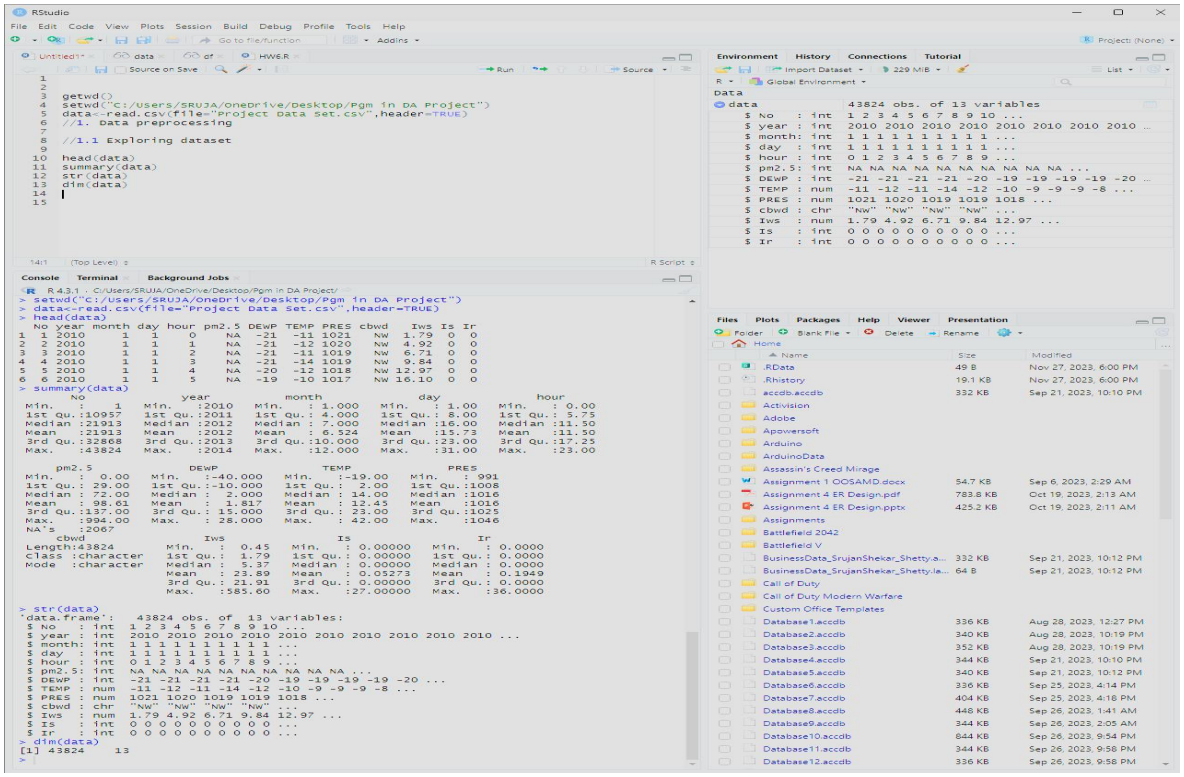
1.Data Preprocessing
1.1 Exploring Data-Set



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.

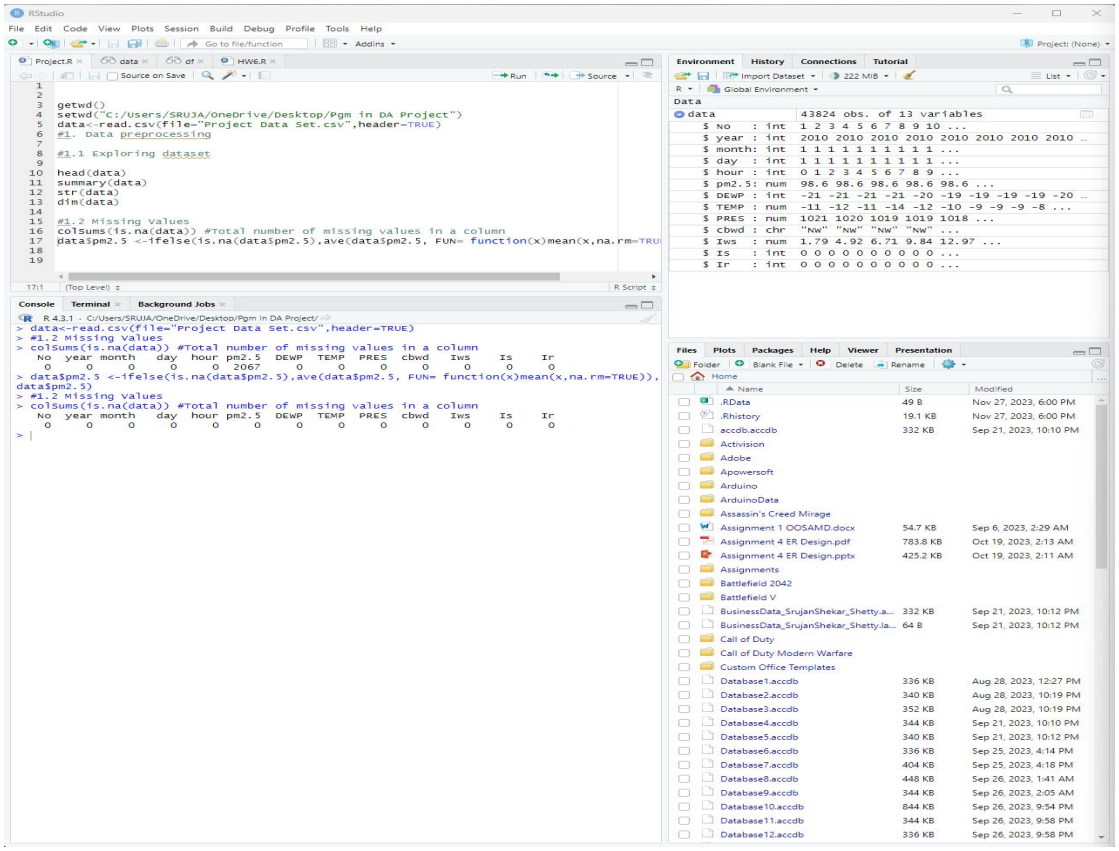
summary(data): This command offers a summary of the dataset's variables' respective statistical data. Depending on the kind of variable, it includes metrics like mean, median, minimum, maximum, and quartiles.

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.

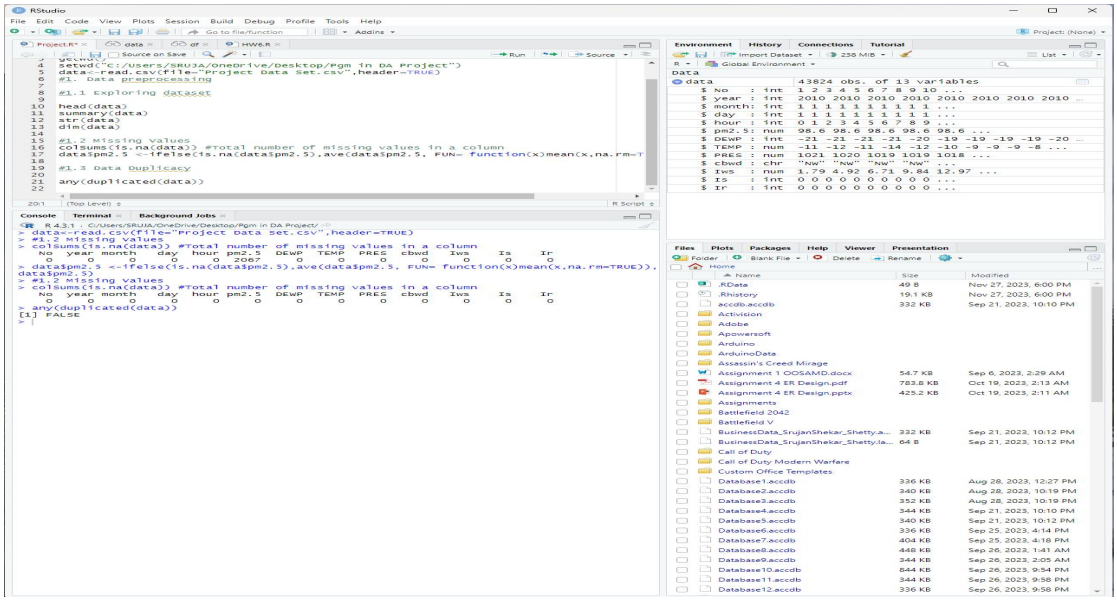
1.DATA-PREPROCESSING

1.2 Missing Values
Filled in them by using the mean.



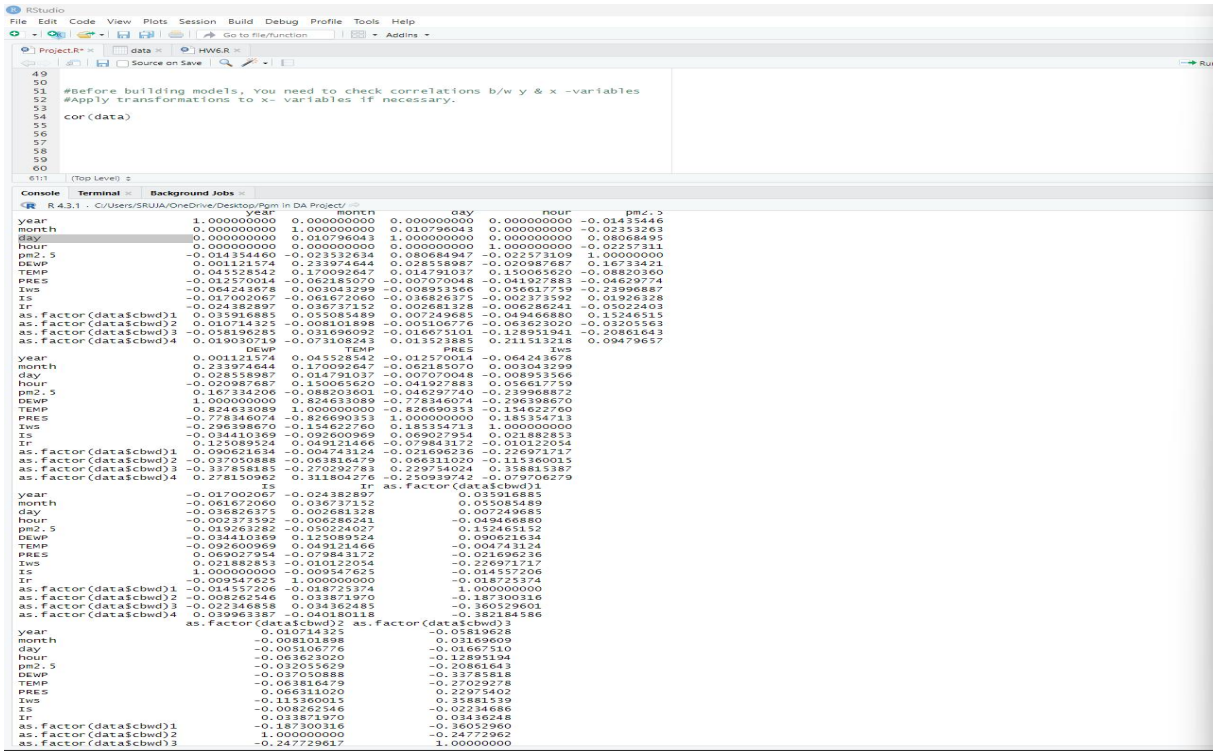
1.DATA-PREPROCESSING

1.3 Duplicate Rows Check



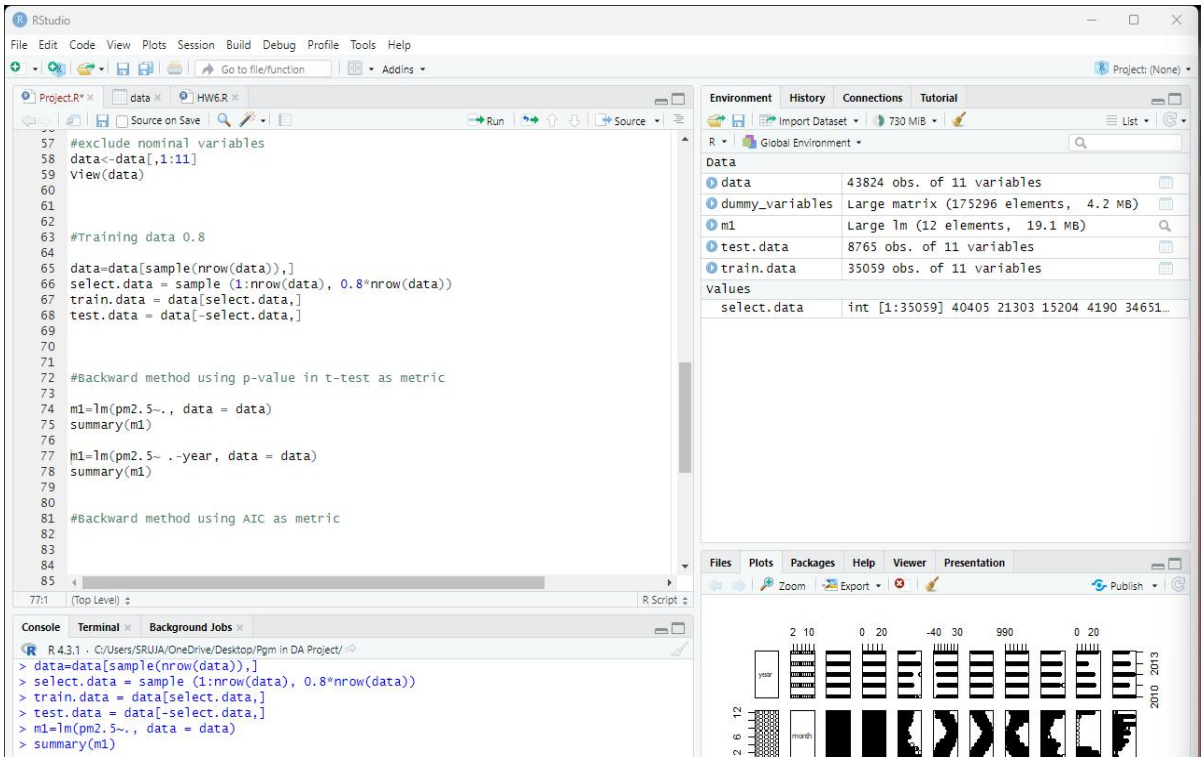
1.DATA-PREPROCESSING

1.4 Correlation of Dataset



1.DATA-PREPROCESSING

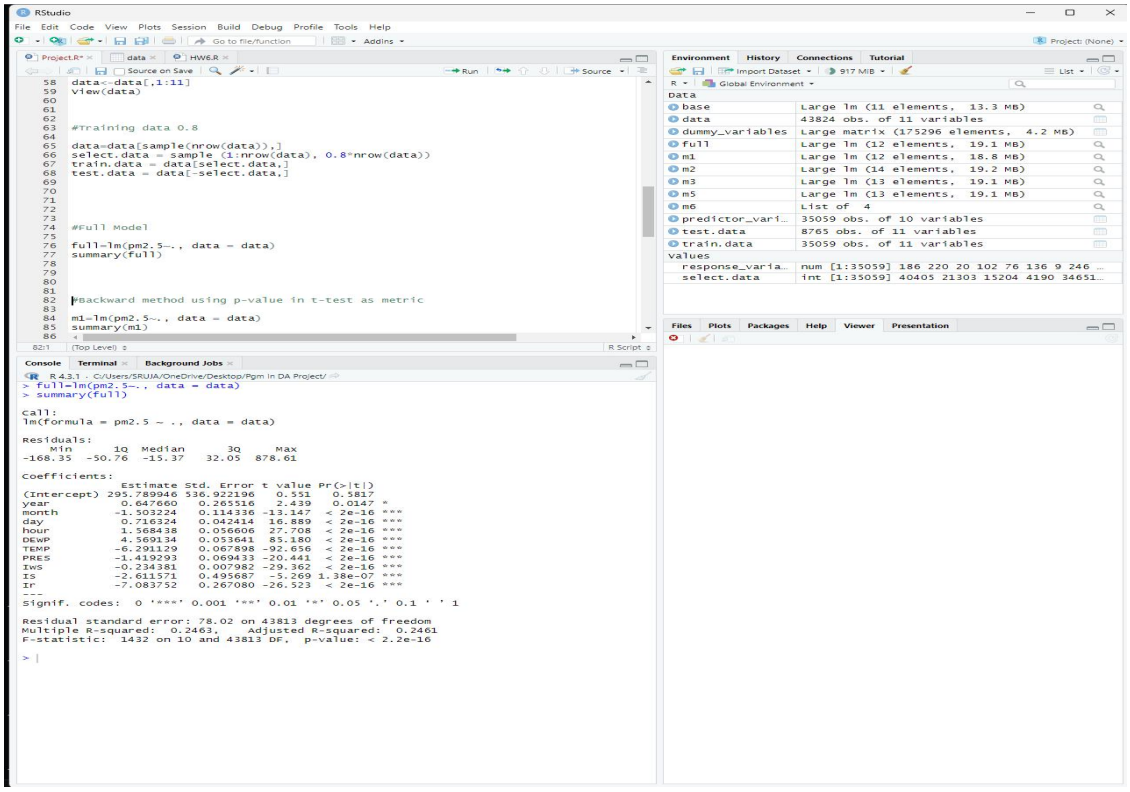
1.5Using hold-out evaluation only, 80% as training



2. LINEAR REGRESSION

2. LINEAR REGRESSION

2.1 Full mode

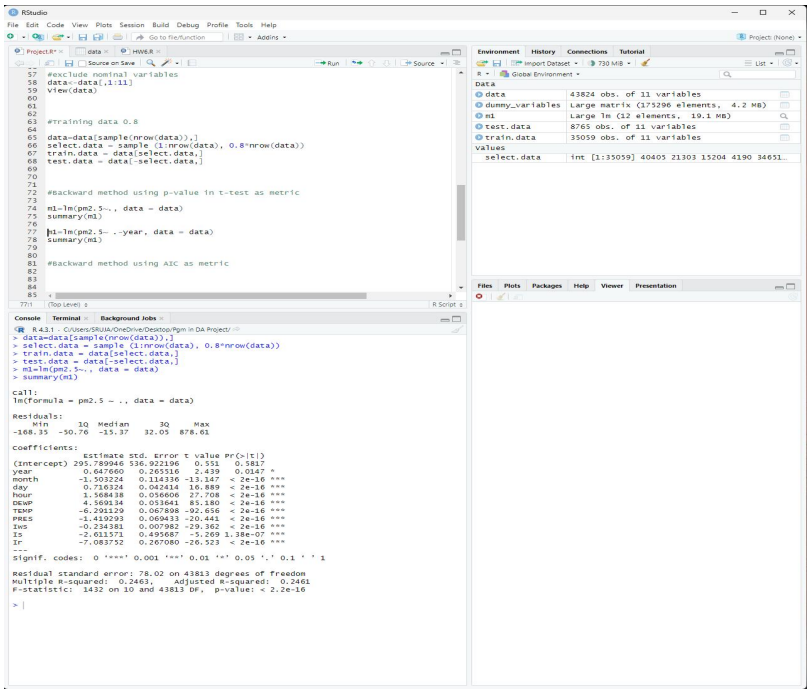


```
data<-data[,1:11]
View(data)


#Training data 0.8

data=data[sample(nrow(data)),]
select.data = sample (1:nrow(data), 0.8*nrow(data))
train.data = data[select.data,]
test.data = data[-select.data,]



#Full Model

full=lm(pm2.5~., data = data)
summary(full)


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

m1=lm(pm2.5~., data = data)
summary(m1)
```

```
> full=lm(pm2.5~., data = data)
> summary(full)

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

Residuals:
    Min      1Q  Median      3Q     Max
-168.35  -50.76  -15.37   32.05  878.61

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 295.789946 536.922196   0.551   0.5817
year          0.647660   0.265516   2.439   0.0147 *
month        -1.503224   0.114336 -13.147  < 2e-16 ***
day           0.716324   0.042414  16.889  < 2e-16 ***
hour          1.568438   0.056606  27.708  < 2e-16 ***
DEWP          4.569134   0.053641  85.180  < 2e-16 ***
TEMP         -6.291129   0.067898 -92.656  < 2e-16 ***
PRES         -1.419293   0.069433 -20.441  < 2e-16 ***
Iws          -0.234381   0.007982 -29.362  < 2e-16 ***
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:  1432 on 10 and 43813 DF,  p-value: < 2.2e-16

>
```

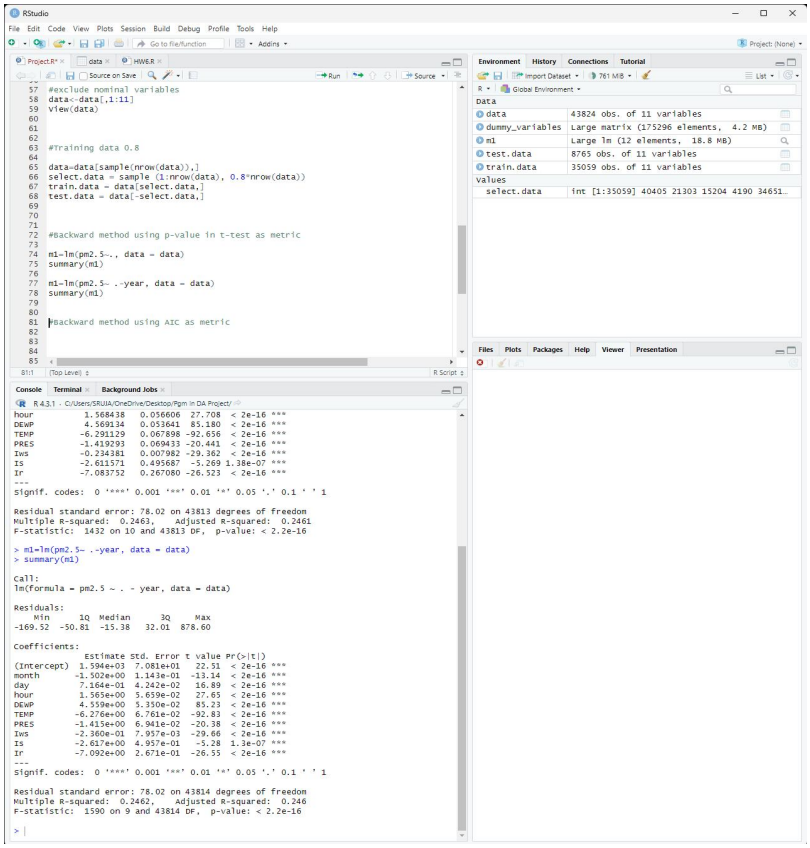
2. LINEAR REGRESSION

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



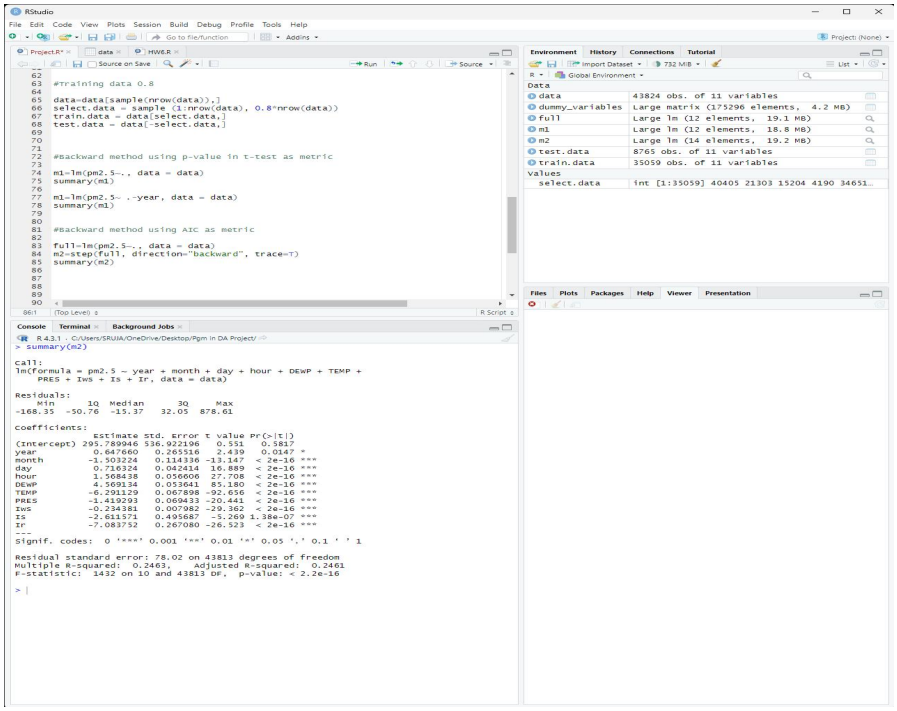
```
#exclude nominal variables
data= data[,1:11]
View(data)


#Training data 0.8

data=data[sample(nrow(data)),]
select.data = sample (1:nrow(data), 0.8*nrow(data))
train.data = data[select.data,]
test.data = data[-select.data,]


#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
```

```
> data=data[sample(nrow(data)),]
> select.data = sample (1:nrow(data), 0.8*nrow(data))
> train.data = data[select.data,]
> test.data = data[-select.data,]
> m1=lm(pm2.5~., data = data)
> summary(m1)

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

Residuals:
    Min      1Q  Median      3Q     Max
-168.35  -50.76  -15.37   32.05  878.61

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 295.789946 536.922196   0.551   0.5817
year          0.647660   0.265516   2.439   0.0147 *
month        -1.503224   0.114336 -13.147  < 2e-16 ***
day           0.716324   0.042414  16.889  < 2e-16 ***
hour          1.568438   0.056606  27.708  < 2e-16 ***
DEWP          4.569134   0.053641  85.180  < 2e-16 ***
TEMP         -6.291129   0.067898 -92.656  < 2e-16 ***
PRES         -1.419293   0.069433 -20.441  < 2e-16 ***
Iws          -0.234381   0.007982 -29.362  < 2e-16 ***
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:  1432 on 10 and 43813 DF,  p-value: < 2.2e-16

>
```



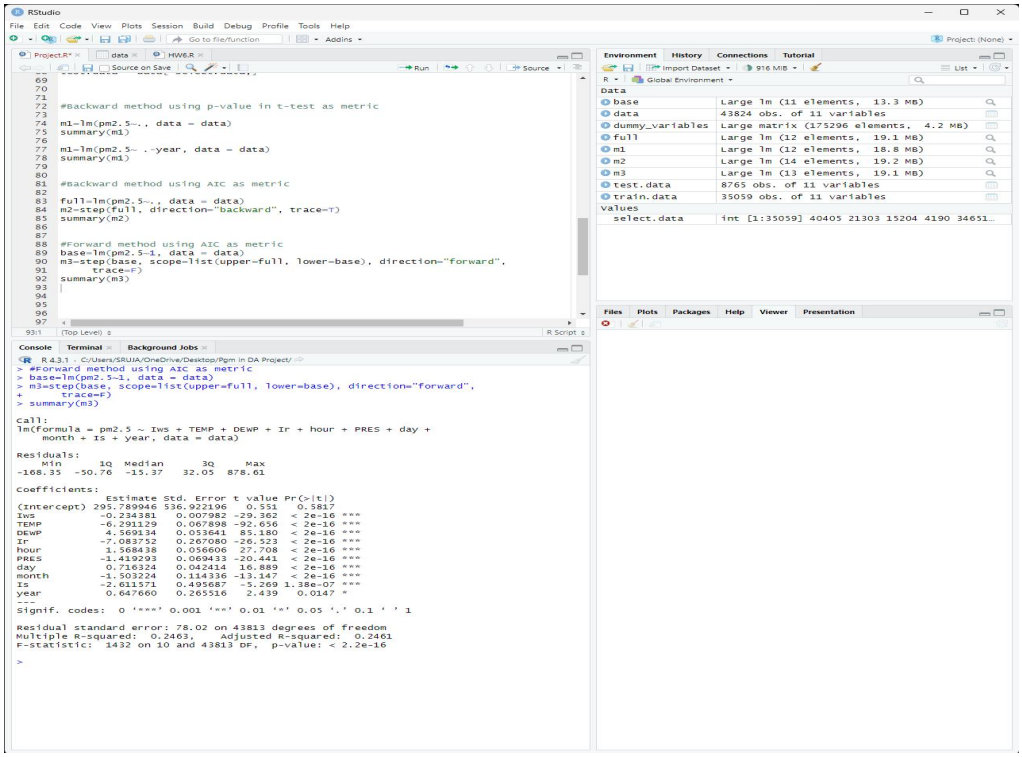
2. LINEAR REGRESSION

2.3 Backward method using AIC as metric



2. LINEAR REGRESSION

2.4 Forward method using AIC as metric

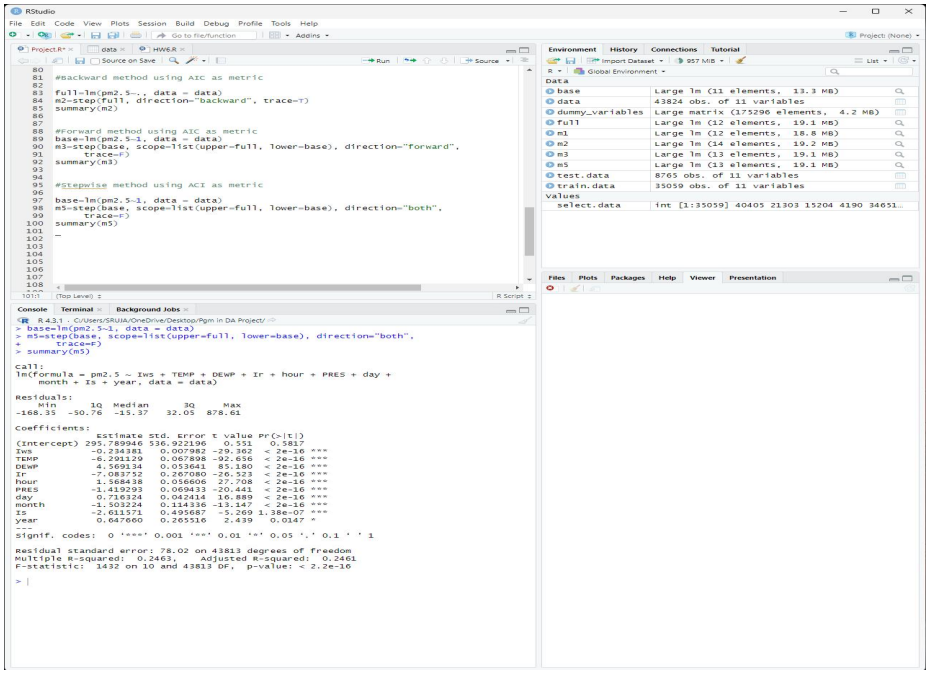


```
#Backward method using p-value in t-test as metric

m1=lm(pm2.5~., data = data)
summary(m1)

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


#Backward method using AIC as metric

full=lm(pm2.5~., data = data)
m2=step(full, direction="backward", trace=T)
summary(m2)


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

```
> #Forward method using AIC as metric
> base=lm(pm2.5~1, data = data)
> m3=step(base, 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  -50.76  -15.37   32.05  878.61

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 295.789946 536.922196   0.551   0.5817
Iws          -0.234381   0.007982 -29.362  < 2e-16 ***
TEMP         -6.291129   0.067898 -92.656  < 2e-16 ***
DEWP          4.569134   0.053641  85.180  < 2e-16 ***
Ir           -7.083752   0.267080 -26.523  < 2e-16 ***
hour          1.568438   0.056606  27.708  < 2e-16 ***
PRES         -1.419293   0.069433 -20.441  < 2e-16 ***
day           0.716324   0.042414  16.889  < 2e-16 ***
month        -1.503224   0.114336 -13.147  < 2e-16 ***
Is           -2.611571   0.495687  -5.269 1.38e-07 ***
year          0.647660   0.265516   2.439   0.0147 *
---
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



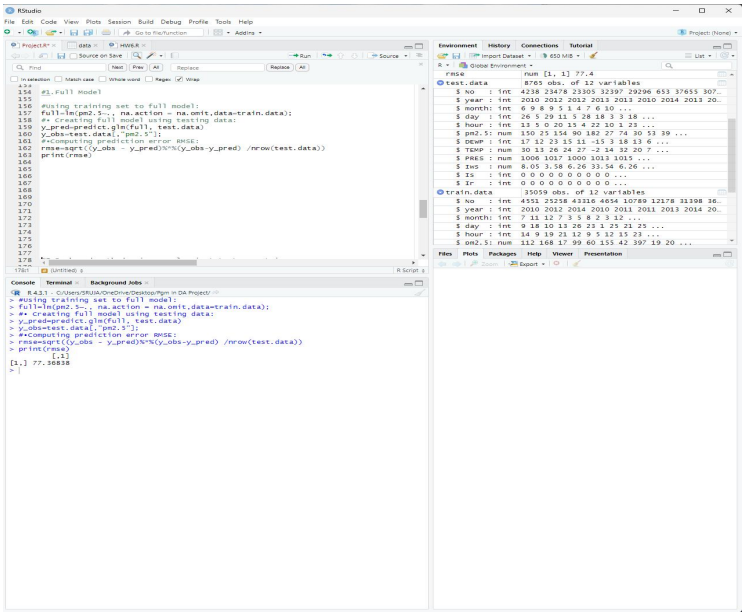
```
#Backward method using AIC as metric

full=lm(pm2.5~., data = data)
m2=step(full, direction="backward", trace=T)
summary(m2)

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


#Stepwise method using ACI as metric

base=lm(pm2.5~1, data = data)
m5=step(base, scope=list(upper=full, lower=base), direction="both",
        trace=F)
summary(m5)
```

```
> base=lm(pm2.5~1, data = data)
> m5=step(base, scope=list(upper=full, lower=base), direction="both",
+         trace=F)
> summary(m5)

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  -50.76  -15.37   32.05  878.61

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 295.789946 536.922196   0.551   0.5817
Iws          -0.234381   0.007982 -29.362  < 2e-16 ***
TEMP         -6.291129   0.067898 -92.656  < 2e-16 ***
DEWP          4.569134   0.053641  85.180  < 2e-16 ***
Ir           -7.083752   0.267080 -26.523  < 2e-16 ***
hour          1.568438   0.056606  27.708  < 2e-16 ***
PRES         -1.419293   0.069433 -20.441  < 2e-16 ***
day           0.716324   0.042414  16.889  < 2e-16 ***
month        -1.503224   0.114336 -13.147  < 2e-16 ***
Is           -2.611571   0.495687  -5.269 1.38e-07 ***
year          0.647660   0.265516   2.439   0.0147 *
---
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

>
```

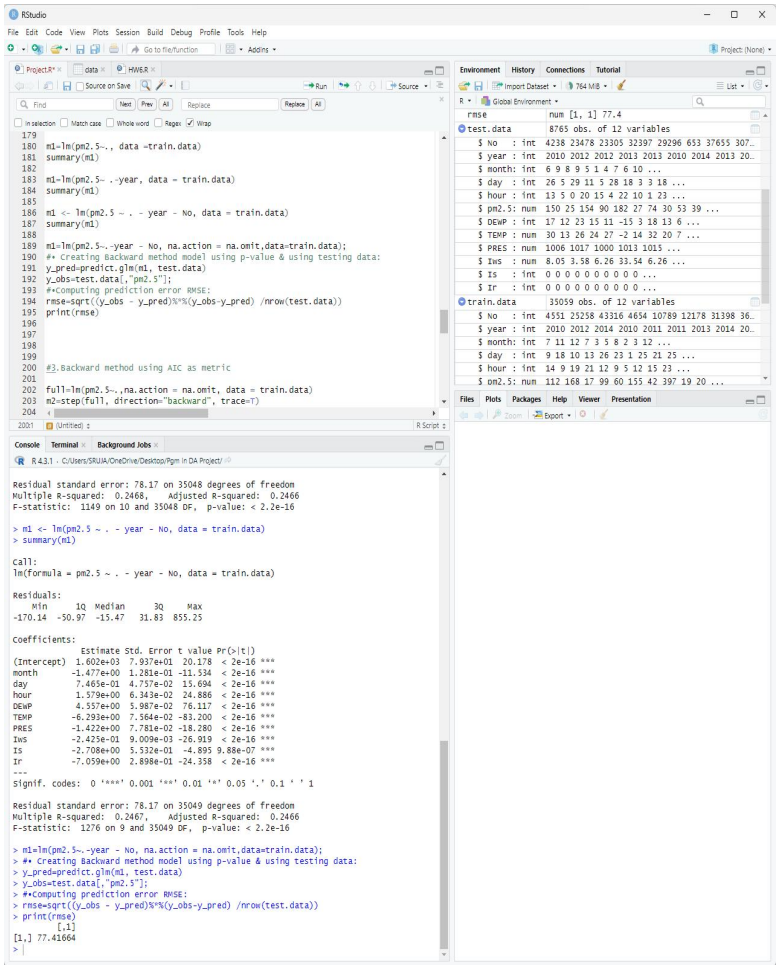
3. POST PROCESSING

3.1 BEST MODEL

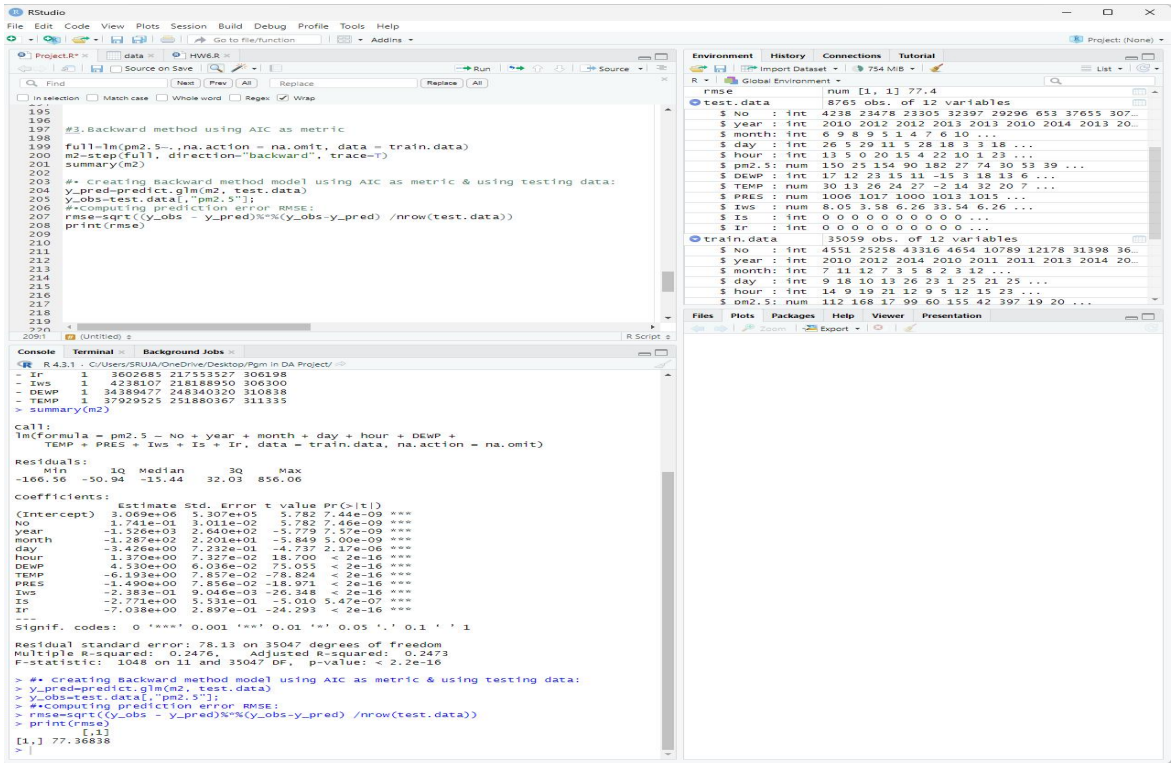
1.Full Model- RMSE :77.36838



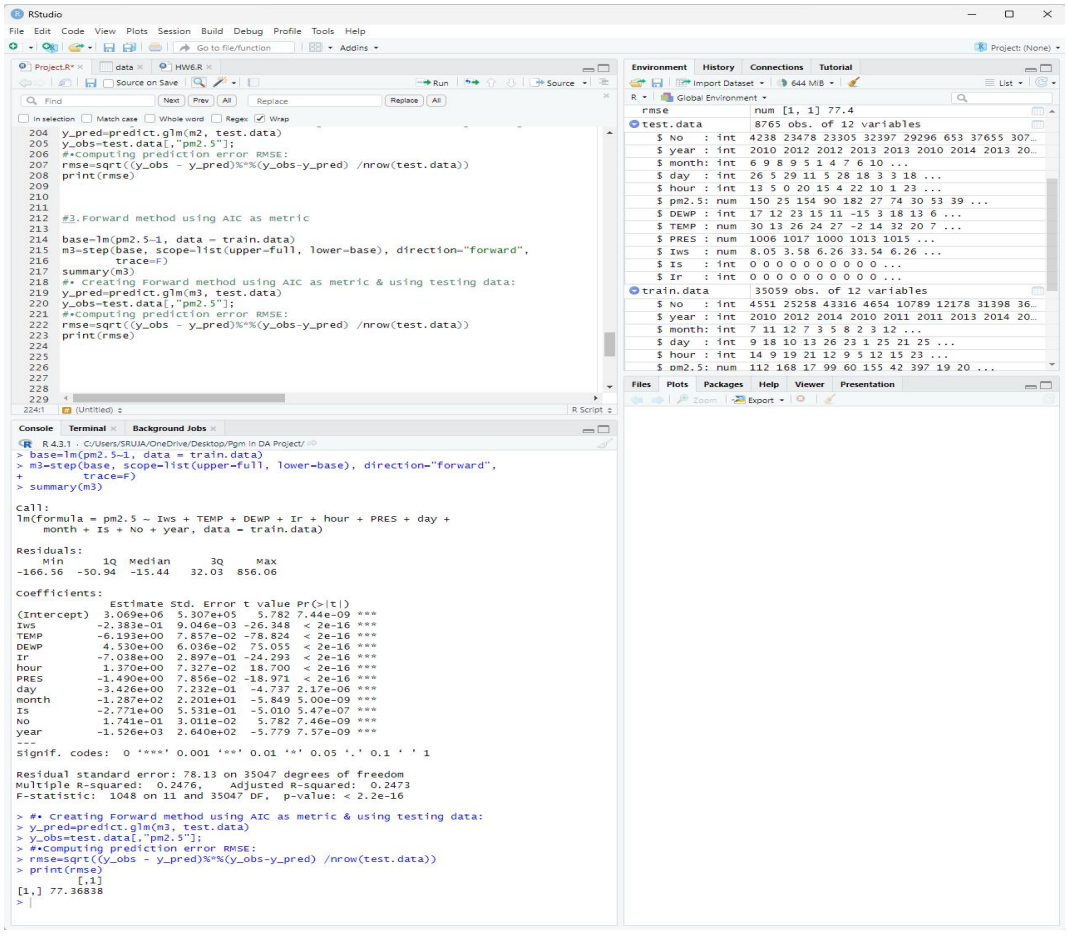
2. Backward method using p-value in t-test as metric- RMSE :77.41664



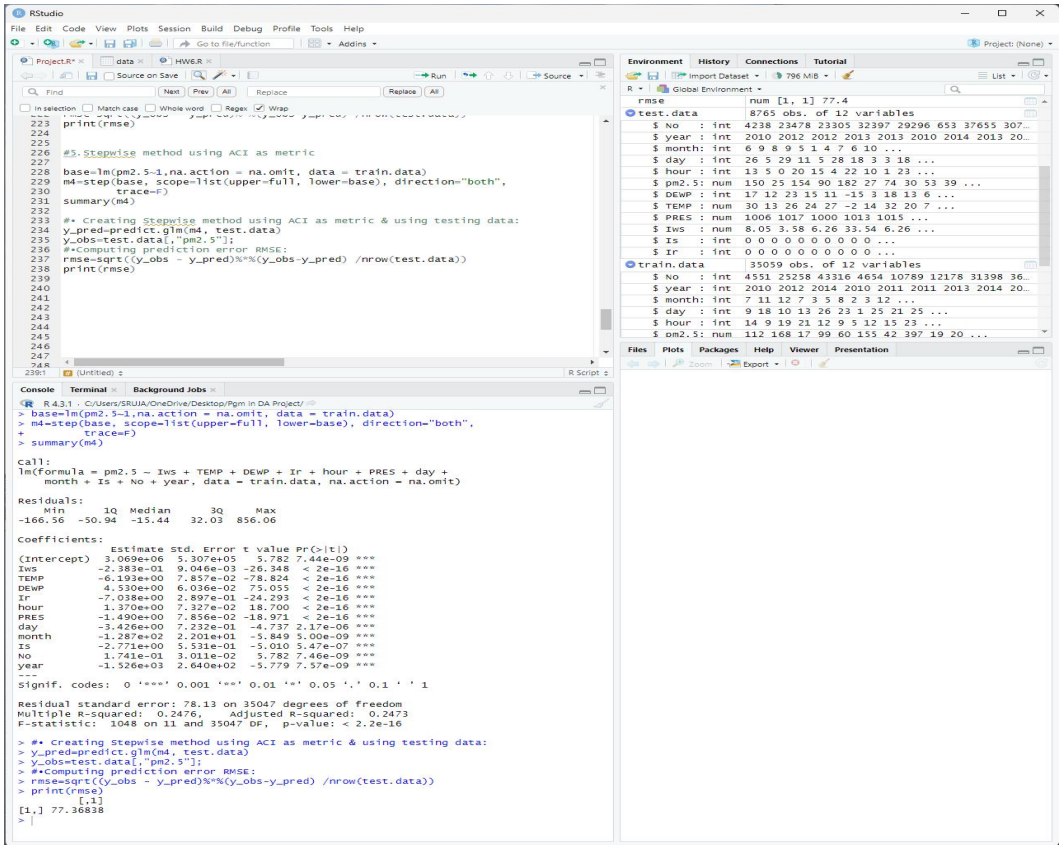
3. Backward method using AIC as metric: RMSE :77.36838



4.Forward method using AIC as metric: RMSE :77.36838



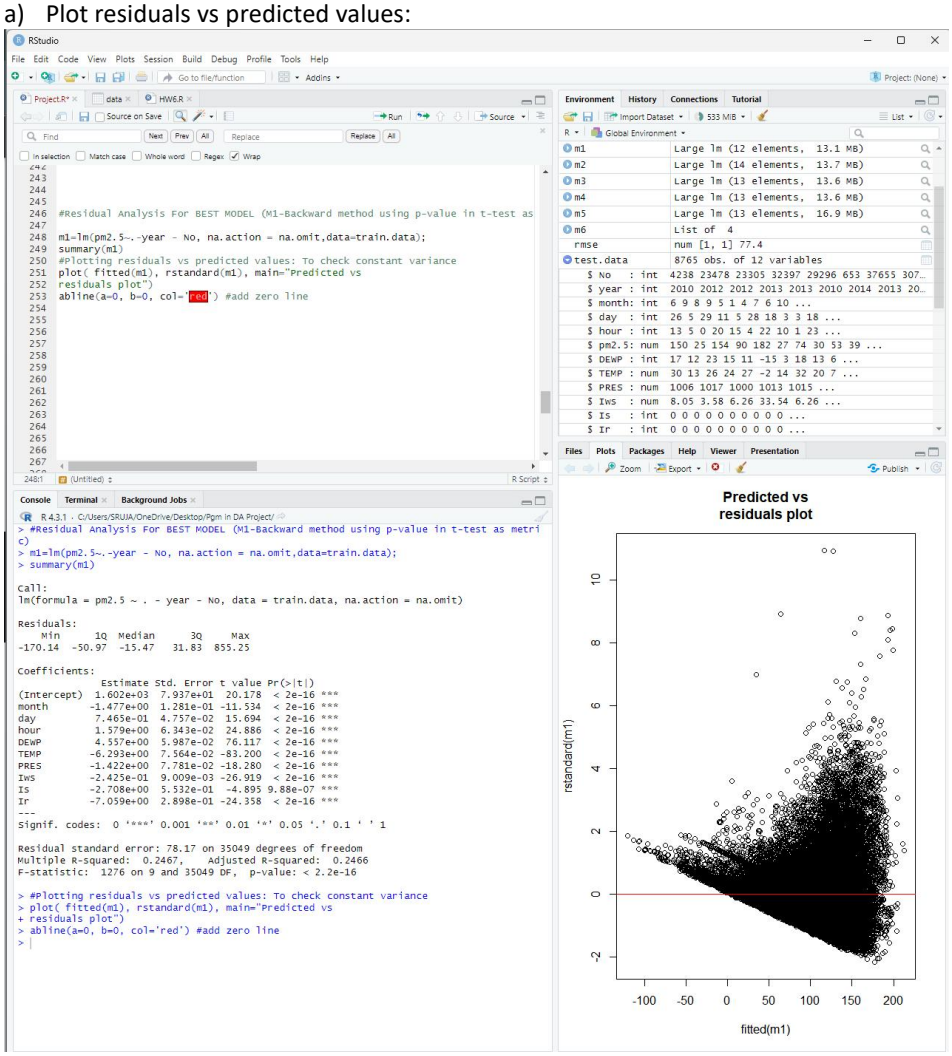
5.Stepwise method using ACI as metric: RMSE :77.36838



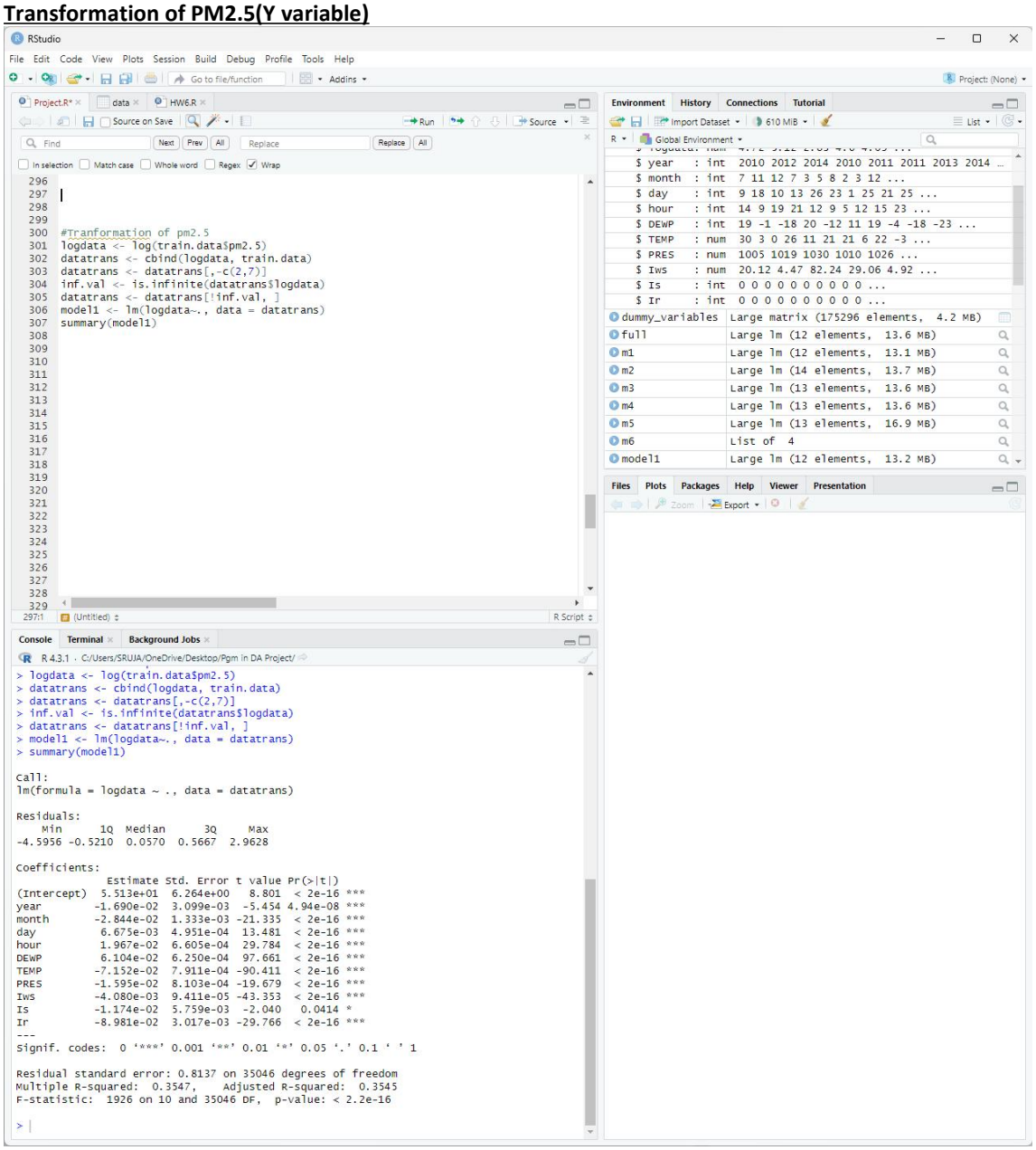
3. POST PROCESSING
3.2MODEL DIAGNOSIS

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
3.2MODEL DIAGNOSIS
Residual Analysis on the Best Model
a) Plot residuals vs predicted values:



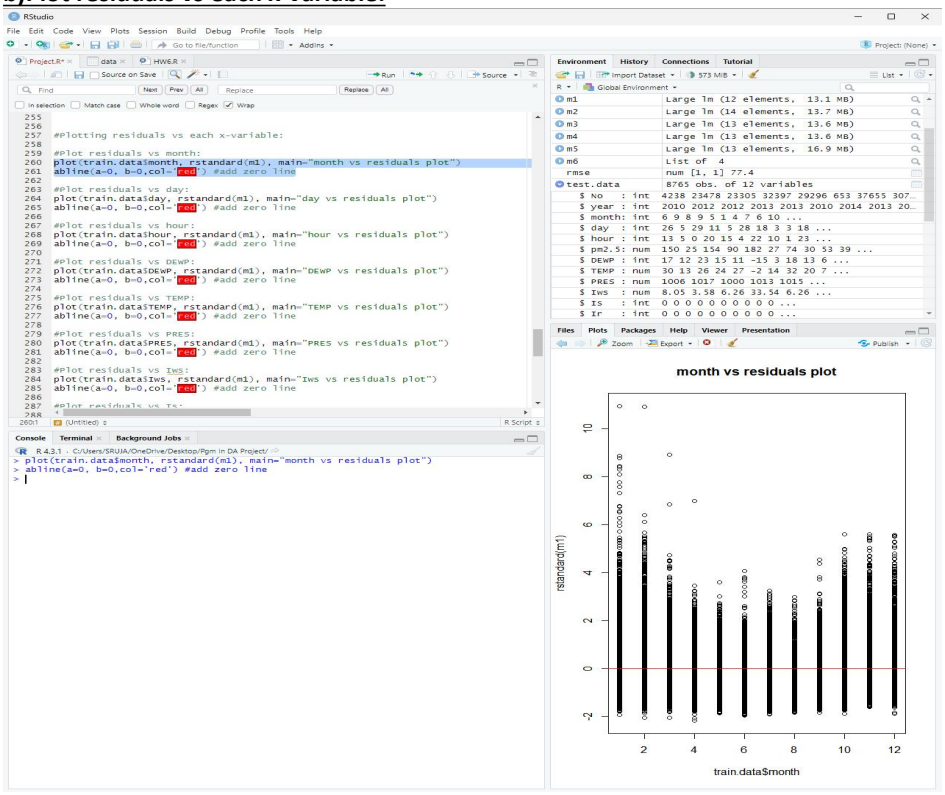
Transformation of PM2.5(Y variable)



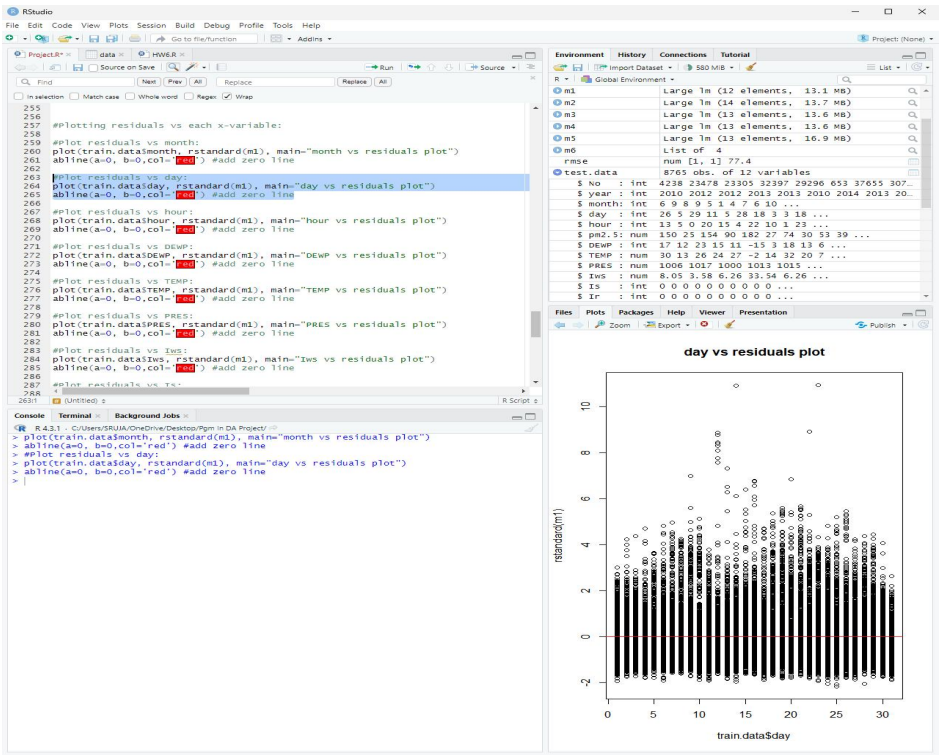
3.2MODEL DIAGNOSIS

Residual Analysis on the Best Mode

b)Plot residuals vs each x-variable:

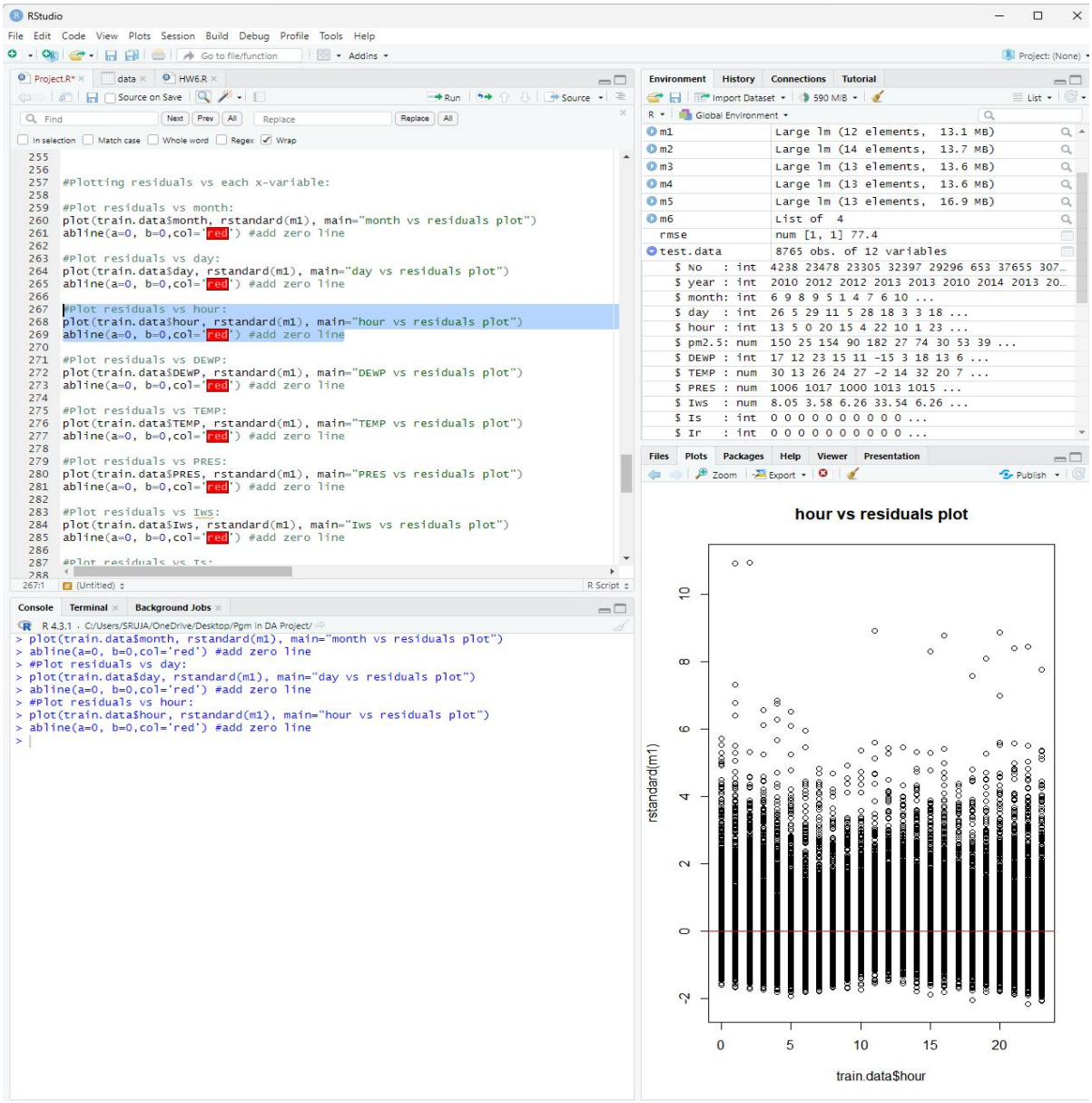


Residuals vs month

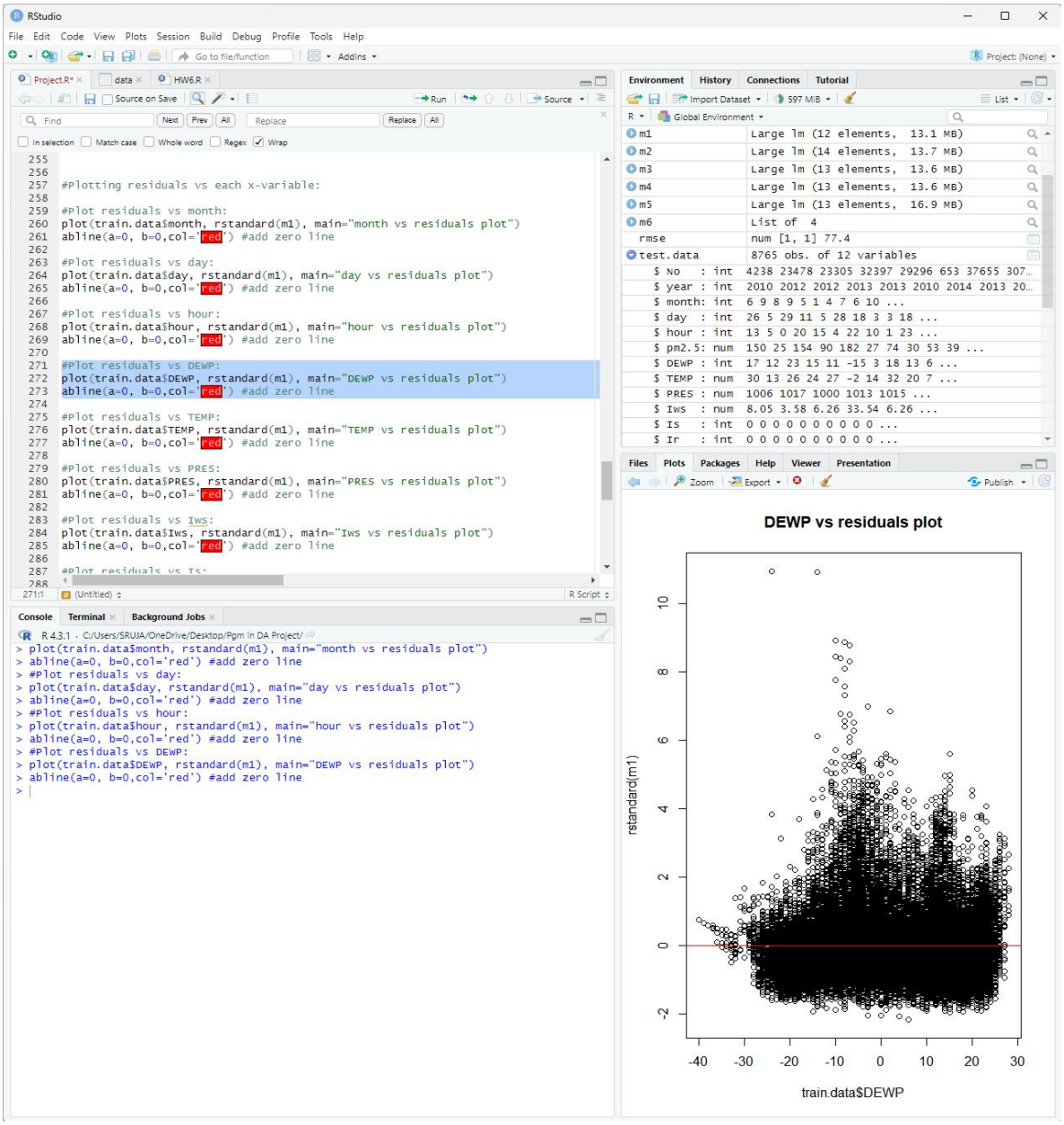


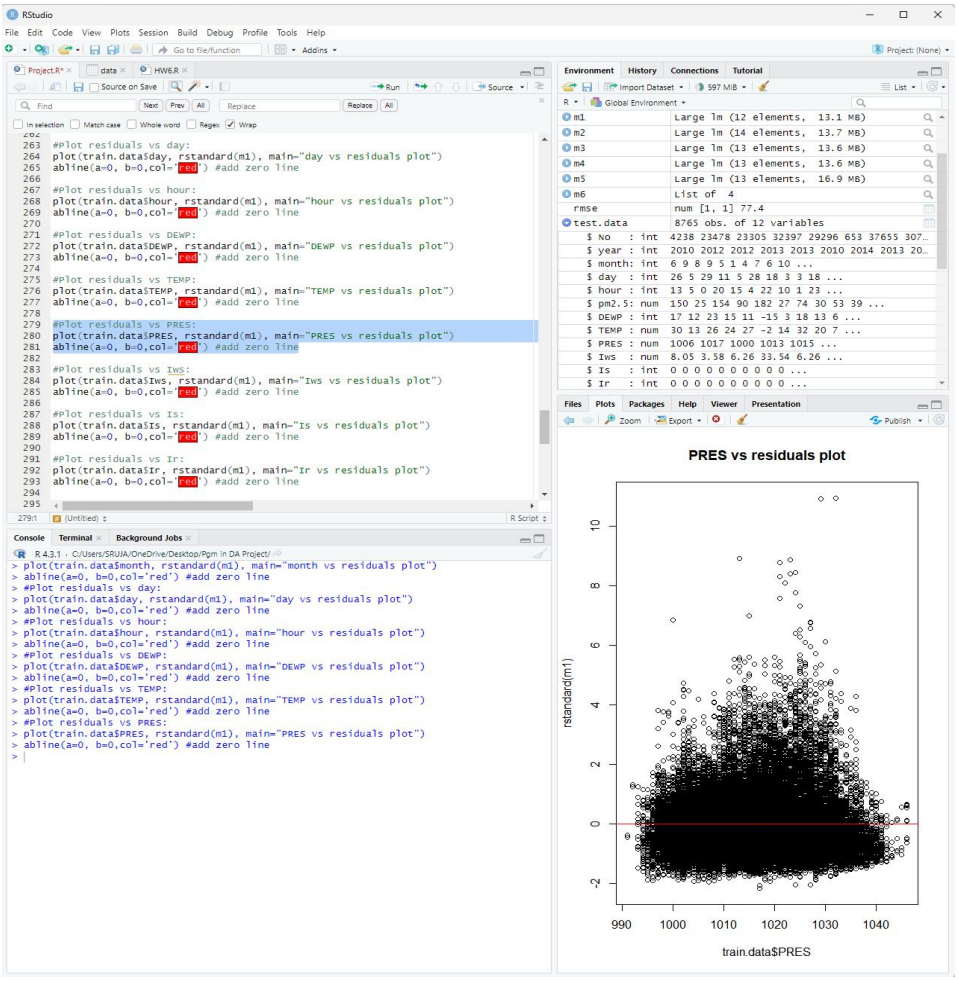
Day vs month

Residuals vs Hour

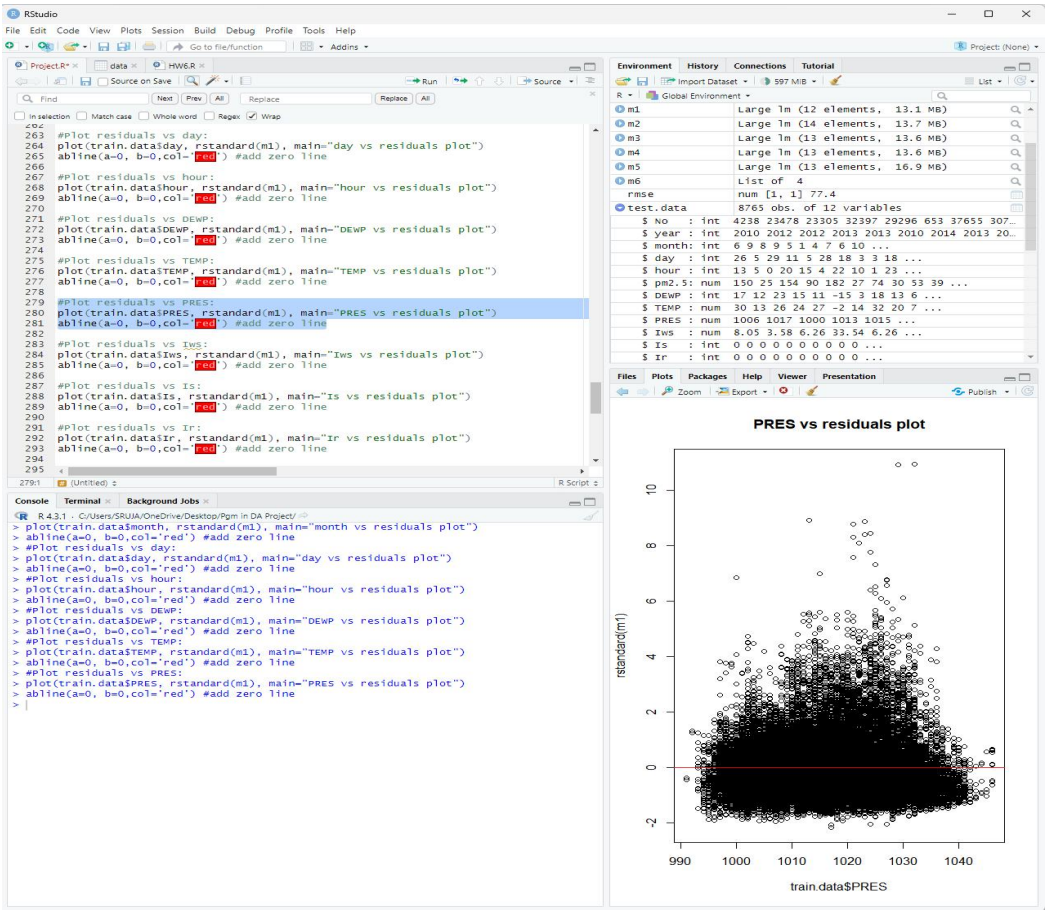


Residuals vs DEWP

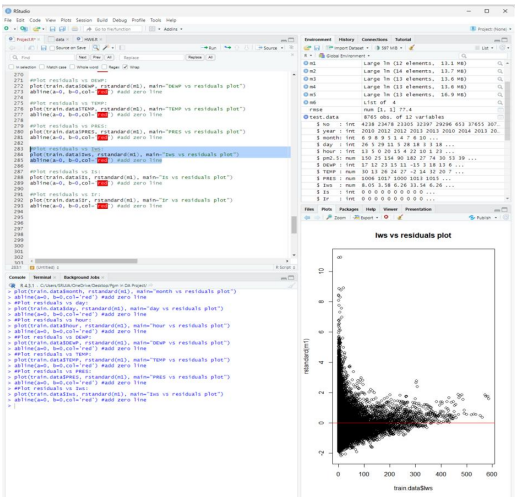




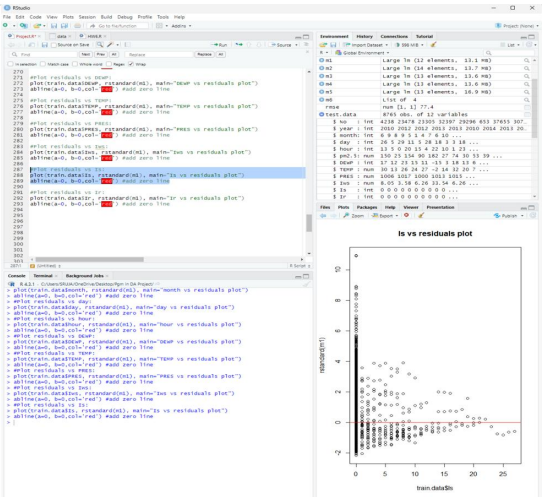
Residuals vs TEMP



Residuals vs PRES



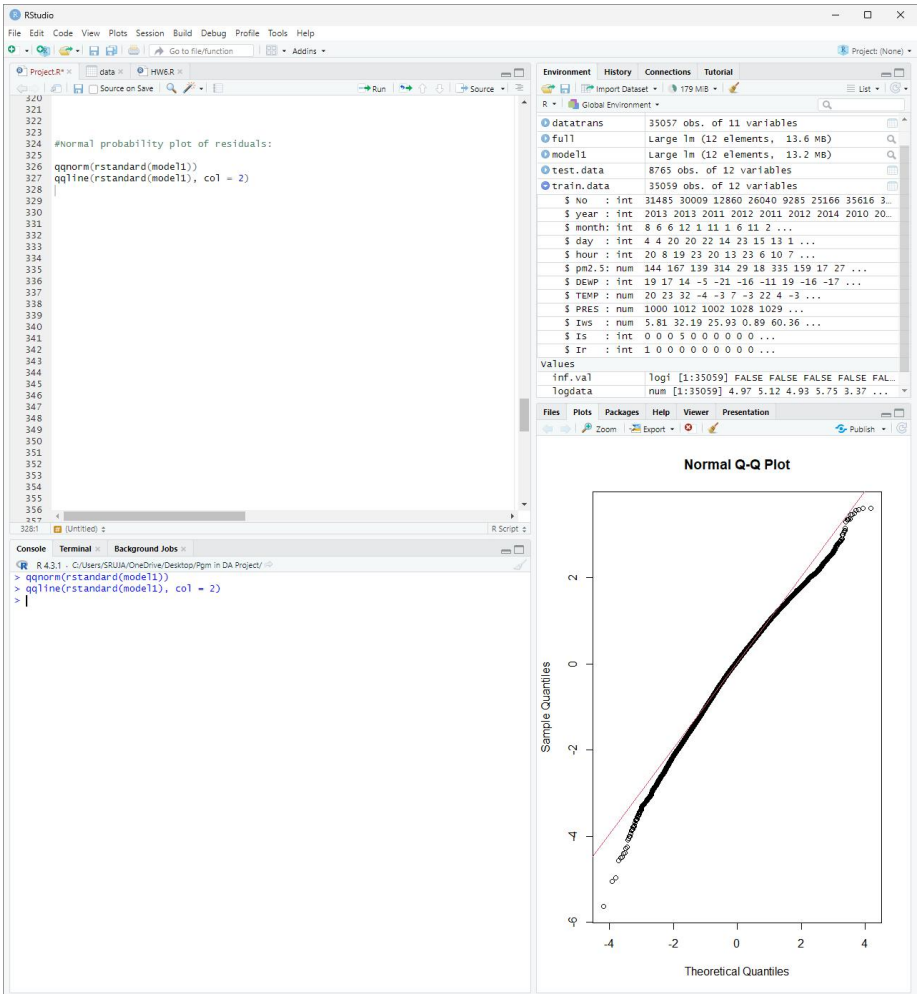
Residuals vs Iws



Residuals vs Is

3.2MODEL DIAGNOSIS

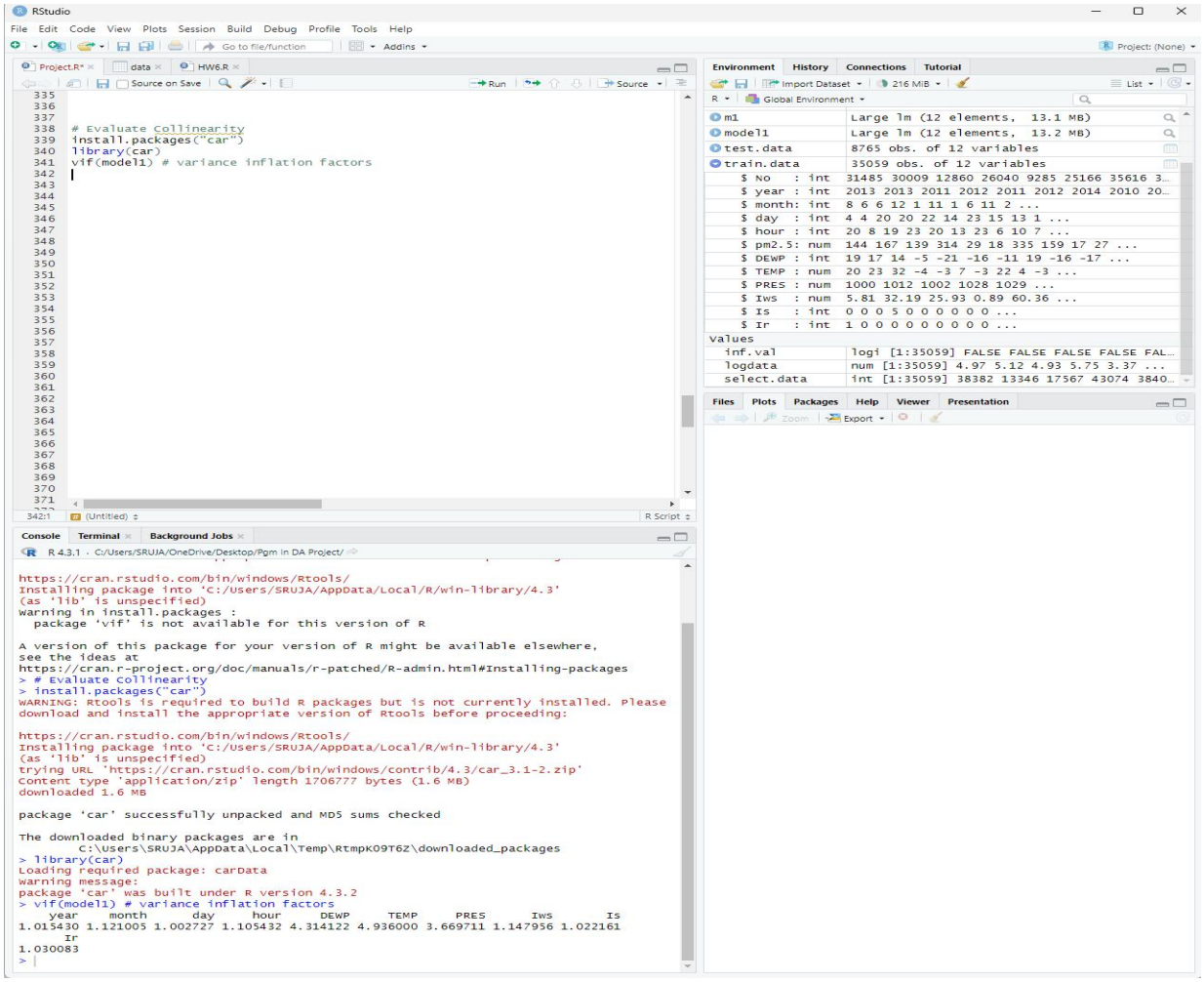
c)Normal probability plot of residuals:To check normality assumption for the error terms

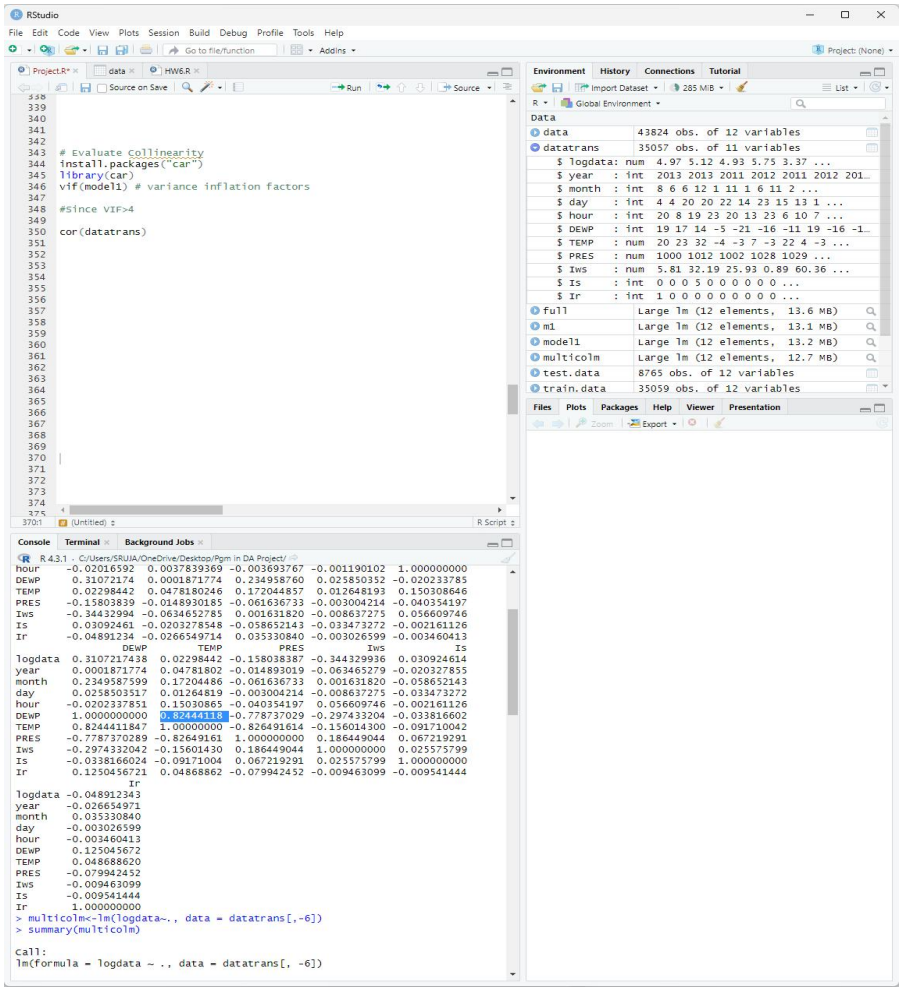


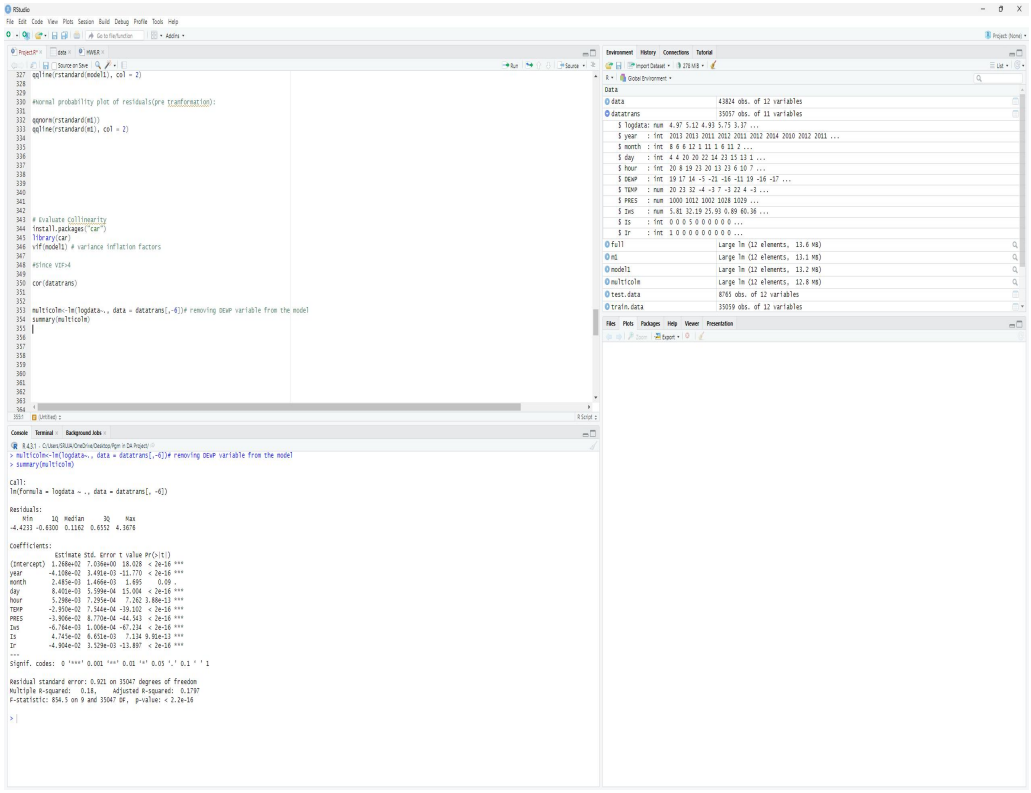
QQ plot after Tranforming Pm2.5 variable.

3.3 IMPROVING MODEL

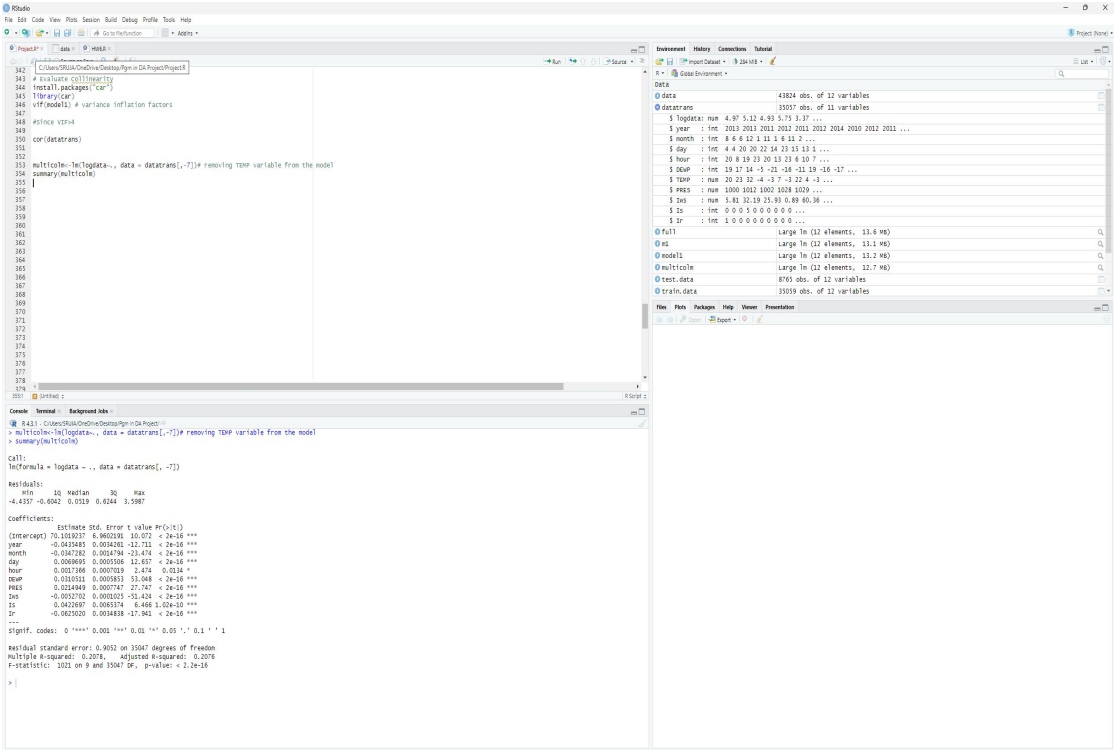
a)Multicollinearity problem




RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
# Evaluate Collinearity
install.packages("car")
library(car)
vif(model1) # variance inflation factors
#since VIF>4
cor(datatrans)
Environment History Connections Tutorial
Data
data 43824 obs. of 12 variables
datatrans 35057 obs. of 11 variables
full Large lm (12 elements, 13.6 MB)
m1 Large lm (12 elements, 13.1 MB)
model1 Large lm (12 elements, 13.2 MB)
multicolm Large lm (12 elements, 12.7 MB)
test.data 8765 obs. of 12 variables
train.data 35059 obs. of 12 variables
Console Terminal Background Jobs
R 4.3.1 · C:/Users/SRUJA/OneDrive/Desktop/Pgm in DA Project/
hour -0.02016592 0.0037839369 -0.003693767 -0.001190102 1.000000000
DEWP 0.31072174 0.0001871774 0.234958760 0.025850352 -0.020233785
TEMP 0.02298442 0.0478180246 0.172044857 0.012648193 0.150308646
PRES -0.15803839 -0.0148930185 -0.061636733 -0.003004214 -0.040354197
Iws -0.34432994 -0.0634652785 0.001631820 -0.008637275 0.056609746
Is 0.03092461 -0.0203278548 -0.058652143 -0.033473272 -0.002161126
Ir -0.04891234 -0.0266549714 0.035330840 -0.003026599 -0.003460413
DEWP TEMP PRES Iws Is
logdata 0.3107217438 0.02298442 -0.158038387 -0.344329936 0.030924614
year 0.0001871774 0.04781802 -0.014893019 -0.063465279 -0.020327855
month 0.2349587599 0.17204486 -0.061636733 0.001631820 -0.058652143
day 0.0258503517 0.01264819 -0.003004214 -0.008637275 -0.033473272
hour -0.0202337851 0.15030865 -0.040354197 0.056609746 -0.002161126
DEWP 1.0000000000 0.82444118 -0.778737029 -0.297433204 -0.033816602
TEMP 0.8244411847 1.00000000 -0.826491614 -0.156014300 -0.091710042
PRES -0.7787370289 -0.82649161 1.000000000 0.186449044 0.067219291
Iws -0.2974332042 -0.15601430 0.186449044 1.000000000 0.025575799
Is -0.0338166024 -0.09171004 0.067219291 0.025575799 1.000000000
Ir 0.1250456721 0.04868862 -0.079942452 -0.009463099 -0.009541444
Ir
logdata -0.048912343
year -0.026654971
month 0.035330840
day -0.003026599
hour -0.003460413
DEWP 0.125045672
TEMP 0.048688620
PRES -0.079942452
Iws -0.009463099
Is -0.009541444
Ir 1.000000000
> multicolm<-lm(logdata~., data = datatrans[,-6])
> summary(multicolm)
Call:
lm(formula = logdata ~ ., data = datatrans[, -6])

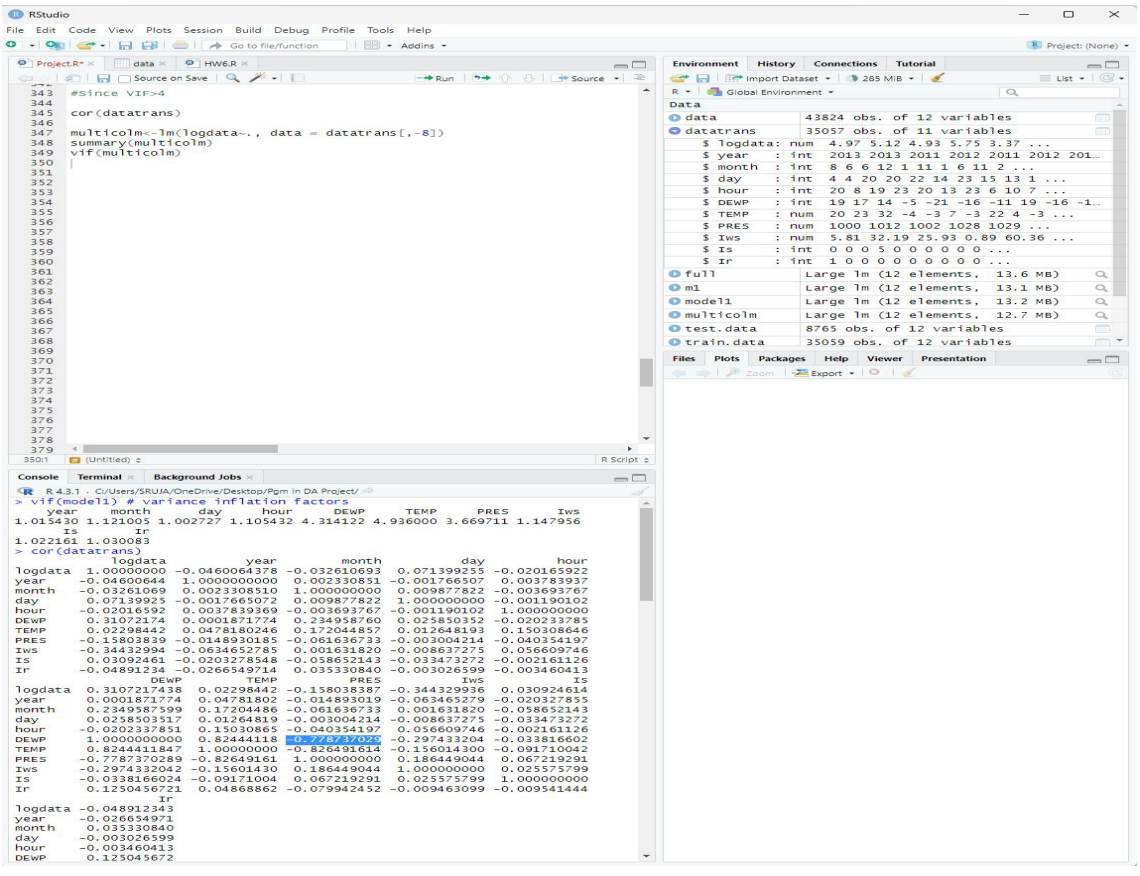

RStudio
qqline(rstandard(model1), col = 2)
#normal probability plot of residuals(pre tranformation):
qqnorm(rstandard(m1))
qqline(rstandard(m1), col = 2)
# Evaluate Collinearity
install.packages("car")
library(car)
vif(model1) # variance inflation factors
#since VIF>4
cor(datatrans)
multicolm<-lm(logdata~., data = datatrans[,-6])# removing DEWP variable from the model
summary(multicolm)
R 4.3.1 · C:/Users/SRUJA/OneDrive/Desktop/Pgm in DA Project/
> multicolm<-lm(logdata~., data = datatrans[,-6])# removing DEWP variable from the model
> summary(multicolm)
Call:
lm(formula = logdata ~ ., data = datatrans[, -6])
Residuals:
Min 1Q Median 3Q Max
-4.4233 -0.0300 0.1162 0.6552 4.3678
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.268e+02 7.036e+00 18.028 < 2e-16 ***
year -4.110e-02 3.491e-03 -11.770 < 2e-16 ***
month 2.485e-03 1.496e-03 1.695 0.09 .
day 8.401e-03 5.599e-04 15.004 < 2e-16 ***
hour 5.298e-03 7.295e-04 7.262 3.88e-13 ***
TEMP -2.950e-02 7.344e-04 -39.102 < 2e-16 ***
PRES -3.906e-02 8.770e-04 -44.543 < 2e-16 ***
Iws -6.764e-03 1.009e-04 -67.234 < 2e-16 ***
Is 4.745e-02 6.651e-03 7.134 9.81e-13 ***
Ir -4.904e-02 3.529e-03 -13.897 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.921 on 35047 degrees of freedom
Multiple R-squared: 0.18, Adjusted R-squared: 0.1797
F-statistic: 854.5 on 9 and 35047 DF, p-value: < 2.2e-16



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

#since VIF>4

cor(datatrans)


multicolm<-lm(logdata~., data = datatrans[,-7])# removing TEMP variable from the model
summary(multicolm)
```

```
> multicolm<-lm(logdata~., data = datatrans[,-7])# removing TEMP variable from the model
> summary(multicolm)

Call:
lm(formula = logdata ~ ., data = datatrans[, -7])

Residuals:
    Min      1Q  Median      3Q     Max
-4.4357 -0.6042  0.0519  0.6244  3.5987

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.1019237  6.9602191  10.072  < 2e-16 ***
year        -0.0435483  0.0034291 -12.711  < 2e-16 ***
month       -0.0347282  0.0014794 -23.474  < 2e-16 ***
day          0.0069695  0.0005506  12.657  < 2e-16 ***
hour         0.0017366  0.0007019   2.474   0.0134 *
DEWP         0.0351511  0.0005853  53.048  < 2e-16 ***
PRES         0.0214949  0.0007747  27.747  < 2e-16 ***
Iws         -0.0052702  0.0001025 -51.424  < 2e-16 ***
Is           0.0422697  0.0065374   6.466 1.02e-10 ***
Ir          -0.0825020  0.0034838 -17.941  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9052 on 35047 degrees of freedom
Multiple R-squared:  0.2078,    Adjusted R-squared:  0.2076
F-statistic:  1021 on 9 and 35047 DF,  p-value: < 2.2e-16
```

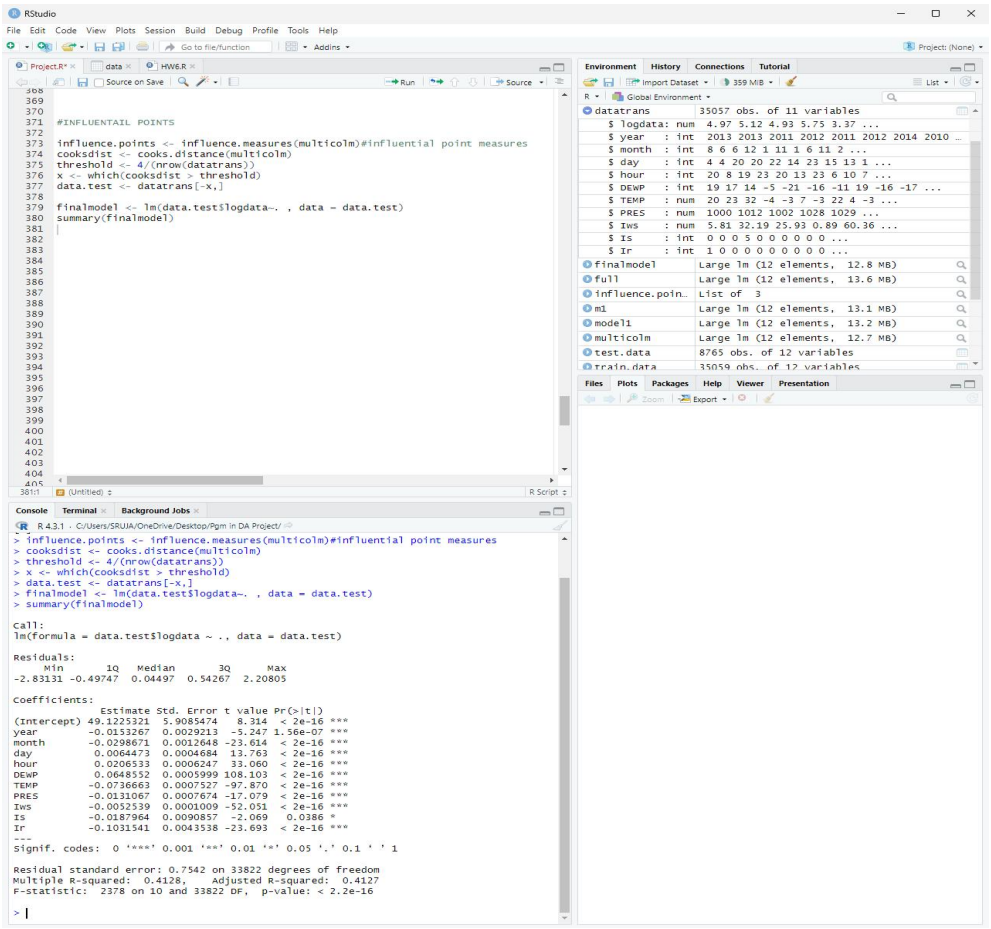


```
#since VIF>4

cor(datatrans)

multicolm<-lm(logdata~., data = datatrans[,-8])
summary(multicolm)
vif(multicolm)
```

```
> vif(model1) # variance inflation factors
    year     month       day      hour      DEWP      TEMP      PRES       Iws
1.015430 1.121005 1.002727 1.105432 4.314122 4.936000 3.669711 1.147956
      Is        Ir
1.022161 1.030083
> cor(datatrans)
              logdata          year         month           day          hour
logdata    1.00000000 -0.0460064378 -0.032610693  0.071399255 -0.020165922
year      -0.04600644  1.0000000000  0.002330851 -0.001766507  0.003783937
month     -0.03261069  0.0023308510  1.000000000  0.009877822 -0.003693767
day        0.07139925 -0.0017665072  0.009877822  1.000000000 -0.001190102
hour      -0.02016592  0.0037839369 -0.003693767 -0.001190102  1.000000000
DEWP       0.31072174  0.0001871774  0.234958760  0.025850352 -0.020233785
TEMP       0.02298442  0.0478180246  0.172044857  0.012648193  0.150308646
PRES      -0.15803839 -0.0148930185 -0.061636733 -0.003004214 -0.040354197
Iws       -0.34432994 -0.0634652785  0.001631820 -0.008637275  0.056609746
Is         0.03092461 -0.0203278548 -0.058652143 -0.033473272 -0.002161126
Ir        -0.04891234 -0.0266549714  0.035330840 -0.003026599 -0.003460413
                DEWP        TEMP        PRES         Iws          Is
logdata    0.3107217438  0.02298442 -0.158038387 -0.344329936  0.030924614
year       0.0001871774  0.04781802 -0.014893019 -0.063465279 -0.020327855
month      0.2349587599  0.17204486 -0.061636733  0.001631820 -0.058652143
day        0.0258503517  0.01264819 -0.003004214 -0.008637275 -0.033473272
hour      -0.0202337851  0.15030865 -0.040354197  0.056609746 -0.002161126
DEWP       1.0000000000  0.82444118 -0.778737029 -0.297433204 -0.033816602
TEMP       0.8244411847  1.00000000 -0.826491614 -0.156014300 -0.091710042
PRES      -0.7787370289 -0.82649161  1.000000000  0.186449044  0.067219291
Iws       -0.2974332042 -0.15601430  0.186449044  1.000000000  0.025575799
Is        -0.0338166024 -0.09171004  0.067219291  0.025575799  1.000000000
Ir         0.1250456721  0.04868862 -0.079942452 -0.009463099 -0.009541444
                  Ir
logdata -0.048912343
year    -0.026654971
month    0.035330840
day     -0.003026599
hour    -0.003460413
DEWP     0.125045672
```

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



EVALUATION STRATEGIES

5-Crossfold evaluation