11/01/2020

Unit - III

Regression :

Definition :

Regression is the measure of the average relationship between two or more variables in terms of original units of data.

Example:

If the sales and advertising are correlated we can find out the expected amount of sales for a given adversti -sing expenditure or the amount needed for attaining the given amount of sales

Lines of Regression

If two variables X and Y are Correlated there exists an association beteween them. We can see that the scatter diagram will be more or less

Concentrated around a Curve. This Curve is called Curve of regression.

If the Curve is a straight line, it is called the line of regression and the regression is a linnear regression.

We shall have two regression lines as the regression line of X and Y and the regression Line of Y and X.

The regression Line of Y and X gives the most Problem Probable value of Y for given values of X and the regression line of X and Y gives the most Probable values of X for given values of Y.

Relation between Correlation Analysis And Regression Analysis...

S. NO	Correlation Analysis	Regression Analysis
1.	Correlation Coefficient γ between x and y is a measure of linear relationship bet -ween X and Y	The Regression coeffici -ents are mathematical measure expressing the average relationship bet -ween the two variables.
2.	The correlation coefficient does not reflect upon the nature of variable (independent or dependent variable).	Regression coefficients reflects on the nature of variable which is dependent variable. In other words, it estima -tes the value of dependent Variable for any given value of independent Variable.
3.	It does not imply cause and effect relationship between the variables under study.	It indicates the cause and effect relationship between the variables. The variables corresponding to cause is taken as independent variable Where as corresponding to eff -ect is taken as dependent vari able
4.	It is a relative measure and is independent of the units of measurement.	Regression coefficient are absolute measures of finding out the relationship between two or more variables.
5.	It indicates the degree of association.	It is used to forecast the nature of dependent variable when the value of independent variable is known.

Uses of Regression Analysis.

1. The cause and effect relations are indicated from the study of regression Analysis.

2. It establishes the rate of change in one variable in terms of the changes in another variable.

3. It is useful in economic analysis as regression equation can determine an increase in the cost of living index for a Particular increase in general Price level.

4. It helps in Prediction and thus it can estimate the value of unknown quantities.

5. It enables us to study the nature of relationship between the variables.

6. It can useful to all natural, social and physical sciences where the data are in functional relationship.

Regression Equations.

(i). Equation of line of regression of Y on X is

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

Where $r \frac{\sigma_y}{\sigma_x}$ is the regression coefficient of Y on X.

(ii). Equation of line of regression of x on y is

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

Where $r \frac{\sigma_x}{\sigma_y}$ is the regression co-efficient of x on Y.

Note :

1. The regression coefficients can be denoted by

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} \quad \text{and} \quad b_{xy} = r \frac{\sigma_x}{\sigma_y}$$

2. The regression co-efficients are obtained by the following expressions for discrete values of x and y

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{n \Sigma x^2 - (\Sigma x)^2}$$

$$b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{n \Sigma y^2 - (\Sigma y)^2}.$$

3. Both the regression lines pass through the Point $(\bar{x}, \bar{y})$ where $\bar{x}$ and $\bar{y}$ are means of x and y respectively.

4. Correlation coefficient is the Geometric mean between the regression coefficients.

$$b_{xy} \cdot b_{yx} = r^2 \Rightarrow r = \pm \sqrt{b_{xy} \cdot b_{yx}}$$

5. If one of the regression coefficients is greater than unity the other must be less than unity.

6. Regression coefficient are independent of the change of origin but not of scale.

7. Both the regression coefficients will have the same sign, they will be either both Positive or both Negative.

The coefficients correlation will have the same sign as that of regression coefficients, if regression coefficients have a Negative sign, r will also have negative sign and if the regression coefficients have a Positive sign, r will also be Positive

Angle between Regression lines.

If θ is the angle between the two regression lines, then

$$\tan\theta = \left(\frac{1 - r^2}{r}\right) \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

Where r, σ_x, σ_y have the usual meaning.

Proof :

Equation of the regression lines of Y on X and X on Y are.

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

Slope of the two lines are

$$m_1 = b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$$

$$m_2 = \frac{1}{b_{xy}} = \frac{1}{r} \cdot \frac{\sigma_y}{\sigma_x}$$

If θ is the angle, then,

$$\tan\theta = \frac{|m_1 - m_2|}{1 + m_1 m_2}$$

$$= \frac{\left| r \cdot \frac{\sigma_y}{\sigma_x} - \frac{1}{r}\frac{\sigma_y}{\sigma_x} \right|}{1 + \left(r \cdot \frac{\sigma_y}{\sigma_x}\right)\left(\frac{1}{r} \cdot \frac{\sigma_y}{\sigma_x}\right)}$$

$$= \frac{\left| r - \frac{1}{r} \right| \frac{\sigma_y}{\sigma_x} \cdot \sigma_x^2}{\sigma_x^2 + \sigma_y^2}$$

$$\tan\theta = \frac{\left| r - \frac{1}{r} \right| \sigma_y \sigma_x}{\sigma_x^2 + \sigma_y^2}$$

Since $r^2 \leq 1$ and σ_x and σ_y are Positive, the anagle between the lines is

$$\tan\theta = \frac{1 - r^2}{r} \cdot \frac{\sigma_x + \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

Note:

(i). Suppose $r = 0$. Then $\tan\theta = \infty$.

$\theta = \frac{\pi}{2} = 90^\circ$

The two regression lines are perpendi-cular to each other and the equations will be

$y = \bar{y}$ and $x = \bar{x}$.

(ii). If $r = \pm 1$, then $\tan\theta = 0 \Rightarrow \theta = 0$ or π

Here the lines of regression coincide. They cannot be parallel since they have a common point $(\bar{x}, \bar{y})$.

Solved Problem 5.12

Calculate the coefficient of correlation and obtain the lines of regression for the following.

X :	1	2	3	4	5	6	7	8	9
Y :	9	8	10	12	11	13	14	16	15

obtain an estimate of y which should correspond to the value $x = 6.2$.

Solution:

x	y	x^2	y^2	xy
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
8	16	64	256	128
9	15	81	225	135
$\Sigma x = 45$	$\Sigma y = 108$	$\Sigma x^2 = 285$	$\Sigma y^2 = 1356$	$\Sigma xy = 597$

$n = 9$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{108}{9} = 12$$

Correlation Coefficient.

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{9 \times 597 - (45)(108)}{\sqrt{9(285) - (45)^2}\sqrt{9(1356) - (108)^2}}$$

$$= \frac{5373 - 4860}{\sqrt{731025 - 81225}\sqrt{16548625 - 18387736}}$$

$$= \frac{513}{1.11700089}$$

$$\boxed{r = 0.95}$$

Regression Coefficient of X on Y

$$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$$

$$= \frac{9 \times 597 - (45 \times 108)}{(9 \times 1356) - (108)^2}$$

$$= \frac{513}{540} \Rightarrow 0.95$$

$$\boxed{b_{xy} = 0.95}$$

Regression Coefficient of Y on X

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{9 \times 597 - 45 \times 108}{9 \times 285 - (45)^2}$$

$$= \frac{513}{540}.$$

$$\boxed{b_{yx} = 0.95}$$

Regression Line of X on Y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 5 = 0.95(y - 12)$$

$$x = 0.95y - (11.4 - 5)$$

$$x = 0.95y - 6.4.$$

Regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 12 = 0.95(x - 5)$$

$$= 0.95x - 4.75$$

$$y = 0.95x + 7.25$$

Value of y corresponding to $x = 6.2$ is

$$y = (0.95 \times 6.2) + 7.25$$

$$= 5.89 + 7.25$$

$$\boxed{y = 13.14}$$

8/10/2020

Illustration 5

a). The following table shows the ages (x) and blood pressure (y) of 8 persons.

x :	52	63	45	36	72	65	47	25
y :	62	53	51	25	79	43	60	33

Obtain the regression equation of y on x and find the expected blood Pressure of a Person who is 49 years old.

Solution:

X	Y	x^2	y^2	XY
52	62	2704	3844	3224
63	53	3969	2809	3339
45	51	2025	2601	2295
36	25	1296	625	900
72	79	5184	6241	5688
65	43	4225	1849	2795
47	60	2209	3600	2820
25	33	625	1089	825
$\Sigma x = 405$	$\Sigma y = 406$	$\Sigma x^2 = 22237$	$\Sigma y^2 = 22658$	$\Sigma xy = 21886$

$$\bar{X} = \frac{\Sigma x}{n} = \frac{405}{8} = 50.625$$

$$\bar{Y} = \frac{\Sigma y}{n} = \frac{406}{8} = 50.75$$

Regression coefficient of Y on X

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{8 \times 21886 - (405)(406)}{8(22237) - (406)^2 \quad 164025}$$

$$= \frac{10658}{13871}$$

$$b_{yx} = 0.768$$

Regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 50.75 = 0.768(50.625 - x)$$

$$= 0.768(354.375)$$

$$y - 50.75 = 0.768x - 38.88$$

$$y = 11.87 + 0.768(49)$$

$$= 49.502.$$

There, the expected bloodpressure of a Person who is 49 years old shall be 49.50

Solved Problem : 5.18

From the following data, find the equations of the regression lines

	Marks in Mathematics	Marks in English
mean	62.5	39
S.D	9.5	10

Co-efficient of correlation between marks in mathematics and English = 0.60.

1. Estimate the marks in English when marks mathematics is 70

2. Estimate the marks on mathematicals corresponding to 54 marks in English

Solution:

Regression Eqnation of Y on X

$$y - \bar{y} = {}^{*}r \frac{\sigma y}{\sigma x} (x - \bar{x})$$

$$y - 67 = 0.8 \times \frac{3.5}{2.5} (x - 65)$$

$$y - 67 = 1.12x - 72.8$$

$$y = 1.12x - 72.8 + 67$$

$$y = 1.12x - 5.8$$

Let marks in mathematics be "x"

Let marks in English be "y"

$$\bar{x} = 62.5, \quad \bar{y} = 39, \quad \sigma x = 9.5, \quad \sigma y = 10, \quad r = 0.60.$$

Regression eqnation of X on Y

$$x - \bar{x} = r \frac{\sigma x}{\sigma y} (y - \bar{y})$$

$$x - 62.5 = 0.60 \times \frac{9.5}{10} (y - 39)$$

$$x - 62.5 = 0.57y - 22.28$$

$$x = 0.57y - 22.23 + 62.5$$

$$x = 0.57y + 40.29 \quad \text{—} (1)$$

Regression equation of Y on X

$$y - \bar{y} = r \frac{\sigma y}{\sigma x} (x - \bar{x})$$

$$y - 39 = 0.6 \times \frac{10}{9.5} (x - 62.5)$$

$$y - 39 = 0.63x - 39.375$$

$$y = 0.63x - 39.375 + 39$$

$$y = 0.63x - 0.375 \quad -(2)$$

(1) When $x = 70$, $y = ?$

From equation (2)

$$y = 0.63x - 0.375$$

$$y = 0.63(70) - 0.375$$

$$y = 44.1 - 0.375$$

$$y = 43.725.$$

From equation (2) when $y = 54$, $x = ?$

$$x = 0.57y + 40.27$$

$$x = 0.57(54) + 40.27$$

$$x = 30.78 + 40.27$$

$$x = 71.05$$

29/10/2020

CURVE FITTING

* The equation of the curve of "Best Fit" which may be most suitable for Predicting the unknown values.

* A Smooth ~~Curve~~ Curve that approximates the above set of Points in known as the approximating curve.

* $y = f(x)$ is called empirical equation.

* This approximating curve is an empirical equation and the method of finding such an approximating curve is called curve fitting.

* The Principle of least Squars provides a fine Procedure of fitting a unique curve to a given data.

The residual at $x = x_i$ is

$$d_i = y_i - f(x_i) = y_i - (ax_i + b), \; i = 1, 2, 3, \cdots n$$

$$E = \sum_{i=1}^{n} d_i^2$$

$$= \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

By the Principle of least Squares,

E is minimum.

$$\frac{\partial E}{\partial a} = 0 \quad \text{and} \quad \frac{\partial E}{\partial b} = 0$$

i.e., $2E[y_i - (ax_i + b)](x_i) = 0$ and $2\Sigma[y_i - (ax_i + b)]$ (1)

i.e., $\sum_{i=1}^{n} (x_i y_i - a x_i^2 - b x_i) = 0$ and $\sum_{i=1}^{n} (y_i - a x_i - b) = 0$

$$a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \quad — (1)$$

and

$$a \sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i \quad — (2)$$

NOTE :

1. Equation (1) and (2) are called normal equation.
2. Dropping suffix i from (1) and (2) the normal equation are

$$a\Sigma x + nb = \Sigma y \quad \text{and} \quad a\Sigma x^2 + b\Sigma a = \Sigma xy.$$

Which are got by taking Σ on both side of $y = ax + b$ and also taking Σ on both sides after multiplying by x on both sides of $y = ax + b$

Example 1 :-

By the method of least Squares find the best fitting straight line to the data given below:

x :	15	10	15	20	25
y :	16	19	23	26	30

Solution :

Let the Straight line be $y = ax + b$.

The normal equations are,

$$a\Sigma x + 5b = \Sigma y$$

$$a\Sigma x^2 + b\Sigma x = \Sigma xy.$$

x	y	x^2	xy
5	16	25	80
10	19	100	190
15	23	225	345
20	26	400	520
25	30	625	750
Total = 45	114	1375	1885

$\Sigma x = 75$, $a = 75$, $a\Sigma x^2 = 1375$, $\Sigma y = 114$, $\Sigma xy = 1885$

The normal equations are:

$$75a + 5b = 114 \quad \text{—①}$$

$$1375a + 75b = 1885 \quad \text{—②}$$

Multiply ① by 15

$$1125a + 75b = 1710 \quad \text{—③}$$

② − ③ gives.

$$250a = 175 \quad \text{or} \quad a = 0.7$$

Hence

$$b = 12.3.$$

Hence the best fitting line is $y = 0.7x + 12.3$

Example 2:

Fit a straight line to the data given below. Also estimate the value of y at $x = 2.5$

x :	0	1	2	3	4
y :	1	1.8	3.3	4.5	6.3

The normal equations are $y = ax + b \rightarrow ①$

$$a\Sigma x + 5b = \Sigma y \longrightarrow ②$$

$$a\Sigma x^2 + b\Sigma x = \Sigma xy \longrightarrow ③$$

We prepare the table for easy use

x	y	x^2	xy
0	1.0	0	0
1	1.8	1	1.8
2	3.3	4	6.6
3	4.5	9	13.5
4	6.3	16	25.2
10	16.9	30	47.1

Substituting in (2) and (3) we get:

$$10a + 5b = 16.9$$

$$30a + 10b = 47.1$$

Solving, we get, $a = 1.33$, $b = 0.42$.

Hence the equation is $y = 1.33x + 0.42$.

y (at $x = 2.5$) $= 1.33 \times 2.5 + 0.42 = 4.045$

Iterative methods:-

* Gauss - seidel method of iteration:-

* Iterative method is a self - correcting method. i.e, any error made in computation is corrected in the subsequent iterations.