**Lab 2**

**Multiple Linear Regression**

Name: Jesto Peter

Title: Multiple Linear Regression

Date: 25/11/2022

Class: 2MSTAT

* **Introduction**

Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows us to estimate how a dependent variable changes as the independent variables change.

Multiple linear regression is used to estimate the relationship between two or more independent variables and one dependent variable. We use multiple linear regression when we want to know how strong the relationship is between two or more [independent variables](https://www.scribbr.com/methodology/independent-and-dependent-variables/#independent) and one dependent variable.

The [correlation](https://www.investopedia.com/terms/c/correlation.asp) coefficient is a statistical measure of the strength of a [linear relationship](https://www.investopedia.com/terms/l/linearrelationship.asp) between two variables. Its values can range from -1 to 1. A correlation coefficient of -1 describes a perfect [negative](https://www.investopedia.com/terms/n/negative-correlation.asp) or [inverse](https://www.investopedia.com/terms/i/inverse-correlation.asp) correlation and a correlation coefficient of +1 describes a perfect positive correlation.

In this particular Lab Session, we will be focusing on the dataset involving 2 Independent variables and 1 Dependent variable.

* **Objective**

In this Lab Session, we use R programming and try to import a dataset of values into R Studio. The dataset contains the scores in the End Semester examination (ESE) and the scores in two preliminary examinations CIA1 and CIA2 for 22 students in a Statistics course.

Here ESE is the Dependent Variable and CIA1 and CIA2 marks are the Independent Variables

We have to:

1. Plot a matrix scatter diagram for the data and find the coefficient of correlation. What do you infer from the scatter plot?
2. Estimate the parameters of a multiple linear regression model and fit a regression line. Interpret the results.
3. Test the significance of the regression coefficients and interpret the results.

* **Procedure**

1. **Plot a matrix scatter diagram for the data and find the coefficient of correlation.**

library(readxl)

data=read\_excel("D:/Regression Analysis/2 LAB (25TH NOV)/Dataset\_MLR.xlsx")

head(data)

## # A tibble: 6 x 6  
## `Student Reg. No.` ESE CIA1 CIA2 ...5 Data Description: ~1  
## <dbl> <dbl> <dbl> <dbl> <lgl> <lgl>   
## 1 2248101 68 78 73 NA NA   
## 2 2248102 75 74 76 NA NA   
## 3 2248103 85 82 79 NA NA   
## 4 2248104 94 90 96 NA NA   
## 5 2248105 86 87 90 NA NA   
## 6 2248106 90 90 92 NA NA   
## # ... with abbreviated variable name  
## # 1: `Data Description:

The above dataset shows the scores in the End Semester examination (ESE) and the scores in two preliminary examinations CIA1 and CIA2 for 22 students in a Statistics course.

attach(data)#converting columns to lists

#We will now remove the unnecessary columns

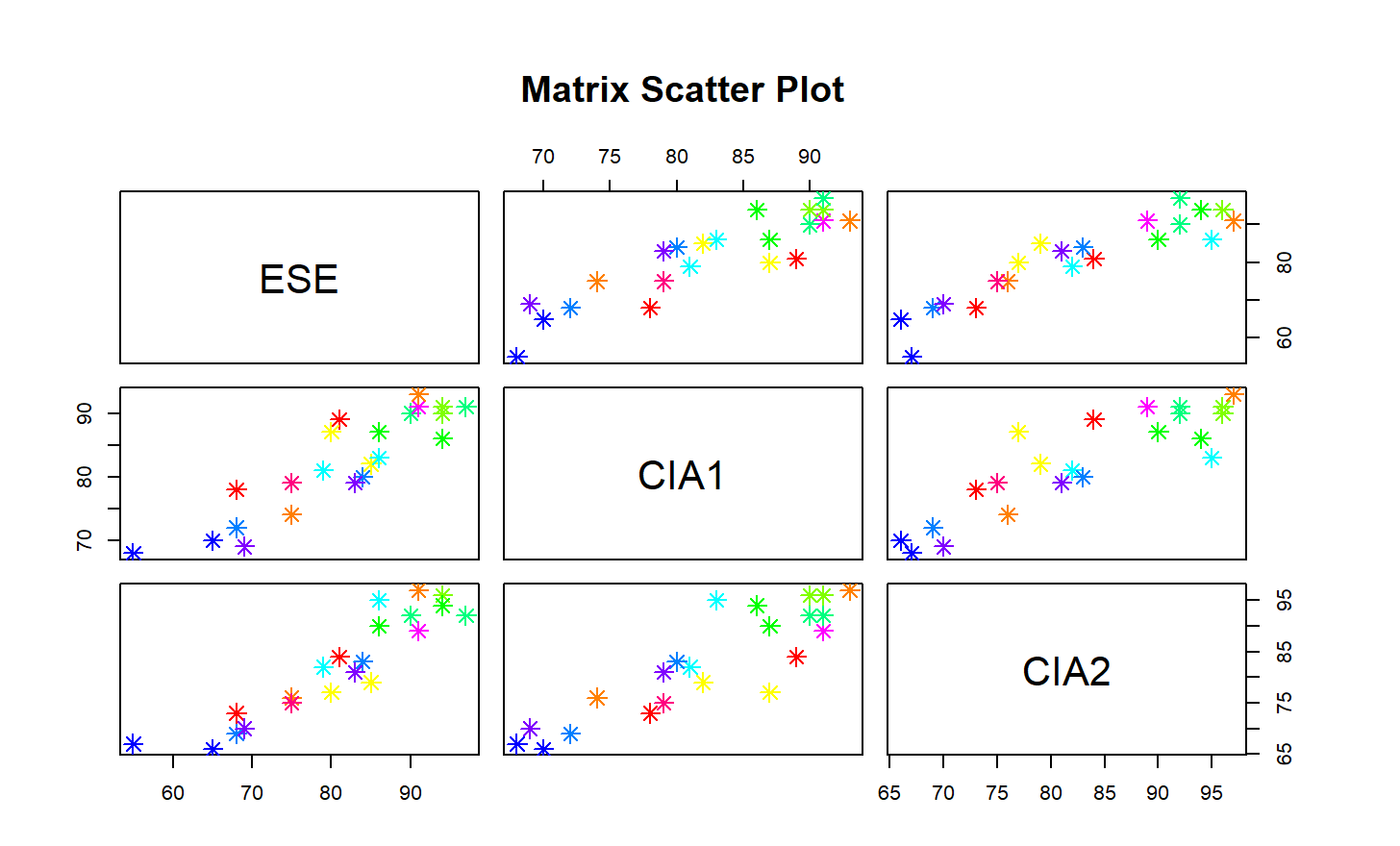
newdata1=data[,-6]

newdata=newdata1[,-5]

n=newdata[-1] #Taking only the 3 necessary columns  
n

## # A tibble: 22 x 3  
## ESE CIA1 CIA2  
## <dbl> <dbl> <dbl>  
## 1 68 78 73  
## 2 75 74 76  
## 3 85 82 79  
## 4 94 90 96  
## 5 86 87 90  
## 6 90 90 92  
## 7 86 83 95  
## 8 68 72 69  
## 9 55 68 67  
## 10 69 69 70  
## # ... with 12 more rows

#Getting a scatter plot matrix  
pairs(n ,col = rainbow(c(12)),pch =8,cex = 1.5,main="Matrix Scatter Plot")



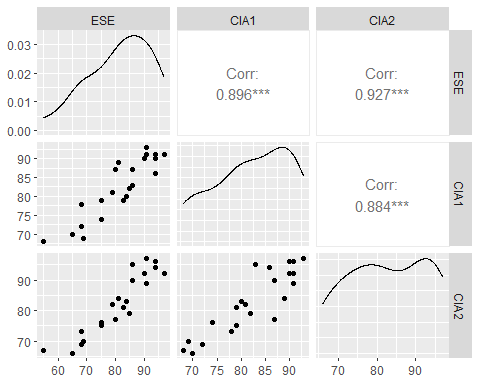
* **Interpretation:** We see that the relationship between all the 3 variable taken 2 at a time has a positive correlation.

#Finding the coefficients of Correaltion  
library(ggplot2)

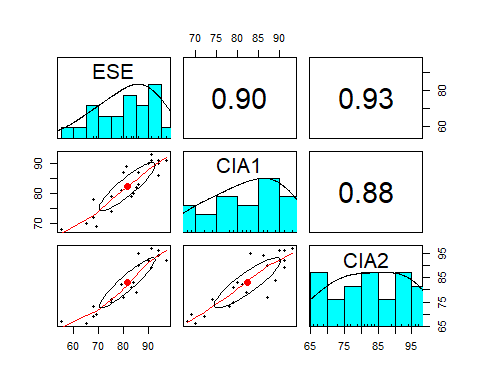
library(GGally)

library(psych)

ggpairs(n)



pairs.panels(n)



summary(model)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.7328 -2.1703 0.3938 2.6443 6.3660   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -14.5005 9.2356 -1.570 0.13290   
## X1 0.4883 0.2330 2.096 0.04971 \*   
## X2 0.6720 0.1793 3.748 0.00136 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.953 on 19 degrees of freedom  
## Multiple R-squared: 0.8863, Adjusted R-squared: 0.8744   
## F-statistic: 74.07 on 2 and 19 DF, p-value: 1.069e-09

* **Interpretation:** We see that there is a Strong Correlation between ESE and CIA1,ESE and CIA2 and, CIA1 and CIA2. Also from the summary we see that the coefficients of correlation obtained is B0=-14.5005 ,X1=0.4883 and X2=0.6720

1. **Estimate the parameters of a multiple linear regression model**

*#Changing the column names*

colnames(n)= c("Y","X1","X2")  
n

## # A tibble: 22 x 3  
## Y X1 X2  
## <dbl> <dbl> <dbl>  
## 1 68 78 73  
## 2 75 74 76  
## 3 85 82 79  
## 4 94 90 96  
## 5 86 87 90  
## 6 90 90 92  
## 7 86 83 95  
## 8 68 72 69  
## 9 55 68 67  
## 10 69 69 70  
## # ... with 12 more rows

*#Forming the model*

model=lm(Y~.,data=n)   
model

##   
## Call:  
## lm(formula = Y ~ ., data = n)  
##   
## Coefficients:  
## (Intercept) X1 X2   
## -14.5005 0.4883 0.6720

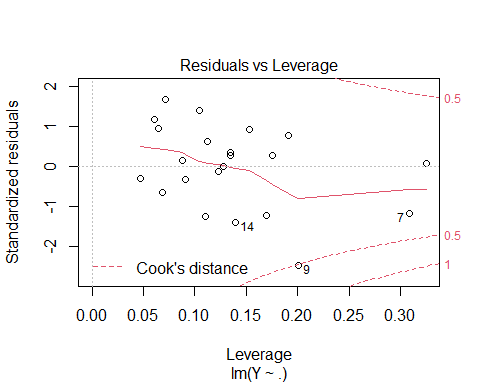
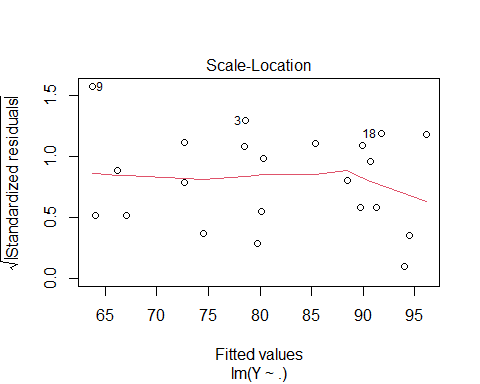
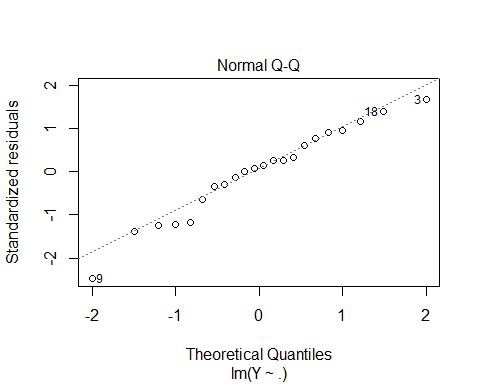
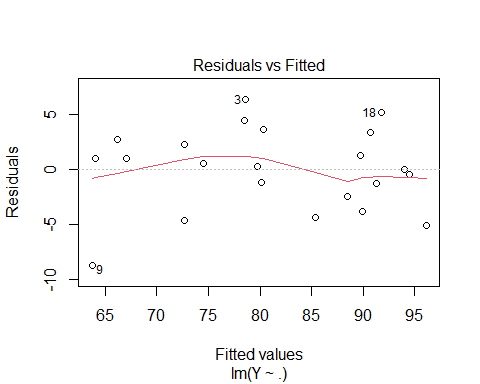
#or  
summary(model)

##   
## Call:  
## lm(formula = Y ~ ., data = n)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.7328 -2.1703 0.3938 2.6443 6.3660   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -14.5005 9.2356 -1.570 0.13290   
## X1 0.4883 0.2330 2.096 0.04971 \*   
## X2 0.6720 0.1793 3.748 0.00136 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.953 on 19 degrees of freedom  
## Multiple R-squared: 0.8863, Adjusted R-squared: 0.8744   
## F-statistic: 74.07 on 2 and 19 DF, p-value: 1.069e-09

#We see that the Adjusted R Squared value of 0.8744 indicates a model that almost perfectly predicts values in the target field.

* **Interpretation:** Hence B0 and Coefficients of X1(CIA1),X2(CIA2) are B0=-14.5005, B1=0.4883, B2=0.6720

#FITTING A REGRESSION LINE  
plot(model)



* **Interpreation:**

#**Residuals Vs Fitted** **graph** tells us that there is a linearity between dependent and independent variables

#**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.

#**Scale-Location graph** tells us whether the variance is constant or not. Since all points are above or below the graph ,so it is constant

#**Residuals Vs Leverage** plot tells us about the infuential points in the graph. We see that 9th, 7th and 14th observed values are the most infuenced.

1. **Testing the significance of Regression Coefficients**

#Testing the significance of the regression coefficients and interpreting the results.  
anova(model)

## Analysis of Variance Table  
##   
## Response: Y  
## Df Sum Sq Mean Sq F value Pr(>F)   
## X1 1 2094.75 2094.75 134.084 4.721e-10 \*\*\*  
## X2 1 219.51 219.51 14.051 0.001361 \*\*   
## Residuals 19 296.83 15.62   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* **Interpretation:**

#For X1, We see that p=4.721e-10 < 0.001, Hence it is statistically significant.

#For X2, We see that p=0.001361 < 0.01 ,Hence it is statistically significant.

* **CONCLUSION**

We saw that there is a positive correlation between all the variables using the scatter plot. Also there is a Strong Correlation between ESE and CIA1,ESE and CIA2 and, CIA1 and CIA2. From the summary we see that the coefficients of correlation obtained is Y=-14.5005 ,X1=0.4883 and X2=0.6720

Also from the summary, we see that the coefficients of correlation obtained is Y=-14.5005 ,X1=0.4883 and X2=0.6720

For X1, we see that p=4.721e-10 < 0.001, hence it is statistically significant.

For X2, We see that p=0.001361 < 0.01 ,Hence it is statistically significant.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_