**Lab 6**

Name: Jesto Peter

Title: Multiple Linear Regression

Date: 06/01/2023

Class: 2MSTAT

* **Objective**

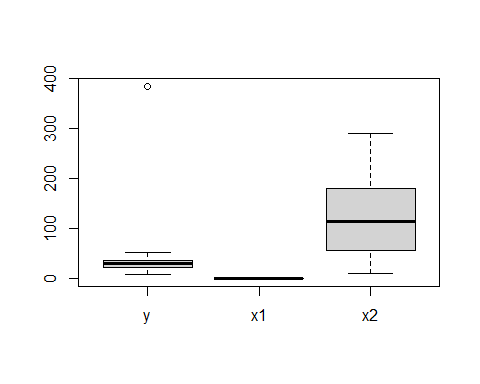
Use the dataset which are available in the text book of *Introduction to Linear Regression Analysis by D.C. Montgomery, E. A. Peck and G. G. Vining, a*nd  
  
1. Fit a linear regression model by identifying the most significant variables in the data.  
2. How do you test for the autocorrelation?

* **Procedure**

cal=read.csv("D:/Regression Analysis Practical/6 LAB (6TH JAN)/data.csv")  
head(cal)

## Clathrate.formation Amount.of.surfactant Time  
## 1 7.5 0.00 10  
## 2 15.0 0.00 50  
## 3 22.0 0.00 85  
## 4 28.6 0.00 110  
## 5 31.6 0.00 140  
## 6 34.0 0.00 170

**#CHECKING FOR OUTLIERS**  
colnames(cal)=c('y','x1','x2')  
  
boxplot(cal)



#Finding the specific Outlier  
boxplot(cal,plot=FALSE)$out

## [1] 385

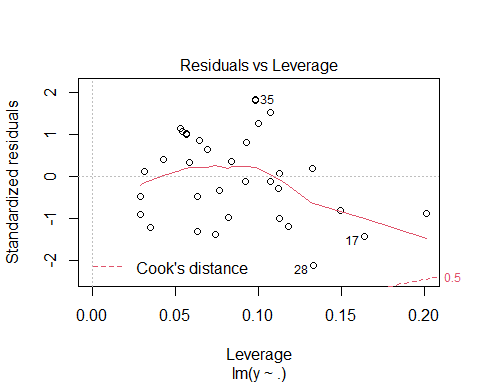
#Outlier is 385  
  
#Removing the outlier  
cal1=subset(cal, y!=385)  
  
boxplot(cal1,plot=FALSE)$out no more outliers in dataset

**#CHECKING FOR MULTICOLLINEARITY USING VIF values**  
mlr=lm(y~.,cal1)  
  
library(mctest)  
imcdiag(mlr)

##   
## Call:  
## imcdiag(mod = mlr)  
##   
##   
## All Individual Multicollinearity Diagnostics Result  
##   
## VIF TOL Wi Fi Leamer CVIF Klein IND1 IND2  
## x1 1.0064 0.9937 0.2108 Inf 0.9968 0.7086 0 0.0301 1  
## x2 1.0064 0.9937 0.2108 Inf 0.9968 0.7086 0 0.0301 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: 0.8437   
##   
## \* use method argument to check which regressors may be the reason of collinearity  
## ===================================

#Conclusion:We see from the imcdiag test that there is no multicollinearity between the independent variables.

**#CHECKING FOR COOKS DISTANCE FOR OUTLIERS**plot(mlr,which=5)



#Conclusion: No outliers exist

**#CHECKING FOR SIGNIFICANCE OF THE REGRESSORS**  
summary(mlr)

##   
## Call:  
## lm(formula = y ~ ., data = cal1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.4229 -4.0402 0.2912 3.8708 8.3321   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.88432 1.68498 6.460 2.89e-07 \*\*\*  
## x1 345.79294 39.99520 8.646 7.02e-10 \*\*\*  
## x2 0.11245 0.01065 10.554 5.96e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.789 on 32 degrees of freedom  
## Multiple R-squared: 0.8437, Adjusted R-squared: 0.8339   
## F-statistic: 86.35 on 2 and 32 DF, p-value: 1.272e-13

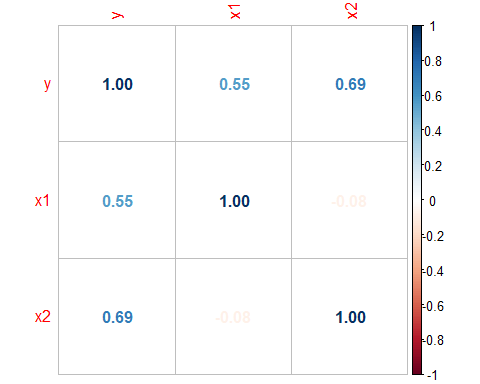
#From p-value of the summary table,we observe that the regreesors are highly significant for the dependent variable.  
#Here both x1 and x2 are Highly Significant  
#The Adjusted R^2 values are really high and good

**#CHECKING THE CORRELATION BETWEEN VARIABLES**library(corrplot)

## Warning: package 'corrplot' was built under R version 4.1.3

## corrplot 0.92 loaded

cor1=round(cor(cal1),2)  
corrplot(cor1,method = "number")



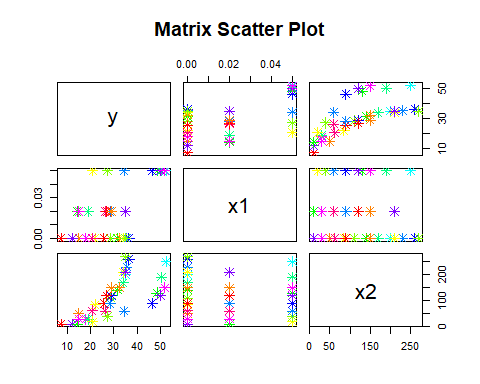
#We observe that there is a moderate correlation between the dependent variable and regressors

**#CHECKING FOR ALL ASSUMPTIONS**  
**#1.NORMALITY OF ERROR TERMS**  
e=residuals(mlr)  
shapiro.test(e)

##   
## Shapiro-Wilk normality test  
##   
## data: e  
## W = 0.97255, p-value = 0.517

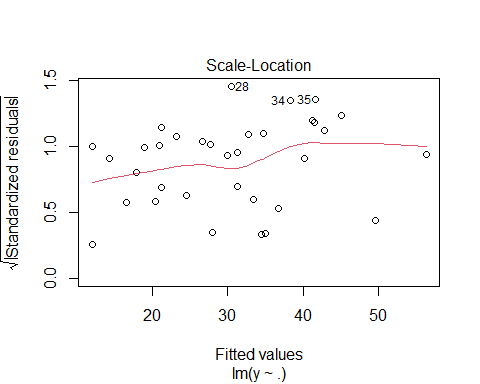
#Conclusion:We see that the p value is 0.2408 which is greater than 0.05 hence we can ACCEPT THE Null Hypothesis, hence the residuals follow Normal Distribution.

**#2. LINEARITY OF VARIABLE**  
pairs(cal1,col = rainbow(c(12)),pch =8,cex = 1.5,main="Matrix Scatter Plot")



#Conclusion: Linearity exists

#**3.CHECKING FOR CONSTANT VARIANCE**  
plot(mlr,which=3)



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

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

## [1] -1.238469e-16

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

**#5. Autocorrelation**  
  
library(lmtest)

## Warning: package 'lmtest' was built under R version 4.1.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.1.3

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

dwtest(mlr)

##   
## Durbin-Watson test  
##   
## data: mlr  
## DW = 0.79902, p-value = 1.215e-05  
## alternative hypothesis: true autocorrelation is greater than 0

**#Conclusion: Autocorrelation exists in the data, since p<0.05**