Unit-2

Correlation and Linear Regression

- **Correlation:-** Quantitative measure of relationship between two variables.
- **Positive / Direct correlation:-** If the increase / decrease in one variable results in increase / decrease in another variable.
  Ex. Income & Expenditure.
- **Negative / Diverse correlation:** If the increase / decrease in one variable results in decrease / increase in another variable.
  Ex. Price & Demand.
- **Karl pearsons coefficient of correlation.**

Correlation coefficient between two random variable's X and Y is denoted by $r(X,Y)$ and is numerical measure of relationship b/w X and Y.

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

where $Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

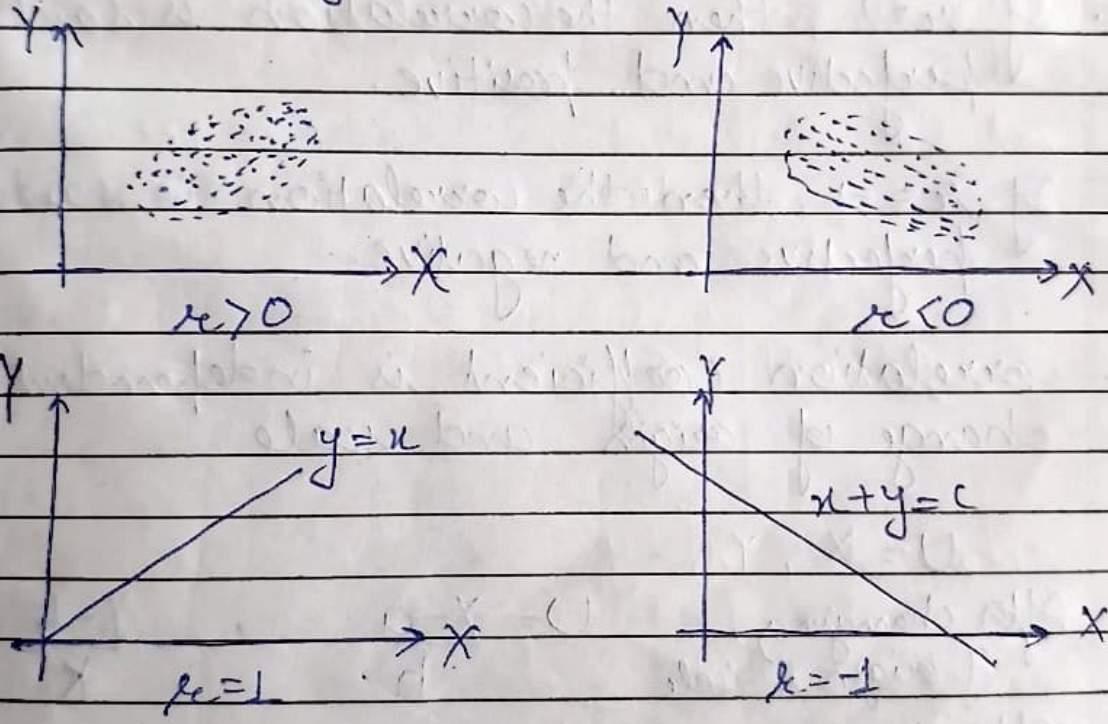
Date.
Page.

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2$$

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y})^2$$

X	Y	(X, Y)
x_1	y_1	(x_1, y_1)
x_2	y_2	(x_2, y_2)
x_3	y_3	(x_3, y_3)
⋮	⋮	⋮
x_i	y_i	(x_i, y_i)
⋮	⋮	⋮
x_n	y_n	(x_n, y_n)

- **Scatter Diagrams:**



* Properties of Correlation :-

1. $r(X, Y)$ is a measure of linear relationship between X and Y.

2. Karl Pearson's correlation is also called product moment correlation coefficient.

3. $-1 \leq r(X, Y) \leq 1$

4. If $r = 1$, then the correlation is said to be perfective and positive.

5. If $r = -1$, then the correlation is said to be perfective and negative.

6. Correlation coefficient is independent change of origin and scale.

$U = X, Y$

After changing, of origin & scale $U = \frac{X-a}{h}, \frac{Y-b}{K}$

~~After changing~~

7. If X and Y are random variables & a, b, c, d are any numbers provided $a \neq 0$ $c \neq 0$ then,

$$r(aX+b, cY+d) = \frac{ac}{|ac|} r(X, Y)$$

$$r(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Date :
Page :

8. If X and Y are independent then $r(X,Y)=0$ but converse is not true.

Q. Calculate the coefficient of correlation b/w X & Y.

$$\sigma_x^2 = \frac{1}{n}\left(\left(\sum x_i^2\right) - \bar{x}^2\right)$$

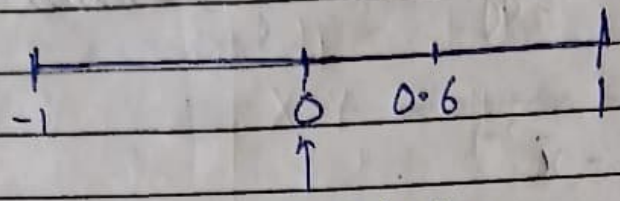
$$\sigma_y^2 = \frac{1}{n}\left(\left(\sum y_i^2\right) - \bar{y}^2\right)$$

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\sum x_i^2 - (\bar{x})^2\right)\left(\frac{1}{n}\sum y_i^2 - (\bar{y})^2\right)}}$$

(Ht. of Father) X	(Ht. of Son) Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4838
72	71	5184	5041	5112
$\Sigma X = 544$	$\Sigma Y = 552$	Σx_i^2	Σy_i^2	$\Sigma x_i y_i$
$\bar{x} = 68$	$\bar{y} = 69$	37628	38132	37560

$$\therefore r(X,Y) = \frac{\frac{1}{8}\times 37628 - 68\times 69}{\sqrt{\left(\frac{1}{8}\times 37628 - (68)^2\right)\left(\frac{1}{8}\times 38132 - (69)^2\right)}} = 0.603$$

1. X — Height of Father; Y — Height of Sons



* Spearman's Rank correlation :-

Let (x_i, y_i) $i = 1, 2, 3, \ldots n$ be the rank of i^{th} individual in two characteristics A and B

then,

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

(Sigma, not sigma)

X (Characteristic A)	Y (Characteristic B)
x_1	y_1
x_2	y_2
x_3	y_3
⋮	⋮

where $d_i = x_i - y_i$

we always have $\sum_{i=1}^{n} d_i = 0$

Example: The ranks of 16 students in maths & Physics are given below. Calculate rank correlation coefficient.

X (Rank in math)	Y (Rank in Physics)	$d_i = x_i - y_i$	d_i^2
1	1	0	0
2	10	-8	64
3	3	0	0
4	4	0	0
5	5	0	0
6	7	-1	1
7	2	5	25
8	6	2	4
9	8	1	1
10	11	-1	1
11	15	-4	16
12	9	3	9
13	14	-1	1
14	12	2	4
15	16	-1	1
16	13	3	9
		$\Sigma d_i = 0$	$\Sigma d_i^2 = 136$

Should be zero always.

3)

Here, $n = 16$,

$$\rho = \frac{1 - 6 \cdot \sum_{i=1}^{n} d_i}{n(n^2-1)} = 1 - \frac{6 \times 136}{16(16^2-1)}$$

$$\rho = 0.8$$

This means

High positive correlation b/w X & Y

Ex. Ten participants in a musical test, were ranked by three judges X, Y, Z. Discuss which pair of judges have common liking in music.

X	Y	Z	d_1 X-Y	d_2 Y-Z	d_3 X-Z	d_1^2	d_2^2	d_3^2
1	3	6	-2	-3	-5	4	9	25
6	5	4	1	1	2	1	1	4
5	8	9	-3	-1	-4	9	1	16
10	4	8	6	-4	2	36	16	4
3	7	1	-4	6	2	16	36	4
2	10	2	-8	8	0	64	64	0
4	2	3	2	-1	1	4	1	1
9	1	10	8	-9	-1	64	81	1
7	6	5	1	1	2	1	1	4
8	9	7	-1	2	1	1	4	1
			0	0	0	200	214	60

We have to find ρ_{XY}, ρ_{YZ}, ρ_{XZ} & check which of these have +ve (positive) sign.

$$\rho_{XY} = 1 - 6 \cdot \frac{\sum_{i=1}^{n} d_1^2}{n(n^2-1)} = 1 - \frac{6 \times 200}{10(10^2-1)} = \frac{-7}{33}$$

$$\rho_{YZ} = 1 - 6 \cdot \frac{\sum_{i=1}^{n} d_2^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10(10^2-1)} = \frac{-49}{165}$$

$$\rho_{XZ} = 1 - 6 \cdot \frac{\sum_{i=1}^{n} d_3^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{10(10^2-1)} = \frac{7}{11}$$

So, X & Z are having common liking

Date :
Page :

* Rank correlation in case of tied ranks or repeated ranks :-

$$\rho = 1 - \frac{6 \cdot \left(\sum_{i=1}^{n} d_i^2 + \text{Correction factor} \right)}{n(n^2-1)}$$

where correction factor $= \frac{m(m^2-1)}{12}$, where m is number of times an item repeated

Ex. Calculate the rank correlation for following data

X	Y	Rank(X)	Rank(Y)	d(X-Y)	d^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16

For 75 as rank we take 2+3 / 2 = 2.5

1
2
3
4
5
6
7
8
9
10

$$\text{Correction factor} = \frac{m_1(m_1^2-1)}{12} + \frac{m_2(m_2^2-1)}{12} + \frac{m_3(m_3^2-1)}{12}$$

(75) $m_1 = 2 \Rightarrow 75 \Rightarrow 2$ times

(64) $m_2 = 3 \Rightarrow 64 \Rightarrow 3$ times

(68) $m_3 = 2 \Rightarrow 68 \Rightarrow 2$ times

Date
Page.

* ~~Linear Regression~~ :

* Regression : It is the study of nature of relationship between the variables so that one is able to predict the value of one variable on the basis of other. In regression analysis one variable is dependent and of another is independent.

* Linear Regression : If the variables in the bivariate distribution are related, we will find that points (x_i, y_i) $i = 1, 2, \ldots, n$ cluster around some curve, that curve is called curve of Regression.

If the curve is straight line, then it is called the linear regression.

Suppose we have n data points (x_i, y_i), $i = 1, 2, \ldots n$

Let the line of regression of Y on X be

$$y = a + bx$$

and line of regression of X on Y be

$$x = c + dy$$

We have line of regression of Y on X $\quad y = a + bx$

$$y = a + bx \quad \text{---(1)}$$

$$\Sigma y = na + b\Sigma x \quad \text{---(2)}$$

$y_1 = a + bx_1$

$y_2 = a + bx_2$

$\vdots$

$y_n = a + bx_n$

$\Sigma y_i = na + b\Sigma x_i$

Date :
Page :

Multiply eq^n ① by x & take summation,

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{---} ③$$

Eq^n ② & ③ are called Normal equations of y on x and are used to determine values of a & b.

Multiply eq^n ② by $\frac{1}{n}$

$$\frac{1}{n}\Sigma y = a + b\frac{1}{n}\Sigma x$$

$$\boxed{\bar{y} = a + b\bar{x}}$$

Line of regression of y on x passes through $(\bar{x}, \bar{y})$.

Now line of regression of x on y,

$$x = c + dy \quad \text{---} ④$$

$$\Sigma x = nc + d\Sigma y \quad \text{---} ⑤$$

Multiply ④ by y & take summation,

$$\Sigma xy = c\Sigma y + d\Sigma y^2 \quad \text{---} ⑥$$

Date / Page /

Equation (5) & (6) are normal equation of X on Y.

Multiply eqn (5) by $\frac{1}{n}$.

$$\frac{1}{n}\Sigma x = c + d\frac{1}{n}\Sigma y$$

$$\bar{x} = c + d\bar{y}$$

⇒ The line of regression of X on Y, also passes through $(\bar{x}, \bar{y})$.

Q. Find the line of regression for the following data

X	Y	X^2	Y^2	XY
1	8	1	64	8
2	7	4	49	14
3	5	9	25	15
4	9	16	81	36
5	11	25	121	55
15	40	55	340	128

Let the line of regression of Y on X be $y = a + bx$

then normal equation of Y on X are:

$$\Sigma y = na + b\Sigma x$$

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

$$\left.\begin{aligned} 40a &= 5a + 15b \\ 128 &= 15a + 55b \end{aligned}\right\} \quad a = \frac{28}{5},\ b = \frac{4}{5}$$

Date.
Page.

∴ Line of regression of Y on X is :-

$$\boxed{y = \frac{2}{5}x + \frac{4}{5}x}$$

Let the line of regression of X on Y be $x = c + dy$ then Normal equation are :-

$$\Sigma x = nc + d\Sigma y$$
$$\Sigma xy = c\Sigma y + d\Sigma y^2$$
$$15 = 5c + 40d$$
$$128 = 40c + 340d$$

$$c = \frac{-1}{5}, \quad d = \frac{2}{5}$$

Equation of line of regression of X on Y is

$$\boxed{x = \frac{-1}{5} + \frac{2}{5}y}.$$

(3,8) ~~$y = \frac{4}{5}$~~

* Another form of regression lines:

The line of regression of Y on X is given by

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

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

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

Also line of regression of X on Y is given by

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

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

where r is correlation coefficient of X & Y

σ_x is standard deviation of X

σ_y is standard deviation of Y.

And b_{xy} & b_{yx} are called regression coefficients.

Date :
Page :

* Properties of Regression coefficient

1. Correlation coefficient r of X & Y is geometric mean between regression coefficient. The sign of r, b_{xy}, b_{yx} is same.

$$r^2 = b_{xy}\, b_{yx}$$

$$r = \pm\sqrt{b_{xy}\, b_{yx}}$$

2. If one of the regression coefficient is greater than unity, the other must be less than unity.

$$-1 \le r \le 1$$

3. $$\left|\frac{b_{xy} + b_{yx}}{2}\right| \ge |r|$$

$$(A.M. \ge G.M.)$$

4. Regression coefficients are independent of change of origin but not of scale.

NOTE: Angle b/w two lines (straight) :

$$y_1 = m_1 x + c_1$$
$$y_2 = m_2 x + c_2$$

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

Date
Page

* Angle between two lines of Regression:

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

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

then,

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

Case-1: If $r=0$, then $\theta = \frac{\pi}{2}$

This means if X and Y are uncorrelated then lines of regression are perpendicular.

Case-2: If $r = \pm 1$, then $\theta = 0$ or $\theta = \pi$

In case of perfect correlation positive or negative the lines of regression coincide.

Example: For variables X and Y,

$\sigma_x^2 = 9$,

Regression equations:

$8X - 10Y + 66 = 0$ (Y on X)

$40X - 18Y = 214$ (X on Y)

Find (1) mean value of X and Y

(2) Correlation coefficients b/w X & Y

(3) Standard deviation of Y.

Date.
Page.

Ans. Given $\sigma_x^2 = 9$

$\sigma_x = 3$

(1) Since $(\bar{x}, \bar{y})$ is point of intersection of lines of regression

$\therefore 8\bar{x} - 10\bar{y} + 66 = 0$

$40\bar{x} - 18\bar{y} = 214$

$\bar{x} = 13, \quad \bar{y} = 17$

(2)

From (1) & (2),

$Y = \frac{8}{10} x + \frac{66}{10}$

$X = \frac{18}{40} y + \frac{214}{40}$

We get, $b_{yx} = \frac{8}{10}$, $b_{xy} = \frac{18}{40}$

Also, $r = \pm \sqrt{b_{xy}\, b_{yx}}$

$r = \pm \sqrt{b_{xy}\, b_{yx}} = \pm \sqrt{\frac{8}{10} \times \frac{18}{40}}$

$r = \pm \frac{3}{5}$

Since r, b_{xy}, b_{yx} have same sign -

$\therefore r = +\frac{3}{5}$

(3) $b_{yx} = r \frac{\sigma_y}{\sigma_x}$ $\quad \sigma_x = 3$ (Given)

$\frac{8}{10} = \frac{3}{5} \times \frac{\sigma_y}{\sigma_x}$ $\quad \therefore \sigma_y = \frac{40}{10} \Rightarrow \boxed{\sigma_y = 4}$

Date
Page

Example: Find the most likely price in Mumbai corresponding to the prices of Rs. 70 in Calcutta from the following data.

	X	Y
	Calcutta	Mumbai
Avg. Price	65	67
S.D (Standard deviation)	2.5	3.5

Correlation between prices in two cities is 0.8.

Sol^n: $X = 70$ $Y = ?$

The line of regression of Y on X is given by

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

$\bar{x} = 65$ $\bar{y} = 67$

$\sigma_x = 2.5$ $\sigma_y = 3.5$

Line of regression of Y on X is:

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

At $x = 70$,

$$Y - 67 = 0.8 \times \frac{3.5}{2.5} (70 - 65)$$

$$Y = 72.6$$

Date.
Page.

1. $r(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$

$\sigma_{xy} = Cov(x,y) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}\right)$

$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \quad \bar{y} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)$

$\sigma_x = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - (\bar{x})^2\right)$

$\sigma_y = \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 - (\bar{y})^2\right)$

2. $\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$

$d_i = x_i - y_i, \quad \sum_{i=1}^{n} d_i = 0$

3. $\rho = 1 - \frac{6\left(\sum_{i=1}^{n} d_i^2 + \text{correction factor}\right)}{n(n^2-1)}$

$\text{Correction factor} = \frac{m(m^2-1)}{12}$

m → no. of times an item repeated

4. Line of Regression of Y on X be $y = a+bx$ & X on Y $x = c+dy$

Normal equations: Y on X:

$\Sigma y = na + b\Sigma x$

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

Normal eqns: X on Y

$\Sigma x = nc + d\Sigma y$

$\Sigma xy = c\Sigma y + d\Sigma y^2$

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

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

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

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

b_{yx} & b_{xy} are regression coefficients

6. Angle b/w two regression lines:-

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