Correlation

In bivariate dist, we may be interested to find out if there is any correlation or covariation b/w the two variables under study. If the change in one variable affects a change in the other variable, the variables are said to be correlated. If the two variables deviate in the same direction, ie, if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be direct or positive. But if inc (dec) in one results in dec (inc) in the other, correlation is said to be diverse/negative. For eg: the correlation b/w (i) heights & weights of a group of persons (ii) the income and expenditure, is positive and the correlation b/w (i) price & demand of the commodity and (ii) the volume & pressure of a perfect gas., is negative.

Correlation is said to be perfect if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

Scatter diagram

It is diagrammatic representation of bivariate data. For a bivariate dist (x_i, y_i) $i = 1, 2, \ldots, n$ if the values of variables X & Y are plotted along x-axis & y-axis resp., in the xy plane, the diagram of dots so obtained is known as scatter diagram.

If the pts are very close to each other, we expect a fairly good amount of correlation b/w the variables & if the pts are scattered widely, a poor correlation is expected. This method is however not suitable if no. of observations is fairly large.

Karl Pearson's coefficient of Correlation

Karl Pearson developed a formula called Correlation Coefficient to measure the intensity or degree of linear relationship b/w two variables.

Correlation coeff b/w two r.v X & Y, denoted by $r(x,y)$ or r_{XY} is a numerical measure of linear relationship b/w them & is defined as

* for non-linear relationship it is not suitable.

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

$$\therefore r_{XY} = \frac{E[(X-\bar{x})(Y-\bar{y})]}{\sqrt{E(X-\bar{x})^2 \, E(Y-\bar{y})^2}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{[E(X^2)-E(X)^2][E(Y^2)-E(Y)^2]}}$$

$$= \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i-\bar{x})^2 \, \frac{1}{n}\Sigma(y_i-\bar{y})^2}} = \frac{\frac{1}{n}\Sigma x_i y_i + \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\Sigma x_i^2 - \bar{x}^2\right)\left(\frac{1}{n}\Sigma y_i^2 - \bar{y}^2\right)}}$$

$x_i - \bar{x} = a_i$

$y_i - \bar{y} = b_i$

$$r_{XY}^2 = \frac{(\Sigma a_i b_i)^2}{\Sigma a_i^2 \, \Sigma b_i^2}$$

As per Schwartz inequality - if a_i, b_i $i = 1, 2, \ldots, n$ are real then

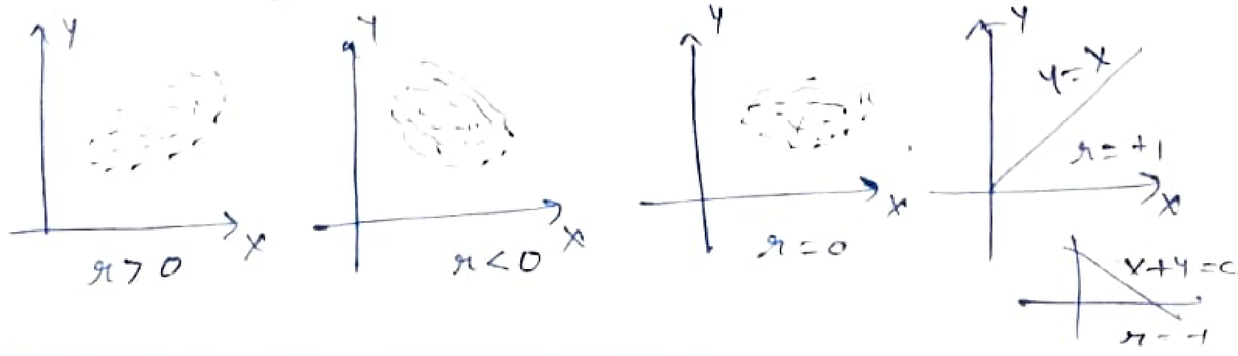
$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \le \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

$$\therefore r^2_{XY} \le 1 \quad \text{or} \quad |r_{XY}| \le 1$$

i.e. $\boxed{-1 \le r_{XY} \le 1}$

If $r = +1$, the correlation is positive & perfect

" $r = -1$, " " negative & perfect.



tion
near

Correlation coeff. is independent of change of Origin and scale.

Let $U = \dfrac{X-a}{h}$, $V = \dfrac{Y-b}{k}$

So $X = a + hU$, $Y = b + kV$. a, b, h, k are constants, $h > 0$, $k > 0$.

To prove $r(X,Y) = r(U,V)$

$E(X) = a + hE(U)$, $E(Y) = b + kE(V)$

$X - E(X) = h(U - E(U))$, $Y - E(Y) = k[V - E(V)]$

$$\begin{aligned} Cov(X;Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E[h(U - E(U))\, k(V - E(V))] \\ &= hk\, E[(U - E(U))(V - E(V))] \\ &= hk\, Cov(U,V) \end{aligned}$$

$\sigma_X^2 = E[X - E(X)]^2 = E(h^2(U - E(U))^2) = h^2\sigma_U^2$

$\Rightarrow \sigma_X = h\sigma_U$ $(h > 0)$

$\sigma_Y^2 = E[Y - E(Y)]^2 = E(k^2(V - E(V))^2) = k^2\sigma_V^2$

$\Rightarrow \sigma_Y = k\sigma_V$, $k > 0$

$$\therefore r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{hk\, Cov(U,V)}{h\sigma_U\, k\sigma_V} = \frac{Cov(U,V)}{\sigma_U \sigma_V}$$

$= r(U,V)$

Cor. If X & Y are r.v's & a, b, c, d are any nos. then $r(aX+b, cY+d) = \dfrac{ac}{|ac|} r(X,Y)$, $a \neq 0$, $c \neq 0$

$Var(aX+b) = a^2 var(X) = a^2\sigma_X^2$, $var(cY+d) = c^2\sigma_Y^2$

$Cov(aX+b, cY+d) = ac\,\sigma_{XY}$

$$\therefore r(aX+b, cY+d) = \frac{Cov(aX+b; cY+d)}{\sqrt{Var(aX+b) \cdot Var(cY+d)}} = \frac{ac\,\sigma_{XY}}{|a|\sigma_X |c|\sigma_Y} = \frac{ac}{|ac|}\frac{\sigma_{XY}}{\sigma_X\sigma_Y} = \frac{ac}{|ac|} r_{XY}.$$

Thm 2 Two independent variables are uncorrelated.

If X & Y are independent variables then

$$Cov(X,Y)=0 \Rightarrow r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = 0$$

Hence two independent variables are uncorrelated.

But the converse of the theorem is not true. ie., two uncorrelated variables may not be independent.

							Total
X	-3	-2	-1	1	2	3	$\Sigma X = 0$
Y	9	4	1	1	4	9	$\Sigma Y = 28$
XY	-27	-8	-1	1	8	27	$\Sigma XY = 0$

$\bar{X} = 0$, $\bar{Y} = \frac{28}{6}$ $\qquad Cov(X,Y) = \frac{1}{n}\Sigma XY - \bar{X}\bar{Y} = 0$

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = 0$$

$\Rightarrow$ X & Y are uncorrelated but $Y = X^2$.

$r_{XY} = 0 \Rightarrow$ absence of any linear relationship b/w X & Y

However there may exist some other form of " " ie. quadratic, cubic or trigonometric.

Ex Calculate the correlation coeff for the following heights (in inches) of fathers (X) & their sons (Y).

X	Y	$U = X-68$	$V = Y-69$	U^2	V^2	UV
65	67	-3	-2	9	4	6
66	68	-2	-1	4	1	2
67	65	-1	-4	1	16	4
67	68	-1	-1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
Total 544	552	0	0	36	44	24

$\bar{U} = \frac{\Sigma U}{n} = 0, \quad \bar{V} = \frac{\Sigma V}{n} = 0$

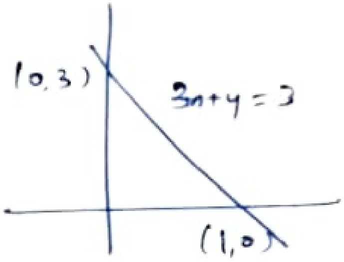
$\text{cov}(U, V) = \frac{1}{n} \Sigma UV - \bar{U}\bar{V} = \frac{1}{8} \times 24 = 3$

$\sigma_U^2 = \frac{1}{n} \Sigma U^2 - \bar{U}^2 = \frac{36}{8} = 4.5$

$\sigma_V^2 = \frac{1}{n} \Sigma V^2 - \bar{V}^2 = \frac{44}{8} = 5.5$

$$r(U,V) = \frac{\text{cov}(U,V)}{\sigma_U \sigma_V} = \frac{3}{\sqrt{4.5 \times 5.5}} = 0.603 = r(x,y)$$

Ex pdf $f(x,y) = \begin{cases} \frac{1}{2}(x+y) & x > 0,\ y > 0,\ 3x+y < 3 \\ 0 & \text{ew} \end{cases}$



$$g(x) = \frac{1}{2} \int_0^{3-3x} (x+y)\, dy = \frac{1}{2} \left(xy + \frac{y^2}{2} \right)_0^{3(1-x)}$$

$$= \frac{1}{2} \left(x(3-3x) + \frac{1}{2} \times 9 (1 + x^2 - 2x) \right)$$

$$= \frac{1}{4} (3x^2 - 12x + 9), \quad 0 < x < 1$$

$$h(y) = \frac{1}{2} \int_0^{1-y/3} (x+y)\, dx = \frac{1}{4} + \frac{y}{3} - \frac{5y^2}{36}, \quad 0 < y < 3$$

$$E(x) = \int_0^1 x\, g(x)\, dx = \frac{1}{4} \int_0^1 x (3x^2 - 12x + 9)\, dx$$

$$= \frac{1}{4} \left(\frac{9x^2}{2} - \frac{12x^3}{3} + \frac{3x^4}{4} \right)_0^1 = \frac{5}{16}$$

$$E(y) = \int_{y=0}^3 y\, h(y)\, dy = \int_{y=0}^3 y \left(\frac{1}{4} + \frac{y}{3} - \frac{5y^2}{36} \right) dy = \frac{21}{16}$$

$$E(xy) = \frac{1}{2} \int_0^1 \int_0^{3-3x} xy (x+y)\, dy\, dx$$

$$= \frac{3}{10}$$

(i) $Cov(x,y) = E(xy) - E(x)E(y) =$

(ii) $Cov(x+2,\ y-3) = E[(x+2)(y-3)] - E(x+2)E(y$

$= E[xy - 3x + 2y - 6] - \cancel{E[x(y) - 3x + 2y - 6]}$

$(E(x)E(y) - 3E(x) + 2E(y)$

$= E$

-6

(iii) $Corr(-2X+3,\ 2y+7) = \dfrac{Cov(-2x+3,\ 2y+7)}{\sqrt{Var(-2x+3)}\ \sqrt{Var(2y+7)}}$

$Var(-2x+3) = Cov(3-2x,\ 3-2x)$

$= (-2)(-2)\ Cov(x,x) = 4\ Cov(x,x) = 4\ Var\ x$

* $Cov(x_1 + x_2, y) = Cov(x_1, y) + Cov(x_2, y)$

LHS $E[(x_1 + x_2 - E(x_1+x_2))(y - E(y))]$

$= E[(x_1 + x_2 - E(x_1) - E(x_2))(y - E(y))]$

$= E[x_1 y - x_1 E(y) + x_2 y - x_2 E(y)^2 - y E(x_1) + E(y)E(x)$

$- y E(x_2) + E(x_2) E(y)]$

$= E(x_1 y) - E(x_1 \mu_y) + E(x_2 y) - E(x_2 \mu_y) - E(y \mu_{x_1}) +$

$\mu_x \mu_y - E(y) \mu_{x_2} + \mu_{x_2} \mu_y$

$= E(x_1 y) - \mu_y E(x_1) + E(x_2 y) - \mu_y E(x_2) - \cancel{\mu_{x_1} E(y)}$

$+ \cancel{\mu_{x_1} \mu_y} - \cancel{E(y) \mu_{x_2}} + \cancel{\mu_{x_2} \mu_y}$

$= E(x_1 y) - E(x_1) E(y) + E(x_2 y) - E(y) E(x_2)$

$= Cov(x_1, y) + Cov(x_2, y)$

* $Cov(x, x) = Var\ x$

$= E[(x - \mu_x)(x - \mu_x)] = E((x - \mu_x)^2) = E(x^2 + \mu_x^2 - 2x\mu_x)$

$= E(x^2) + \mu_x^2 - 2\mu_x^2 = E(x^2) - E(x)^2 = Var\ x$

π $\text{var}(X+Y) = \text{Var } X + \text{var } Y + 2\,\text{Cov}(X,Y)$

?s

$$= \text{Cov}(X+Y,\ X+Y)$$
$$= \text{Cov}(X,X) + \text{Cov}(X,Y) + \text{Cov}(Y,X) + \text{Cov}(Y,Y)$$
$$= \text{Var } X + 2\,\text{Cov}(X,Y) + \text{Var } Y.$$

Rank Correlation

Sometimes the actual numerical value of X & Y may not be available but the positions of the actual values arranged in order of merit (ranks) may be available.

Spearman's rank correlation coefficient -

Let $d_i = U_i - V_i$ where U_i represents rank of n values of X

V_i " " " of Y

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

* $\sum d_i = \sum (u_i - v_i) = \sum u_i - \sum v_i = n(\bar{u} - \bar{v}) = 0$

$(\because \bar{u} = \bar{v})$

as $\bar{u} = \dfrac{1+2+3+\dots+n}{n} = \bar{v}$

Ex The ranks of 16 students in Maths & Science are as follows. Calculate the rank correlation coeff. for proficiencies of in Maths & Science

Ranks in Maths (U)	1	2	3	4	5	6	7	8	9	10	11
Rank in Science (V)											

S.No	M	S	U	V	d=U-V	d_i^2
1	78	84	4	3	1	1
2	36	51	9	9	0	0
3	98	91	1	1	0	0
4	25	60	10	7	3	9
5	75	68	5	4	1	1
6	82	62	3	6	-3	9
7	90	86	2	2	0	0
8	62	58	7	8	-1	1
9	65	63	6	5	1	1
10	39	77	8	10	-2	4
						26

$$r_{uv} = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 26}{10(99)} = 0.8424$$

Tied Ranks

If some individuals receive the same rank, then each of these individuals is assigned a common rank, which is the arithmetic means of ranks.

Ex Calculate rank correlation coeff. -

Expenditure X	Profit	R_x	R_y	$d_i =$	d_i^2
10	6	8	8	0	0
15	25	4	2.5	1.5	2.25
14	12	6	5	1	1
25	18	1	4	-3	9
14	25	6	2.5	3.5	12.25
14	40	6	1	5	25
20	10	3	6	-3	9
20	7	2	7	-5	25
					83.50

Am of 2 & 3

$$r = 1 - \frac{6\left(\sum d_i^2 + \underbrace{\frac{m_i(m^2-1)}{12} + \cdots}_{\text{Correction factor}}\right)}{n(n^2-1)}$$

m is the frequency of rank

6 is repeated 3 times

2.5 " " 2 times

$$r = 1 - \frac{6\left(83.50 + \frac{3(9-1)}{12} + \frac{2(4-1)}{12}\right)}{8(64-1)}$$

$$= 1 - \frac{6\left\{83.50 + \frac{3\times 8}{12} + \frac{2\times 3}{12}\right\}}{8\times 63} = -0.024$$

(-ve association)

Ex

X	Y	R_x	R_y	$d_i^2 = (R_x - R_y)^2$
68	62	4	5	1
64	58	6	7	1
75	68	2.5	3.5	1
50	45	9	10	1
64	81	6	1	25
80	60	1	6	25
75	68	2.5	3.5	1
40	48	10	9	1
55	50	8	8	0
64	74	6	2	16
				$\Sigma d_i^2 = 72$

$$c = \frac{2(4-1)}{12} + \frac{3(9-1)}{12} + \frac{2(4-1)}{12} = 3$$

$$r = 1 - \frac{6(72+3)}{10\times 99} = \frac{1-6\times 75}{990} = 0.545$$

Regression (stepping back towards the average).

Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.
In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called dependent variable & the variable which influences the values is called or is used for prediction is called independent variable (or regressor)

Linear regression

If the variables in a bivariate distribution are related, we will find the points in the scatter diagram will cluster around some curve called the 'curve of regression'.
If the curve is a straight line, it is called the line of regression & then there is a linear regression b/w the variables, o/w regression is curvilinear.
When two variables are linearly correlated -

(i) If X is treated as independent variable, then the regression line is called regression line of Y on X.

(ii) If Y - - - - . . X on Y.

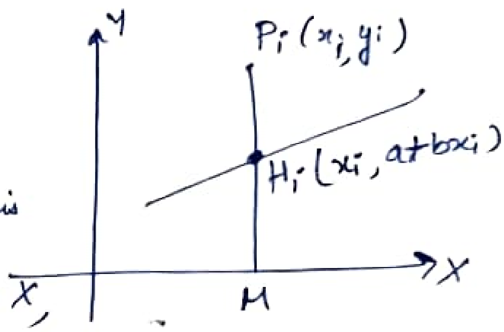
(iii) When there is either perfect +ve or -ve correlation ($r = \pm 1$) the regression lines will coincide, i.e. only one line will be there.

(iv) The farther the two regression lines from each other, the lesser is the degree of correlation and the nearer the lines, the higher is the degree of correlation.

) If variables are independent i.e. $r = 0$, then the gression lines are at right angles.

Regression lines cut each other at the pt of average of X and Y.

The lines of regression is the line which gives the best estimate of the value of one variable for any specific value of the other variable. Thus the line of regression is the line of 'best fit' and is obtained by the principle of least squares.



Let $P_i(x_i, y_i)$ be any general point in the scatter plot. Draw $P_iM \perp$ x-axis meeting, the line of regression of Y on X,

$$Y = a + bX, \quad (1)$$

→ represents the family of st. lines for different values of a & b.

To determine 'a' & 'b' so that (1) is the line of 'best fit'.

in $H_i(x_i, a + bx_i)$

$$P_iH_i = P_iM - H_iM$$
$$= y_i - (a + bx_i)$$

is called the error of estimate or the residual for y_i.

According to the principle of least square, to determine 'a' & 'b' so that

the sum of the squares of the deviation of actual Y values from computed Y is least.

$$E = \sum_{i=1}^{n} P_iH_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \text{ is minimum} \quad \text{or } \sum(\ )^2 \text{ is min.}$$

From the principle of max/min,

$$\frac{\partial E}{\partial a} = 0 = -2\sum_{i=1}^{n}(y_i - a - bx_i) \Rightarrow \sum y_i = na + b\sum x_i \quad (2)$$

$$\frac{\partial E}{\partial b} = 0 = -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) \Rightarrow \sum x_i y_i = a\sum x_i + b\sum x_i^2 \quad (3)$$

② & ③ are the normal equations for estimating 'a' & 'b' simultaneously.

Once 'a' & 'b' are obtained, eq ① is the line of best fitting to the given set of pts (x_i, y_i), $i = 1, \cdots, n$,

② on dividing by n give $\bar{y} = a + b\bar{x}$, — ④

∴ the line of regression of Y on X passes through $(\bar{x}, \bar{y})$.

Now

$$\mu_{11} = \text{Cov}(X, Y) = \frac{1}{n} \sum x_i y_i - \bar{x}\bar{y}$$

$$\text{or } \frac{1}{n} \sum x_i y_i = \mu_{11} + \bar{x}\bar{y} \quad — ⑤$$

Also $$\sigma_X^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

$$\text{or } \frac{1}{n} \sum x_i^2 = \sigma_X^2 + \bar{x}^2 \quad — ⑥$$

Dividing ③ by n & using ⑤ & ⑥,

$$\mu_{11} + \bar{x}\bar{y} = a\bar{x} + b(\sigma_X^2 + \bar{x}^2) \quad — ⑦$$

Multiplying ④ by $\bar{x}$ & subtracting from ⑦,

$$\mu_{11} = b\sigma_X^2 \quad \Rightarrow \quad b = \frac{\mu_{11}}{\sigma_X^2}$$

Since 'b' is the slope of the line of reg. of Y on X.

& " the line of reg passes through $(\bar{x}, \bar{y})$,

its equation is

$$Y - \bar{y} = b(X - \bar{x}) = \frac{\mu_{11}}{\sigma_X^2}(X - \bar{x})$$

$$\Rightarrow Y - \bar{y} = \frac{r \sigma_X}{\sigma_Y}(X - \bar{x})$$

Starting with the equation $X = A + BY$ and proceeding similarly,

$$X - \bar{x} = \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{y})$$

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

- Both the lines of reg, Y on X & X on Y, passes through $(\bar{x}, \bar{y})$.
  i.e. the mean value $(\bar{x}, \bar{y})$ can be obtained by as the point of intersection of the two regression lines.

- line of reg. of Y on X is used to predict or estimate the value of Y for any given value of X; i.e., The estimate so obtained will be best in the sense that it will have the minimum possible error as defined by the principle of least square.
  It has been obtained by minimizing the sum of the squares of the errors parallel to the Y-axis while reg. eqn of X on Y is obtained by minimizing the sum of squares of error parallel to X axis.

<u>Reg. coefficients</u>

'b' – the slope of the line of reg of Y on X is also called the coeff of reg of Y on X. It represents the increment in the value of Y corresponding to unit change in the value of indep. variable X.

$$b_{YX} = \text{Reg. coeff of } Y \text{ on } X = \frac{\mu_{11}}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X}$$

Similarly, $$b_{XY} = \text{'' } X \text{ on } Y = \frac{\mu_{11}}{\sigma_Y^2} = r\frac{\sigma_X}{\sigma_Y}.$$

Properties of reg. coefficients

a) Corr. coeff is the geometric mean b/w reg. coefficients

$$b_{xy} \times b_{yx} = r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x} = r^2$$

$$\text{or } r^2 = \pm\sqrt{b_{xy} \times b_{yx}}$$

* If b is +ve, r is +ve

If b is -ve, r is -ve

Since $r = \frac{\mu_{11}}{\sigma_x \sigma_y}$, $b_{yx} = \frac{\mu_{11}}{\sigma_x^2}$ and $b_{xy} = \frac{\mu_{11}}{\sigma_y^2}$.

b) If one of reg. coeff is >1, the other must be <1.

Let $b_{yx} > 1$.

$$\Rightarrow \frac{1}{b_{yx}} < 1$$

$$r^2 \le 1 \Rightarrow b_{xy}\, b_{yx} \le 1$$

$$\therefore b_{xy} \le \frac{1}{b_{yx}} < 1$$

c) Reg. coeff are independent of change of origin but not of scale.

Let $U = \frac{X-a}{h}$, $V = \frac{Y-b}{k}$

$$\Rightarrow X = a + hU, \quad Y = b + Vk$$

a, b $h > 0$, $k > 0$

$$Cov(X,Y) = hk\, Cov(U,V)$$

$$\sigma_x^2 = h^2\sigma_U^2, \quad \sigma_y^2 = k^2\sigma_V^2$$

$$\therefore b_{yx} = \frac{Cov(X,Y)}{\sigma_x^2} = \frac{hk\, Cov(U,V)}{h^2\sigma_U^2} = \frac{k}{h}\frac{Cov(U,V)}{\sigma_U^2}$$

$$= \frac{k}{h} b_{VU}$$

Similarly $b_{xy} = \frac{h}{k} b_{UV}$

fficients

Angle b/w two lines of Regression

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

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

imply slopes as $r\frac{\sigma_y}{\sigma_x}$ & $\frac{\sigma_y}{r\sigma_y}$ resp.

If θ is the acute angle b/w the 2 lines of reg, then

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

$\because r^2 \le 1$

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

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

* If $r=0$, $\tan\theta=\infty \Rightarrow \theta=\pi/2$
i.e. if the two variables are uncorrelated, the lines of reg. become $\perp$ to each other.

* If $r=\pm 1$, $\tan\theta = 0 \Rightarrow \theta = 0$ or π,
i.e. the two lines of reg. either coincide or they are parallel to each other. But since both lines pass through $(\bar{x},\bar{y})$, they can't be $\parallel$. Hence in this case, the two lines coincide.

Ex find the regression equation

X	Y	$x = X - \bar{x}$	$y = Y - \bar{y}$	x^2	y^2	xy
6	9	0	1	0	1	0
2	11	-4	3	16	9	-12
10	5	4	-3	16	9	-12
4	8	-2	0	4	0	0
8	7	2	-1	4	1	-2
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 40$	$\Sigma y^2 = 20$	$\Sigma xy = -26$

$\Sigma X = 30$

$\bar{x} = E(X) = 6, \quad \bar{y} = E(Y) = 8$

$\sigma_x^2 = \frac{1}{n}\Sigma x^2 - \left(\frac{1}{n}\Sigma x\right)^2 = \frac{40}{5} - \left(\frac{0}{7}\right) = 8$

$\sigma_y^2 = \frac{1}{n}\Sigma y^2 - \left(\frac{1}{n}\Sigma y\right)^2 = \frac{20}{5} - \left(\frac{0}{5}\right) = 4$

$Cov(X,Y) = \frac{1}{n}\Sigma xy - \frac{\Sigma x}{n} \cdot \frac{\Sigma y}{n} = \frac{-26}{5} - \frac{0}{5} \times \frac{0}{5} = \frac{-26}{5}$

$= -5.2$

Reg lines of Y on X is

$y - \bar{y} = \frac{Cov(X,Y)}{\sigma_x^2}(x - \bar{x})$

$y - 8 = \frac{-5.2}{8}(x - 6)$

$y = 11.9 - 0.65x$

Similarly $x - \bar{x} = \frac{Cov(X,Y)}{\sigma_y^2}(y - \bar{y})$

$x - 6 = \frac{-5.2}{4}(y - 8)$

$x = 16.4 - 1.3y$

If we use $\Sigma y_i = a \Sigma x_i + nb$

$$\Sigma x_i y_i = a \Sigma x_i^2 + b \Sigma x_i$$

X	Y	XY	X^2	Y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	32	49
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma XY = 214$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 340$

Reg. eq. of Y on X is

$$y = a + bx$$

$\Sigma y = na + b\Sigma x$ or $40 = 5a + 30b$

$\Sigma xy = a\Sigma x + b\Sigma x^2$ or $214 = 30a + 220b$

$$\Rightarrow y = 11.9 - 0.65x$$

Reg. eq of X on Y is

$$x = a + by$$

$\Sigma x = na + b\Sigma y$ or $30 = 5a + 40b$

$\Sigma xy = a\Sigma y + b\Sigma y^2$ or $214 = 40a + 340b$

$$\therefore x = 16.4 - 1.3y$$

Ex var $x = 9$

Reg. eq. $8x - 10y + 66 = 0$

& $40x - 18y = 214$

(i) Mean values of x & y

Reg. lines passes through $\bar{x}$ & $\bar{y}$

$$\left.\begin{array}{l} 8\bar{x} - 10\bar{y} = -66 \\ 40\bar{x} - 18\bar{y} = 214 \end{array}\right] \quad \bar{x} = 13 \ \& \ \bar{y} = 17$$

(ii) Corr. coeff b/w X & Y.

Reg. lines of Y on X $\quad 8x - 10y + 66 = 0$

$$\Rightarrow y = \frac{8x}{10} + \frac{66}{10}$$

Reg line X on Y $\quad 40x - 18y = 214$

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

Reg coeff of Y on X $= 8/10$

" X on Y $= 18/40$

$$r^2 = \frac{8}{10} \times \frac{18}{40} = \frac{9}{25}$$

$$r = \pm\frac{3}{5} = \pm 0.6$$

$\therefore r = 0.6$ ($\because b_{xy}$ & b_{yx} both +ve)

(iii) Standard deviation of Y.

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

$$\frac{8}{10} = \frac{3}{5} \times \frac{\sigma_y}{3} \Rightarrow \sigma_y = 4$$

Obt

x

Obtain the equations of lines of reg.

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

$$y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

$\sigma_{xy} \to E(XY) - E(X)E(Y)$

$\bar{y} \to E(Y)$

$\sigma_x^2 \to \operatorname{Var} X = E(X^2) - E(X)^2$

X	Y	U = X − 4	V = Y − 12	U^2	V^2	UV
1	9	−3	−3	9	9	9
2	8	−2	−4	4	16	8
3	10	−1	−2	1	4	2
4	12	0	0	0	0	0
5	11	1	−1	1	1	−1
6	13	2	1	4	1	2
7	14	3	2	9	4	6
		0	−7	28	35	26

$\bar{X} = E(X) = E(U) + 4 = 4$

$\bar{Y} = E(Y) = E(V) + 12 = \frac{-7}{7} + 12 = 11$

$\operatorname{Var} U = \operatorname{Var}(X - 4) = \operatorname{var} X$

$\operatorname{Var} U = \frac{\Sigma U^2}{n} - \left(\frac{\Sigma U}{n}\right)^2 = \frac{28}{7} - 0 = 4$

$\operatorname{Var} Y = \operatorname{Var} V = \frac{\Sigma V^2}{n} - \left(\frac{\Sigma V}{n}\right)^2 = \frac{35}{7} - \left(\frac{-7}{7}\right)^2 = 5 - 1 = 4$

$E(UV) = E[(X-4)(Y-12)] = E(XY - 12X - 4Y + 48)$

$= E(XY) - 12E(X) - 4\underbrace{E(Y)}_{11} + 48$ as $E(X) = 4$

$\frac{\Sigma UV}{n} = E(XY) - 44$ or $E(XY) = \frac{26}{7} + 44$

$= 334/7$

$$y - \cancel{y}_{11} = \frac{\frac{334}{7} - 4\times 11}{4} (x-4)$$

$$y - 11 = \frac{334 - 308}{28} (x-4)$$

$$y - 11 = \frac{13}{14} (x-4)$$

$$x - 4 = \frac{\cancel{26}}{44} \frac{13}{22} (y-11)$$

Ex Given that $x = 4y + 5$ ($\to x$ on y), $y = kx + 4$ (y on x) are the reg. lines. Show that $0 \le k \le 1/4$

If $k = \frac{1}{16}$, find the means of x & y & r_{xy}.

$$b_{yx} = k, \quad b_{xy} = 4$$

$$r^2 = b_{yx}\, b_{xy} = 4k$$

$$-1 \le r \le 1 \quad \Rightarrow \quad 0 \le r^2 \le 1$$

$$0 \le 4k \le 1$$

$$0 \le k \le 1/4$$

If $k = \frac{1}{16}$ $\quad r_{xy}^2 = 4 \times \frac{1}{16} = \frac{1}{4}$

$$r_{xy} = \pm 1/2$$

But both b_{xy} & b_{yx} are +ve

$$\therefore r_{xy} = 1/2$$

$$x = 4y + 5, \quad y = \frac{x}{16} + 4$$

$$16y = x + 64$$

$$\left.\begin{array}{l} x = 4y + 5 \\ x = 16y - 64 \end{array}\right] \Rightarrow \quad y = \frac{69}{12} = 5.75$$

$$x = 4 \times \frac{\overset{23}{69}}{12} + 5 = 28$$

$x = 28$, $y = 5.75$