LAB-5

Linear regression and Multiple Linear Regression

Description:-

Regression analysis can be defined as the process of developing a mathematical model that can be used to predict one variable by using another variable or variables. This section first covers the key concepts of two common approaches to data analysis: *graphical data analysis* and *correlation analysis* and then introduces the two main types of regression: *linear regression* and *non-linear regression*. The section also introduces a number of *data transformations* and explains how these can be used in regression analysis.

When you have worked through this section, you should be able to:

- Distinguish between a dependent variable and an independent variable and analyse data using graphical means.
- Examine possible relationships between two variables using graphical analysis and correlation analysis.
- Develop simple linear regression models and use them as a forecasting tool.
- Understand polynomial functions and use non-linear regression as a forecasting tool.
- Appreciate the importance of data transformations in regression modelling.

Assumptions :-

There are **four principal assumptions** which justify the use of linear regression models for purposes of inference or prediction:

(i) linearity and additivity of the relationship between dependent and independent variables:

(a) The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed.

(b) The slope of that line does not depend on the values of the other variables.

(c) The effects of different independent variables on the expected value of the dependent variable are additive.

(ii) statistical independence of the errors (in particular, no correlation between consecutive errors in the case of time series data)

(iii) homoscedasticity (constant variance) of the errors

(a) versus time (in the case of time series data)

(b) versus the predictions

(c) versus any independent variable

(iv) normality of the error distribution.

If any of these assumptions is violated (i.e., if there are nonlinear relationships between dependent and independent variables or the errors exhibit correlation, heteroscedasticity, or non-normality), then the forecasts, confidence intervals, and scientific insights yielded by a regression model may be (at best) inefficient or (at worst) seriously biased or misleading.

Problem 1: The following table shows the scores (X) of 10 students on Zoology test and scores (Y) on Botony test .The maximum score in each test was 50.Obtain least square equation of line of regression of X on Y. If it is known that the score of a student in Botony is 28,Estimate his/her score in Zoology.

X	34	37	36	32	32	36	35	34	29	35
Y	37	37	34	34	33	40	39	37	36	35

R code

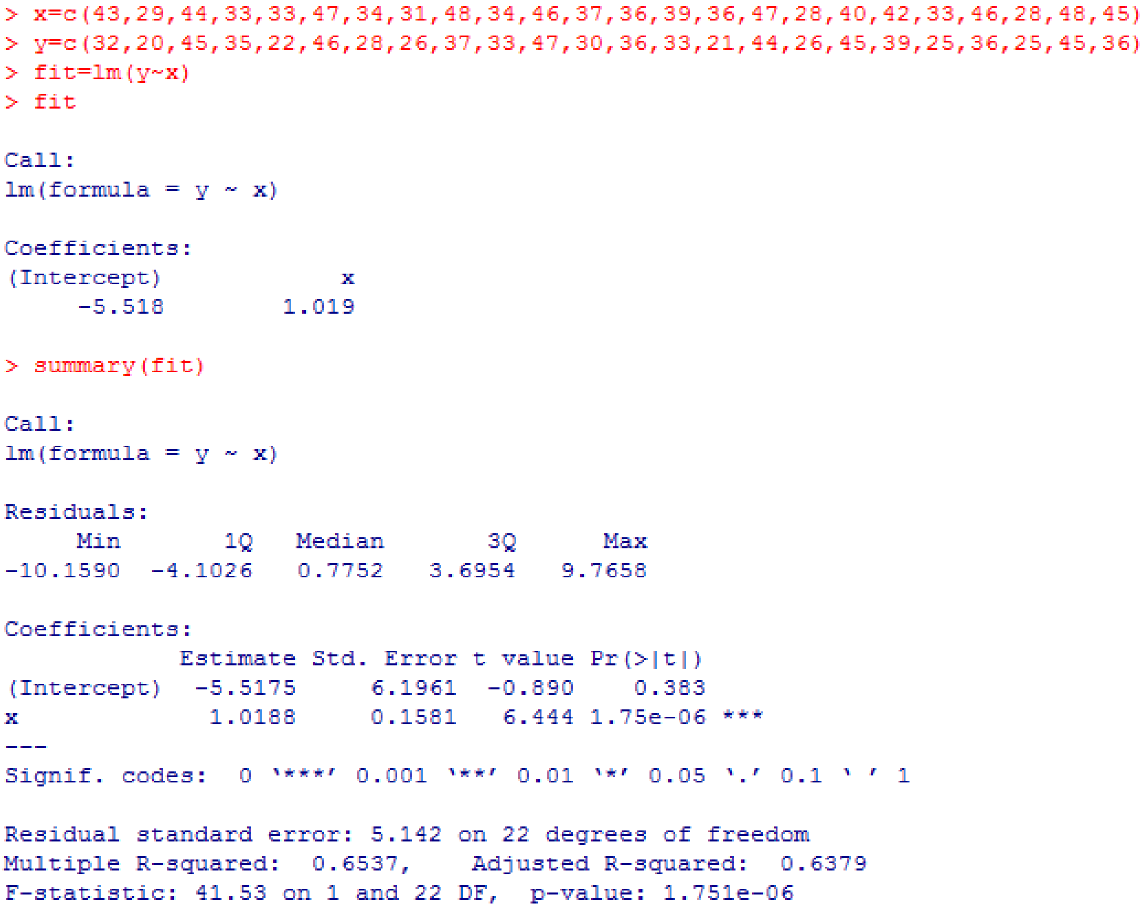
```
> x=c(34,37,36,32,32,36,35,34,29,35)
> y=c(37,37,34,34,33,40,39,37,36,35)
> fit=lm(x~y)
> fit
Call:
lm(formula = x ~ y)
Coefficients:
(Intercept)          y
   18.9167       0.4167
```

The equation of the line of regression of X and Y is X=18.9167+0.4167Y.
The required score of the student in Zoology is 30.58333

Problem 2 :- The following data pertain to the resistance in (ohms) and the failure times (minutes) of 24 overloaded resistors.

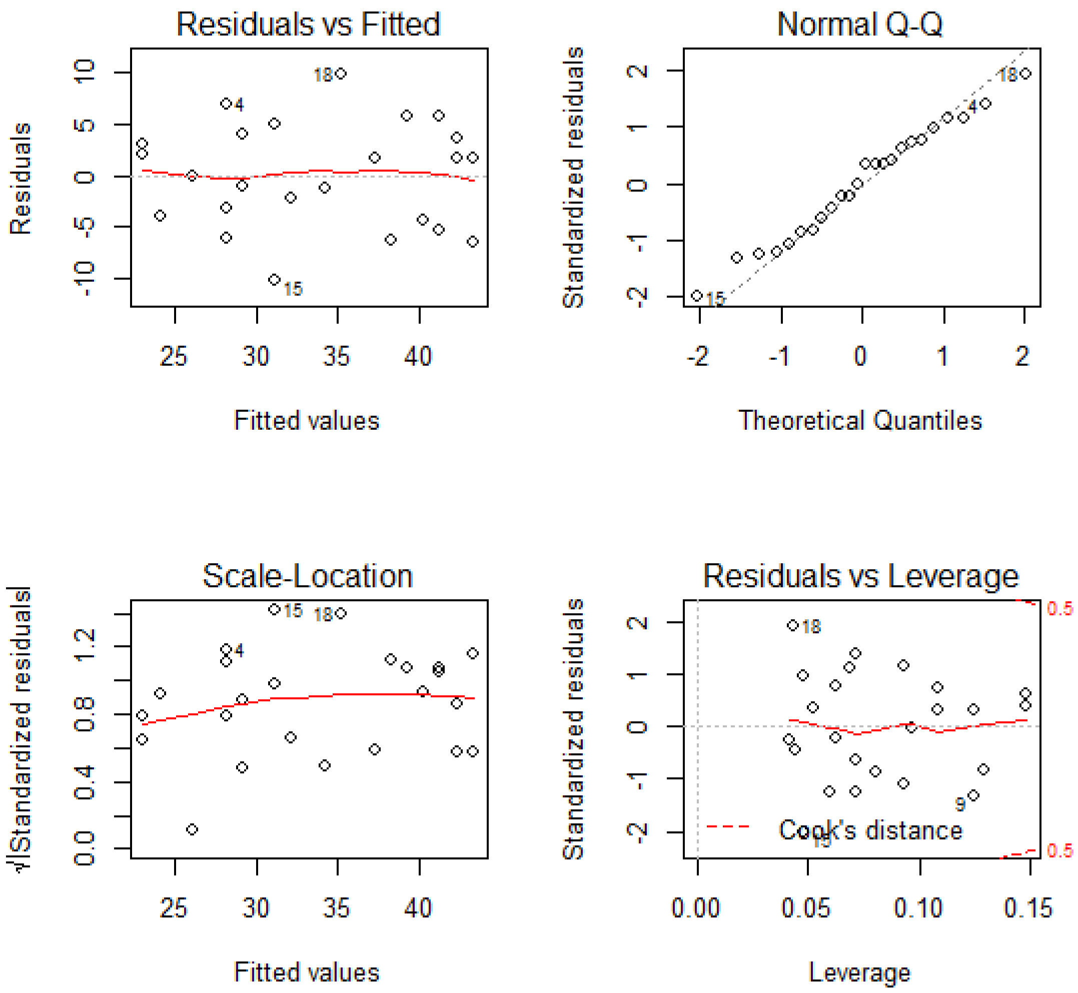
Resistance(x)	***43***	***29***	***44***	***33***	***33***	***47***	***34***	***31***	***48***
	34	***46***	***37***	***36***	***39***	***36***	***47***	***28***	***40***
	42	***33***	***46***	***28***	***48***	***45***			
Failure time(y)	***32***	***20***	***45***	***35***	***22***	***46***	***28***	***26***	***37***
	33	***47***	***30***	***36***	***33***	***21***	***44***	***26***	***45***
	39	***25***	***36***	***25***	***45***	***36***			

R code:-



```
> x=c(43,29,44,33,33,47,34,31,48,34,46,37,36,39,36,47,28,40,42,33,46,28,48,45)
> y=c(32,20,45,35,22,46,28,26,37,33,47,30,36,33,21,44,26,45,39,25,36,25,45,36)
> fit=lm(y~x)
> fit

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x  
     -5.518        1.019  

> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.1590  -4.1026   0.7752   3.6954   9.7658 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -5.5175     6.1961  -0.890    0.383    
x             1.0188     0.1581   6.444 1.75e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 5.142 on 22 degrees of freedom
Multiple R-squared:  0.6537,	Adjusted R-squared:  0.6379 
F-statistic: 41.53 on 1 and 22 DF,  p-value: 1.751e-06
```

```
> par(mfrow=c(2,2));
> plot(fit)
> par(mforw=c(1,1));
```



Diagnostic plots

Problem 3: The sale of a Product in lakhs of rupees(Y) is expected to be influenced by two variables namely the advertising expenditure X1 (in’OOO Rs) and the number of sales persons(X2) in a region. Sample data on 8 Regions of a state has given the following results

Area	**Y**	**X1**	**X2**
1	**110**	**30**	**11**
2	**80**	**40**	**10**
3	**70**	**20**	**7**
4	**120**	**50**	**15**
5	**150**	**60**	**19**
6	**90**	**40**	**12**

7	70	20	8
8	120	60	14

Code:-

```
> Y=c(110,80,70,120,150,90,70,120)
> X1=c(30,40,20,50,60,40,20,60)
> X2=c(11,10,7,15,19,12,8,14)
> input_data=data.frame(Y,X1,X2)
> input_data
    Y X1 X2
1 110 30 11
2  80 40 10
3  70 20  7
4 120 50 15
5 150 60 19
6  90 40 12
7  70 20  8
8 120 60 14
> RegModel <- lm(Y~X1+X2, data=input_data)
> RegModel

Call:
lm(formula = Y ~ X1 + X2, data = input_data)

Coefficients:
(Intercept)           X1           X2
    16.8314      -0.2442       7.8488

> summary(RegModel)

Call:
lm(formula = Y ~ X1 + X2, data = input_data)
```

```
Residuals:
     1      2      3      4      5      6      7      8 
14.157  -5.552   3.110  -2.355  -1.308 -11.250  -4.738   7.936 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  16.8314    11.8290   1.423   0.2140  
X1           -0.2442     0.5375  -0.454   0.6687  
X2            7.8488     2.1945   3.577   0.0159 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.593 on 5 degrees of freedom
Multiple R-squared:  0.9191,    Adjusted R-squared:  0.8867 
F-statistic:  28.4 on 2 and 5 DF,  p-value: 0.001862
```

Interpretation :

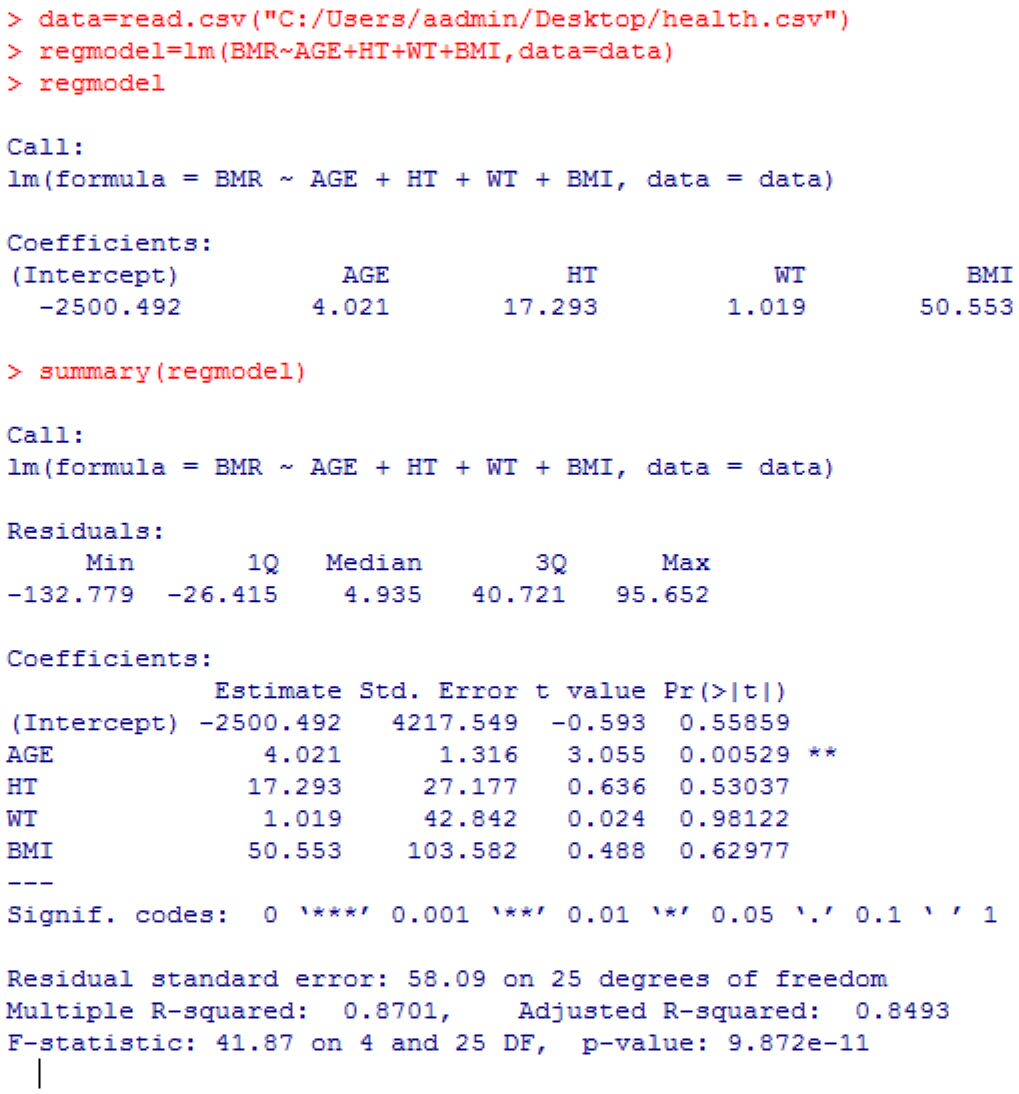
Now the regression the regression model is

$$Y = 16.834 - 0.2442 * X1 + 7.8488 * X2$$

Since R^2 is 0.9593 and the ANOVA shows that the F-ratio is significant, this model can be taken as good-fit in explaining the sales interms of the other two variables.

Problem 4 :(Health.csv) Let us develop a multiple regression model of BMR on the variables age, HT, WT and BMI and interpret the data

Code:-



```
> data=read.csv("C:/Users/aadmin/Desktop/health.csv")
> regmodel=lm(BMR~AGE+HT+WT+BMI,data=data)
> regmodel

Call:
lm(formula = BMR ~ AGE + HT + WT + BMI, data = data)

Coefficients:
(Intercept)          AGE           HT           WT          BMI
  -2500.492        4.021       17.293        1.019       50.553

> summary(regmodel)

Call:
lm(formula = BMR ~ AGE + HT + WT + BMI, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-132.779  -26.415    4.935   40.721   95.652

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2500.492   4217.549  -0.593  0.55859
AGE             4.021      1.316   3.055  0.00529 **
HT             17.293     27.177   0.636  0.53037
WT              1.019     42.842   0.024  0.98122
BMI            50.553    103.582   0.488  0.62977
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58.09 on 25 degrees of freedom
Multiple R-squared:  0.8701,    Adjusted R-squared:  0.8493
F-statistic: 41.87 on 4 and 25 DF,  p-value: 9.872e-11
```

Interpretation:-

Now the Regression model can be stated as

$$BMR = -2500.492 + 4.021(age) + 17.293(HT) + 1.1019 + 50.553(BMI)$$

R^2 is 0.8701 ,which is about 87% of BMR can be explained in terms of age HT,WT and BMI of a person through this linear model, we also see that all the explanatory variables have positive relationship with BMR. These regression coefficient are how ever not statistically significant except that of age, though the F-test in ANOVA shows that the overall regression is significant at 0.01 level(p-value is almost zero).The meaning of the regression coefficient can be understood as follows

if the age increases by 4.021 at fixed values of the other factors like HT,WT and BMI.

Problem 5:(Agriculturedata.csv)

Write the model and interpret about that model for the fallowing Code:

R code:-

```
>input_data<-read.csv('C:/Users/10526/Desktop/Moksha_New/Agriculturedata.csv')
>input_data
>summary(input_data)
>cor(input_data[,c("Net_Agricultural_Output","Population_Active_in_Agriculture","Fertilizer_Consumption","Number_of_Tractors_in_Agriculture")], use="complete.obs")
>RegModel.2 <- lm(Net_Agricultural_Output~Population_Active_in_Agriculture+Fertilizer_Consumption, data=input_data)
>summary(RegModel.2)
>plot(RegModel.2)
```

Practice problems :-

1. ***For the given details viz. Sector wise Number of Factories, Productive Capital, No. of Employees, Total Output and Net Value Added – Fit the Multiple Regression and interpret your result. Assume the variables as Dependent and Independent according to your requirement/description. File Name: Ex 3 data file.***

2. ***Use the Life Satisfaction dataset to fit the regression equation. File Name: Ex 1 and 4 data file.***