# LAB 4: Multiple Regression (Part I)

**THEORY:** Multiple regression is a generalization of linear regression. In this case there will be more than one independent variable and one dependent variable. It is of the form

Where *ai* s are called the regression coefficients. In general, the model with *k* independent variable can be expressed as

Where *Y*= (*y1, y2......yn*)’ is a n x 1 vector of n observations on dependent variable and

is n x k matrix of *n* observations of *k* independent variables, *A=*[*a1......ak*]’ is a k x 1 vector of regression coefficients and *c* = [*c1*, *c*2, ....*cn*] is n x 1 vector of random error. If intercept term is present, take first column of X to be (1,1,…,1)’. To find the regression coefficients, the normal equation is given by

If it is assumed that rank(*X*) =*k*, then *X’X* is positive definite and inverse exists. Then

There are various measures to check the goodness of fit. Multiple R is the correlation coefficient between the actual and predicted values of y in case of multiple regression. It represents the strength of the linear relationship. If it is closer to 1, it implies the relationship is highly linear. It is the square root of r squared. R squared gives the measure of explained variation by the linear relationship so obtained. Adjusted R squared adjusts for the number of terms in the model. Hence it can be considered as a better measure as compared to R squared. Standard error of regression gives an estimate of standard deviation of the error.

ANOVA is a type of statistical testing widely used to statistically analyse data. Here df stands for degree of freedom. SS stands for sum of squares. MS stands for mean square. Now df corresponding to regression is number of independent variable and df corresponding to residual is nothing difference between number of observations and number of independent variable. SS corresponding to regression represents explained variation and is given by sum of squares of deviation of predicted value from the mean value. SS corresponding to the residual represents unexplained variation and is given by sum of squares of difference between the actual and predicted value. The MS is obtained by dividing SS by corresponding df. The F statistic is calculated as the ratio of MS corresponding to regression and MS corresponding to residuals. In the second table, the intercept and regression coefficients are given. Alongwith it, the standard error in calculation of regression coefficients, corresponding t- statistic and confidence limits are given.

**Ex. 1:** The following contains data on three variables that were collected in an

observational study in a semiconductor manufacturing plant. In this plant, the finished semiconductor is wire-bonded to a frame. The variables reported are pull strength (a measure of the amount of force required to break the bond), the wire length, and the height of the die. We would like to find a model relating pull strength to wire length and die height. Fit a multiple regression model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Observation Number** | **Pull strength y** | **Wire Length x1** | **Die Length x2** |
| 1 | 9.95 | 2 | 50 |
| 2 | 24.45 | 8 | 110 |
| 3 | 31.75 | 11 | 120 |
| 4 | 35 | 10 | 550 |
| 5 | 25.02 | 8 | 295 |
| 6 | 16.86 | 4 | 200 |
| 7 | 14.38 | 2 | 375 |
| 8 | 9.6 | 2 | 52 |
| 9 | 24.35 | 9 | 100 |
| 10 | 27.5 | 8 | 300 |
| 11 | 17.08 | 4 | 412 |
| 12 | 37 | 11 | 400 |
| 13 | 41.95 | 12 | 500 |
| 14 | 11.66 | 2 | 360 |
| 15 | 21.65 | 4 | 205 |
| 16 | 17.89 | 4 | 400 |
| 17 | 69 | 20 | 600 |
| 18 | 10.3 | 1 | 585 |
| 19 | 34.93 | 10 | 540 |
| 20 | 46.59 | 15 | 250 |
| 21 | 44.88 | 15 | 290 |
| 22 | 54.12 | 16 | 510 |
| 23 | 56.63 | 17 | 590 |
| 24 | 22.13 | 6 | 100 |
| 25 | 21.15 | 5 | 400 |

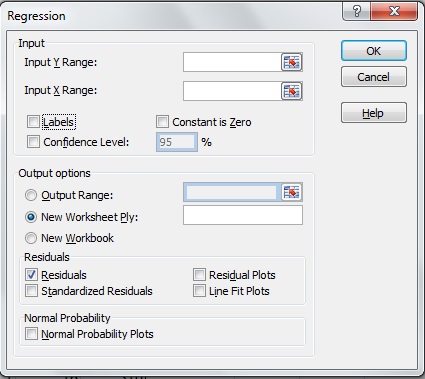
**PROCEDURE:**

* Go to files, open a new EXCEL workbook
* Enter the data in the workbook.
* To fit a multiple regression model using principle of least squares, define X as with first column as ones, second column as x1 and third column as x2.
* Compute X’ by typing = TRANSPOSE(select all cells of matrix X), press CTRL+SHIFT+ ENTER. Note that TRANSPOSE() is used to get transpose.
* Compute (X’X)-1 by typing = MINVERSE((select cells under X’, select cells under X)), press CTRL+SHIFT+ENTER.
* Compute ((X’X)-1X’) by typing =MMULT(select cells under (X’X)-1, select cells under X’),press CTRL+SHIFT+ENTER.
* Now . So, compute A by typing = MMULT(select cells under ((X’X)-1X’), select cells under y), press CTRL+SHIFT+ENTER.
* Therefore y’ = XA. So compute y’ by typing =MMULT(select cells under X, select cells under A), press CTRL+SHIFT+ENTER.
* Compute residual by typing = (select 1st cell of y)-(select 1st cell of y’), press ENTER. Then select, drag and press ENTER. Dragging helps in copying the same formula.
* Compute average of Y by typing =AVERAGE(select cells under Y), press ENTER.
* Compute Multiple R by typing =CORREL(select cells under Y, select cells under y’), press ENTER. Clearly, CORREL() computes the correlation.
* In order to use Data tools, select tab data and click on Data Analysis. A pop up window will appear as shown in the Fig 4.1.1

# fig4.1.1.jpg

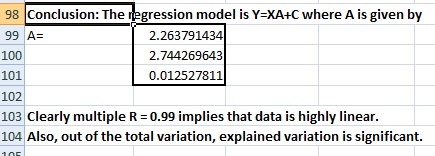
**Fig. 4.1.1**

* Choose Regression and press OK. A new pop up window will appear as shown in the Fig 4.1.2
* In Input Y range, select Y values including the label.
* In Input X range, select x1 and x2 values including the label.
* Since Labels are included, select label.
* In Output range, select some empty space in the worksheet.
* In Residuals, select residuals.
* Press OK.

****

**Fig. 4.1.2**

Therefore, the final conclusion is shown in Fig. 4.1.3



**Fig. 4.1.3**