Airquality1

GS Kumbhare

09/08/2020

Airquality

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 consists 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

First we load the dataset from r directory.

datasets::airquality

## Ozone Solar.R Wind Temp Month Day  
## 1 41 190 7.4 67 5 1  
## 2 36 118 8.0 72 5 2  
## 3 12 149 12.6 74 5 3  
## 4 18 313 11.5 62 5 4  
## 5 NA NA 14.3 56 5 5  
## 6 28 NA 14.9 66 5 6  
## 7 23 299 8.6 65 5 7  
## 8 19 99 13.8 59 5 8  
## 9 8 19 20.1 61 5 9  
## 10 NA 194 8.6 69 5 10  
## 11 7 NA 6.9 74 5 11  
## 12 16 256 9.7 69 5 12  
## 13 11 290 9.2 66 5 13  
## 14 14 274 10.9 68 5 14  
## 15 18 65 13.2 58 5 15  
## 16 14 334 11.5 64 5 16  
## 17 34 307 12.0 66 5 17  
## 18 6 78 18.4 57 5 18  
## 19 30 322 11.5 68 5 19  
## 20 11 44 9.7 62 5 20  
## 21 1 8 9.7 59 5 21  
## 22 11 320 16.6 73 5 22  
## 23 4 25 9.7 61 5 23  
## 24 32 92 12.0 61 5 24  
## 25 NA 66 16.6 57 5 25  
## 26 NA 266 14.9 58 5 26  
## 27 NA NA 8.0 57 5 27  
## 28 23 13 12.0 67 5 28  
## 29 45 252 14.9 81 5 29  
## 30 115 223 5.7 79 5 30  
## 31 37 279 7.4 76 5 31  
## 32 NA 286 8.6 78 6 1  
## 33 NA 287 9.7 74 6 2  
## 34 NA 242 16.1 67 6 3  
## 35 NA 186 9.2 84 6 4  
## 36 NA 220 8.6 85 6 5  
## 37 NA 264 14.3 79 6 6  
## 38 29 127 9.7 82 6 7  
## 39 NA 273 6.9 87 6 8  
## 40 71 291 13.8 90 6 9  
## 41 39 323 11.5 87 6 10  
## 42 NA 259 10.9 93 6 11  
## 43 NA 250 9.2 92 6 12  
## 44 23 148 8.0 82 6 13  
## 45 NA 332 13.8 80 6 14  
## 46 NA 322 11.5 79 6 15  
## 47 21 191 14.9 77 6 16  
## 48 37 284 20.7 72 6 17  
## 49 20 37 9.2 65 6 18  
## 50 12 120 11.5 73 6 19  
## 51 13 137 10.3 76 6 20  
## 52 NA 150 6.3 77 6 21  
## 53 NA 59 1.7 76 6 22  
## 54 NA 91 4.6 76 6 23  
## 55 NA 250 6.3 76 6 24  
## 56 NA 135 8.0 75 6 25  
## 57 NA 127 8.0 78 6 26  
## 58 NA 47 10.3 73 6 27  
## 59 NA 98 11.5 80 6 28  
## 60 NA 31 14.9 77 6 29  
## 61 NA 138 8.0 83 6 30  
## 62 135 269 4.1 84 7 1  
## 63 49 248 9.2 85 7 2  
## 64 32 236 9.2 81 7 3  
## 65 NA 101 10.9 84 7 4  
## 66 64 175 4.6 83 7 5  
## 67 40 314 10.9 83 7 6  
## 68 77 276 5.1 88 7 7  
## 69 97 267 6.3 92 7 8  
## 70 97 272 5.7 92 7 9  
## 71 85 175 7.4 89 7 10  
## 72 NA 139 8.6 82 7 11  
## 73 10 264 14.3 73 7 12  
## 74 27 175 14.9 81 7 13  
## 75 NA 291 14.9 91 7 14  
## 76 7 48 14.3 80 7 15  
## 77 48 260 6.9 81 7 16  
## 78 35 274 10.3 82 7 17  
## 79 61 285 6.3 84 7 18  
## 80 79 187 5.1 87 7 19  
## 81 63 220 11.5 85 7 20  
## 82 16 7 6.9 74 7 21  
## 83 NA 258 9.7 81 7 22  
## 84 NA 295 11.5 82 7 23  
## 85 80 294 8.6 86 7 24  
## 86 108 223 8.0 85 7 25  
## 87 20 81 8.6 82 7 26  
## 88 52 82 12.0 86 7 27  
## 89 82 213 7.4 88 7 28  
## 90 50 275 7.4 86 7 29  
## 91 64 253 7.4 83 7 30  
## 92 59 254 9.2 81 7 31  
## 93 39 83 6.9 81 8 1  
## 94 9 24 13.8 81 8 2  
## 95 16 77 7.4 82 8 3  
## 96 78 NA 6.9 86 8 4  
## 97 35 NA 7.4 85 8 5  
## 98 66 NA 4.6 87 8 6  
## 99 122 255 4.0 89 8 7  
## 100 89 229 10.3 90 8 8  
## 101 110 207 8.0 90 8 9  
## 102 NA 222 8.6 92 8 10  
## 103 NA 137 11.5 86 8 11  
## 104 44 192 11.5 86 8 12  
## 105 28 273 11.5 82 8 13  
## 106 65 157 9.7 80 8 14  
## 107 NA 64 11.5 79 8 15  
## 108 22 71 10.3 77 8 16  
## 109 59 51 6.3 79 8 17  
## 110 23 115 7.4 76 8 18  
## 111 31 244 10.9 78 8 19  
## 112 44 190 10.3 78 8 20  
## 113 21 259 15.5 77 8 21  
## 114 9 36 14.3 72 8 22  
## 115 NA 255 12.6 75 8 23  
## 116 45 212 9.7 79 8 24  
## 117 168 238 3.4 81 8 25  
## 118 73 215 8.0 86 8 26  
## 119 NA 153 5.7 88 8 27  
## 120 76 203 9.7 97 8 28  
## 121 118 225 2.3 94 8 29  
## 122 84 237 6.3 96 8 30  
## 123 85 188 6.3 94 8 31  
## 124 96 167 6.9 91 9 1  
## 125 78 197 5.1 92 9 2  
## 126 73 183 2.8 93 9 3  
## 127 91 189 4.6 93 9 4  
## 128 47 95 7.4 87 9 5  
## 129 32 92 15.5 84 9 6  
## 130 20 252 10.9 80 9 7  
## 131 23 220 10.3 78 9 8  
## 132 21 230 10.9 75 9 9  
## 133 24 259 9.7 73 9 10  
## 134 44 236 14.9 81 9 11  
## 135 21 259 15.5 76 9 12  
## 136 28 238 6.3 77 9 13  
## 137 9 24 10.9 71 9 14  
## 138 13 112 11.5 71 9 15  
## 139 46 237 6.9 78 9 16  
## 140 18 224 13.8 67 9 17  
## 141 13 27 10.3 76 9 18  
## 142 24 238 10.3 68 9 19  
## 143 16 201 8.0 82 9 20  
## 144 13 238 12.6 64 9 21  
## 145 23 14 9.2 71 9 22  
## 146 36 139 10.3 81 9 23  
## 147 7 49 10.3 69 9 24  
## 148 14 20 16.6 63 9 25  
## 149 30 193 6.9 70 9 26  
## 150 NA 145 13.2 77 9 27  
## 151 14 191 14.3 75 9 28  
## 152 18 131 8.0 76 9 29  
## 153 20 223 11.5 68 9 30

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 variables 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 it can be seen that 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()

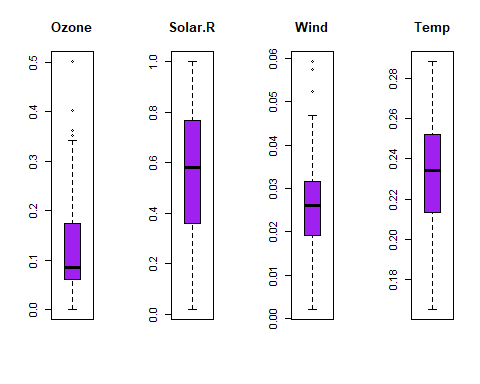
library(corrplot)

## corrplot 0.84 loaded

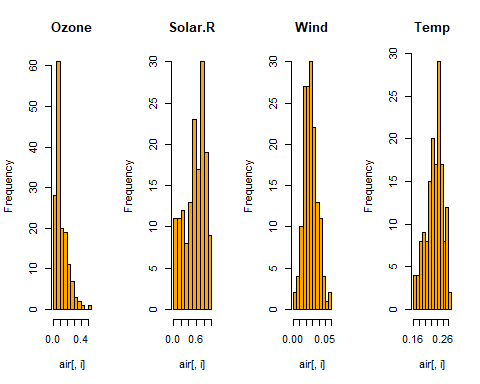
univariate analysis

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. There can be va Histogram

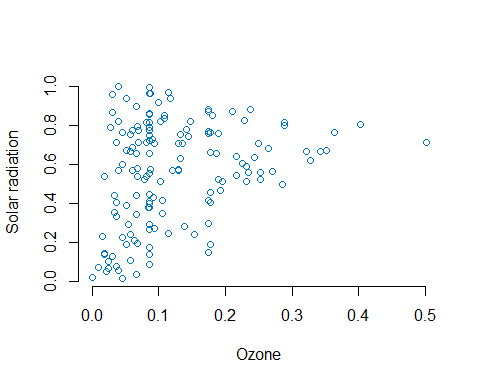
par(mfrow=c(1,4))  
for(i in 1:4) {  
 hist(air[,i], main=names(air)[i],  
 col = c("orange"))  
}



Multivariate Analysis

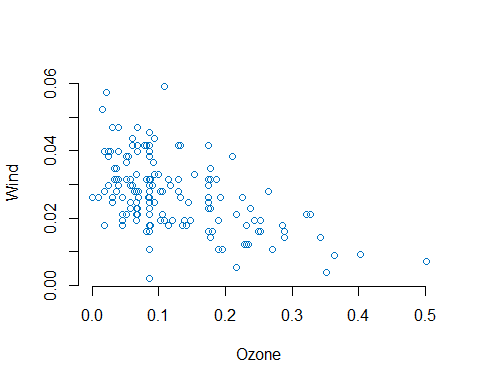
In multi variate analysis we would be using scatter plot to do multivariate analysis our first model is Ozone vs Solar radiation scatter box.

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

 Model 2

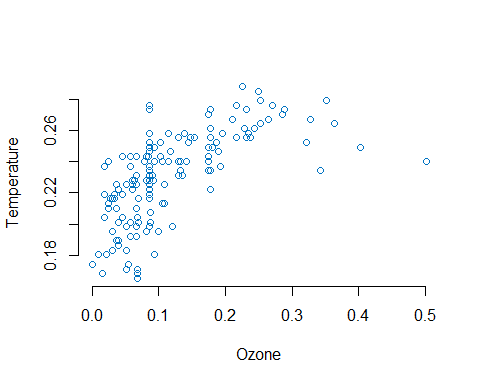
Ozone vs Wind

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



Model 3 Ozone vs Temperature

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



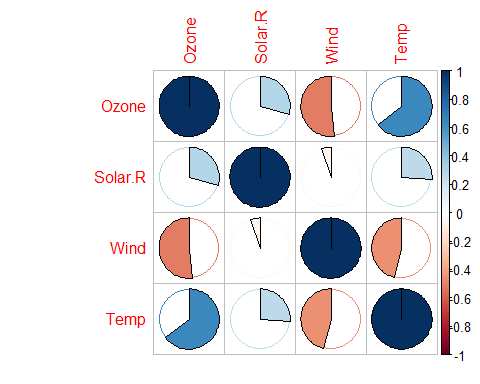
Correlation

The following table and plot shows 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

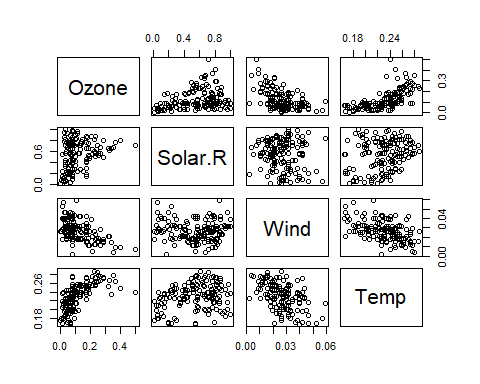
corrplot(cor(air[,1:4]), method = "pie")



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

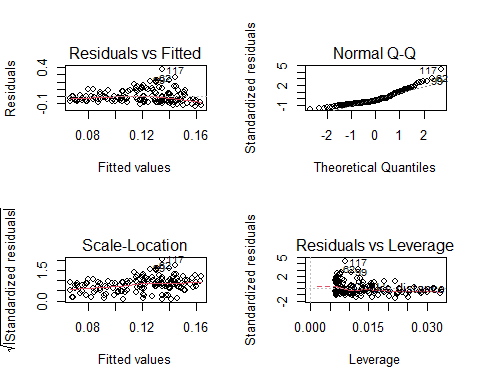
While progressing through linear regression we will use forward selection Method. Our first model will have Ozone as dependent variable and Solar radiation as independent variable.

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

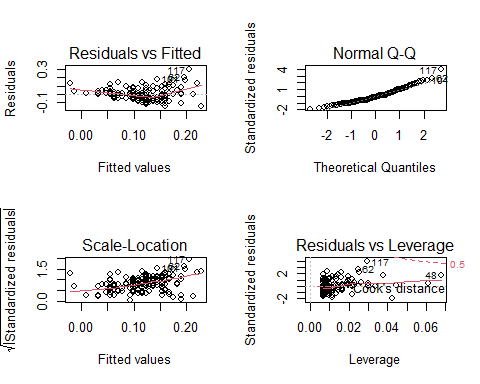
From the model above we can see that every single percentage increase in Solar radiation our Ozone increases by 0.098.

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.
8. 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

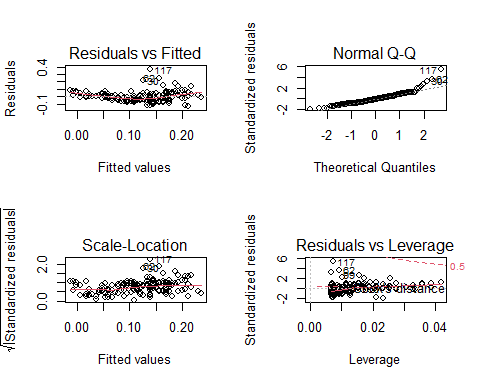
1. Residual is near 0 which means Ozone to Wind residual is symmetrical.
2. The average wind speed is 0.236 Units to that of Ozone.
3. Every 1 unit increase in Wind speed the Ozone decreases by 4.43 Unit and vice versa.
4. If we re run the model there can be diffrence of 0.016 Units of Ozone.
5. Our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

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)

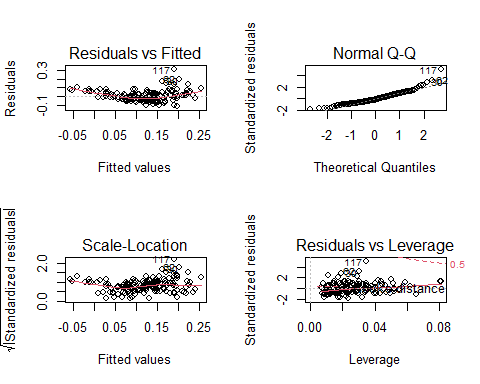
##   
## Call:  
## lm(formula = Ozone ~ Temp, data = air)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.12590 -0.04709 -0.00644 0.03172 0.36293   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.3464 0.0452 -7.664 2.03e-12 \*\*\*  
## Temp 2.0187 0.1943 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.

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. resudual 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

Conclusion

Through this Analysis we did 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 bivariate analysis. We learned to build histogram and boxplots for variables. We learned to plot correlation plots.