Y_i) = \beta_0 + \beta_1 log(x_{i,1}) + \beta_2 log(x_{i,2}) + \beta_3 log(x_{i,2}^2) + \varepsilon_i$.
We observe from **Table 3** and **Figure 3** below that the distribution of the residuals is symmetric and centered around 0 : we observe an major improvement on the QQ-plot regarding the minimum and maximum tails, the heteroscedascity disappeared and the assumption of linearity of the model got stronger. Also, the R squared improved : 66.12% of the change in the response variable is explained by this fitted model and the residual standard error is reasonably low. Here again, the cook's distance plot shows some outliers data points, nonetheless they are not so concerning to require the help of robust estimators.

Model :	$log(Y_i) =$	β_0	$+ \beta_1 log(x_{i,1})$	$+ \beta_2 log(x_{i,2})$	$+ \beta_3 log(x_{i,3}^2) + \varepsilon_i$
Residuals :	Min:-2.03287	1Q:-0.33664	Median:-0.00693	3Q:0.35801	Max:1.48026
Residual standard error :	0.5509	R^2 :	0.6612	p-value :	$< 2.2e^{-16}$

Table 3: Main results for the boxcox model [corresponds to R-output 8]

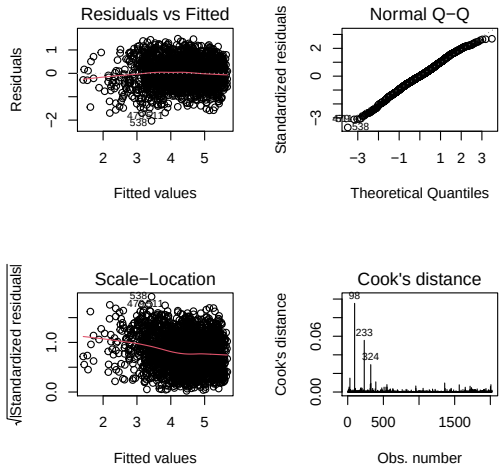


Figure 3: Residual plots for the final model [corresponds to R-output 10]

Furthermore, we notice that $\frac{\partial \log \hat{y}}{\partial \log x_{i,1}} \approx 0.64$. Here $\hat{\beta}_1$ represents the elasticity, that is an increase of 1% in LogNoxem causes on average an approximately increase of 0.64% in LogNox (all else remaining the same). Likewise, $\frac{\partial E[\log Y]}{\partial \log x_{i,2}} = -0.41 - 2 * 0.26 \approx -0.93$ implies that 1% increase in LogWs results in a 0.93% decrease in LogNox. Thus, both LogNoxem and LogWs, have an impact on the variations of LogNox.

Lastly, the **Table 4** below, outputs the confidence intervals for the estimates of this final model. Those are thigh and do not contain the 0 value, thus, the error is relatively low and we can reject the hypothesis $H_0 = 0$.

	2.5%	97.5%
Intercept	-0.111	0.241
logNoxem	0.619	0.667
logWs	-0.460	-0.363
I($\log W_s^2$)	-0.298	-0.226

Table 4: Confidence intervals for the final model's estimates [corresponds to R-output 9]

1.2.4 Comparisons and recommended model

During the last steps, we fitted three different models and are recommending the last one (final model). Here is an simplified summarizing table to have an overview of our findings :

	Full model	BoxCox model	Final model
Linearity	bad	acceptable	good
Gaussian residuals	bad	acceptable	good
Variance	heteroscedasticity	constant	constant
R^2	42.7%	61.35%	66.12%
RSE	73.89	0.5883	0.5509

Table 5: main results for the three models

Without doubt, the final model is the one we recommend to the institute. Indeed, the assumptions of linearity and Gaussian residuals are respected and the variance is constant. Also, the full model presents the highest R^2 and the lowest residual standard error. Notice that the RSE is the standard deviation of the residuals, thus, a smaller RSE means predictions are better.
Thus, we don't need further analysis to draw a conclusion : we recommend the final model fitted above.

1.3 Test for the recommended model

In order to answer properly to the question: how would you test the null hypothesis that the effect of two regression coefficients are the same ? Assume that we want to test the following null hypothesis (Verbeek, 2012):

$$H_o : \beta_{K-J+1} = \cdots = \beta_k = 0$$

The easiest test procedure is to compare the sum of squared of the residuals of the full model with the sum of square of the restricted model (which is the model with the last J regressors omitted). Denote the residual sum of squares of the full model by S_1 and that of the restricted model by S_0. If the null hypothesis is correct one would expect that the sum of squares with the restrictions imposed is only slightly larger than in the unrestricted case. It can be proved that a t-statistic is:

$$F = \frac{(S_0 - S_1)/J}{S_1/(N-K)} \sim F(J, N-J)$$

The above statistic, under the null hypothesis, follows a fisher distribution with $(J, N-J)$ degrees of freedom.

Thus, the task of testing, can be performed using the command 'Linearhypothesis' in the package 'car' in R.

The **R-output 12** and **R-output 13** (in the appendix) output interesting results about the different impacts of the variables LogNoxem and LogWs on the response variable. We can affirm that, in the final model, the magnitude of LogNoxen in terms of explanatory power, is significantly different than the magnitude of LogWs.

It is important to tackle the multiple test problem: the p-values of the t-test for each estimate in a model are not reliable to establish whether a coefficient is significant or not. They don't take into account the uncertainty related with the model selection process. A more reliable way to test whether an estimate is zero or not would be to test the differences in terms of residual sum of squares. The test on the single coefficient are showed in the **R-output 14** **R-output 15** and **R-output 16** (in the appendix). We conclude that the null hypothesis ($H_0 : \beta_i = 0$) for $i = 1, 2, 3$ is rejected in all three tests.

1.4 Report for the institute

The proposed model is based on logarithmic transformations of the variables, the reason is to be found in the initial strong variability in the data.

The logarithmic transformation was aimed precisely at reducing this variability and make some multiplicative dependence present between the variables linear. It seems to be reasonable to model the NOx measured near the lake as a function on the NOx emitted by the motorway and the wind speed. On the contrary, the humidity does not seem to have an impact on the variations of Nox measured near the lake and it has no interactions with the wind either.

Here are the main impacts we noticed on the variations on NOx near the lake :

- An increase of 1% in Noxem results approximately in an increase of 0.64% in Nox.
- An increase of 1% in Ws results approximately in a non linear decrease of 0.93% in Nox.

NB : These effects yield only under the assumption that all other variables are held constant.

All in all, the level of NOx measured near the lake seems to be strongly influenced by both the level of NOx measured near the motorway and the wind speed whilst it is not explained by the variations in absolute humidity.
Also, the reader have to keep in mind that this model is based on a one year sample. That is, those explanations can explain the measurement of NOx taken in that specific year only. Indeed, we observed a slight time dependency in the residuals in the data during the first and last months of the year. This might be due to extreme temperatures or some other uncontrolled factors. Although it is not excessively high and thus does not cause much concern, we recommend carrying out the next measurements in a more controlled design if possible.

1.5 Appendix

```
Linear hypothesis test

Hypothesis:
LogNoxem - LogWs - I(LogWs^2) = 0

Model 1: restricted model
Model 2: LogNox ~ LogNoxem + LogWs + I(LogWs^2)

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   2019 1609.74
2   2018  612.35  1    997.39 3286.9 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4: Test to compare LogNoxem and LogWs - **R-output 12**

```
Linear hypothesis test

Hypothesis:
LogNoxem - I(LogWs^2) = 0

Model 1: restricted model
Model 2: LogNox ~ LogNoxem + LogWs + I(LogWs^2)

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   2019 1134.18
2   2018  612.35  1    521.83 1719.7 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: Test to compare LogNoxem and LogWs^2 – **R-output 13**

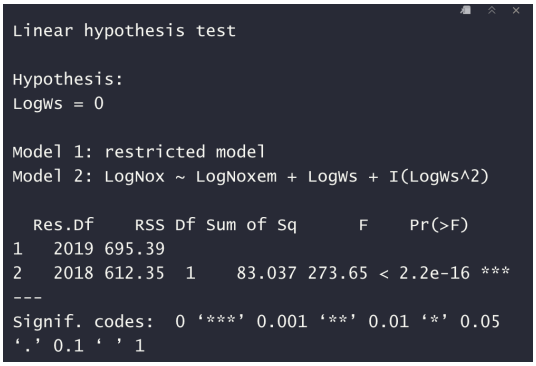
```
Linear hypothesis test

Hypothesis:
LogNoxem = 0

Model 1: restricted model
Model 2: LogNox ~ LogNoxem + LogWs + I(LogWs^2)

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1   2019 1457.07
2   2018  612.35  1    844.72 2783.8 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1
```

Figure 6: Test on LogNoxem- **R-output 14**



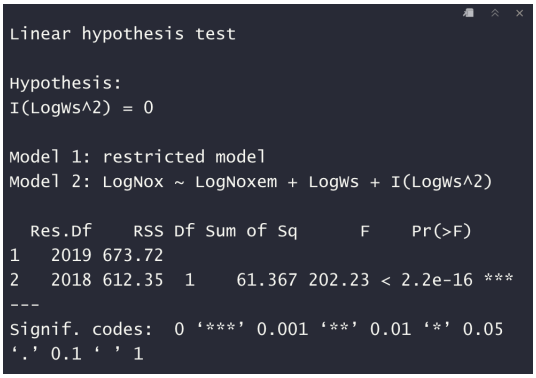
```
Linear hypothesis test

Hypothesis:
LogWs = 0

Model 1: restricted model
Model 2: LogNox ~ LogNoxem + LogWs + I(LogWs^2)

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2019 695.39
2   2018 612.35  1    83.037 273.65 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1
```

Figure 7: Test on LogWs- **R-output 15**



```
Linear hypothesis test

Hypothesis:
I(LogWs^2) = 0

Model 1: restricted model
Model 2: LogNox ~ LogNoxem + LogWs + I(LogWs^2)

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2019 673.72
2   2018 612.35  1    61.367 202.23 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
‘.’ 0.1 ‘ ’ 1
```

Figure 8: Test on LogWs^2 – **R-output 16**