1. Introduction

PM2.5 particles are tiny particles that remain in the air and are produced from combustion processes, industrial operations, and vehicle exhaust, as well as from dust and pollen. PM2.5 predictions play an important role in benefiting society by promoting public health, increasing environmental concern, helping urban planning etc. These contributions lead to a reduction in health issues related to air pollution.

The dataset "Beijing PM2.5 Data" contains air quality data for Beijing, China, specifically the concentration of PM2.5 particulate matter.

2. Objective

The project aims to understand how the variables in "Beijing PM2.5 Data" affect the pm2.5 concentration. For this, we choose pm2.5 as Dependent Variable and all the other variables except pm2.5 as Independent variables.

3. Gathering Data

1) Source: The data is from the U.S. Embassy in Beijing, where hourly air quality readings were documented. The dataset period is from Jan 1st, 2010 to Dec 31st, 2014. Missing data are denoted as NA. There are 43824 instances and 13 attributes.

2) Variables: The dataset contains the following variables:

No: A unique identifier for each row.

Year: The year of the observation.

Month: The year of the observation.

Day: The year of the observation.

Hour: The year of the observation.

pm2.5: The concentration of PM2.5 particulate matter (measured in micrograms per cubic meter).

DEWP: The dew point temperature in Celsius.

TEMP: The temperature in Celsius.

PRES: The air pressure in hPa.

Cbwd: Combined wind direction

Iws: Cumulated wind speed

Is: Cumulated hours of snow

Ir: Cumulated hours of rain

Types of Variables in the Dataset

```
'data.frame':   43824 obs. of  13 variables:
 $ No    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ year : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
 $ month: int  1 1 1 1 1 1 1 1 1 1 ...
 $ day  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ hour : int  0 1 2 3 4 5 6 7 8 9 ...
 $ pm2.5: int  NA NA NA NA NA NA NA NA NA NA ...
 $ DEWP : int  -21 -21 -21 -21 -20 -19 -19 -19 -19 -20 ...
 $ TEMP : num  -11 -12 -11 -14 -12 -10 -9 -9 -9 -8 ...
 $ PRES : num  1021 1020 1019 1019 1018 ...
 $ cbwd : chr  "NW" "NW" "NW" "NW" ...
 $ Iws  : num  1.79 4.92 6.71 9.84 12.97 ...
 $ Is   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Ir   : int  0 0 0 0 0 0 0 0 0 0 ...
```

3.1. Data Cleaning

We need to validate and verify the data before further processing. **pm2.5** attribute contains NA values, hence the NA values are removed manually. Since the attribute **No** doesn't contribute to the analysis it is also removed from the dataset

'cbwd' attributes contain values NE(NorthEast) , NW(NorthWest), SE(SouthEast) and cv. Compare to the other directions cv seems to be odd, hence cv is substituted to the value SW(SouthWest)

Now the cleaned dataset has variables as below:

```
'data.frame':   41757 obs. of  12 variables:
 $ pm2.5: int  994 980 972 886 858 852 845 824 810 805 ...
 $ year : int  2012 2010 2012 2013 2013 2013 2013 2013 2013 2013 ...
 $ month: int  1 2 1 1 1 1 1 1 1 1 ...
 $ day  : int  23 14 23 12 12 12 12 12 12 12 ...
 $ hour : int  1 1 2 20 22 21 16 19 17 23 ...
 $ DEWP : int  -24 -14 -24 -8 -10 -9 -7 -8 -7 -10 ...
 $ TEMP : num  -12 -7 -12 -7 -9 -8 -2 -7 -4 -9 ...
 $ PRES : num  1032 1029 1032 1023 1024 ...
 $ cbwd : chr  "NW" "cv" "NW" "cv" ...
 $ Iws  : num  4.92 0.89 8.05 1.34 0.89 0.89 8.95 0.89 9.84 1.79 ...
 $ Is   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Ir   : int  0 0 0 0 0 0 0 0 0 0 ...
```

3.2. Summary

Summary of the values contained in each column of the dataset

```
> summary(Mydata)
     pm2.5             year          month             day             hour            DEWP       
 Min.   :  0.00   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   : 0.0   Min.   :-40.00  
 1st Qu.: 29.00   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 5.0   1st Qu.:-10.00  
 Median : 72.00   Median :2012   Median : 7.000   Median :16.00   Median :12.0   Median :  2.00  
 Mean   : 98.61   Mean   :2012   Mean   : 6.514   Mean   :15.69   Mean   :11.5   Mean   :  1.75  
 3rd Qu.:137.00   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:18.0   3rd Qu.: 15.00  
 Max.   :994.00   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.   :23.0   Max.   : 28.00  
      TEMP            PRES          cbwd                Iws               Is          
 Min.   :-19.0   Min.   : 991   Length:41757       Min.   :  0.45   Min.   : 0.00000  
 1st Qu.:  2.0   1st Qu.:1008   Class :character   1st Qu.:  1.79   1st Qu.: 0.00000  
 Median : 14.0   Median :1016   Mode  :character   Median :  5.37   Median : 0.00000  
 Mean   : 12.4   Mean   :1016                      Mean   : 23.87   Mean   : 0.05534  
 3rd Qu.: 23.0   3rd Qu.:1025                      3rd Qu.: 21.91   3rd Qu.: 0.00000  
 Max.   : 42.0   Max.   :1046                      Max.   :565.49   Max.   :27.00000  
       Ir         
 Min.   : 0.0000  
 1st Qu.: 0.0000  
 Median : 0.0000  
 Mean   : 0.1949  
 3rd Qu.: 0.0000  
 Max.   :36.0000  
```

4. Diagnostics

Regression diagnostics includes several methods and analyses used to assess the effectiveness, assumptions, and dependability of regression models. Their purpose is to evaluate how well a regression model fits the data, identify issues or violations of assumptions, and help the enhancement and adjustment of the model. Regression diagnostics are important for confirming the results and ensuring how much the regression analysis is reliable.

Common techniques and measures used in regression diagnostics include:

Residual analysis: Residuals are the difference between the observed values and the predicted values obtained from the regression model. Analyzing residuals helps identify deviations, unequal variances (heteroscedasticity), and outliers. Residual plots, such as scatterplots where residuals against fitted values or independent variables are plotted, help to make inferences.

Normality and heteroscedasticity tests: Regression models frequently assume that residuals are normally distributed with constant variance (homoscedasticity).

Goodness-of-fit measures: Various measurements check the overall quality of fit of a regression model, including the coefficient of determination (R-squared), adjusted R-squared, root mean square error (RMSE), Mean Absolute Error (MAE) or Akaike information criterion (AIC), among others. These metrics measure the amount of variance in the model and help in model comparison.

Linearity: Assumes that there is a linear connection between the predictor variables (independent variables) and the response variable (dependent variable) being analyzed. It assumes that the relationship can be accurately described by a straight line or a linear combination of the independent variables.

Correlation: The association between the predictor variables (independent variables) and the response variable (dependent variable) within a regression model. It measures the severity and direction of the linear connection between these variables.

4.1. Correlation Matrix

We find the Correlation between the Independent variables and the dependent variable. Below is the correlation matrix showing the association between pm2.5 and other independent variables.

```
> print(cor_matrix)
            pm2.5          year         month           day          hour         DEWP        TEMP
pm2.5  1.00000000 -0.0146901999 -0.0240687836  0.0827884927 -0.0231164430  0.171423272 -0.09053400
year  -0.01469020  1.0000000000 -0.0024521591 -0.0001027102  0.0002000588  0.007298028  0.05565572
month -0.02406878 -0.0024521591  1.0000000000  0.0069009497 -0.0005427315  0.234491983  0.17213525
day    0.08278849 -0.0001027102  0.0069009497  1.0000000000  0.0003268606  0.033536769  0.02287140
hour  -0.02311644  0.0002000588 -0.0005427315  0.0003268606  1.0000000000 -0.021783832  0.14944294
DEWP   0.17142327  0.0072980285  0.2344919827  0.0335367692 -0.0217838316  1.000000000  0.82382123
TEMP  -0.09053400  0.0556557207  0.1721352533  0.0228713977  0.1494429386  0.823821233  1.00000000
PRES  -0.04728231 -0.0134661457 -0.0663170285 -0.0104967856 -0.0418312957 -0.777722121 -0.82690281
Iws   -0.24778445 -0.0682777137  0.0146635756 -0.0049436552  0.0588653488 -0.293105921 -0.14961252
Is     0.01926558 -0.0195492097 -0.0628832535 -0.0374488001 -0.0024547276 -0.034925232 -0.09478480
Ir    -0.05136871 -0.0262978320  0.0388739475 -0.0001019233 -0.0087407540  0.125340756  0.04954445
             PRES          Iws            Is            Ir
pm2.5 -0.04728231 -0.247784449  0.019265576 -0.0513687055
year  -0.01346615 -0.068277714 -0.019549210 -0.0262978320
month -0.06631703  0.014663576 -0.062883254  0.0388739475
day   -0.01049679 -0.004943655 -0.037448800 -0.0001019233
hour  -0.04183130  0.058865349 -0.002454728 -0.0087407540
DEWP  -0.77772212 -0.293105921 -0.034925232  0.1253407561
TEMP  -0.82690281 -0.149612519 -0.094784798  0.0495444536
PRES   1.00000000  0.178871492  0.070537123 -0.0805322089
Iws    0.17887149  1.000000000  0.022630317 -0.0091569394
Is     0.07053712  0.022630317  1.000000000 -0.0097638617
Ir    -0.08053221 -0.009156939 -0.009763862  1.0000000000
```

Independent Variables	Correlation with pm2.5
year	Negative correlation with pm2.5
month	Negative correlation with pm2.5
day	Positive correlation with pm2.5
hour	Negative correlation with pm2.5
DEWP	Positive correlation with pm2.5
TEMP	Negative correlation with pm2.5
PRES	Negative correlation with pm2.5
Iws	Negative correlation with pm2.5
Is	Positive correlation with pm2.5
Ir	Negative correlation with pm2.5

DEWP has a high positive correlation with pm2.5 and **Iws** have a high negative correlation with pm2.5

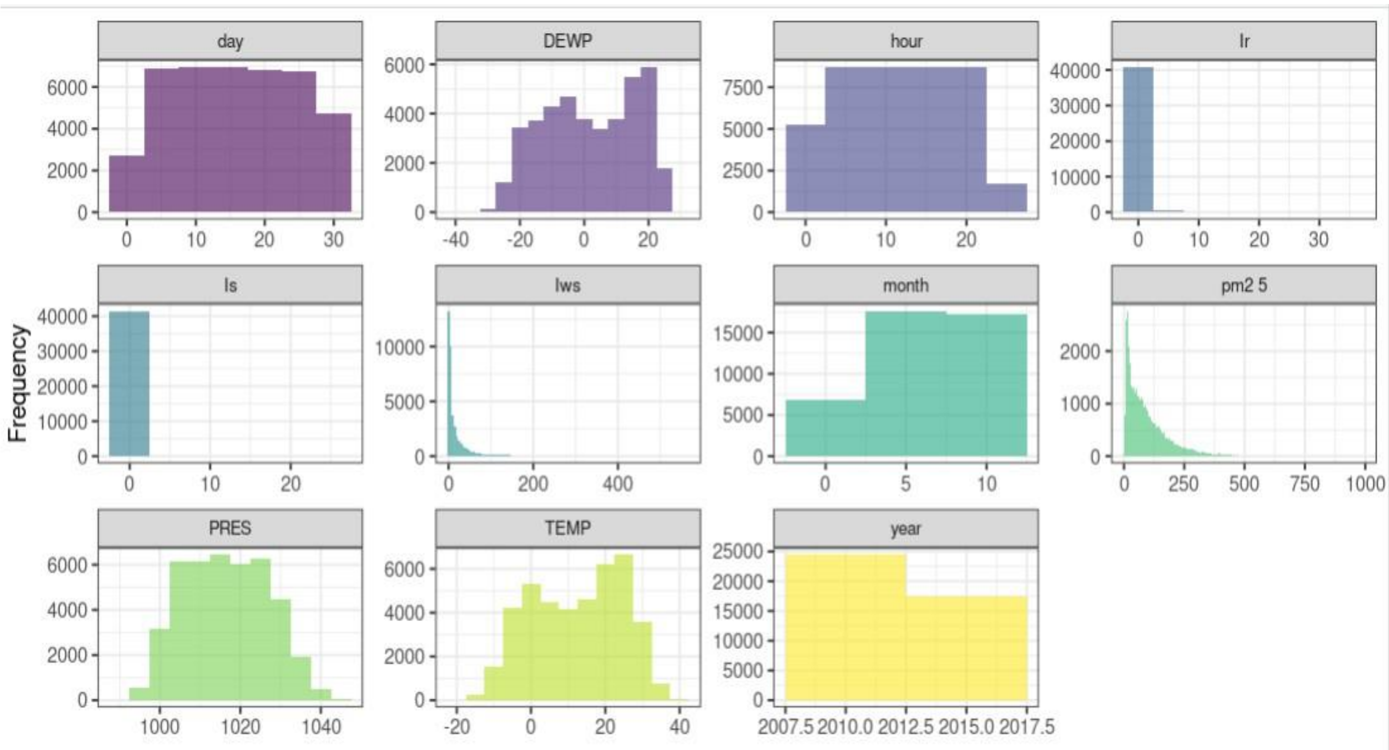
With the help of code, the best correlation is found, and it seems to be with **Iws**

Values	
best_correlated_index	8L
best_correlated_variable	"Iws"

```
> cor(Mydata$pm2.5,Mydata$Iws)
[1] -0.2477844
>
```

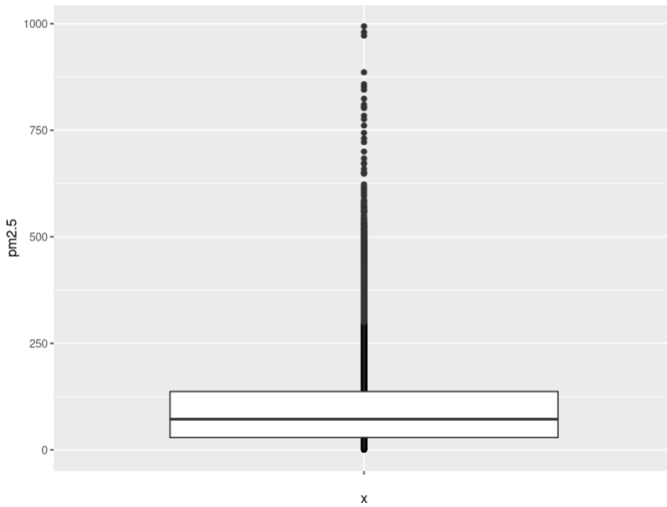
4.2. Normality

Histogram is plotted for all the variables except the categorical variable cbwd. The histogram of the dependent variable pm2.5 is right skewed, and the distribution is not normal. Hence the normality assumption is violated



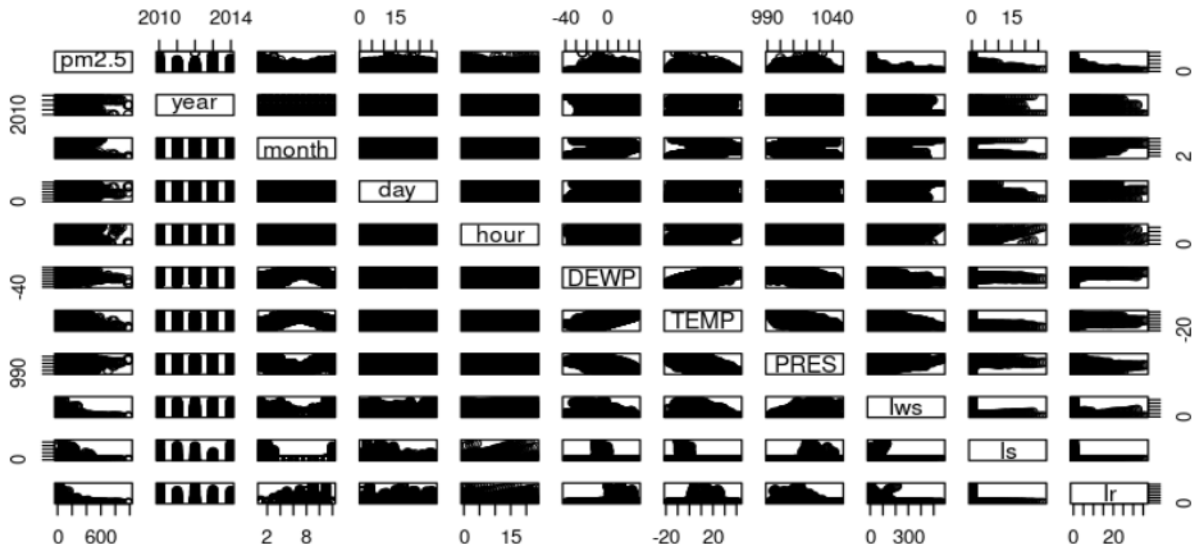
4.3. Outliers

By plotting the boxplots we can check if there are outliers in the data. Outliers can refer to data points that significantly deviate from the majority of observations. These outliers can occur if there are extreme values or errors in the data. The boxplot displays points outside the whiskers which means there are outliers for the pm2.5 variable



4.4 Linearity

Scatterplots are plotted to check if the data is showing a linear relationship between the dependent and independent variables. When working with multiple variables, it is often useful to generate a matrix of scatter plots to visualize the relationships between variables. This scatterplot matrix, also called a pairwise scatter plot, displays each variable against all others, enabling the analysis of correlations. In R, you can create a scatterplot matrix using the pairs() function.pm2.5 doesn't seem to have a linear relationship with any of the independent variables, hence we need to model non-linearity.



4.5 Homoscedasticity

Homoscedasticity is a concept in regression analysis that assumes a constant variance for the residuals, which are the discrepancies between the actual and predicted values. It means the dispersion of the residuals remains consistent across all levels of the independent variables, otherwise, the variability of the residuals remains uniform throughout the entire range of the predictor variables. Homoscedasticity is violated in our dataset.

5. Initial Modelling

5.1. Building First Model

While building the first model we will consider the simple linear regression where the dependent

variable and an independent is modelled.

Below is the equation for Simple Linear Regression :

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Y is the Dependent variable to be predicted

β_0 is the y-Intercept

β_1 is the slope for the independent variable

X is the independent variable

ε is the error term

The first model is the small model where the Dependent variable pm2.5 is modelled with each independent variable one by one.

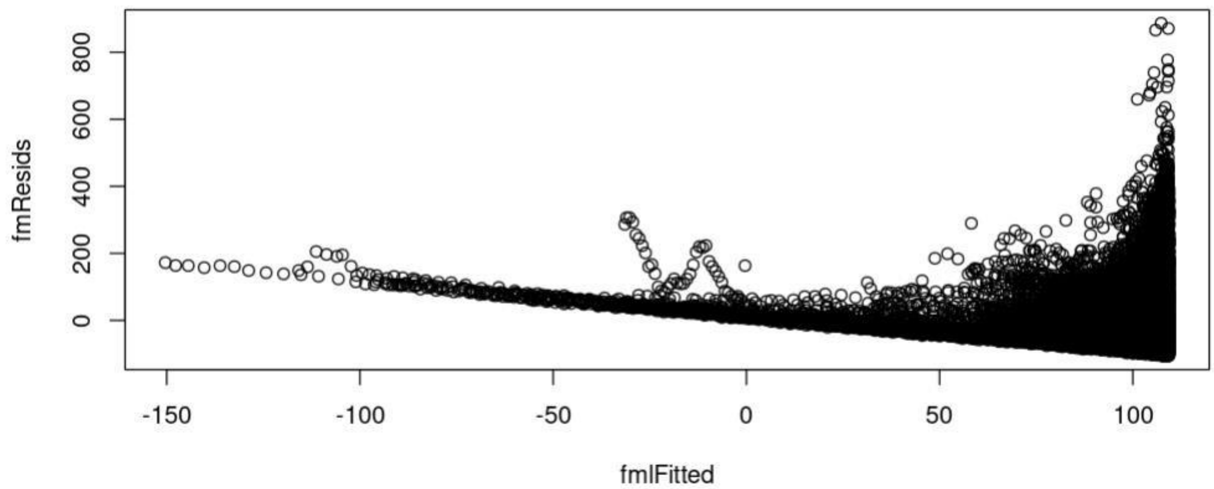
When the Dependent variable **pm2.5** is modelled with the independent variable **Iws**, we get the maximum Adjusted R-squared.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.584514   0.484288  226.28   <2e-16 ***
Iws          -0.459690   0.008796  -52.26   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89.18 on 41755 degrees of freedom
Multiple R-squared:  0.0614,	Adjusted R-squared:  0.06137
F-statistic:  2731 on 1 and 41755 DF,  p-value: < 2.2e-16
```

We Train a linear model with response = pm2.5 and a single predictor **Iws**, with 10-fold cross-validation and the RMSE and MAE are as below:

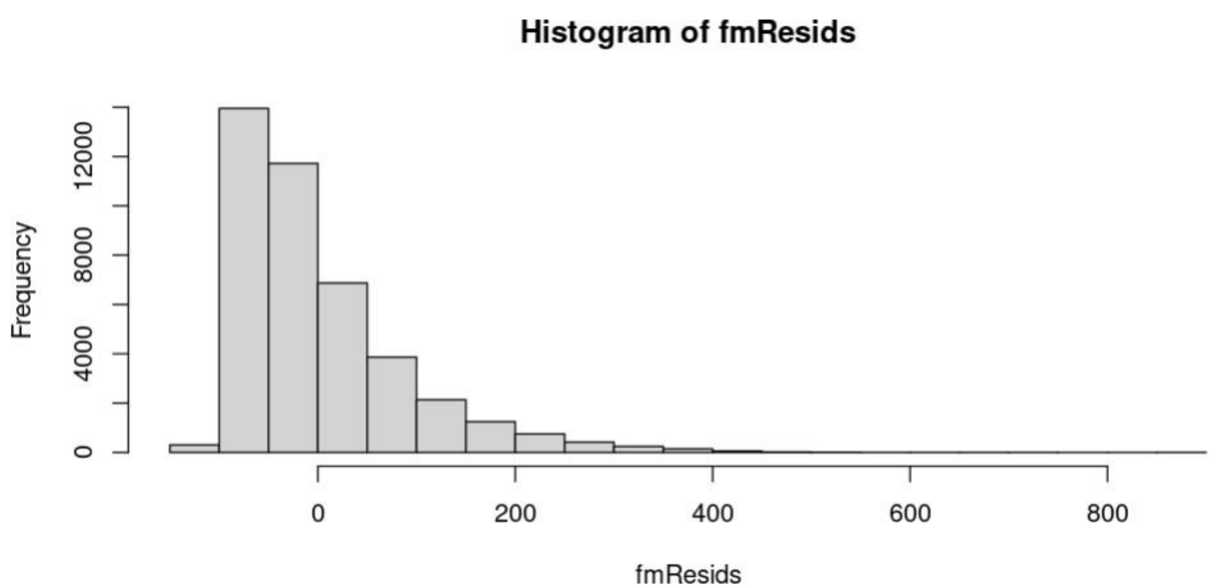
```
RMSE      Rsquared    MAE
89.15138  0.06162772  66.20343
```

Now the Fitted vs Residual plot is plotted for the First model and it looks like the below:



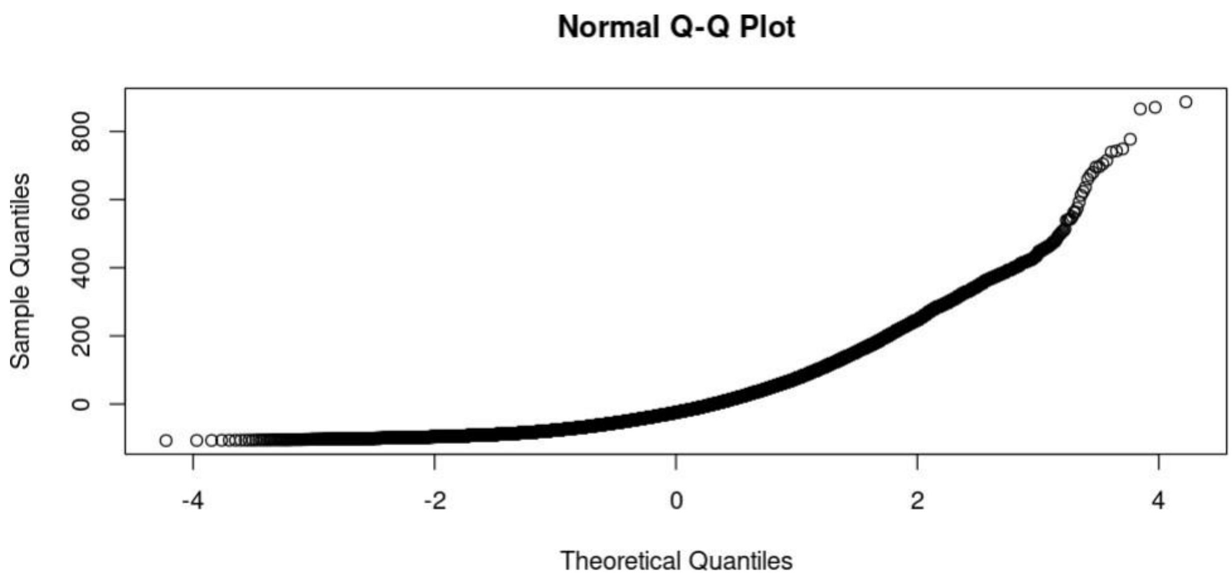
The plot is clustered and doesn’t follow Homoscedasticity.

Now let’s see the histogram for the first model residual



The distribution is not normal and may lead to Inaccurate confidence intervals.

Below is the qqplot for the first model Residuals



There are outliers for this model.

5.2. Building Second Model

While building the second model we will consider the simple linear regression where the dependent variable and all the independent variables are modelled.

Below is the equation for Simple Linear Regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

Y is the Dependent variable to be predicted

β_0 is the y-Intercept

β_1 is the slope for the independent variable

X is the independent variable

ε is the error term

- First, we model with all the predictors except the time-series variables like the year, month, hour and day

pm2.5~ DEWP + TEMP + PRES + Iws + Is + Ir + cbwd

- Then, we model with all the predictors except the categorical variable,cbwd

pm2.5~year + month + day + hour + DEWP + TEMP + PRES + Iws + Is + Ir

- Then, we model with all the predictors except both time-series variables like the year, month, hour, date and categorical variable cbwd

pm2.5~ DEWP + TEMP + PRES + Iws + Is + Ir

When all these steps are done the Adjusted R-squared seems to be low.

- Later when modelled Dependent variable pm2.5 with all the Independent variables

pm2.5~year + month + day + hour + DEWP + TEMP + PRES + Iws + Is + Ir + cbwd

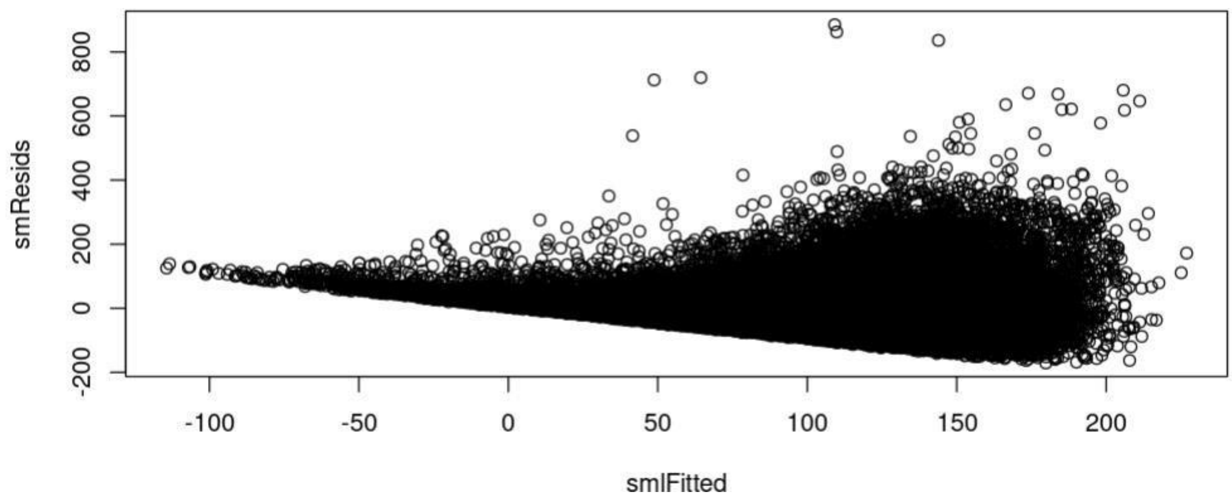
Now the Adjusted R-squared seems to be improved

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 736.229614 552.789859   1.332   0.1829
year          0.538228   0.273754   1.966   0.0493 *
month        -1.059233   0.119093  -8.894  < 2e-16 ***
day           0.764684   0.043736  17.484  < 2e-16 ***
hour          1.293366   0.059603  21.700  < 2e-16 ***
DEWP          4.408548   0.056351  78.234  < 2e-16 ***
TEMP         -6.632793   0.070476 -94.114  < 2e-16 ***
PRES         -1.621562   0.071413 -22.707  < 2e-16 ***
Iws          -0.201792   0.008714 -23.157  < 2e-16 ***
Is           -3.406946   0.499126  -6.826 8.86e-12 ***
Ir           -6.481966   0.275829 -23.500  < 2e-16 ***
cbwdNE      -25.413164   1.416113 -17.946  < 2e-16 ***
cbwdNW      -27.147094   1.168523 -23.232  < 2e-16 ***
cbwdSE        0.486569   1.093073   0.445   0.6562
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 78.38 on 41743 degrees of freedom
Multiple R-squared:  0.2753,	Adjusted R-squared:  0.275
F-statistic:  1220 on 13 and 41743 DF,  p-value: < 2.2e-16
```

Now we train a linear model with response = pm2.5 and every predictor, with 10-fold cross-validation and the RMSE and MAE are as below:

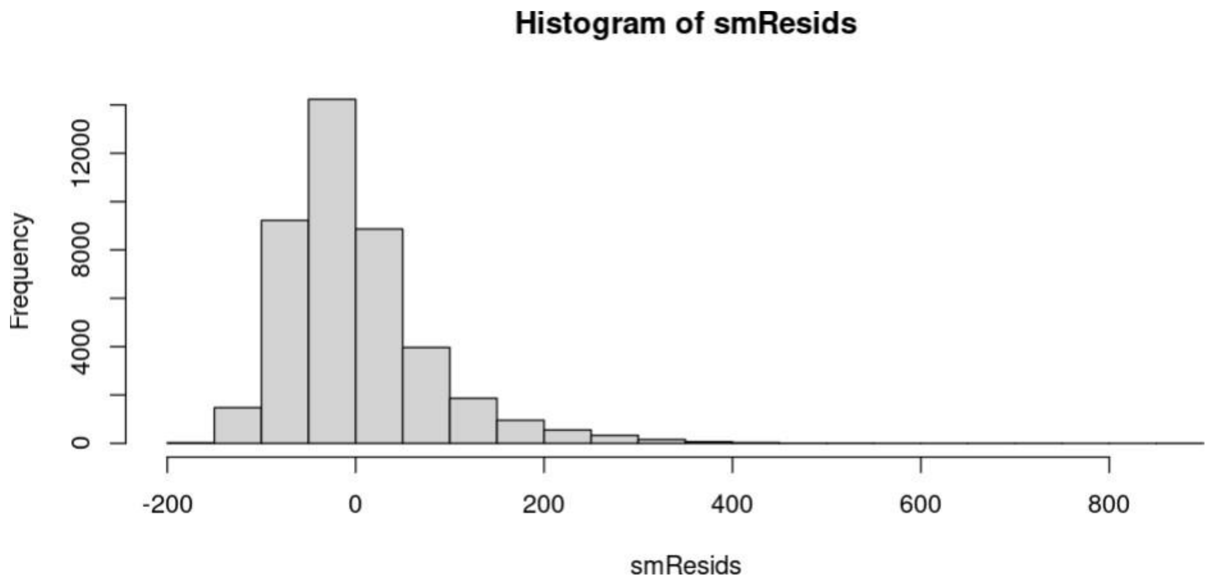
```
RMSE      Rsquared   MAE
78.35729  0.2753475  56.77377
```

Now the Fitted vs Residual plot is plotted for the second model and it looks like below:



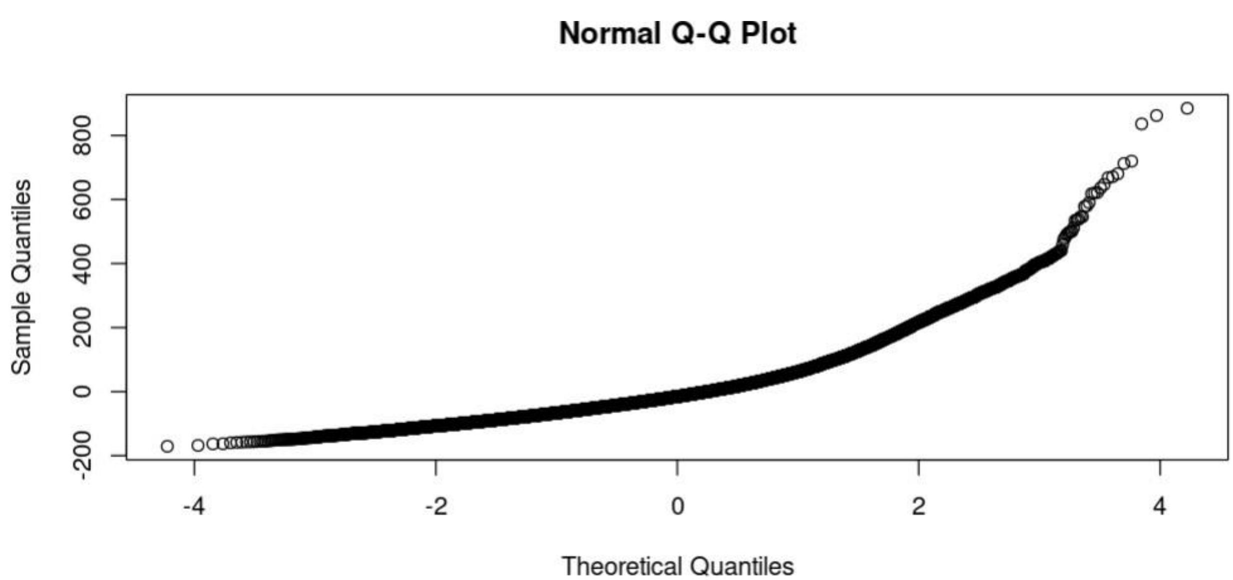
The plot is clustered and doesn't follow Homoscedasticity.

Now let's see the histogram for the second model residual



The distribution is not normal too and may lead to Inaccurate confidence intervals.

Below is the qqplot for the first model Residuals



There are outliers for this model too.

6. Model Selection

We know that the relationship between Dependent and Independent variables should be linear. But in the above cases, the linearity assumption is violated.

So to deal with the non-linearity we use the techniques below:

6.1. Polynomial Regression

Polynomial regression is a technique that helps to model nonlinearity between explanatory variables and a response variable. It fits the model to a curved line rather than fitting it to a straight line.

In polynomial regression, the relationship between an independent variable (x) and a dependent variable (y) is conveyed through the equation below:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n$$

In this equation

β_0, β_1 ….. β_n are the coefficients that find out the shape of the curve

Here in the case of the pm2.5 prediction, we add the squared independent variable one by one to check the better Adjusted R-squared.

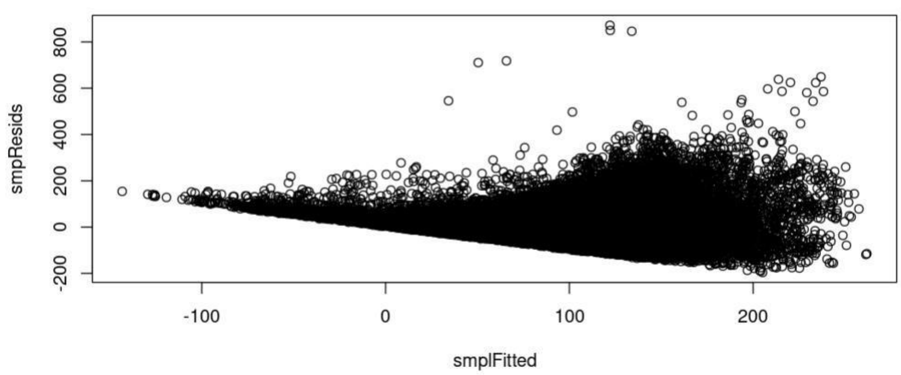
When the Independent variable **month** is squared and added we get the better Adjusted R-squared which is as below:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.927e+03  5.371e+02   5.449  5.1e-08 ***
year         -3.172e-01  2.657e-01  -1.194    0.233    
month        -6.190e+01  1.167e+00 -53.057  < 2e-16 ***
I(month^2)    4.475e+00  8.539e-02  52.406  < 2e-16 ***
day           6.731e-01  4.240e-02  15.875  < 2e-16 ***
hour          6.849e-01  5.889e-02  11.631  < 2e-16 ***
DEWP          5.350e+00  5.747e-02  93.102  < 2e-16 ***
TEMP         -4.182e+00  8.275e-02 -50.541  < 2e-16 ***
PRES         -1.957e+00  6.947e-02 -28.167  < 2e-16 ***
Iws          -1.631e-01  8.473e-03 -19.251  < 2e-16 ***
Is           -4.607e+00  4.840e-01  -9.517  < 2e-16 ***
Ir           -5.909e+00  2.674e-01 -22.099  < 2e-16 ***
cbwdNE       -2.273e+01  1.373e+00 -16.559  < 2e-16 ***
cbwdNW       -2.668e+01  1.132e+00 -23.567  < 2e-16 ***
cbwdSE       -1.004e+00  1.059e+00  -0.948    0.343    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 75.92 on 41742 degrees of freedom
Multiple R-squared:   0.32,	Adjusted R-squared:  0.3198 
F-statistic:  1403 on 14 and 41742 DF,  p-value: < 2.2e-16
```

Now we Train a linear model with response = pm2.5 and a single predictor month, with 10-fold cross-validation and the RMSE and MAE are as below:

```
 RMSE      Rsquared     MAE
 92.02483  0.001173242  68.82687
```

Now the Fitted vs Residual plot is plotted for the model where polynomial regression is introduced and it looks like the below:



The plot is clustered and doesn't follow Homoscedasticity.

Now Feature Selection is done for this model

Feature selection is the procedure of selecting a subset of meaningful variables from a larger set of variables within a dataset.

Forward feature selection begins with an empty set first and adds features one by one, taking into account their performance. The goal is to discover the ideal subset of variables that maximizes the predictive capability.

Whereas, the backward feature starts with a complete set of variables and removes them based on their performance one by one. The goal is to identify the essential variables while simplifying the model in the process.

1) Forward feature selection:

While doing the forward feature selection in the model where polynomial regression is used, we can see there are no variables which is asked to be removed:

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
pm2.5 ~ year + month + I(month^2) + day + hour + DEWP + TEMP +
    PRES + Iws + Is + Ir + cbwd

Final Model:
pm2.5 ~ year + month + I(month^2) + day + hour + DEWP + TEMP +
    PRES + Iws + Is + Ir + cbwd


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                      41742  240592054 361603.5
> |
```

2) Backward feature selection:

While doing the backward feature selection in the model where polynomial regression is used, we can see the variable **year** is asked to be removed:

```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
pm2.5 ~ year + month + I(month^2) + day + hour + DEWP + TEMP +
    PRES + Iws + Is + Ir + cbwd

Final Model:
pm2.5 ~ month + I(month^2) + day + hour + DEWP + TEMP + PRES +
    Iws + Is + Ir + cbwd


    Step Df Deviance Resid. Df Resid. Dev      AIC
1                        41742  240592054 361603.5
2 - year  1 8215.274     41743  240600269 361602.9
> |
```

After removing the variable year from the model, we still can't see an improvement in Adjusted R-squared

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.291e+03  7.111e+01  32.218   <2e-16 ***
month       -6.181e+01  1.164e+00 -53.082   <2e-16 ***
I(month^2)   4.469e+00  8.523e-02  52.431   <2e-16 ***
day          6.733e-01  4.240e-02  15.879   <2e-16 ***
hour         6.882e-01  5.883e-02  11.698   <2e-16 ***
DEWP         5.354e+00  5.737e-02  93.332   <2e-16 ***
TEMP        -4.194e+00  8.211e-02 -51.084   <2e-16 ***
PRES        -1.959e+00  6.943e-02 -28.221   <2e-16 ***
Iws         -1.625e-01  8.458e-03 -19.214   <2e-16 ***
Is          -4.601e+00  4.840e-01  -9.506   <2e-16 ***
Ir          -5.907e+00  2.674e-01 -22.091   <2e-16 ***
cbwdNE      -2.271e+01  1.373e+00 -16.544   <2e-16 ***
cbwdNW      -2.661e+01  1.131e+00 -23.537   <2e-16 ***
cbwdSE      -9.791e-01  1.059e+00  -0.925    0.355    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.92 on 41743 degrees of freedom
Multiple R-squared:   0.32,	Adjusted R-squared:  0.3198 
F-statistic:  1511 on 13 and 41743 DF,  p-value: < 2.2e-16
```

6.2. Adding Interaction Terms (Additivity)

In regression analysis, the inclusion of interaction terms, which is also referred to as considering additivity, is a method employed to capture how two or more variables jointly influence the dependent variable. This technique enables the examination and modelling of relationships between variables that are not simply additive in nature.

Here in the case of the pm2.5 prediction, we add the interaction term one by one along with the polynomial to check the better Adjusted R-squared.

When the Independent variable **month** is squared and added **TEMP** as the Interaction term we get the better Adjusted R-squared which is as below:

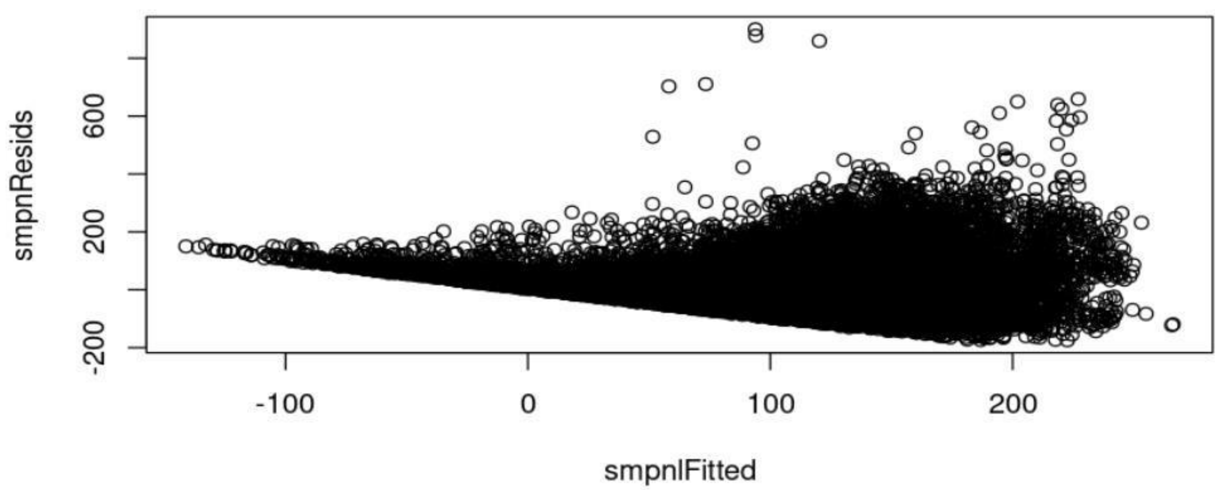
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.760e+03  5.320e+02   5.189 2.12e-07 ***
year        -5.561e-01  2.633e-01  -2.113   0.0346 *  
month       -7.013e+01  1.191e+00 -58.886  < 2e-16 ***
I(month^2)   5.135e+00  8.768e-02  58.564  < 2e-16 ***
day          4.243e-01  4.289e-02   9.892  < 2e-16 ***
hour         6.910e-01  5.833e-02  11.847  < 2e-16 ***
DEWP         6.077e+00  6.236e-02  97.451  < 2e-16 ***
TEMP        -1.387e+00  1.278e-01 -10.856  < 2e-16 ***
PRES        -1.301e+00  7.254e-02 -17.940  < 2e-16 ***
Iws         -1.678e-01  8.394e-03 -19.995  < 2e-16 ***
Is          -4.625e+00  4.794e-01  -9.647  < 2e-16 ***
Ir          -5.999e+00  2.649e-01 -22.649  < 2e-16 ***
cbwdNE      -2.151e+01  1.360e+00 -15.814  < 2e-16 ***
cbwdNW      -2.463e+01  1.123e+00 -21.929  < 2e-16 ***
cbwdSE      -2.637e+00  1.051e+00  -2.510   0.0121 *  
month:TEMP  -3.828e-01  1.342e-02 -28.522  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75.19 on 41741 degrees of freedom
Multiple R-squared:  0.333,	Adjusted R-squared:  0.3328 
F-statistic:  1389 on 15 and 41741 DF,  p-value: < 2.2e-16
```

Now Train a linear model with response = pm2.5 and a single predictor TEMP, with 10-fold cross-validation, we get the RMSE and MAE as below

```
RMSE      Rsquared     MAE
91.66245  0.008368761  68.91715
```

Now the Fitted vs Residual plot is plotted for the model where polynomial regression and interaction term is added and it looks like below:

The plot is clustered and doesn't follow Homoscedasticity.



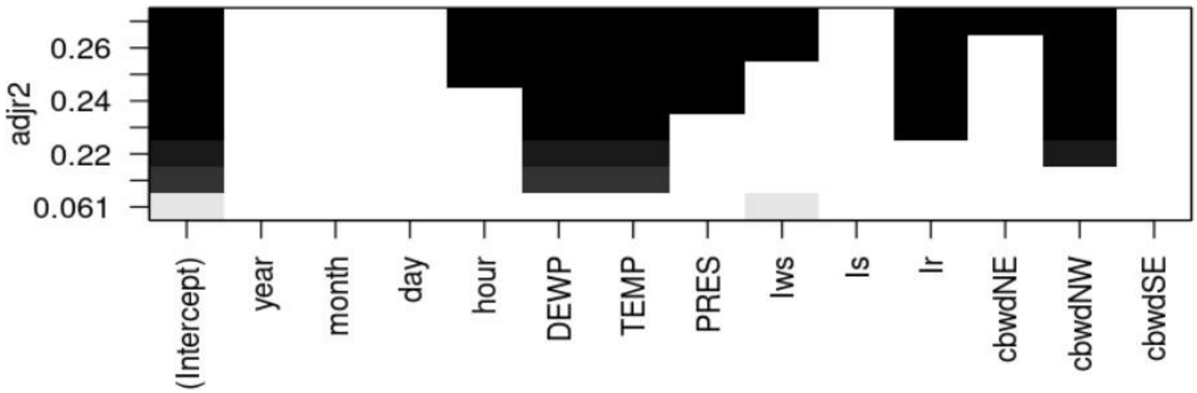
The plot is clustered and doesn't follow Homoscedasticity.

When feature selection is applied, no variables are asked to remove from the model

6.3. Subset Selection

Subset selection is a method used in feature selection where a smaller set of features is chosen from a larger set. The objective of subset selection is to pinpoint the most informative and significant features that have the greatest impact on the predictive accuracy or comprehension of a model.

We use the best subset selection to find the "best" model for pm2.5 and plot a graph of this using the adj2 scale, it looks like this:



The top model seems to be DEWP, TEMP and cbwd

Now we Train a linear model with response = pm2.5 and best predictors, with 10-fold cross-validation and the RMSE and MAE are as below:

```
RMSE      Rsquared  MAE
80.69105  0.231624  58.60049
```

7. Prediction and Summary

We are checking here, if the values of **Iws** are 10,20 and 30 what is the value of pm2.5 going to be?

Considering Iws since the Adjusted R-squared was high while performing simple linear regression. The predicted values are

```
> Iws_predictions <- data.frame(Iws = c(10, 20, 30))
> predict(fm8,Iws_predictions)
        1         2         3
104.98762 100.39072  95.79382
> 
```

so as the Iws increases ,pm2.5 value decreases(Negative Correlation)

Now we want the range - 95% confident that the predicted value will fall within this range.

```
> prediction_interval
        fit       lwr       upr
1 104.98762 104.09944 105.87579
2 100.39072  99.53273 101.24871
3  95.79382  94.93191  96.65573
> 
```

So we are 95% confident that the predicted value will fall within the above range.

After training several linear models with response = pm2.5 and other predictors, with 10-fold cross-validation we can conclude that the second model (response = pm2.5 and all the predictors) is better with RMSE and MAE values 78.35729 and 56.77377 respectively and with Adjusted R-squared 0.2753475 comparing to other models. Even though this model is not the best and optimum, compared to other models it seems to have the least RMSE, MAE and better Adjusted R-squared.