Regression Analysis of Greenhouse gas emissions

Anup Bhutada
Master's in Data Science
University of Colorado
Boulder, CO
anbh1796@colorado.edu

Shouvik Sengupta
Master's in Data Science
University of Colorado
Boulder, CO
shse8233@colorado.edu

Uday Gadge
Master's in Data Science
University of Colorado
Boulder, CO
udga1318@colorado.edu

Ayush Pandey
Master's in Data Science
University of Colorado
Boulder, CO
aypa2130@colorado.edu

ABSTRACT

Monitoring the greenhouse gas emissions can be a time and resource consuming task for a lot of the industries. Both periodic measurements and continuous emission monitoring systems can be cost intensive as they would require constant monitoring of the sensors used to measure the gas emissions. Research was published by the Turkish journal of Electrical and Computer Science department to tackle this issue. The research proposes a predictive emission monitoring system (PEM) that predicts the emissions of CO and NOx using calibrated equipment established on site. The researchers set a benchmark for the predictive models built and this project is an attempt to apply statistical methods for the data. [1]

KEYWORDS

Linear regression, predictive modeling, greenhouse gas emissions, periodic measurements, continuous emission monitoring systems, sensors, gas emissions, predictive emission monitoring system, CO, NOx, calibrated equipment, benchmark, statistical methods

1 Introduction

Increase in energy demands has caused irreparable damages to the environment. One of the major concerns is emission of greenhouse gasses mainly Carbon monoxide and nitrogen oxides. Due to the concerns associated with climate change, the Paris convention on climate change was adopted by 196 countries. The main goal was to limit the emission of greenhouse gasses by imposing heavy taxes on factories responsible for large emissions of greenhouse gasses.

These gasses collectively termed as "air pollutants" are produced largely by factories in the combustion process. There has been a lot of research done on predicting the emissions of these gasses using both statistical and machine learning techniques instead of calculating empirical measurements which saves valuable time and resources. We plan on doing an analysis on the same area of research.

We plan on using the data provided by UCI (University of California, Irvine)[2] which has data collected from eleven sensor measures aggregated over an hour from a gas turbine located in Turkey's north western region for the purpose of studying flue gas emissions, namely CO and NOx. The data has attributes of a power plant that are most likely associated with CO and NOX emissions. Our research is focused on doing a multivariate regression analysis on this dataset to predict the air pollutant emissions and extract insights from the process. We plan on applying the statistical tools to build models, analyze, interpret and diagnose them. The data is experimental in nature, so there are chances of error in measurements collected from these sensors.

We are basing our research on an existing research conducted by Department of Computer Engineering at the Scientific and Technological Research Council of Türkiye titled "Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS[1], where they developed random forest models and meta - ELM to predict the same and presented a benchmark. We plan on using the paper for references on theory pertaining to the subject matter and the benchmark as something to strive for in our regression analysis.

2. Gas Turbine CO and NOx Emission Data Set

The dataset provided by UCI machine learning repository[2] consists of hourly average measures of eleven sensors from a gas turbine located in the northern region of Turkey for the purpose of studying the greenhouse gas emissions NO_x and CO. The nine features used to predict the emissions can be put into two categories - ambient variables (temperature, pressure, humidity) and process variables (Turbine energy yield, Gas turbine exhaust pressure). The abbreviations of these variables are described in table [1] .

2.1 Exploratory Data Analysis

In this project, we start off by exploring the distributions of the 9 features and the 2 dependent variables (CO and NO_x emissions). The histograms in Fig.[1] and boxplots in Fig.[2] give a brief overview of the distribution of these features.

Variable(Abbr.)	Unit	Min	Max	Mean
Ambient Temperature (AT)	C	6.23	37.10	17.71
Ambient Pressure (AP)	mbar	985.85	1036.56	1013.07
Ambient humidity (AH)	(%)	24.08	100.20	77.87
Air filter difference pressure (AFDP)	mbar	2.09	7.61	3.93
Gas turbine exhaust pressure (GTEP)	mbar	17.70	40.72	25.56
Turbine inlet temperature (TIT)	C	1000.85	1100.89	1081.43
Turbine after temperature (TAT)	C	511.04	550.61	546.16
Compressor discharge pressure (CDP)	mbar	9.85	15.16	12.06
Turbine energy yield (TEY)	MWH	100.02	179.50	133.51
Carbon monoxide (CO)	mg/m3	0.00	44.10	2.37
Nitrogen oxides (NOx)	mg/m3	25.90	119.91	65.29

Table 1: Variable abbreviation and description

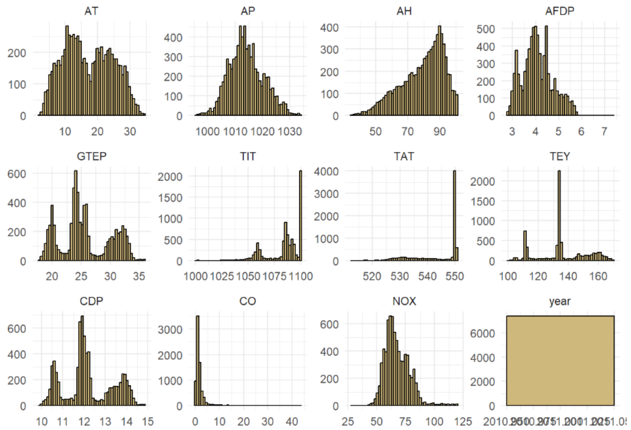


Figure 1: Histogram plot

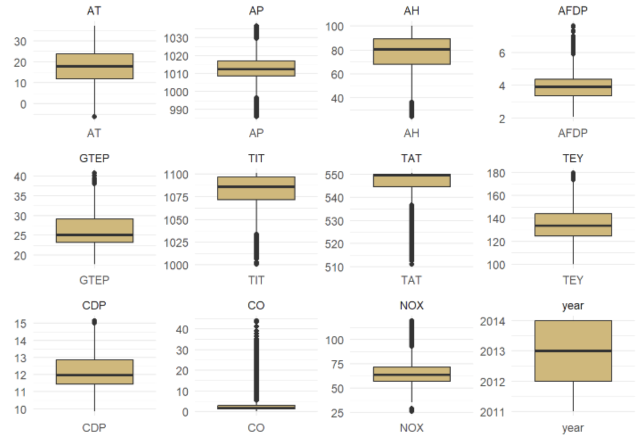


Figure 2: Box plot

The next step in analysis was to understand the correlation between features. The pair plots Fig.[3] and correlation plots Fig.[4] for CO and NO_x reveal some correlations between not only the dependent and independent features but also among the independent features. These plots suggest high correlation among TIT, TEY, CDP, GTEP. Based on this observation, CDP was removed from regression analysis as it has an almost perfect correlation with TEY and GTEP. We find that CO has decent correlation with some of the features - AFD, GTEP, TIT, TEY, CDP suggesting a better prediction possibility while NO_x has lesser correlations.

The correlation between two features is measured as

$$corr(X,Y) = \frac{Cov(X,Y)}{\sigma^x \sigma^y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

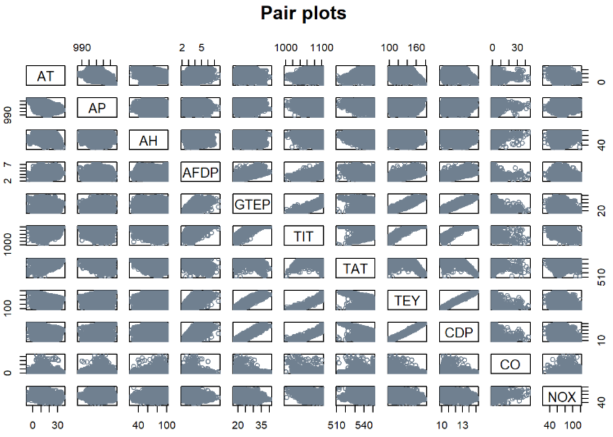
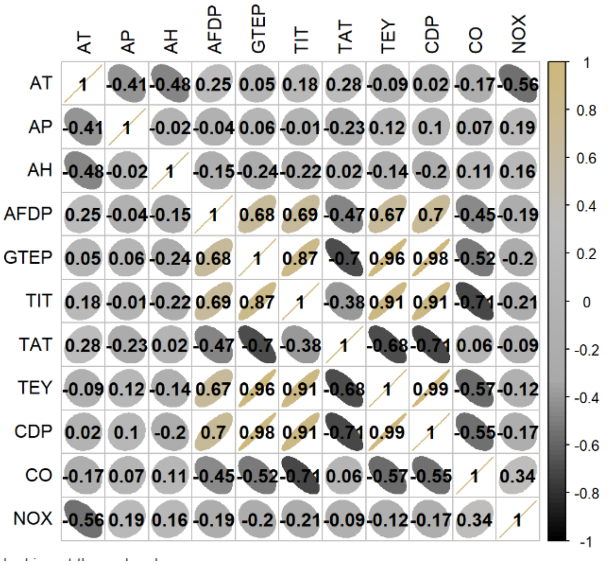


Figure 3: Pair plot



	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
AT	1	-0.41	-0.48	0.25	0.05	0.18	0.28	-0.09	0.02	-0.17	-0.56
AP	-0.41	1	-0.02	-0.04	0.06	-0.01	-0.23	0.12	0.1	0.07	0.19
AH	-0.48	-0.02	1	-0.15	-0.24	-0.22	0.02	-0.14	-0.2	0.11	0.16
AFDP	0.25	-0.04	-0.15	1	0.68	0.69	-0.47	0.67	0.7	-0.45	-0.19
GTEP	0.05	0.06	-0.24	0.68	1	0.87	-0.7	0.96	0.98	-0.52	-0.2
TIT	0.18	-0.01	-0.22	0.69	0.87	1	-0.38	0.91	0.91	-0.71	-0.21
TAT	0.28	-0.23	0.02	-0.47	-0.7	-0.38	1	-0.68	-0.71	0.06	-0.09
TEY	-0.09	0.12	-0.14	0.67	0.96	0.91	-0.68	1	0.99	-0.57	-0.12
CDP	0.02	0.1	-0.2	0.7	0.98	0.91	-0.71	0.99	1	-0.55	-0.17
CO	-0.17	0.07	0.11	-0.45	-0.52	-0.71	0.06	-0.57	-0.55	1	0.34
NOX	-0.56	0.19	0.16	-0.19	-0.2	-0.21	-0.09	-0.12	-0.17	0.34	1

Figure 4: Correlation Matrix

3. Procedure

We proceed with basic linear regression and check the assumptions and use different methods to fix the violated assumptions for both CO and NO_x.

3.1 PEM for CO

3.1.a. Linear Model for CO

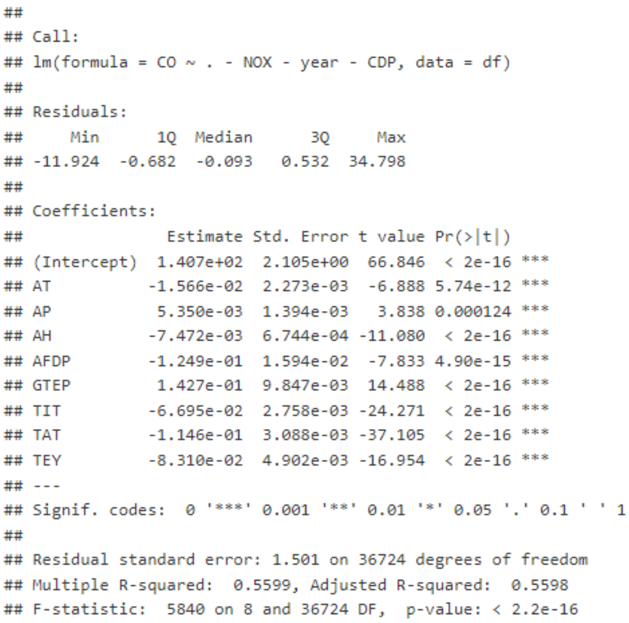
The fitted model for CO using all predictors gave an adjusted R-squared value of 0.51. The p-value for predictors indicated that all features are significant in predicting the emission of CO. The p-values are from a t-test whose null hypothesis is that the parameter β associated with the feature is not significantly different from 0. A p-value less than the critical value, usually 0.05 means that we reject the null hypothesis and conclude that beta is significantly different from 0 and it has an impact on the model. As for the assumptions of linear regression, the model fails certain assumptions. The Fig.[6] indicate a violation of normality and homoscedasticity to a certain extent. However, the biggest violation is linearity as seen in the fitted vs actual plot Fig.[7] which has a parabolic curve instead of a linear curve.

The plot indicates a relationship

$$y \sim (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)^2$$

By taking a square root we obtained a linear relationship

$$\sqrt{y} \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$



```
##
## 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
```

Figure 5: Model summary of Linear model of CO

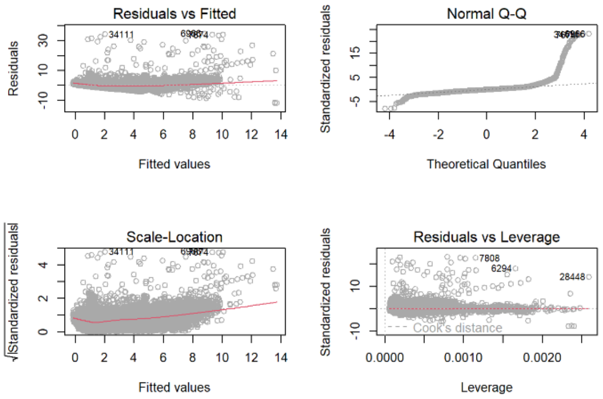


Figure 6: Diagnostic Plot of CO

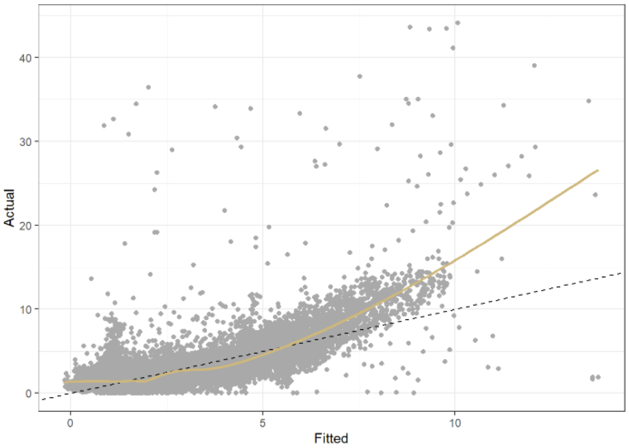
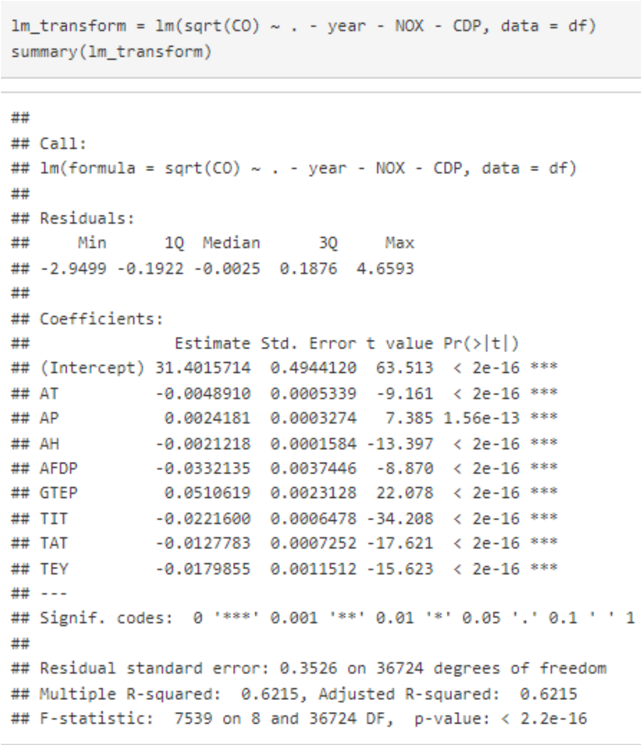


Figure 7: Actual vs Fitted plot of CO

3.1.b. Transformation

The transformed regression model with sqrt(CO) with the features gave a significant improvement in R-squared to a 0.62. The p-values suggest that all features are significant. The diagnostic plots show an improvement in the normality assumption with the homoscedasticity slightly violated. The fitted vs actual plot shows a linear relationship now. To fix homoscedasticity issue we went with weighted least squares(WLS)



```
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
```

Figure 8: Model summary of transformed Linear model of CO

3.1.c. Weighted Least Squares

One of the assumptions of ordinary least squares that we have attempted in the previous two sections is homoscedasticity or constant variance for the residuals. The ordinary least squares method assumes that the error covariance matrix is of the form $\text{cov}(\bar{\epsilon}) = \sigma^2 I(n)$ where all diagonal terms indicate the variance of errors and off diagonals represent the covariance between the error terms. The diagonal terms being σ^2. When this assumption is violated the estimates for the linear model are not unbiased. One way to fix this problem is by using a weighted least squares model that takes into account the different variances of error terms.

To estimate the weights for data points we need to estimate the variances of error terms. We used the pairplot for CO against some of the correlated predictors Fig.[9] to find the predictor that can explain the variance in sqrt(CO) (the response). The plots show a similar relationship between CO and some of the predictors (CDP, TEY, GTEP, AFDP). The plot shows that as the values of one of these variables (TEY say) increases, the spread of CO decreases. Therefore we chose TEY to estimate the variance of sqrt(CO) for the weighted least squares model

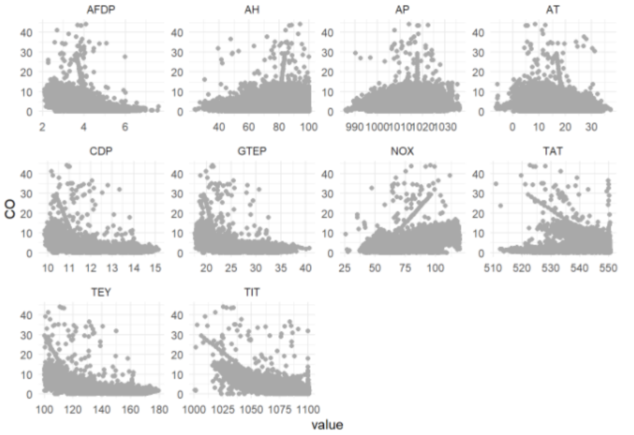
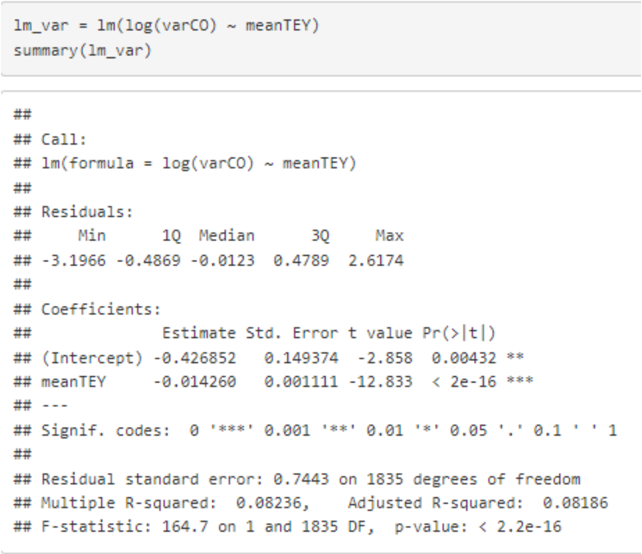


Figure 9: Pairplot for CO against correlated features

We ordered the data based on TEY values and then split the data points into groups of 20. The fig shows the plot of mean of TEY vs log of variance of sqrt(CO) in each of these groups. The log transformation was used since the relation appeared to be exponential in the pair plots. It is seen that for the most part the variance decreases as the mean TEY increases. A linear regression model was fit using mean TEY as the predictor and variance of sqrt(CO) as the response variable. The resulting model and the variable t-tests are shown in the Fig.[10].



```
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
```

Figure 10: Log Transformed Model for estimating response variance

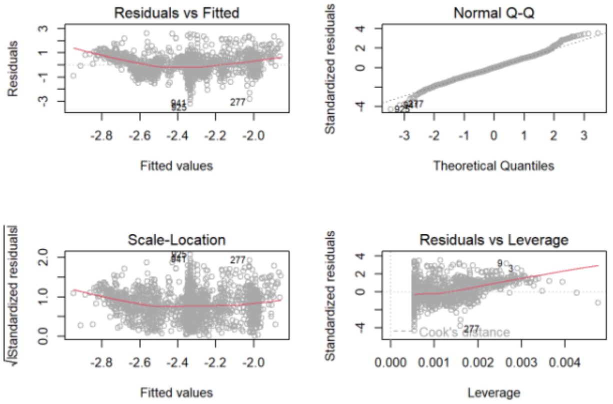
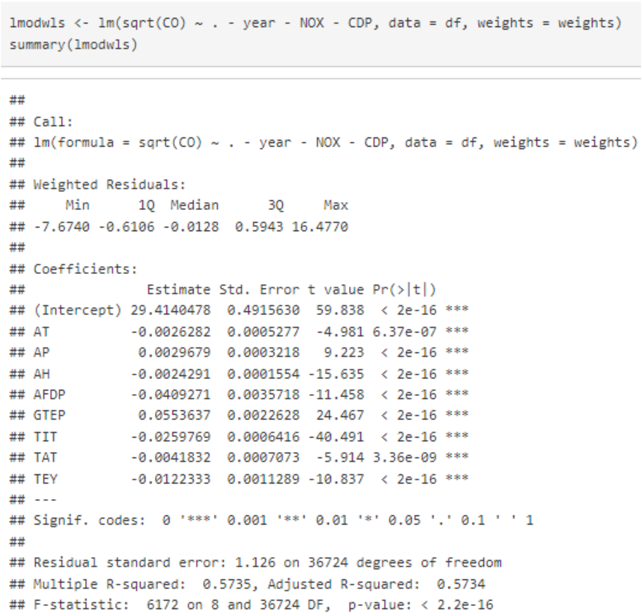


Figure 11: Diagnostic plots of model to estimate response variance

The diagnostic plots for the model are shown in Fig.[11]. The results show an adjusted coefficient of determination of 0.081 with high significance for the predictor. The diagnostics plots show that assumptions of linearity, normality, homoscedasticity and independence are satisfied. Using this model, the estimates for variances for all values of the response were computed and the reciprocal of these values was used to estimate the weights. The weighted least squares model fit using these weights is shown in fig. The adjusted coefficient of determination for this model is 0.57 and all the variables are highly significant. The diagnostic plots for this model Fig.[13] show that the assumptions of linearity and homoscedasticity look satisfactory. The distribution of residuals is closer to normal as compared to the linear model.



```
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
```

Figure 12: Model Summary of WLS model

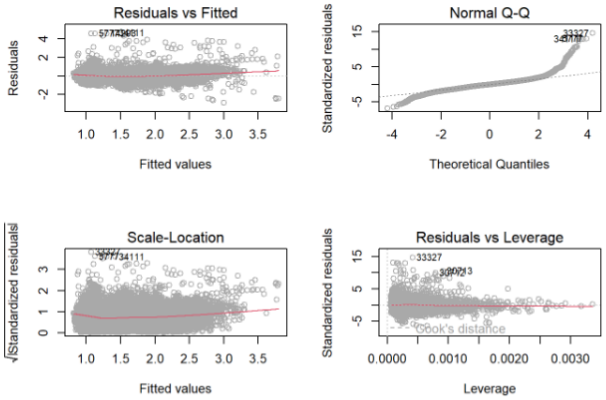


Figure 13: Diagnostic plot for WLS model of CO

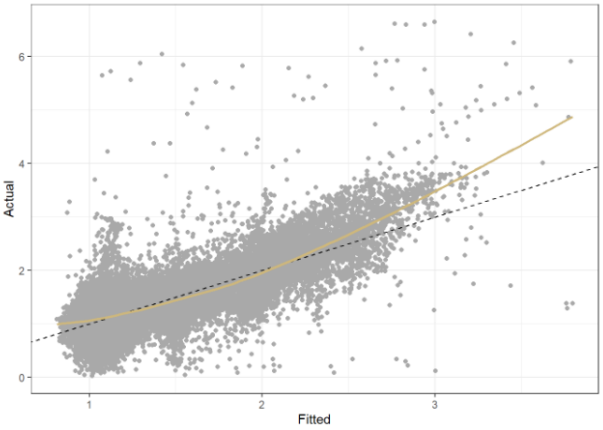
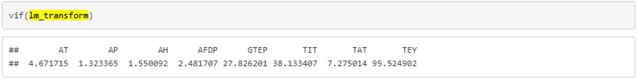


Figure 14: Actual vs Fitted Plot for WLS model of CO

3.1.d. Multicollinearity and model Selection

One other assumption of linear regression is that the predictors are independent of each other and are not linearly related to each other. One way to check this assumption is calculating the Variance Inflation Factors (VIF) of the predictors. It is given by, $VIF = \frac{1}{1-{R_j}^2}$, where ${R_j}^2$ is the R-squared value of the model used to predict the feature using other predictors.

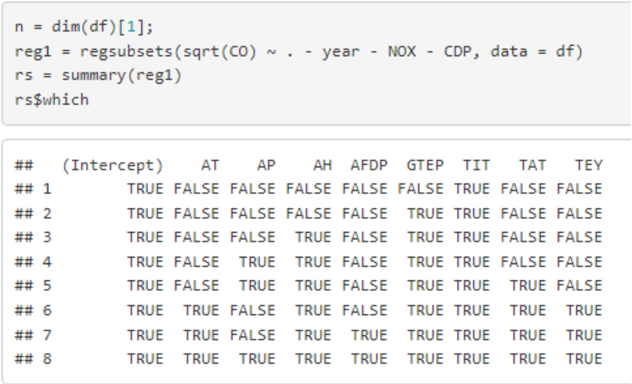
A high value of VIF would indicate a high correlation of the feature with the other predictors. The square root transformed model had high VIF values for the features used. To tackle this problem, we reduce the dimensionality of the data by picking features from the existing data.



```
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
```

We started off by getting the subset of features that give minimum RSS for the size. *Regsubsets* computes the RSS value of all possible combinations of i features for i=1,2,...p and picks the model with the least RSS for each i. This is valid because we are comparing models on the same data and same number of features. The resulting summary Fig.[15] indicates the best options for each number of features



```
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
```

Figure 15: model selection using regsubsets

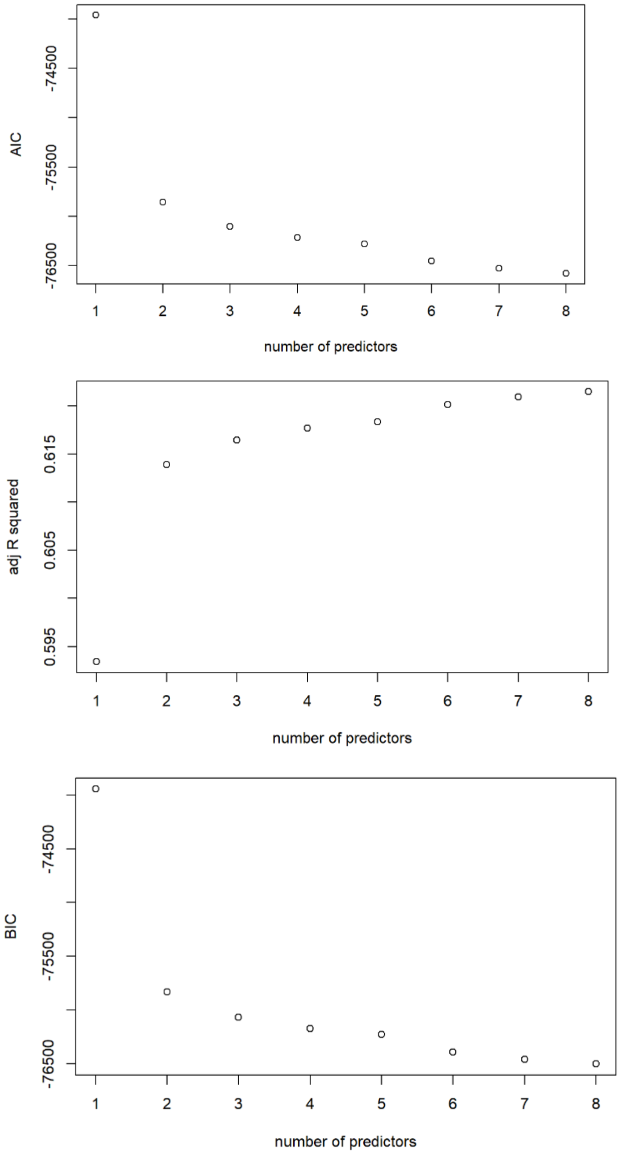
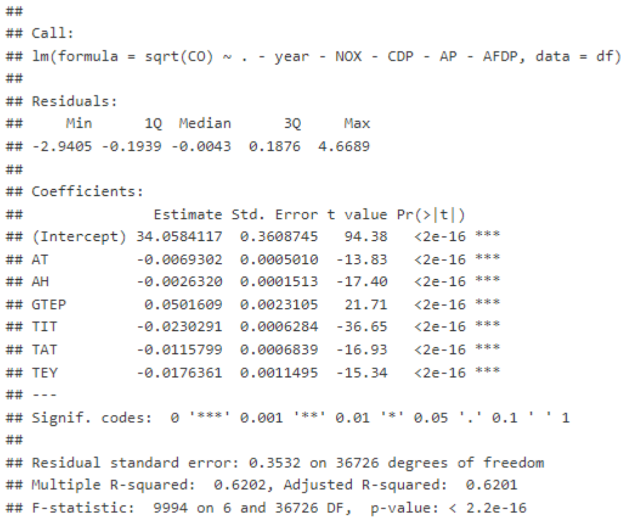


Figure 16: AIC, Adj. R-Squared, BIC vs no. of predictors plot

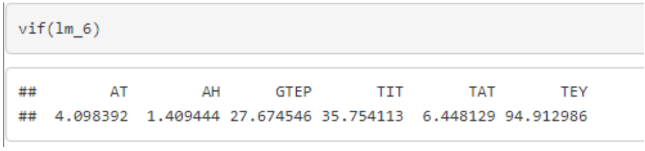
We calculated the values of AIC, BIC and adjusted R^2 for all the best models of different sizes to pick the model that has lower dimensionality but explains most of the variance in the data. These metrics compute the closeness of the model to the true model. The three plots Fig.[16] indicate that a model with 6 features obtained from the summary Fig.[17]



```
##
## 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
```

Figure 17: Model summary of reduced 6 feature model

has more or less the same metrics as the model with 8 features. This model with the 6 features had an adjusted R^2 of 0.6201. The recalculated VIF values are lower but are still significant. The diagnostic plots did not vary much.



```
vif(lm_6)
```

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

3.1.e. Ridge Regression

High VIF valued model could lead to high variance in the parameters estimated. To fix this we used Ridge Regression after normalizing the data which penalizes the parameters by adding bias to minimize variance. The cost function is changed from RSS to RSS + $\lambda\Sigma\beta_i^2$.

However in this case, increasing λ did not affect the parameters Fig.[18]

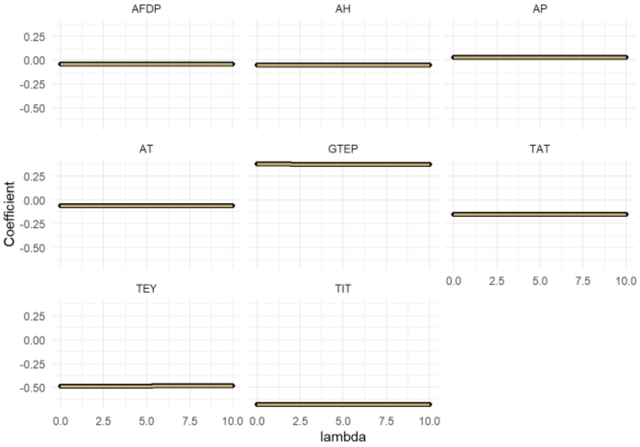


Figure 18: Coefficient vs lambda plot for ridge regression of CO

This could be because the parameters were low enough to begin with. However, the best lambda in this case turned out to be 3.8 and we used this to finalize the model.

We later tested this model against the 6 features transformed model with a train test split and checking the MSPE on the test set.

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

Table 2: Metrics Comparison of different models of CO

3.1.f. Autocorrelated Errors

The diagnostic plots for the linear model with all variables showed a slight pattern in the residual vs fitted values plot. This calls for a check on the correlation between the error terms. The data points were sorted with respect to the response variable to plot the successive residuals Fig.[19].

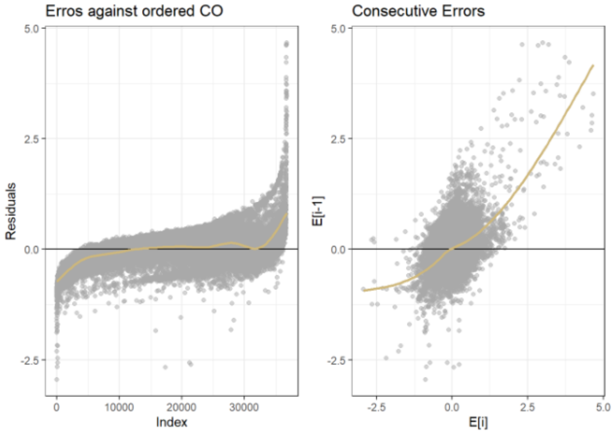


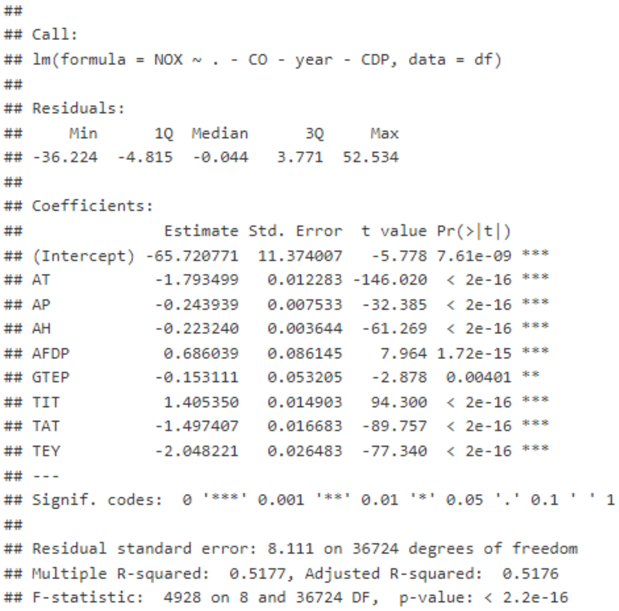
Figure 19: Correlated errors plot

The resulting plot showed evidence of autocorrelation in the error terms. In a future study, advanced models such as Generalized least squared and Generalized Additive models can be used to fix this issue.

3.2 PEM for NO_x

3.2.a. Linear Model for NO_x

The fitted model for NO_x using all predictors indicated that GTEP was insignificant. The model without GTEP gave an adjusted R-squared value of 0.51. The p-value for predictors indicated that all features are significant in predicting the emission of NO_x. As for the assumptions of linear regression, the model fails certain assumptions. The Fig.[21] indicate a violation of normality.



```
##
## 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
```

Figure 20: Model summary of Linear Model of NOx

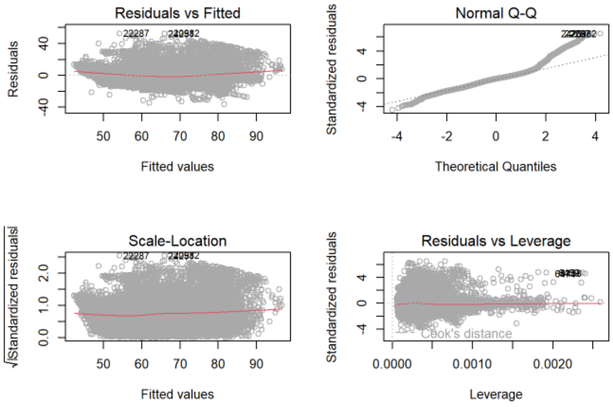


Figure 21: Diagnostic Plot for Linear Model of NOx

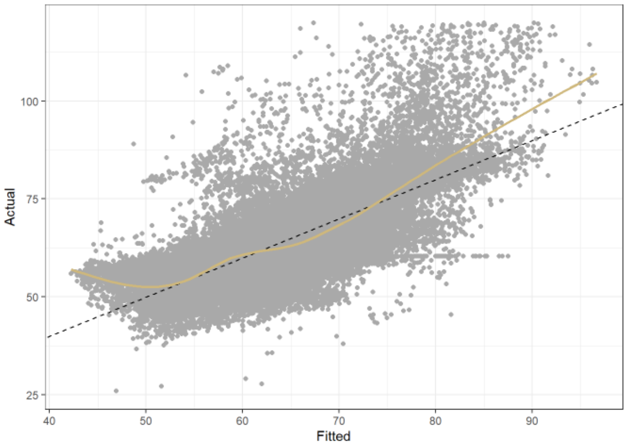


Figure 22: Actual vs Fitted plot of NOx

The fitted vs actual plot Fig.[22] shows a violation of linearity to an extent. The plot suggests an exponential relationship.

The plot indicates a relationship

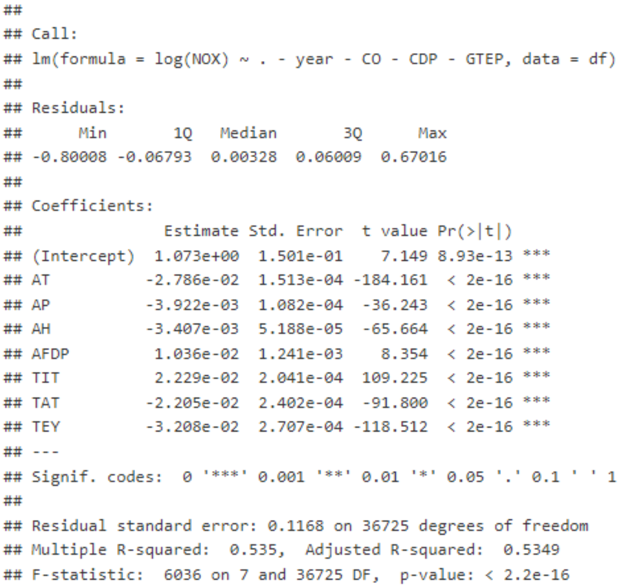
$$y{\sim}e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}$$

By taking a log transformation we obtained a linear relationship

$$log(y){\sim}\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p$$

3.2.b. Transformation

The transformed regression model with log(NO$_x$) with the features gave a slight improvement in R-squared to a 0.53. The p-values suggest that all features (except GTEP) are significant. The diagnostic plots show an improvement in the normality assumption and linearity is valid with the transformation as evident in the diagnostic plots Fig.[24] and fitted vs actual plot Fig.[25]



```
##
## 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
```

Figure 23: Model summary of transformed Linear Model of NOx

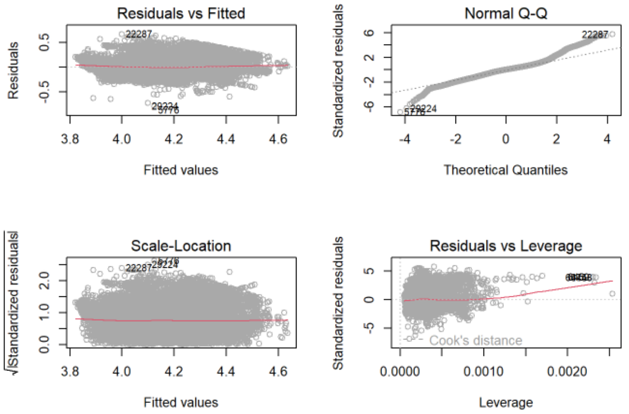


Figure 24: Diagnostic Plot for Linear Model of NOx

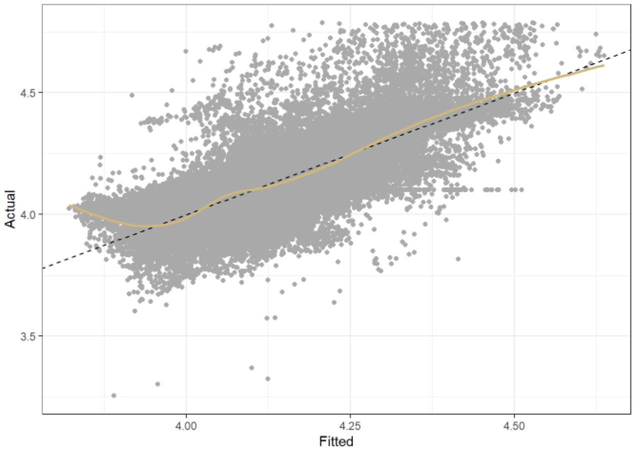
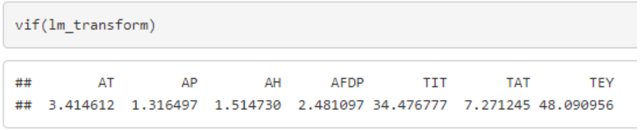


Figure 25: Actual vs Fitted plot of NOx

3.2.c. Multicollinearity and model selection

Similar to the analysis of CO, the variance inflation factor values for all predictors

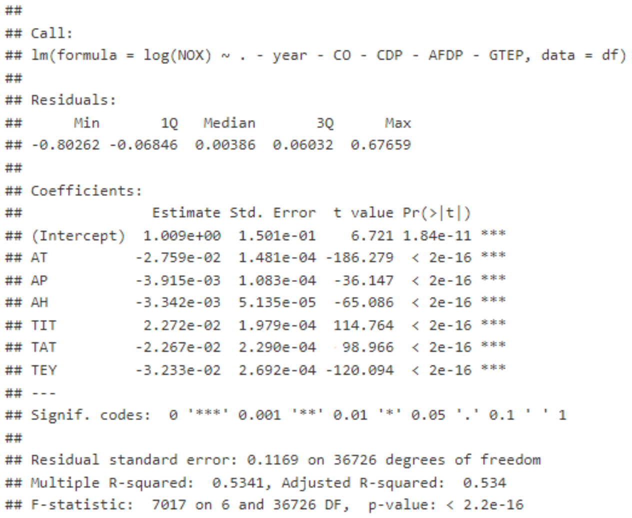


```
vif(lm_transform)
```

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

for the log transformed model showed a high correlation between predictors for TIT and TEY. To account for this multicollinearity, model selection was performed using *Regsubsets* as explained in section **3.1.d**. The resulting plots comparing the models with different number of predictors using AIC, BIC and adjusted coefficient of determination are shown in fig. All these metrics are seen to have very similar values for 6 and 7 predictors. Since the existing multicollinearity required that a correlated predictor be removed, the model with 6 predictors suggested that AFDP can be removed. This was contrary to expectations as the aim was to remove one of the correlated predictors. But the results of model selection showed a significant drop for removing any other predictor.

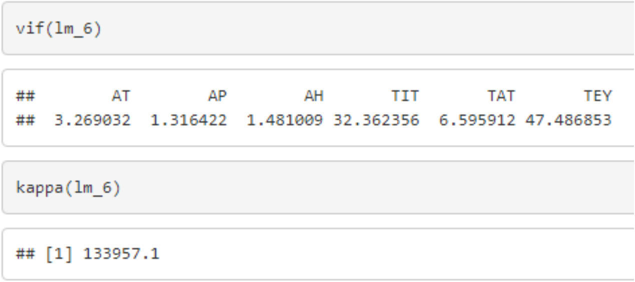
The new linear model, using log transformation of NOX as the response and the 6 variables selected above as the predictors, is shown in Fig.[26].



```
## 
## 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
```

Figure 26: Model Summary of 6 predictor model of NOx

The results show an adjusted coefficient of determination for this model to be 0.534 and all predictors as highly significant. The variance inflation factors for the predictors (fig [27])were only changed very slightly. Another method to help fix this multicollinearity is to use ridge regression.



```
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
```

Figure 27: variance inflation factors for 6 predictor model of NOx

3.2.d. Ridge Regression

We implemented a ridge regression model in an attempt to reduce the variance but the best possible for λ was 0.03 which is too low and there was no impact on the parameters as seen in the fig [28]

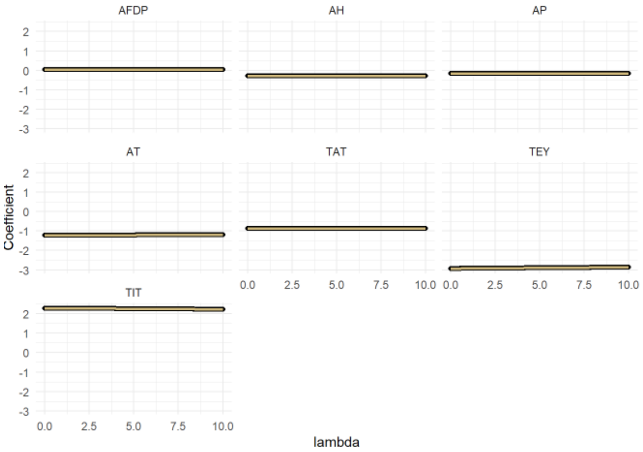


Figure 28: Coefficient vs lambda plot for ridge regression of NOx

Model	R^2	Adj R^2	MAE	MSPE
6 predictor model	0.5341	0.534	4.5341	54.34
Ridge Regression	-	-	5.73	68.55

Table 3: Metrics Comparison of different models of NOx

4. Evaluation and Results

After fixing the assumptions of linear regression we tested different possibilities for CO and NO_x concluded in the earlier sections. The results are formulated in the table[3].

5. Conclusion and Future work

Through this paper we test the statistical significance of a predictive emission monitoring system (PEM). The paper we based this research on did sophisticated neural networks. Considering we achieved comparable results with statistical modeling which is more interpretable and explainable would mean that the 9 features used can very well be used to set up a PEM. There could be more uses to this model like building confidence intervals for each predictors to set up triggers when the emissions exceed the limit set up.

In Future we can refine these models using generalized least squares and generalized additive models as there were still some assumptions violated in our analysis.

REFERENCES

[1] KAYA, HEYSEM; TÜFEKCİ, PINAR; and UZUN, ERDİNÇ (2019) "Predicting CO and NOx emissions from gas turbines: novel data and a benchmark PEMS," *Turkish Journal of Electrical Engineering and Computer Sciences*: Vol. 27: No. 6, Article 53. https://doi.org/10.3906/elk-1807-87

[2] Heysem Kaya, Department of Information and Computing Sciences, Utrecht University, 3584 CC, Utrecht, The Netherlands Email: h.kaya'@' uu.nl https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set