Airquality

Assignment 1

GS Kumbhare

08/08/2020

Introduction

Regression is a statistical method used in finance, investing and other disciplines that attempt to determine the strength and character of the relationship between one dependent variable to that of other independent variables. Through this assignment we try to find relationship of Ozone layer to that of other independent variables like Solar radiation, Wind speed and Temperature. We study how regression analysis works and learn about dependent and independent variables.

Our dataset consists of Airquality data for the city of New York from the year 1973. The data was collected for a period of 5 month, from the month of May to the month of September. Our variables consist of 1. Ozone in PPM 2. Solar radiation in PPM 3. Wind speed in Miles per hour 4. Temperature in Farenheit 5. Month 6. Days of month

Cleaning

Firstly we load the dataset from r directory.

```
datasets::airquality
```

We first analyse structure of Airquality dataset obtained from r dataset directory.

```
str(airquality)
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

We find that there are number of na values in the structure of 153 observations. Let us calculate number of na values in each variable and filter them accordingly.

```
colSums(is.na(airquality))
```

```
##   Ozone Solar.R    Wind    Temp   Month     Day
##      37       7       0       0       0       0
```

With the above table number of na values or missing values in our dataset.

Next, we remove na values so that our dataset is ready for next step of

```
air= airquality
#Monthly mean to Ozone
for (i in 1:nrow(air)){
  if(is.na(air[i, "Ozone"])){
    air[i,"Ozone"]<- mean(air[which(air[,"Month"]==air[i,"Month"]),"Ozone"],na.rm = TRUE)
  }
}
#Monthly mean to solar. R
for (i in 1:nrow(air)){
  if(is.na(air[i, "Solar.R"])){
    air[i,"Solar.R"]<- mean(air[which(air[,"Month"]==air[i,"Month"]),"Solar.R"],na.rm = TRUE)
  }
}
summary(air)
```

```
##      Ozone          Solar.R          Wind            Temp
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00
##  1st Qu.: 21.00   1st Qu.:120.0   1st Qu.: 7.400   1st Qu.:72.00
##  Median : 29.44   Median :194.0   Median : 9.700   Median :79.00
##  Mean   : 40.85   Mean   :185.5   Mean   : 9.958   Mean   :77.88
##  3rd Qu.: 59.12   3rd Qu.:256.0   3rd Qu.:11.500   3rd Qu.:85.00
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
##      Month            Day
##  Min.   :5.000   Min.   : 1.0
##  1st Qu.:6.000   1st Qu.: 8.0
##  Median :7.000   Median :16.0
##  Mean   :6.993   Mean   :15.8
##  3rd Qu.:8.000   3rd Qu.:23.0
##  Max.   :9.000   Max.   :31.0
```

We removed na from the dataset.

Normalization

Our dataset has varying range. Ozone is in scale of PPM, Solar radiation is in range of PPM, Temp is scale of Farenheit, and wind in scale of km/hr. As our data set has variying range and we normalize the dataset for better fit.

```
normal<- function(x){
  return((x-min(x))/(max(x)-min(x)))
}
air<- normal(air)
str(air)
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : num  0.1201 0.1051 0.033 0.0511 0.0679 ...
##  $ Solar.R: num  0.568 0.351 0.444 0.937 0.541 ...
##  $ Wind   : num  0.0192 0.021 0.0348 0.0315 0.0399 ...
##  $ Temp   : num  0.198 0.213 0.219 0.183 0.165 ...
##  $ Month  : num  0.012 0.012 0.012 0.012 0.012 ...
##  $ Day    : num  0 0.003 0.00601 0.00901 0.01201 ...
```

Libraries

We load required libraries for our regression analysis.

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

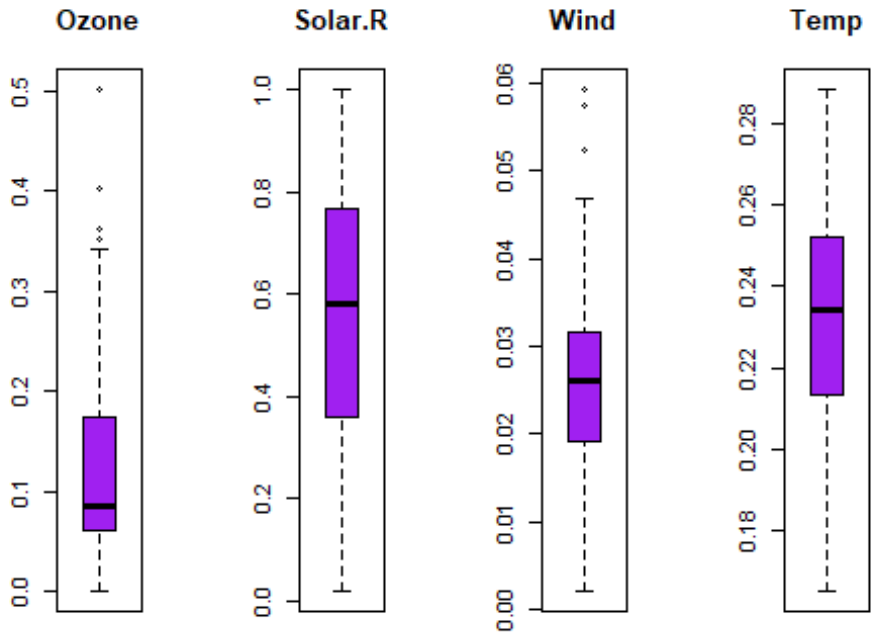
```
## v tibble  3.0.2     v purrr   0.3.4
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(corrplot)

## corrplot 0.84 loaded
```

A. Univariate Analysis

1. Box plot

```
par(mfrow=c(1,4))
for(i in 1:4) {
   boxplot(air[,i], main=names(air)[i],
  col = c("purple"))
}
```

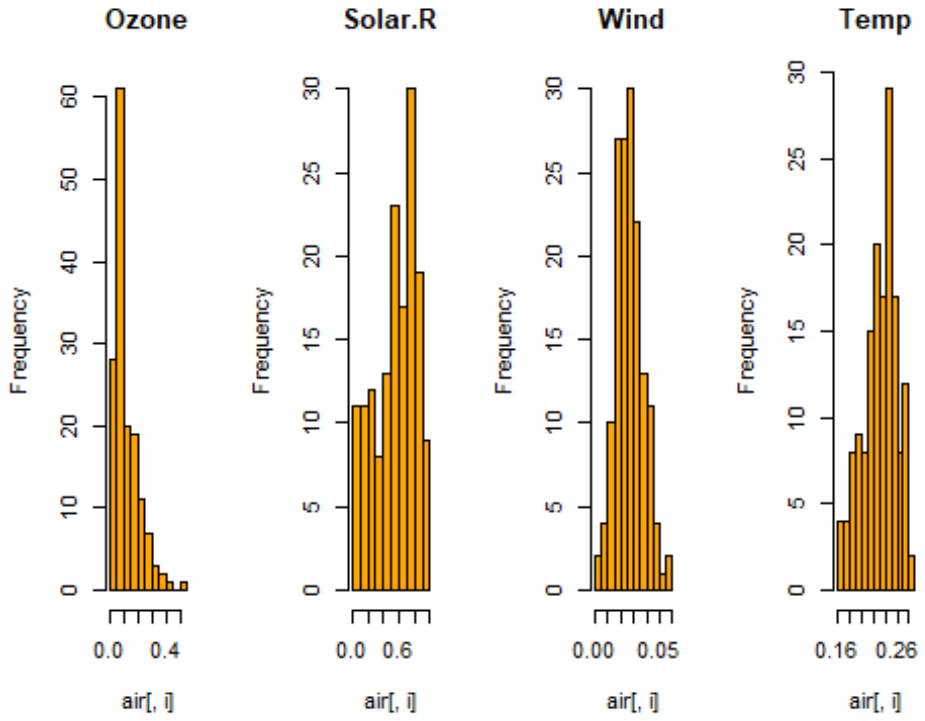


1. Solar radiation, wind and Temperature boxplots are almost evenly distributed.

2. Ozone boxplot is unevenly distributed.

Histogram

```
par(mfrow=c(1,4))
for(i in 1:4) {
   hist(air[,i], main=names(air)[i],
  col = c("orange"))
}
```

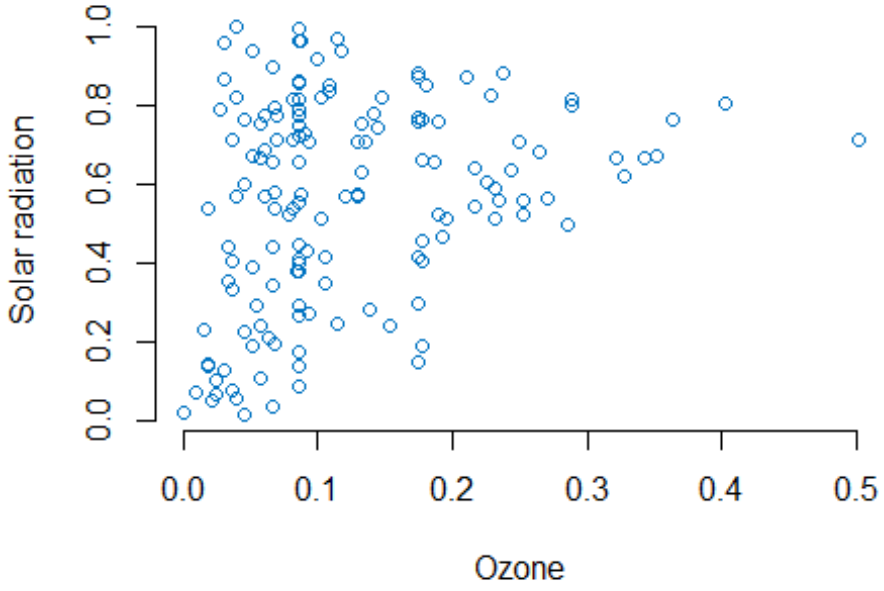


Multivariate Analysis

In multi variate analysis we would be using scatter plot to analyse our first model is Ozone vs Solar radiation scatter box.

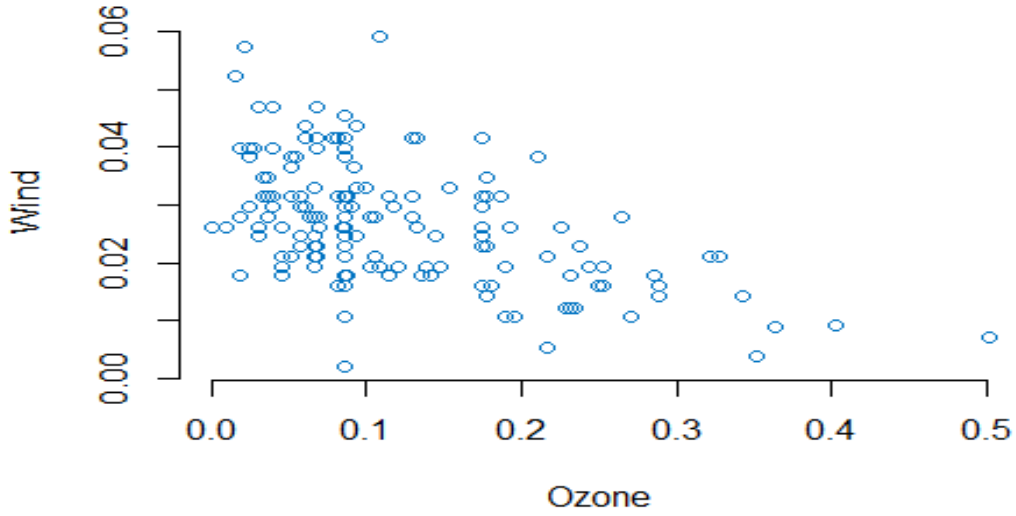
1. Model 1 Ozone Vs Solar Radiation

```
plot(x = air$Ozone, y = air$Solar.R, frame = FALSE,
     xlab = "Ozone", ylab = "Solar radiation",
     col =  "#0073C2FF")
```



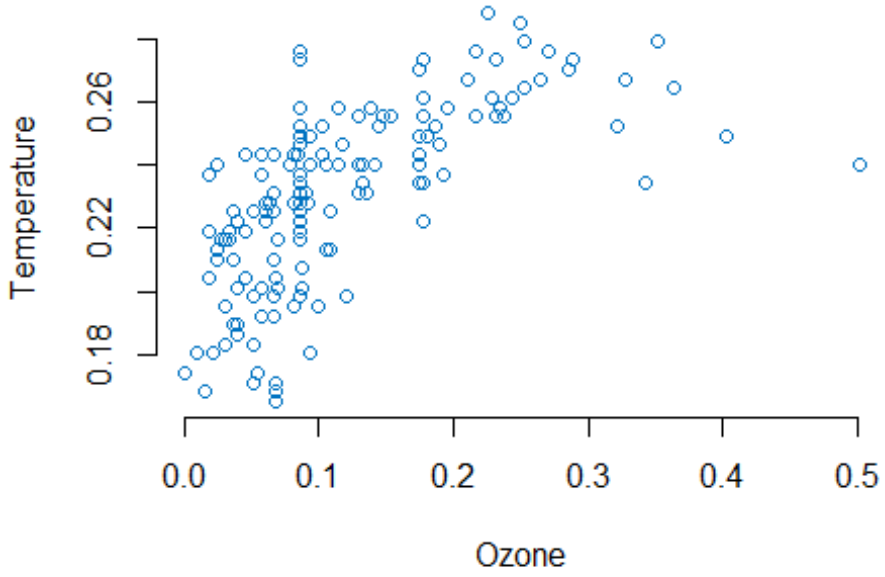
2. Model 2 Ozone vs Wind

```
plot(x = air$Ozone, y = air$Wind, frame = FALSE,
     xlab = "Ozone", ylab = "Wind",
      col =  "#0073C2FF")
```



3. Model 3 Ozone vs Temperature

```
plot(x = air$Ozone, y = air$Temp, frame = FALSE,
     xlab = "Ozone", ylab = "Temperature",
      col =  "#0073C2FF")
```



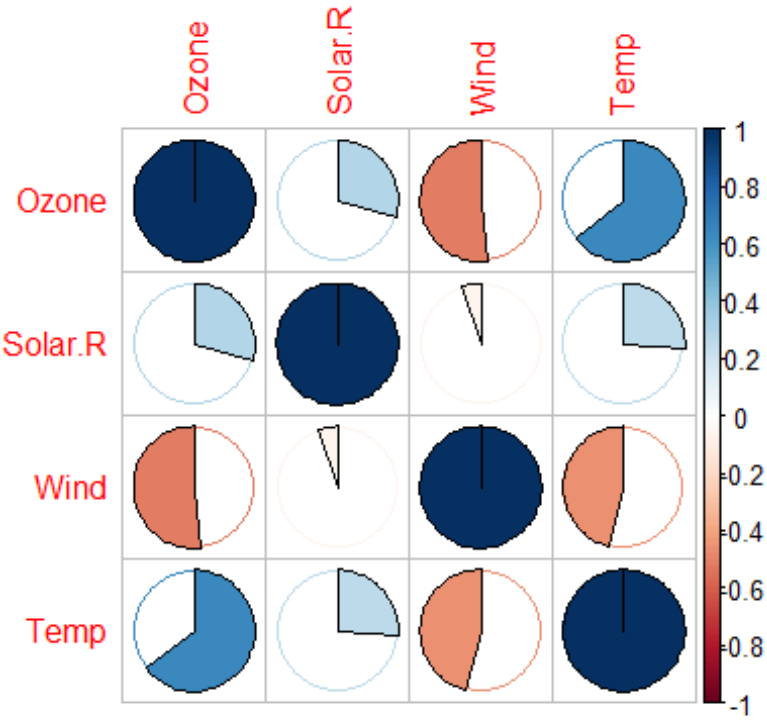
Correlation

The following table and plot show correlation between variables.

```
cor(air[,1:4])
```

```
##             Ozone     Solar.R       Wind       Temp
## Ozone   1.0000000  0.29280514 -0.51675044  0.6456381
## Solar.R 0.2928051  1.00000000 -0.05237183  0.2619312
## Wind   -0.5167504 -0.05237183  1.00000000 -0.4579879
## Temp    0.6456381  0.26193117 -0.45798788  1.0000000
```

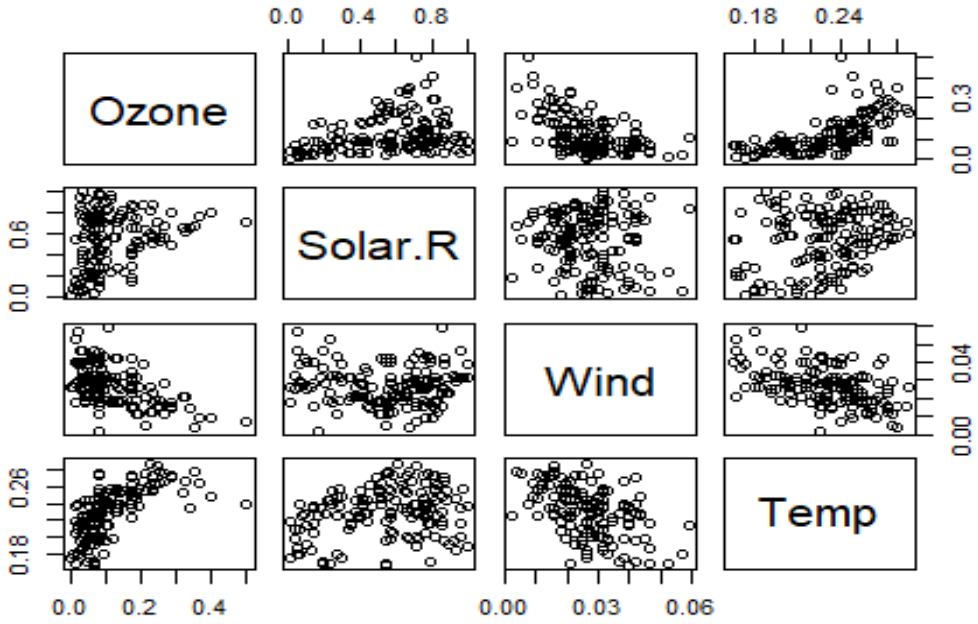
```
corrplot(cor(air[,1:4]), method = "pie")
```



From the above corr plot we can see that Ozone, Wind and Temperature are highly correlated.

Overall Plot

```
plot(air[,1:4])
```



Regression analysis
linear regression

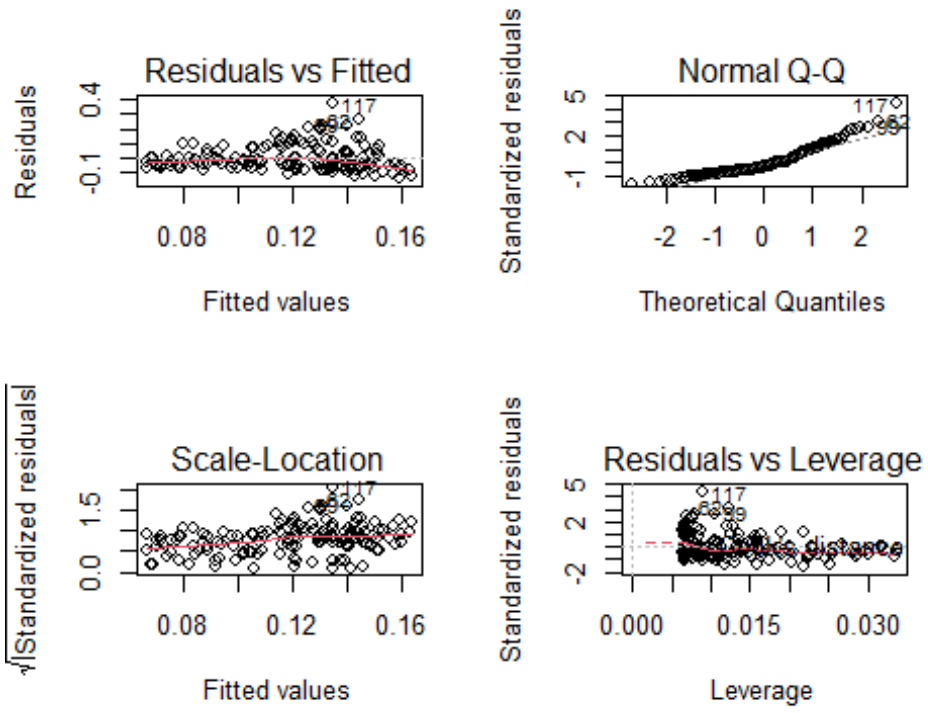
While progressing through linear regression we will use forward selection Method. Our first model will have Ozone as dependent variable and Solar radiation as independent variable.

1. Model 1 Ozone vs Solar radiation Linear regression

```
modelLm1<- lm(Ozone~ Solar.R, data = air)
print(modelLm1)
```

```
##
## Call:
## lm(formula = Ozone ~ Solar.R, data = air)
##
## Coefficients:
## (Intercept)      Solar.R
##     0.06509      0.09849
```

```
par(mfrow = c(2,2))
plot(modelLm1)
```




```
summary(modelLm1)
```

```
##
## Call:
## lm(formula = Ozone ~ Solar.R, data = air)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12941 -0.06012 -0.02274  0.04524  0.36631
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.06509    0.01606   4.054 8.05e-05 ***
## Solar.R      0.09849    0.02617   3.763  0.00024 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08526 on 151 degrees of freedom
## Multiple R-squared:  0.08573,    Adjusted R-squared:  0.07968
## F-statistic: 14.16 on 1 and 151 DF,  p-value: 0.0002398
```

From the model above we can see that every single percentage increase in Solar radiation our Ozone increases by 0.098.

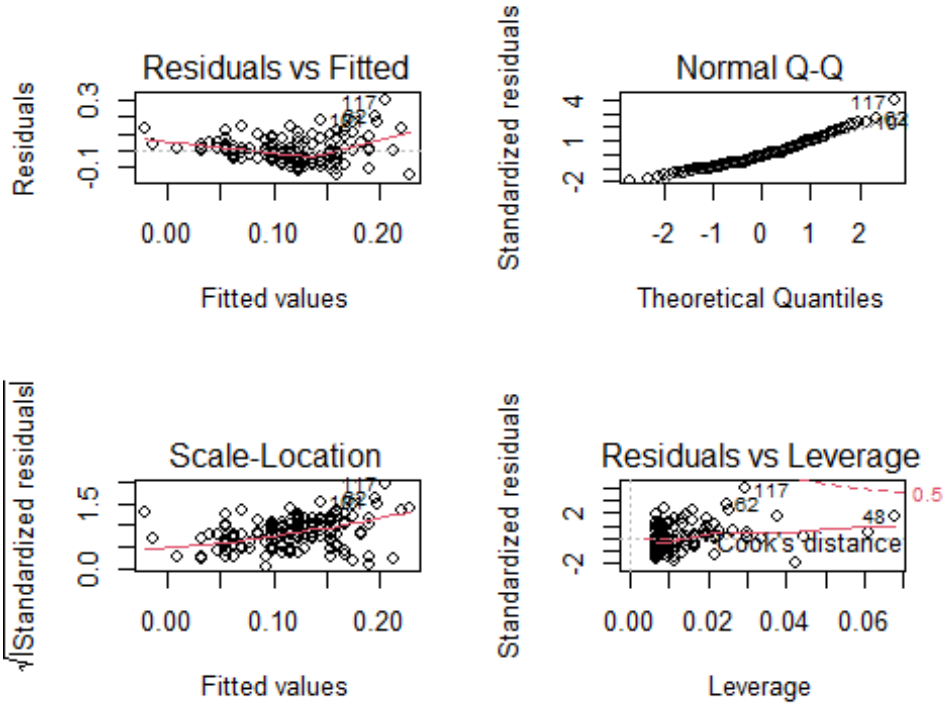
1. Residual is near 0 which means Ozone to Solar radiation residual is symmetrical.
2. The average Solar radiation is 0.065 Units to that of Ozone.
3. Every 1 unit increase in Solar radiation the Ozone increases by 0.098 Unit and vice versa.
4. If we re run the model there can be difference of 0.016 Units of Ozone.
5. Our p value is significantly small thus we can reject null hypothesis.
6. Our residual standard error is 0.085. We can say that percentage rate is 130.76%.
7. R^2 is 0.085 or 8.5% variaance found which is relatively small. Solar radiation is not a strong predictor variable for Ozone.

2. Model 2 Ozone vs Solar radiation

```
modelLm2<- lm(Ozone~Wind,data= air)
print(modelLm2)
```

```
##
## Call:
## lm(formula = Ozone ~ Wind, data = air)
##
## Coefficients:
## (Intercept)         Wind
##      0.2364      -4.3410
```

```
par(mfrow = c(2,2))
plot(modelLm2)
```



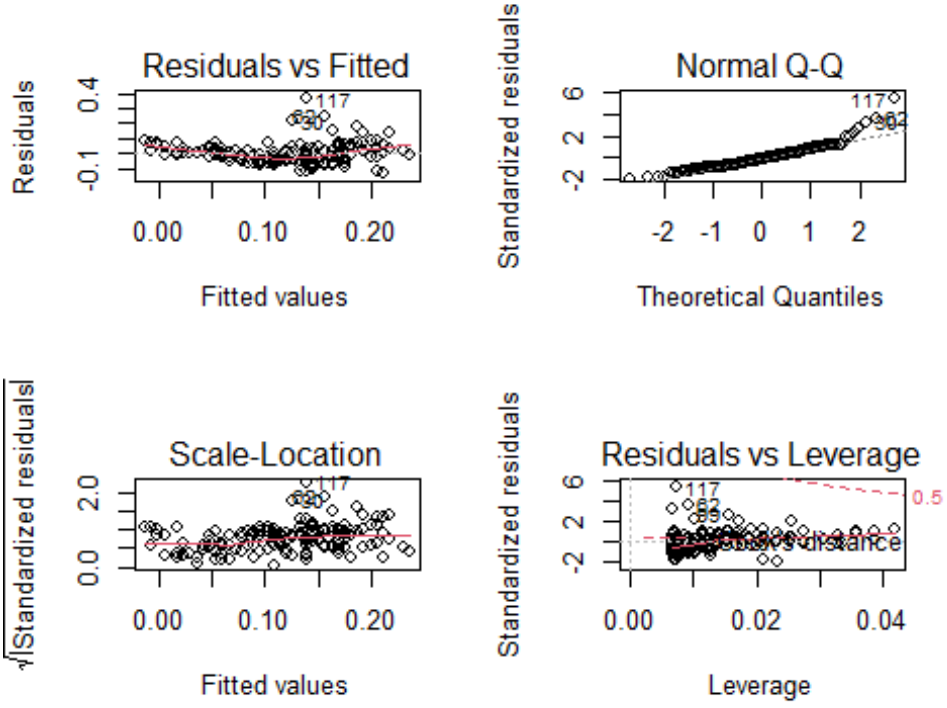

```
summary(modelLm2)

##
## Call:
## lm(formula = Ozone ~ Wind, data = air)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14190 -0.05215 -0.01311  0.04523  0.29634
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.23644    0.01691  13.982  < 2e-16 ***
## Wind        -4.34099    0.58528  -7.417 8.03e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07634 on 151 degrees of freedom
## Multiple R-squared:  0.267,  Adjusted R-squared:  0.2622
## F-statistic: 55.01 on 1 and 151 DF,  p-value: 8.034e-12
```

1. Residual is near 0 which means Ozone to Wind residual is symmetrical.
2. The average wind speed is 0.236 Units to that of Ozone.
3. Every 1 unit increase in Wind speed the Ozone decreases by 4.43 Unit and vice versa.
4. If we re run the model there can be diffrence of 0.016 Units of Ozone.

5. Our p value is significantly small thus we can reject null hypothesis.
6. Our residual standard error is 0.076. We can say that percentage rate is 32.2%.
7. R^2 is 0.267 or 26.7% variaance found which is relatively small. This Wind is not a strong predictor variable for Ozone.

3. Model 3 Ozone vs Temperature

```
modelLm3<- lm(Ozone~Temp,data= air)
print(modelLm3)

##
## Call:
## lm(formula = Ozone ~ Temp, data = air)
##
## Coefficients:
## (Intercept)        Temp
##     -0.3464       2.0187

par(mfrow = c(2,2))
plot(modelLm3)
```



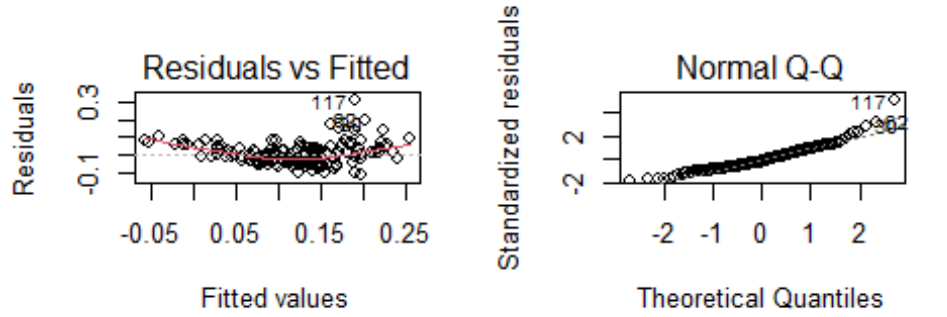

```
summary(modelLm3)

##
## Call:
## lm(formula = Ozone ~ Temp, data = air)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12590 -0.04709 -0.00644  0.03172  0.36293
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3464     0.0452  -7.664 2.03e-12 ***
## Temp          2.0187     0.1943  10.389  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06809 on 151 degrees of freedom
## Multiple R-squared:  0.4168, Adjusted R-squared:  0.413
## F-statistic: 107.9 on 1 and 151 DF,  p-value: < 2.2e-16
```

1. Residual is near 0 which means Ozone to Temperature residual is symmetrical.
2. Every 1 unit increase in Temperature the Ozone Increases by 2.018 Unit and vice versa.
3. If we re run the model, standard error difference can be of 0.194 Units of Ozone.
4. Our p value is significantly small thus we can reject null hypothesis.
5. Our residual standard error is 0.068. We can say that percentage rate is 3.38%.
6. R^2 is 0.416 or 41.6% variance found which is relatively bigger than the other predictor. This shows that Temperature is a strong predictor variable for Ozone.

4. Model 4 Ozone with all predictor

```
modelLm4<- lm(Ozone~ Solar.R + Wind +Temp,data= air)
print(modelLm4)

##
## Call:
## lm(formula = Ozone ~ Solar.R + Wind + Temp, data = air)
##
## Coefficients:
## (Intercept)      Solar.R         Wind         Temp
##    -0.18296      0.05182     -2.46058      1.47311

par(mfrow = c(2,2))
plot(modelLm4)
```




```
summary(modelLm4)

##
## Call:
## lm(formula = Ozone ~ Solar.R + Wind + Temp, data = air)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.11234 -0.04372 -0.01333  0.03648  0.31142
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.18296    0.05564  -3.288  0.00126 **
## Solar.R      0.05182    0.02024   2.560  0.01145 *
## Wind        -2.46058    0.54876  -4.484 1.45e-05 ***
## Temp         1.47311    0.21135   6.970 9.56e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06343 on 149 degrees of freedom
## Multiple R-squared:  0.5007, Adjusted R-squared:  0.4906
## F-statistic:  49.8 on 3 and 149 DF,  p-value: < 2.2e-16
```

In our last model we have all the predictors. The observations made are

1. R^2 is 0.5007 or 50.07% variance. This shows that all the three predictors together have strong impact on Ozone layer concentration.
2. Our p value is significanlty small or near 0 which shows that we can reject null hypothesis and accept the model.
3. Residual standard error is 0.063 and the percentage rate is 34.5% for the model. The error rate is significantly high.
4. The model with all the independent variables has high error rate.

Prediction

```
predy <- predict(modelLm3, air, interval="predict", level=.95) + predict(modelLm1, air, interval="predict", level=.95) + predict(modelLm2, air, interval="predict", level=.95)
summary(predy)
```

```
##       fit             lwr              upr        
##  Min.   :0.0752   Min.   :-0.38751   Min.   :0.5379  
##  1st Qu.:0.2963   1st Qu.:-0.16086   1st Qu.:0.7535  
##  Median :0.3668   Median :-0.08861   Median :0.8225  
##  Mean   :0.3590   Mean   :-0.09776   Mean   :0.8158  
##  3rd Qu.:0.4301   3rd Qu.:-0.02566   3rd Qu.:0.8858  
##  Max.   :0.5682   Max.   : 0.10922   Max.   :1.0272
```

```
conf <- predict(modelLm3, air, interval="confidence", level=.95) + predict(modelLm1, air, interval="confidence", level=.95) + predict(modelLm2, air, interval="confidence", level=.95)

summary(conf)
```

```
##       fit             lwr              upr        
##  Min.   :0.0752   Min.   :-0.01359   Min.   :0.1640  
##  1st Qu.:0.2963   1st Qu.: 0.24611   1st Qu.:0.3477  
##  Median :0.3668   Median : 0.32463   Median :0.4092  
##  Mean   :0.3590   Mean   : 0.30904   Mean   :0.4090  
##  3rd Qu.:0.4301   3rd Qu.: 0.38108   3rd Qu.:0.4733  
##  Max.   :0.5682   Max.   : 0.50251   Max.   :0.6339
```

```
conf[1]== predy[1]
```

```
## [1] TRUE
```

Our Model 1,2 and 3 has fit for the predictor.

Conclusion

Through this Analysis we learnt to perform regression analysis for different predictors of Ozone. We found that Temperature is fittest predictor of Ozone layer. We learned to analyse different predictor with univariate and multivariate analysis. We learned to build histogram and boxplots for variables. We learned to plot correlation plots.