**Lab 4**

Name: Jesto Peter

Title: Multiple Linear Regression

Date: 09/12/2022

Class: 2MSTAT

* **Introduction**

Linear regression makes several assumptions about the data, such as :

1. **Linearity of the data**. The relationship between the predictor (x) and the outcome (y) is assumed to be linear.
2. **Normality of residuals**. The residual errors are assumed to be normally distributed.
3. **Homogeneity of residuals variance**. The residuals are assumed to have constant variance (**homoscedasticity**)
4. **Independence of residuals error terms**.

You should check whether or not these assumptions hold true. Potential problems include:

1. **Non-linearity** of the outcome - predictor relationships
2. **Heteroscedasticity**: Non-constant variance of error terms.
3. **Presence of influential values** in the data that can be:
   * Outliers: extreme values in the outcome (y) variable
   * High-leverage points: extreme values in the predictors (x) variable

All these assumptions and potential problems can be checked by producing some diagnostic plots visualizing the residual errors.

* **Objective**

Select a suitable data of your choice on Agriculture.

* 1. Check the data for general linear model assumptions
  2. Fit a linear regression model
  3. Draw the residual plots and interpret them.
* **Procedure**

#importing the dataset  
#The data set is the yield of Rice in Karnataka ,Kodagu District over the years  
agri=read.csv("D:/Regression Analysis/4 LAB (9TH DEC)/district.csv")  
agri

## Dist.Code Year State.Code State.Name Dist.Name RICE.YIELD..Kg.per.ha.  
## 1 82 1966 5 Karnataka Kodagu / Coorg 1604.65  
## 2 82 1967 5 Karnataka Kodagu / Coorg 1454.55  
## 3 82 1968 5 Karnataka Kodagu / Coorg 1891.89  
## 4 82 1969 5 Karnataka Kodagu / Coorg 1726.68  
## 5 82 1970 5 Karnataka Kodagu / Coorg 1634.57

#Only considering necessary columns  
agr=agri[-c(1:5)]  
colnames(agr)=c("Y","X1","X2")  
head(agr)

## Y X1 X2  
## 1 1604.65 43.0 69.0  
## 2 1454.55 44.0 64.0  
## 3 1891.89 44.4 84.0  
## 4 1726.68 46.1 79.6  
## 5 1634.57 45.7 74.7  
## 6 1939.91 46.6 90.4

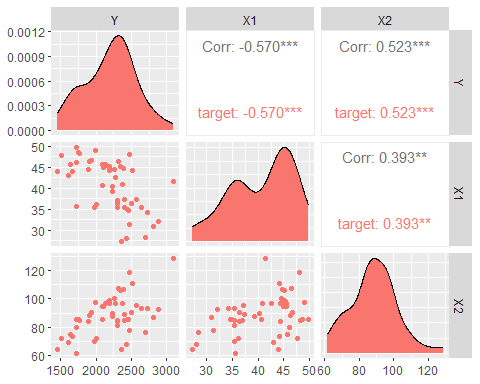
**Use of Multiple Linear Regression**

Here Yield(Y) is the Dependent Variable and the other 2 columns are the regressors. We will be using MLR for the Agricultural Dataset to predict and see how the Yield of the Crop is affected by the Production and Area of the Rice Crop production.

#Finding the correlation matrix and identifying the highly correlated variables.  
library(GGally)

ggpairs(agr,aes(colour='target'),pch=21,main="Yield of Rice")

## Warning in warn\_if\_args\_exist(list(...)): Extra arguments: "pch", "main" are  
## being ignored. If these are meant to be aesthetics, submit them using the  
## 'mapping' variable within ggpairs with ggplot2::aes or ggplot2::aes\_string.



#Fitting the model  
mlr=lm(Y~.,agr)  
mlr

##   
## Call:  
## lm(formula = Y ~ ., data = agr)  
##   
## Coefficients:  
## (Intercept) X1 X2   
## 2366.58 -54.95 23.44

#Finding the summary of the model  
summary(mlr)

##   
## Call:  
## lm(formula = Y ~ ., data = agr)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -124.978 -22.709 -8.862 14.703 110.846   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2366.5794 50.0748 47.26 <2e-16 \*\*\*  
## X1 -54.9452 1.1289 -48.67 <2e-16 \*\*\*  
## X2 23.4392 0.5001 46.87 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 44.98 on 49 degrees of freedom  
## Multiple R-squared: 0.9853, Adjusted R-squared: 0.9847   
## F-statistic: 1639 on 2 and 49 DF, p-value: < 2.2e-16

#**Interpretation:** We see that the R2 Squared value is 0.9853, hence we can see it is a very good model.

We see from the correlation matrix that there is a weak correlation between the independent variables. Hence we find the VIF values to see if we can include the variables X1 and X2 in the model.

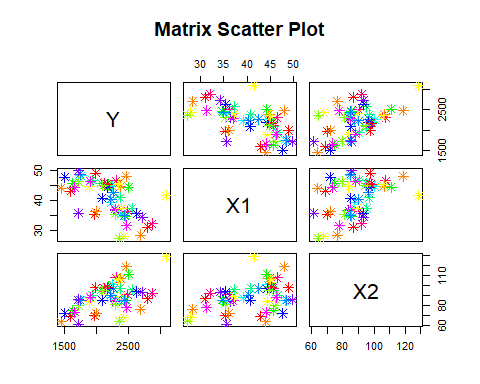
library("mctest")  
imcdiag(mlr)

##   
## Call:  
## imcdiag(mod = mlr)  
##   
##   
## All Individual Multicollinearity Diagnostics Result  
##   
## VIF TOL Wi Fi Leamer CVIF Klein IND1 IND2  
## X1 1.1831 0.8453 9.1525 Inf 0.9194 0.0434 0 0.0169 1  
## X2 1.1831 0.8453 9.1525 Inf 0.9194 0.0434 0 0.0169 1  
##   
## 1 --> COLLINEARITY is detected by the test   
## 0 --> COLLINEARITY is not detected by the test  
##   
## \* all coefficients have significant t-ratios  
##   
## R-square of y on all x: 0.9853   
##   
## \* use method argument to check which regressors may be the reason of collinearity  
## ===================================

**Interpretation:** We see that VIF values are less than 5 , so X1 and X2 can be included.

1. **Checking the Data for the general linear model assumptions.**

#1. Checking Linearity between Y and X1,X2  
pairs(agr ,col = rainbow(c(12)),pch =8,cex = 1.5,main="Matrix Scatter Plot")



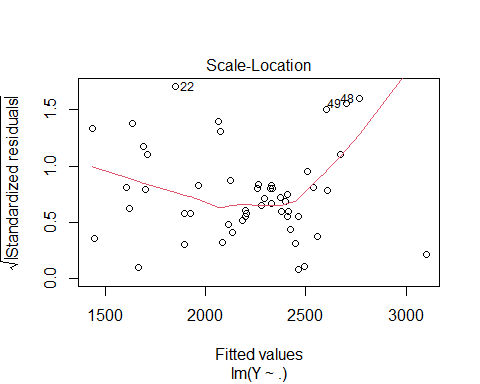
#**Interpretation**: We see Y and X1 are Negatively Correlated. Y and X2 is positively related.

###########  
#2. Checking the E(e)=0  
  
#if the sum of error terms terms =0 then expected value of error =0  
e=residuals(mlr)  
sum(e)

## [1] 2.087219e-14

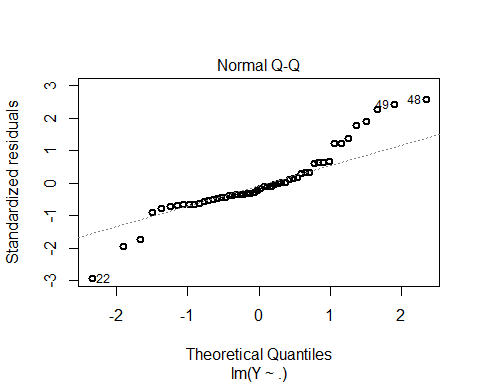
#**Interpretation**: Since sum of error term is approximately 0 ,we have E(e)=0

###########  
#3. Checking Homoscedasticity i.e V(e)= constant  
plot(mlr,which=3)



#**Interpretation**: We see that the observations are equally distributed above and below the regression line, hence it has a constant Variance

###########  
#4.The error terms follow Normal(0,v(X))  
  
plot(mlr,which=2,lwd=2)



#**Interpretation**: We see that the ovservation lie almost on the regression line and hence they follow normality.

#5.The error terms are not CORELATED  
attach(agr)  
#finding error terms  
e=residuals(mlr)# error  
  
#Errors are not correlated if Sum of (error\*Independent Variable)=0  
sum(e\*X1)#approximately 0

## [1] 1.866951e-12

sum(e\*X2) #approximately 0

## [1] 5.684342e-12

#**Interpretation**: Hence ei's are uncorrelated.  
#Also if we have our model as Homoscedasstic then automatically we have error terms to be uncorrelated

**2. Fitting a Linear Regression Model**

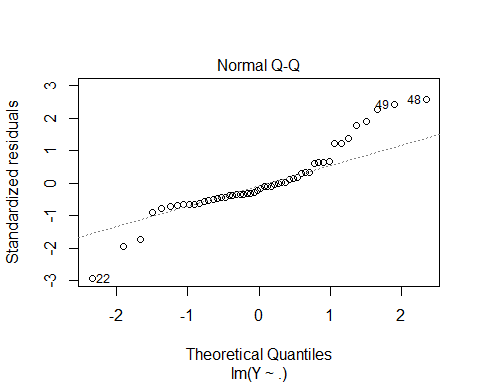
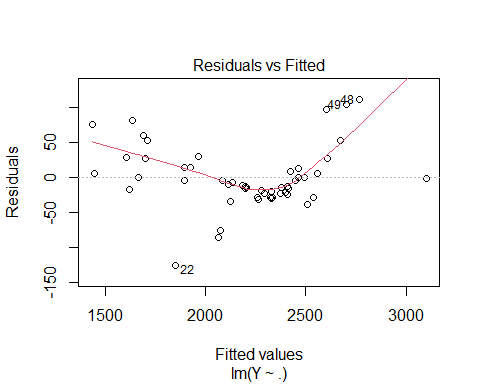
#Fitting the model  
mlr=lm(Y~.,agr)  
mlr

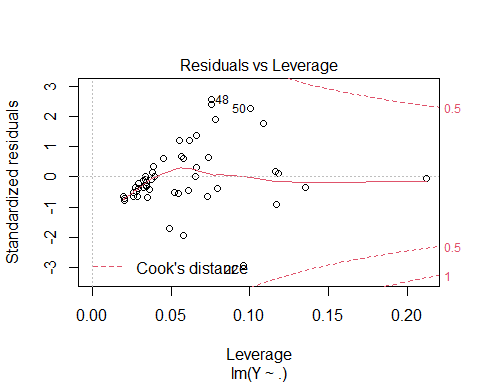
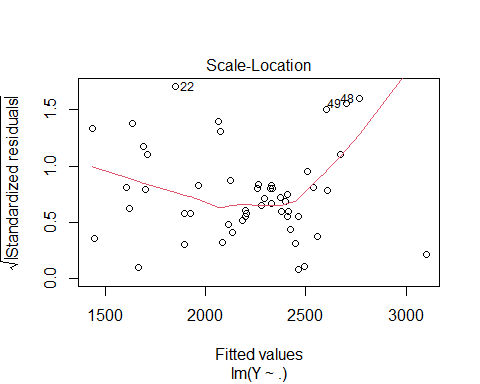
##   
## Call:  
## lm(formula = Y ~ ., data = agr)  
##   
## Coefficients:  
## (Intercept) X1 X2   
## 2366.58 -54.95 23.44

**Conclusion:** The Coefficients are B0=2366.58, B1=-54.95, B2=23.44

**3. Drawing the residual plots and interpretting them.**

plot(mlr)





* **Residuals Vs Fitted** graph tells us that there is a constant variance and that linearity exists, since the data is distributed approximately above and below the regression line
* **Normal Q-Q Graph** is used for testing the Assumption for Normality of Errors. It uses Quantiles to plot the graph. The errors follow Normal Distribution in our case as the line and plotted points almost coincide. So the errors follow normal distribution.
* **Scale-Location** graph tells us whether the variance is constant or not. Since all points are approximately above or below the graph, so it is constant
* **Residuals Vs Leverage** plot tells us about the influential points in the graph. We see that 22nd, 48th and 50th observed values are the most influenced.
* **Conclusion**

In this lab session, we took a data set that showed the yield of Rice in Karnataka, Kodagu District, over the years.

We had the dependent variable as “Yield” and the regressors as “Production” and “Area of Production”. We saw from the correlation matrix that the dependent variables (Yield) were moderately related to the regressor “Production” in a positive way and negatively related to the “Area of Production”.

We thereafter checked the assumptions for General Linear Model, which we found was approximately satisfied.

We fitted the linear model for our dataset and finally plotted the Residual plots and interpreted their conclusions.