19MSMS26 (Anshika Saxena)

**Part A:-**

I am using 2nd dataset i.e. air quality dataset in datasets library for my analysis.

**Data Description:-**

This data has Daily air quality measurements of New York which was recorded for May to September 1973.This data contains 153 observations for 6 variables.

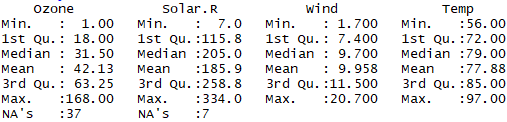
**Aim: -** The aim of this analysis is to study the relationship mean ozone concentration with Solar.R, Wind and temperature. We are also interested if there is any seasonal effect present on variables Ozone, Solar.R, Wind, and Temperature with respect to categorical variable Month.

**Data Exploration:-**

The description of observed variables is as follows-

| **Variable Name** | **Variable Type** | **Unit** | **Description** |
| --- | --- | --- | --- |
| Ozone | numeric | parts per billion | mean Ozone concentration |
| Solar.R | numeric |  | Solar Radiation |
| Wind | numeric | miles per hour | average wind speed |
| Temp | numeric | Fahrenheit | maximum daily temperature |
| Month | Categorical(level=5) |  | Month of observation |
| Day | Categorical(level=31) |  | day of month |

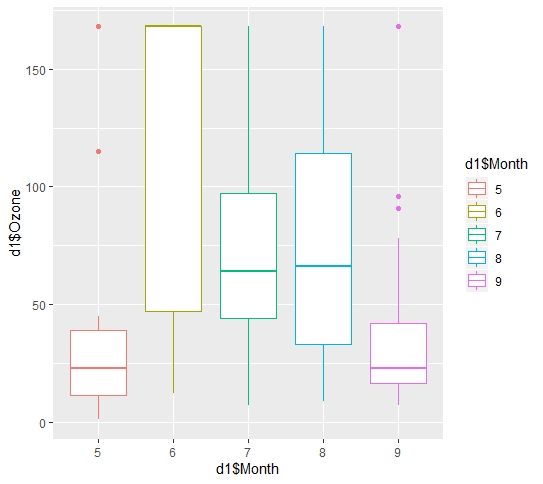
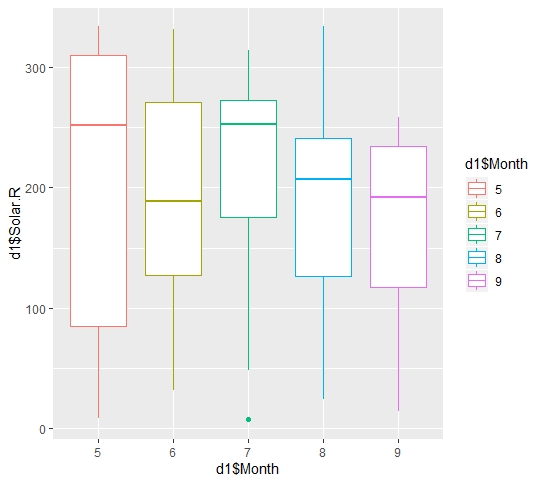
Let’s look at the descriptive summary of dataset for checking if any missing values or null values are present in the dataset:

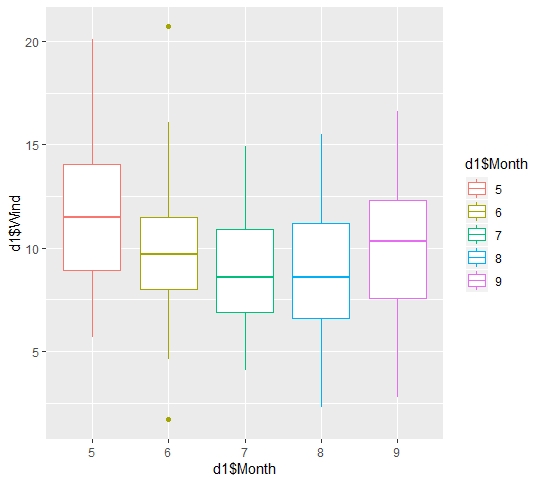
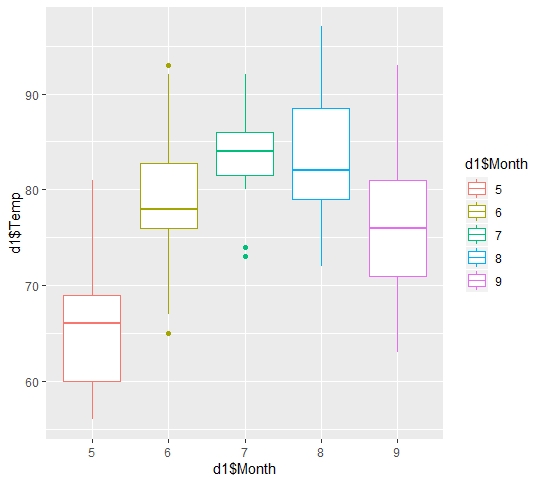


Here we can see that data has some missing values for factors Ozone and Solar.R respectively.

One more thing to notice that there is huge difference between the median value and the maximum value for the factors Ozone, Solar.R. This indicates that some outliers (extreme values) may present in the data.

We are interested in knowing that is there any season effect present on variables Ozone, Solar.R, Wind and temperature with respect of categorical variable month?

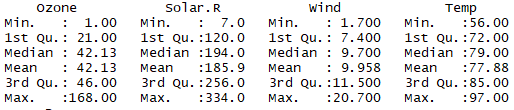




So the above three plots showed that there is seasonal effect present on variables Ozone, Solar.R, temperature and wind with respect of categorical variable Month.

Missing data imputation:-

As we have seen earlier that data contains some missing values for two variables, so the suitable treatments for missing values are either we can remove these values from data or we can replace these values with some suitable value. I have replaced missing values with mean value for each variable containing missing values.



So from the above table we observed that missing values are handled and now, there is no missing value in the dataset .Notice that the median value for both the variables is different from earlier reported value because we replaced missing values with mean value.

Fitting Regression Model:-

Since we are interested in studying the relationship of mean ozone concentration with the other variables so the correlation coefficients between ozone and other variables are as follows-

Solar.R Wind Temp

Ozone 0.30296951 -0.53093584 0.6087420

So the multiple linear regression model will be fitted for analysis. Population model is -

**Y=**

Where,

Y is ozone

x1 is Solar.R, x2 is Wind, x3 is emp.

.

To fit the regression model, we used method of Least Square to estimate the intercept and regression coefficients. And the fitted regression model based on data is:-

^

**Model 1**:-Y= -38.22315 + 0.05775\*x1 -2.71725\*x2 + 1.24126\*x3

After checking the model adequacy of model 1 we found that the errors were not following normal distribution. So we used the square root transformation on Y and refitted the model as model2.

^

**Model 2**:-Y1= 0.005220\*x1-0.198067\*x2+ 0.091741\*x3

Where,

Y1=

**Significance of Intercept and regression coefficients:-**

**For model1**

95% confidence interval for β0 is (-0.90934359,-75.53696551).So even if all the regressors take value zero then also mean ozone concentration cannot be negative, but this interval does not include value zero. Hence, β0 is insignificant.

As the p-value is much less than 0.05 for β1, β2 and β3.  Hence there is a significant relationship between Y and the variables x1, x2 and x3.

**Refitted Reduced Model is-**

^

Y= 0.05998\*x1 -3.45138\*x2 +0.84295\*x3

**For model 2:-**

As the p-value is much less than 0.05 for β1, β2 and β3.Hence there is a significant relationship between Y and the variables x1, x2 and x3.

Model adequacy check:-

**For model 1**-

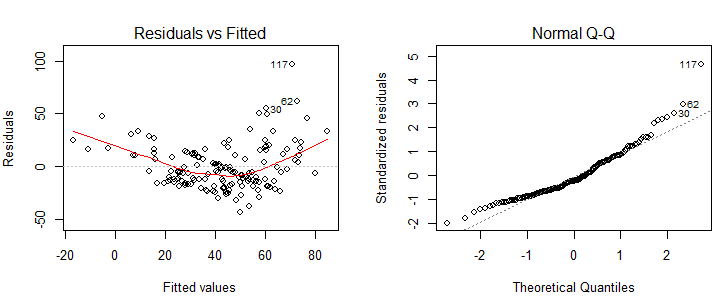
The percentage of variance explained by regressors can be measured using coefficient of determination.

And adjusted ==0.8281

Where,

N->total sample size

p->number of regressors

Which means this model captures 82% variation of data and suggests a good fit. But we should not conclude anything about goodness of model and do the residual analysis.

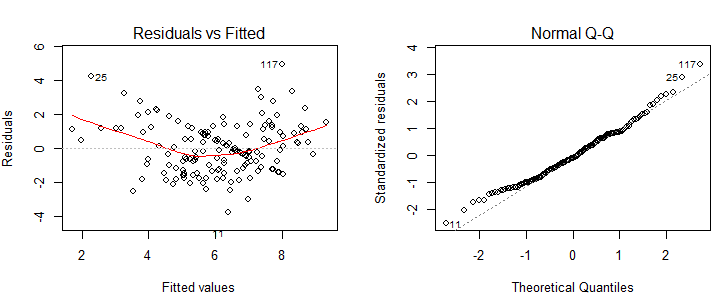
Residual vs. fitted plot is showing non linear pattern and from the qq–plot we observed that there are some observations which do not lie on the normal qq-plot line. And from the Shapiro-test with test statistic W = 0.92785, p-value = 5.667e-07, we confirmed that errors are not following normality assumption. Now we need use some transformation and refit the model.

**Model 2**:-

The coefficient of determination for this model = 0.9489

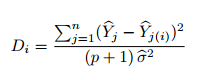
And the adjusted == 0.9479

So this model suggests a better fit than model 1 and explained about 94% variation of data. But we should do residual analysis to get a clear picture.



So Residual vs fitted plot does not show any discrepancy now and is almost constant band around residual=0 line, which means errors has constant variance. And errors seem to lie on the normal line approximately. And from the Shapiro-test with test-statistics W = 0.9835 and p-value = 0.06433,we confirmed that errors are normal. Hence this model is satisfying all the assumptions.

Outliers Diagnostics for model 2:-

 Cook’s distance is calculated by removing the ith data point from the model and refitting the regression. **It summarizes how much all the values in the regression model change when the ith observation is removed.**The formula for Cook’s distance is:  
[](https://www.statisticshowto.com/wp-content/uploads/2014/12/cooks-d.png)

Where,

^

Yj->predicted value of j-th observation, when ith observation is included

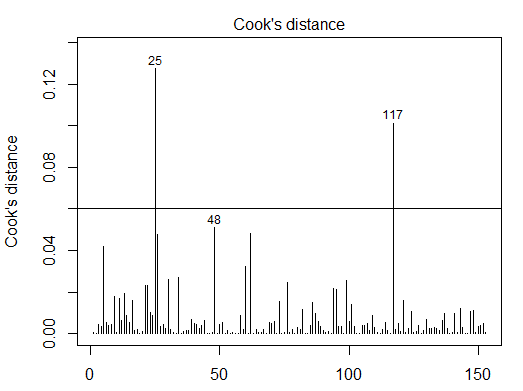
^

Yj (i) ->predicted value of j-th observation, when i-th observation is excluded.

P->number of regressors

^

(Estimator of error variance (Mean square error)) = 0.009873333



So I have taken the cutoff point as 0.06 because almost all the distances fall below this point. There are two values for which cook’s distance is outside this cutoff line and huge. Hence these observations are influential points (outliers).and If we remove these two observations from data and refit the model then adjusted = 0.9531 which greater than model 2.

Conclusion:-

The regression model2 is one adequate model for this data, since it does not violate any necessary assumptions of multiple linear regression. It has the power to explain most of variability in the data. The regression coefficients are significant and most of the variability for Ozone can be explained by these regressors. There are two outliers which are influential and have high cook’s distance compare to rest of the observations in data. But these observations do not impact the model in larger extent and adjusted differs by less than 1%.

**Part B:-**

**2. Problem on Solar data**

1. Population regression model:-

Where,

Y-> total heat flux y (kilowatts)

->intercept

->coefficient of X4

−> The radial deflection of the deflected rays

To fit the model we used method of least square estimates for estimating and the estimates are

^

= 607.103

^

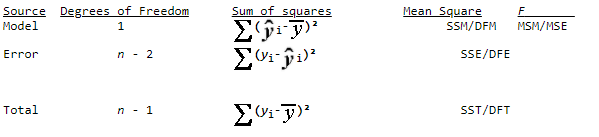
= -21.402

The fitted simple linear regression model based on x4 is-

^

yi=607.103 -21.402 \*x4i; i=1(1) n

(b)Anova table for the simple linear regression model is as follows:-



Where,

Yi->i-th observation

^

Yi->predicted value for i-th observation

y\_bar->sample mean of Y

And computed anova table for the fitted model is

Df Sum Sq Mean Sq F value

model 1 10578.7 10578.7 69.609

Error 27 4103.2 152

Total 28 5161.9

Test for Significance of regression:-

For testing the hypothesis that

H0:-β1 = 0

Against H1:β1= 0

The "F" column provides a statistic i.e. F=69.609>F-tab (1, 27, 0.05) =4.21.Hence we reject null hypothesis and conclude that the fitted regression model is significant.

(c).

A 100(1 − 𝛼) % confidence interval will be estimated for 𝛽1 using the formula



Where,

J=1(1) p

n->number of observations

p->number of regressors

^

Standard error (= 2.565

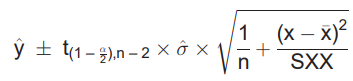
α->significance level(0.05)

Hence a 95% confidence interval on the slope is (-16.1390,-26.66592).

(d)

0.7205(computed using anova table)

(e) A 100(1 – α ) % confidence interval for the mean response given x is calculated using this expression,

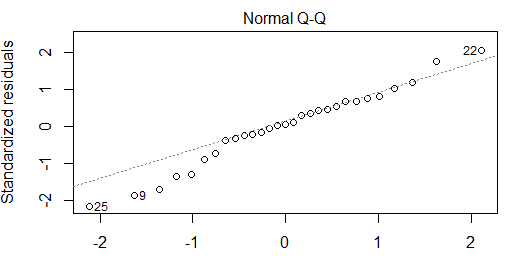


Here,

=

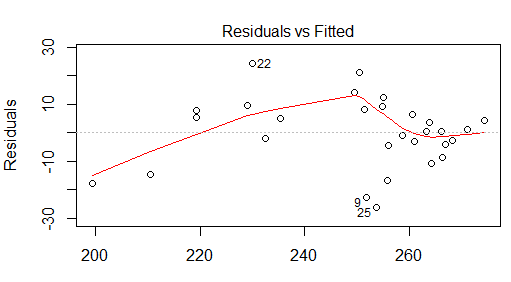
Hence 95% confidence intervals on the mean heat flux when the radial deflection is 16.5 milli-radians (249.1468, 258.7787).

(f)



From this plot we observed that some observations are deviation from the normal probability line. But data is small so we cannot be 100% sure from looking at this plot that errors are not normal. We should confirm about normality of errors assumption using any normality test, since data has very less observations.

(g)



In the above plot, this pattern is not indicating any non-linear trend but we observed that this pattern is indicating that variance of the errors is not constant. So the homoscedasticity assumption for errors does not seem to satisfy.