# LAB 3: Linear Regression

**THEORY:** If the points in the scatter plot lie almost in a straight line then we say that X is related to Y linearly. This linear interdependence between the two variables or data sets is quantified by the measure called as correlation coefficient. The value lies between -1 and 1. If the absolute value of correlation coefficient is closer to 1 implies that X and Y are linearly dependent, whereas correlation close to zero implies existence of hardly any linear dependence of X and Y. In such cases, the points will not be in a straight line in the scatter plot. The sign signifies the direction of linear dependence.

So, once if it is established that there is some kind of linear dependence between X and Y, then we would like to find the equation of the line that describes the data perfectly. This best fit line is called the line of regression. The line that gives the best possible value of Y in terms of X is called as the line of regression of Y on X. Similarly, the line that gives the best possible value of X in terms of Y is called as the line of regression of X on Y. The coefficients of the lines of regression are computed using the principle of least squares.

Consider the line of regression of *y* on *x*. Let the equation satisfied by the data points be of the form

*y=a* + *bx.*

By the principle of least squares, we get the normal equations as follows

∑y = a n+ b∑x

∑xy = a ∑x + b∑xx

This can be written in the matrix form as

.

Clearly, we can find *a* and *b* by solving the system of equations. The *b* is called the regression coeffieicnt of *y* on *x*. The correlation coefficient *r* is given by

where are standard deviation of *x* and *y* respectively.

Note that the regression coefficient is nothing but the slope of the line and *a* is the intercept. Once the regression line of *y* on *x* is obtained, we can predict values of *y* by substituting values of x in the equation. The sum of square of difference of actual value and the predicted value is called the RSS and the TSS is nothing but the sum of square of deviation y from its mean value. RSS gives a measure of dispersion of data points around the regression line. The difference between TSS and RSS gives the sum of explained variation.

**Ex. 1:** The data set contains marketing budget and sales of certain company for past 17 years. Build a simple regression model to understand how the marketing budget impacts on sales. Plot the scatter plot and regression equation of y on x, correlation coefficient, TSS, RSS without using data analysis tools as well using data analysis tools.

|  |  |
| --- | --- |
| **Marketing Budget (X) (In lakhs)** | **Actual Sales(Y) (In crores)** |
| 127.4 | 10.1 |
| 364.4 | 21.4 |
| 150 | 10 |
| 128.7 | 9.6 |
| 285.9 | 17.4 |
| 200 | 12.5 |
| 303.3 | 20 |
| 315.7 | 21 |
| 169.8 | 14.7 |
| 104.9 | 10.1 |
| 297.7 | 21.5 |
| 256.4 | 16.6 |
| 249.1 | 17.1 |
| 323.1 | 20.7 |
| 223 | 15.5 |
| 235 | 13.5 |
| 200 | 12.5 |

**Solution:**

* Go to files, open a new EXCEL workbook
* Enter the data in the workbook
* Select data under x and y, select tab insert, choose scatter plot.
* Let the regression equation be given by *y* = *a* +*bx*
* So, *a =*type =INTERCEPT(select data in y column, select data in x column), press ENTER. As the word suggest, INTERCEPT() computes the y intercept of the line corresponding to the data
* *b*= type =SLOPE(select data in y column, select data in x column), press ENTER.

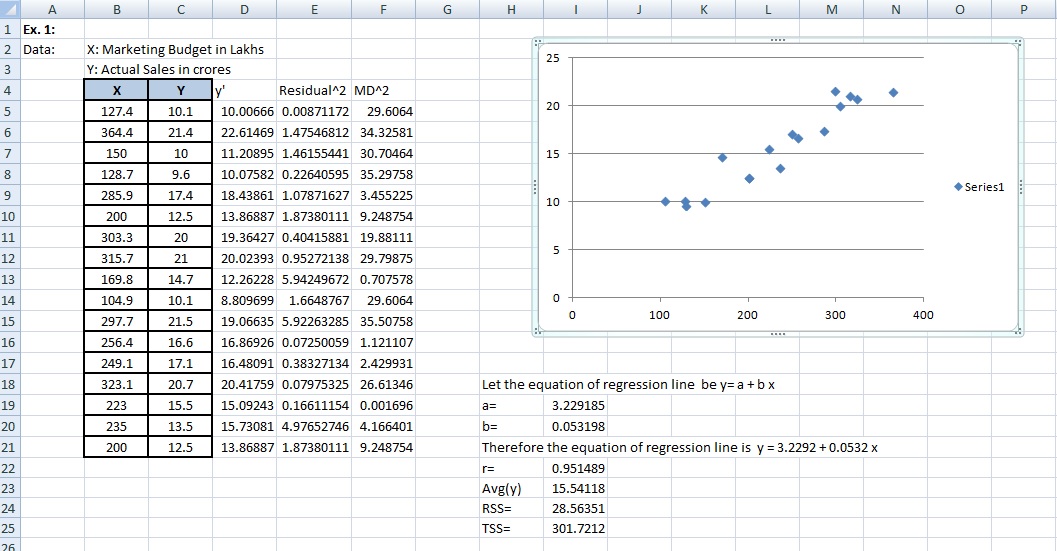
Therefore the equation of regression line is given by

*y=*0.053*x+*3.229

Similarly, SLOPE() evaluates the slope.

* Correlation coefficient r= type =correl(select data in y column, select data in x column), press ENTER.
* Avg(y)=type =AVERAGE(select data in y column), press ENTER
* Compute predicted value of y= y’ using the regression equation
* Compute the square of residual (y-y’)^2. This can be done by typing =(select 1st cell under y-select 1st cell under y’)^2,press ENTER. Then select that cell and drag till the last cell of that column, press ENTER.
* Compute the square of deviation from the mean. This can be done by typing =((select cells under y)-select cell containing avg(y))^2,press ENTER.
* Compute RSS by typing=SUM(cells under residual squares), press ENTER.
* Compute TSS by typing=SUM(cells under square of deviation from the mean), press ENTER

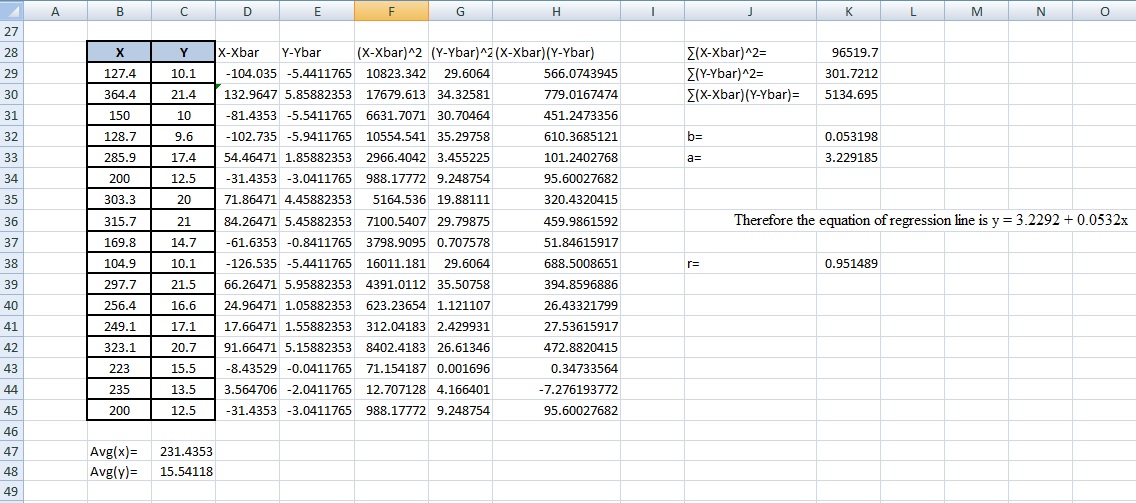
.

****

**Fig 3.1.1**

We can solve this problem using method of least squares.

* Enter the data.
* Compute Avg(x) by typing average(select all cells under x), press ENTER.
* Similarly, compute Avg(y).
* Compute deviation of x from its mean X-Xbar by typing=(select the 1st cell under x)-type alphabet of the column containing average of x followed by the $ and the row no, press ENTER. Then drag the entire column.
* Compute deviation of y from its mean Y-Ybar.
* Compute (X-Xbar)^2 by typing =(select a cell under X-Xbar)^2, press ENTER. Then drag. In same way compute (Y-Ybar)^2
* Compute (X-Xbar)(Y-Ybar) by typing = select 1st cell under X-Xbar followed by a \* and then select the 1st cell under Y-Ybar, press ENTER. Then drag.
* Compute sum of(X-Xbar)^2, (Y-Ybar)^2, (X-Xbar)(Y-Ybar)
* Compute b by typing = (sum of (X-Xbar)(Y-Ybar))/(sum of (X-Xbar)^2)
* Compute a by typing = -b\*(select cell containing average(x))+(select cell containing average(y)), press ENTER.
* Therefore the equation of regression line is y = 3.2292 + 0.0532x
* Compute correlation coefficient r by typing =(select cell containing (sum of (X-Xbar)(Y-Ybar))/sqrt((select cell containing (sum of(X-Xbar)^2)\* (select cell containing (sum of(Y-Ybar)^2)), press ENTER.

****

**Fig 3.1.2**

This completes our problem.

**Ex. 2:** An article in the Journal of Environmental Engineering (1989, Vol. 115(3), pp. 608–619) reported the results of a study on the occurrence of sodium and chloride in surface streams in central Rhode Island. The following data are chloride concentration y (in milligrams per liter) and roadway area in the watershed x (in percentage).

(a) Fit a simple linear regression model with y = green liquor Na2 S concentration and x = production. Find an estimate of σ2 . Draw a scatter diagram of the data and the resulting least squares fitted model.

(b) Find the fitted value of y corresponding to x = 910 and the associated residual.

(c) Find the mean green liquor Na2 S concentration when the production rate is 950 tons per day.

|  |  |
| --- | --- |
| **x** | **y** |
| 0.19 | 4.4 |
| 0.15 | 9.6 |
| 0.57 | 9.7 |
| 0.7 | 10.6 |
| 0.67 | 10.8 |
| 0.63 | 10.9 |
| 0.47 | 11.8 |
| 0.7 | 12.1 |
| 0.6 | 14.3 |

|  |  |
| --- | --- |
| **x** | **y** |
| 0.78 | 14.7 |
| 0.81 | 15 |
| 0.78 | 14.3 |
| 0.69 | 19.2 |
| 1.3 | 23.1 |
| 1.05 | 27.4 |
| 1.06 | 27.7 |
| 1.74 | 31.8 |
| 1.62 | 39.5 |

**Ex. 3:** The coefficient of thermal expansion of steel α is given at discrete values of temperature.Develop a simple linear regression model for the given data between temperature and α

|  |  |
| --- | --- |
| **Temperature (T) 0F** | **Coefficient of Thermal expansion in (α) in/in 0F** |
| 80 | 6.46E-06 |
| 60 | 6.36E-06 |
| 40 | 6.24E-06 |
| 20 | 6.12E-06 |
| 0 | 6.00E-06 |
| -20 | 5.86E-06 |
| -40 | 5.72E-06 |
| -60 | 5.58E-06 |
| -80 | 5.43E-06 |
| -100 | 5.28E-06 |
| -120 | 5.09E-06 |
| -140 | 4.91E-06 |
| -160 | 4.72E-06 |
| -180 | 4.52E-06 |
| -200 | 4.30E-06 |
| -220 | 4.08E-06 |
| -240 | 3.83E-06 |
| -260 | 3.58E-06 |
| -280 | 3.33E-06 |
| -300 | 3.07E-06 |
| -320 | 2.76E-06 |
| -340 | 2.45E-06 |
| -340 | 2.45E-06 |