Factors Influencing the NO2 Concentration in Air

Summary

This is an analytical study that seeks to explore the different factors that impact the log of nitrogen dioxide concentration in the air. Multiple regression model was used to examine the relationship between the log of no2 concentration and different explanatory variables. In addition to that, 'all subsets regression' was used to compare the performance of different models, after quantifying the performance using Baeysian Information Criterion (BIC) values. The reuslts of this study indicate that log of no2 concentration is positively correlated with the number of cars per hour on a particular road (p-value: < 2e-16), temperature difference between 2 and 5 metres above ground in degrees Celsius (p-value: 2.52e-07), wind direction measured in degrees between 0 and 360 (p-value: 0.0417) and, increment in day number from October 1, 2001 (p-value: 0.0179) Furthermore, the log of no2 concentration in air was also found to be negatively correlated with the air temperature 2 metres above ground in degrees Celsius (p-value: 1.23e-07) and, square root of wind speed measured in metres/second (p-value: < 2e-16)

Background

Nitrogen dioxide concentration in the environment is an important concern that needs to be investigated. This is because increased concentration of nitrogen dioxide in air has been associated with many different health problems. Several scientific studies have been conducted to explore the different factors that might contribute to a higher concentration of no2 in air. A study conducted by Donnelly and colleagues found a strong impact of wind direction and wind speed on the concentration of nitrogen dioxide in air(Donnelly et al.). Another study by Joanna in 2018 that looked at the concentration of no2 in air in a city in Polland, found the concentration of no2 in air to be strongly related with both traffic volume and wind speed (Kamińska).

The current analytical study seeks to explore the relationship between log of no2 concentration in air and other traffic volume and meterological factors. The purpose of this study is to find the different factors that might be able to explain the variability in the log of no2 concentration in air, in addtion to factors such as wind speed, traffic volume, etc. The data for this analytical study is taken from a Norwegian study on air pollution, which was conducted at Alnabru in Oslu, Norway from October 2001 to August 2003. The original dataset consists of 12 different variables and 500 observations. However, all the variables were not used for the purposes of the analyses in the current study. This is because some of the variables in the original dataset are categorical variants of other numerical variables. Therefore, the categorical variables that contained redundant information, which was better explained and expressed in more detail by their numerical counterparts, were excluded from the analyses. The response variable for this analysis is the logarithm of nitrogen dioxide particle concentration. Below is a list of the explanatory variables that were used in the analyses:

• Logarithm of number of cars per hour at a particular road (lcph)

• Air temperature 2 metres above ground in degrees Celsius (temp)

• Wind speed in metres/second (wind_speed)

• Temperature difference between 2 and 5 metres above ground in degrees Celsius (temp_diff)

• Wind direction expressed in degrees between 0 and 360 (wind_dir)

• Hour of the day at the time of observation (hour_day)

• Day number from October 1, 2001 (day_number)

Below is a list of the categorical variables that were excluded because they contained the same information as their numerical counterparts. Since these categorical variables did not contribute any additional information about the variability in log of no2 concentration in air (response variable), they were not used for the analyses in this study.

• Categorical version of the 'hour_day' variable (AM_PM)

• Categorical counterpart of 'temp_diff' variable (temp_diff_bin)

• Categorical variant of the 'wind_dir' variable (wind_dir_bin)

• Categorical for the 'wind_speed' variable (wind_speed_bin)

Methods:

This study uses multiple regression model to look at the relationship between the log of no2 concentration (response variable) and other explanatory variables. Multiple regression model allows us to map the relationship among differnt varuables, involving multiple explanatory variables effectively. Multiple regression model involves the following conditions/assumptions that should be satisfied:

(Refer to 'ggpairs plot before transformation' and 'ggpairs plot after transformation' in Appendix,)

Linearity: This condition was not satisfied intially in the original data. Linearity seemed to be a significant issue specifically for 'lcph' and 'wind_speed' variables. This implies that some transformation of data to help satisfy the linearity condition should be considered.

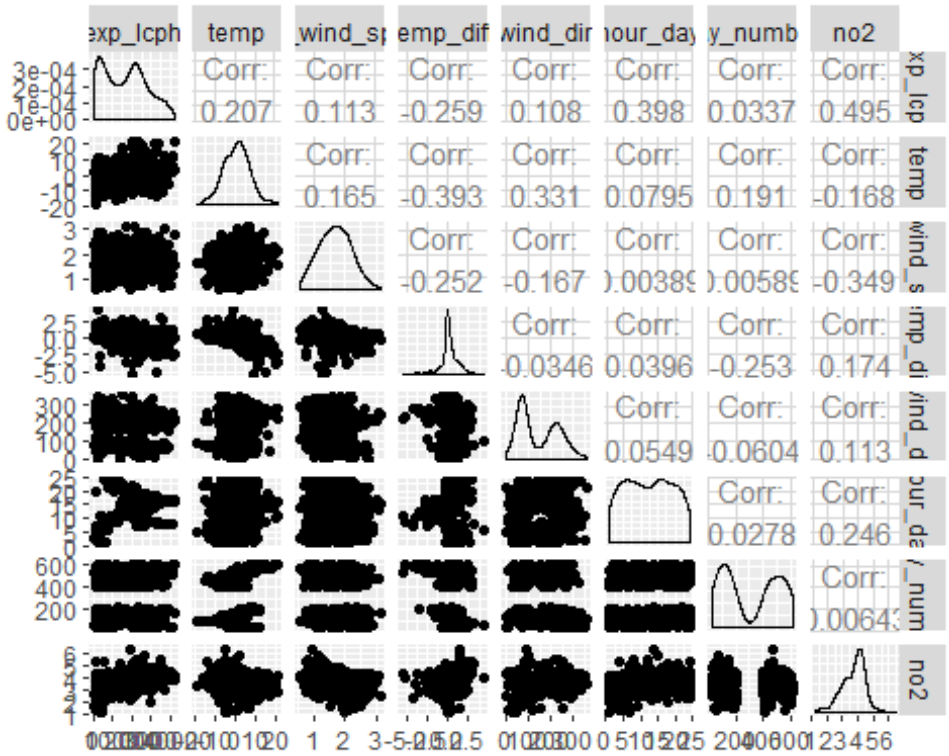
Outliers/Leverage Points: As seen in the spread of data points in the ggpairs plot, there were some potential outliers/ high influence leaverage points which need further investigation. This detailed investifation is included later in the report.

Equal Variance: In 'ggpairs plot before transformation', the density plot for 'lcph' was left skewed and the density plot for 'wind_speed' was right skewed. Therefore, equal variance condition was not satisfied for 'lcph' and 'wind_speed' variables. Equal variance for all the other variables looked fine.

Multicollinearity: Since none of the correlation coefficients in the 'ggpairs plot' were greater than 0.5, multicollinearity between the different candidate explanatory variables did not seem to be an issue. However, a double check on multicollinearity was also performed by calculating the VIF values.

Independence: The observations in the dataset seem to be independent. However, there is not absolute certainity about the indepedndece of observations because there is no detailed information available on how exactly the observations were recorded.

After determining the transformations needed in the original dataset, following transformations were performed: 'lcph' was tranformed by using the exponential function and the tranformed variable was named 'exp_lcph' 'wind_speed' was transformed by taking square root and the transformed variable was named 'sqrt_wind_speed' After transformation, the conditions of linearity and equal variance were better satisfied as seen in the ggpairs plot after tranformation of variables. The 'ggpairs plot' after tranformations of the two variables, 'lcph' and 'wind_speed' is attached below.



Various visualization tools such as 'ggapirs plot' were used to examine the different characteristics of the dataset. To check for the distribution of observations for each of the variables, the residual plots for each of the explanatory variables were also examined (Refer to appendix: residual plots). After transformation of 'lcph' and 'wind_speed' variables, their residual plots looked more centered at 0, which implies that the conditions of linerity and equal variance were better satisfied after transformation of these variables. The new transformed variables are named 'exp_lcph' and 'sqrt_wind_speed' respectively.

Additionally, some diagnostic checks were performed to look for potential outliers/ high leverage points. To do that, plots for 'studentised residuals' and 'cook's distance' were also constructed (Refer to 'diagnostic plots' in Appendix). Any data-points that were farther from the cluster of observations in the 'studentised resduals' and 'cook's distance' plots were noted as suspicious observations. These observations could be potential outliers or high influence points which could impact the results of our analyses. The plots for studentised residuals and cook's distance are included in the appendix. The observations that were far from the other cluster of data points are thus flagged as suspicious are 125, 156, 371, 372, 429, 240 and 320. Another verstion of the 'nitrogen' dataset was created after excluding the suspicious observations, to see if the relationship between the log of no2 concentration and other explanantory variables remained consistent after the potential high influence points were excluded from the analyses.

Since multicollinearity negatively impacts the precision of our estimates, VIF (variance inflation factor) values for all the coefficient estimates were calculated. To do this, a model was fit with log of no2 concentration as the response variable and all the other variables (transformed) as explantory variables. No interaction terms were included in this model. Below is a summary of all the VIFs for all the explanatory variables.

```
##        exp_lcph            temp sqrt_wind_speed       temp_diff
##        1.328434        1.417418        1.143606        1.383365
##        wind_dir        hour_day      day_number
##        1.223034        1.228061        1.109006
```

Since The VIF values for all the variables in the model were quite small (less than 2), there were no potential issues with multicollinearity among the candidate explanatory variables.

Furthermore, all subsets regression was used to look at different candidate models and to select models that had the best performance. The performance of different models was quantified by calculating the BIC (Baeysian Information Criterion) values. A smaller BIC value corresponds to better performance of a model as compared to a larger BIC value. Similar BIC values for multiple models indicates similar performance across those models. This method allows the selection of a model that best represents the relationship between the different variables in the dataset, without overfitting. This was done twice, once without the excluding any observations and later after the exclusion of all the suspicious observations that were considered potential outliers/ high influence points.

Results:

Below is a summary of what is included in the top three models before exclusion of any suspicious observations. The BIC values and the explanantory variables included (in adtion to the 'no2' as response variable) in the these models are also listed. This summary result is from all subsets regression **before excluding any observations**. More detailed results are included in the appendix.

Models 4, 5 and 6 had roughly equal performance.

Model 4: explanatory variables included were -> exp_lcph, temp, sqrt_wind_speed, temp_diff (BIC = -300.70) Model 5: explanatory variables included were -> exp_lcph, temp, sqrt_wind_speed, temp_diff, day_number (BIC = -299.01) Model 6: explanatory variables included were -> exp_lcph, temp, sqrt_wind_speed, temp_diff, wind_dir, day_number (BIC = -297.01)

Below is a summary of what is included in the top three models after exclusion of all the suspicious observations. The BIC values and the explanantory variables included (in adtion to the 'no2' as response variable) in the these models are also listed. This summary result is from all subsets regression **after excluding any observations**. More detailed results are included in the appendix.

Models 4, 5 and 6 have roughly equal performance.

Model 4: exp_lcph, temp, sqrt_wind_speed, temp_diff (BIC = -305.04) Model 5: exp_lcph, temp, sqrt_wind_speed, temp_diff, day_number (BIC = -304.73) Model 6: exp_lcph, temp, sqrt_wind_speed, temp_diff, wind_dir, day_number (BIC = -302.80)

The summaries from the top three models from each of the analyses, before and after the exclusion of suspicious observations are included in detail in the appendix. The results from the summaries of these models are as follows: There is a consistent and strong evidence that:

As the number of cars per hour at a particular road increase, the log of concentration of nitrogen dioxide (particle) also increases. The log of concentration of nitrogen dioxide (particle) is 4.036e-04 times higher with a one unit increase in the number of cars per hour as a particular road (p-value: <2e-16), assuming that all the other variables are kept constant.

As the air temperature 2 metres above ground (degrees Celsius) increase, the log of concentration of nitrogen dioxide (particle) decreases. The log of concentration of nitrogen dioxide is 2.352e-02 times lower with a one unit increase in the air temperature 2 metres above ground (degrees Celsius) (p-value: 1.23e-07), assuming that all the other variables are kept constant.

As the square root of wind speed (metres/second) increase, the log of concentration of nitrogen dioxide (particle) decreases. The log of concentration of nitrogen dioxide (particle) is 4.715e-01 times lower with a one unit increase in the square root of wind speed (metres/second) (p-value: <2e-16), assuming that all the other variables are kept constant.

As the temperature difference between 2 and 5 metres above ground (degrees Celsius) increase, the log of concentration of nitrogen dioxide (particle) also increases. The log of concentration of nitrogen dioxide (particle) is 1.368e-01 times higher with a one unit

There is moderate evidence that:

As the wind direction increases (degrees between 0 and 360), the log of concentration of nitrogen dioxide (particle) also increases. The log of concentration of nitrogen dioxide

(particle) is 2.352e-02 times higher with a one unit increase in the wind direction (degrees between 0 and 360) (p-value: 0.0417), assuming that all the other variables are kept constant.

As the day number from October 1, 2001 increase, the log of concentration of nitogen dioxide (particle) also increases. The log of concentration of nitrogen dioxide (particle) is 2.990e-04 times higher with a one unit increase in the day number from October 1, 2001 (p-value: 0.0179), assuming that all the other variables are kept constant.

Discussion

The results of this current study were consistent with the findings from the previous research studies. The findings of this analytical study were similar to what was found in the research study done by Donelly and colleagues. Both these studies established some sort of relationship between log no2 concentration in air with both wind speed and wind direction. However, in the present study, sqruare root of wind speed was more strongly correlated with the log of no2 concentration in air, than wind direction. While square root of wind speed was negatively correlated with log of no2 concentration, wind direction had a positive correlation. The findings of this study were also consistent with the findings of Joanne's research study that found no2 concentration to be related with both wind speed and traffic volume. Similarly, in this study, we found a positive correlation between log of no2 concentration and number of cars at a prticular road per hour. In addition to these factors, this study also found log of no2 concentration to be related with some temperature dependent factors such as air temperature 2 metres above ground (negative correlation) and temperature difference between 2 and 5 metres above ground (positive correlation). Furthermore, as the number of days from October 1, 2001 increased, the log of no2 concentration in air also increased, indicating an no2 concentration increase in air with time. The association of these different factors with log of no2 concentration provides a useful insight into the different influence factors of no2 concentration in air. Information about the different factors that impact no2 concentration would be useful in exploring these factors individually in detail. This could further help in lowering the level of no2 concentration in air.

The findings from this study suggest that most of the variability in log of no2 concentration is explanied by the square root of wind speed, temperature 2 metres above ground, temperature difference between 2 and 5 metres above ground, and the number of cars per hour at a particular road. The results about the correlations between the log of no2 concentration in air and different factors discussed above did not change when the suspicious observations (outliers/high influence points) were excluded from the analyses. This indicates that the findings of these analyses are consistent and do not depend on the details of the analyses. Based on these findings, useful future experiments could involve studying each of the factors that have been shown to be correlated with log no2 concentration, more closely. This would also allow us to exlore if manipulations in each of those factors could help in decreasing the overall no2 concentration in air.

References

Donnelly, Aoife, et al. “Application of Nonparametric Regression Methods to Study the Relationship between NO2 Concentrations and Local Wind Direction and Speed at Background Sites.” Science of The Total Environment, vol. 409, no. 6, 2011, pp. 1134–1144., doi:10.1016/j.scitotenv.2010.12.001.

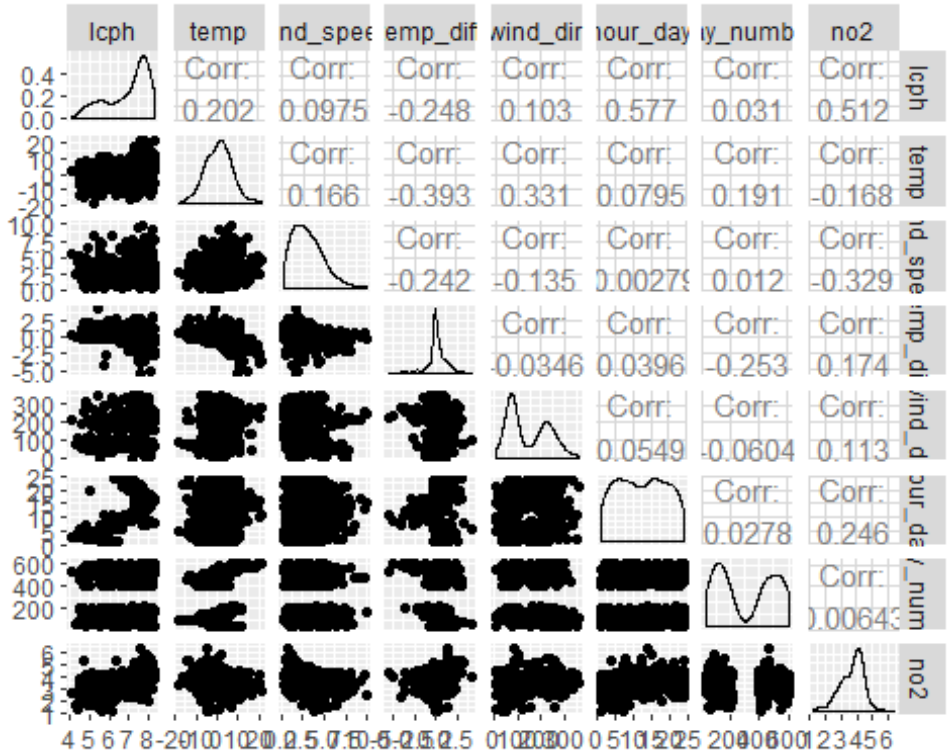
Kamińska, Joanna. “Probabilistic Forecasting of Nitrogen Dioxide Concentrations at an Urban Road Intersection.” Sustainability, vol. 10, no. 11, 2018, p. 4213., doi:10.3390/su10114213.

Data source: http://lib.stat.cmu.edu/datasets/

Appendix

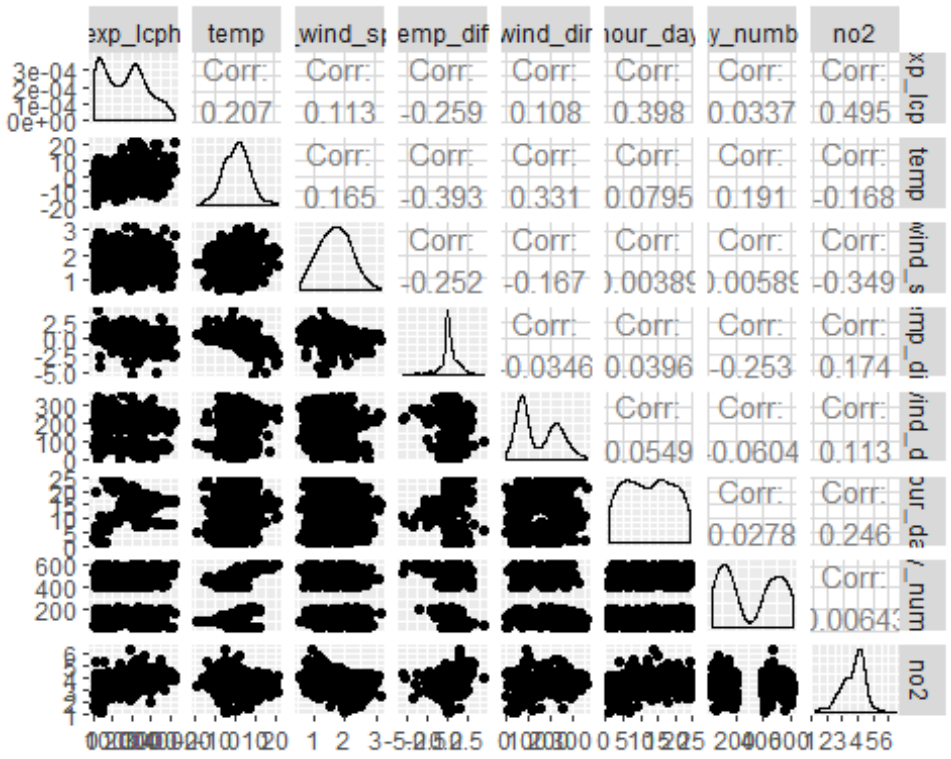
ggpairs plot before any transformation of variables

Below I am including ggpairs plots, with ‘no2’ as the response variable and all the variables (being considered) as explanatory variables.



ggpairs plot after transformation of variables

Below is a ggpairs plot after the following transformations in the variables are made: ‘lcph’ tranformed by using exponential function and named ‘exp_lcph’ ‘wind_speed’ transformed after taking the square root and named ‘sqrt_wind_speed’

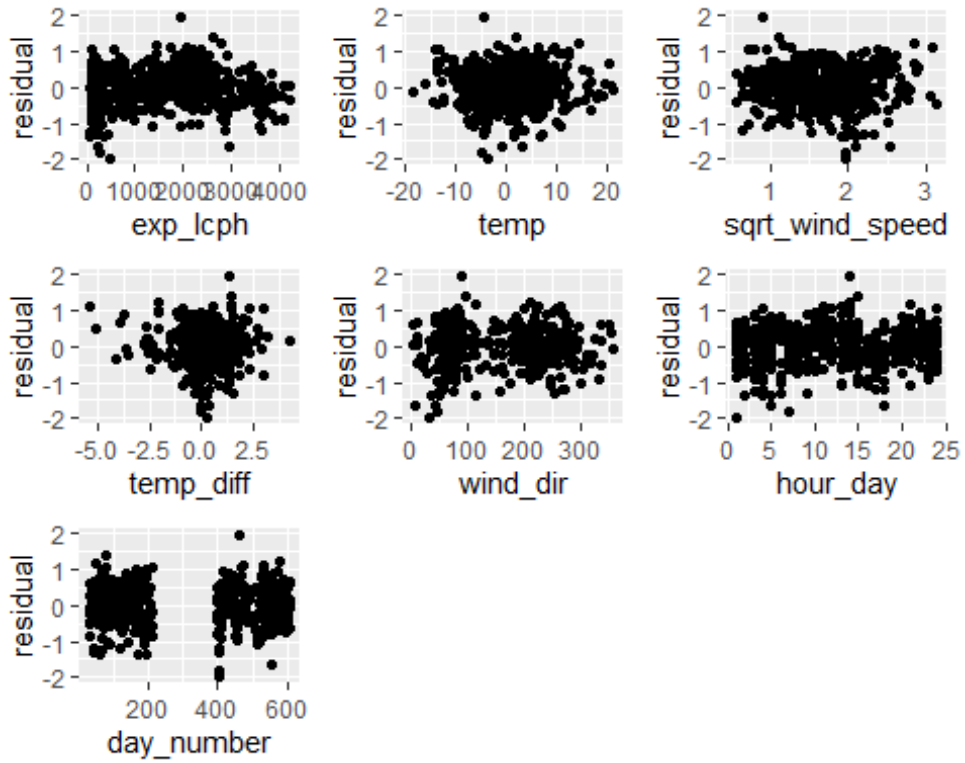


Residual plots

Below is a model with no2 as the response variable and all the other variables as explanatory variables (after tranformation).

```
##
## Call:
## lm(formula = no2 ~ exp_lcph + temp + sqrt_wind_speed + temp_diff +
##     wind_dir + hour_day + day_number, data = nitrogen)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94594 -0.33817  0.03111  0.38314  1.90634
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.6409123  0.1244923  29.246  < 2e-16 ***
## exp_lcph          0.0004017  0.0000243  16.528  < 2e-16 ***
## temp             -0.0235637  0.0043926  -5.364 1.25e-07 ***
## sqrt_wind_speed -0.4714311  0.0498840  -9.451  < 2e-16 ***
## temp_diff         0.1358994  0.0265796   5.113 4.55e-07 ***
## wind_dir          0.0006279  0.0003077   2.041   0.0418 *
## hour_day          0.0007296  0.0039215   0.186   0.8525
## day_number        0.0002978  0.0001261   2.361   0.0186 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.5377 on 492 degrees of freedom
## Multiple R-squared:  0.4939, Adjusted R-squared:  0.4867
## F-statistic:  68.6 on 7 and 492 DF,  p-value: < 2.2e-16
```

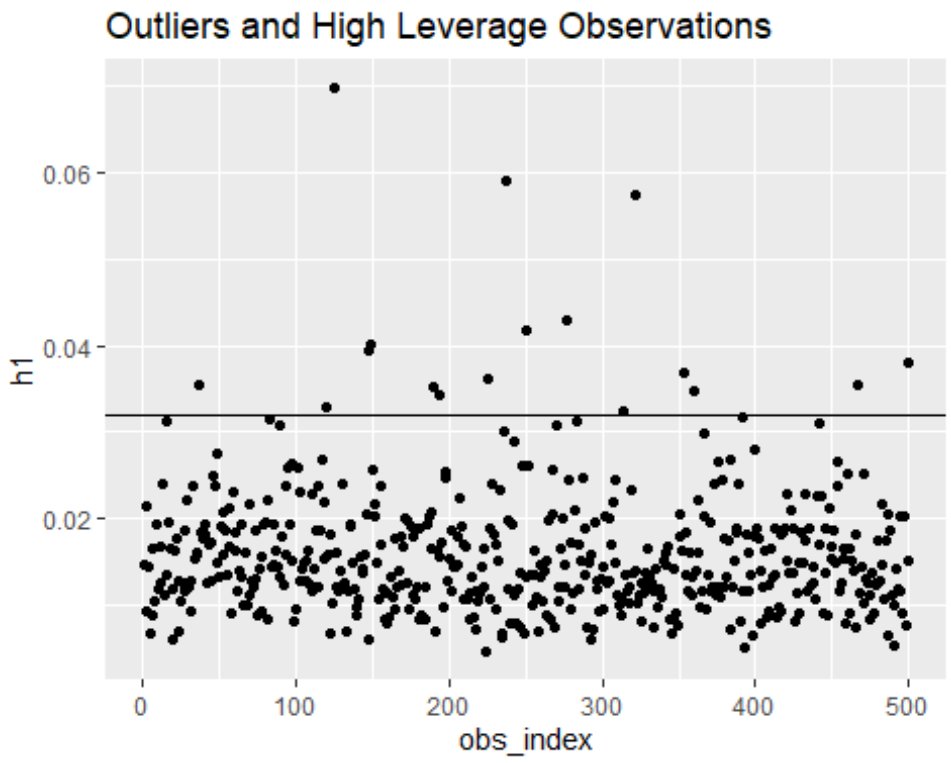


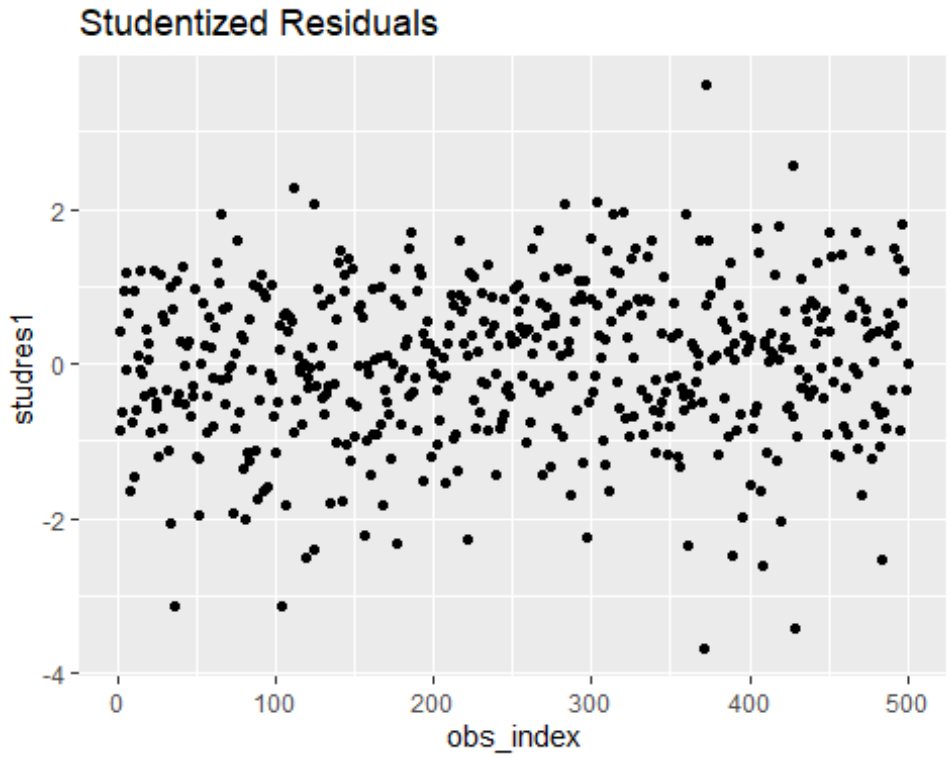
The residual plots for all the explanatory variables above look roughly centered at 0. However, there might be some potential outliers as some of the data-points in the residual plots above are far from the center cluster of observations.

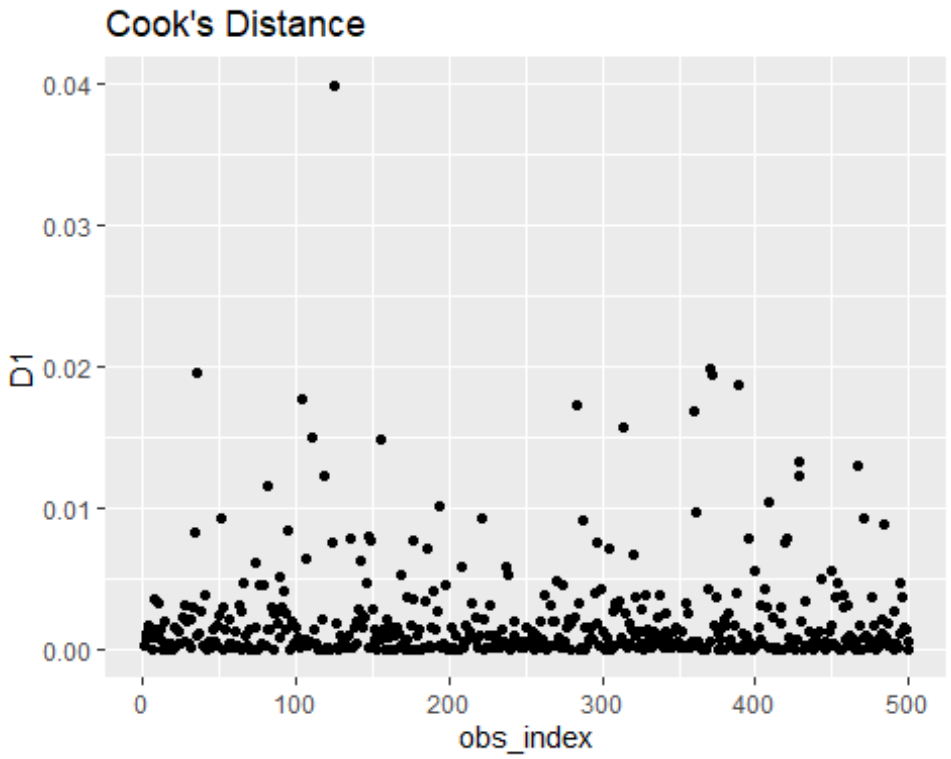
Diagnostic plots for identifying potential outliers and high levearge points

Diagnostic plots such as plots for 'studentised residuals' and 'cook's distance' are included below:

```
## [1] 0.032
```


Outliers and High Leverage Observations
h1
0.06
0.04
0.02
0
100
200
300
400
500
obs_index


Studentized Residuals
studres1
2
0
-2
-4
0
100
200
300
400
500
obs_index



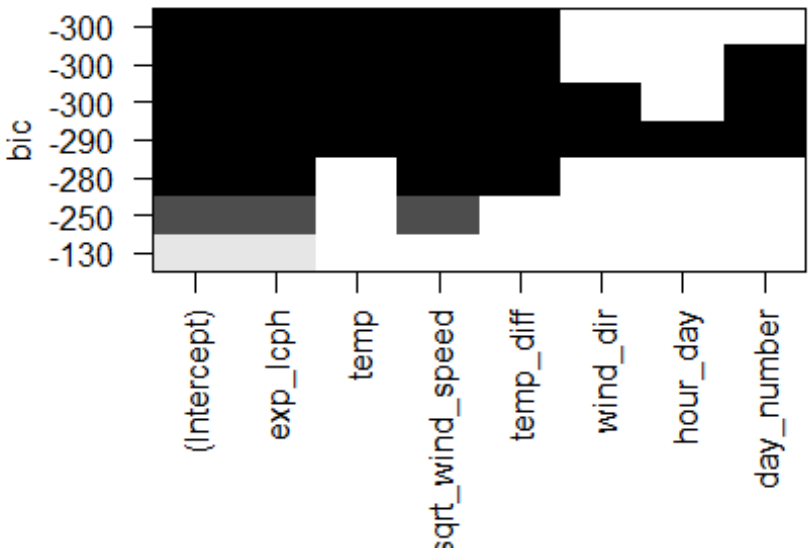
Model Summary after excluding suspicious observations

```
## # A tibble: 493 x 19
##            no2        lcph          temp wind_speed temp_diff wind_dir
##          <dbl>       <dbl>         <dbl>      <dbl>     <dbl>    <dbl>
##  1 3.71844     7.6912        9.200000000        4.8      -0.1     74.4
##  2 3.10009000  7.69894       6.4                3.5      -0.3     56
##  3 3.31419     4.812180000  -3.7                0.9      -0.1    281.3
##  4 4.38826     6.95177      -7.2                1.7       1.2     74
##  5 4.3464      7.51806      -1.3                2.6      -0.1     65
##  6 4.160440000 7.67183       2.6                1.6       0.3    224.2
##  7 4.0127700   5.52545      -7.9                1.6       0.3    211.9
##  8 2.15176     4.68213      -4.100000000        3.8      -0.1     63.1
##  9 3.157       7.15618     -12.7                5.2      -0.1     64.5
## 10 2.37955     4.74493      -1.6                3         0.4     58.3
## # ... with 483 more rows, and 13 more variables: hour_day <dbl>,
## #   day_number <dbl>, AM_PM <chr>, temp_diff_bin <chr>,
## #   wind_dir_bin <chr>, wind_speed_bin <chr>, sqrt_wind_speed <dbl>,
## #   exp_lcph <dbl>, residual <dbl>, obs_index <int>, h1 <dbl>,
## #   studres1 <dbl>, D1 <dbl>

##
## Call:
## lm(formula = no2 ~ exp_lcph + temp + sqrt_wind_speed + temp_diff +
##     wind_dir + hour_day + day_number, data = nitrogen_minus_suspicious)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1.7003 -0.3340  0.0331  0.3821  1.3728
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.609e+00  1.195e-01  30.189  < 2e-16 ***
## exp_lcph         3.992e-04  2.336e-05  17.086  < 2e-16 ***
## temp            -2.446e-02  4.207e-03  -5.814 1.11e-08 ***
## sqrt_wind_speed -4.440e-01  4.816e-02  -9.219  < 2e-16 ***
## temp_diff        1.430e-01  2.622e-02   5.453 7.92e-08 ***
## wind_dir         6.065e-04  2.952e-04   2.054  0.04048 *
## hour_day        -3.177e-04  3.752e-03  -0.085  0.93255
## day_number       3.230e-04  1.209e-04   2.672  0.00779 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5132 on 485 degrees of freedom
## Multiple R-squared:  0.5045, Adjusted R-squared:  0.4974
## F-statistic: 70.55 on 7 and 485 DF,  p-value: < 2.2e-16
```

All subsets regression without excluding any high influence/outliers observation

Below are the results from all subsets regression. These results will help me decide which models have roughly similar performance, and then pick a model that best represents the data.



```
## Subset selection object
## Call: regsubsets.formula(no2 ~ exp_lcph + temp + sqrt_wind_speed +
##     temp_diff + wind_dir + hour_day + day_number, data = nitrogen)
## 7 Variables  (and intercept)
##                 Forced in Forced out
## exp_lcph            FALSE      FALSE
## temp                FALSE      FALSE
## sqrt_wind_speed     FALSE      FALSE
## temp_diff           FALSE      FALSE
## wind_dir            FALSE      FALSE
## hour_day            FALSE      FALSE
## day_number          FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          exp_lcph temp sqrt_wind_speed temp_diff wind_dir hour_day
## 1  ( 1 ) "*"      " "  " "             " "       " "      " "
## 2  ( 1 ) "*"      " "  "*"             " "       " "      " "
## 3  ( 1 ) "*"      " "  "*"             "*"       " "      " "
## 4  ( 1 ) "*"      "*"  "*"             "*"       " "      " "
## 5  ( 1 ) "*"      "*"  "*"             "*"       " "      " "
## 6  ( 1 ) "*"      "*"  "*"             "*"       "*"      " "
## 7  ( 1 ) "*"      "*"  "*"             "*"       "*"      "*"
##          day_number
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"

## [1] -128.1289 -246.2105 -284.8711 -300.6967 -299.0130 -297.0104 -290.8310
```

Below are the summaries of the top 3 models (corresponding to the lowest BIC values) selected from the candidate models above

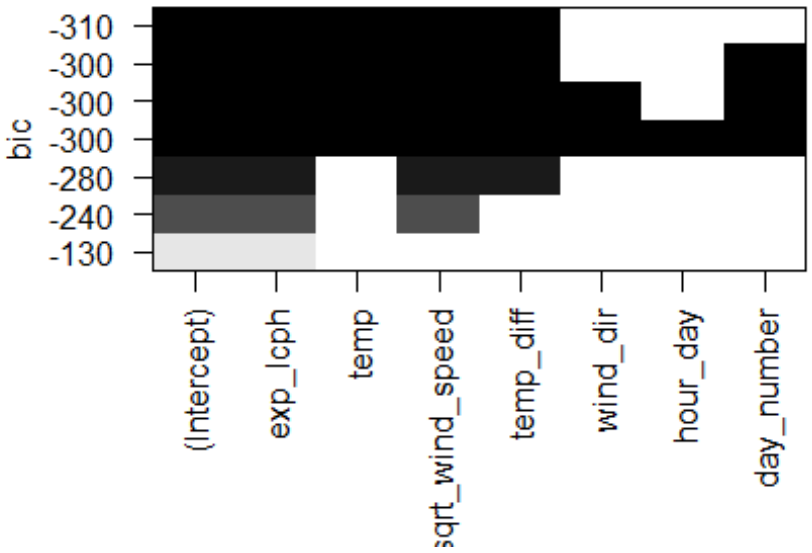
```
##
## Call:
## lm(formula = no2 ~ exp_lcph + temp + sqrt_wind_speed + temp_diff,
##     data = nitrogen)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96253 -0.32580  0.02497  0.37341  1.92608
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.877e+00  9.152e-02  42.358  < 2e-16 ***
## exp_lcph         4.048e-04  2.213e-05  18.288  < 2e-16 ***
## temp            -1.924e-02  4.073e-03  -4.723 3.03e-06 ***
```

```
## sqrt_wind_speed -5.024e-01  4.866e-02 -10.325  < 2e-16 ***
## temp_diff        1.276e-01  2.570e-02   4.964 9.51e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5408 on 495 degrees of freedom
## Multiple R-squared:  0.485,  Adjusted R-squared:  0.4808
## F-statistic: 116.5 on 4 and 495 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = no2 ~ exp_lcph + temp + sqrt_wind_speed + temp_diff +
##     day_number, data = nitrogen)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99385 -0.30999  0.02978  0.38651  1.87131
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.778e+00  1.023e-01  36.924  < 2e-16 ***
## exp_lcph         4.068e-04  2.208e-05  18.428  < 2e-16 ***
## temp            -2.021e-02  4.085e-03  -4.948 1.03e-06 ***
## sqrt_wind_speed -4.954e-01  4.860e-02 -10.193  < 2e-16 ***
## temp_diff        1.393e-01  2.621e-02   5.317 1.60e-07 ***
## day_number       2.654e-04  1.252e-04   2.121   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5389 on 494 degrees of freedom
## Multiple R-squared:  0.4896, Adjusted R-squared:  0.4845
## F-statistic: 94.78 on 5 and 494 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = no2 ~ exp_lcph + temp + sqrt_wind_speed + temp_diff +
##     wind_dir + day_number, data = nitrogen)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.95211 -0.33547  0.02766  0.38627  1.90590
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.646e+00  1.207e-01  30.209  < 2e-16 ***
## exp_lcph         4.036e-04  2.206e-05  18.292  < 2e-16 ***
## temp            -2.352e-02  4.383e-03  -5.367 1.23e-07 ***
## sqrt_wind_speed -4.715e-01  4.983e-02  -9.462  < 2e-16 ***
## temp_diff        1.368e-01  2.615e-02   5.229 2.52e-07 ***
## wind_dir         6.278e-04  3.074e-04   2.042   0.0417 *
```

```
## day_number       2.990e-04  1.258e-04   2.376   0.0179 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5372 on 493 degrees of freedom
## Multiple R-squared:  0.4939, Adjusted R-squared:  0.4877
## F-statistic: 80.19 on 6 and 493 DF,  p-value: < 2.2e-16
```

All subsets regression after excluding the outliers and high influence points

Below are the results for all subsets regression after excluding the potential outliers / suspicious values.




```
## Subset selection object
## Call: regsubsets.formula(no2 ~ exp_lcph + temp + sqrt_wind_speed +
##     temp_diff + wind_dir + hour_day + day_number, data =
nitrogen_minus_suspicious)
## 7 Variables  (and intercept)
##                 Forced in Forced out
## exp_lcph            FALSE      FALSE
## temp                FALSE      FALSE
## sqrt_wind_speed     FALSE      FALSE
## temp_diff           FALSE      FALSE
## wind_dir            FALSE      FALSE
## hour_day            FALSE      FALSE
## day_number          FALSE      FALSE
## 1 subsets of each size up to 7
```

```
## Selection Algorithm: exhaustive
##          exp_lcph temp sqrt_wind_speed temp_diff wind_dir hour_day
## 1  ( 1 ) "*"      " "  " "            " "       " "      " "
## 2  ( 1 ) "*"      " "  "*"            " "       " "      " "
## 3  ( 1 ) "*"      " "  "*"            "*"       " "      " "
## 4  ( 1 ) "*"      "*"  "*"            "*"       " "      " "
## 5  ( 1 ) "*"      "*"  "*"            "*"       " "      " "
## 6  ( 1 ) "*"      "*"  "*"            "*"       "*"      " "
## 7  ( 1 ) "*"      "*"  "*"            "*"       "*"      "*"
##          day_number
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"

## [1] -127.9703 -241.7518 -284.9793 -305.0423 -304.7299 -302.8023 -296.6090
```

Below are the summaries of the top 3 models (corresponding to the lowest BIC values), from the candidate models above, after the exclusion of suspicious observations

```
##
## Call:
## lm(formula = no2 ~ exp_lcph + temp + sqrt_wind_speed + temp_diff,
##     data = nitrogen)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96253 -0.32580  0.02497  0.37341  1.92608
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.877e+00  9.152e-02  42.358  < 2e-16 ***
## exp_lcph         4.048e-04  2.213e-05  18.288  < 2e-16 ***
## temp            -1.924e-02  4.073e-03  -4.723 3.03e-06 ***
## sqrt_wind_speed -5.024e-01  4.866e-02 -10.325  < 2e-16 ***
## temp_diff        1.276e-01  2.570e-02   4.964 9.51e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5408 on 495 degrees of freedom
## Multiple R-squared:  0.485,  Adjusted R-squared:  0.4808
## F-statistic: 116.5 on 4 and 495 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = no2 ~ exp_lcph + temp + sqrt_wind_speed + temp_diff +
##     day_number, data = nitrogen)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99385 -0.30999  0.02978  0.38651  1.87131
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.778e+00  1.023e-01  36.924  < 2e-16 ***
## exp_lcph         4.068e-04  2.208e-05  18.428  < 2e-16 ***
## temp            -2.021e-02  4.085e-03  -4.948 1.03e-06 ***
## sqrt_wind_speed -4.954e-01  4.860e-02 -10.193  < 2e-16 ***
## temp_diff        1.393e-01  2.621e-02   5.317 1.60e-07 ***
## day_number       2.654e-04  1.252e-04   2.121   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5389 on 494 degrees of freedom
## Multiple R-squared:  0.4896, Adjusted R-squared:  0.4845
## F-statistic: 94.78 on 5 and 494 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = no2 ~ exp_lcph + temp + sqrt_wind_speed + temp_diff +
##     day_number + wind_dir, data = nitrogen)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.95211 -0.33547  0.02766  0.38627  1.90590
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.646e+00  1.207e-01  30.209  < 2e-16 ***
## exp_lcph         4.036e-04  2.206e-05  18.292  < 2e-16 ***
## temp            -2.352e-02  4.383e-03  -5.367 1.23e-07 ***
## sqrt_wind_speed -4.715e-01  4.983e-02  -9.462  < 2e-16 ***
## temp_diff        1.368e-01  2.615e-02   5.229 2.52e-07 ***
## day_number       2.990e-04  1.258e-04   2.376   0.0179 *
## wind_dir         6.278e-04  3.074e-04   2.042   0.0417 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5372 on 493 degrees of freedom
## Multiple R-squared:  0.4939, Adjusted R-squared:  0.4877
## F-statistic: 80.19 on 6 and 493 DF,  p-value: < 2.2e-16
```