

## 1. Data Preprocessing

## 1.1 Exploring Data-Set

The screenshot shows an RStudio interface with several open panes:

- Code Editor:** Displays R code for data processing, including reading CSV files, setting working directories, and performing statistical summaries.
- Console:** Shows the output of the R code, including summary statistics for variables like month, day, and hour.
- Environment:** Shows the global environment with objects like data, month, year, and various summary statistics.
- Data View:** Displays the raw data frame with 43824 observations and 13 variables.
- Plots:** Shows a histogram of the 'month' variable.
- Packages:** Lists available packages such as tidyverse, dplyr, and lubridate.
- Help:** Provides help documentation for various R functions.
- Viewer:** Shows the raw text of the assignment files.
- Presentation:** Shows the assignment files in a presentation format.

```
head(data)  
summary(data)  
str(data)  
dim(data)
```

`head(data)`: This command gives a brief overview of the data by displaying the top few rows of your dataset.

`summary(data)`: This command offers a summary of the dataset's variables' respective statistical data. Depending on the kind of variable, it includes metrics like mean, median, minimum, maximum, and quartiles.

`str(data)`: This command displays the first few values of each variable together with the data types, giving the dataset's structure. It helps comprehend the many kinds of variables in your dataset.

`dim(data)`: This command provides with the dataset's number of rows and columns.

## 1.DATAPREPROCESSING

### 1.2 Missing Values

Filled in them by using the mean.

The screenshot shows the RStudio interface. In the top-left pane, there is an R script with the following code:

```
1 getwd("C:/Users/SRUJA/OneDrive/Desktop/Pgm in DA Project")
2 setwd("C:/Users/SRUJA/OneDrive/Desktop/Pgm in DA Project")
3 data.read.csv(file="Project Data Set.csv",header=TRUE)
4 #Data preprocessing
5
6 #1.1 Exploring dataset
7
8 head(data)
9 summary(data)
10 str(data)
11 dim(data)
12
13 #1.2 Missing values
14 colSums(is.na(data)) #Total number of missing values in a column
15 data$pm2.5 <- ifelse(is.na(data$pm2.5),ave(data$pm2.5, FUN= function(x)mean(x,na.rm=TRUE)))
16 data$pm2.5
17
18 #1.3 Duplicate values
19 colSums(is.na(data)) #Total number of missing values in a column
20 colSums(is.na(data))
21 data<- ifelse(is.na(data$pm2.5),ave(data$pm2.5, FUN= function(x)mean(x,na.rm=TRUE)))
22 data$pm2.5
```

In the top-right pane, the "Environment" tab shows the data frame structure:

```
$ No : int 1 2 3 4 5 6 7 8 9 10 ...
$ year : int 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 : num 98.6 98.6 98.6 98.6 98.6 ...
$ DEWP : int -21 -21 -21 -21 -20 -19 -19 -19 -19 -20 ...
$ TEMP : num -11 -12 -11 -14 -12 -10 -9 -9 -9 -8 ...
$ 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 ...
```

The bottom-right pane shows a file browser with a list of files and their details.

## 1.DATAPREPROCESSING

### 1.3 Duplicate Rows Check

The screenshot shows the RStudio interface. In the top-left pane, there is an R script with the following code:

```
1 getwd("C:/Users/SRUJA/OneDrive/Desktop/Pgm in DA Project")
2 setwd("C:/Users/SRUJA/OneDrive/Desktop/Pgm in DA Project")
3 data.read.csv(file="Project Data Set.csv",header=TRUE)
4 #Data preprocessing
5
6 #1.1 Exploring dataset
7
8 head(data)
9 summary(data)
10 str(data)
11 dim(data)
12
13 #1.2 Missing values
14 colSums(is.na(data)) #Total number of missing values in a column
15 data$pm2.5 <- ifelse(is.na(data$pm2.5),ave(data$pm2.5, FUN= function(x)mean(x,na.rm=TRUE)))
16 data$pm2.5
17
18 #1.3 Data duplicates
19 any(duplicated(data))
20
21 any(duplicated(data))
22
23 [1] FALSE
```

In the top-right pane, the "Environment" tab shows the data frame structure:

```
$ No : int 1 2 3 4 5 6 7 8 9 10 ...
$ year : int 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 : num 98.6 98.6 98.6 98.6 98.6 ...
$ DEWP : int -21 -21 -21 -21 -20 -19 -19 -19 -19 -20 ...
$ TEMP : num -11 -12 -11 -14 -12 -10 -9 -9 -9 -8 ...
$ 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 ...
```

The bottom-right pane shows a file browser with a list of files and their details.

## 1.DATA-PREPROCESSING

### 1.4 Correlation of Dataset

```

49
50
51 #Before building models, You need to check correlations b/w y & x- variables
52 #Apply transformations to x- variables if necessary.
53
54 cor(data)
55
56
57
58
59
60
61:1 (Top Level) : 

```

The screenshot shows the RStudio interface with the code for calculating the correlation matrix of the dataset. The console output displays the correlation matrix for all variables, including year, month, day, hour, pm2.5, DEWP, TEMP, Iws, and various factor variables (as.factor(data\$cbw)). The matrix shows high correlations between variables like pm2.5 and DEWP, and lower correlations between categorical factors.

## 1.DATA-PREPROCESSING

### 1.5 Using hold-out evaluation only, 80% as training

```

57 #Exclude nominal variables
58 data<-data[,1:11]
59 view(data)
60
61
62
63 #Training data 0.8
64
65 data=data[sample(nrow(data)),]
66 select.data = sample (1:nrow(data), 0.8*nrow(data))
67 train.data = data[select.data,]
68 test.data = data[-select.data,]
69
70
71
72 #Backward method using p-value in t-test as metric
73
74 m1=lm(pm2.5~., data = data)
75 summary(m1)
76
77 m1=lm(pm2.5~ .-year, data = data)
78 summary(m1)
79
80
81 #Backward method using AIC as metric
82
83
84
85
77:1 (Top Level) : 

```

The screenshot shows the RStudio interface with the code for splitting the dataset into training and testing sets using a 0.8/0.2 ratio. It also includes comments for a backward selection process using t-test p-values and AIC. To the right, there is a histogram of the 'year' variable, which shows a distribution from 2010 to 2013.

## 2. LINEAR REGRESSION

### 2. LINEAR REGRESSION

#### 2.1 Full mode

The screenshot shows the RStudio interface with the following details:

- Project View:** Shows a project named "HWLR" containing a script file "R HWLR.R".
- Code Editor:** Displays the R script "R HWLR.R" which includes code for data splitting, fitting a full model, and printing its summary.
- Environment View:** Shows the global environment with objects like base, data, day, hour, month, predictor.var, response.var, select.data, train.data, and values.
- Console View:** Shows the output of the R script, including the summary of the full model. The output indicates a call to lm with formula pm2.5 ~ ., data = data. The residuals show a normal distribution. The coefficients table includes intercept, year, month, day, hour, month\*year, DEMO, TEMP, PRES, IWS, IS, and IR. The p-values for all variables except the intercept are significant at the 0.05 level. The R-squared value is 0.2461.

### 2. LINEAR REGRESSION

#### 2.2 Backward method using p-value in t-test as metric

The screenshot shows the RStudio interface with the following details:

- Project View:** Shows a project named "HWLR" containing a script file "R HWLR.R".
- Code Editor:** Displays the R script "R HWLR.R" which includes code for data splitting, fitting a full model, and then using the backward method to remove variables based on AIC until the model is significant.
- Environment View:** Shows the global environment with objects like base, data, day, hour, month, predictor.var, response.var, select.data, train.data, and values.
- Console View:** Shows the output of the R script, including the summary of the final model. The output indicates a call to lm with formula pm2.5 ~ ., data = data. The residuals show a normal distribution. The coefficients table is identical to the full model, with significant p-values for all variables except the intercept. The R-squared value is 0.2461.

```

File Edit Code View Plots Session Build Debug Profile Tools Help
Project RStudio Source Save Run Source Environment History Connections Tutorial
File Edit Code View Plots Session Build Debug Profile Tools Help
Project RStudio Source Save Run Source Environment History Connections Tutorial
57 #exclude nominal variables
58 data<-data[,-1:11]
59 Vtew(data)
60
61
62 #training data 0.8
63 data<-data[sample(nrow(data),]
64 select.data = sample (1:nrow(data), 0.8*nrow(data))
65 train.data = data[select.data,]
66 test.data = data[-select.data,]
67
68
69
70
71 #backward method using p-value in t-test as metric
72 m1=lm(pm2.5~., data = data)
73 summary(m1)
74
75 m1=lm(pm2.5~.-year, data = data)
76 summary(m1)
77 m1=lm(pm2.5~.-year, data = data)
78 summary(m1)
79
80
81 #backward method using AIC as metric
82
83
84
85
86
87
88
89
89 <-- (Top Level) >

```

R Script:

```

R 4.3.1 - C:\Users\SRUJA\OneDrive\Desktop\Pgm in DA Project>
hour 1.568438 0.056006 27.708 < 2e-16 ***
DEWP -4.591209 0.067498 -85.156 < 2e-16 ***
TEMP -4.591209 0.067498 -85.156 < 2e-16 ***
PRES -1.419293 0.069433 -20.441 < 2e-16 ***
Iws -0.234381 0.007982 -28.362 < 2e-16 ***
IS -2.360e-01 7.957e-02 -29.66 < 2e-16 ***
IR -7.093752 0.267080 -26.523 < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 78.02 on 43813 degrees of freedom
Multiple R-squared: 0.2463, Adjusted R-squared: 0.2461
F-statistic: 1542 on 10 and 43813 DF, p-value: < 2.2e-16
> summary(m1)

Call:
lm(formula = pm2.5 ~ . - year, data = data)

Residuals:
    Min      1Q   Median      3Q     Max 
-169.32 -50.81 -15.38  32.01  878.60 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.594e+03 7.081e+01 22.51 < 2e-16 ***
month      -1.568e-01 4.242e-02 -35.81 < 2e-16 ***
day        -7.164e-01 4.242e-02 16.89 < 2e-16 ***
hour       1.565e-02 5.659e-02 27.65 < 2e-16 ***
DEWP      -4.590e-02 5.330e-02 -85.16 < 2e-16 ***
TEMP      -4.590e-02 5.330e-02 -85.16 < 2e-16 ***
PRES      -1.435e-00 6.941e-02 -20.38 < 2e-16 ***
Iws      -2.360e-01 7.957e-02 -29.66 < 2e-16 ***
IS      -2.360e-01 7.957e-02 -29.66 < 2e-16 ***
IR      -7.092e-00 2.671e-01 -26.55 < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 78.02 on 43814 degrees of freedom
Multiple R-squared: 0.2462, Adjusted R-squared: 0.246 
F-statistic: 1542 on 10 and 43814 DF, p-value: < 2.2e-16
>

```

## 2. LINEAR REGRESSION

### 2.3 Backward method using AIC as metric

```

File Edit Code View Plots Session Build Debug Profile Tools Help
Project RStudio Source Save Run Source Environment History Connections Tutorial
File Edit Code View Plots Session Build Debug Profile Tools Help
Project RStudio Source Save Run Source Environment History Connections Tutorial
62 #training data 0.8
63 data<-data[sample(nrow(data),]
64 select.data = sample (1:nrow(data), 0.8*nrow(data))
65 train.data = data[select.data,]
66 test.data = data[-select.data,]
67
68
69
70
71 #backward method using p-value in t-test as metric
72 m1=lm(pm2.5~., data = data)
73 summary(m1)
74
75 m1=lm(pm2.5~.-year, data = data)
76 summary(m1)
77
78
79
80 #backward method using AIC as metric
81
82 full=lm(pm2.5~., data = data)
83 full=full[order(full$AIC, direction="backward"), trace=T]
84 summary(full)
85
86
87
88
89
89 <-- (Top Level) >

```

R Script:

```

R 4.3.1 - C:\Users\SRUJA\OneDrive\Desktop\Pgm in DA Project>
pm2.5 ~ year + month + day + hour + DEWP + TEMP +
PRES + Iws + IS + IR, data = data)

Residuals:
    Min      1Q   Median      3Q     Max 
-168.35 -50.76 -15.37  32.05  878.60 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 295.789946 536.922196  0.551  0.5817
year        -1.503224 0.114330 -13.147 < 2e-16 ***
month       -1.503224 0.114330 -13.147 < 2e-16 ***
day         -1.588438 0.056006 -27.708 < 2e-16 ***
hour        1.588438 0.056006 27.708 < 2e-16 ***
DEWP        4.569134 0.053643 85.180 < 2e-16 ***
TEMP        4.569134 0.053643 85.180 < 2e-16 ***
PRES       -1.419293 0.069433 -20.441 < 2e-16 ***
Iws        -2.611571 0.495687 -5.269 1.38e-07 ***
IS          -2.611571 0.495687 -5.269 1.38e-07 ***
IR        -7.083752 0.267080 -26.523 < 2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 78.02 on 43813 degrees of freedom
Multiple R-squared: 0.2463, Adjusted R-squared: 0.2461
F-statistic: 1542 on 10 and 43813 DF, p-value: < 2.2e-16
>

```

## 2. LINEAR REGRESSION

#### 2.4 Forward method using AIC as metric

The screenshot shows the RStudio interface with two panes. The left pane displays an R script for a linear regression model, and the right pane shows the resulting console output.

**Script Content:**

```
#backward method using p-value in t-test as metric
m1=lm(pm2.5~., data = data)
summary(m1)
m1=lm(pm2.5~year, data = data)
summary(m1)

#backward method using AIC as metric
m2=fullModel, direction="backward", trace=T)
summary(m2)

#Forward method using AIC as metric
base=lm(pm2.5~1, data = data)
m3=stepAIC, scope=list(upper=full, lower=base), direction="forward",
trace=F)
summary(m3)

#selecting variables
select.data = int [1:35059] 04045 21303 15204 4190 34651...
```

**Console Output:**

```
R433 : C:\Users\SAU\OneDrive\Dataset\Rgm in DA Project<-->
> #Forward method using AIC as metric
> base=lm(pm2.5~1, data = data)
> m3=stepAIC, scope=list(upper=full, lower=base), direction="forward",
+   trace=F)
> summary(m3)

Call:
lm(formula = pm2.5 ~ Iws + TEMP + DEWP + Ir + hour + PRES + day +
month + Is + year, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-168.35 -50.70 -15.37  32.05  878.01 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 295.789946 536.922196  0.551  0.5813    
Iws         -0.334383  0.0077882 -29.262 < 2e-16 ***
TEMP        -0.000129  0.0001898 -92.040 < 2e-16 ***
DEWP        -4.569134  0.035641  85.180 < 2e-16 ***
Ir          -0.000104  0.0001040 -96.000 < 2e-16 ***
hour        1.568438  0.056606  27.708 < 2e-16 ***
PRES        -1.419293  0.069433 -20.441 < 2e-16 ***
day         -0.000104  0.0001040 -96.000 < 2e-16 ***
month       -1.503224  0.114336 -13.147 < 2e-16 ***
Is          0.047660  0.495516  -0.239 1.375e-01    
year        0.047660  0.495516  -0.239 1.375e-01    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 78.02 on 43813 degrees of freedom
Multiple R-squared:  0.2463   Adjusted R-squared:  0.2461 
F-statistic: 1432 on 10 and 43813 DF,  p-value: < 2.2e-16
```

## 2. LINEAR REGRESSION

## 2.5 Stepwise method using ACI as a metric

The screenshot shows the RStudio interface with the following details:

- Project**: Pgm in DA Project
- Code Editor**: Shows R code for a stepwise regression model. The code includes imports, data loading, and a stepwise function call.
- Environment**: Shows the global environment with variables like `m1`, `m2`, `m3`, `base`, `test.data`, and `train.data`.
- Console**: Displays the output of the R code, including the fitted model, coefficients, residuals, and a summary of the results.
- Plots**: No plots are visible in this screenshot.

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
  Go to function... Source on Save Run Source
Project Pgm in DA Project Data HW6.R
80 #backward method using AIC as metric
81
82 full<-lm(p2.5~., data = data)
83 m2<-step(full, direction="backward", trace=T)
84 summary(m2)
85
86 #forward method using AIC as metric
87 base<-lm(p2.5~1, data = data)
88 m3<-step(base, scope=list(upper=full, lower=base), direction="forward",
89 trace=F)
90 summary(m3)
91
92
93 #Stepwise method using AIC as metric
94
95 base<-lm(p2.5~1, data = data)
96 m5<-step(base, scope=list(upper=full, lower=base), direction="both",
97 trace=F)
98 summary(m5)
99
100 -
101
102 -
103
104
105
106
107
108
109
110
111 (Top Level) >
```

**Console** output:

```
R 4.3.1 -- "Curtains!" 2021-05-17 on 2021-05-17 running on Windows 10 x64
> library(tidyverse)
> data <- read_csv("Pgm in DA Project\\Data\\HW6.csv")
> m5<-step(base, scope=list(upper=full, lower=base), direction="both",
+ trace=F)
> summary(m5)

Call:
lm(formula = p2.5 ~ Iws + TEMP + DEWP + Ir + hour + PRES + day +
  month + year + Is + year, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-166.35 -50.76 -13.57  32.07  878.61 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 295.78994  336.322196   0.853  0.5817    
Iws          -6.29112   0.067988 -92.658 < 2e-16 ***
TEMP         -6.29112   0.067988 -92.658 < 2e-16 ***
DEWP         -6.29112   0.067988 -92.658 < 2e-16 ***
Ir           -7.083752  0.267080 -26.523 < 2e-16 ***
hour        -1.419293  0.069433 -20.444 < 2e-16 ***
PRES        -0.716324  0.042414 -16.889 < 2e-16 ***
day          -1.419293  0.069433 -20.444 < 2e-16 ***
month       -2.611571  0.495687 -5.269 1.8e-07 ***
year         0.647900  0.285315  2.243  0.0265    
Is           1.185450  0.495687  2.381  0.0175 *  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Individual t-and adjusted R-squared on last step of Tukey's
multiple R-squared:  0.24463;  Adjusted R-squared:  0.2461
F-statistic:  1432 on 10 and 43813 DF,  p-value: < 2.2e-16
```

### 3. POST PROCESSING

### 3.1 BEST MODEL

## 1. Full Model- RMSE :77.36838

2. Backward method using p-value in t-test as metric- RMSE :77.41664

RStudio

Edit Code View Plots Session Build Debug Profile Tools Help

Go to function ▾ Addins ▾

Project R HW8.R

data HW8.R

Source

Find Next Prev All Replace Replace All

Source

File Source Save As

rmse

checkboxes: Match case Whole word Page Wrap

179

180 m1=lm(pm2.5~., data = train.data)

181 summary(m1)

182

183 m1=lm(pm2.5~.-year, data = train.data)

184 summary(m1)

185

186 m1=~ lm(pm2.5 ~ . - year ~ No, data = train.data)

187 summary(m1)

188

189 m1=lm(pm2.5~.-year ~ No, na.action = na.omit,data=train.data);

190 ## Creating Backward method model using p-value & using testing data:

191 y\_obs= test[, "pm2.5"]

192 test[, "obs\_pm2.5"] = pm2.5

193 ##computing prediction error RMSE:

194 rmse=sqrt((y\_obs - y\_pred)^2/(y\_obs-y\_pred)) / nrow(test.data))

195 print(rmse)

196

197

198

199

200 #3.Backward method using AIC as metric

201

202 full=lm(pm2.5~.,na.action = na.omit, data = train.data)

203 m2=step(full, direction="backward", trace=T)

204

2001 [1] "Untitled" R Script

Console Terminal Background Jobs

R 4.3.1 C:\Users\SHRIJA\OneDrive\Desktop\Rgm\Da Project

Residual standard error: 78.17 on 35048 degrees of freedom

Multiple R-squared: 0.2468, Adjusted R-squared: 0.2466

F-statistic: 1149 on 10 and 35048 DF, p-value: < 2.2e-16

> m1 <- lm(pm2.5 ~ . - year ~ No, data = train.data)

> summary(m1)

Call:

lm(formula = pm2.5 ~ . - year ~ No, data = train.data)

Residuals:

Min 1Q Median 3Q Max

-170.14 -50.97 -15.47 31.83 855.25

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.603e+03	7.937e+01	20.178	< 2e-16 ***
month	-1.477e+00	1.281e+01	-11.534	< 2e-16 ***
day	7.465e-01	4.757e-01	15.694	< 2e-16 ***
hour	1.579e+00	6.343e-02	24.886	< 2e-16 ***
DEWP	-4.192e-01	1.281e-01	-3.258	0.0011 ***
TEMP	-6.293e+00	7.584e-02	-83.200	< 2e-16 ***
PRES	-1.422e+00	7.781e-02	-18.280	< 2e-16 ***
WIND	-2.425e-01	9.009e-03	-26.919	< 2e-16 ***
IS	-2.708e+00	5.532e-01	-4.895	9.88e-07 ***
IT	-7.059e+00	2.898e-01	-24.358	< 2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 78.17 on 35048 degrees of freedom

Multiple R-squared: 0.2467, Adjusted R-squared: 0.2466

F-statistic: 1276 on 9 and 35049 DF, p-value: < 2.2e-16

> m1=lm(pm2.5~.-year ~ No, na.action = na.omit,data=train.data);

> ## Creating Backward method model using p-value & using testing data:

> y\_pred=predict(lm(m1, test.data))

> y\_obs=test[, "pm2.5"]

> ##computing prediction error RMSE:

> rmse=sqrt((y\_obs - y\_pred)^2/(y\_obs-y\_pred)) / nrow(test.data))

> print(rmse)

[1] 77.41664

> |

Environment History Connections Tutorial

rmse num [1, 1] 77.4

test.data num 8765 obs. of 12 variables

\$ No : int 4238 23478 23305 32397 27996 633 37655 307 ...

\$ year : year 2010 2011 2012 2013 2010 2014 2013 2010 ...

\$ month : int 9 10 11 12 1 2 3 4 ...

\$ hour : int 26 3 29 11 5 28 18 3 18 ...

\$ day : int 13 5 25 10 20 15 4 22 24 ...

\$ pm2.5: num 150 25 154 90 182 27 74 30 53 39 ...

\$ DEWP : int 17 12 23 15 11 -15 3 18 13 16 ...

\$ TEMP : num 30 13 26 24 27 -2 14 32 20 7 ...

\$ PRES : num 1006 1017 1000 1013 1015 ...

\$ IS : num 8.05 3.58 6.26 33.54 6.26 ...

\$ LS : int 0 0 0 0 0 0 0 0 0 0 ...

\$ IR : int 0 0 0 0 0 0 0 0 0 0 ...

train.data 35059 obs. of 12 variables

\$ No : int 4551 25258 43316 4654 10789 12178 31398 36 ...

\$ year : year 2010 2012 2014 2010 2011 2013 2014 2010 ...

\$ month : int 7 11 12 7 3 5 8 2 3 12 ...

\$ day : int 9 18 10 13 26 23 1 25 21 25 ...

\$ hour : int 14 9 19 21 12 9 5 12 15 23 ...

\$ pm2.5: num 112 168 17 99 60 155 42 397 19 20 ...

Files Plots Packages Help Viewer Presentation

### 3. Backward method using AIC as metric: RMSE :77.36838

```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Project(R) data HWRLR
Find Source On Save Run Replace Source
In selection Match case Whole word Regex Wrap
195
196 #3. Backward method using AIC as metric
197 full=lm(pm2.5~.,na.action = na.omit, data = train.data)
198 m2=stepAIC(full,direction="backward",trace=T)
199 summary(m2)
200
201 ## creating backward method model using AIC as metric & using testing data:
202 y_pred=predict.glm(m2,test.data)
203 y_obs=test.data[,"pm2.5"]
204 rmse=sqrt((y_obs - y_pred)^2/(y_obs-y_pred)) /nrow(test.data)
205 print(rmse)
206
207 rmse=sqrt((y_obs - y_pred)^2/(y_obs-y_pred)) /nrow(test.data)
208 print(rmse)
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
738
739
739
740
741
742
743
744
745
746
747
748
749
749
750
751
752
753
754
755
756
757
758
759
759
760
761
762
763
764
765
766
767
768
769
769
770
771
772
773
774
775
776
777
778
779
779
780
781
782
783
784
785
786
787
788
789
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
809
810
811
812
813
814
815
816
817
818
819
819
820
821
822
823
824
825
826
827
828
829
829
830
831
832
833
834
835
836
837
838
839
839
840
841
842
843
844
845
846
847
848
849
849
850
851
852
853
854
855
856
857
858
859
859
860
861
862
863
864
865
866
867
868
869
869
870
871
872
873
874
875
876
877
878
879
879
880
881
882
883
884
885
886
887
888
889
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
909
910
911
912
913
914
915
916
917
918
919
919
920
921
922
923
924
925
926
927
928
929
929
930
931
932
933
934
935
936
937
938
939
939
940
941
942
943
944
945
946
947
948
949
949
950
951
952
953
954
955
956
957
958
959
959
960
961
962
963
964
965
966
967
968
969
969
970
971
972
973
974
975
976
977
978
979
979
980
981
982
983
984
985
986
987
988
989
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2369
2370
2371
2372
2373
23
```

5. Stepwise method using ACI as metric: RMSE :77.36838

### **3. POST PROCESSING**

---

### **3.2 MODEL DIAGNOSIS**

**Residual Analysis is considered for the following :**

### To Validate the constant variance

### To Validate the linearity relationship

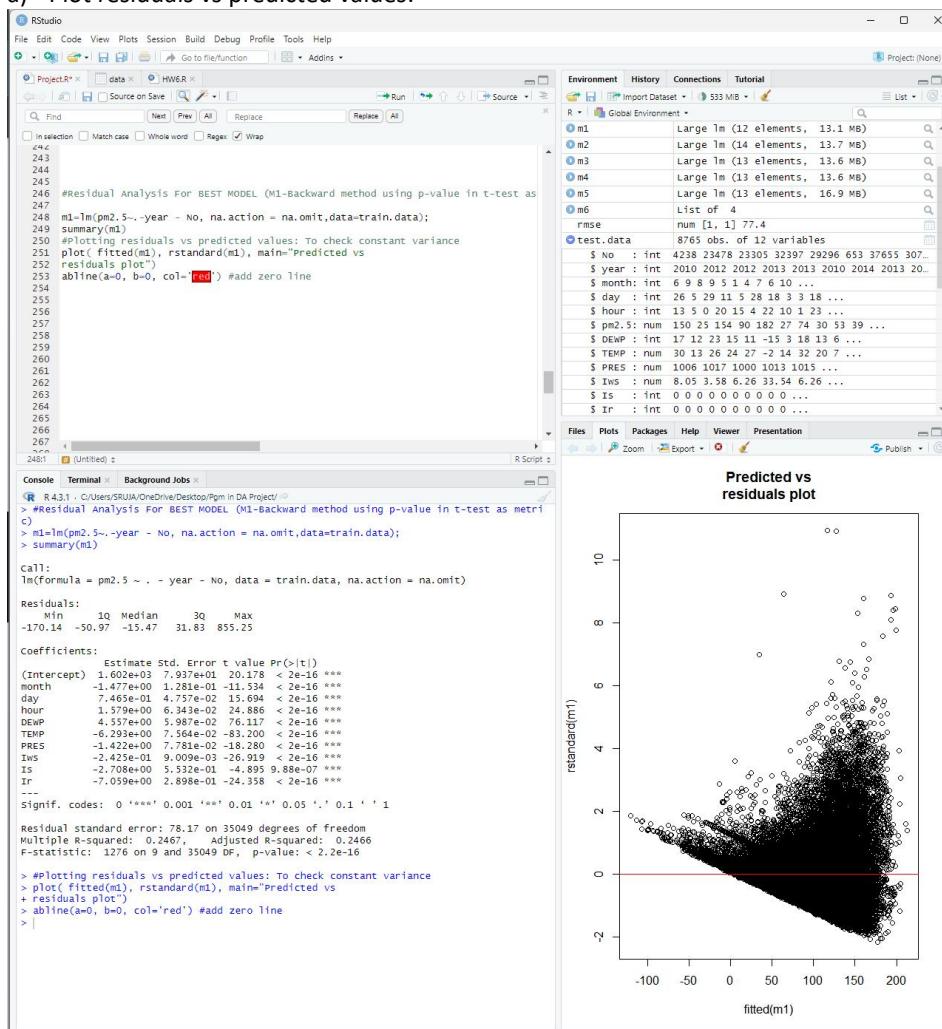
### To Validate normal distribution of residuals

### 3. POST PROCESSING

### 3.2 MODEL DIAGNOSIS

### Residual Analysis on the Best Model

a) Plot residuals vs predicted values:



## Transformation of PM2.5(Y variable)

The screenshot shows the RStudio interface with the following components:

- Top Bar:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project Panel:** Shows 'Project.R' and 'data' under 'Source on Save'.
- Code Editor:** Displays R code for transforming the PM2.5 variable. Lines 296-329 show the transformation process, and line 297 contains a comment '#Transformation of pm2.5'. Lines 330-331 show the creation of 'logdata' and 'datatrans' objects. Lines 332-333 show the removal of infinite values. Lines 334-335 show the creation of 'datatrans' and 'model1' objects. Line 336 shows the summary of 'model1'.
- Environment Tab:** Shows the global environment with various variables like 'logdata', 'year', 'month', 'day', 'hour', 'DEWP', 'TEMP', 'PRES', 'Iws', 'Is', 'Ir', 'dummy\_variables', 'full', 'm1', 'm2', 'm3', 'm4', 'm5', 'm6', and 'model1'.
- Console Tab:** Displays the R session history. It includes the R version (R 4.3.1), the working directory (C:/Users/SRUA/Desktop/Pgm in DA Project), and the execution of the transformation code. The output shows the 'Call:', 'Residuals:', 'Coefficients:', and 'Signif. codes:' sections of the lm() function results. It also provides the residual standard error, multiple R-squared, adjusted R-squared, F-statistic, and p-value.

```

296
297  |
298
299
300 #Transformation of pm2.5
301 logdata <- log(train.data$pm2.5)
302 datatrans <- cbind(logdata, train.data)
303 datatrans <- datatrans[, -c(2,7)]
304 inf.val <- is.infinite(datatrans$logdata)
305 datatrans <- datatrans[!inf.val, ]
306 model1 <- lm(logdata ~ ., data = datatrans)
307 summary(model1)
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329  |

297:1 (Untitled) : R Script

Console Terminal × Background Jobs ×
R 4.3.1 · C:/Users/SRUA/Desktop/Pgm in DA Project/ ·
> logdata <- log(train.data$pm2.5)
> datatrans <- cbind(logdata, train.data)
> datatrans <- datatrans[, -c(2,7)]
> inf.val <- is.infinite(datatrans$logdata)
> datatrans <- datatrans[!inf.val, ]
> model1 <- lm(logdata ~ ., data = datatrans)
> summary(model1)

Call:
lm(formula = logdata ~ ., data = datatrans)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5956 -0.5210  0.0570  0.5667  2.9628 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.513e+01 6.264e+00  8.801 < 2e-16 ***
year        -1.690e-02 3.099e-03 -5.454 4.94e-08 ***
month       -2.844e-02 1.333e-03 -21.335 < 2e-16 ***
day         6.675e-03 4.951e-04 13.481 < 2e-16 ***
hour        1.967e-02 6.605e-04 29.784 < 2e-16 ***
DEWP        6.104e-02 6.250e-04 97.661 < 2e-16 ***
TEMP        -7.152e-02 7.911e-04 -90.411 < 2e-16 ***
PRES        -1.595e-02 8.103e-04 -19.679 < 2e-16 ***
Iws          -4.080e-03 9.411e-05 -43.353 < 2e-16 ***
Is           -1.174e-02 5.759e-03 -2.040  0.0414 *  
Ir          -8.981e-02 3.017e-03 -29.766 < 2e-16 ***

Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

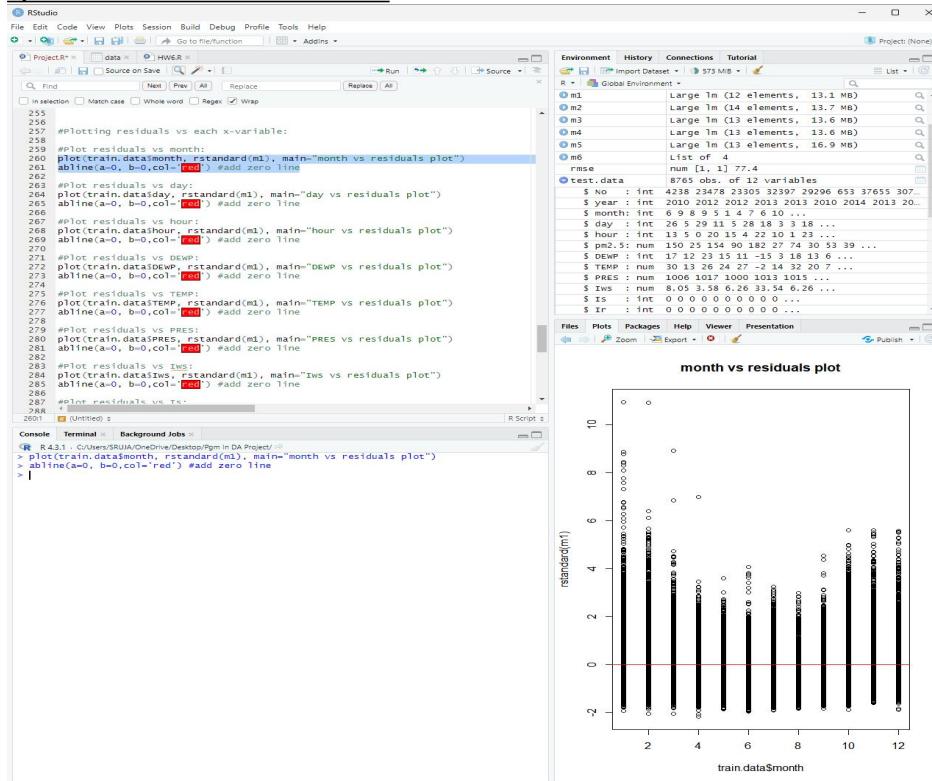
Residual standard error: 0.8137 on 35046 degrees of freedom
Multiple R-squared:  0.3547, Adjusted R-squared:  0.3545 
F-statistic: 1926 on 10 and 35046 DF,  p-value: < 2.2e-16

```

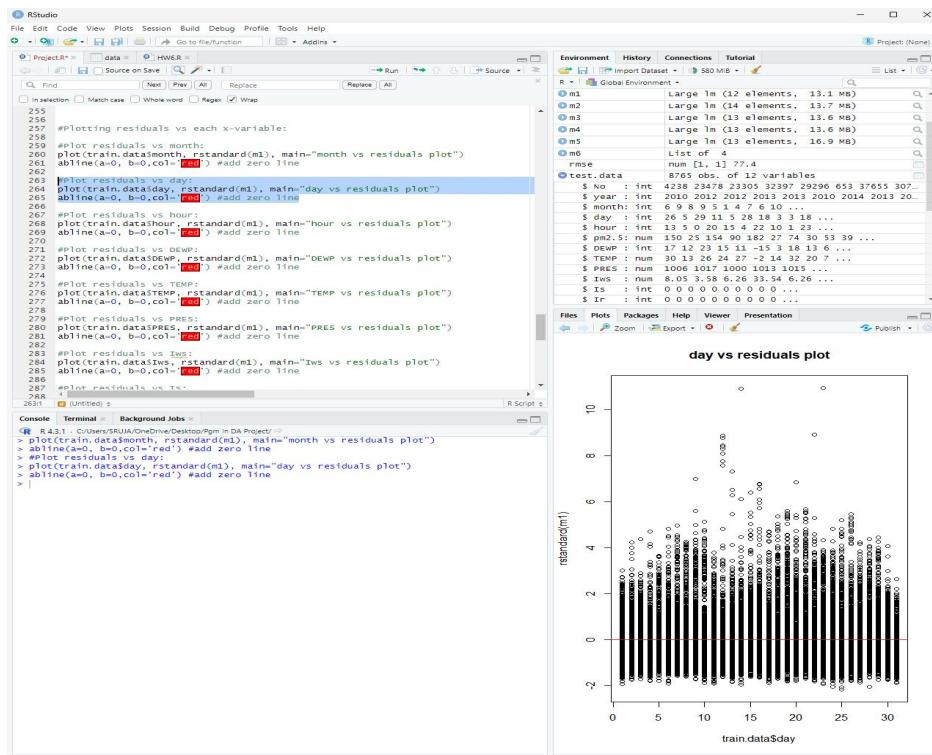
### 3.2 MODEL DIAGNOSIS

#### Residual Analysis on the Best Mode

##### b) Plot residuals vs each x-variable:

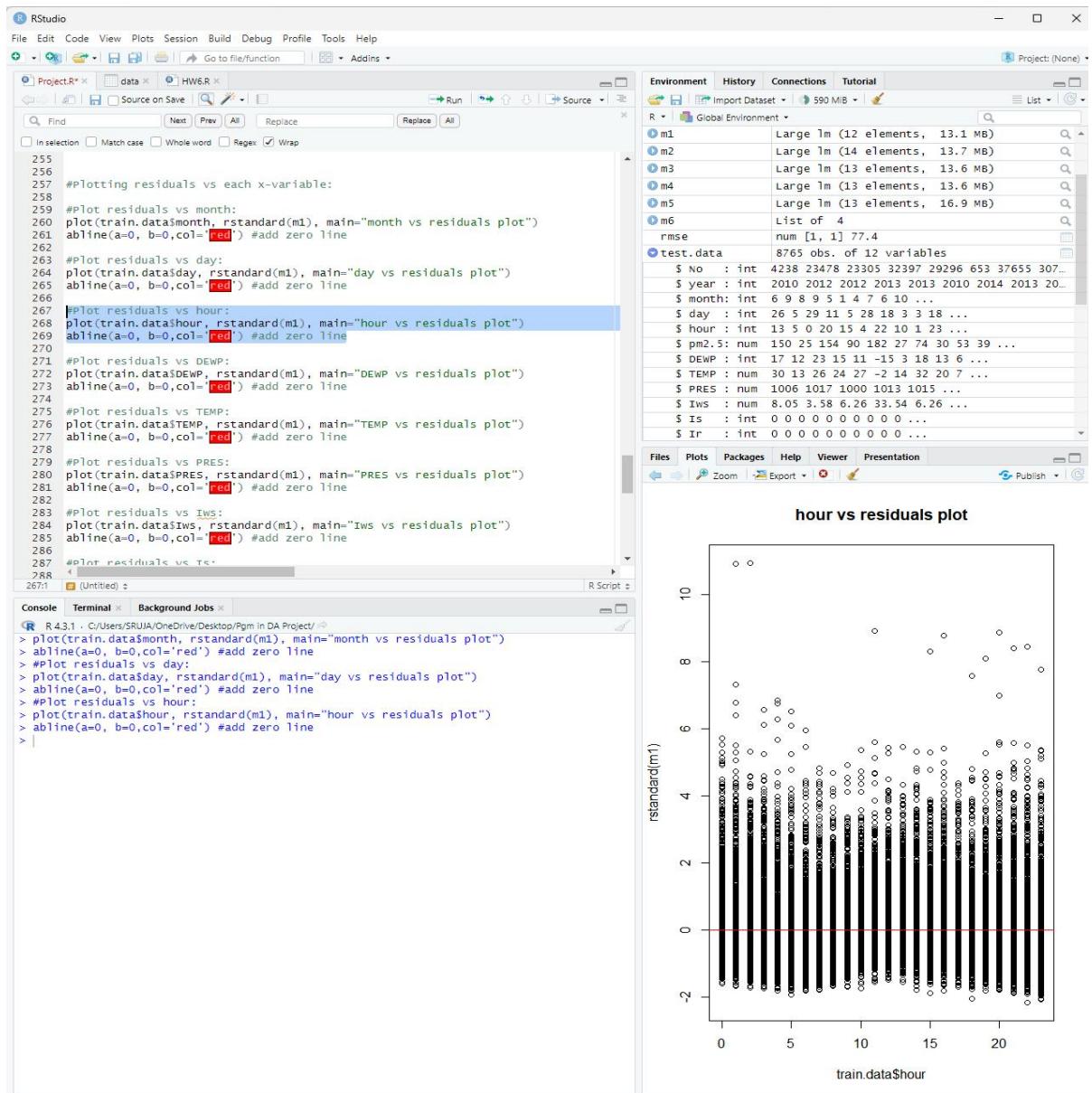


Residuals vs month

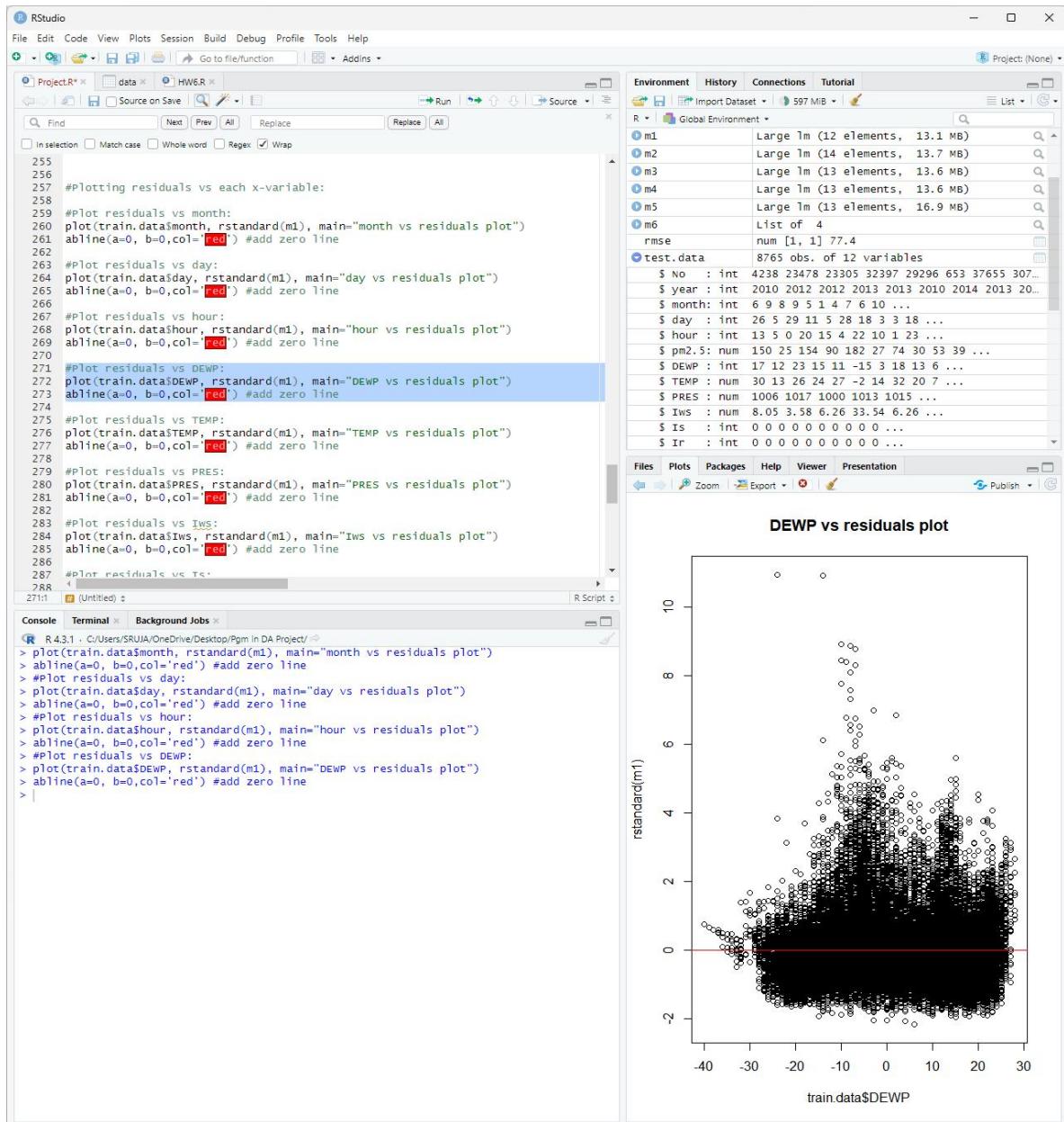


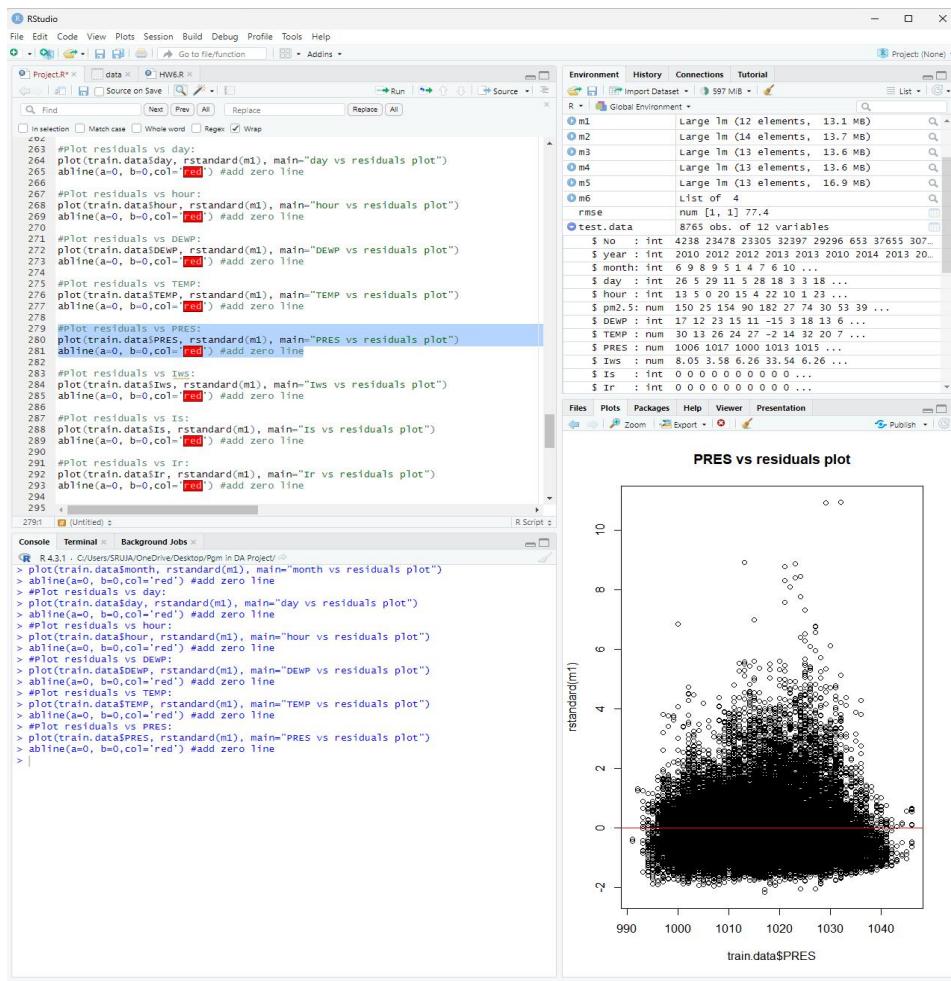
Day vs month

## Residuals vs Hour

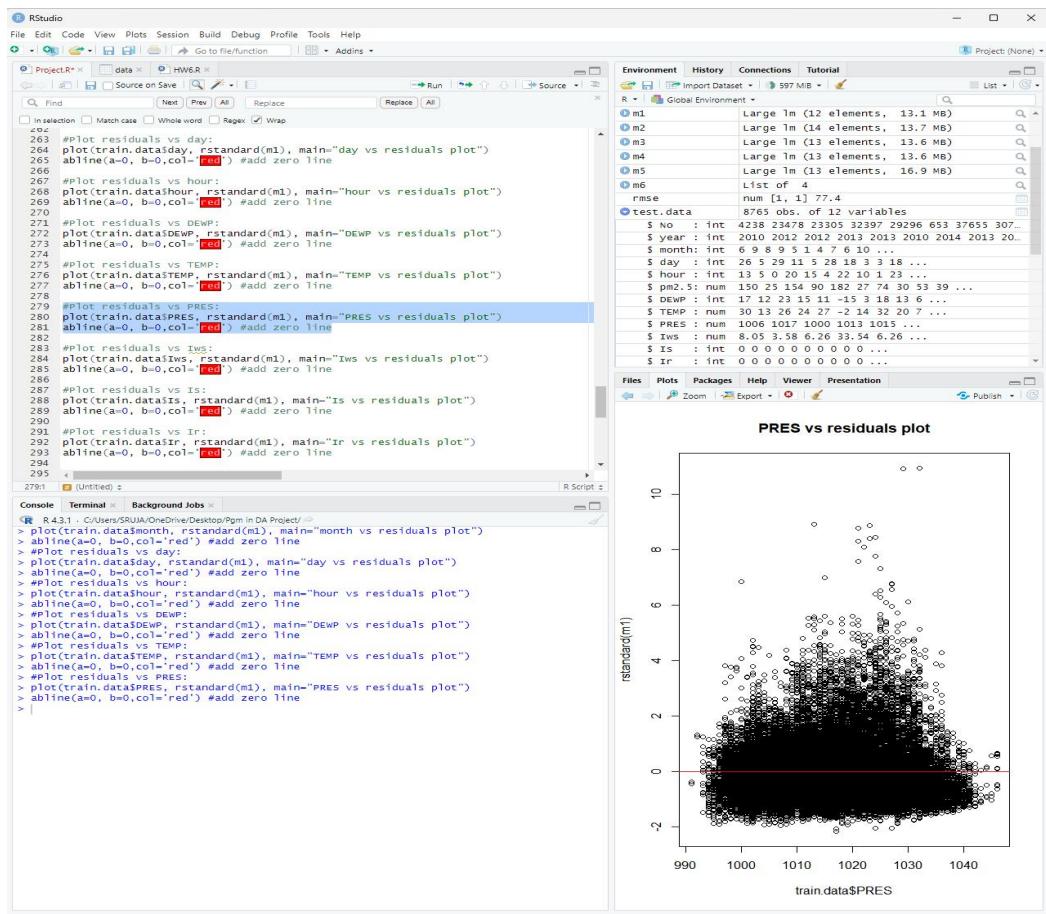


## Residuals vs DEWP

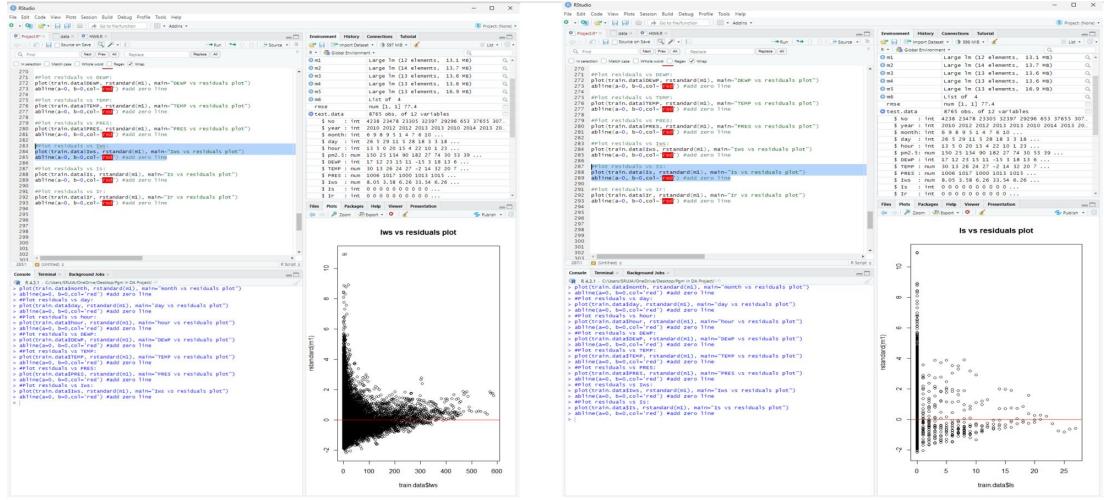




Residuals vs TEMP



## Residuals vs PRES

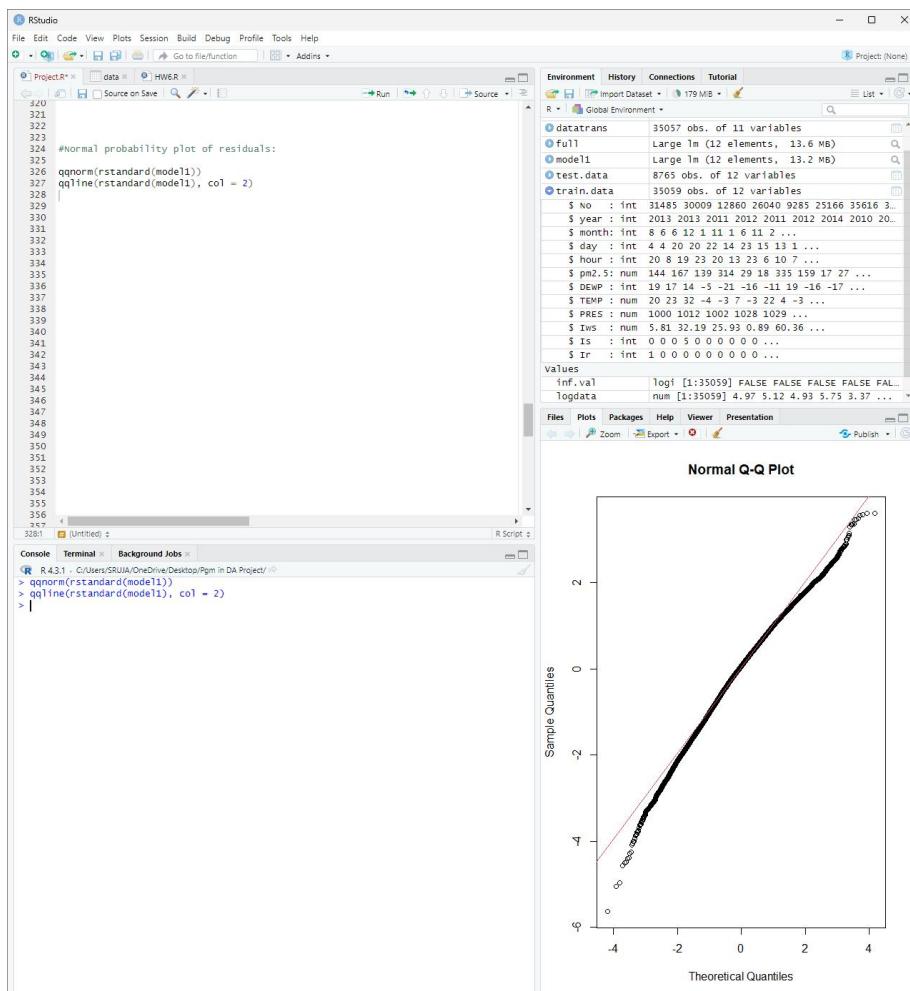


## Residuals vs Iws

## Residuals vs Is

### 3.2 MODEL DIAGNOSIS

c) Normal probability plot of residuals: To check normality assumption for the error terms



### **3.3 IMPROVING MODEL**

#### a) Multicollinearity problem

The screenshot shows the RStudio interface with the following details:

- Project View:** Shows a project named "HW6.R" containing a file "data.R".
- Code Editor:** Displays the following R code:

```
335
336
337
338 # Evaluate Collinearity
339 install.packages("car")
340 library(car)
341 vif(model1) # variance inflation factors
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
```

- Environment View:** Shows the global environment with the following objects and their descriptions:

  - m1: Large lm (12 elements, 13.1 MB)
  - model1: Large lm (12 elements, 13.2 MB)
  - test.data: 8765 obs. of 12 variables
  - train.data: 35059 obs. of 12 variables

Object	Description
\$ No	int 31485 30009 12860 26040 25166 35616 3...
\$ year	int 2013 2013 2011 2012 2011 2012 2014 2010 20...
\$ month	int 8 6 6 12 1 11 1 6 11 2 ...
\$ day	int 4 4 20 20 22 14 23 15 13 1 ...
\$ hour	int 20 8 19 23 20 13 23 6 10 7 ...
\$ pm2.5	num 144 167 139 314 29 18 335 159 17 27 ...
\$ DEWP	int 19 17 14 -5 -21 -16 -11 19 -16 -17 ...
\$ TEMP	num 20 23 32 -4 -3 7 -3 22 4 -3 ...
\$ PRES	num 1000 1012 1002 1028 1029 ...
\$ IWS	num 5.81 32.19 25.93 0.89 60.36 ...
\$ IS	int 0 0 0 5 0 0 0 0 0 0 0 ...
\$ IR	int 1 0 0 0 0 0 0 0 0 0 0 ...

- Console View:** Shows the R session output, including package installation and the execution of the provided R code.

The screenshot shows the RStudio interface with two main panes. The left pane contains an R script titled 'Untitled.R' with code related to collinearity analysis using the 'car' package. The right pane shows the 'Environment' tab with a list of objects and their details.

**R Script:**

```
# Evaluate Collinearity
install.packages("car")
library(car)
vif(model1) # variance inflation factors

#Since VIF>4

cor(datatrans)
```

**Environment:**

Object	Type	Description
data	43824 obs. of 12 variables	
datatrans	35057 obs. of 11 variables	
logdata	num 4.97 5.12 4.93 5.75 3.37 ...	
\$ year	: int 2013 2013 2011 2012 2011 2012 2011 ...	
\$ month	: int 8 6 12 11 1 11 2 ...	
\$ day	: int 4 4 20 20 22 14 23 15 13 1 ...	
\$ hour	: int 20 8 19 23 20 13 23 16 10 7 ...	
\$ DEWP	: int 19 17 14 -5 -21 -16 -11 19 -16 -1 ...	
\$ TEMP	: num 20 23 32 -4 -3 7 -3 22 4 -3 ...	
\$ PRES	: num 1000 1012 1002 1028 1029 ...	
\$ Iws	: num 5.81 32.19 25.93 0.89 60.36 ...	
\$ Is	: int 0 0 0 0 0 0 0 0 0 0 0 ...	
\$ Ir	: int 1 0 0 0 0 0 0 0 0 0 0 ...	
Full	Large lm (12 elements, 13.6 MB)	
m1	Large lm (12 elements, 13.1 MB)	
model1	Large lm (12 elements, 13.2 MB)	
multicoll	Large lm (12 elements, 12.7 MB)	
test.data	8765 obs. of 12 variables	
train.data	35059 obs. of 12 variables	

The screenshot shows the RStudio interface with the following components:

- File Edit Code View Plot Session Build Debug Profile Tools Help**: The top menu bar.
- Project**: A tree view of the project structure.
- Code Editor**: The main workspace where the R script is written. The code includes data loading, model fitting, and diagnostic plots.
- Data View**: A detailed view of the "airquality" dataset, showing variables like Ozone, Solar.R, Wind, Temp, and Pressure over time.
- Console**: The bottom pane showing the R command history and output, including summary statistics, coefficient tables, and diagnostic plots.

Key code snippets from the script:

```
# Load libraries
library(ggfortify)
library(dplyr)
library(ggplot2)
library(gridExtra)
library(kableExtra)
library(rms)
library(caret)
library(lmtest)
library(sandwich)
library(foreign)
library(tidyverse)
library(multicollinearity)

# Load data
airquality <- read.csv("C:/Users/.../airquality.csv")
airquality$Month <- as.factor(airquality$Month)
airquality$Temp <- as.numeric(airquality$Temp)
airquality$Wind <- as.numeric(airquality$Wind)
airquality$Ozone <- as.numeric(airquality$Ozone)

# Data visualization
ggplot(airquality, aes(x = Month, y = Ozone)) +
  geom_boxplot() +
  geom_jitter()

# Model fitting
model1 <- lm(Ozone ~ Solar.R + Wind + Temp + Month + Pressure, data = airquality)
summary(model1)
```

Output from the console:

```
Call:
lm(formula = Ozone ~ ., data = airquality[, -c(1)])
```

Residuals:

Min	Q1	Median	Q3	Max
-4.433	-0.820	0.182	0.455	4.367

Coefficients:

(Intercept)	estimate	std. error	t value	Pr(> t )
(Intercept)	7.339e+00	18.028	4.0e-16	***
Solar.R	-4.125e-01	1.050	-4.0e-15	
Wind	1.489e-01	1.495	1.0e-15	
Temp	5.402e-03	1.459e-04	3.7e-16	***
Month	1.388e-01	1.495	9.0e-15	***
Pressure	-2.859e-02	7.344e-04	3.9e-16	***

Signif. codes: 0 '\*\*\*\*' 0.001 '\*\*\*' 0.01 '\*\*' 0.05 '\*' 0.1 ' ' 1

Residual standard error: 0.921 on 3007 degrees of freedom

Multiple R-squared: 0.18, Adjusted R-squared: 0.1797

F-statistic: 83.4 on 9 and 3004 DF, p-value: < 2.2e-16

The screenshot shows the RStudio interface with a large R script in the top pane and its execution output in the bottom pane.

**Script Content:**

```
#> 342 library(caret)
#> 343 install.packages("car")
#> 344 library(car)
#> 345 lm(logdata~., data = datatrans[, -?]) # removing TEMP variable from the model
#> 346 summary(lm(logdata~.,
#> 347   data = datatrans[, -?]))
```

**Output Content:**

```
Call:
lm(formula = logdata ~ ., data = datatrans[, -?])

Residuals:
    Min      Q1      Median      Q3      Max 
-4.4337 -0.6942  0.0319  0.6244  3.5987 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 70.151037  6.902191 10.212 2e-16 ***
year        -0.034164  0.003426 -12.711 2e-16 ***
month       -0.000104  0.000104 -1.000 0.3140    
day         -0.000693  0.000358 11.657 2e-16 ***
hour        0.001766  0.000719  2.475 0.0334 *  
PRES        0.021449  0.000774  27.747 2e-16 ***
DWS        -0.000104  0.000104 -1.000 0.3140    
EIS        -0.042697  0.003137 -13.463 2e-16 *** 
IR          -0.030203  0.003488 -17.943 2e-16 *** 
...
signif. codes:  0 '****' 0.001 '**' 0.05 '*' 0.1 ' ' 1

residual standard error: 0.052 on 13047 degrees of freedom
Multiple R-squared:  0.2076, Adjusted R-squared:  0.2076 
F-statistic: 12021 on 9 and 13047 DF,  p-value: < 2.2e-16
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: HWR

Source on Save Run Source

343 #Since VIF>4

344

345 cor(datatrans)

346

347 multicollm<-lm(logdata~, data = datatrans[,-8])

348 summary(multicollm)

349 vif(multicollm)

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

1134

1135

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

1242

1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

1296

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316

1317

1318

1319

1320

1321

1322

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

1350

1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

1404

1405

1406

1407

1408

1409

1410

1411

1412

1413

1414

1415

1416

1417

1418

1419

1420

1421

1422

1423

1424

1425

1426

1427

1428

1429

1430

1431

1432

1433

1434

1435

1436

1437

1438

1439

1440

1441

1442

1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457

1458

1459

1460

1461

1462

1463

1464

1465

1466

1467

1468

1469

1470

1471

1472

1473

1474

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

1486

1487

1488

1489

1490

1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

1506

1507

1508

1509

1510

1511

1512

1513

1514

1515

1516

1517

1518

1519

1520

1521

1522

1523

1524

1525

1526

1527

1528

1529

1530

1531

1532

1533

1534

1535

1536

1537

1538

1539

1540

1541

1542

1543

1544

1545

1546

1547

1548

1549

1550

1551

1552

1553

1554

1555

1556

1557

1558

1559

1560

1561

1562

1563

1564

1565

1566

1567

1568

1569

1570

1571

1572

1573

1574

1575

1576

1577

1578

1579

1580

1581

1582

1583

1584

1585

1586

1587

1588

1589

1590

1591

1592

1593

1594

1595

1596

1597

1598

1599

1600

1601

1602

1603

1604

1605

1606

1607

1608

1609

1610

1611

1612

1613

1614

1615

1616

1617

1618

1619

1620

1621

1622

1623

1624

1625

1626

1627

1628

1629

1630

1631

1632

1633

1634

1635

1636

1637

1638

1639

1640

1641

1642

1643

1644

1645

1646

1647

1648

1649

1650

1651

1652

1653

1654

1655

1656

1657

1658

1659

1660

1661

1662

1663

1664

1665

1666

1667

1668

1669

1670

1671

1672

1673

1674

1675

1676

1677

1678

1679

1680

1681

1682

1683

1684

1685

1686

1687

1688

1689

1690

1691

1692

1693

1694

1695

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

1723

1724

1725

1726

1727

1728

1729

1730

1731

1732

1733

1734

1735

1736

1737

1738

1739

1740

1741

1742

1743

1744

1745

1746

1747

1748

1749

1750

1751

1752

1753

1754

1755

1756

1757

1758

1759

1760

1761

1762

1763

1764

1765

1766

1767

1768

1769

1770

1771

1772

1773

1774

1775

1776

1777

1778

1779</p

### b) Influential points (removing)

Influential points are observations that significantly affect the fitted model; these are usually outliers. In the event that they are eliminated, the parameter estimations change.

The screenshot shows the RStudio interface with the following details:

- File Bar:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project Bar:** Project R\*, data, HWLR.R, Source On Save, Run, Source.
- Code Editor:** Contains R code for identifying and removing influential points from a dataset named datatrans. The code includes:
  - Reading the dataset datatrans.
  - Calculating Cook's distance for each row.
  - Setting a threshold for influential points.
  - Identifying rows where Cook's distance exceeds the threshold.
  - Fitting a final model using the remaining data.
  - Summarizing the final model.
- Environment View:** Shows the global environment with objects like datatrans, finalmodel, full, influence.poin, m1, model1, multicolm, test.data, and train.data.
- Console View:** Displays the R code execution and the resulting regression output. The output includes:
  - Call: lm(formula = data.test\$logdata ~ ., data = data.test)
  - Residuals:

	Min	1Q	Median	3Q	Max
	-2.83131	-0.49747	0.04497	0.54267	2.20805

  - Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	49.1225321	5.3085474	8.314	< 2e-16 ***
year	-0.0153267	0.0029213	-5.247	1.56e-07 ***
month	-0.0298671	0.0012648	-23.614	< 2e-16 ***
day	0.0064473	0.0004547	13.763	< 2e-16 ***
hour	0.000162840	0.000162840	1.000	0.317
DWP	0.0648532	0.0005999	108.103	< 2e-16 ***
TEMP	-0.0736663	0.0007527	-97.873	< 2e-16 ***
PRES	-0.0131067	0.0007674	-17.073	< 2e-16 ***
IWS	-0.0187964	0.0009857	-19.526	< 2e-16 ***
IS	-0.0187964	0.0009857	-19.526	< 2e-16 ***
IR	-0.1031541	0.0043538	-23.693	< 2e-16 ***

  - Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1
  - Residual standard error: 0.7542 on 33822 degrees of freedom
  - Multiple R-squared: 0.4128, Adjusted R-squared: 0.4127
  - F-statistic: 2378 on 10 and 33822 DF, p-value: < 2.2e-16

## EVALUATION STRATEGIES

### 5-Crossfold evaluation

The screenshot shows the RStudio interface with the following components:

- Project View:** Shows a project named "HW6.R" containing a file "HW6.R".
- Code Editor:** Displays the R script "HW6.R" with code for 5-fold cross-validation using the caret package. The code includes setting a seed, loading data, training a model, and printing a summary.
- Environment View:** Shows the global environment with various objects listed, such as data.test, datatrans, finalmodel, full, influence.poin., m1, model, model1, multicollm, test.data, train.control, train.data, cooksdist, inf.val, logdata, select.data, threshold, and x.
- Console View:** Shows the R session output. It starts with the R version (R 4.3.1), setting a seed (set.seed(123)), and running the same script. The output includes:
  - 33833 samples, 10 predictor
  - No pre-processing
  - Resampling: Cross-Validated (5 fold)
  - Summary of sample sizes: 27067, 27067, 27065, 27067, 27066
  - Resampling results:

RMSE	R squared	MAE
0.754288	0.4125214	0.6085027

  - Tuning parameter 'intercept' was held constant at a value of TRUE