Regression

- Regression Analysis measures the nature and extent of the relationship between two or more variables, thus enables us to make predictions.
- Regression is the measure of the average relationship between two or more variables.

Utility of Regression

- Degree & Nature of relationship
- Estimation of relationship.
- Prediction.
- Useful in Economic and Business Research.

Difference Between Correlation & regression:

- Degree & Nature of Relationship.
  - Correlation is a measure of degree of relationship between X &Y
  - Regression studies the nature of relationship between the variables so that one may be able to predict the value of one variable on the basis of another.
- Cause and Effect Relationship
  - Correlation does not always assume cause and effect relationship between two variables. The independent variable is the cause and dependent variable is the effect.
- Prediction
  - Correlation doesn't help in making predictions.
  - Regression enable us to make predictions using regression line.
- Symmetric
  - Correlation coefficients are symmetrical . i.e., $r_{xy} = r_{yx}$.
  - Regression coefficients are not symmetrical. i.e., $b_{xy} \neq b_{yx}$.
- Origin & scale
  - Correlation coefficient is independent of change of origin and scale.
  - Regression coefficient is independent of change of origin but not of scale.

Angle between two lines of regression:

If θ is the angle between the lines: $\tan\theta = \left[\frac{1-r^2}{r}\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)\right]$. As $|r| \le 1$, based on r values, regression lines may be parallel or perpendicular.

Case (i): If r = 0, Regression lines are perpendicular to each other.
Case (ii): If $r = \pm 1$, two lines of regression either coincide or parallel to each other.

Type of regression analysis:

- Simple and Multiple regression
- Linear and Nonlinear regression
- Partial and Total regression

In simple Linear Regression , we have Regression lines, Regression equations, and Regression coefficients.

Regression Lines:

- Regression lines show the average relationship between two variables. It is also called line of best fit
- If two variables X and Y are given, then there are two regression lines.
    - Regression line X on Y
      $X = a + bY$
      $\left(X - \overline{X}\right) = b_{xy}\left(Y - \overline{Y}\right)$
    - Regression line Y on X
      $Y = a + bX$
    - $\left(Y - \overline{Y}\right) = b_{yx}\left(X - \overline{X}\right)$

Regression Coefficients:
There are two types of regression coefficients.

- Regression coefficient Y on X

$$b_{yx} = \frac{r\sigma_y}{\sigma_x}$$

- Regression coefficient X on Y

$$b_{xy} = \frac{r\sigma_x}{\sigma_y}$$

Properties of Regression Coefficients:

- Coefficient of correlation is the geometric mean of the regression coefficients. i.e., $r = \sqrt{b_{xy}.b_{yx}}$
- Both the regression coefficients must have the same algebraic sign.
- Coefficient of correlation must have the same sign as that of the regression coefficients.
- Both the regression coefficients can not be greater than unity.
- Arithmetic mean of the regression coefficients is equal to or greater than the correlation coefficient r.

- Regression coefficient is independent of change of origin but not of scale.
- Nature of Regression of lines.
  - If $r = \pm 1$, two lines of regression either coincide or parallel to each other.
  - If r = 0, Regression lines intersect each other at 90^0
  - Nearer the regression lines to each other, the greater will be the degree of correlation.
  - If the regression lines rise from left to right upward, then the correlation is positive.
  - If the regression lines from left to right downward, then the correlation is negative

Least square method:

To fit a parabola $y = a + bx + cx^2$

Form the normal equations:

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Solve these simultaneous equations for a,b, and c

Substitute the values

Regression equations using Normal equations (Least square method)

Under this regression equations can be calculated by solving normal equations.

For regression equations Y on X: Y = a + bX

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma X\ Y = a\ \Sigma X + b\Sigma X^2$$

Another method:

$$\left(Y - \bar{Y}\right) = b_{yx}\left(X - \bar{X}\right)$$

$$b_{yx} = \left(\frac{N\sum XY - \sum X\sum Y}{N\sum X^2 - \left(\sum X\right)}\right)$$

Here a is the y- intercept, indicates the minimum value of Y for X=0

And b is the slope of the line, indicates the absolute increase in Y for a unit increase in X

Under this method, regression equations can be calculated by solving two normal equations:

For regression equations Y on X: Y = a + bX

$\Sigma Y = Na + b\Sigma X$

$\Sigma X\ Y = a\ \Sigma X + b\Sigma X^2$

1. Calculate the regression equation of X on Y using method of least squares.

X:	1	2	3	4	5
Y:	2	5	3	8	7

X	Y	XY	Y^2
1	2	2	4
2	5	10	25
3	3	9	9
4	8	32	64
5	7	35	49
15	25	88	151

$\Sigma X = Na + b\Sigma Y$

$\Sigma X\ Y = a\ \Sigma Y + b\Sigma Y^2$

$15 = 5a + 25b$ (1)

$88 = 25a + 151\ Y^2$ (2)

Solving (1) and (2)

We get **a = 0.5 and b= 0.5**

∴ Regression line X on Y is **X = 0.5 + 0.5 Y**

2. Given the following data:

   $N = 8,\ \sum x = 21,\ \sum x^2 = 99, \sum y = 4, \sum y^2 = 68,\ \sum xy = 36$

   Using the values , find:

   Regression Equation of y on x.

   Regression equation of x on y.

   Value of y when x = 10.

   Value of x when y = 2.5

$$\left(Y-\overline{Y}\right)=b_{yx}\left(X-\overline{X}\right)$$

$$b_{yx}=\left(\frac{N\sum XY-\sum X\sum Y}{N\sum X^{2}-\left(\sum X\right)^{2}}\right)$$

$$\overline{Y}=\frac{\Sigma Y}{N}=\frac{4}{8}=0.5;\quad \overline{X}=\frac{\Sigma X}{N}=\frac{21}{8}=2.625$$

$$b_{yx}=\left(\frac{8\times 36-21\times 4}{8\times 99-21^{2}}\right)=0.581$$

$$\left(Y-0.5\right)=0.581\left(X-2.625\right)$$

When X= 10, Y=5.81-1.0251=4.785

$$Y=0.581X-1.0251$$

$$\left(X-\overline{X}\right)=b_{xy}\left(Y-\overline{Y}\right)$$

$$b_{xy}=\left(\frac{N\sum XY-\sum X\sum Y}{N\sum Y^{2}-\left(\sum Y\right)^{2}}\right)$$

$$\overline{Y}=\frac{\Sigma Y}{N}=\frac{4}{8}=0.5;\quad \overline{X}=\frac{\Sigma X}{N}=\frac{21}{8}=2.625$$

$$b_{xy}=\left(\frac{8\times 36-21\times 4}{8\times 68-4^{2}}\right)=0.386$$

$$\left(X-2.625\right)=0.386\left(Y-0.5\right)$$

whenY -= 2.5, X = 0.965+2.432=3.397

$$X=0.386Y+2.432$$