

## LINEAR REGRESSION AND MULTIPLE LINEAR REGRESSIONS

### *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

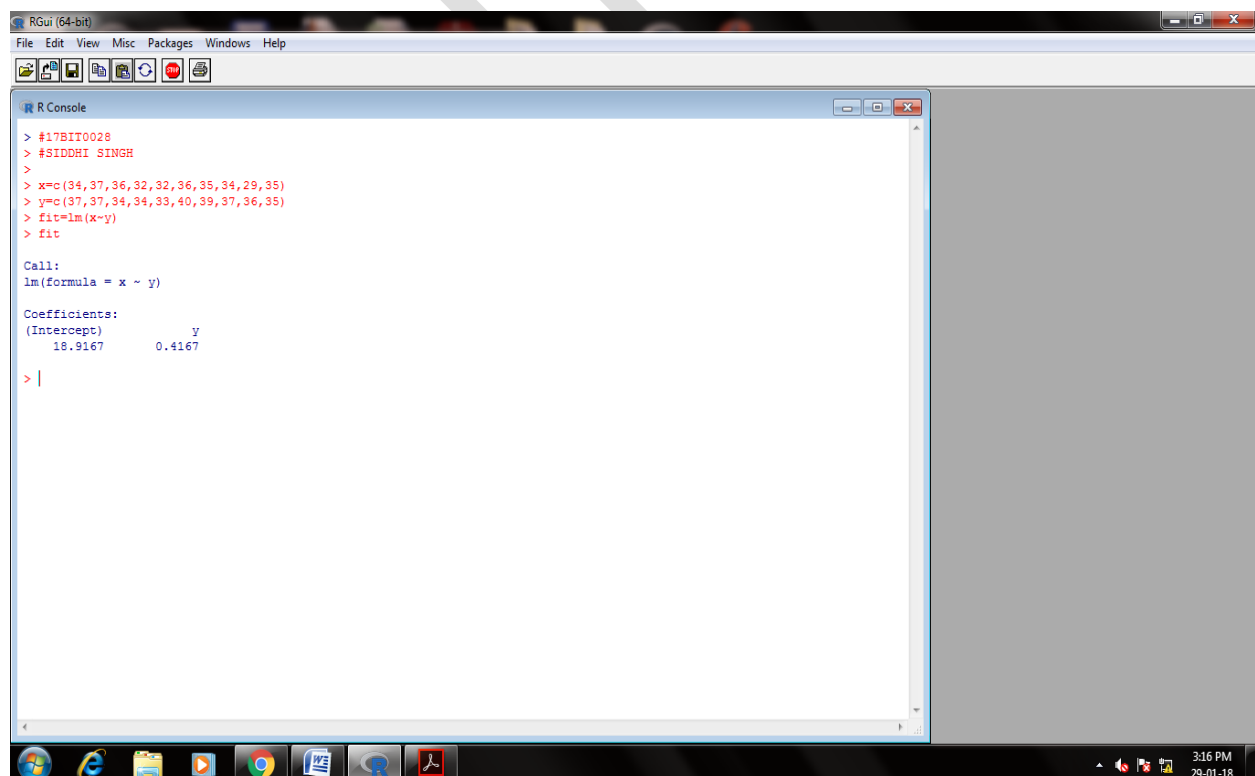
(c) versus any independent variable

**(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



```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console
> #17BIT0028
> #SIDDIHI SINGH
>
> 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

> |
```

**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			

```

RGui (64-bit)
File Edit View Misc Packages Windows Help

>
> #17BIT0028
> #SIDDIH SINGH
> x=c(49,29,44,33,33,47,34,31,48,34,46,37,36,39,36,47,28,40,42,33,46,28,48,45)
> 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
    -2.3883      0.9317

> summary(fit)

Call:
lm(formula = y ~ x)

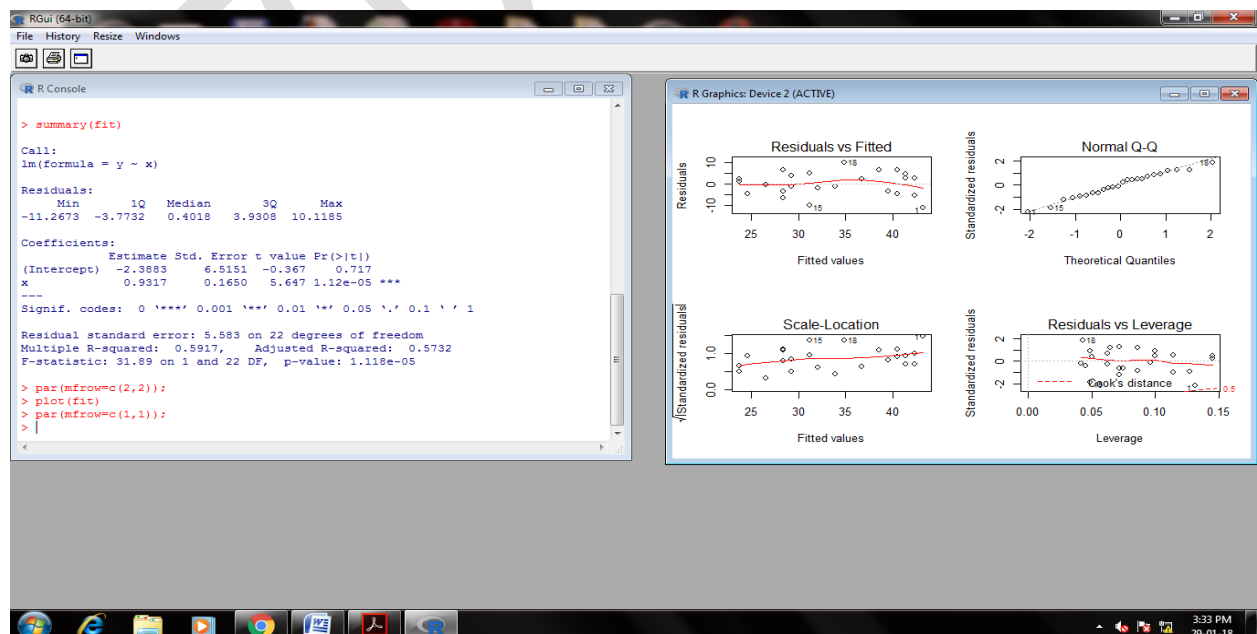
Residuals:
    Min       1Q   Median       3Q      Max
-11.2673  -3.7732   0.4018   3.9308  10.1185

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.3883     6.5151  -0.367   0.717
x              0.9317     0.1650   5.647 1.12e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.583 on 22 degrees of freedom
Multiple R-squared:  0.5917,    Adjusted R-squared:  0.5732
F-statistic: 31.89 on 1 and 22 DF,  p-value: 1.118e-05

>

```



**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'000 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

```

RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console
> 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

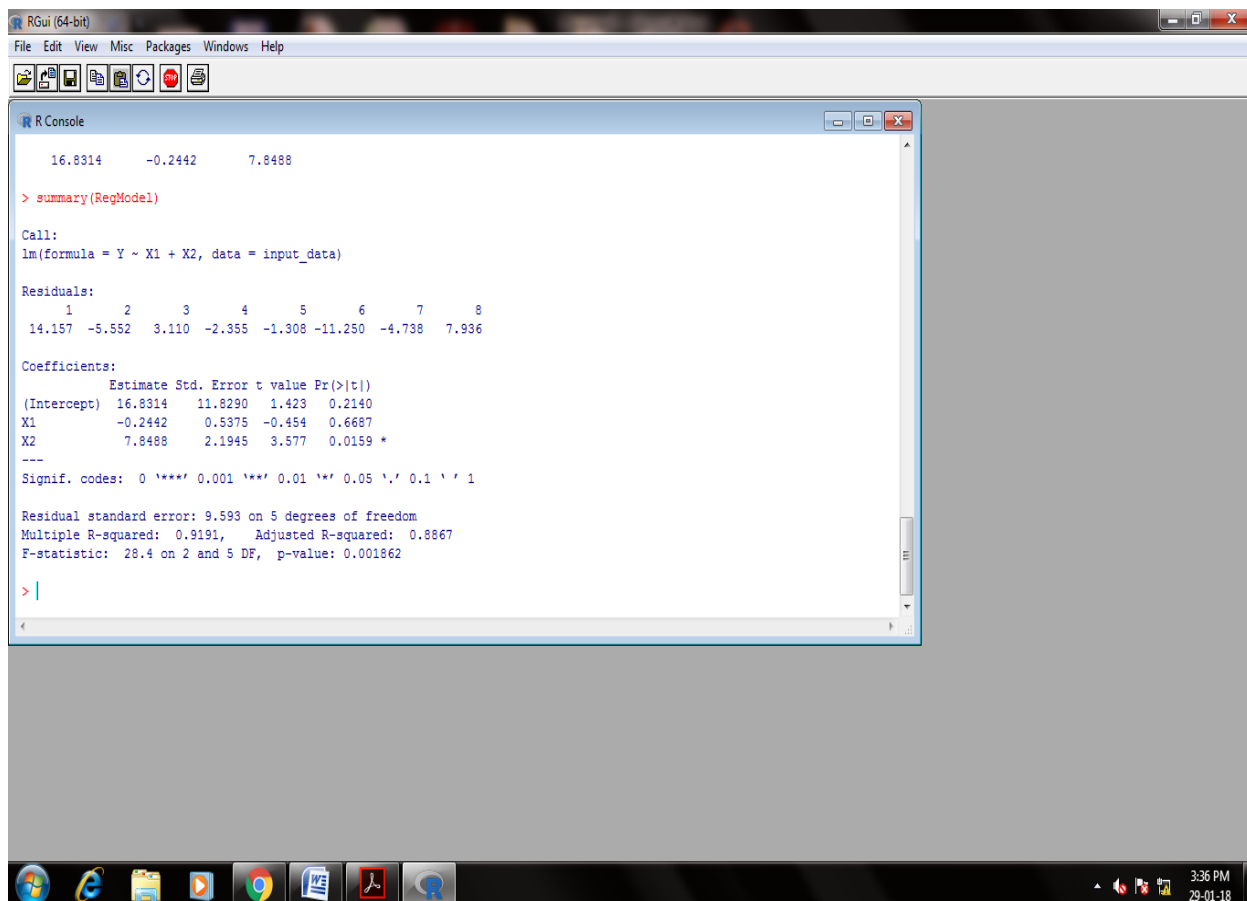
> summary(RegModel)

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:

```



```
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**

```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console
> #SIDHI SINGH
> input_data=read.csv("C:\\Users\\Administrator\\Desktop\\HEALTH.csv")
> input_data
  BMR AGE HT WT BMI
1 1459.3 21 158 51.0 20.43
2 1474.6 21 152 52.0 22.51
3 1413.4 22 159 48.0 18.99
4 1451.6 23 157 50.5 20.49
5 1551.1 24 157 57.0 23.12
6 1597.0 25 153 60.0 25.63
7 1352.2 25 156 44.0 18.08
8 1466.9 25 158 51.5 20.63
9 1581.1 26 159 57.0 22.55
10 1535.8 27 158 56.0 22.43
11 1505.2 27 154 54.0 22.77
12 1566.4 28 158 58.0 23.23
13 1581.7 29 161 59.0 22.76
14 1558.8 29 159 57.5 22.74
15 1453.2 30 157 49.5 20.08
16 1470.6 30 156 51.0 20.96
17 1505.4 32 155 54.0 22.48
18 1528.6 35 158 56.0 22.43
19 1569.2 37 154 59.5 25.09
20 1482.2 38 157 52.0 21.10
21 1401.0 40 159 45.0 17.80
22 1493.8 41 160 53.0 20.70
23 1447.4 44 156 49.0 20.13
24 1459.0 46 155 50.0 20.81
25 1470.6 49 150 51.0 22.67
26 1098.7 18 158 41.0 16.42
27 1201.6 19 159 48.0 18.99
28 1157.5 19 152 45.0 19.48
29 1054.6 20 146 38.0 17.83
30 1157.5 21 155 45.0 18.73
> regmodel=lm(BMR~AGE+HT+WT+BMI,data=data)
> regmodel
```

```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console
27 1201.6 19 159 48.0 18.99
28 1157.5 19 152 45.0 19.48
29 1054.6 20 146 38.0 17.83
30 1157.5 21 155 45.0 18.73
> 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
  234.8941    0.2173   -1.3544    0.6712    3.6808

> summary(regmodel)

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

Residuals:
    1     2     3     4     5     6     7     8     9    10
-52.735 -2.515  2.729  6.599 33.159 55.359 -16.557 -2.175 -35.750 11.886

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  234.8941    224.9131   1.044   0.344
AGE           0.2173     1.0010   0.217   0.837
HT          -1.3544     1.7578  -0.771   0.476
WT           0.6712     1.5578   0.431   0.685
BMI           3.6808     7.3110   0.503   0.636

Residual standard error: 41.71 on 5 degrees of freedom
Multiple R-squared:  0.2157,    Adjusted R-squared:  -0.4117
F-statistic: 0.3439 on 4 and 5 DF,  p-value: 0.8385

> |
```

### **Interpretation:-**

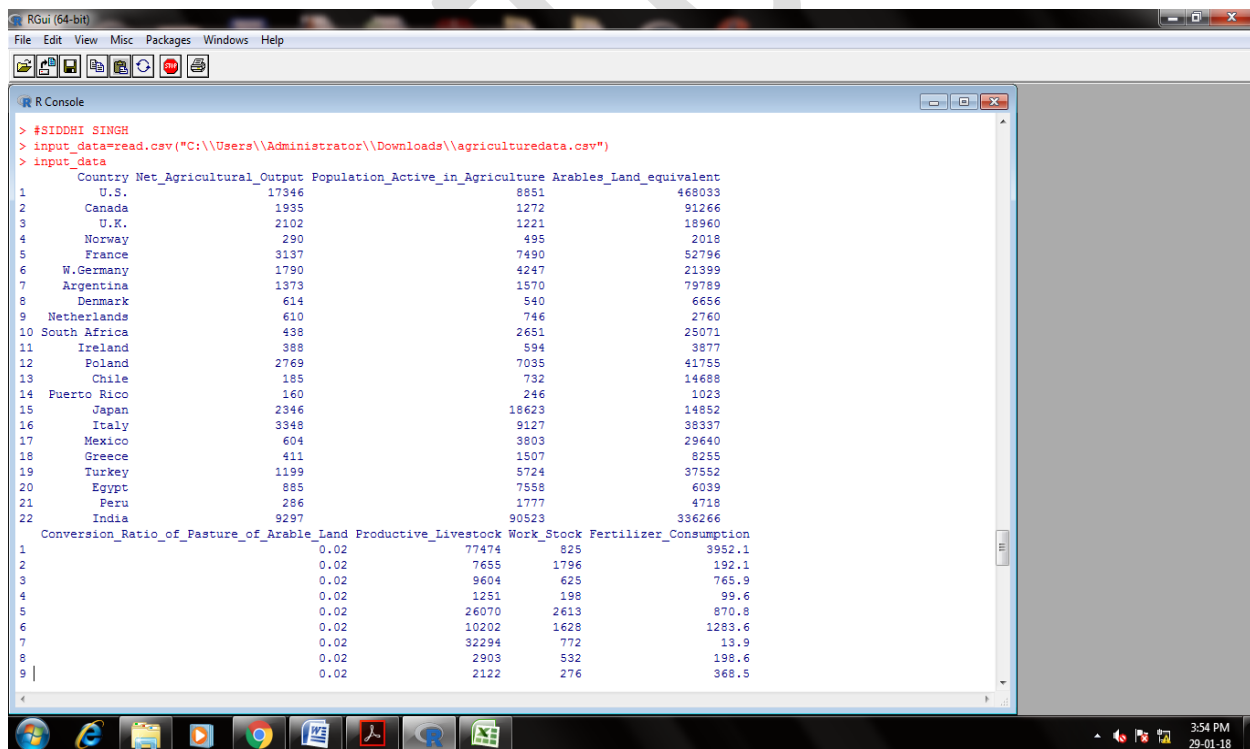
Now the Regression model can be stated as

$$BMR = -2500.492 + 4.021(\text{age}) + 17.293(\text{HT}) + 1.1019 + 50.553(\text{BMI})$$

$R^2$  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 however 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 following Code:**



```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console
> #SIDDHI SINGH
> input_data=read.csv("C:\\Users\\Administrator\\Downloads\\agriculturedata.csv")
> input_data
  Country Net_Agricultural_Output Population_Active_in_Agriculture Arables_Land_equivalent
1      U.S.             17346                8851                468033
2    Canada              1935                 1272                 91266
3       U.K.              2102                 1221                 18960
4    Norway               290                  495                  2018
5     France             3137                 7490                 52796
6 W.Germany             1790                 4247                 21399
7  Argentina            1373                 1570                 79789
8   Denmark              614                  540                  6656
9 Netherlands            610                  746                  2760
10 South Africa          438                 2651                 25071
11   Ireland              388                  594                  3877
12   Poland             2769                 7035                 41755
13    Chile              185                  732                 14688
14 Puerto Rico           160                  246                  1023
15    Japan             2346                 18623                 14852
16    Italy             3348                 9127                 38337
17 Mexico               604                 3803                 29640
18 Greece               411                 1507                  8255
19 Turkey              1199                 5724                 37552
20 Egypt                885                 7558                  6039
21 Peru                 286                 1777                  4718
22 India                9297                 90523                 336266
  Conversion_Ratio_of_Pasture_of_Arable_Land Productive_Livestock Work_Stock Fertilizer_Consumption
1              0.02              77474              825              3952.1
2              0.02              7655              1796              192.1
3              0.02              9604              625              765.9
4              0.02              1251              198              99.6
5              0.02             26070             2613              870.8
6              0.02             10202             1628             1283.6
7              0.02             32294              772              13.9
8              0.02             2903              532              198.6
9              0.02             2122              276              368.5
```



```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console

22      Number_of_Tractors_in_Agriculture      0.05      83328      75373      64.3
1      3550000
2      367828
3      308540
4      9506
5      122624
6      109776
7      25000
8      12257
9      15950
10     39500
11     9480
12     14500
13     6000
14     2150
15     1810
16     50590
17     32000
18     2869
19     3959
20     5400
21     2400
22     7500

> summary(input_data)
      Country  Net_Agricultural_Output  Population_Active_in_Agriculture  Arables_Land_equivalent
Argentina: 1   Min.   : 160.0          Min.   : 246.0          Min.   : 1023
Canada: 1     1st Qu.: 417.8          1st Qu.: 864.8          1st Qu.: 6193
Chile: 1      Median: 1042.0         Median: 2214.0         Median: 20180
Denmark: 1    Mean    : 2341.5         Mean    : 8015.1         Mean    : 59352
Egypt: 1     3rd Qu.: 2285.0         3rd Qu.: 7376.2         3rd Qu.: 40901
France: 1     Max.    :17346.0         Max.    :90523.0         Max.    :468033
(Other): :16
Conversion_Ratio_of_Pasture_of_Arable_Land  Productive_Livestock  Work_Stock  Fertilizer_Consumption
Min.   :0.01000          Min.   : 336          Min.   : 60.0  Min.   : 6.30
1st Qu.:0.02000          1st Qu.: 2750         1st Qu.: 552.2  1st Qu.: 50.77

Number_of_Tractors_in_Agriculture
Min.   : 1810
1st Qu.: 5550
Median : 13378
Mean    : 213620
3rd Qu.: 47818
Max.    :3550000

> cor(input_data[,c("Net_Agricultural_Output", "Population_Active_in_Agriculture", "Fertilizer_Consumption", "Number_of_Tractors_in_Agr")])
      Net_Agricultural_Output  Population_Active_in_Agriculture  Fertilizer_Consumption  Number_of_Tractors_in_Agriculture
Net_Agricultural_Output      1.0000000          0.473733479          0.82630111          0.861061078
Population_Active_in_Agriculture  0.4737335          1.000000000          -0.01069871          -0.007925136
Fertilizer_Consumption      0.8263011          -0.010698707          1.000000000          0.933395590
Number_of_Tractors_in_Agriculture  0.8610611          -0.007925136          0.933395590          1.000000000
```

```
RGui (64-bit)
File Edit View Misc Packages Windows Help

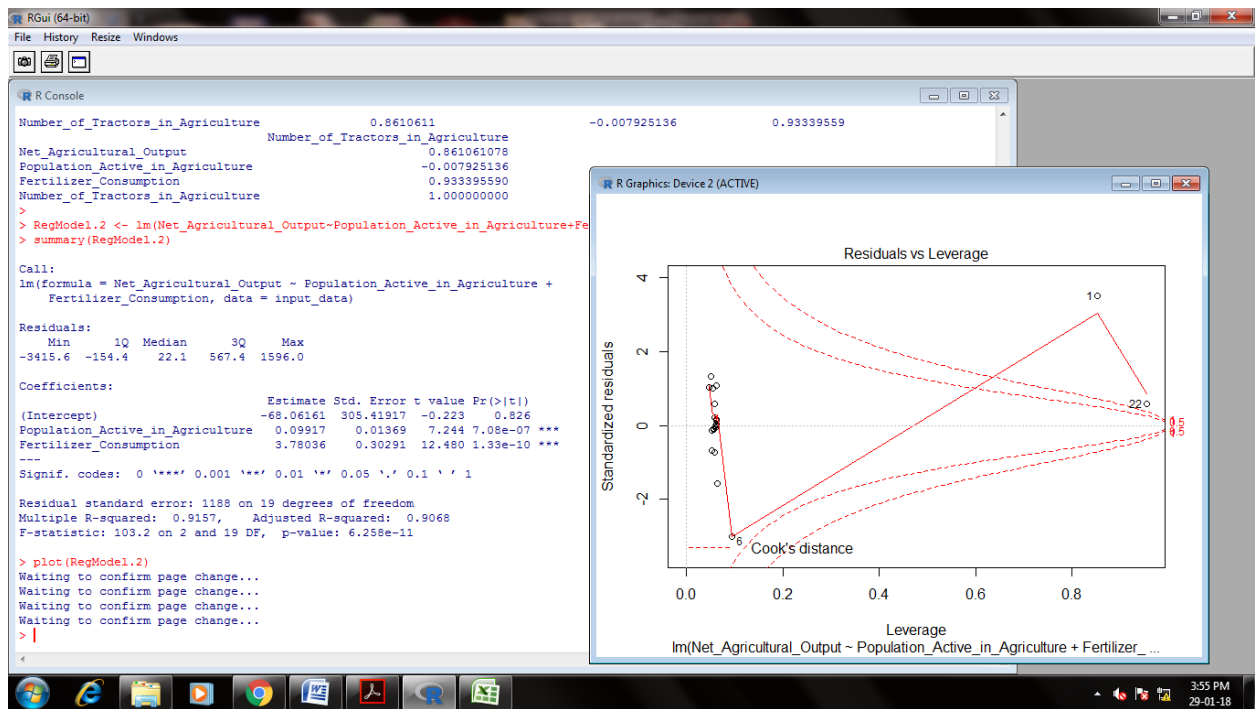
R Console

> summary(input_data)
      Country  Net_Agricultural_Output  Population_Active_in_Agriculture  Arables_Land_equivalent
Argentina: 1   Min.   : 160.0          Min.   : 246.0          Min.   : 1023
Canada: 1     1st Qu.: 417.8          1st Qu.: 864.8          1st Qu.: 6193
Chile: 1      Median: 1042.0         Median: 2214.0         Median: 20180
Denmark: 1    Mean    : 2341.5         Mean    : 8015.1         Mean    : 59352
Egypt: 1     3rd Qu.: 2285.0         3rd Qu.: 7376.2         3rd Qu.: 40901
France: 1     Max.    :17346.0         Max.    :90523.0         Max.    :468033
(Other): :16
Conversion_Ratio_of_Pasture_of_Arable_Land  Productive_Livestock  Work_Stock  Fertilizer_Consumption
Min.   :0.01000          Min.   : 336          Min.   : 60.0  Min.   : 6.30
1st Qu.:0.02000          1st Qu.: 2750         1st Qu.: 552.2  1st Qu.: 50.77
Median :0.02000          Median : 7061         Median : 983.0  Median : 98.55
Mean    :0.02318          Mean    :14801         Mean    :4877.1  Mean    :427.13
3rd Qu.:0.02000          3rd Qu.:12846         3rd Qu.: 2301.8  3rd Qu.: 362.95
Max.    :0.05000          Max.    :83328         Max.    :75373.0  Max.    :3952.10

Number_of_Tractors_in_Agriculture
Min.   : 1810
1st Qu.: 5550
Median : 13378
Mean    : 213620
3rd Qu.: 47818
Max.    :3550000

> cor(input_data[,c("Net_Agricultural_Output", "Population_Active_in_Agriculture", "Fertilizer_Consumption", "Number_of_Tractors_in_Agr")])
      Net_Agricultural_Output  Population_Active_in_Agriculture  Fertilizer_Consumption  Number_of_Tractors_in_Agriculture
Net_Agricultural_Output      1.0000000          0.473733479          0.82630111          0.861061078
Population_Active_in_Agriculture  0.4737335          1.000000000          -0.01069871          -0.007925136
Fertilizer_Consumption      0.8263011          -0.010698707          1.000000000          0.933395590
Number_of_Tractors_in_Agriculture  0.8610611          -0.007925136          0.933395590          1.000000000
```





# LAB-5

## LINEAR REGRESSION AND MULTIPLE REGRESSIONS

17BIT0028

SIDDHI SINGH

Lab-5

28/11/18

classmate  
Date  
Page

### Problem 1: 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: 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, 23, 49,
        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
```