Module-3

Correlation and Regression

Correlation and Regression - Rank correlation, partial and multiple correlation; Multiple regression

Correlation means that between two series or group of data there exist some casual connection. ie Correlation analysis deal with the association between two or more variables. and attemps to determine the 'degree of relationship' between variables.

"Correlation is an analysis of the covariance between two or more variables."

Types of correlation:-

(1) Positive or Negative
(2) Simple, partial and multiple
(3) Linear and non-linear

Positive or Negative correlation:-

Whether correlation is positive or negative, would depend upon the direction of change of variable. If both the variable are ~~variable~~ varing in the same direction ie if one of the variable is increasing and the other on an average is also increasing. or if one of the variable is decreasing and other is also decreasing then its called positive correlation. On the other hand, if variables are varing in oposite directio ie one is ↑ and other is ↓ then. correlation is said to be negative correlation.

Ex:- X: 10 12 15 18 20
Y: 15 20 22 25 37] → +ive correlation

X: 100 90 60 40 30
Y: 10 20 30 40 50] → Negative correlation

(2) Simple, Partial and multiple correlation:-

The distinction b/w these three correlation is based upon the no. of variables studied.

when two variables are studied, it is problem of simple correlation.

when three or more variables are studied it is a problem of either multiple or partial correlation. In multiple correlation three or more variables are studied simultaneously. Ex:- If one studies the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilizes used it is a problem of multiple correlation.

on the other hand, in partial correlation we recognise more than two variables but consider only two variables to be influencing each other, the effect of other influencing variables keeping constant. Ex:- If in above ex. if we limit our analysis to yield of rice and amount of rainfall, when a certain average daily temp existed, it becomes prob of partial correlation.

(3) Linear and ~~to~~ Non-linear (curvilinear) ~~Regres~~ Correlation :-

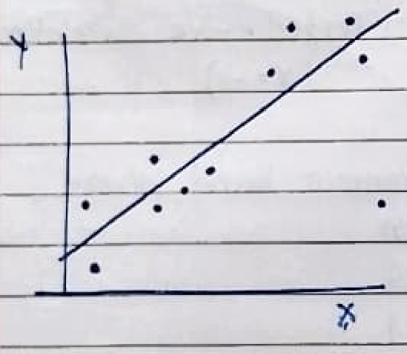
It is based on the ratio of change b/w the variables.
If the amount of change in one variable tends to be a constant ratio to the amount of change in the other variable the correlation is linear.

Ex

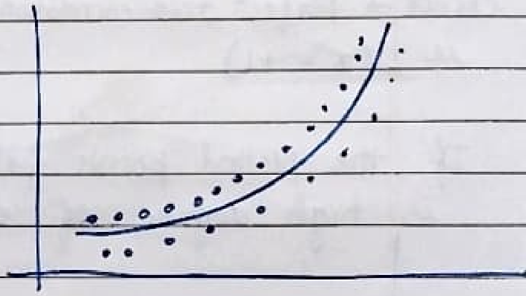
x	10	20	30	40	50	ratio is same
Y	70	140	210	280	350	

The graph of such correlation is straight line.

If amount of change in one variable does not bear a constant ratio to the amount of change in other variable, its nonlinear correlation.



positive Linear Regression.

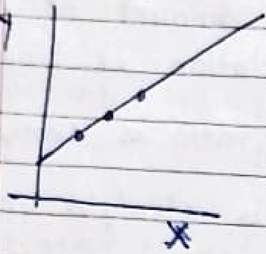


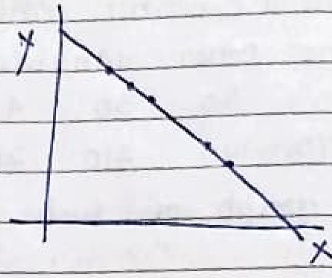
curvi-linear regression.

Method of Studying Correlation:-

1. Scatter diagram method
2. Graphic Method.
3. Karl- Pearson's Coefficient of Correlation
4. Spearman's Rank correlation

① Scatter diagram method :- The greater the scatter of the plotted points on the chart, the lesser is the relationship between two variables.
The more closely points come on a straight line, the higher degree of relationship.

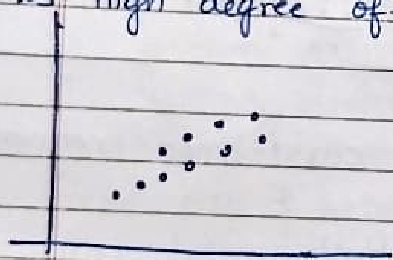


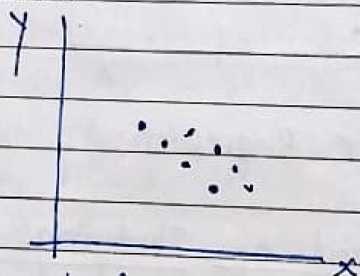


If all point lie on a st. line from lower left hand to upper right corner → perfect +ive correlation ie ($r = +1$)

If all points lie on a st. line from upper left corner to lower right corner → Perfect –ive correlation ($r = -1$).

If the plotted points fall in narrow band, there is high degree of correlation

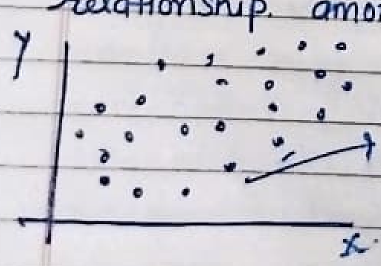


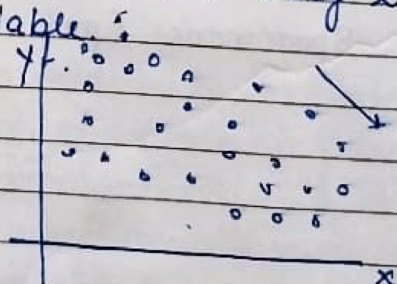


high degree of +ive Correlation

high degree of –ive correlation

If point are widely scattered ie it indicates ~~every~~ little relationship among variable.

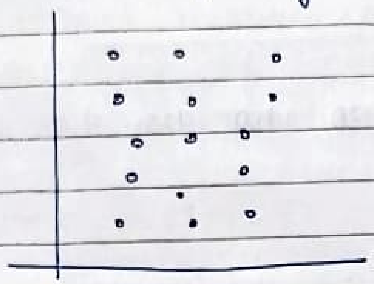




Low degree of positive correlation

Low degree of Negative correlation.

If plotted points lie on st. line parallel to x axis it shows absence of any relationship ie ($r=0$)

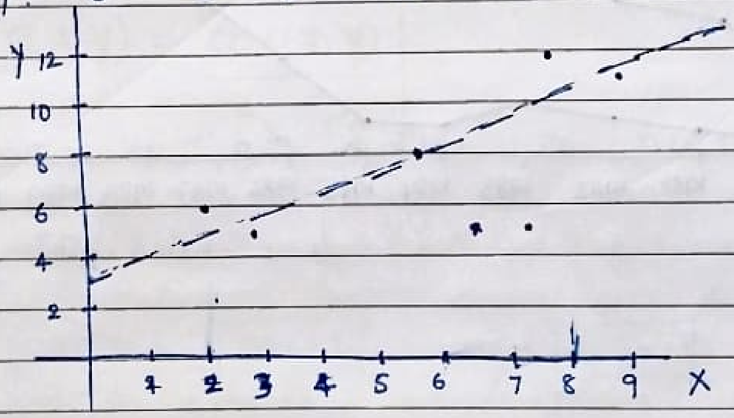


No correlation

~~The~~ Scatter diagram method, gives rough idea whether or not variables are related but cannot establish exact ~~degree~~ degree of correlation.

Ex:-

X :	2	3	7	6	8	9
Y :	6	5	5	8	12	11



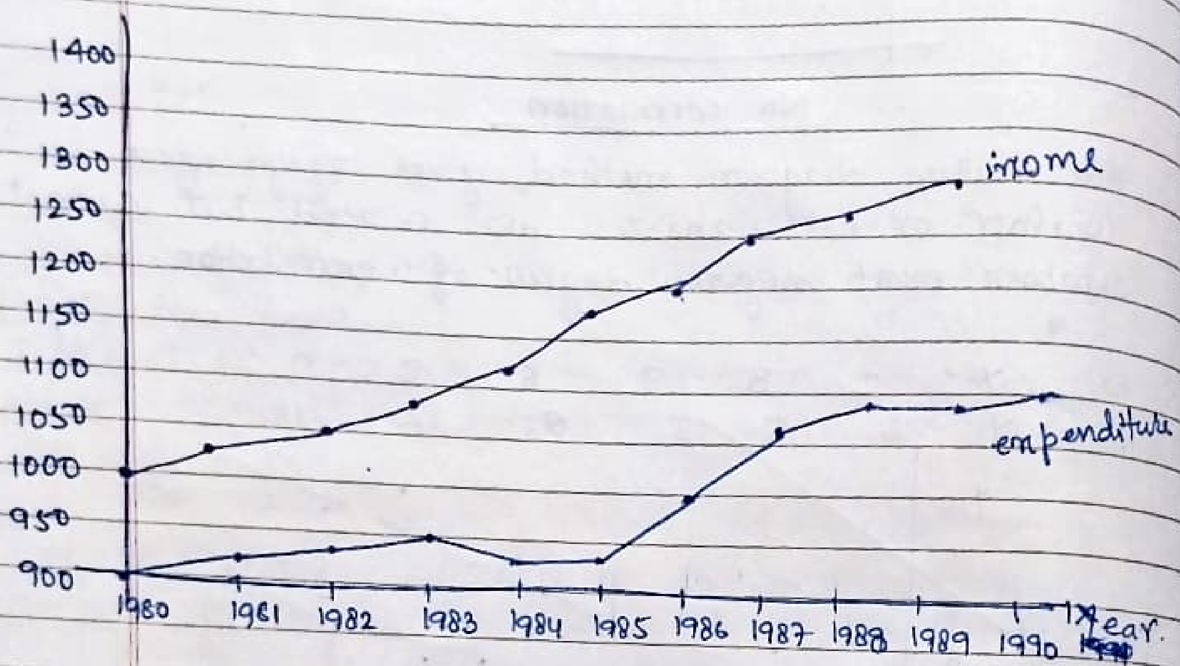
Graphic Method:- Individual values of two variables are plotted on the graph. Thus we obtain two curve, one for x variable and another for Y variable. By examining the direction and closeness of two variables, we can infer whether they are related or not.

If both curve are moving in same direction, correlation is said to be +ive.

If curves are moving in oppositie direction, correlation is negative.

eg:-

Year	1981	1982	1983	1984	1985	1986	1987	1988	1989
(average income)	1000	1020	1050	1080	1120	1180	1200	1250	1260
(averge Expenditure)	900	910	930	950	920	940	1000	1060	1080



(3) Karl Pearson's Coefficient of Correlation :-

Karl Pearson's method popularly known as Pearsonian coeff. of correlation. It is most widely used in pratice, denoted by r.

$$\left[r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} \right]$$

This method is to be applied only where the deviation of items are taken from actual mean

$$\left[r(x,y) = \frac{\operatorname{cov}(x,y)}{\sigma x \ \sigma y} \right]$$

$$\operatorname{cov}(x,y) = \frac{1}{N}\Sigma xy - \bar{x}\bar{y} = \frac{1}{N}\Sigma xy - \frac{\Sigma x}{N}\frac{\Sigma y}{N}$$

$$\sigma x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}} \qquad \sigma y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{N}}$$

$$r(x,y) = \frac{\frac{1}{N}\Sigma xy - \bar{x}\bar{y}}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}\sqrt{\frac{\Sigma(y-\bar{y})^2}{N}}}$$

$= \Sigma xy -$ [illegible]

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

this method is to be applied when the deviation are taken from actual mean and not from assumed mean.

The value of correlation coeff. obtained by above formula shall always lies b/w ±1.

when $r = +1$, it means there is Perfect positive correlation between the variables

when $r = -1$, it means there is perfect negative correlation between the variables.

when $r = 0$, it mean there is no relationship b/w two variables.

We normally get values, which lie between +1 and -1 such as 0.8 and -0.4.

The coeff of correlation describes not only magnitude of correlation but its direction Thus 0.8 means that correlation is positive as sign is +ive.

S'lly -0.4 means correlation is negative.

the formula is

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

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

Correlation coeff. when deviation are taken from assumed mean:-

$$r = \frac{N \Sigma dx \Sigma dy - \Sigma dx \Sigma dy}{\sqrt{N \Sigma dx^2 - (\Sigma dx)^2} \sqrt{N \Sigma dy^2 - (\Sigma dy)^2}}$$

Where $d_x = X - A_x$, A_x be assumed mean for X series

$d_y = Y - A_y$, A_y be assumed mean for Y series.

Q1 Calculate the Karl pearson's coeff of correlation from the following data!

X:	6	8	12	15	18	20	24	28	31
Y:	10	12	15	15	18	25	22	26	28

Sol^n $\bar{x} = \frac{\Sigma x}{N} = \frac{162}{9} = 18$

$\bar{y} = \frac{\Sigma y}{N} = \frac{171}{9} = 19$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
x	y	x−18	y−19	$(x-18)^2$	$(y-19)^2$	(x−18)(y−19)
6	10	−12	−9	144	81	108
8	12	−10	−7	100	49	70
12	15	−6	−4	36	16	24
15	15	−3	−4	9	16	12
18	18	0	−1	0	1	0
20	25	2	6	4	36	12
24	22	6	3	36	9	18
28	26	10	7	100	49	70
31	28	13	9	169	81	117
162	171	0	0	598	338	431

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

$$= \frac{431}{\sqrt{598}\sqrt{338}}$$

$$\left[r = +.959\right]$$

data

X : 78 89 96 69 59 79 68 61

Y : 125 137 156 112 107 136 123 108

Q: Find Karl pearson's coeff of Correlation b/w the values of X and Y. Assume 69 and 112 as the assume mean of X & Y are respectively.

solution:- $A_x = 69$ $A_y = 112$

$$d_x = X - A_x = X - 69$$

$$d_y = Y - A_y = Y - 112$$

$$* \quad r = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \Sigma d_y}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}} \sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

$$\text{or} \quad r = \frac{N \Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2} \sqrt{N \Sigma d_y^2 - (\Sigma d_y)^2}}$$

X	Y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
X	Y	X-69	Y-112	$(X-69)^2$	$(Y-112)^2$	$(X-69)(Y-112)$
8	125	9	13	81	169	117
9	137	20	25	400	625	500
6	156	+27	44	729	1936	1188
9	112	0	0	0	0	0
9	107	-10	-5	100	25	50
9	136	10	24	100	576	240
8	123	-1	11	1	121	-11
1	108	-8	-4	64	16	32
=8		47	108	1475	3468	2116

$$r = \frac{8 \times 2116 - 47 \times 108}{\sqrt{8 \times 1475 - (47)^2} \sqrt{8 \times 3468 - (108)^2}} = \frac{+11852}{\sqrt{9591}\sqrt{16080}}$$

$$= \frac{11852}{97.93 \times 126.80} = .9544$$

Q Compute Karl Pearson's coeff. of correlation b/w per capita national income and per capita consumer Expenditure from data given below:-

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Per capita National come	249	251	248	252	258	269	271	272	280	275
Per Capita consumer expenditure	237	238	236	240	245	255	254	252	258	251

Ans .967

Q: The following table gives the distributions of the total population and those who are wholly and partially blind among them. Find out if there any relation b/w age and blindness.

Age	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Person's (thousand)	100	60	40	36	24	11	6	3
Blind	55	40	40	40	36	22	18	15

Soln

For comparison, we must determine no. of blinds in terms of common denominator say 1 lakh. The first value will remain 55 as 55 persons are blind out of 100 thousand i.e. out of 1 lakh. The second and other value should also be obtained from 1 lakh.

out of 60,000, → 40 blind

out of 100000 → $\frac{40 \times 100000}{60000} = \frac{400}{6} = 66.6 \approx 67$

Illy other values should be obtained

Age	Mid point x	$X-A$ $X-35$	$d_x = \frac{X-A}{10}$	d_x^2
0-10	5	-30	-3	9
10-20	15	-20	-2	4
20-30	25	-10	-1	1
30-40	35	0	0	0
40-50	45	10	1	1
50-60	55	20	2	4
60-70	65	30	3	9
70-80	75	40	4	16

$\Sigma d_x = 4 \quad \Sigma d_x^2 = 44$

Y	$d_y = Y-185$	d_y^2	$d_x d_y$
55	-130	16900	+390
67	-118	13924	+236
100	-85	7225	85
111	-74	5476	0
150	-35	1225	-35
200	+15	225	30
300	115	13,225	345
500	315	99,225	1260

$\Sigma d_y = 3 \quad \Sigma d_y^2 = 157425 \quad \Sigma d_x d_y = 2,311$

$$r = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \Sigma d_y}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}}\sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

$$r = \frac{2113 - \frac{4 \times 3}{8}}{\sqrt{42}\sqrt{157423.88}} = .898$$

For Grouped data:-

$$r = \frac{\sum f\,dx\,dy - \dfrac{\sum f\,dx \sum f\,dy}{N}}{\sqrt{\sum f\,dx^2 - \dfrac{(\sum f\,dx)^2}{N}}\sqrt{\sum f\,dy^2 - \dfrac{(\sum f\,dy)^2}{N}}}$$

Rank correlation coefficient:-

Q: If x, y and z are uncorrelated random variables with zero means & S.D 5, 12, 9 resp. If $U = X + Y$, $V = Y + Z$, find correlat coeff b/w U and V.

Ans
$$r = \frac{E(UV) - E(U)\,E(V)}{\sigma_U\,\sigma_V}$$

$E(U) = E(X+Y) = E(X) + E(Y) = 0$
$E(V) = E(Y+Z) = 0$
$E(UV) = E((X+Y)(Y+Z))$

$= E(XY + Y^2 + XZ + YZ)$

$E(UV) = E(XY) + E(Y^2) + E(XZ) + E(YZ)$ —— ①

given $\sigma_x = 5$ $\sigma_y = 12$ $\sigma_z = 9$

since X and Y are uncorrelated
ie $E(XY) = E(X) \cdot E(Y) = 0$

Now $\sigma_x = 5$

$\sigma_x^2 = E(X^2) - (E(X))^2$, $E(X) = 0$

ie $E(X^2) = 25$

slly $\sigma_y^2 = 144 = E(Y^2)$, $\sigma_z^2 = 81 = E(Z^2)$

ie $E(UV) = 0 + 144 + 0 + 0$

$\boxed{E(UV) = 144}$

$E(U) = E(X+Y) = E(X) + E(Y) = 0$
Slly $E(V) = 0$.

ie $E(U)\cdot E(V) = 0$.

Now $\sigma_U = \sqrt{E(U^2) - [E(U)]^2}$

$$E(U^2) = E(X^2 + Y^2 + 2XY)$$
$$= E(X^2) + E(Y^2) + 2E(XY)$$

$E(U^2) = 25 + 144 = 169 \Rightarrow \sigma_U = 13$

Slly. $E(V^2) = 144 + 81 = 225 \Rightarrow \sigma_V = 25$

$$r = \frac{144}{13 \times 25} = \frac{48}{65}$$ Ans.

Q:- If joint pdf of (X, Y) is given by,

$f(x,y) = x + y$, $\cancel{0 \le x \le y \le 1}$ find

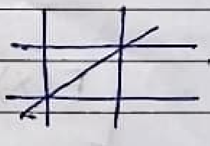
$0 \le x \le 1$

$0 \le y \le 1$

r_{xy} $(\rho(x,y))$

Solⁿ $\cancel{r_{xy}} = \rho(X,Y) = \dfrac{E(XY) - E(X)\cdot E(Y)}{\sigma_X \ \sigma_Y}$

$$E(XY) = \int_0^1 \int_0^1 xy \, f(x,y) \, dx \, dy$$



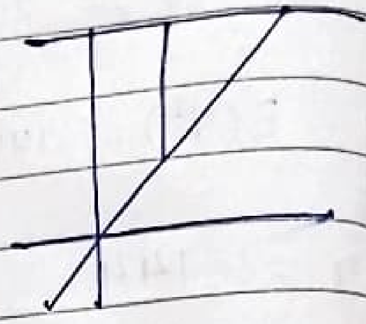
$$= \int_0^1 \int_0^1 xy\,(x+y)\, dx\, dy$$

$$\boxed{E(XY) = \frac{1}{3}}$$

$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$ $\quad E(Y) = \int_{-\infty}^{\infty} y f_Y(y)\, dy$.

Now

$$f_x(x) = \int_0^1 f(x,y)\,dy$$



$$f_x(x) = \int_0^1 (x+y)\,dy$$

$$= xy + y^2/2 \Big|_0^1 = x + 1/2$$

$$f_x(x) = x + 1/2, \quad 0 \le x \le 1$$

$$f_y(y) = \int_0^1 f(x,y)\,dx = y + \frac{1}{2}, \quad 0 \le y \le 1$$

$$\left[E(X) = \int_0^1 x\, f_x(x)\,dx = \frac{1}{3} + \frac{1}{4} = \frac{7}{12} \right]$$

$$\left[E(Y) = \int_0^1 y\, f_y(y)\,dy = 7/12 \right]$$

Now

$$E(X^2) = \int_0^1 x^2 f_x(x)\,dx = \int_0^1 x^2\left(x + \frac{1}{2}\right) dx$$

$$E(X^2) = \frac{5}{12}$$

Slly $E(Y^2) = 5/12$

$$\sigma x = \sqrt{E(x^2) - (E(x))^2} = \sqrt{\frac{5}{12} - \left(\frac{7}{12}\right)^2}$$

$$\sigma x = \sqrt{\frac{5}{12} - \frac{49}{144}} = \sqrt{\frac{60-49}{144}}$$

$$\sigma x = \frac{\sqrt{11}}{12}$$

Slly $\sigma y = \frac{\sqrt{11}}{12}$

so $\rho(x,y) = \frac{E(xy) - E(x).E(y)}{\sigma_x \sigma_y} = \frac{\frac{1}{3} - \frac{49}{144}}{\frac{\sqrt{11}}{12} \frac{\sqrt{11}}{12}}$

$$\boxed{\rho(x,y) = \frac{-1}{11}}$$

Q The independent r.v x and y have pdf given by

$$f_x(x) = \begin{cases} 4ax, & 0 \le x \le 1 \\ 0, & \text{o.w} \end{cases} \qquad f_y(y) = \begin{cases} 4by, & 0 \le y \le 1 \\ 0, & \text{o.w} \end{cases}$$

find the correlation coefficient.

Soln:- $E(x) = \int_{-\infty}^{\infty} x\, f(x)\, dx = \int_0^1 4ax^2 dx = 4a/3$

$E(Y) = \frac{4b}{3}$

Since x and Y are independent

$f(x,y) = f(x)\, f(y) = 16ab\,xy, \quad 0 \le x \le 1, \; 0 \le y \le 1$

$E(xY) = \int_0^1 \int_0^1 xy\, f(x,y)\, dx\, dy$

$E(xY) = \frac{16ab}{9}$

classmate
Date
Page

$$\operatorname{cov}(x, y) = E(xy) - E(x) \cdot E(y)$$

$$= \frac{16ab}{9} - \frac{4a}{3} \cdot \frac{4b}{3}$$

$$= 0$$

ie $\boxed{\rho(x, y) = 0}$

x and y are independent and no. correlation them.