Module – III Correlation and Regression

Correlation

Suppose two variables x and y are related in such a way that an increase in the value of one of these variables is accompanied by an increase or decrease in the value of the other. Such a relationship is called correlation.

If the values of x and y increase or decrease together, then we say that x and y are positively correlated. If the value of y decreases as the value of x increases or vice versa, then we say that x & y are negatively correlated.

The numerical measure of correlation between two variables x & y is known as Pearson's coefficient of correlation and is defined as

$$\boxed{r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}} \longrightarrow (1)$$

where $X = x - \bar{x}$ and $Y = y - \bar{y}$

and $\bar{x} = \frac{\Sigma x_i}{n}$; $\bar{y} = \frac{\Sigma y_i}{n}$; $\sigma_x^2 = \frac{\Sigma (x-\bar{x})^2}{n}$

$\sigma_y^2 = \frac{\Sigma (y-\bar{y})^2}{n}$, employing these expression in (1)

we get alternate formula

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

Alternative formula for the correlation

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x \sigma_y}$$

proof: Let $z = x - y$

$$\frac{\Sigma z}{n} = \frac{\Sigma x}{n} - \frac{\Sigma y}{n}$$

or $\bar{z} = \bar{x} - \bar{y}$

Hence $(z - \bar{z}) = (x - \bar{x}) - (y - \bar{y})$

Squaring both side, taking summation and dividing by n, we've

$$\frac{\Sigma (z - \bar{z})^2}{n} = \left[\frac{\Sigma (x - \bar{x})^2}{n}\right] + \left[\frac{\Sigma (y - \bar{y})^2}{n}\right] - 2\frac{\Sigma (x - \bar{x})(y - \bar{y})}{n}$$

$$\Rightarrow \sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$$

$$\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$$

$$\Rightarrow \sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2 = 2r\sigma_x\sigma_y$$

or $$\boxed{r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}}$$

Problems

(1) Find the correlation coefficient for the following data:

x	1	2	3	4	5
y	2	5	3	8	7

Soln:- Here $n=5$, we find that $\bar{x} = \frac{\Sigma x_i}{n} = \frac{1}{5}[1+2+3+4+5]$

$$\boxed{\bar{x} = 3}$$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{1}{5}[2+5+3+8+7]$$

$$\boxed{\bar{y} = 5}$$

We now prepare the following table

x_i	$X_i = x-\bar{x}$	X_i^2	y_i	$Y_i = y-\bar{y}$	Y_i^2	XY
1	-2	4	2	-3	9	6
2	-1	1	5	0	0	0
3	0	0	3	-2	4	0
4	1	1	8	3	9	3
5	2	4	7	2	4	4
		$\Sigma X_i^2 = 10$			$\Sigma Y_i^2 = 26$	13

Now we find that

$$\gamma = \frac{\Sigma XY}{\sqrt{\Sigma X_i^2 \, \Sigma Y_i^2}} = \frac{13}{\sqrt{10 \times 26}}$$

$$\boxed{\gamma = 0.806}$$

2) The following table gives the intelligence ratio (I.R) and engineering ability (E.A) of 10 students obtained on the basis of psychological tests. calculate the coefficient of correlation.

I.R (x) :	105	104	102	101	100	99	98	96	93	92
E.A (y) :	101	103	100	98	95	96	104	92	97	94

Soln:- Here $n = 10$, $\bar{x} = \frac{\Sigma x_i}{n} = \frac{990}{10} = 99$

$$\bar{y} = \frac{\Sigma y_i}{n} = \frac{980}{10} = 98$$

We now prepare the following table

x_i	$X_i = x_i - \bar{x}$	X_i^2	y_i	$Y_i = Y_i - y$	Y_i^2	$X_i Y_i$
105	6	36	101	3	9	18
104	5	25	103	5	25	25
102	3	9	100	2	4	6
101	2	4	98	0	0	0
100	1	1	95	-3	9	-3
99	0	0	96	-2	4	0
98	-1	1	104	6	36	-6
96	-3	9	92	-6	36	18
93	-6	36	97	-1	1	6
92	-7	49	94	-4	16	28
		170			140	92

We've

$$r = \frac{\Sigma X_i Y_i}{\sqrt{\Sigma X_i^2 \, \Sigma Y_i^2}} = \frac{92}{\sqrt{170 \times 140}} = 0.59$$

3. Employ the formula $\gamma = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x \sigma_y}$ do determine the correlation coefficient 'γ' for the following data :

x	92	89	87	86	83	77	71	63	53	50
y	86	83	91	77	68	85	52	82	37	57

Soln:- Here $n = 10$, $\bar{x} = \dfrac{751}{10} = 75.1$; $\bar{y} = \dfrac{718}{10} = 71.8$

Let $z_i = x_i - y_i$, $\bar{z} = \dfrac{1}{10} \Sigma z_i = \dfrac{1}{10} \Sigma (x_i - y_i)$

$$\bar{z} = \frac{33}{10} = 3.3$$

$$\sigma_x^2 = \frac{1}{n} \Sigma x_i^2 - (\bar{x})^2 = \frac{1}{10} \left[\{ (92)^2 + (89)^2 + (86)^2 + (83)^2 + (77)^2 + (71)^2 + (63)^2 + (53)^2 + (50)^2 \} \right] - (75.1)^2$$

$$= \frac{58487}{10} - (75.1)^2 = 208.69$$

$$\sigma_y^2 = \frac{1}{n} \Sigma y_i^2 - (\bar{y})^2 = \frac{54390}{10} - (71.8)^2 = 283.76$$

$$\sigma_{x-y}^2 = \sigma_z^2 = \frac{1}{n} \Sigma z_i^2 - (\bar{z})^2 = \frac{1485}{10} - (3.3)^2 = 137.61$$

Now $\gamma = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x \sigma_y} = \dfrac{208.69 + 283.76 - 137.61}{2 \times \sqrt{208.69 \times 283.76}}$

$$\boxed{\gamma = 0.729}$$

Regression

Suppose we are given n pairs of values (x_1, y_1) (x_2, y_2) .... (x_n, y_n) of two variables x & y. If we fit a straight line to this data by taking x as indepen dent variable and y as dependent variable, then the straight line obtained is called the line of regression of y on x. Similarly, if we fit a straight line to the data by taking y as independent variable and x as dependent variable, the line obtained is the line of regression of x on y.

We've Line of regression of y on x is

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

Line of regression of x on y is

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

Note :: We've alternative formula

(i) $r = \sqrt{(\text{coeff of } x)(\text{coeff of } y)}$ for coefficient of correlation

(ii) $Y = \dfrac{\Sigma XY}{\Sigma X^2} \cdot X$ & $X = \dfrac{\Sigma XY}{\Sigma Y^2} Y$ for lines of regression

Prove that if 'θ' is the angle between the lines of regression then $\tan\theta = \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\left(\frac{1-r^2}{r}\right)$

Pf:- W.K.T if 'θ' is acute, the angle b/w the lines $y = m_1 x + c_1$ & $y = m_2 x + c_2$ is given by

$$\tan\theta = \frac{m_2 - m_1}{1 + m_1 m_2} \longrightarrow (1)$$

We've

$$\left.\begin{aligned} m_1 &= r\frac{\sigma_y}{\sigma_x} \;; \quad m_2 = r\frac{\sigma_x}{\sigma_y} \\ &\qquad\qquad\quad = \frac{1}{r}\cdot\frac{\sigma_y}{\sigma_x} \end{aligned}\right\} \rightarrow (2)$$

Sub (2) in (1)

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

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

1. Find the coefficient of correlation and obtain the lines of regression for the following data:

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

Obtain an estimate for y which corresponds to $x = 6.2$

Sol:- We've $n = 9$, $\bar{x} = \frac{1}{n} \Sigma x_i = \frac{45}{9} = 5$; $\bar{y} = \frac{108}{9} = 12$

We prepare the following table

x_i	$X_i = x_i - \bar{x}$	X_i^2	y_i	$Y_i = y_i - \bar{y}$	Y_i^2	$X_i Y_i$
1	-4	16	9	-3	9	12
2	-3	9	8	-4	16	12
3	-2	4	10	-2	4	4
4	-1	1	12	0	0	0
5	0	0	11	-1	1	0
6	1	1	13	1	1	1
7	2	4	14	2	4	4
8	3	9	16	4	16	12
9	4	16	15	3	9	12
		60			60	57

Thus correlation coefficient is

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X_i^2 \, \Sigma Y_i^2}} = \frac{57}{60} = 0.95$$

Also $\sigma_x^2 = \frac{\Sigma X_i^2}{n} = \frac{60}{9} = 6.6667 \Rightarrow \sigma_x = \sqrt{6.6667}$

$= 2.5819$

$\sigma_y^2 = \frac{\Sigma Y_i^2}{n} = \frac{60}{9} = 6.6667$ $\qquad = \sigma_y$

Thus line of regression of y on x

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

$$y - 12 = 0.95 \cdot \frac{2.5819}{2.5819}(x - 5)$$

$$y - 12 = 0.95x - 4.75$$

$$\Rightarrow y = 0.95x - 4.75 + 12$$

$$\boxed{y = 0.95x + 7.25} \Rightarrow \boxed{y_{x=6.2} = 13.14}$$

Line of regression of x on y is

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

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

$$\Rightarrow \boxed{x = 0.95y - 6.4}$$

(2) Obtain the lines of regression and hence find the coefficient of correlation for the data:

x	1	3	4	2	5	8	9	10	13	15
y	8	6	10	8	12	16	16	10	32	32

Soln: Here $n = 10$, $\bar{x} = \frac{70}{10} = 7$; $\bar{y} = 15$

We've $X = x - \bar{x}$; $Y = y - \bar{y}$

We now prepare the foll. table

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
1	8	−6	−7	36	49	42
3	6	−4	−9	16	81	36
4	10	−3	−5	9	25	15
2	8	−5	−7	25	49	35
5	12	−2	−3	4	9	6
8	16	1	1	1	1	1
9	16	2	1	4	1	2
10	10	3	−5	9	25	−15
13	32	6	17	36	289	102
15	32	8	17	64	289	136
				204	818	360

We've line of regression of y on x

$$Y = \frac{\Sigma XY}{\Sigma X^2} \cdot X$$

$$y - \bar{y} = \frac{\Sigma XY}{\Sigma X^2} \cdot (x - \bar{x})$$

$$y - 15 = \frac{360}{204} (x - 7)$$

$$\boxed{y = 1.76x + 2.68}$$

Also line of regression of x on y is

$$x - \bar{x} = \frac{\Sigma XY}{\Sigma Y^2} (y - \bar{y})$$

$$x - 7 = \frac{360}{818} (y - 15)$$

$$\boxed{x = 0.44y + 0.4}$$

Thus

correlation coefficient $= r = \sqrt{(\text{coeff of } x)(\text{coeff of } y)}$

$$r = \sqrt{(1.76)(0.44)}$$

$$\boxed{r = 0.8813}$$

(8) Given

	x-Series	y-Series
Mean	18	100
SD	14	20

& $r = 0.8$

Write down the equation of the lines of regression only and hence find the most probable value when $x = 70$

Soln: By data $\bar{x} = 18$, $\bar{y} = 100$

$\sigma_x = 14$ $\sigma_y = 20$

Line of regression of y on x is

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

$$y - 100 = 0.8 \cdot \frac{20}{14}(x - 18)$$

$$\Rightarrow y = 1.14x + 79.48$$

when $x = 70$; $\boxed{y = 159.28}$

Line of regression of x on y is

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

$$x - 18 = 0.8 \frac{14}{20}(y - 100)$$

$$\boxed{x = 0.56y - 3.8}$$

(4) Given $8x - 10y + 66 = 0$, $40x - 18y = 214$ are two regression lines. Find the means of x's & y's and correlation coefficient. Find σ_y if $\sigma_x = 3$?

Soln:- We know that lines of regression passes through $\bar{x}$ & $\bar{y}$ $\Rightarrow$

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

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

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

Regression coefficients are

$8x - 10y + 66 = 0$; $40x - 18y = 214$

$10y = 8x + 66$; $40x = 18y + 214$

$\boxed{y = (0.8)x + 6.6}$; $\boxed{x = (0.45)y + 5.35}$

Thus coefficient of correlation $\Big\} = r = \sqrt{(\text{Coeff of } x)(\text{Coeff of } y)}$

$$= \sqrt{(0.8)(0.45)}$$

$$\boxed{r = 0.6}$$

By data $\sigma_x = 3$; $r\dfrac{\sigma_y}{\sigma_x}$ = Regression line y on x

$$0.6 \cdot \frac{\sigma_y}{3} = 0.8$$

$$\Rightarrow \sigma_y = \frac{0.8 \times 3}{0.6} = \underline{\underline{4}}$$

(5) If the coefficient of correlation between two variables x & y are 0.5, acute angle b/w their lines of regression is $\tan^{-1}(3/5)$. S.T $\sigma_y = 2\sigma_x$ or $\sigma_x = 2\sigma_y$

Sol: Give $r = 0.5$, $\theta = \tan^{-1}\left(\frac{3}{5}\right) \Rightarrow \tan\theta = \frac{3}{5}$

We've

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

Since $r = \frac{1}{2}$

$$\frac{3}{5} = \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\left(\frac{3}{4} \cdot 2\right)$$

$$\Rightarrow \frac{1}{5} = \frac{\sigma_x \sigma_y}{2(\sigma_x^2 + \sigma_y^2)}$$

$$\Rightarrow 2\sigma_x^2 + 2\sigma_y^2 = 5\sigma_x \sigma_y$$

$$\Rightarrow 2\sigma_x^2 - 5\sigma_x\sigma_y + 2\sigma_y^2 = 0$$

$$(2\sigma_x - \sigma_y)(\sigma_x - 2\sigma_y) = 0$$

$$\Rightarrow \boxed{2\sigma_x = \sigma_y}$$

or $\boxed{\sigma_x = 2\sigma_y}$

(6) In a bivariate distribution $\sigma_x = \sigma_y$ and the angle b/w the regression lines is $\tan^{-1}(3)$. Find the correlation coefficient.

Sol:- Given $\theta = \tan^{-1}(3) \Rightarrow \tan\theta = 3$ & $\sigma_x = \sigma_y$

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

$$3 = \frac{\sigma_x^2}{2\sigma_x^2}\left(\frac{1-r^2}{r}\right)$$

$$\Rightarrow 6r = 1 - r^2$$

or $r^2 + 6r - 1 = 0$

$r = 0.1623$ or $r = -6.6123$

But $|r| \leq 1 \Rightarrow \boxed{r = 0.1623}$

(7) If x & y are random variables with SD's, σ_x & σ_y. It was found that random variables $x+y$, $x-y$, $2x+y$ resp'ly have variance 15, 11, 29. compute SD's of x & y resp'ly and also the coefficient of correlation.

Soln: By data $\sigma^2_{x+y} = 15$, $\sigma^2_{x-y} = 11$, $\sigma^2_{2x+y} = 29$

we've $\sigma^2_{ax+by} = a^2\sigma_x^2 + b^2\sigma_y^2 + 2rab\sigma_x\sigma_y \rightarrow (1)$

Taking $(a,b) = (1,1)$ $(1,-1)$ & $(2,1)$
$\{x+y \quad x-y \quad 2x+y\}$

Eqn (1) $\Rightarrow$ $\sigma_x^2 + \sigma_y^2 + 2r\sigma_x\sigma_y = 15 \rightarrow (2)$

$\sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y = 11 \rightarrow (3)$

$4\sigma_x^2 + \sigma_y^2 + 4r\sigma_x\sigma_y = 29 \rightarrow (4)$

Eqns (2) + (3) gives

$2\sigma_x^2 + 2\sigma_y^2 = 26$

$\sigma_x^2 + \sigma_y^2 = 13 \rightarrow (5)$

Eqns (2) x (3) + (4) gives

$6\sigma_x^2 + 3\sigma_y^2 = 51$

$2\sigma_x^2 + \sigma_y^2 = 17 \rightarrow (6)$

on solving (5) & (6) $\sigma_x^2 = 4$; $\sigma_y^2 = 9$

$\boxed{\sigma_x = 2 \ ; \ \sigma_y = 3}$

sub these values in (2)

$4 + 9 + 2r \cdot 2 \times 3 = 15$

$\boxed{r = 0.17}$