REGRESSION ANALYSIS

MEANING OF REGRESSION:

The dictionary meaning of the word Regression is ‘Stepping back’ or ‘Going back’. Regression is the measures of the average relationship between two or more variables in terms of the original units of the data. And it is also attempts to establish the nature of the relationship between variables that is to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting.

DIFFERENCE BETWEEN REGRESSION & CORRELATION

Correlation	Regression
Correlation coefficient (r) between x & y is a measure of direction & degree of linear relationship between x & y;	bxy & byx are mathematical measures expressing the average relationship between the two variables
It does not imply cause & effect relationship between the variables.	It indicates the cause & effect relationship between variables.
It indicates the degree of association	It is used to forecast the nature of dependent variable when the value of independent variable is know

Importance of Regression Analysis

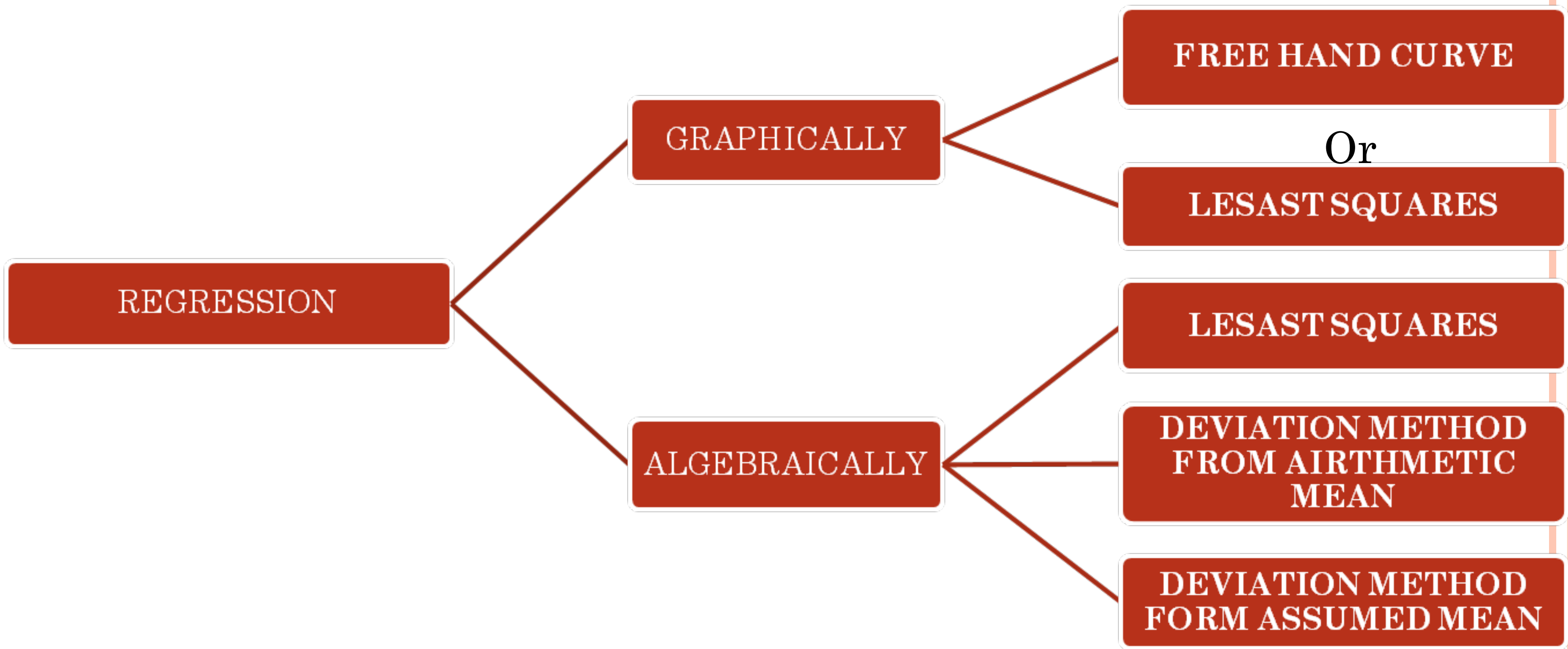
Regression analysis helps in three important ways :-

- It provides estimate of values of dependent variables from values of independent variables.
- It can be extended to 2or more variables, which is known as multiple regression.
- It shows the nature of relationship between two or more variable.

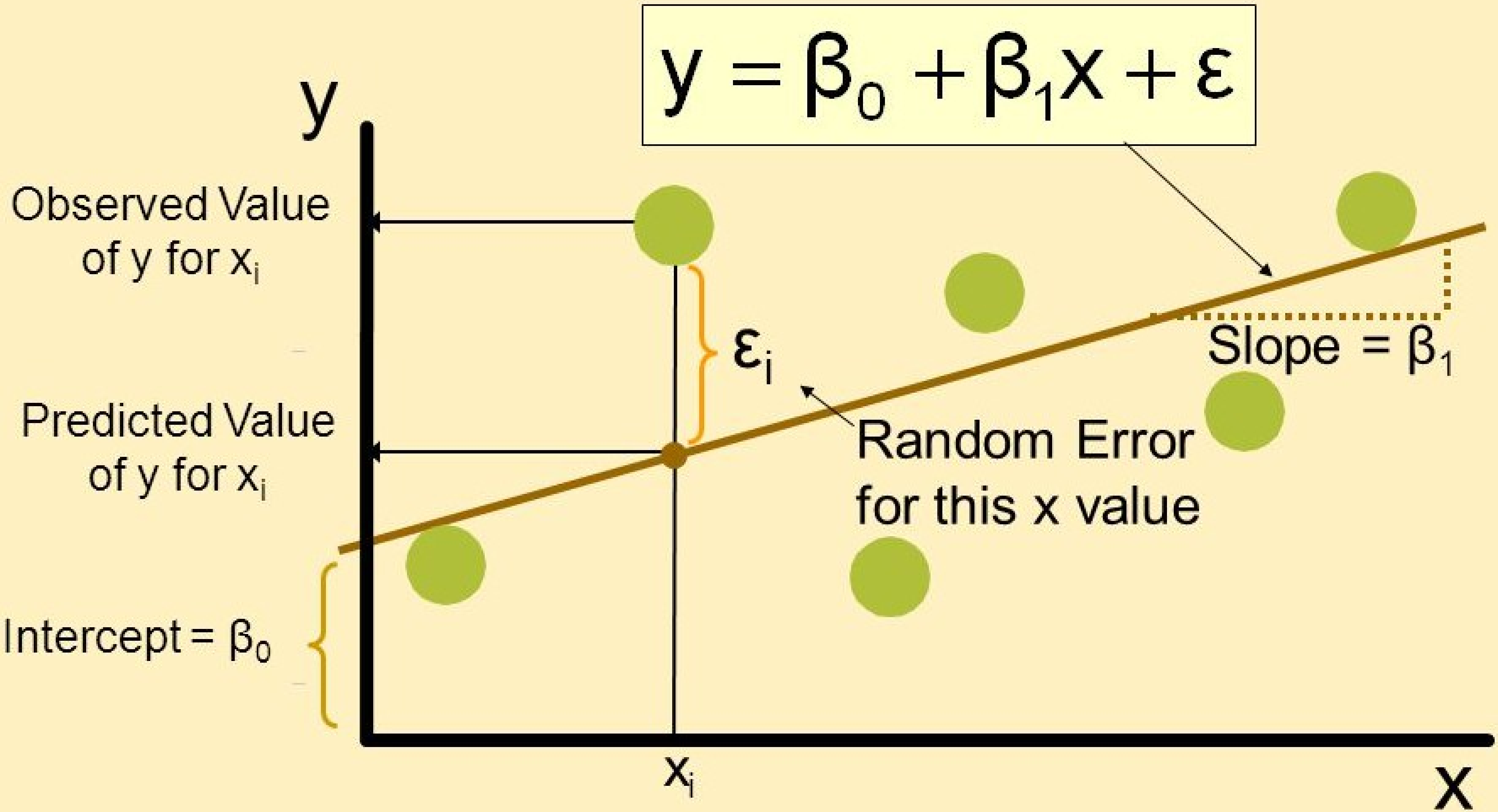
USE IN ORGANIZATION

In the field of business regression is widely used. Businessman are interested in predicting future production, consumption, investment, prices, profits, sales etc. So the success of a businessman depends on the correctness of the various estimates that he is required to make. It is also use in sociological study and economic planning to find the projections of population, birth rates. death rates etc.

METHODS OF STUDYING REGRESSION:



Population Linear Regression



Algebraically method-:

1.Least Square Method-:

The regression equation of X on Y is :

$$X = a + bY$$

Where,

X=Dependent variable

Y=Independent variable

The regression equation of Y on X is:

$$Y = a + bX$$

Where,

Y=Dependent variable

X=Independent variable

And the values of a and b in the above equations are found by the method of least of Squares-reference . The values of a and b are found with the help of normal equations given below:

(I)

$$\sum X = na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

(II)

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Example1-: From the following data obtain the two regression equations using the method of Least Squares.

X	3	2	7	4	8
Y	6	1	8	5	9

Solution-:

X	Y	XY	X^2	Y^2
3	6	18	9	36
2	1	2	4	1
7	8	56	49	64
4	5	20	16	25
8	9	72	64	81
$\sum X = 24$	$\sum Y = 29$	$\sum XY = 168$	$\sum X^2 = 142$	$\sum Y^2 = 207$

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Substitution the values from the table we get

29=5a+24b…………………(i)
168=24a+142b
84=12a+71b………………..(ii)

Multiplying equation (i) by 12 and (ii) by 5

348=60a+288b………………(iii)
420=60a+355b………………(iv)

By solving equation(iii)and (iv) we get

a=0.66 and b=1.07

By putting the value of a and b in the **Regression equation Y on X** we get

Y=0.66+1.07X

Now to find the regression equation of X on Y ,
The two normal equation are

$$\sum X = na + b\sum Y$$
$$\sum XY = a\sum Y + b\sum Y^2$$

Substituting the values in the equations we get

24=5a+29b………………………(i)
168=29a+207b…………………..(ii)

Multiplying equation (i)by 29 and in (ii) by 5 we get

a=0.49 and b=0.74

Substituting the values of a and b in the **Regression equation X and Y**

$$X=0.49+0.74Y$$

2.Deaviation from the Arithmetic mean method:

The calculation by the least squares method are quit cumbersome when the values of X and Y are large. So the work can be simplified by using this method.
The formula for the calculation of Regression Equations by this method:

Regression Equation of Y on X- $(X-\bar{X})=b_{xy}(Y-\bar{Y})$
Regression Equation of X on Y- $(Y-\bar{Y})=b_{yx}(X-\bar{X})$

Where, b_{xy} and b_{yx} = Regression Coefficient

$$b_{xy}=\frac{\sum xy}{\sum y^2} \quad \text{and} \quad b_{yx}=\frac{\sum xy}{\sum x^2}$$

Example2-: from the previous data obtain the regression equations by Taking deviations from the actual means of X and Y series.

X	**3**	**2**	**7**	**4**	**8**
Y	6	1	8	5	9

Solution-:

X	**Y**	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	**xy**
3	6	-1.8	0.2	3.24	0.04	-0.36
2	1	-2.8	-4.8	7.84	23.04	13.44
7	8	2.2	2.2	4.84	4.84	4.84
4	5	-0.8	-0.8	0.64	0.64	0.64
8	9	3.2	3.2	10.24	10.24	10.24
$\sum X = 24$	$\sum Y = 29$	$\sum x = 0$	$\sum y = 0$	$\sum x^2 = 26.8$	$\sum y^2 = 38.8$	$\sum xy = 28.8$

Regression Equation of X on Y is

$$(X-\bar{X})=b_{xy}(Y-\bar{Y})$$

$$b_{xy}=\frac{\sum xy}{\sum y^2}$$

$$X-4.8=\frac{28.8}{38.8}(Y-5.8)$$

$$X-4.8=0.74(Y-5.8)$$

$$X=0.74Y+0.49 \quad \text{.............(I)}$$

Regression Equation of Y on X is

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

$$b_{yx}=\frac{\sum xy}{\sum x^2}$$

$$Y-5.8=\frac{28.8}{26.8}(X-4.8)$$

$$Y-5.8=1.07(X-4.8)$$

$$Y=1.07X+0.66 \quad \text{.............(II)}$$

It would be observed that these regression equations are same as those obtained by the direct method .

3.Deviation from Assumed mean method-:

When actual mean of X and Y variables are in fractions ,the calculations can be simplified by taking the deviations from the assumed mean.

The Regression Equation of X on Y-:

$$(X-\bar{X}) = b_{xy}(Y-\bar{Y})$$

The Regression Equation of Y on X-:

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

But , here the values of b_{xy} and b_{yx} will be calculated by following formula:

$$b_{xy} = \frac{N\sum d_x d_y - \sum d_x \sum d_y}{N\sum d_y^2 - \left(\sum d_y\right)^2}$$

$$b_{yx} = \frac{N\sum d_x d_y - \sum d_x \sum d_y}{N\sum d_x^2 - \left(\sum d_x\right)^2}$$

Example-: **F**rom the data given in previous example calculate regression equations by assuming 7 as the mean of X series and 6 as the mean of Y series.

Solution-:

X	**Y**	**Dev. From assu. Mean 7 (d_x)=X-7**	d_x^2	**Dev. From assu. Mean 6 (d_y)=Y-6**	d_y^2	$d_x d_y$
3	6	-4	16	0	0	0
2	1	-5	25	-5	25	+25
7	8	0	0	2	4	0
4	5	-3	9	-1	1	+3
8	9	1	1	3	9	+3
$\sum X = 24$	$\sum Y = 29$	$\sum d_x = -11$	$\sum d_x^2 = 51$	$\sum d_y = -1$	$\sum d_y^2 = 39$	$\sum d_x d_y = 31$

$$\bar{X} = \frac{\sum X}{N} \Rightarrow \bar{X} = \frac{24}{5} = 4.8 \qquad \bar{Y} = \frac{\sum Y}{N} \Rightarrow \bar{Y} = \frac{29}{5} = 5.8$$

The Regression Coefficient of X on Y-:

$$b_{xy} = \frac{N\sum d_x d_y - \sum d_x \sum d_y}{N\sum d_y^2 - \left(\sum d_y\right)^2}$$

$$b_{xy} = \frac{5(31) - (-11)(-1)}{5(39) - (-1)^2}$$

$$b_{xy} = \frac{155 - 11}{195 - 1}$$

$$b_{xy} = \frac{144}{194}$$

$$b_{xy} = 0.74$$

The Regression equation of X on Y-: $(X - \bar{X}) = b_{xy}(Y - \bar{Y})$

$$(X - 4.8) = 0.74(Y - 5.8)$$

$$X = 0.74Y + 0.49$$

The Regression coefficient of Y on X-:

$$b_{yx} = \frac{N\sum d_x d_y - \sum d_x \sum d_y}{N\sum d_x^2 - \left(\sum d_x\right)^2}$$

$$b_{yx} = \frac{5(31) - (-11)(-1)}{5(51) - (-11)^2}$$

$$b_{yx} = \frac{155 - 11}{255 - 121}$$

$$b_{yx} = \frac{144}{134}$$

$$b_{yx} = 1.07$$

The Regression Equation of Y on X-:

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

$$(Y - 5.8) = 1.07(X - 4.8)$$

$$Y = 1.07X + 0.66$$

It would be observed the these regression equations are same as those obtained by the least squares method and deviation from arithmetic mean .

The Standard Error of Estimate

- The **standard error of estimate** measures the scatter, or dispersion, of the observed values around the line of regression
- Formulas used to compute the standard error:

$$SE_{YX} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{n-2}} \quad \text{or} \quad SE_{YX} = \sqrt{\frac{\Sigma Y^2 - a(\Sigma Y) - b(\Sigma XY)}{n-2}}$$

$$SE_{XY} = \sqrt{\frac{\Sigma(X - X_c)^2}{n-2}} \quad \text{or} \quad SE_{XY} = \sqrt{\frac{\Sigma X^2 - a(\Sigma X) - b(\Sigma XY)}{n-2}}$$

THE COEFFICIENT OF DETERMINATION

It is the primary way we can measure the extent or strength of the association that exists between two variables x & y . Because we have used a sample of points to develop regression lines .

It is denoted by r^2

THE COEFFICIENT OF DETERMINATION

$$\text{Total Variation} = \text{Explained Variation} + \text{Unexplained Variation}$$

$$\sum (Y - \overline{Y})^2 = \sum (Y_c - \overline{Y})^2 + \sum (Y - Y_c)^2$$

$$\text{Coefficient of Determination} = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$r^2 = \frac{SSR}{SST}$$

Where SSR $\Rightarrow$ Regression Sum of Square

SST $\Rightarrow$ Total Sum of Square

Coefficient of Determination

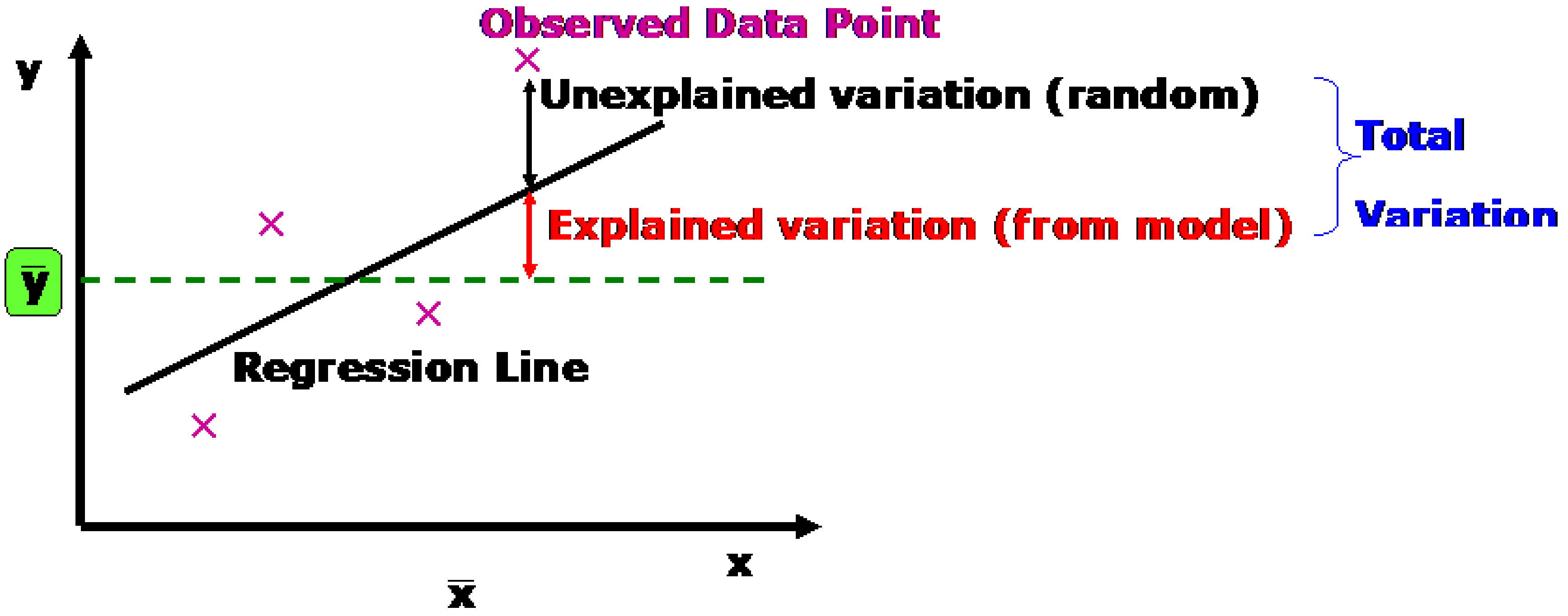
- Relationship Among SST, SSR, SSE

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

where:

SST = total sum of squares
SSR = sum of squares due to regression
SSE = sum of squares due to error


Observed Data Point
y
Unexplained variation (random)
Total
Variation
Explained variation (from model)
ȳ
Regression Line
x
x̄