NOx Emissions Analysis

Junyong Tan

Explanatory Analysis

Before we start fitting any models, it is useful to first look at how our variables interact with each other as this will help us make model assumptions and detect outliers.

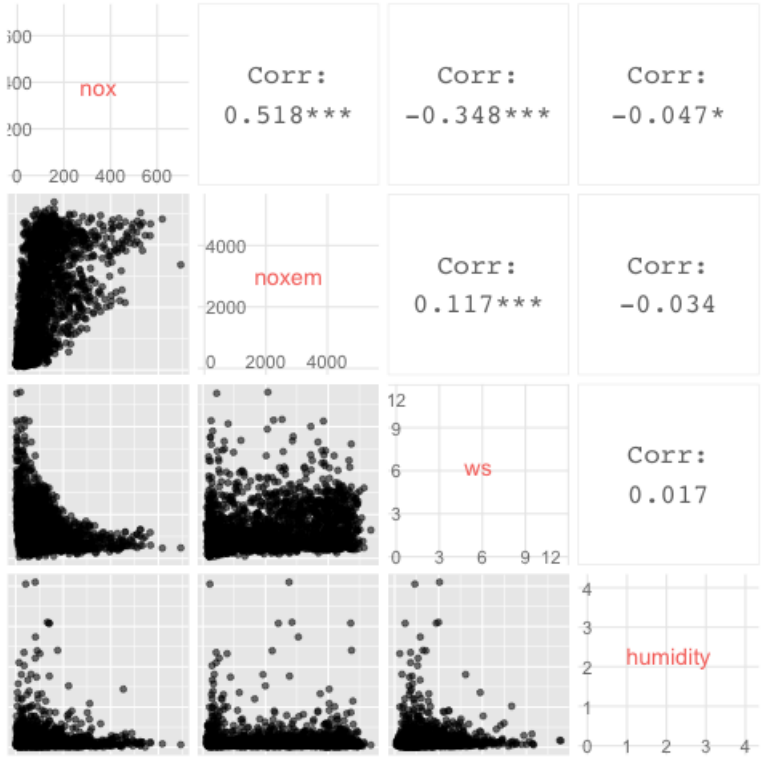

Figure 1: There is no linear dependencies between any of our variables

From above, we can clearly see that "nox" shares a non-linear relationship with each of the 3 potential explanatory variables, this prompts the need to transform our data before fitting any models to explore linear interaction between the variables. We have gone with the choice of log transformation to see if the results are better. As log changes can be interpreted as percentage changes, this is particularly useful for this data set because we don't have the units for "noxem". Lastly, we can see that there are a number of extreme values for each variable and the frequency of such values is much higher for "humidity", thus we will look out for the effects of these values on our models.

Below is the matrix plot of our data set after applying the log transformation:

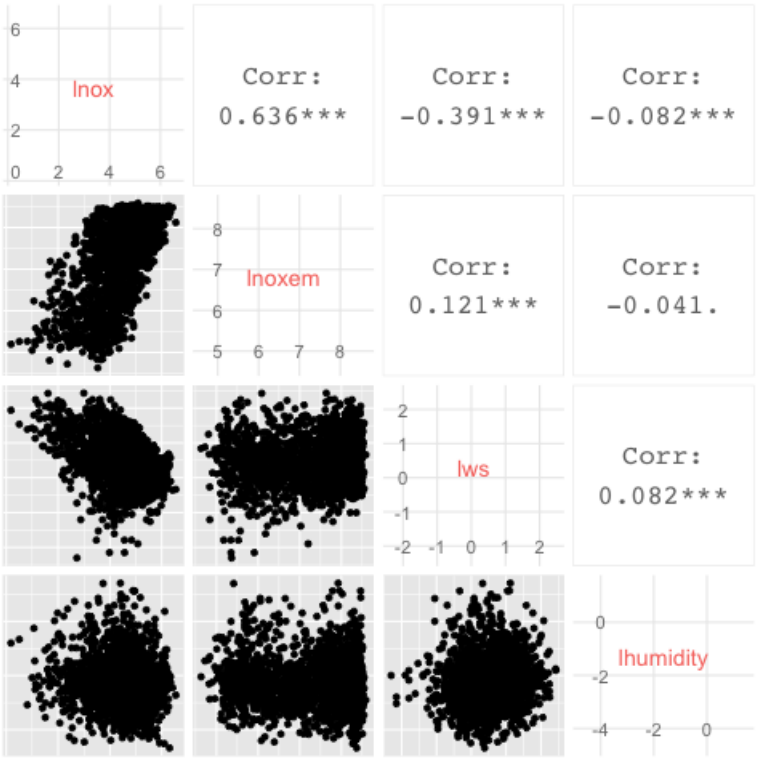

Figure 2: "log(nox)" appears to have linear dependencies with "log(noxem)" and "log(ws)"

By applying the log transformation to our data set, we have introduced some linearity patterns to our data, namely "log(nox)" vs "log(noxem)" and "log(nox)" vs "log(ws)". With the data transformed, we can be a bit more confident with our assumption of fitting a linear model with "nox" as the response variable.

Model A

- For a baseline model, we propose the following:

$$log(nox)_i = \alpha + \beta_1 log(noxem)_i + \beta_2 log(ws)_i + \beta_3 log(humidity)_i + \epsilon_i$$
$$\text{where } \epsilon_i \sim_{iid} N(0, \sigma^2) \text{ and } i = 1, \ldots, 2022$$

- Fitting such model, we have the following R output and diagnostic plots:

```
Call:
lm(formula = lnox ~ lnoxem + lws + lhumidity, data = df)

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
lnoxem       0.64729    0.01280  50.570   <2e-16 ***
lws         -0.65460    0.01899 -34.462   <2e-16 ***
lhumidity   -0.01576    0.01478  -1.067    0.286
```

```
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .

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
```

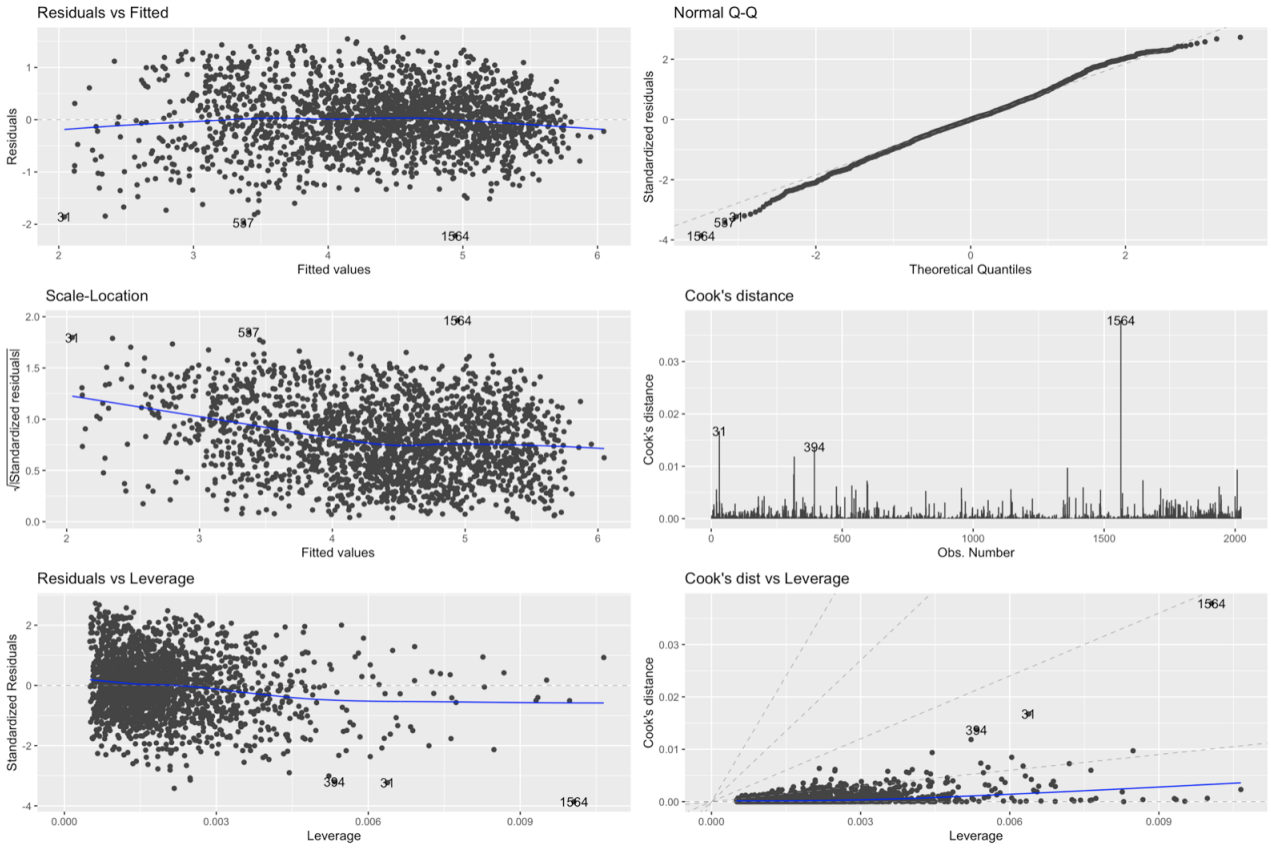

Figure 3: Diagnostic plots of "logmodel"

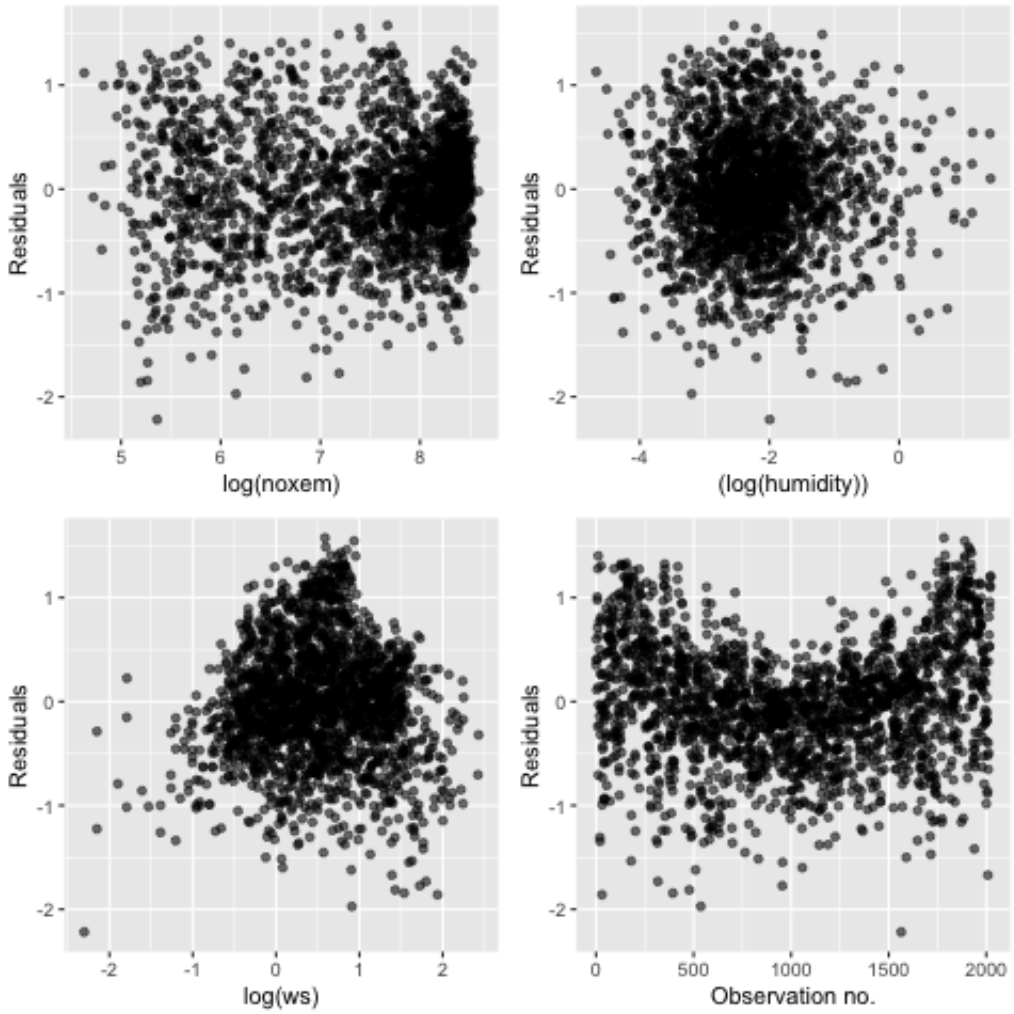

Figure 4: Residuals vs explanatory variable plots

- Comments on the output and plots:

  1. F-test has significant p-value which indicates dependence of "log(nox)" on at least one of the 3 explanatory variables.
  2. $R^2_{adjusted} = 0.6269$ whilst not low, it is not ideal. The BIC value is 3552.864.
  3. "log(humidity)" has large t-test p-value which indicates that "log(nox)" doesn't depend on "log(humidity)".
  4. "Residual vs Fitted" plot looks random and the QQ-plot follows a normal distribution except at the tail values.
  5. "Residuals vs log(noxem)" and "Residuals vs log(humidity)" look random, but that is not the case for residuals vs "log(ws)". To deal with this, we may want to consider another transformation for "ws".
  6. The outlier detection plots show three high-leverage points: 1564, 31, 384 which we will try to address.
  7. The scale-location plot indicates the presence of heteroskedasticity. The residual plot against observation order also suggests serial correlation. This is expected as measurements from the same device or within the same day may not be independent. We try to address this below.

Model B

- Our next model aims to address the shortcomings of Model A.
- We notice from both the matrix plots and the results from the previous model that log(nox) does not have a significant association with log(humidity). We omit it from this model because we prefer simpler models to complex ones.
- We make a plot of $log(nox)$ vs. $log(noxem)$ by ws category. We categorize ws as follows: $ws < 2$ m/s as low, 2 m/s <= ws <= 4 m/s as middle and $ws > 4$ m/s as high. The relationship seems linear: higher wind speeds are associated with lower log(nox) values on average. This plot also suggests that the bias and slope coefficients may differ depending on the wind speed category. This model thus includes dummy variable terms for wind speed and interaction terms between the continuous explanatory variables to capture this.

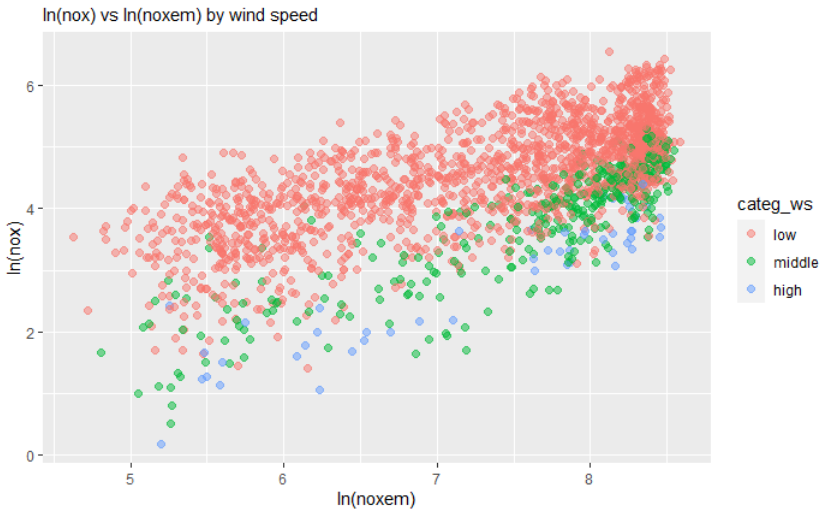

Figure 5: Scatter plot of response variable vs log(noxem)

- We make a similar plot for log(nox) vs. log(ws) by noxem category. We categorize noxem as follows: noxem < 1600 as low, 1600 <= noxem <= 3200 as middle and noxem > 3200 as high. We see similar evidence to the previous plot. To adjust the constants we include dummy variable terms for the noxem category too.

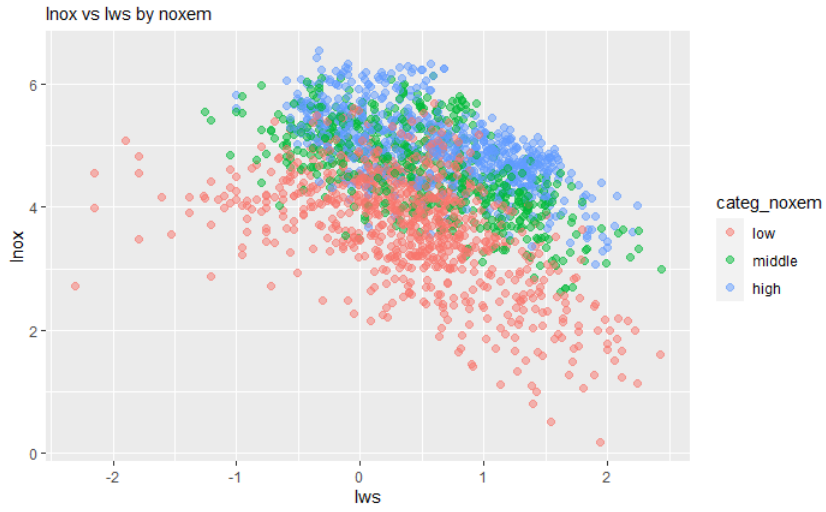

Figure 6: Scatter plot of response variable vs log(ws)

- Given that the residual plot of Model A against $log(ws)$ suggests that the residuals have a non-constant variance with respect to values of $log(ws)$, we include a $log(ws)^2$ term to capture this non-linear effect.
- Finally we add a quarter of the year indicator to address that the model shows higher residual values at high and low observation "nox". i.e. at the start of the observation year and end of the observation year.
- The model equation and corresponding R output is below:

$$\begin{aligned} log(nox)_i = \alpha + \beta_1 log(noxem)_i + \beta_2 log(ws)_i + \beta_3 log(noxem)_i log(ws)_i + \beta_4 log(ws)_i^2 \\ + \beta_5 D_{mid,ws,i} + \beta_6 D_{high,ws,i} + \beta_7 D_{mid,noxem,i} + \beta_8 D_{high,noxem,i} \\ + \beta_9 D_{quarter_2,i} + \beta_{10} D_{quarter_3,i} + \beta_{11} D_{quarter_4,i} + \epsilon_i \\ \text{where } \epsilon_i \sim_{iid} N(0, \sigma^2) \text{ and } i = 1, \ldots, 2022 \end{aligned}$$

```
Call:
lm(formula = lnox ~ lws * lnoxem + lws2 + categ_ws + categ_noxem +
    I(obs_no > n_quarter & obs_no <= 2 * n_quarter + 1) + I(obs_no >
    2 * n_quarter + 1 & obs_no <= 3 * n_quarter) + I(obs_no >
    3 * n_quarter), data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.71187 -0.30846  0.02852  0.33784  1.26774

Coefficients:
Estimate                Value       Std. Error  t value     Pr(>|t|)

(Intercept)           0.934585      5.298       0.176397     1.30e-07 ***
lws                  -1.334072     -10.744      0.124167    < 2e-16   ***
lnoxem                0.541015      19.711      0.027448    < 2e-16   ***
lws2                 -0.302471     -11.808      0.025615    < 2e-16   ***
categ_wsmiddle       -0.042400     -0.957       0.044287     0.1601
categ_wshigh         -0.123925     -1.405       0.088195     0.3385
```

```
categ_noxemmiddle     0.082542      1.695      0.048706    0.0903    .
categ_noxemhigh       0.145696      2.424      0.060109    0.0154    *
I(quarter_2)        - 0.414728    - 12.790     0.032425    < 2e-16   ***
I(quarter_3)        - 0.351063    - 10.699     0.032814    < 2e-16   ***
I(quarter_4)          0.007791      0.239      0.032635    0.8113
lws:lnoxem            0.137267      7.904      0.017367    4.42e-15  ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .

Residual standard error: 0.5108 on 2010 degrees of freedom
Multiple R-squared:  0.7098,    Adjusted R-squared:  0.7082
F-statistic:  447 on 11 and 2010 DF,  p-value: < 2.2e-16
```

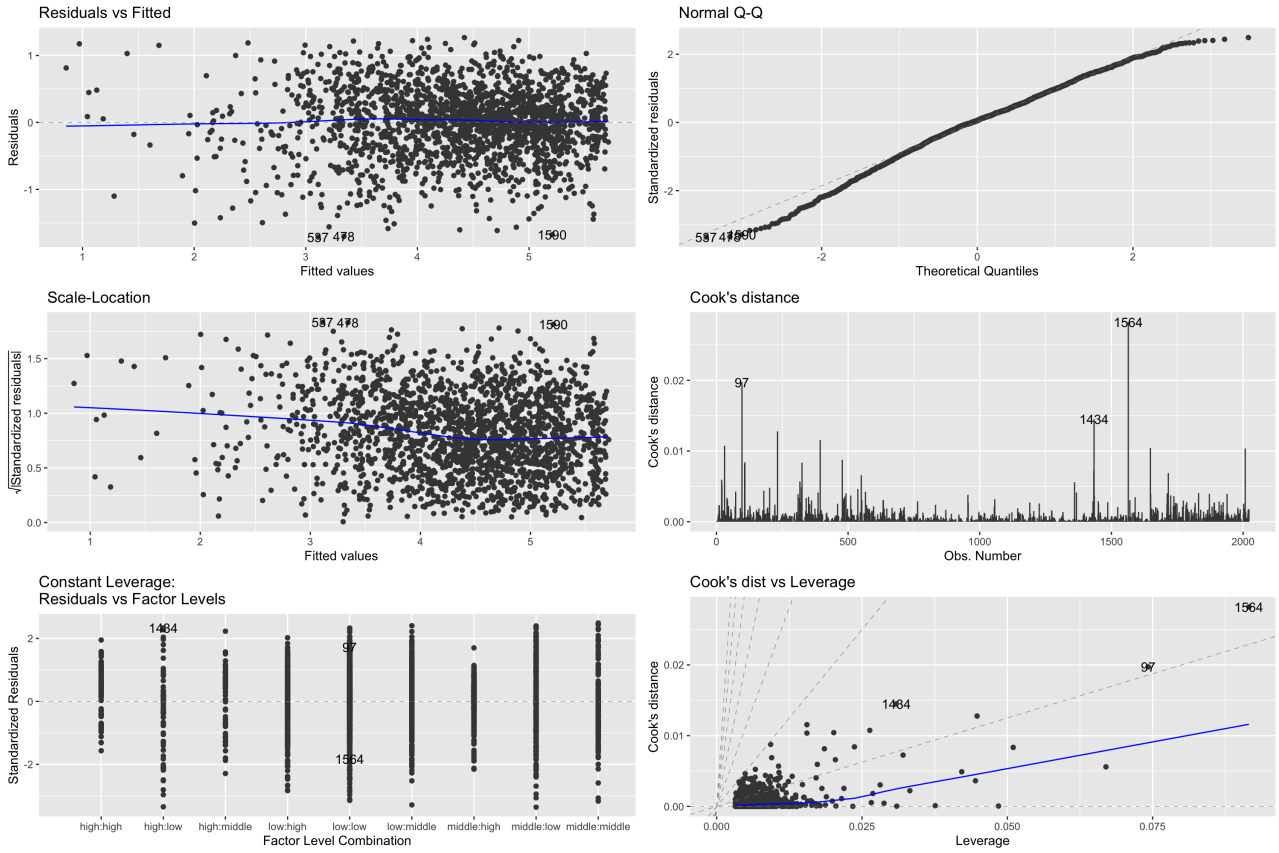

Figure 7: Diagnostic plots for Model B

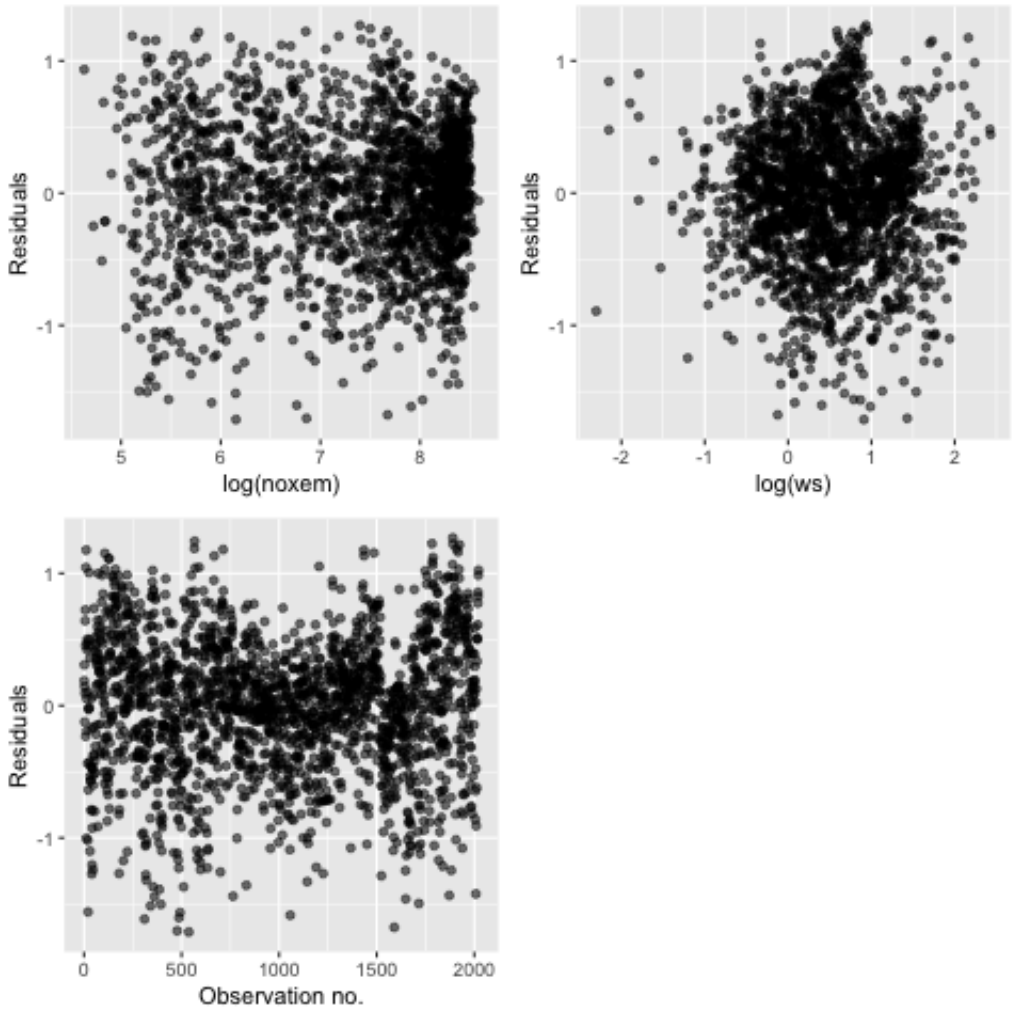

Figure 8: Residuals vs explanatory variable plots

- The results are an improvement over the previous results. The adjusted R-squared goes up from 0.62 to 0.70. The BIC value improves from 3552.864 to 3108.229.
- The residuals vs. fitted plot shows no discernible pattern.
- The q-q plot is approximately normal except at the tails. There are observations which deviate from the normal distribution especially at the lower quantiles which pose a worry. This is addressed below.
- The outlier detection plots shows three points with high Cook's distance and/or high leverage: 1564, 97, and 1464, which we will look into below. Note that apart from 1564 these are different from before.
- The cut points for the categorical variables involve a choice of threshold. Though we see the same patterns with the other thresholds we tried, in general results will differ if the categorical variables are defined differently.
- The residual vs. observation order plot looks flatter than before. This may be due to the quarterly indicators or the other added variables to the model.

Model C

- We run the model from above using robust M-estimation with a bi-squared weighting function.

- This allows us to check whether the previous results were being influenced by outliers.
- This is justified as there were three observations (1564, 97, and 1464) which had either high-leverage and/or high Cook's distance.
- We are also concerned about whether the points at the tails of the distribution in the q-q plot are unduly influencing our results.
- We run this model and get the results and diagnostics given below. We see that the results are not substantially different. This suggests that our previous results were not being unduly influenced by outliers.
- The BIC measure has deteriorated from 3108.229 to 3114.651, but this is not an issue as we are not comparing models with different variables.
- The residuals plotted against the fitted values look randomly scattered. This suggests that heteroskedascity is less of a concern.
- The normal plot has not improved at the tails. This remains an important limitation of our model.
- We note from the summary table of weights given above that the lowest weight is 0.12. The median and mean of the weights are 0.95 and 0.90 respectively. We note the lowest weight points returned by this regression and can make them available to the client to see whether these points are valid or whether they are data collection errors.
- The two key equations in interpreting the regression outputs are:

  1. $-1.2569 + 0.1271log(noxem) - 0.6188log(ws)$
  2. $0.5290 + 0.1271log(ws)$

  which are the partial derivatives of $log(nox)$ with respect to $log(ws)$ and $log(noxem)$ respectively, and they represent the % change in nox for 1% increase in ws and $noxem$ respectively, at the given levels of those variables.

```
Call: rlm(formula = lnox ~ lws * lnoxem + lws2 + categ_ws + categ_noxem +
    I(obs_no > n_quarter & obs_no <= 2 * n_quarter + 1) + I(obs_no >
    2 * n_quarter + 1 & obs_no <= 3 * n_quarter) + I(obs_no >
    3 * n_quarter), data = df, psi = psi.bisquare)
Residuals:
    Min      1Q  Median      3Q     Max
-1.7682 -0.3263  0.0150  0.3157  1.2680

Coefficients:
                      Value    Std. Error  t value
(Intercept)            1.0560   0.1787      5.9099 ***
lws                   -1.2569   0.1258     -9.9934 ***
lnoxem                 0.5290   0.0278     19.0263 ***
lws2                  -0.3094   0.0259    -11.9247 ***
categ_wsmiddle        -0.0425   0.0449     -0.9474
categ_wshigh          -0.0934   0.0893     -1.0453
categ_noxemmiddle      0.0807   0.0493      1.6354
categ_noxemhigh        0.1641   0.0609      2.6948 **
I(quarter_2)          -0.4314   0.0328    -13.1340 ***
I(quarter_3)          -0.3720   0.0332    -11.1921 ***
I(quarter_4)          -0.0185   0.0331     -0.5611
lws:lnoxem             0.1271   0.0176      7.2244 ***
```

```
---
Signif. codes 'p-value': 0 *** 0.001 ** 0.01 * 0.05 .
Residual standard error: 0.4714 on 2010 degrees of freedom
Robust regression weight summary
    Residual              Weights
    Min.   :-1.76816   Min.   :0.1290
    1st Qu.:-0.32631   1st Qu.:0.8655
    Median : 0.01500   Median :0.9590
    Mean   :-0.02106   Mean   :0.9026
    3rd Qu.: 0.31572   3rd Qu.:0.9914
    Max.   : 1.26805   Max.   :1.0000
```

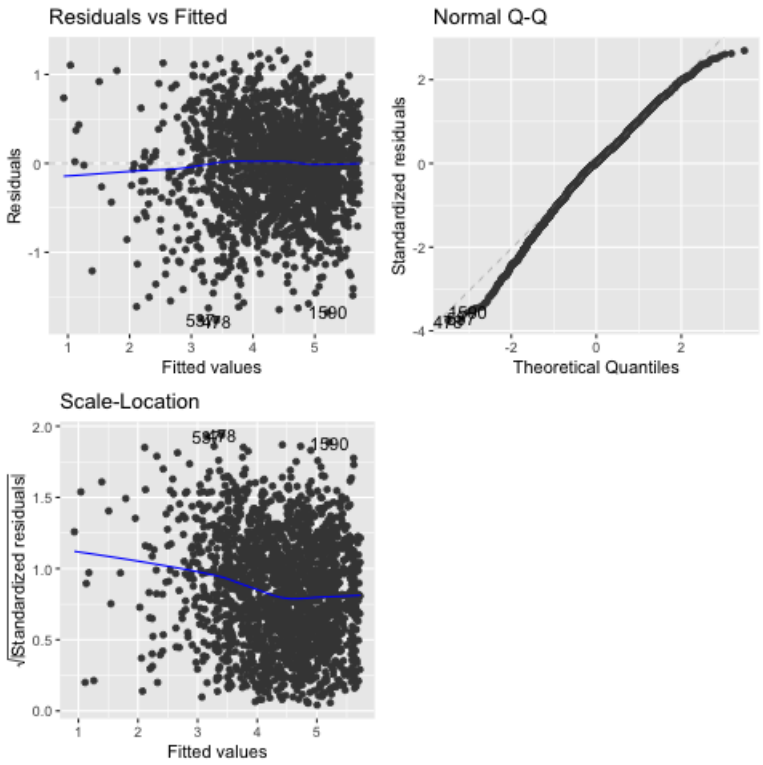

Figure 9: Diagnostic plots for Model C

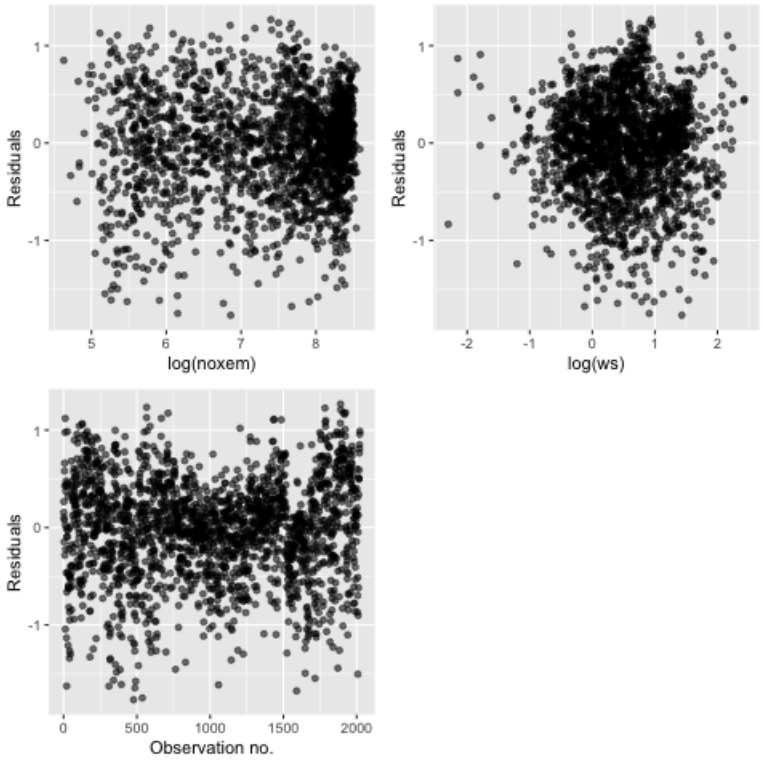

Figure 10: Residuals vs explanatory variable plots

Recommended Model

Model C is our final recommended model. It downweights potential outliers and prevents them from unduly influencing the analysis. The residuals look randomly scattered when plotted against the fitted values, and the BIC value seems reasonable. The results show that $log(ws)$ and $log(ws)^2$ both have strongly statistically significant effects on $log(nox)$, suggesting a non-linear relationship between changes in ws and changes in nox. $log(noxem)$ also has a strongly statistically significant effect on $log(nox)$. The interaction $log(noxem) * log(ws)$ is strongly significant, suggesting that the effect of $log(noxem)$ depends on the value of $log(ws)$ and vice versa. The quarter 2 and quarter 3 dummies are statistically significant, suggesting that the time of year influences $log(nox)$. There is an also an effect of the highest category wind-speed dummy on $log(noxem)$, but this does not have as high magnitude of a t-statistic as the other significant coefficients. The major limitation of our model is that the qq-plot suggests there are observations with very low or very high residuals which are unlikely to have come from a normal distribution.

Testing Coefficients

We can use the F-test to test the null hypothesis of same coefficients. In the following we first implement the F-test to show whether the coefficients of quarter two and three dummy variables are the same.
Unrestricted model

$$\begin{aligned} log(nox)_i = \alpha + \beta_1 log(noxem)_i + \beta_2 log(ws)_i + \beta_3 log(noxem)_i log(ws)_i + \beta_4 log(ws)_i^2 \\ + \beta_5 D_{mid,ws,i} + \beta_6 D_{high,ws,i} + \beta_7 D_{mid,noxem,i} + \beta_8 D_{high,noxem,i} \\ + \beta_9 D_{quarter_2,i} + \beta_{10} D_{quarter_3,i} + \beta_{11} D_{quarter_4,i} + \epsilon_i \end{aligned}$$

A model restricted by the null hypothesis, $\mathbf{H_0} : \beta_{\mathbf{9}} = \beta_{\mathbf{10}}$

$$\begin{aligned} log(nox)_i = \alpha + \beta_1 log(noxem)_i + \beta_2 log(ws)_i + \beta_3 log(noxem)_i log(ws)_i + \beta_4 log(ws)_i^2 \\ + \beta_5 D_{mid,ws,i} + \beta_6 D_{high,ws,i} + \beta_7 D_{mid,noxem,i} + \beta_8 D_{high,noxem,i} \\ + \beta_9 (D_{quarter_2,i} + D_{quarter_3,i}) + \beta_{11} D_{quarter_4,i} + \epsilon_i \end{aligned}$$

We denote the residual sum of squares of each model by RSS^U (unrestricted) and RSS^R (restricted). We compute the F-statistic as follows:

$$F = \frac{(RSS^R - RSS^U)/d}{(RSS^U/Degreeof Freedom)} = \frac{(527.13 - 526.06)/1}{(526.06/(2022 - 12))} = 4.0883$$

Since $F < F^{0.05}_{1,2010} = 4.7472$, we fail to reject the null hypothesis and conclude that the coefficients of dummy variables for quarters two and three are not different. We now implement for quarters three and four.
The unrestricted model is same as before, and a model restricted by the null hypothesis, $\mathbf{H_0} : \beta_{\mathbf{10}} = \beta_{\mathbf{11}}$ is:

$$\begin{aligned} log(nox)_i = \alpha + \beta_1 log(noxem)_i + \beta_2 log(ws)_i + \beta_3 log(noxem)_i log(ws)_i + \beta_4 log(ws)_i^2 \\ + \beta_5 D_{mid,ws,i} + \beta_6 D_{high,ws,i} + \beta_7 D_{mid,noxem,i} + \beta_8 D_{high,noxem,i} \\ + \beta_9 D_{quarter_2,i} + \beta_{10} (D_{quarter_3,i} + D_{quarter_4,i}) + \epsilon_i \end{aligned}$$

$$F = \frac{(RSS^R - RSS^U)/d}{(RSS^U/Degreeof Freedom)} = \frac{(557.14 - 526.06)/1}{((526.06/(2022 - 12))} = 118.75$$

which is much higher than the critical value, hence we reject the null hypothesis. Therefore we can conclude that the coefficients for the dummy variables for quarters three and four are significantly different but quarters two and three are not. This test procedure can be applied to any pair, or even higher number of coefficients.

Report

We have analysed the impact of Nitric Oxide (NOx) emissions by cars on the motorway, wind speed and humidity on the NOx concentration in the air. Some relationships we have found are:

- Humidity has no effect on the NOx concentration in the air.
- NOx concentration is affected seasonally. The concentration was on average higher in the first and fourth quarters than the second and third quarters, holding other variables constant. This relationship is strongly significant. But because the data only covers a single year, there is no guarantee that this relationship will hold every year.
- There are also strongly significant relationships between the NOx concentration, wind speed and the sum of NOx emission cars on the motorway. The interpretation is more complex:

    1. The faster the wind speed, the lower the NOx concentration assuming other variables remain constant. Also faster wind speed resulted in faster rate of decrease in NOx concentration level. However this rate of fall in the concentration was positively affected by current level of the sum of NOx emission of cars.
    2. The higher the sum of NOx emission of cars, the higher the NOx concentration, assuming other variables remain constant. And this rate of increase in NOx concentration is positively affected by wind speed, that is, higher current wind speed results in higher rate of change in concentration due to the increase in the sum of NOx emission of cars.

    Therefore NOx concentration is affected by wind speed through two channels, a negative direct impact and a positive indirect impact through the sum of NOx emission of cars, however the latter impact is relatively smaller in magnitude, hence we believe it is safe to conclude that wind speed negatively affects NOx concentration level.

Our model can be used to predict the the percentage change in *noxem* associated with a 1% change in wind-speed or sum of NOx emission at any fixed values of the other variables. We demonstrate some estimations of our model below:

- Given the current sum of NOx emission of cars is 500, and the current wind speed is 0.5 m/s, a 1% increase in wind speed, on average, reduces the NOx concentration ppb by 0.0391%.
- Given the current sum of NOx emission of cars is 500, and the current wind speed is 2 m/s, a 1% increase in wind speed, on average, reduces the NOx concentration ppb by 0.8964%, which is much lower than the above case with slower wind.
- Given the current sum of NOx emission of cars is 4000, and the current wind speed is 0.5 m/s, a 1% increase in wind speed, on average, increases the NOx concentration ppb by 0.2262%.
- Given the current wind speed is 0.5 m/s, a 1% increase in the sum of NOx emission of cars, on average, increases the NOx concentration ppb by 0.4411%.
- Given the current wind speed is 2 m/s, a 1% increase the sum of NOx emission of cars, on average, increases the NOx concentration ppb by 0.6179%.
- The expected proportionate decrease in NOx concentration in the second observation quarter relative to the first quarter is 0.3499.
- The expected proportionate decrease in NOx concentration in the third observation quarter relative to the first quarter is 0.3106.
- The expected proportionate decrease in NOx concentration in the fourth observation quarter relative to the first quarter is 0.0183.

Although we do not doubt the robustness of our model, due to the highly variable and noisy data, the predictions based on our model may lack precision. Two suggestions for future data collection efforts for improved modeling are:

- Sensor ID, date, and time of each measurement
- Collect if possible the no. of cars traveling on the highway each day
- Provide the units of each variable clearly stated in every dataset