Predictive Emission Monitoring System

By: Anup, Ayush, Shouvik, Uday

Content

- Introduction
- Univariate Analysis
- Linear Model
- Transformed Model
- Weighted Least Square Model
- Model Selection

Introduction

With the increase in energy demands, the environment has suffered irreparable damages due to the emission of greenhouse gases, mainly carbon monoxide and nitrogen oxides. To limit these emissions, the Paris Convention on Climate Change was adopted by 196 countries.

we plan to analyze the area of research that predicts the emissions of these gases using statistical and machine learning techniques. Our focus is on conducting a multivariate regression analysis on the data provided by UCI, which contains eleven sensor measures aggregated over an hour from a gas turbine located in Turkey's northwestern region.

Our research aims to predict air pollutant emissions and extract insights from the process. We will be basing our research on an existing study titled "Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS'' by the Department of Computer Engineering at the Scientific and Technological Research Council of Türkiye.

Our goal is to develop new regression models and benchmark them against the existing ones.

Context

Gas turbine located in Turkey's north western region for the purpose of studying flue gas emissions

Predict the emission of greenhouse gases based on features measured via sensors

Using statistical methods to achieve the benchmark.

TÜBİTAK
Turkish Journal of Electrical Engineering & Computer Sciences
http://journals.tubitak.gov.tr/elektrik/
Research Article
Turk J Elec Eng & Comp Sci
(2019) 27: 4783 – 4796
© TÜBİTAK
doi:10.3906/elk-1807-87

Predicting CO and NO_x emissions from gas turbines: novel data and a benchmark PEMS

Heysem KAYA*, Pınar TÜFEKCİ, Erdinç UZUN
Department of Computer Engineering, Çorlu Faculty of Engineering, Namık Kemal University, Tekirdağ, Turkey

Received: 10.07.2018 • **Accepted/Published Online:** 08.07.2019 • **Final Version:** 26.11.2019

Abstract: Predictive emission monitoring systems (PEMS) are important tools for validation and backing up of costly continuous emission monitoring systems used in gas-turbine-based power plants. Their implementation relies on the availability of appropriate and ecologically valid data. In this paper, we introduce a novel PEMS dataset collected over five years from a gas turbine for the predictive modeling of the CO and NO_x emissions. We analyze the data using a recent machine learning paradigm, and present useful insights about emission predictions. Furthermore, we present a benchmark experimental procedure for comparability of future works on the data.

Key words: Predictive emission monitoring systems, CO, NO_x, exhaust emission prediction, gas turbines, extreme learning machine, database

1. Introduction

The increasing demand for energy had a double negative impact on the environment both through increasing deforestation and increasing carbon and flue gas emissions. Thanks to the increased environmental awareness in both public opinion and political circles, the Paris Convention on Climate Change[1] was adopted by 196 participating nations. The convention aimed to reduce the global greenhouse emissions and the signing parties have already passed rigorous environmental laws, including tax implementation for carbon emissions[2].

Being the core issue of the convention, air pollution poses a vital threat. The term "air pollutant" covers all substances which may harm living beings. The combustion processes of fossil fuels used in power plants and vehicles comprise the major portion of air pollution. NO_x ($NO_x = NO_2 + NO$) are considered the primary pollutants of the atmosphere, since they are responsible for environmental problems such as photochemical smog, acid rain, tropospheric ozone, ozone layer depletion, and eventually global warming [1]. In addition to these environmental catastrophes, they cause various health problems in humans exposed to high concentrations of these gases [1].

An important source of harmful pollutants (NO_x and CO) released in the atmosphere is the combustion process in the power industry [1, 2]. Therefore, there is a special concern on reducing the emissions from power plants. These emissions are restricted within certain limits by rigid environmental rules in different parts of the world [3]. European Union (EU) restricts the flue gas emissions (i.e. NO_x, CO, and dust) by Large Combustion

*Correspondence: kaya.heysem@gmail.com

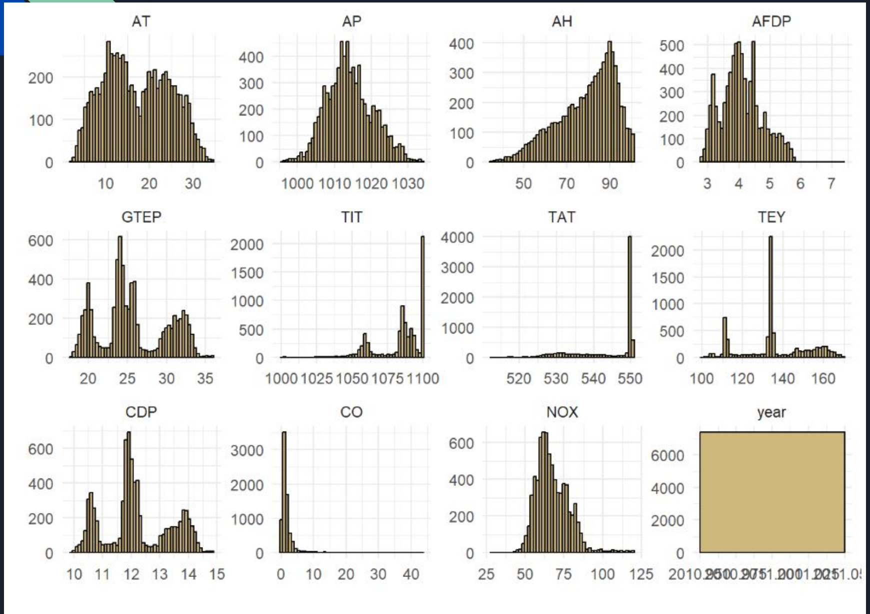
[1] Framework Convention on Climate Change (2015). Adoption of the Paris Agreement [online]. Website http://unfccc.int/resource/docs/2015/cop21/eng/109r01.pdf [accessed 26 July 2019].
[2] Carbon Tax Center [online]. Website http://www.carbontax.org/ [accessed 26 July 2019].

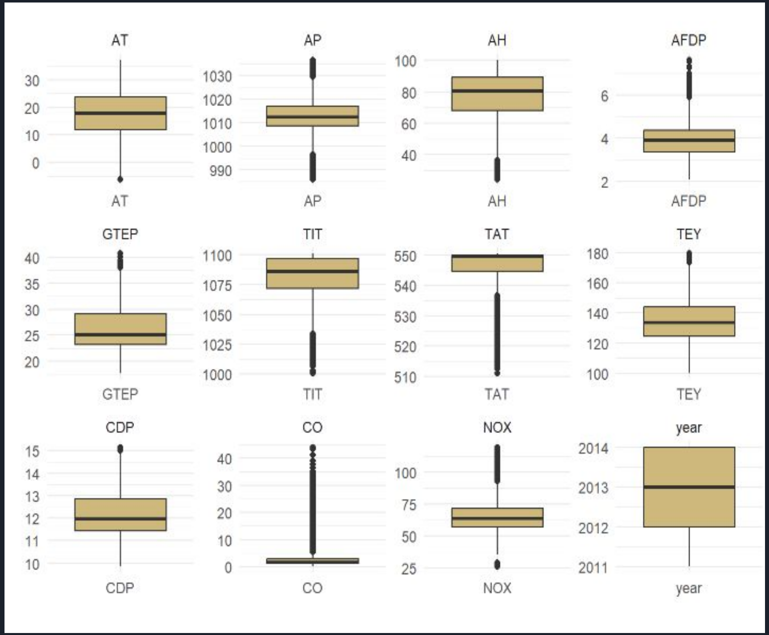
This work is licensed under a Creative Commons Attribution 4.0 International License.

Data

AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
4.5878	1018.7	83.675	3.5758	23.979	1086.2	549.83	134.67	11.898	0.32663	81.952
4.2932	1018.3	84.235	3.5709	23.951	1086.1	550.05	134.67	11.892	0.44784	82.377
3.9045	1018.4	84.858	3.5828	23.990	1086.5	550.19	135.10	12.042	0.45144	83.776
3.7436	1018.3	85.434	3.5808	23.911	1086.5	550.17	135.03	11.990	0.23107	82.505
3.7516	1017.8	85.182	3.5781	23.917	1085.9	550.00	134.67	11.910	0.26747	82.028
3.8858	1017.7	83.946	3.5824	23.903	1086.0	549.98	134.67	11.868	0.23473	81.748

Univariate Analysis





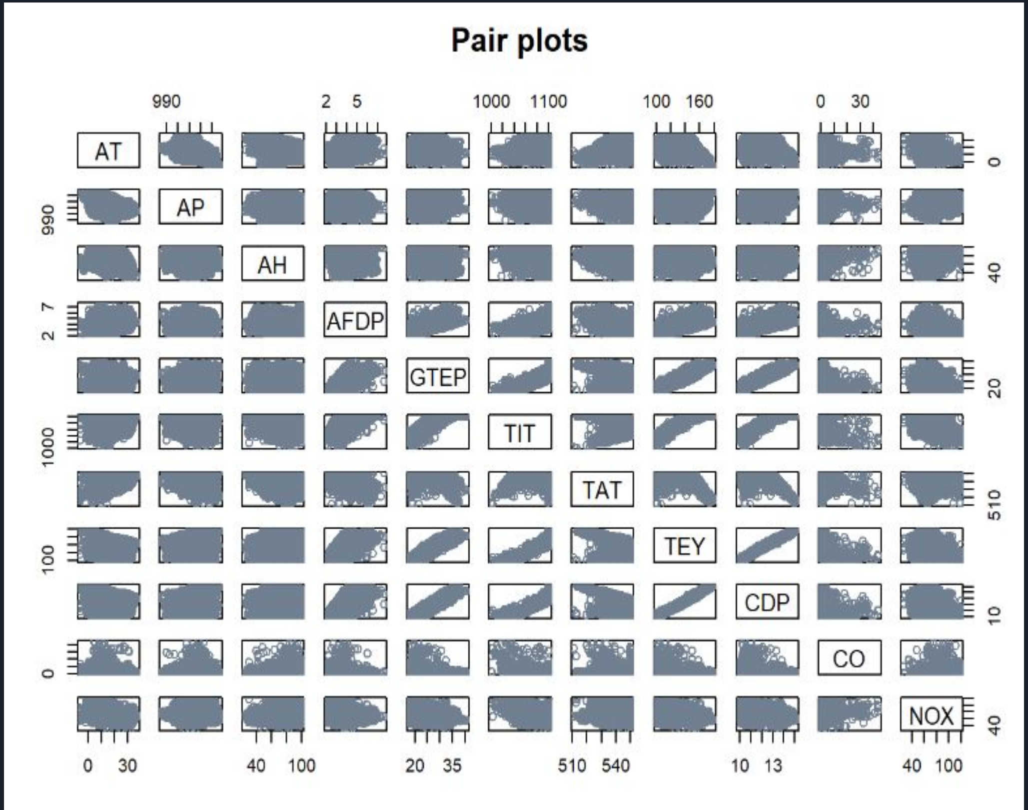
From the histogram plot, we can see that predictors like AP follow normal distribution where as predictor like AH is right skewed. Dependent variable CO is also right skewed transforming which can be beneficial.

The box plot shows that there are some outlier for predictors like AFDP,TIT,TAT. There are too many outliers so deleting could lead to loss of data. These are not significant either.

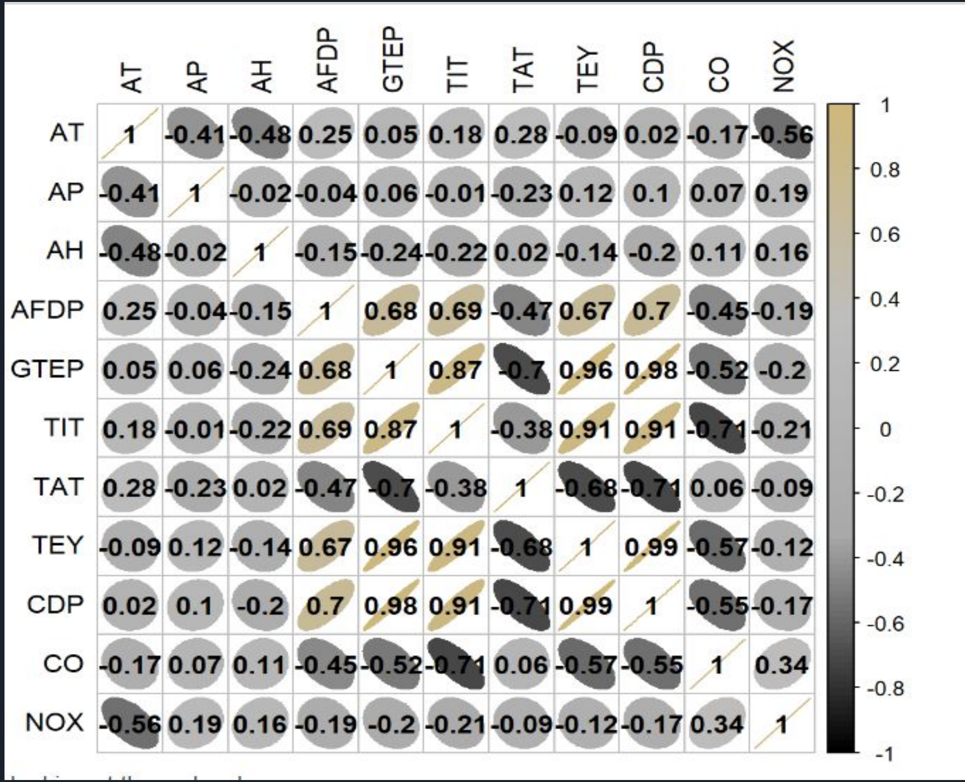
Pair Plots

From the pair plot , it very hard to infer any pattern as the plot looks like a blob for NOX vs other predictor.

For CO as response we can see the pattern for TEY and CDP features.



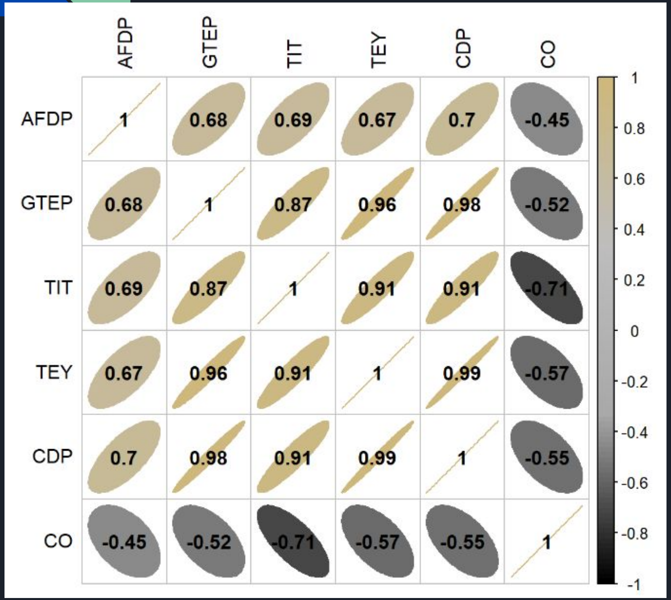
Correlation between Predictors and response Variable

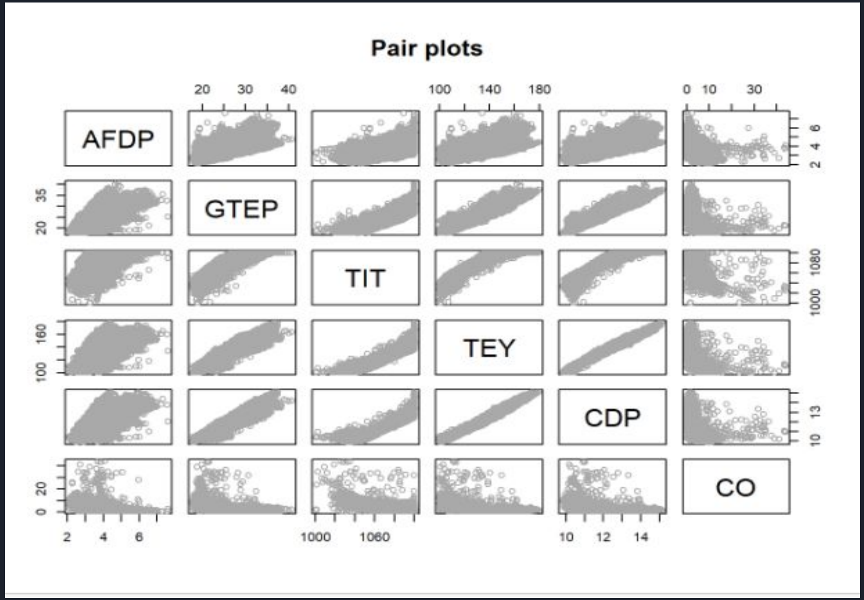


This plots shows there is high correlation between some features GTEP and TAT, CDP and TAT etc.

Analysis for CO

Multivariate Analysis





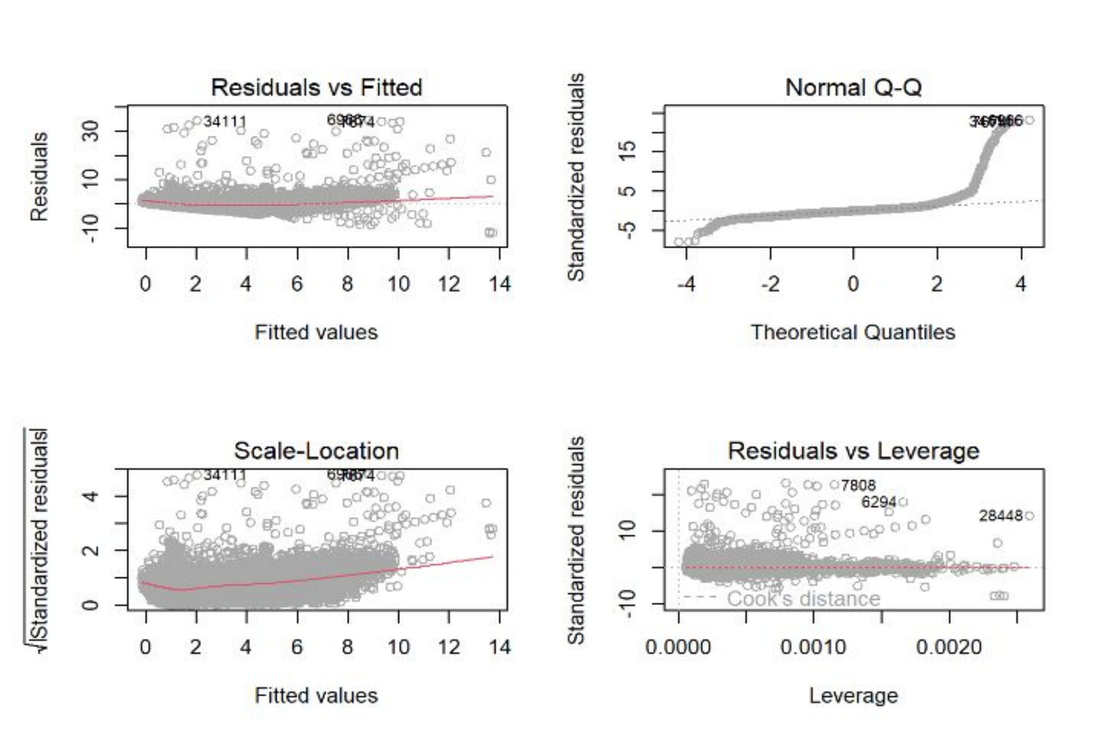
The pair plot doesn't show much pattern for features except for response CO with predictor GTEP and CDP. From the correlation matrix some correlations with the predictors could be observed for CO. There are also some high correlations observed between the independent features (Multicollinearity), some measures will be taken later on accordingly. One being CDP is removed as it is highly correlated with TEY and GTEP.

Linear Model

```
##
## Call:
## lm(formula = CO ~ . - NOX - year - CDP, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.924  -0.682  -0.093   0.532  34.798
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.407e+02  2.105e+00  66.846  < 2e-16 ***
## AT          -1.566e-02  2.273e-03  -6.888 5.74e-12 ***
## AP           5.350e-03  1.394e-03   3.838 0.000124 ***
## AH          -7.472e-03  6.744e-04 -11.080  < 2e-16 ***
## AFDP        -1.249e-01  1.594e-02  -7.833 4.90e-15 ***
## GTEP         1.427e-01  9.847e-03  14.488  < 2e-16 ***
## TIT         -6.695e-02  2.758e-03 -24.271  < 2e-16 ***
## TAT         -1.146e-01  3.088e-03 -37.105  < 2e-16 ***
## TEY         -8.310e-02  4.902e-03 -16.954  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.501 on 36724 degrees of freedom
## Multiple R-squared:  0.5599, Adjusted R-squared:  0.5598
## F-statistic:  5840 on 8 and 36724 DF,  p-value: < 2.2e-16
```

The linear model has all predictors significant. The model gave an Adjusted R-squared of 0.5598 which means it explains the variance decently.

Diagnostic Plot for Linear Model



These plots shows that gauss markov assumption is violated for normality,linearity and homoscedasticity, so we will try model transformation to see if it helps with normality,linearity and homoscedasticity. Homoscedasticity violation is not as severe.

Actual vs Fitted Diagnostic Plot

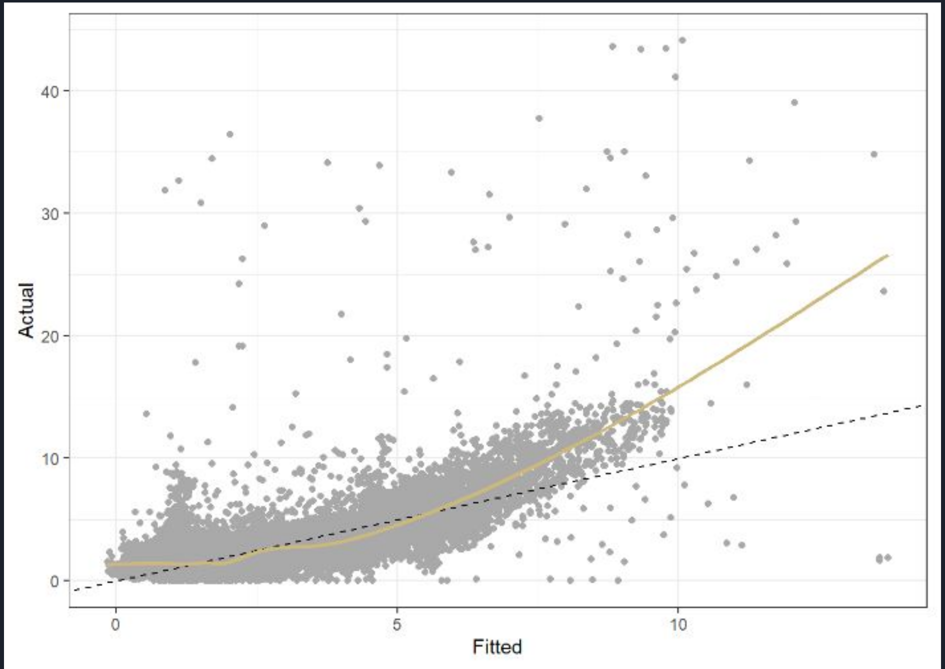
From the previous plots, some key violations of linear regression are Linearity and Normality.

Fitted vs Actuals shows parabolic curve, so relation looks like

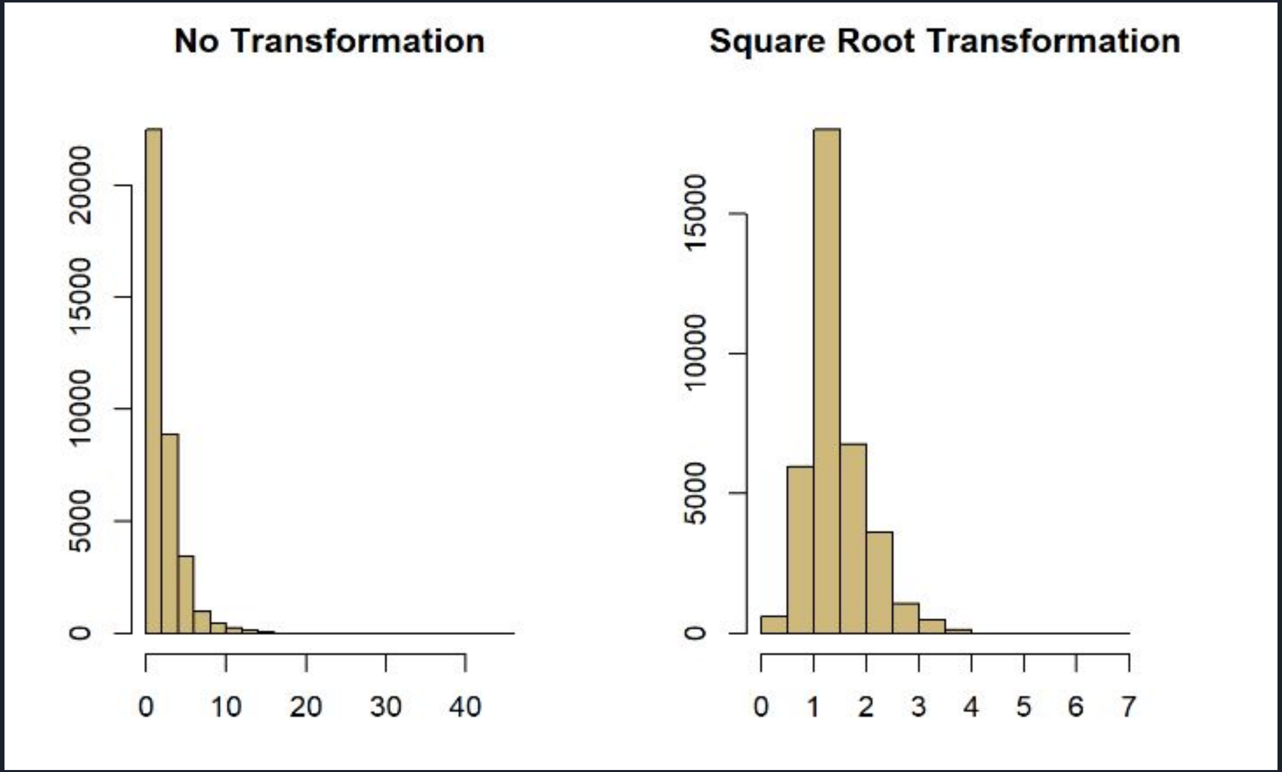
$$y=(\beta0+\beta1x1+...\beta pxp)^2$$

Square root might make this linear.

We need to check how this affects normality.



Comparison of Transformed vs Vanilla Model



As we can see from the plot, Square Root transformation of CO response variable looks normal.

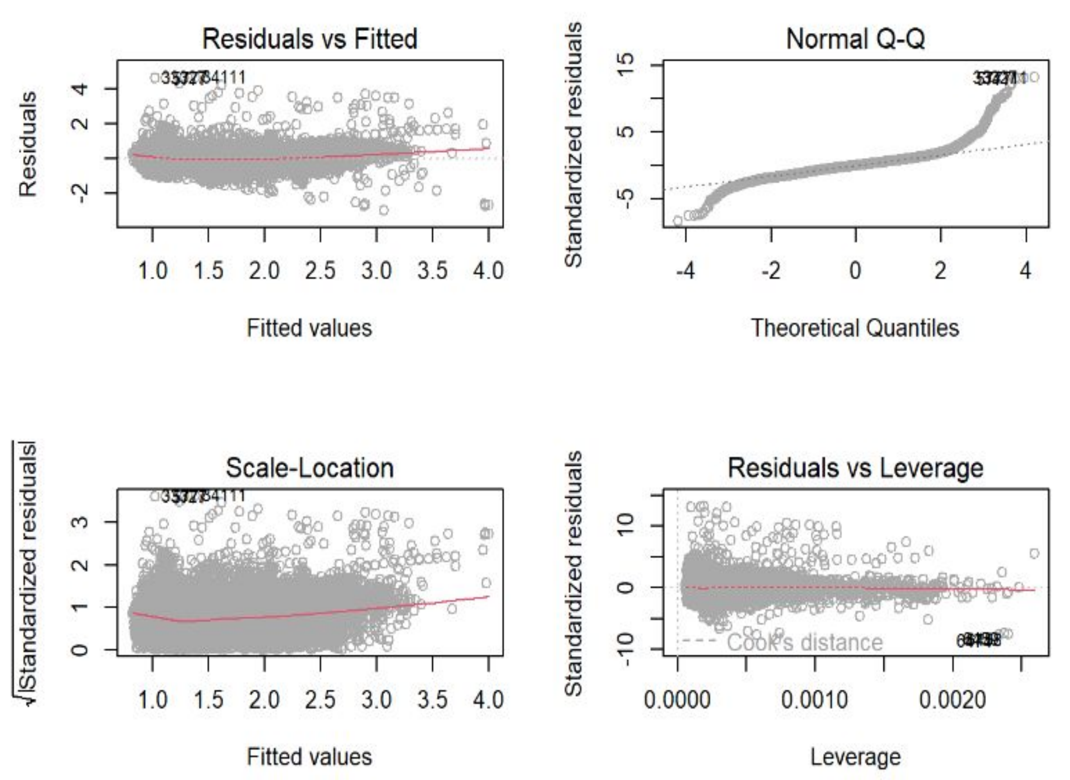
Square Root Transformed Model

The Square Root Transformed model has all predictors significant. The model gave an Adjusted R-squared of 0.6215 which is better than the Adjusted R-square of Linear model.

```
lm_transform = lm(sqrt(CO) ~ . - year - NOX - CDP, data = df)
summary(lm_transform)
```

```
##
## Call:
## lm(formula = sqrt(CO) ~ . - year - NOX - CDP, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9499 -0.1922 -0.0025  0.1876  4.6593
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.4015714  0.4944120  63.513  < 2e-16 ***
## AT          -0.0048910  0.0005339  -9.161  < 2e-16 ***
## AP           0.0024181  0.0003274   7.385 1.56e-13 ***
## AH          -0.0021218  0.0001584 -13.397  < 2e-16 ***
## AFDP        -0.0332135  0.0037446  -8.870  < 2e-16 ***
## GTEP         0.0510619  0.0023128  22.078  < 2e-16 ***
## TIT         -0.0221600  0.0006478 -34.208  < 2e-16 ***
## TAT         -0.0127783  0.0007252 -17.621  < 2e-16 ***
## TEY         -0.0179855  0.0011512 -15.623  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3526 on 36724 degrees of freedom
## Multiple R-squared:  0.6215, Adjusted R-squared:  0.6215
## F-statistic:  7539 on 8 and 36724 DF,  p-value: < 2.2e-16
```

Diagnostic Plot for Squared Root Transformed Model



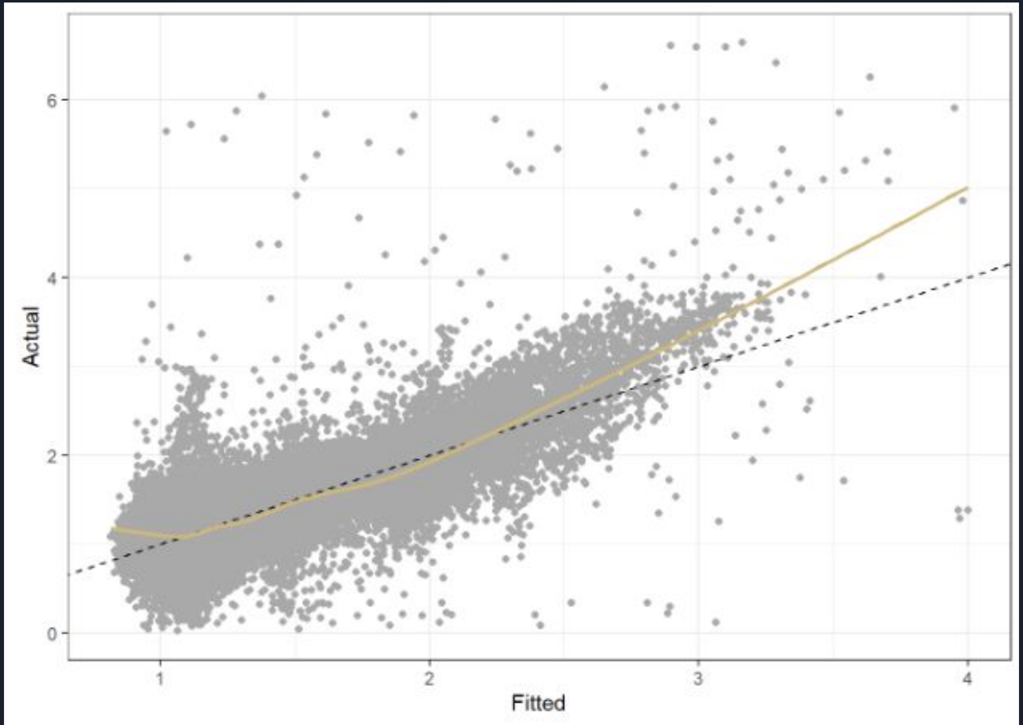
As we can see in these graphs that transformed model helped with normality upto an extent but homoscedasticity is still violated.

Actual vs Fitted Diagnostic for Transformed Plot

Linearity is improved compared to Linear Model. Normality is still an issue.

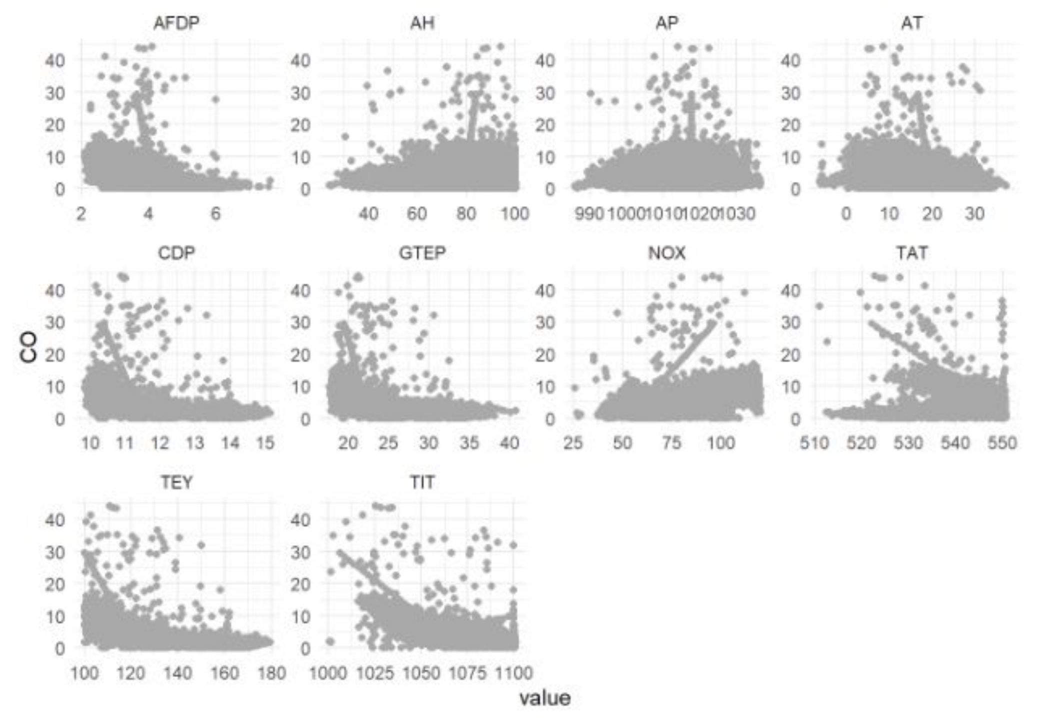
Homoscedasticity looks violated upto an extent.

The only measure that can be used to compare the pre transformed and post transformed model is Rsquare and Adjusted Rsquare.



Weighted Least Squares

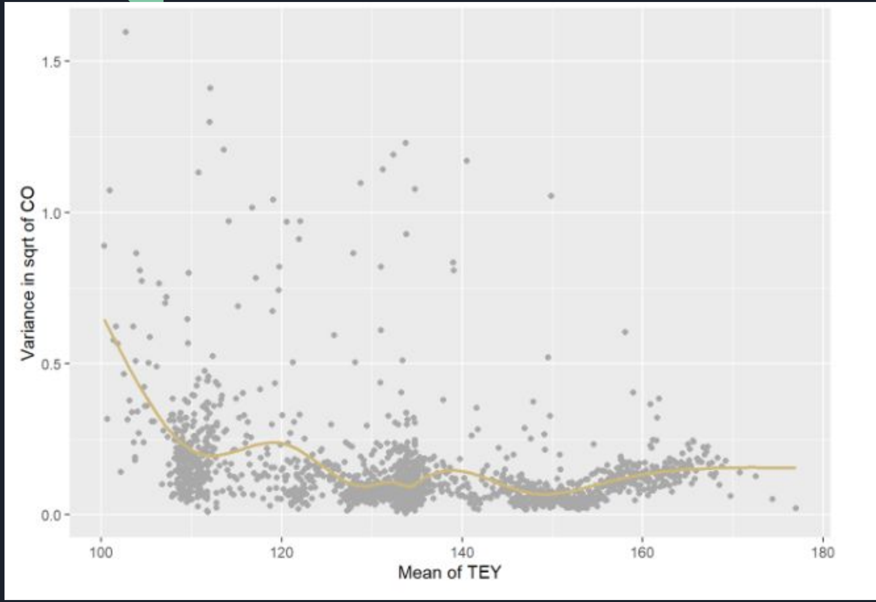
We can try WLS as the data looks slightly heteroscedastic.

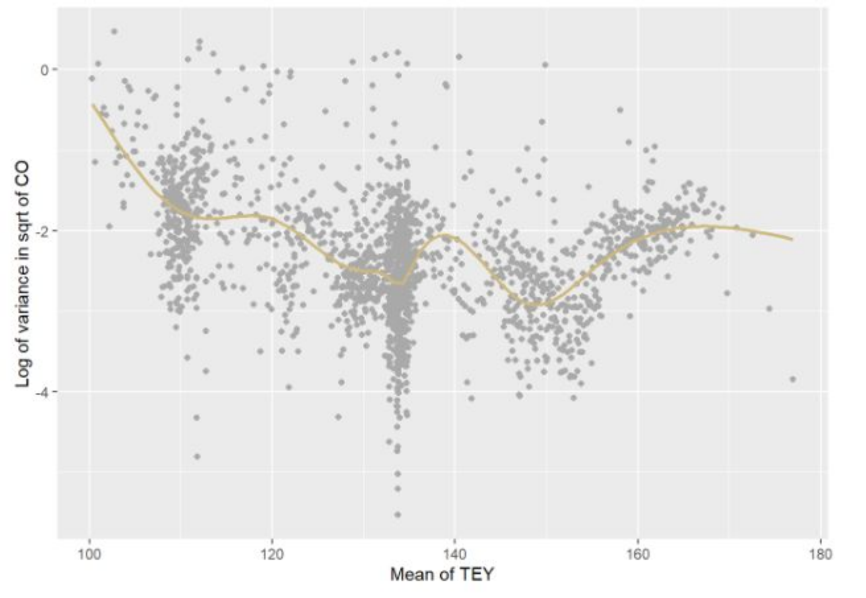


Looking at the response CO against all the predictor to decide the feature that can give an estimate on variance in CO.

The ‘TEY’ feature is chosen for further analysis.

Variance prediction for WLS





Log transformation for variance

Log Transformed Model for estimating response variance

The log transformed model for variance of sqrt(CO) against meanTEY is built to estimate the variance has all predictors significant.

The model gave an Adjusted R-squared of 0.08186 which is pretty low.

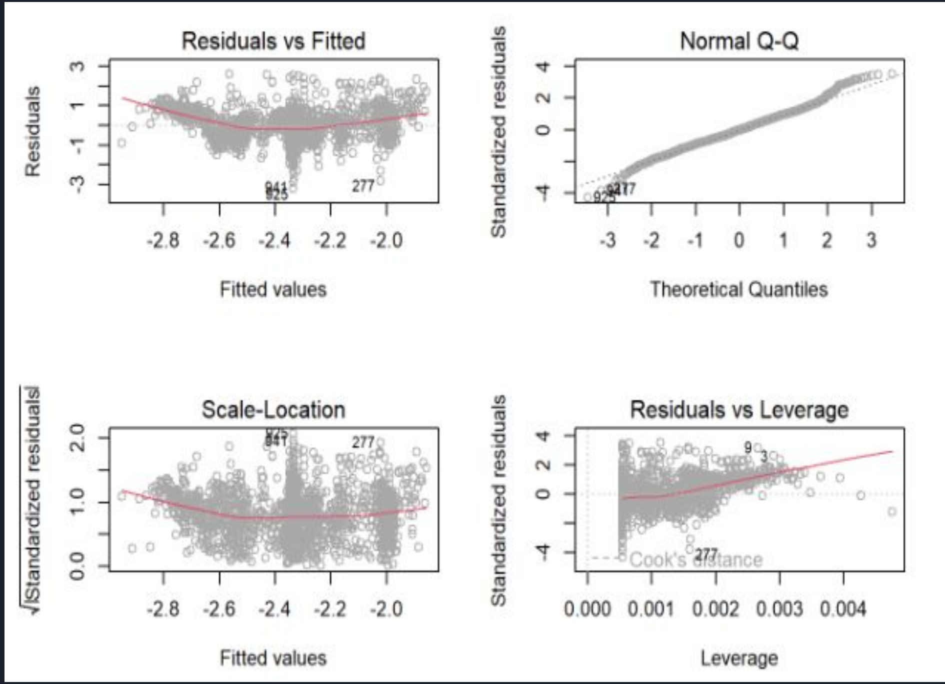
```
lm_var = lm(log(varCO) ~ meanTEY)
summary(lm_var)
```

```
##
## Call:
## lm(formula = log(varCO) ~ meanTEY)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1966 -0.4869 -0.0123  0.4789  2.6174
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.426852   0.149374  -2.858  0.00432 **
## meanTEY     -0.014260   0.001111 -12.833  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7443 on 1835 degrees of freedom
## Multiple R-squared:  0.08236,	Adjusted R-squared:  0.08186
## F-statistic: 164.7 on 1 and 1835 DF,  p-value: < 2.2e-16
```

Diagnostic Plot for Log Transformed var(sqrt(CO)) vs mean(TEY)

This is the plot of linear model of Variance of CO as response and mean of TEY as predictor.

It's a fairly decent model to use to estimate variance as it doesn't as severely violate the assumptions.



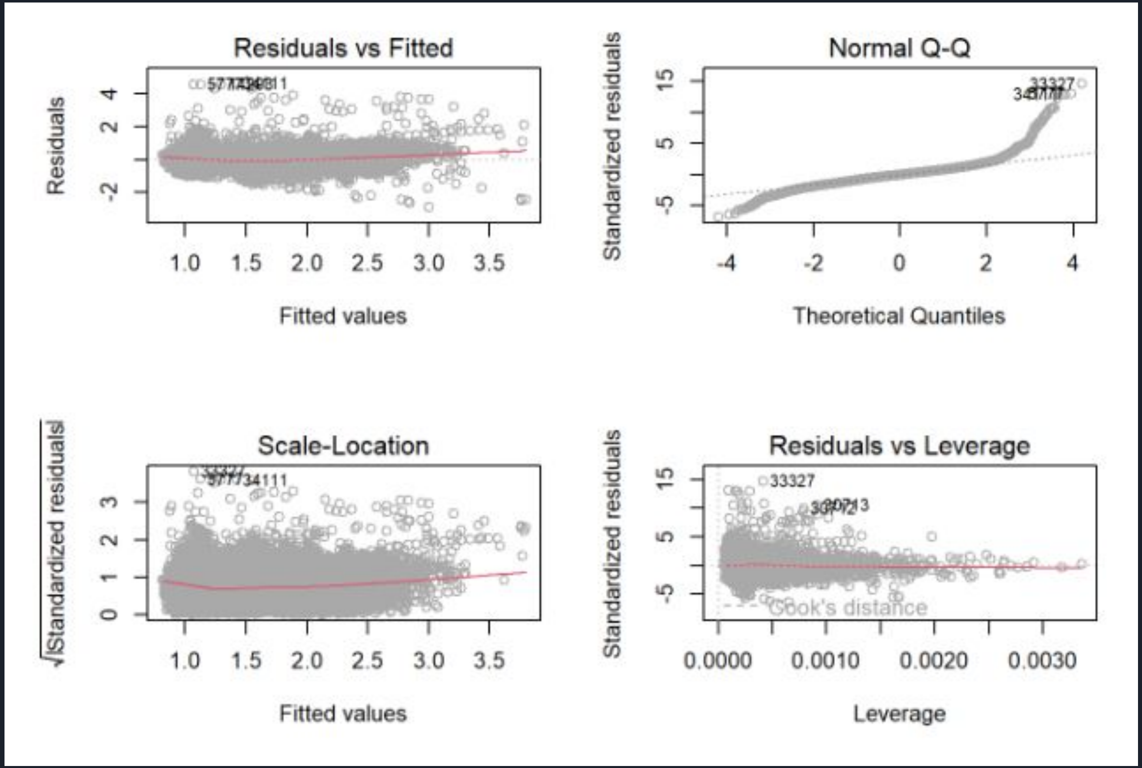
Square Root Transformed WLS Model

The Square root transformed WLS model has all predictors significant. The model gave an Adjusted R-squared of 0.5734 which means it is very close to the transformed linear regression model in explaining variance.

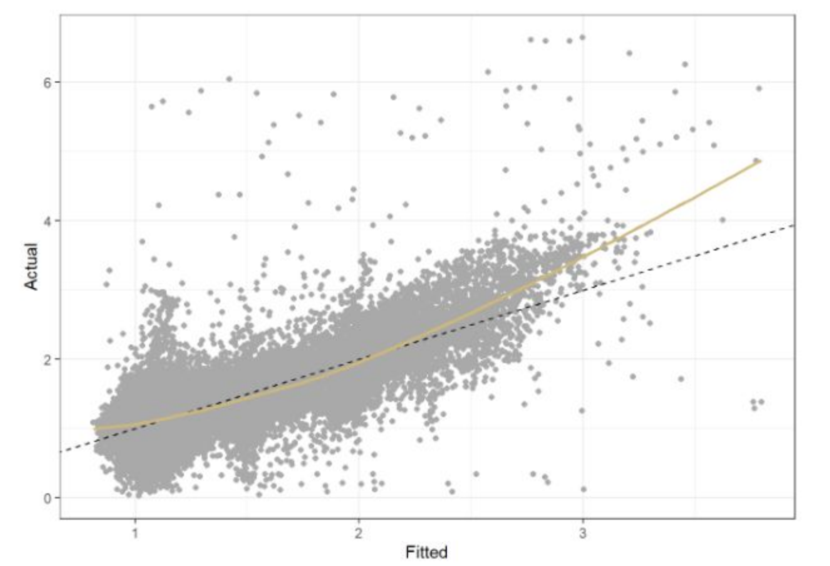
```
lmodwls <- lm(sqrt(CO) ~ . - year - NOX - CDP, data = df, weights = weights)
summary(lmodwls)
```

```
##
## Call:
## lm(formula = sqrt(CO) ~ . - year - NOX - CDP, data = df, weights = weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6740 -0.6106 -0.0128  0.5943 16.4770
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.4140478  0.4915630  59.838  < 2e-16 ***
## AT          -0.0026282  0.0005277  -4.981 6.37e-07 ***
## AP           0.0029679  0.0003218   9.223  < 2e-16 ***
## AH          -0.0024291  0.0001554 -15.635  < 2e-16 ***
## AFDP        -0.0409271  0.0035718 -11.458  < 2e-16 ***
## GTEP         0.0553637  0.0022628  24.467  < 2e-16 ***
## TIT         -0.0259769  0.0006416 -40.491  < 2e-16 ***
## TAT         -0.0041832  0.0007073  -5.914 3.36e-09 ***
## TEY         -0.0122333  0.0011289 -10.837  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.126 on 36724 degrees of freedom
## Multiple R-squared:  0.5735, Adjusted R-squared:  0.5734
## F-statistic:  6172 on 8 and 36724 DF,  p-value: < 2.2e-16
```

Diagnostic Plot for WLS Model



Homoscedasticity is fixed. Normality is still violated to an extent but has a low R^2 value.



As we can see from the plot that fitted values are almost linear to actual values.

Summary of WLS weights

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6.380	9.039	10.317	10.545	11.958	19.817

Conclusion: We see that WLS has fixed heteroscedasticity, but the R^2 value has dropped by a few points.

Multi-collinearity Check

```
vif(lm_transform)
```

```
##        AT        AP        AH      AFDP      GTEP       TIT       TAT       TEY
##  4.671715  1.323365  1.550092  2.481707 27.826201 38.133407  7.275014 95.524902
```

The features have really high VIF.

```
kappa(lm_transform)
```

```
## [1] 137094.9
```

The condition number is also very high. The model can have high variance

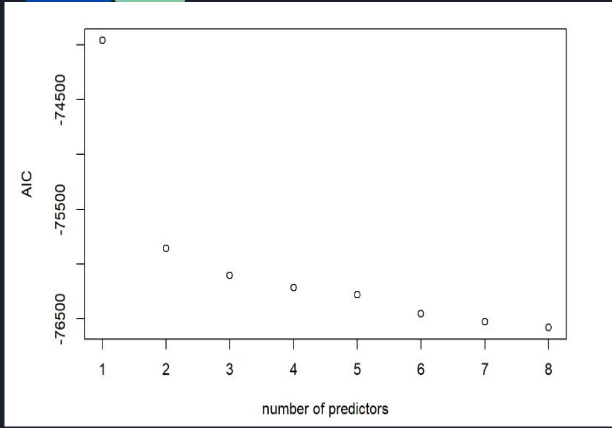
Model Selection

We can lower multi-collinearity by performing model selection on it.

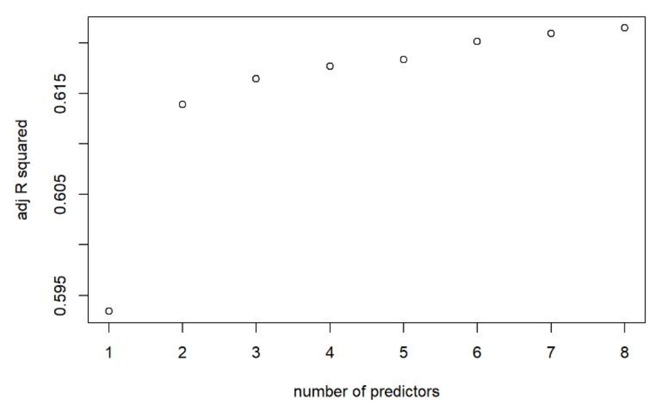
```
n = dim(df)[1];
reg1 = regsubsets(sqrt(CO) ~ . - year - NOX - CDP, data = df)
rs = summary(reg1)
rs$which
```

```
##   (Intercept)    AT    AP    AH  AFDP  GTEP  TIT   TAT   TEY
## 1        TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE
## 2        TRUE FALSE FALSE FALSE FALSE  TRUE TRUE FALSE FALSE
## 3        TRUE FALSE FALSE  TRUE FALSE  TRUE TRUE FALSE FALSE
## 4        TRUE FALSE  TRUE  TRUE FALSE  TRUE TRUE FALSE FALSE
## 5        TRUE FALSE  TRUE  TRUE FALSE  TRUE TRUE  TRUE FALSE
## 6        TRUE  TRUE FALSE  TRUE FALSE  TRUE TRUE  TRUE  TRUE
## 7        TRUE  TRUE FALSE  TRUE  TRUE  TRUE TRUE  TRUE  TRUE
## 8        TRUE  TRUE  TRUE  TRUE  TRUE  TRUE TRUE  TRUE  TRUE
```

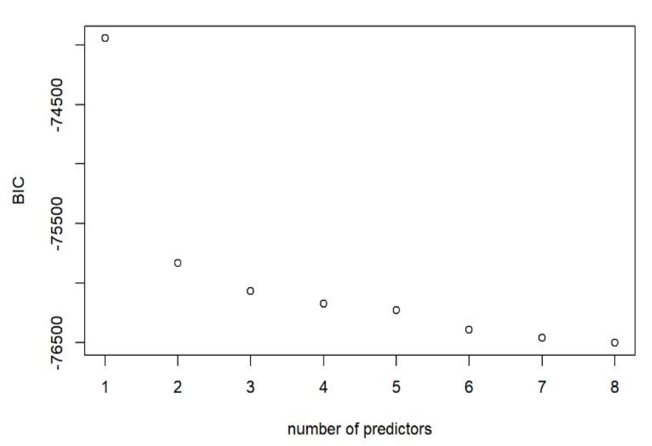
Model Selection



The AIC is lowest for 8 number of predictors



The Adjusted R-squared is highest for 8 number of predictors



The BIC is lowest for 8 number of predictors

Model Selection

Let try a reduced model with 6 predictors as the AIC, BIC and Adjusted R-squared isn't too different from 8 predictors model.

The 6 predictors model as shown gives an adjusted R-squared of 0.6201 which is lower than 8 predictor model but not by much.

However, we were also able to reduce VIF and kappa values.

```
##
## Call:
## lm(formula = sqrt(CO) ~ . - year - NOX - CDP - AP - AFDP, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9405 -0.1939 -0.0043  0.1876  4.6689
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.0584117  0.3608745   94.38   <2e-16 ***
## AT          -0.0069302  0.0005010  -13.83   <2e-16 ***
## AH          -0.0026320  0.0001513  -17.40   <2e-16 ***
## GTEP         0.0501609  0.0023105   21.71   <2e-16 ***
## TIT         -0.0230291  0.0006284  -36.65   <2e-16 ***
## TAT         -0.0115799  0.0006839  -16.93   <2e-16 ***
## TEY         -0.0176361  0.0011495  -15.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3532 on 36726 degrees of freedom
## Multiple R-squared:  0.6202, Adjusted R-squared:  0.6201
## F-statistic:  9994 on 6 and 36726 DF,  p-value: < 2.2e-16
```

```
vif(lm_6)
```

```
##        AT        AH      GTEP       TIT       TAT       TEY
##  4.098392  1.409444 27.674546 35.754113  6.448129 94.912986
```

```
kappa(lm_6)
```

```
## [1] 123964
```

Removal of further features would decrease these metrics. So the next step would be to use regularization to get an low variance model.

Regularization

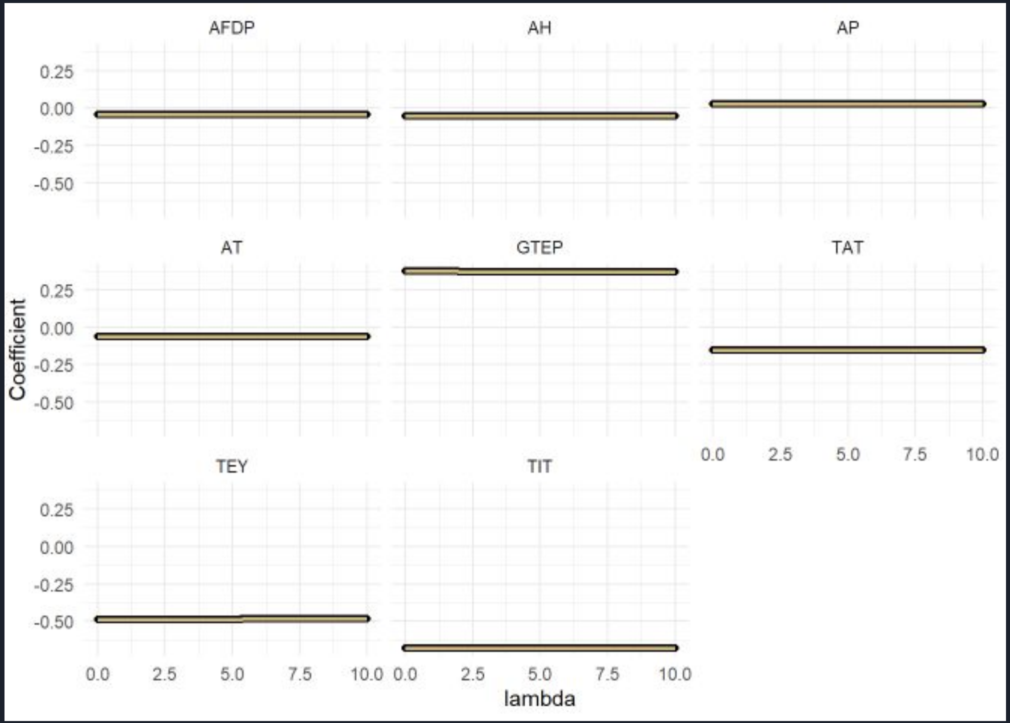
Trying ridge regression to fix multicollinearity..

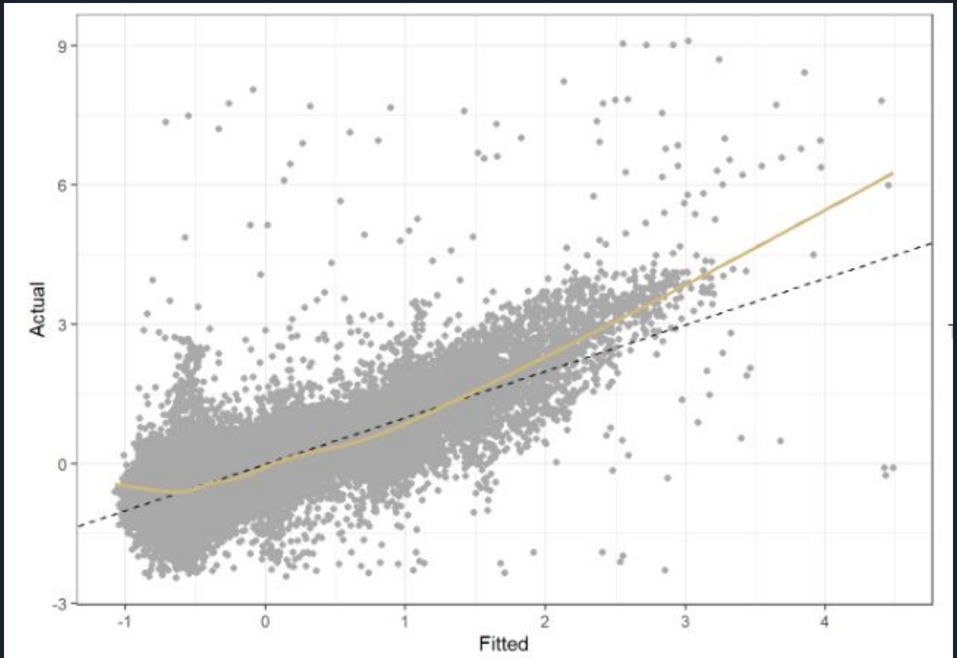
From the plots we can see that there is almost no change in coefficients with increase in lambda.

Using the select function we get best value for lambda as 3.8

```
(mod <- select(lm_ridge))
```

```
## modified HKB estimator is 2.602046
## modified L-W estimator is 3.654323
## smallest value of GCV  at 3.8
```





The regularization did not change the parameters by much.

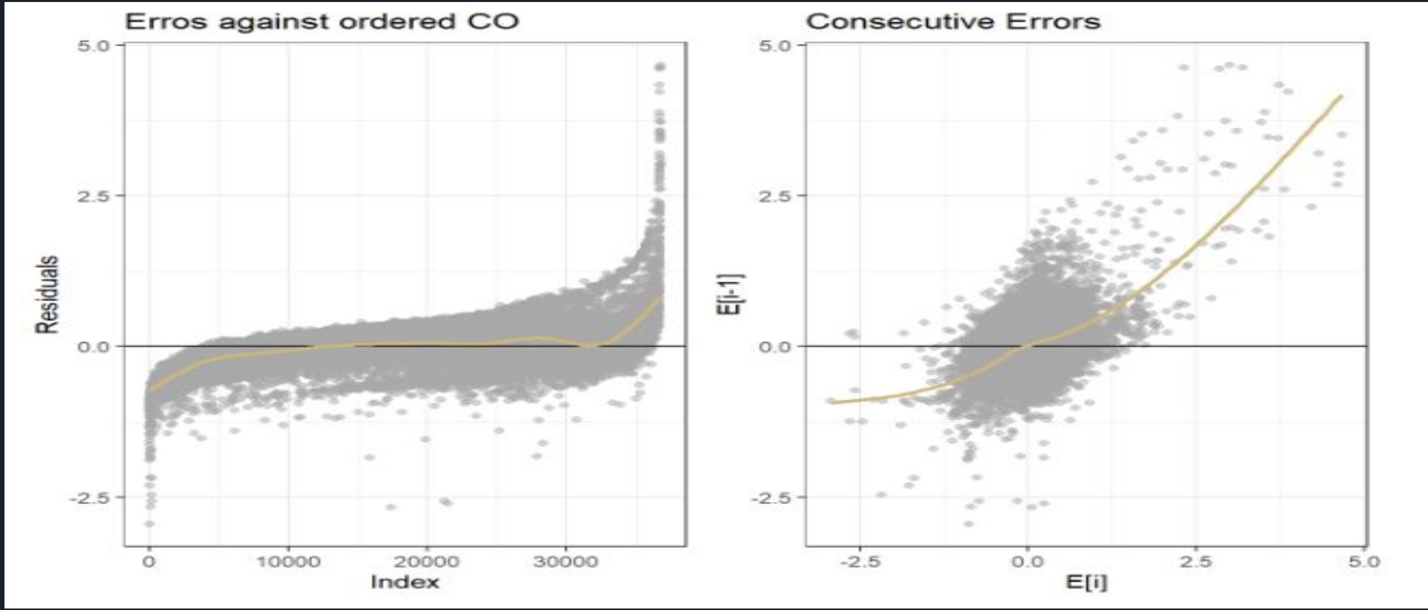
We can compare the MSPE and MAE to compare this model to the previous one.

Metrics Comparison

Model	MAE	MSPE
6 predictor model	0.5441176	1.495244
Ridge Regression	0.5360759	1.473822

			Averaging		Random forest		Meta-ELM	
Task	K	C	MAE	R^2	MAE	R^2	MAE	R^2
CO	512	0.1	1.05	0.43	1.05	0.55	1.34	0.31
	512	1	1.14	0.37	**0.93**	**0.58**	1.26	0.32
NO_x	512	0.01	**7.91**	**0.64**	11.29	0.53	24.05	0.00
	2048	1	10.77	0.16	11.91	0.12	6.64×10^4	0.00

Correlated errors



These graphs shows the correlated error as there is some pattern in Residuals vs index (ordered by CO).

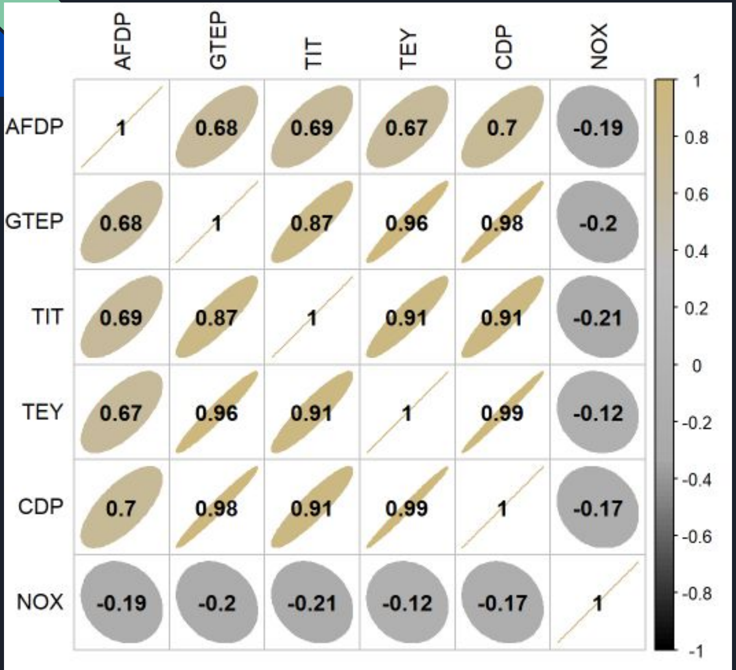
Further Scope

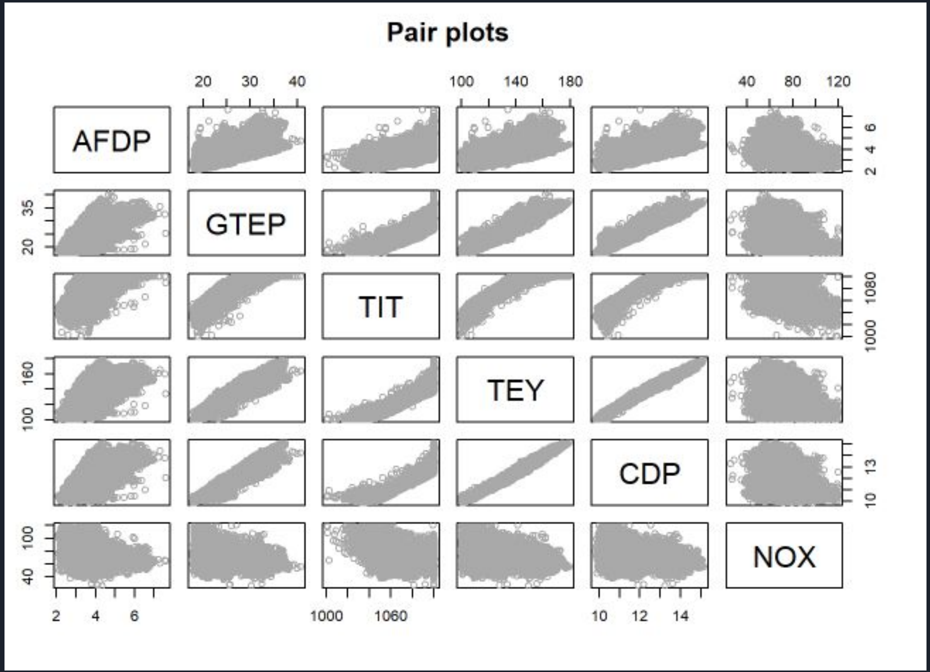
GLS to fix auto correlated errors but there is no temporal element. We can try the matrix way but the data is too large. We can downsample and try it.

For normality violation, we can not do GLMs because the distribution is not any known distribution. The only option left is GAMs

Analysis for NOX

Multivariate Analysis





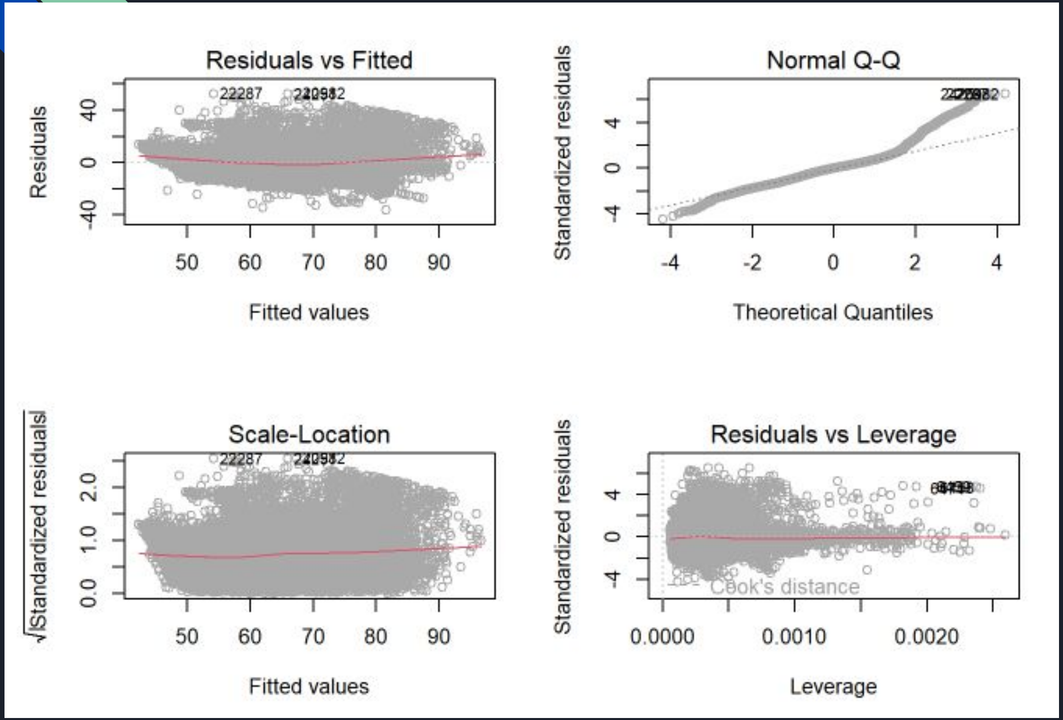
The pair plot doesn't show much pattern for features for response NOX. There are also some high correlations observed between the independent features (Multicollinearity), some measures will be taken later on accordingly. One being CDP is removed as it is highly correlated with TEY.

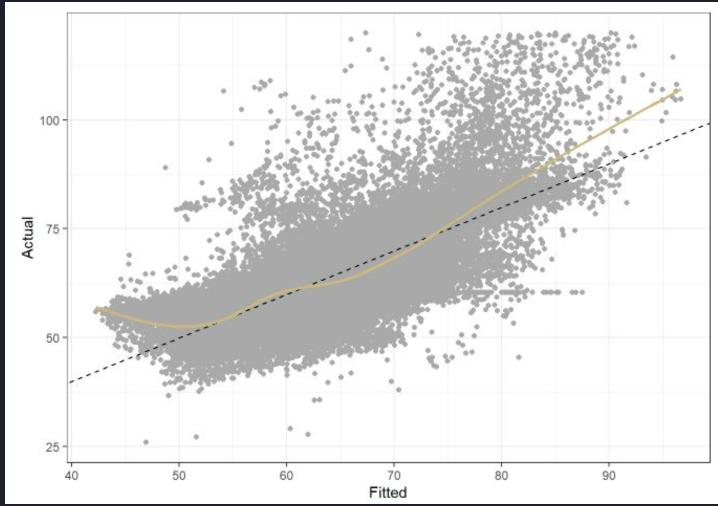
Linear Model

```
##
## Call:
## lm(formula = NOX ~ . - CO - year - CDP, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.224  -4.815  -0.044   3.771  52.534
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -65.720771  11.374007   -5.778 7.61e-09 ***
## AT           -1.793499   0.012283 -146.020  < 2e-16 ***
## AP           -0.243939   0.007533  -32.385  < 2e-16 ***
## AH           -0.223240   0.003644  -61.269  < 2e-16 ***
## AFDP          0.686039   0.086145    7.964 1.72e-15 ***
## GTEP         -0.153111   0.053205   -2.878  0.00401 **
## TIT           1.405350   0.014903   94.300  < 2e-16 ***
## TAT          -1.497407   0.016683  -89.757  < 2e-16 ***
## TEY          -2.048221   0.026483  -77.340  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.111 on 36724 degrees of freedom
## Multiple R-squared:  0.5177, Adjusted R-squared:  0.5176
## F-statistic:  4928 on 8 and 36724 DF,  p-value: < 2.2e-16
```

The linear model has all predictors significant. The model gave an Adjusted R-squared of 0.5176 which means it explains the variance moderately well.

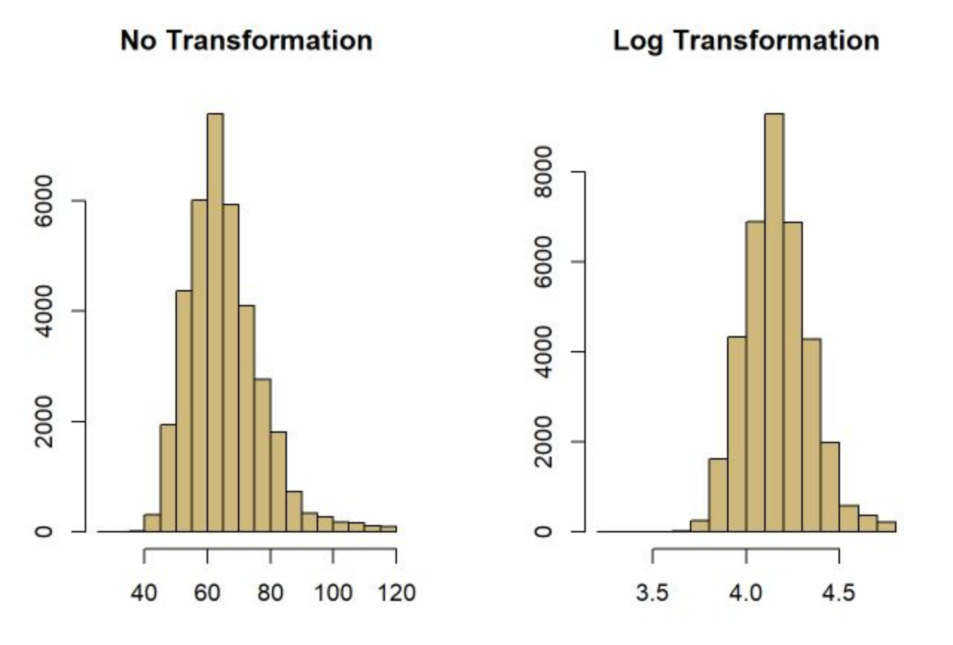
Diagnostic Plot for Linear Model





These plots shows that gaussian markov assumption is violated for normality.

Transformation



After log transformation the distribution of Log Transformed NOX looks normal.

Log Transformed Model

The log transformed model has all predictors significant except GTEP. The model gave an Adjusted R-squared of 0.5349 which is better than the Adjusted R-square of vanilla model.

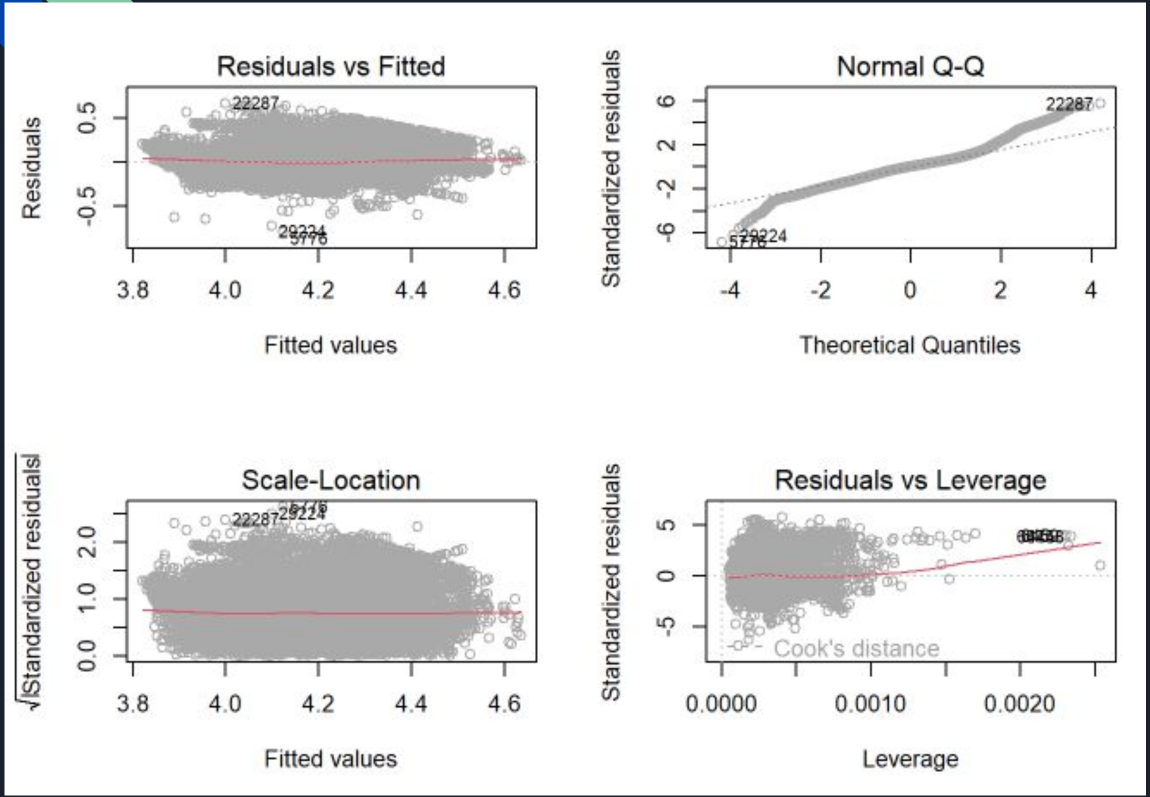
```
##
## Call:
## lm(formula = log(NOX) ~ . - year - CO - CDP, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79999 -0.06803  0.00330  0.06008  0.67087
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1.157e+00  1.638e-01    7.061 1.68e-12 ***
## AT          -2.774e-02  1.769e-04 -156.786  < 2e-16 ***
## AP          -3.932e-03  1.085e-04  -36.241  < 2e-16 ***
## AH          -3.417e-03  5.248e-05  -65.104  < 2e-16 ***
## AFDP         1.034e-02  1.241e-03    8.333  < 2e-16 ***
## GTEP        -9.759e-04  7.664e-04   -1.273    0.203
## TIT          2.221e-02  2.147e-04  103.462  < 2e-16 ***
## TAT         -2.206e-02  2.403e-04  -91.806  < 2e-16 ***
## TEY         -3.174e-02  3.815e-04  -83.192  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1168 on 36724 degrees of freedom
## Multiple R-squared:  0.535,  Adjusted R-squared:  0.5349
## F-statistic:  5282 on 8 and 36724 DF,  p-value: < 2.2e-16
```

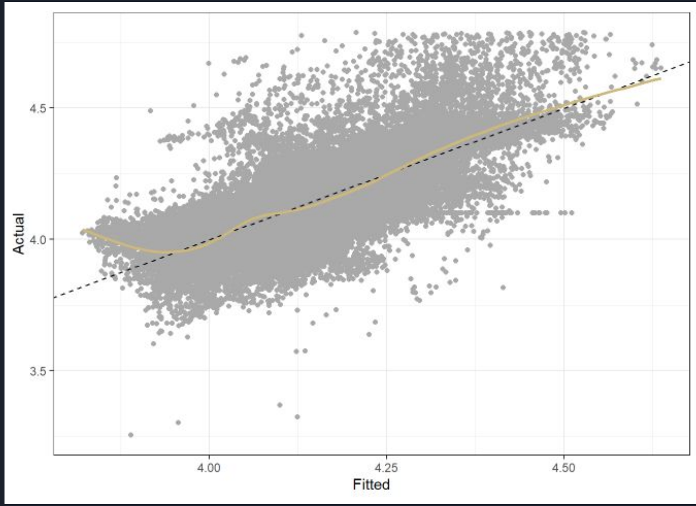
Log Transformed Model

After removing the insignificant feature the model still gave Adjusted R-squared of 0.5349 which is better than the Adjusted R-square of vanilla model.

```
##
## Call:
## lm(formula = log(NOX) ~ . - year - CO - CDP - GTEP, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80008 -0.06793  0.00328  0.06009  0.67016
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1.073e+00  1.501e-01    7.149 8.93e-13 ***
## AT          -2.786e-02  1.513e-04 -184.161  < 2e-16 ***
## AP          -3.922e-03  1.082e-04  -36.243  < 2e-16 ***
## AH          -3.407e-03  5.188e-05  -65.664  < 2e-16 ***
## AFDP         1.036e-02  1.241e-03    8.354  < 2e-16 ***
## TIT          2.229e-02  2.041e-04  109.225  < 2e-16 ***
## TAT         -2.205e-02  2.402e-04  -91.800  < 2e-16 ***
## TEY         -3.208e-02  2.707e-04 -118.512  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1168 on 36725 degrees of freedom
## Multiple R-squared:  0.535,  Adjusted R-squared:  0.5349
## F-statistic:  6036 on 7 and 36725 DF,  p-value: < 2.2e-16
```

Diagnostic Plot for Log Transformed Model





Normality is still violated but not as much.

Multicolinearity Check

```
vif(lm_transform)
```

```
##        AT        AP        AH      AFDP       TIT       TAT       TEY 
##  3.414612  1.316497  1.514730  2.481097 34.476777  7.271245 48.090956
```

The features have really high VIF.

```
kappa(lm_transform)
```

```
## [1] 129631.8
```

The kappa value is also too high. The variance of the model can be high.

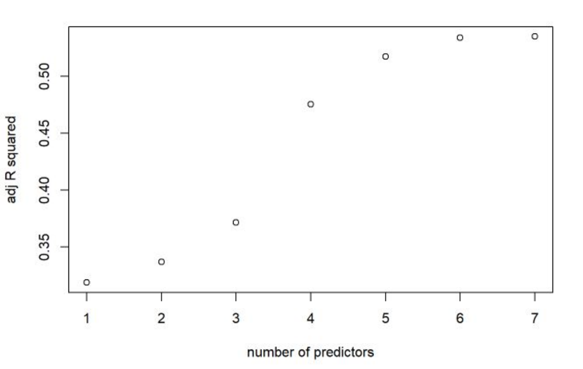
Model Selection

We will improve log transformed model with removed GTEP by performing model selection on it.

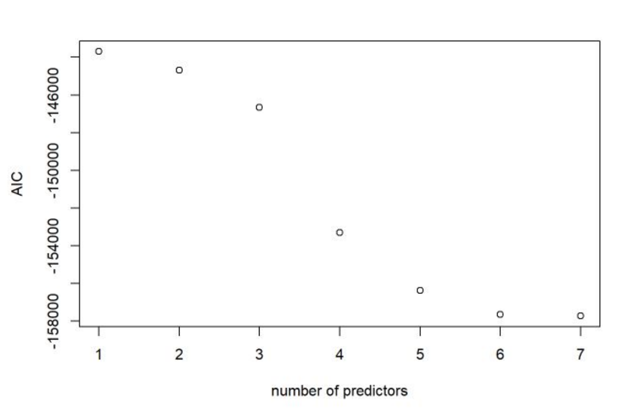
```r
n = dim(df)[1];
reg1 = regsubsets(log(NOX) ~ . - year - CO - CDP - GTEP, data = df)
rs = summary(reg1)
rs$which
```

```
##   (Intercept)   AT    AP    AH  AFDP   TIT   TAT   TEY
## 1        TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 2        TRUE TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
## 3        TRUE TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE
## 4        TRUE TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE
## 5        TRUE TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE
## 6        TRUE TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## 7        TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

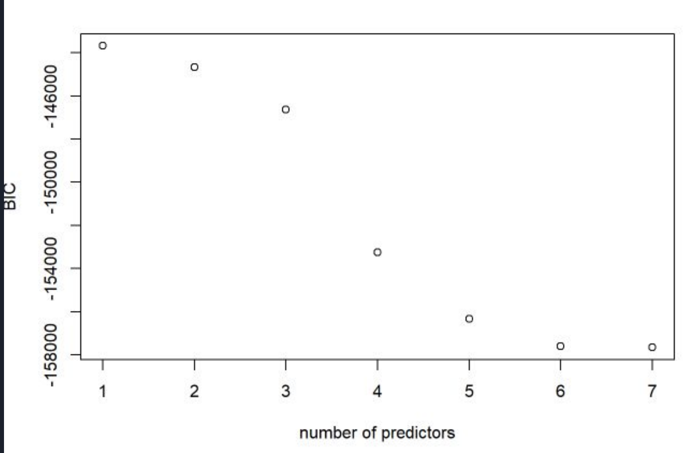
Model Selection



The Adjusted R-squared is highest for 6 or 7 number of predictors



The AIC is lowest for 6 or 7 number of predictors



The BIC is lowest for 6 or 7 number of predictors

Model Selection

Let try a reduced model with 6 predictors as the AIC, BIC and Adjusted R-squared isn't too different from 7 predictors model. Also less complex model is better.

The 6 predictors model as shown gives an adjusted R-squared of 0.534 which is slightly lower than 7 predictor model.

However, we were also able to reduce VIF but kappa value increased.

```
##
## Call:
## lm(formula = log(NOX) ~ . - year - CO - CDP - AFDP - GTEP, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80262 -0.06846  0.00386  0.06032  0.67659
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1.009e+00  1.501e-01    6.721 1.84e-11 ***
## AT          -2.759e-02  1.481e-04 -186.279  < 2e-16 ***
## AP          -3.915e-03  1.083e-04  -36.147  < 2e-16 ***
## AH          -3.342e-03  5.135e-05  -65.086  < 2e-16 ***
## TIT          2.272e-02  1.979e-04  114.764  < 2e-16 ***
## TAT         -2.267e-02  2.290e-04   98.966  < 2e-16 ***
## TEY         -3.233e-02  2.692e-04 -120.094  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1169 on 36726 degrees of freedom
## Multiple R-squared:  0.5341, Adjusted R-squared:  0.534
## F-statistic:  7017 on 6 and 36726 DF,  p-value: < 2.2e-16
```

```
vif(lm_6)
```

```
##        AT        AP        AH       TIT       TAT       TEY
##  3.269032  1.316422  1.481009 32.362356  6.595912 47.486853
```

```
kappa(lm_6)
```

```
## [1] 133957.1
```

Removal of further features would decrease the metrics. So the next step would be to use regularization to get an low variance model.

Regularization

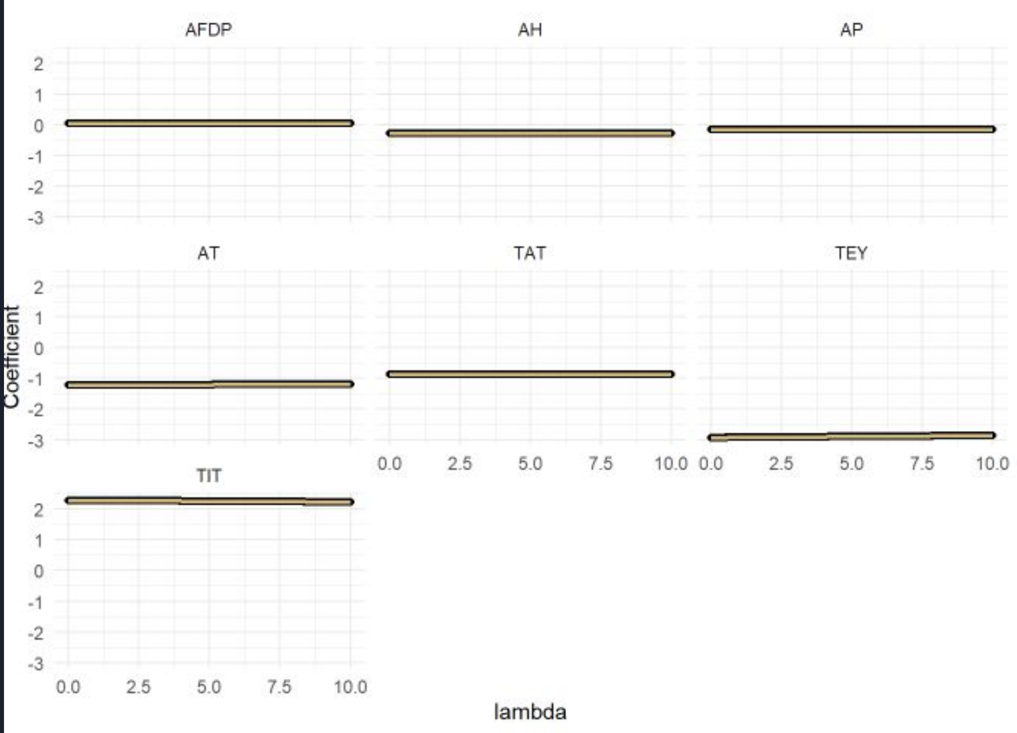
Trying ridge regression to fix multicollinearity..

From the plots we can see that there is almost no change in coefficients with increase in lambda.

Using the select function we get best value for lambda as 0.03.

```
(mod <- select(lm_ridge))
```

```
## modified HKB estimator is 0.1443557
## modified L-W estimator is 4.34683
## smallest value of GCV  at 0.03
```



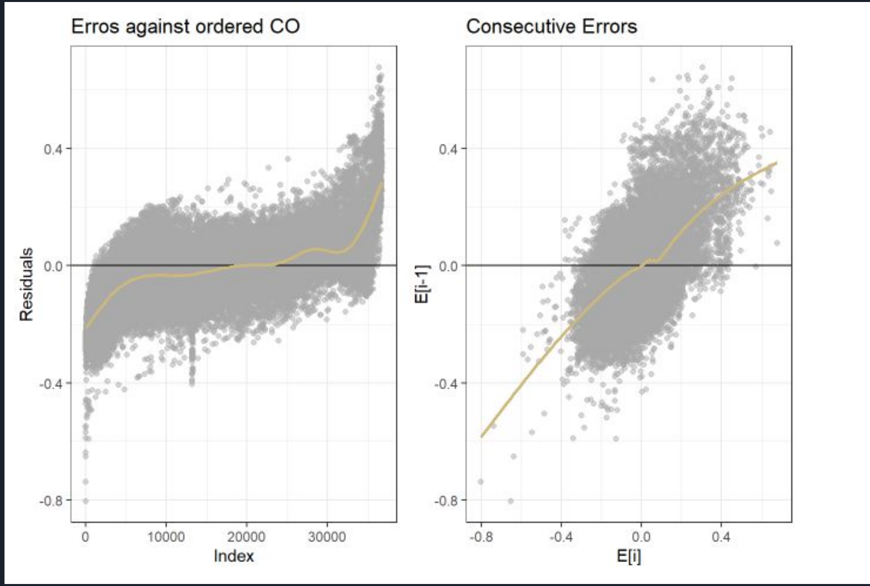
The lambda 0.03 is so low that regularization will not work. The MSPE and MAE could be compared to choose the better fit model.

Metrics Comparison

Model	R^2	Adj R^2	MAE	MSPE
6 predictor model	0.5341	0.534	4.534151	54.34963
Ridge Regression	-	-	5.73	68.55

			Averaging		Random forest		Meta-ELM	
Task	K	C	MAE	R^2	MAE	R^2	MAE	R^2
CO	512	0.1	1.05	0.43	1.05	0.55	1.34	0.31
	512	1	1.14	0.37	**0.93**	**0.58**	1.26	0.32
NO_x	512	0.01	**7.91**	**0.64**	11.29	0.53	24.05	0.00
	2048	1	10.77	0.16	11.91	0.12	6.64×10^4	0.00

Correlated errors



These graphs shows the correlated error as there is a clear pattern in Residuals vs index.

Future Scope: We can try GLS but the dataset doesn't have a temporal component so it isn't as easy. Another solution could be trying GAMs.