Unit 2 Correlation and Linear regression

Part 1: CORRELATION

Correlation Analysis

Correlation analysis is a statistical tool used to measure the strength of the linear relationship between two variables and compute their association.

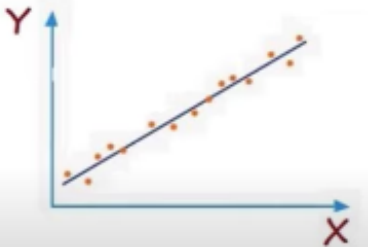
Kinds of Correlations:

can be either a positive, a negative, or no correlation.

Positive Correlation: If the values of the two variables deviate (moves) in the same direction, i.e., if the increase (decrease) in the values of one variable results, on an average, in a corresponding increase (decrease) in the values of the other variables

Examples:

1) Height and weights.
2) Family income and expenditure on luxury items.
3) Price and supply of a commodity



X	2	5	8	10
Y	18	25	34	51

X increases; Y also increases ------- Positive Correlation

Negative or Inverse Correlation: If the values of the two variables deviate (moves) in the opposite direction, i.e., if the increase (decrease) in the values of one variable results, on an average, in a corresponding decrease (increase) in the values of the other variables.

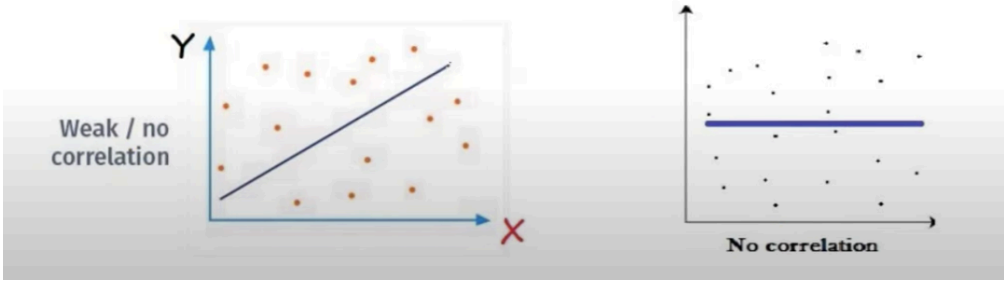
Example:

1) Price and demand of a commodity.
2) Volume and pressure of a perfect gas. ($PV = constant$)

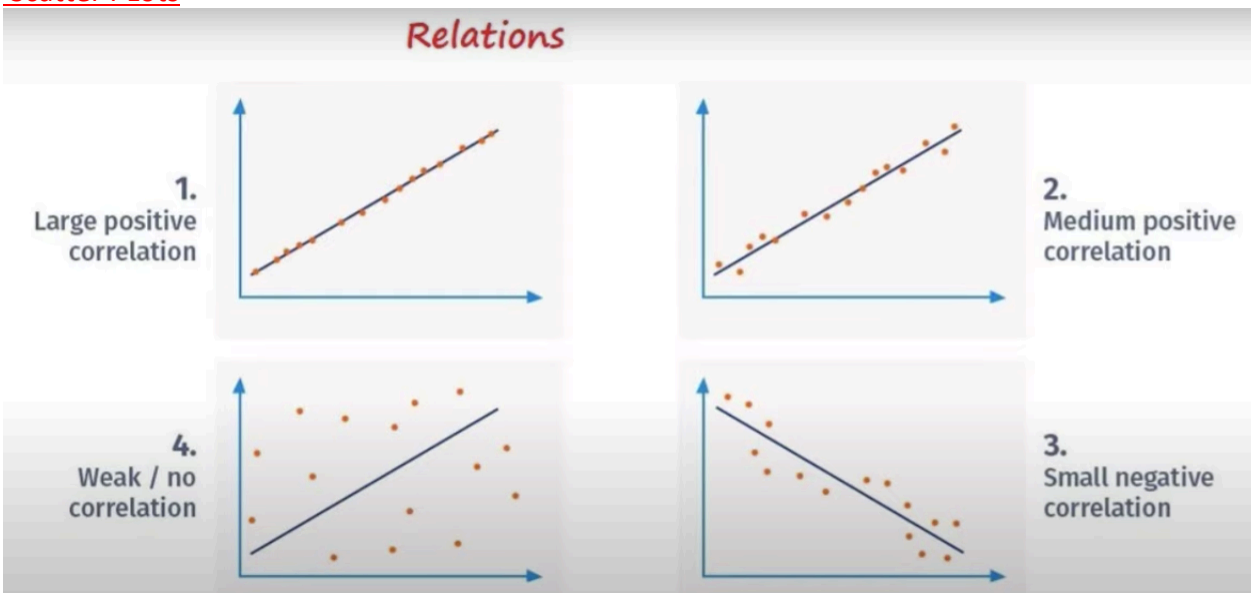
X	8	4	3	1
Y	8	10	15	25

X decreases; Y increases; ----- Negative Correlation

Weak/Zero correlation: It exists when one variable does not affect the other. For example, there is no correlation between the number of years of school a person has attended and the letters in his/her name.

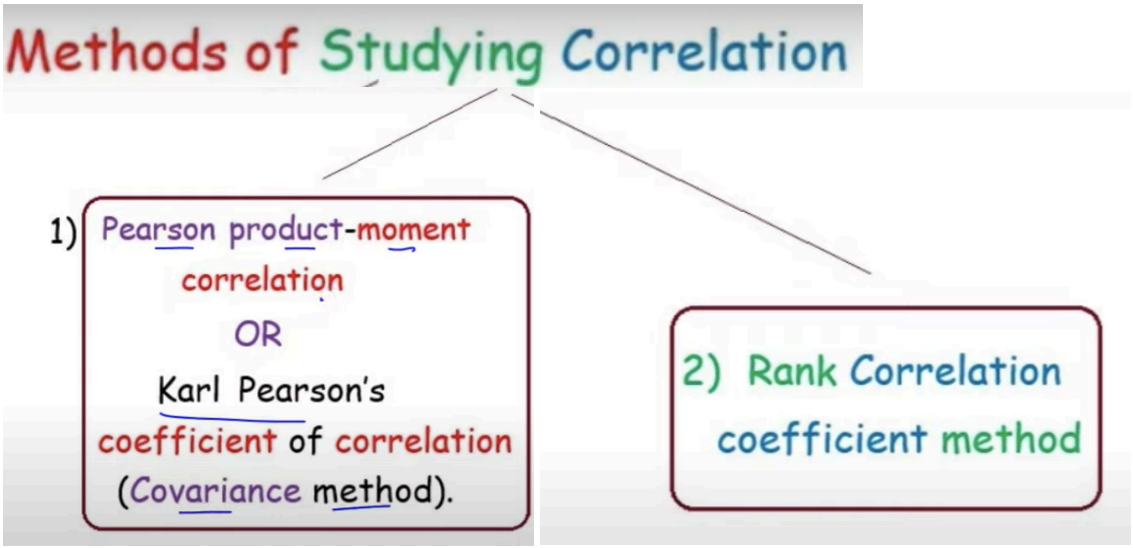


Scatter PLots



Remark:

- If the correlation coefficient of two variables is zero, there is no linear relationship between them.
- However, this is only for a linear relationship. It is possible that the variables have a strong curvilinear relationship (non-linear relationship).



Karl Pearson's coefficient of correlation

(Covariance method).

Karl Pearson's correlation coefficient between X and Y.

It is denoted by **r**(**X**, **Y**) or $\mathbf{r}_{XY}$ or simply **r**.

Let X and Y be random variables with covariance σ_{XY} and standard deviations σ_X and σ_Y, respectively. The correlation coefficient of X and Y is

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

or

Correlation coefficient is defined as

$$r = \frac{cov(X,Y)}{\sqrt{Var(X) \times Var(Y)}}$$

OR

$$r_{XY} = \frac{E\{[X - E(X)][Y - E(Y)]\}}{\sqrt{E\{X - E(X)\}^2 E\{Y - E(Y)\}^2}}$$

Rules for finding Covariance

1st method

$$cov(X,Y) = E(XY) - E(X)E(Y)$$

$$Var(X) = E(X^2) - \left(E(X)\right)^2$$

$$Var(Y) = E(Y^2) - \left(E(Y)\right)^2$$

2nd method

$$cov(X,Y) = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{n}$$

$$Var(X) = \frac{\sum(X - \bar{X})^2}{n}$$

$$Var(Y) = \frac{\sum(Y - \bar{Y})^2}{n}$$

OR

$$r_{XY} = \frac{\frac{1}{n}\Sigma x_i\, y_i - \frac{1}{n}\Sigma x_i \cdot \frac{1}{n}\Sigma y_i}{\sqrt{\left\{\frac{1}{n}\Sigma x_i^2 - \left(\frac{1}{n}\Sigma x_i\right)^2\right\}\left\{\frac{1}{n}\Sigma y_i^2 - \left(\frac{1}{n}\Sigma y_i\right)^2\right\}}}$$

$$r_{XY} = \frac{n\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{\left\{n\Sigma x^2 - (\Sigma x)^2\right\}\left\{n\Sigma y^2 - (\Sigma y)^2\right\}}}$$

Example: Calculate the coefficient of correlation between X and Y from the following data:

Summation of product deviation of X and Y from their respective mean is 122.

	Series	
	X	Y
No. of pairs	15	15
Mean	25	18
Sum of Squares of deviations from mean	$\sum(X-\bar{X})^2 = 136$	$\sum(Y-\bar{Y})^2 = 138$

Solution: Given $\sum(X-\bar{X})(Y-\bar{Y}) = 122$

$$Cov(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n} = \frac{122}{15}$$

$$Var(X) = \frac{\sum(X-\bar{X})^2}{n} = \frac{136}{15}$$

$$Var(Y) = \frac{\sum(Y-\bar{Y})^2}{n} = \frac{138}{15}$$

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X) \times Var(Y)}}$$

$$= \frac{122/15}{\sqrt{\frac{136}{15} \times \frac{138}{15}}}$$

$$= 0.8918$$

Example 1 *Compute the coefficient of correlation between X and Y, using the following data:*

X	:	1	3	5	7	8	10
Y	:	8	12	15	17	18	20

Solution

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	8	1	64	8
3	12	9	144	36
5	15	25	225	75
7	17	49	289	119
8	18	64	324	144
10	20	100	400	200
34	90	248	1446	582

Thus $n = 6$

$$\Sigma x_i = 34,\ \Sigma y_i = 90$$

$$\Sigma x_i^2 = 248,\ \Sigma y_i^2 = 1446$$

$$\Sigma x_i y_i = 582$$

$$r_{XY} = \frac{n\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{\left\{n\Sigma x^2 - (\Sigma x)^2\right\}\left\{n\Sigma y^2 - (\Sigma y)^2\right\}}}$$

$$= \frac{6 \times 582 - 34 \times 90}{\sqrt{\left\{6 \times 248 - (34)^2\right\}\left\{6 \times 1446 - (90)^2\right\}}}$$

$$= \frac{432}{\sqrt{332 \times 576}} = 0.9879$$

Properties of Correlation Coefficient

1. $-1 \le r_{XY} \le 1$

or $|\text{Cov}(X, Y)| \leq \sigma_X \cdot \sigma_Y$.

Note *When $0 < r_{XY} \leq 1$, the correlation between X and Y is said to be positive or direct.*

When $-1 \leq r_{XY} \leq 0$, the correlation is said to be negative or inverse.
When $-1 \leq r_{XY} \leq -0.5$ or $0.5 \leq r_{XY} \leq 1$, the correlation is assumed to be high, otherwise the correlation is assumed to be poor.

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

   i.e., If $U = \frac{X-a}{h}$ and $V = \frac{Y-b}{k}$, where $h, k > 0$, then $r_{XY} = r_{UV}$.

Note *If X and Y take considerably large values, computation of r_{XY} will become difficult. In such problems, we may introduce change of origin and scale and compute r using the above property.]*

3. Two independent RV's X and Y are uncorrelated, but two uncorrelated RV's need not be independent.

Two independent variables are uncorrelated.

If X and Y are independent variables, then

$$\text{Cov}(X, Y) = 0$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

NOTE: If $r = 0$, there is no linear relationship between them. However, this is only for a linear relationship. It is possible that the variables have a strong curvilinear relationship (non-linear relationship).

3 Example: Consider the series

X	-2	-1	0	1	2
Y	4	1	0	1	4
XY	-8	-1	0	1	8

Then E(X)=0 ; E(Y) = 2 ; E(XY) = 0

$$r = \frac{cov(X,Y)}{\sqrt{Var(X) \times Var(Y)}}$$

Cov(XY) = E(XY)-E(X)E(Y)

$r = 0$,

(no linear relation between X & Y)

We easily found a relation which is NON-LINEAR $Y = X^2$

Example 4 *Compute the coefficients of correlation between X and Y using the following data:*

X	:	65	67	66	71	67	70	68	69
Y	:	67	68	68	70	64	67	72	70

Solution We effect change of origin in respect of both X and Y. The new origins are chosen at or near the average of extreme values. Thus we take $\frac{65+71}{2} = 68$ as the new origin for X and $\frac{64+72}{2} = 68$ as the new origin for Y. viz., we put $u_i = (x_i - 68)$ and $v_i = y_i - 68$ and find r_{UV}.

$X = x_i$	$Y = y_i$	$u_i = x_i - 68$	$v_i = y_i - 68$	u_i^2	v_i^2	$u_i v_i$
65	67	−3	−1	9	1	3
67	68	−1	0	1	0	0
66	68	−2	0	4	0	0
71	70	3	2	9	4	6
67	64	−1	−4	1	16	4
70	67	2	−1	4	1	−2
68	72	0	4	0	16	0
69	70	1	2	1	1	2
	Total	−1	2	29	39	13

$$r_{XY} = r_{UV} = \frac{n\Sigma uv - \Sigma u \cdot \Sigma v}{\sqrt{\{n\Sigma u^2 - (\Sigma u)^2\}\{n\Sigma v^2 - (\Sigma v)^2\}}}$$

$$= \frac{8 \times 13 - (-1) \times 2}{\sqrt{(8 \times 29 - 1)(8 \times 39 - 4)}} = \frac{106}{\sqrt{231 \times 308}} = 0.3974$$

REMARK It is advised to calculate the correlation coefficient by \arbitrary origin method rather than by the direct method; since the latter leads to much simpler arithmetical calculations

EXAMPLE 5

A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results:

$n = 25, \Sigma X = 125, \Sigma X^2 = 650, \Sigma Y = 100, \Sigma Y^2 = 460, \Sigma XY = 508$

It was, however, later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values were

X	Y
8	12
6	8

Obtain the correct value of correlation coefficient.

Solution.

Corrected $\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$

Corrected $\Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$

Corrected $\Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Corrected $\Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

Corrected $\Sigma XY = 508 - 6\times14 - 8\times6 + 8\times12 + 6\times8 = 520$

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{1}{25}\times 125 = 5, \quad \bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{25}\times 100 = 4$$

$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\bar{Y} = \frac{1}{25}\times 520 - 5\times 4 = \frac{4}{5}$$

$$\sigma_X^2 = \frac{1}{n}\Sigma X^2 - \bar{X}^2 = \frac{1}{25}\times 650 - (5)^2 = 1$$

$$\sigma_Y^2 = \frac{1}{n}\Sigma Y^2 - \bar{Y}^2 = \frac{1}{25}\times 436 - 16 = \frac{36}{25}$$

$$\therefore \quad \text{Corrected } r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{4}{5}}{1 \times \frac{6}{5}} = \frac{2}{3} = 0.67$$

EXAMPLE 6

If the independent random variables X and Y have the variances 36 and 16 respectively, find the correlation coefficient between (X + Y) and (X – Y).

Solution Let $U = X + Y$ and $V = X - Y$

$$E(U) = E(X) + E(Y);\ E(V) = E(X) - E(Y)$$

$$E(UV) = E(X^2 - Y^2) = E(X^2) - E(Y^2)$$

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

$$E(V^2) = E(X^2) + E(Y^2) - 2E(XY)$$

$$C_{UV} = E(UV) - E(U) \cdot E(V)$$

$$= E(X^2) - E(Y^2) - \{E^2(X) - E^2(Y)\}$$

$$= [E(X^2) - E^2(X)] - [E(Y^2) - E^2(Y)]$$

$$= \sigma^2_X - \sigma^2_Y = 36 - 16 = 20$$

$$\sigma^2_U = E(U^2) - E^2(U)$$

$$= \{E(X^2) + E(Y^2) + 2E(XY)\} - \{E^2(X) + E^2(Y) + 2E(X) \cdot E(Y)\}$$

$$= [E(X^2) - E^2(X)] + [E(Y^2) - E^2(Y)] + 2[E(XY) - E(X) \cdot E(Y)]$$

$$= 36 + 16 + 2 \times 0$$

[$\because$ X and Y are independent and hence uncorrelated]

$$= 52$$

Similarly, $\sigma^2_V = 52$

Now $$r_{UV} = \frac{C_{UV}}{\sigma_U \cdot \sigma_V} = \frac{20}{52} = \frac{5}{13}$$

Some extra questions based on definition and properties of correlation coefficient

7

Example: The covariance of two perfectly correlated variable X and Y is 0.96. Find the S.D. of X and Y if it is known that variance of X and Y are in the ratio of 4:9.

Solution: Cov(X,Y) = 0.96 ; $\sigma_X = ?$; $\sigma_Y = ?$ $\frac{\sigma_X^2}{\sigma_Y^2} = \frac{4}{9}$

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

For perfect correlation; $r = 1 \text{ or } -1$

Since Cov(X,Y) is positive, thus $r = 1$

$$1 = \frac{0.96}{\sigma_X \sigma_Y}$$

$$\Rightarrow \sigma_X \sigma_Y = 0.96$$

$$\frac{\sigma_X^2}{\sigma_Y^2} = \frac{4}{9} \quad \Rightarrow \sigma_X = \frac{2}{3}\sigma_Y$$

$$\Rightarrow \frac{2}{3}\sigma_Y^2 = 0.96$$

$$\Rightarrow \sigma_Y = 1.2$$

Therefore, $\sigma_X = 0.8$

8

Example: Calculate the coefficient of correlation between X and Y from the following data:

Summation of product deviation of X and Y from their respective mean is 122.

	Series	
	X	Y
No. of pairs	15	15
Mean	25	18
Sum of Squares of deviations from mean	$\sum(X-\bar{X})^2 = 136$	$\sum(Y-\bar{Y})^2 = 138$

Solution: Given $\sum(X-\bar{X})(Y-\bar{Y}) = 122$

$$Cov(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n} = \frac{122}{15}$$

$$Var(X) = \frac{\sum(X-\bar{X})^2}{n} = \frac{136}{15}$$

$$Var(Y) = \frac{\sum(Y-\bar{Y})^2}{n} = \frac{138}{15}$$

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X) \times Var(Y)}}$$

$$= \frac{122/15}{\sqrt{\frac{136}{15} \times \frac{138}{15}}}$$

$$= 0.8918$$

9

Example: If X and Y are two random variables then show that the correlation coefficient between them is $\frac{\sigma_X^2+\sigma_Y^2-\sigma_{X-Y}^2}{2\sigma_X\sigma_Y}$

Solution:

$$\sigma_{X-Y}^2 = Var(X-Y)$$

$$= Var(X) + Var(Y) - 2Cov(X,Y)$$

$$= \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X\sigma_Y$$

Use $Var(aX+bY) = a^2V(X) + b^2V(Y) + 2abCov(X,Y)$

$$\Rightarrow \quad r = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2\sigma_X\sigma_Y}$$

10

Example: If X and Y are two correlated random variables with the same variance and if r is the correlation coefficient between X and Y, find the correlation coefficient between X and X+Y.

Target

$$r(U,V) = \frac{Cov(U,V)}{\sqrt{Var(U)Var(V)}}$$

Solution: Given that $Var(X) = Var(Y) = k$

and $r = \frac{Cov(X,Y)}{\sqrt{Var(X)\times Var(Y)}} = \frac{Cov(X,Y)}{k}$

Let $U = X; \quad V = X + Y$

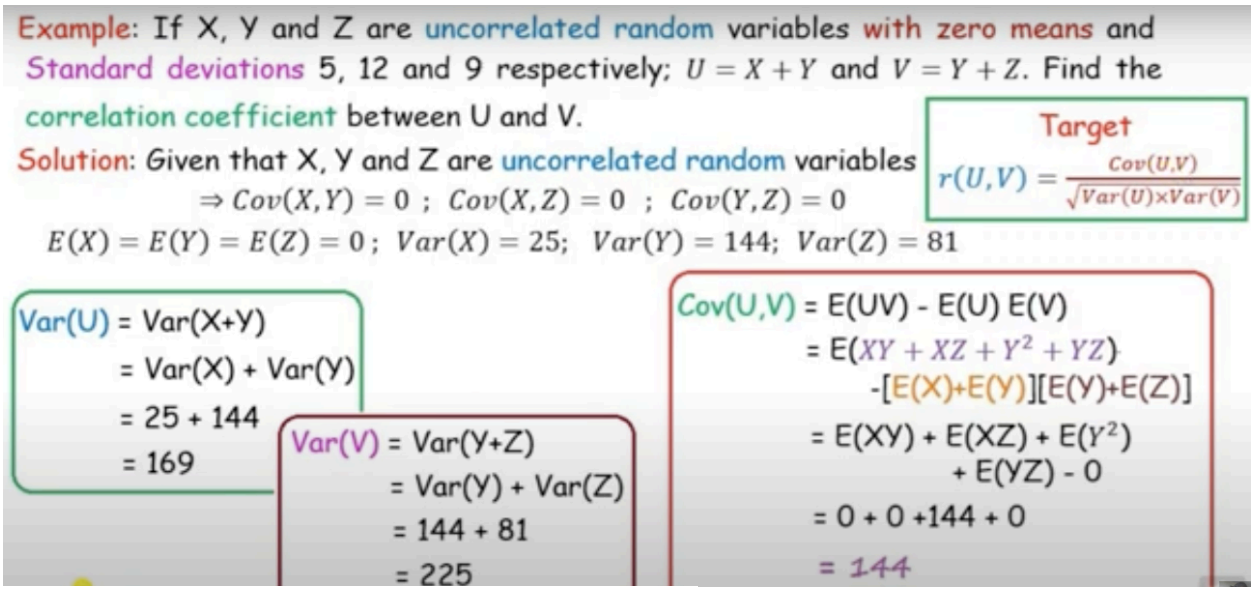
$$Cov(U,V) = Cov(X, X+Y)$$
$$= E[X(X+Y)] - E(X)\,E(X+Y)$$
$$= E(X^2 + XY) - E(X)[E(X)+E(Y)]$$
$$= E(X^2) + E(XY) - [E(X)]^2 - E(X)\,E(Y)$$
$$= Var(X) + E(XY) - E(X)\,E(Y)$$
$$= k + Cov(X,Y)$$
$$= k + rk$$
$$= k\,(1+r)$$

$$Var(U) = Var(X) = k$$

$$Var(V) = Var(X+Y)$$
$$= Var(X)+Var(Y)+2Cov(X,Y)$$
$$= k+k+2rk$$
$$= 2k(1+r)$$

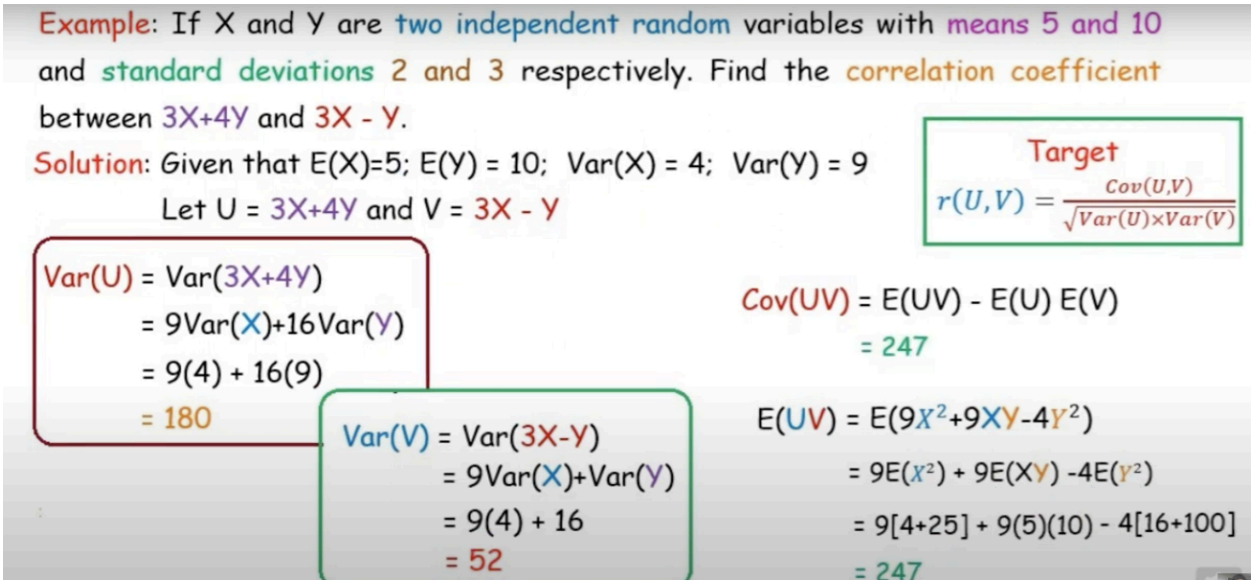
Thus, $r(U,V) = \frac{k(1+r)}{\sqrt{k \times 2k(1+r)}}$

$$= \sqrt{\frac{1+r}{2}}$$



Hence, $r(U,V) = \frac{144}{13 \times 25}$

12.



$Cov(UV) = E(UV) - E(U)\,E(V)$

$E(U) = 3E(X) + 4E(Y)$

$= 15 + 40 = 55$

$E(V) = 3E(X) - E(Y)$

$= 15 - 10 = 5$

$\therefore\ Cov(UV) = 247 - (55)(5) = -28$

Hence,

$$r(U,V) = \frac{-28}{\sqrt{180 \times 52}}$$

13

Example: Let X and Y be jointly distributed with the correlation coefficient $\frac{1}{2}$. The variance of X and Y is 4 and 9. Find Var(2X - 4Y + 3).

Solution: Given that $r(X,Y) = \frac{1}{2}$; $Var(X) = 4$; $Var(Y) = 9$

Var(2X - 4Y + 3)

= Var (2X - 4Y) + Var(3)

= Var (2X - 4Y) + 0 since Var(3) = 0

= 4Var(X) + 16 Var(Y) + 2(2)(-4) Cov(X,Y)

= 4(4) + 16 (9) -16 (1/2)(6)

= 112

Use $Var(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X,Y)$

14

Example: X and Y are two random variables with variances σ_X^2 and σ_Y^2 respectively and "r" is the correlation coefficient between them. If U = X+kY and $V = X + \frac{\sigma_X}{\sigma_Y}Y$, find the value of k so that U and V are uncorrelated.

Solution: For U and V are uncorrelated

$$\Rightarrow r(U,V) = 0$$

$$\Rightarrow Cov(U,V) = 0$$

$$\Rightarrow E(UV) - E(U)\,E(V) = 0$$

$$E(U) = E(X) + kE(Y)$$

$$E(V) = E(X) + \frac{\sigma_X}{\sigma_Y}E(Y)$$

$$E(UV) = E\left(X^2 + kXY + \frac{\sigma_X}{\sigma_Y}XY + \frac{k\sigma_X}{\sigma_Y}Y^2\right)$$

$$= E(X^2) + \left(k + \frac{\sigma_X}{\sigma_Y}\right)E(XY) + \frac{k\sigma_X}{\sigma_Y}E(Y^2)$$

$$= \left[\sigma_X^2 + (E(X))^2\right] + \left(k + \frac{\sigma_X}{\sigma_Y}\right)E(XY) + \frac{k\sigma_X}{\sigma_Y}\left[\sigma_Y^2 + (E(Y))^2\right]$$

$$E(U)E(V) = (E(X))^2 + \left(k + \frac{\sigma_X}{\sigma_Y}\right)E(X)E(Y) + \frac{k\sigma_X}{\sigma_Y}(E(Y))^2$$

Now $E(UV) = E(U)\,E(V)$

$$\Rightarrow \left[\sigma_X^2 + (E(X))^2\right] + \left(k + \frac{\sigma_X}{\sigma_Y}\right)E(XY) + \frac{k\sigma_X}{\sigma_Y}\left[\sigma_Y^2 + (E(Y))^2\right] = (E(X))^2 + \left(k + \frac{\sigma_X}{\sigma_Y}\right)E(X)E(Y) + \frac{k\sigma_X}{\sigma_Y}(E(Y))^2$$

$$\Rightarrow \sigma_X^2 + \frac{k\sigma_X}{\sigma_Y}\sigma_Y^2 + \left(k + \frac{\sigma_X}{\sigma_Y}\right)Cov(X,Y) = 0$$

$$\Rightarrow \sigma_X^2 + k\sigma_X\sigma_Y + \left(k + \frac{\sigma_X}{\sigma_Y}\right)r\sigma_X\sigma_Y = 0$$

$$\Rightarrow (1+r)\sigma_X^2 + k(1+r)\sigma_X\sigma_Y = 0$$

$$\Rightarrow (1+r)\sigma_X(\sigma_X + k\sigma_Y) = 0$$

Since $Cov(X,Y) = E(XY) - E(X)E(Y)$

Use $r = \frac{Cov(X,Y)}{\sqrt{Var(X)\times Var(Y)}}$

perf. Neg

Since $r \neq -1, \sigma_X, \sigma_Y > 0$, thus

$$\sigma_X + k\,\sigma_Y = 0$$

$$\Rightarrow k = -\frac{\sigma_X}{\sigma_Y}$$

15

Example: The covariance of two perfectly correlated variable X and Y is 0.96. Find the S.D. of X and Y if it is known that variance of X and Y are in the ratio of 4:9.

Solution: Cov(X,Y) = 0.96 ; σ_X =?; σ_Y = ? $\frac{\sigma_X^2}{\sigma_Y^2} = \frac{4}{9}$

For perfect correlation; $r = 1$ or -1

Since Cov(X,Y) is positive, thus $r = 1$

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

$$1 = \frac{0.96}{\sigma_X\sigma_Y}$$

$$\Rightarrow \sigma_X\sigma_Y = 0.96$$

$$\frac{\sigma_X^2}{\sigma_Y^2} = \frac{4}{9} \Rightarrow \sigma_X = \frac{2}{3}\sigma_Y$$

$$\Rightarrow \frac{2}{3}\sigma_Y^2 = 0.96$$

$$\Rightarrow \sigma_Y = 1.2$$

Therefore, $\sigma_X = 0.8$

16.

Example: Calculate the coefficient of correlation between X and Y from the following data:

Summation of product deviation of X and Y from their respective mean is 122.

	Series	
	X	Y
No. of pairs	15	15
Mean	25	18
Sum of Squares of deviations from mean	$\sum(X-\bar{X})^2 = 136$	$\sum(Y-\bar{Y})^2 = 138$

Solution: Given $\sum(X-\bar{X})(Y-\bar{Y}) = 122$

$$Cov(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n} = \frac{122}{15}$$

$$Var(X) = \frac{\sum(X-\bar{X})^2}{n} = \frac{136}{15}$$

$$Var(Y) = \frac{\sum(Y-\bar{Y})^2}{n} = \frac{138}{15}$$

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X) \times Var(Y)}}$$

$$= \frac{122/15}{\sqrt{\frac{136}{15} \times \frac{138}{15}}}$$

$$= 0.8918$$

Rank Correlation

A group of n individuals is arranged in order of merit or proficiency in possession of two characteristics A and B. These ranks in the two characteristics will in general be different.

For example, if we consider the relation between intelligence and beauty it is not necessary that a beautiful individual is intelligent also

Spearman's formula for the rank correlation coefficient.

or

Spearman's rank correlation coefficient

To compute the Spearman's rank correlation coefficient (ρ) between the two variables by

(i)

$$\rho = 1 - 6\frac{\sum d^2}{n(n^2-1)}$$

When there is a no tie during the rank

when there is a tie during the rank

(ii)

$$\rho = 1 - 6\frac{\left[\sum d^2 + \sum \frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

To $\sum d^2$ we add $\frac{m(m^2-1)}{12}$ for each value repeated, where m is the number of times a value occurs.

Tied Ranks. If some of the individuals recei.ve the same rank in a ranking or merit, they are said to be tied. Let us suppose that **m**

of the individuals, say, **(k** + 1)th, **(k** + 2)th, ... , **(k** + m)th are tied. Then each of these **m** individuals is assigned a common rank, which is the arithmetic mean of the ranks **k+1,** k +2, **..... k+m.**

i.e common rank value= [(k+1) + (k +2)+ **..... +(k+m)]/ m**

for example

two terms 4th and 5th have same value then common rank value will be (4+5)/2 = 4.5

Lets say three entries 7th , 8th and 9th have same value then Common rank assigned to them is (7+8+9)/3 =8

Example: The following data gives the HDL and LDL cholesterol levels of 7 adults in a locality

HDL(X)	36	39	23	31	33	51	45
HDL(Y)	80	72	101	90	98	70	50

Compute the rank correlation coefficient (ρ)

$$\rho = 1 - 6\frac{\sum d^2}{n(n^2-1)}$$

$$= 1 - 6 \times \frac{108}{7(7^2-1)}$$

$$= -0.928$$

HDL (X)	HDL (Y)	Rank of X (R_X)	Rank of Y (R_Y)	$d = R_X - R_Y$	d^2
36	80	4	4	0	0
39	72	5	3	2	4
23	101	1	7	-6	36
31	90	2	5	-3	9
33	98	3	6	-3	9
51	70	7	2	5	25
45	50	6	1	5	25
Total					108

Example 10·18. *Obtain the rank correlation coefficient for the following data:*

X	:	68	64	75	50	64	80	75	40	55	64
Y	:	62	58	68	45	81	60	68	48	50	70

X	Y	Rank X (x)	Rank Y (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
				$\Sigma d = 0$	$\Sigma d^2 = 72$

$$\rho = 1 - \frac{6\left[\Sigma d^2 + \frac{5}{2} + \frac{1}{2}\right]}{n(n^2 - 1)} =$$

$$1 - \frac{6(72 + 3)}{10 \times 99} = 0{\cdot}545$$

Example: Compute the rank correlation coefficient for the following grades of 12 students selected at random

Mathematics grade	85	83	87	84	88	82	90	86	88	87	89	90
Economic Grade	88	86	88	86	90	86	88	85	92	86	88	91

Solution:

(X)	(Y)	Rank of X (R_X)	Rank of Y (R_Y)	$d = R_X - R_Y$	d^2
85	88	4	7.5	-3.5	12.25
83	86	2	3.5	-1.5	2.25
87	88	6.5	7.5	-1.0	1
84	86	3	3.5	-0.5	0.25
88	90	8.5	10	-1.5	2.25
82	86	1	3.5	-2.5	6.25
90	88	11.5	7.5	4.0	16
86	85	5	1	4.0	16
88	92	8.5	12	-3.5	12.25
87	86	6.5	3.5	3.0	9
89	88	10	7.5	2.5	6.25
90	91	11.5	11	0.5	0.25
Total					84

$$r_{sp} = 1 - 6\frac{\left[\sum d^2 + \sum \frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

$$= 1 - \frac{6}{12(12^2-1)}\left[84 + \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{4(4^2-1)}{12} + \frac{4(4^2-1)}{12}\right]$$

$$= -0.666$$

Example 12 *Ten competitors in a beauty contest were ranked by three judges as follows:*

	Competitors									
Judges	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A:	6	5	3	10	2	4	9	7	8	1
B:	5	8	4	7	10	2	1	6	9	3
C:	4	9	8	1	2	3	10	5	7	6

Discuss which pair of judges have the nearest approach to common taste of beauty.

Solution

Rank by A (U)	*Rank by B (V)*	*Rank by C (W)*	$d_1 = U - V$	$d_2 = V - W$	$d_3 = U - W$	d_1^2	d_2^2	d_3^2
6	5	4	1	1	2	1	1	4
5	8	9	−3	−1	−4	9	1	16
3	4	8	−1	−4	−5	1	16	25
10	7	1	3	6	9	9	36	81
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
1	3	6	−2	−3	−5	4	9	25
					Total:	157	214	158

$$r_{UV} = 1 - \frac{6\,\Sigma d_1^2}{n\left(n^2-1\right)} = 1 - \frac{6\times 157}{10\times 99} = 0.0485$$

$$r_{VW} = 1 - \frac{6\,\Sigma d_2^2}{n\left(n^2-1\right)} = 1 - \frac{6\times 214}{10\times 99} = -0.2970$$

$$r_{UW} = 1 - \frac{6\,\Sigma d_3^2}{n\left(n^2-1\right)} = 1 - \frac{6\times 158}{10\times 99} = 0.0424$$

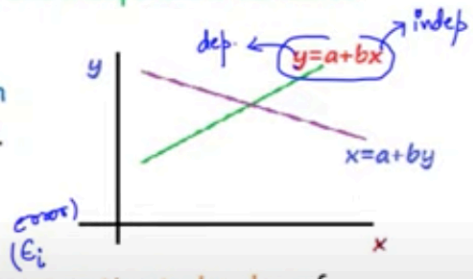
Since r_{UV} is maximum, the judges A and B may be considered to have common taste of beauty to some extent compared to other pairs of judges.

Unit 2 MTH302
Part II : Linear Regression with properties

Regression Lines:

A regression line is a graphic technique to show the functional relationship between two variables X and Y, i.e., dependent and independent variables.

It is a line which shows average relationship between two variables X and Y. Thus, this is a line of average.



This is also called estimating lines, as it gives the average estimated value of dependent variable (Y) for any given value of independent variable (X).

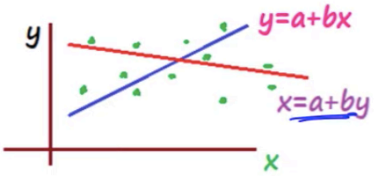
Regression Equations/Estimating lines:

There are two algebraic expression of regression lines,

1) the regression equation of X on Y

$$X = a + bY$$

which shows the variation in the values of X for given changes in Y.

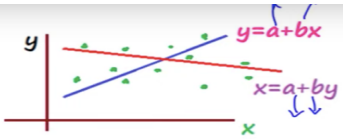


2) the regression equation of Y on X

$$Y = a + bX$$

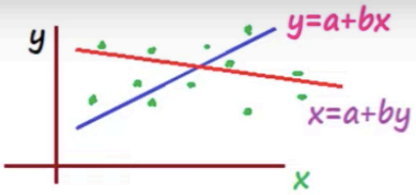
which shows the variation in the values of Y for given changes in X.

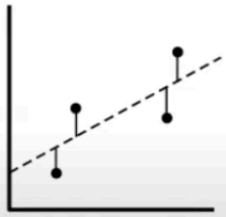
The objective of these regression lines is to fit the data on the lines.

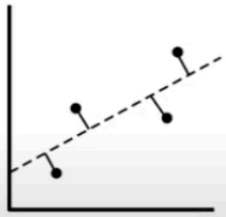


For this, we need to estimate unknown parameters a and b

A mathematical procedure for finding the best-fitting curve to a given set of points by **minimizing the sum of the squares** of the offsets ("the residuals") of the points from the curve




vertical offsets


perpendicular offsets

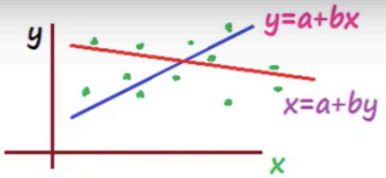
Note: In practice, the *vertical offsets* from a line are almost always minimized instead of the perpendicular offsets.

For Y on X; the regression equation is

$$Y = a + bX$$

For any given X, an estimated value Y_e of Y is

$$Y_e = a + bX$$



By principle of least squares, we minimize the residual errors i.e.,

$$E = \sum(Y - Y_e)^2$$

$$= \sum(Y - a - bX)^2$$

For minimization; we have

$$\frac{\partial E}{\partial a} = 0 \Rightarrow -2\sum(Y - a - bX) = 0$$

$$\frac{\partial E}{\partial b} = 0 \Rightarrow -2\sum(Y - a - bX)X = 0$$

$$\Rightarrow \sum(Y - a - bX) = 0 \ ;$$

$$\sum(XY - aX - bX^2) = 0$$

$$\Rightarrow \sum Y = na + b\sum X \ ;$$

$$\sum XY = a\sum X + b\sum X^2$$

These equations are called as NORMAL EQUATIONS.

Hence,

for Y on X; the regression equation is

$$Y = a + bX$$

Its normal equations are

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Similarly, $a \to 1$, $b \to Y$

For X on Y; the regression equation is

$$X = a + bY$$

The Normal equations are

$$\sum X = na + b\sum Y$$

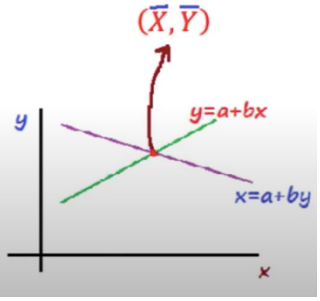
$$\sum XY = a\sum Y + b\sum Y^2$$

Note:

the point of intersection of the two-regression lines

$Y = a + bX$ and $X = a + bY$

gives the MEAN of the X and Y.



For Y on X line, regression line

$Y = a + bX$,

slope is b, so call as regression coefficient of Y on X and is denoted as b_{yx}.

b_{yx}

Similarly,

For X on Y,

$X = a + bY$

b_{xy} represent the regression coefficient.

Expression of b_{xy} & b_{yx}

for Y on X; the regression equation is

$$Y = a + bX$$

Its normal equations are

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

After solving, we get "a" and "b"

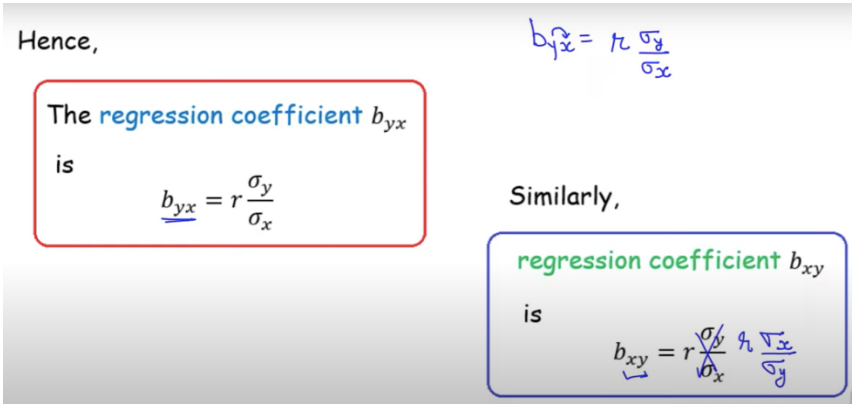
$$b_{yx} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$= \frac{\frac{\sum XY}{n} - \frac{\sum X}{n}\frac{\sum Y}{n}}{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$$

$$= \frac{E(XY) - E(X)E(Y)}{E(X^2) - (E(X))^2}$$

$$= \frac{cov(X,Y)}{Var(X)}$$

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



Properties of Regression Coefficients.

1.Correlation coefficient is the geometric mean between the regression coefficients.

$$b_{XY} \times b_{YX} = r\frac{\sigma_X}{\sigma_Y} \times r\frac{\sigma_Y}{\sigma_X} = r^2$$

$$r = \pm\sqrt{b_{XY} \times b_{YX}}$$

The sign of r is the same as that of b_{xy} and b_{yx}

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

3. Arithmetic. mean of the, regression coefficients is greater than the correlation coefficient r, provided r>0

$$\frac{1}{2}(b_{YX} + b_{XY}) \geq r$$

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

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

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

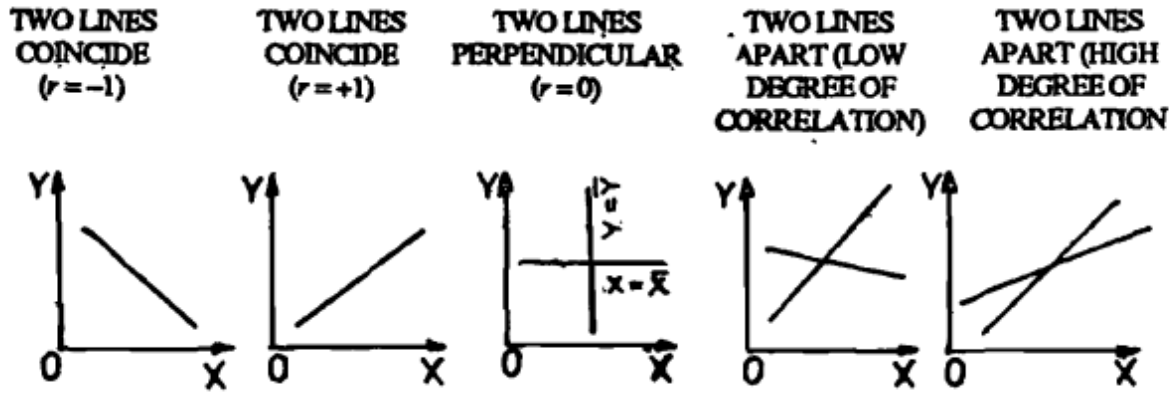
And

$$b_{XY} = (h/k)\, b_{UV}$$

5.If the two variables are uncorrelated (r=0) , the lines of regression become perpendicular to each other

in the case of perfect correlation ($r = \pm 1$) positive or negative, the two lines of regression coincide.

But since both the lines of regression pass through the point $(\bar{x}, \bar{y})$, they cannot be parallel.



Case (i). ($r = 0$). If $r = 0$, $\tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$

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

Thus if the two variables are uncorrelated, the lines of regression become perpendicular to each other

Example 1 *In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible: Variance of X = 1. The regression equations are $3x + 2y = 26$ and $6x + y = 31$. What were (i) the mean values of X and Y ? (ii) the standard deviation of Y ? and (iii) the correlation coefficient between X and Y ?*

Solution

(i) Since the lines of regression intersect at $(\bar{x}, \bar{y})$, we have $3\bar{x} + 2\bar{y} = 26$ and $6\bar{x} + \bar{y} = 31$

Solving these equations, we get $\bar{x} = 4$ and $\bar{y} = 7$.

(ii) Which of the two equations is the regression equation of Y on X and which one is the regression equation of X on Y are not known.

Let us tentatively assume that the first equation is the regression line of X on Y and the second equation is the regression line of Y on X. Based on this assumption, the first equation can be re-written as

$$x = -\frac{2}{3}y + \frac{26}{3} \quad (1)$$

and the other as $$y = -6x + 31 \quad (2)$$

Then $b_{XY} = -\frac{2}{3}$ and $b_{YX} = -6$

$\therefore$ $$r_{XY}^2 = b_{XY} \times b_{YX} = 4$$

$\therefore$ $r_{XY} = -2$, which is absurd.

Hence our tentative assumption is wrong.

$\therefore$ The first equation is the regression line of Y on X and re-written as

$$y = -\frac{3}{2}x + 13 \quad (3)$$

The second equation is the regression line of X on Y and re-written as

$$x = -\frac{1}{6}y + \frac{31}{6} \quad (4)$$

Hence the correct $b_{YX} = -\frac{3}{2}$ and the correct $b_{XY} = -\frac{1}{6}$

$\therefore$ $$r_{XY}^2 = b_{YX} \cdot b_{XY} = \frac{1}{4}$$

$\therefore$ $$r_{XY} = -\frac{1}{2}$$ ($\because$ both b_{YX} and b_{XY} are negative)

(iii) Now $$\frac{\sigma_Y^2}{\sigma_X^2} = \frac{b_{YX}}{b_{XY}} = \frac{-\frac{3}{2}}{-\frac{1}{6}} = 9$$

$\therefore$ $$\sigma_Y^2 = 9 \times \sigma_x^{\,2} = 9$$

$\therefore$ $$\sigma_Y = 3$$

2.

Example: The following results were declared in Physics and Mathematics in B.Tech examination. Find

(i) Regression lines

(ii) Regression coefficients

(iii) Estimate the value of Y when X = 40

(iv) Estimate the value of X when Y = 20.

	Scores in Physics (X)	Score in Mathematics (Y)
Mean	30	40
S.D.	10	20

Karl Pearson's coefficient of correlation between X and Y is 0.4.

Solution:

The regression line of Y on X is

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$
$$\Rightarrow Y - 40 = b_{yx}(X - 30)$$
$$\Rightarrow Y - 40 = 0.8(X - 30)$$
$$\Rightarrow Y = 0.8X + 16$$

The regression line of X on Y is

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.4\left(\frac{10}{20}\right) = 0.2$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.4\left(\frac{20}{10}\right) = 0.8$$

The regression line of Y on X is

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$
$$\Rightarrow Y - 40 = b_{yx}(X - 30)$$
$$\Rightarrow Y - 40 = 0.8(X - 30)$$
$$\Rightarrow Y = 0.8X + 16$$

When X = 40, then Y = 48

When Y = 20, then X = 26

The regression line of X on Y is

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$
$$\Rightarrow X - 30 = 0.2(Y - 40)$$
$$\Rightarrow X = 0.2Y + 22$$

3.

Example: Estimate X when Y=10, if the two lines of regressions are

$$X = -\frac{1}{18}Y + \lambda; \quad Y = -2x + \mu$$

$b_{yx} = -2$ $b_{xy} = -\frac{1}{18}$

where (λ, μ) are unknown and the mean of the distribution is at $(-1, 2)$. Find r, λ, μ.

Solution: Since mean (-1, 2) passes through regression lines

$$\Rightarrow -1 = -\frac{2}{18} + \lambda \quad ; \quad 2 = 2 + \mu$$

$$\Rightarrow \lambda = -\frac{8}{9} \quad ; \quad \mu = 0$$

$r = \sqrt{-\frac{1}{18} \times -2} = -\frac{1}{3}$

Regression line of X on Y is

$$X = -\frac{1}{18}Y - \frac{8}{9}$$

When $Y = 10$;

then $X = -\frac{1}{18}(10) - \frac{8}{9} = -\frac{13}{9}$

4.

Example: By using the following data, find the two lines of regression and Karl-Pearson's coefficient of correlation.

$$\sum X = 250\,;\sum Y = 300;\sum XY = 7900\,;\sum X^2 = 6500\,;\sum Y^2 = 10000, N = 10$$

Solution:

Regression line of Y on X is

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$$

Regression line of X on Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{cov(X,Y)}{\sigma_X^2} = \frac{E(XY) - E(X)E(Y)}{E(X^2) - (E(X))^2} = \frac{\frac{7900}{10} - \left(\frac{250}{10}\right)\left(\frac{300}{10}\right)}{\frac{6500}{10} - \left(\frac{250}{10}\right)^2} = 1.6$$

Regression line of Y on X is

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$\Rightarrow Y - 30 = 1.6(X - 25)$$

$$\Rightarrow Y = 1.6X - 10$$

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y} = \frac{cov(X,Y)}{\sigma_Y^2} = \frac{E(XY) - E(X)E(Y)}{E(Y^2) - (E(Y))^2} = \frac{\frac{7900}{10} - \left(\frac{250}{10}\right)\left(\frac{300}{10}\right)}{\frac{10000}{10} - \left(\frac{250}{10}\right)^2} = 0.4$$

Regression line of X on Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$\Rightarrow X - 25 = 0.4(Y - 30)$$

$$\Rightarrow X = 0.4Y + 13$$

$r^2 = b_{xy} \cdot b_{yx} = 0.4 \times 1.6 = 0.64$

$r = 0.8$

Standard Error of Estimate of Y

Although we use the regression line of Y on X to predict the value of Y corresponding to a specified value of X we may also use it to estimate the value of Y corresponding to an observed value of $X = x_i$, say. The value of Y estimated in this manner need not, in general, be equal to the corresponding observed value

of Y, namely, y_i. Hence the difference between Y and Y_E is called *the error of estimate of Y*. This error will vary from one observed value to the other and a random variable. The standard deviation of this RV $(Y - Y_E)$ is called *the standard error of estimate of Y* and denoted by S_Y.

The standard error of estimate of Y is S_Y

$$S^2_Y = = (1 - r^2) \sigma^2_Y \text{ or } S_Y = \sqrt{1 - r^2}\, \sigma_Y \qquad (1)$$

Similarly, the standard error of estimate of X, denoted by S_X is given by

$$S^2_X = (1 - r^2) \sigma^2_X \text{ or } S_X = \sqrt{1 - r^2}\, \sigma_X \qquad (2)$$

Example 5 *Find the standard error of estimate of Y on X and of X on Y from the following data:*

X:	1	2	3	4	5
Y:	2	5	9	13	14

Solution

x	y	x^2	y^2	xy
1	2	1	4	2
2	5	4	25	10
3	9	9	81	27
4	13	16	169	52
5	14	25	196	70
15	43	55	475	161

$$r_{XY} = \frac{n\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{\left\{n\Sigma x^2 - (\Sigma x)^2\right\}\left\{n\Sigma y^2 - (\Sigma y)^2\right\}}}$$

$$= \frac{5 \times 161 - 15 \times 43}{\sqrt{\left\{5 \times 55 - (15)^2\right\}\left\{5 \times 475 - (43)^2\right\}}}$$

$$= \frac{160}{\sqrt{50 \times 526}} = 0.9866$$

$$\sigma^2_X = \frac{1}{n}\Sigma x^2 - \left(\frac{1}{n}\Sigma x\right)^2$$

$$= \frac{1}{5} \times 55 - \left(\frac{1}{5} \times 15\right)^2 = 2$$

$\therefore$ $\sigma_X = 1.4142$

$$\sigma^2_Y = \frac{1}{n}\Sigma y^2 - \left(\frac{1}{n}\Sigma y\right)^2$$

$$= \frac{1}{5} \times 475 - \left(\frac{1}{5} \times 43\right)^2$$

$$= 21.04$$

$\therefore$ $\sigma_Y = 4.5869$

$$S_Y = \sqrt{1 - r^2} \cdot \sigma_Y = \sqrt{1 - (0.9866)^2} \times 4.5869$$

$$= 0.7484$$

$$S_X = \sqrt{1 - r^2} \cdot \sigma_X = \sqrt{1 - (0.9866)^2} \times 1.4142$$

$$= 0.2307$$

Example: In partially destroyed laboratory record relating to correlation data, the following results are legible,

$\sigma_x^2 = 9$, Regression equations $8X - 10Y + 66 = 0;\ 40X - 18Y = 214$

Find (i) identify which line is of Y on X and X on Y (ii) mean of X and Y

(iii) S.D. of Y (iv) co-efficient of correlation between X and Y.

Solution: (i) Regression equations are

$$8X - 10Y + 66 = 0; \quad 40X - 18Y = 214$$

$$\Rightarrow Y = \frac{8}{10}X + \frac{66}{10} \quad ; \quad X = \frac{18}{40}Y + \frac{214}{40}$$

$$\Rightarrow b_{YX} = \frac{8}{10} \quad ; \quad b_{XY} = \frac{18}{40}$$

Thus,

Regression line of Y on X is $Y = \frac{8}{10}X + \frac{66}{10}$

Regression line of X on Y is $X = \frac{18}{40}Y + \frac{214}{40}$

(ii) Mean is the point of intersection of regression lines

$$8X - 10Y + 66 = 0;$$

$$40X - 18Y = 214$$

On solving, we get X = 13 & Y = 17

Thus, E(X) = 13 and E(Y) = 17

(iii) Correlation coefficient is

$$r = \sqrt{b_{XY} \times b_{YX}}$$

$$= \sqrt{\frac{8}{10} \times \frac{18}{40}}$$

$$= 0.6$$

Find σ_Y.

Using $b_{YX} = r\frac{\sigma_Y}{\sigma_X}$

$$\Rightarrow \frac{8}{10} = 0.6\frac{\sigma_Y}{3}$$

$$\Rightarrow \sigma_Y = 4$$

7

Example: If the two regression coefficients are 0.8 and 1.2, what would be the value of coefficient of correlation?

Solution: Given that $b_{xy} = {}^{+}0.8$; $b_{yx} = {}^{+}1.2$

b_{yx} ; b_{xy} & r have same sign

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$$= \pm\sqrt{0.8 \times 1.2}$$

$$= \pm 0.98$$

8

Example: Find the coefficient of correlation from the following two regression equations: 3Y-2X-10=0 and 2Y-50-X=0. Also, estimate the value of Y when X=0.

Solution: The regression lines are

$$3Y - 2X - 10 = 0 \text{ and } 2Y - 50 - X = 0$$

Y on X: $Y = \frac{2}{3}X + \frac{10}{3}$; $b_{yx} = \frac{2}{3}$

X on Y: $X = 2Y - 50$; $b_{xy} = 2$

Target $r = \sqrt{b_{XY} \times b_{YX}}$

X on Y Y on X

$r = \sqrt{4/3} \notin [-1, 1]$

$= 1.\ldots > \notin [-1, 1]$

The regression lines are

$$3Y - 2X - 10 = 0 \text{ and } 2Y - 50 - X = 0$$

$$\Rightarrow X = \frac{3}{2}Y - 5 \quad ; \quad Y = \frac{1}{2}X + 25$$

$$\Rightarrow b_{XY} = \frac{3}{2} \quad ; \quad b_{YX} = \frac{1}{2}$$

$$r = \sqrt{b_{XY} \times b_{YX}} = \sqrt{\frac{3}{2} \times \frac{1}{2}} = \frac{\sqrt{3}}{2} = 0.87$$

When X=0 then using regression line of Y on X, we get

$$Y = \frac{1}{2}(0) + 25 = 25$$

9

Example: The regression coefficient of regression equation of X on Y is 2.4 and of Y on X is 0.8. Are the regression coefficients consistent?

Solution: Given that $b_{XY} = 2.4$; $b_{YX} = 0.8$

Now, $r = \sqrt{b_{XY} \times b_{YX}}$

$= \sqrt{2.4 \times 0.8}$

$= \sqrt{1.92}$

$= 1.3856 \notin [-1, 1]$

Thus, the given regression coefficients are not consistent.

10

Example: Find the mean values of the variables X and Y and correlation coefficient for the following regression equations

$2Y - X - 50 = 0$; $3Y - 2X - 10 = 0$

Solution: Mean is the point of intersection of regression lines

$2Y - X - 50 = 0$; $3Y - 2X - 10 = 0$

On solving, we get

$X = 130$ & $Y = 90$

Thus, E(X) = 130 and E(Y) = 90

$Y = \frac{1}{2}X + 25$ $\quad X = \frac{3}{2}Y - \frac{10}{2}$

$b_{XY} = \frac{3}{2}$ $\quad b_{YX} = \frac{1}{2}$

$r = \sqrt{b_{YX}\, b_{XY}}$

$= \sqrt{\frac{1}{2} \times \frac{3}{2}}$

$= \sqrt{\frac{3}{4}} \in [0, 1)$

$= \frac{\sqrt{3}}{2}$

11

Example: For 50 students of a class the regression equation of marks in Statistics (X) on marks in Mathematics (Y) is $3Y - 5X + 180 = 0$. The mean mark in mathematics is 44 and variance of marks in Statistics is 9/16th of the variance of marks in Mathematics. Find the mean marks in Statistics and the coefficient of correlation between the marks in two subjects.

Solution: Given that E(Y) = 44 ; $\sigma_X^2 = \frac{9}{16}\sigma_Y^2$

To find $E(X)$ and r

Since mean of X and Y passes through regression lines

Thus, $3\bar{Y} - 5\bar{X} + 180 = 0$

$\Rightarrow 3(44) - 5\bar{X} + 180 = 0$

$\Rightarrow \bar{X} = 62.4$

Regression line of X on Y is

$3Y - 5X + 180 = 0$

$\Rightarrow X = \frac{3}{5}Y + 36$

Thus, $b_{XY} = \frac{3}{5}$

$\Rightarrow r\frac{\sigma_X}{\sigma_Y} = \frac{3}{5}$

$\Rightarrow r\left(\frac{3}{4}\right) = \frac{3}{5}$

$\Rightarrow r = \frac{4}{5}$ ✓

12

Example: Consider the following information about series X and Y. The coefficient of correlation between X and Y is +0.8. Find out the most probable value of Y if X is 70 and most probable value of X if Y is 90.

	Series X	Series Y
Mean	18	100
S.D.	14	20

Solution: When X = 70; then

$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$

Regression line of Y on X is

$Y - \bar{Y} = b_{YX}(X - \bar{X})$

$\Rightarrow Y - 100 = 0.8\left(\frac{20}{14}\right)(70 - 18)$

$\Rightarrow Y = 159.28$

When Y = 90, then

$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$

Regression line of X on Y is

$X - \bar{X} = b_{XY}(Y - \bar{Y})$

$\Rightarrow X - 18 = 0.8\left(\frac{14}{20}\right)(90 - 100)$

$\Rightarrow X = 12.40$

13

Example: Given that the means of X and Y are 65 and 67, their standard deviations are 2.5 and 3.5 respectively and the correlation coefficient between them is 0.8.

(i) Write down the regression lines.

(ii) Obtain the best estimate of X when Y=70.

(iii) Using the estimated value of X as the given value of X, estimate the corresponding value of Y.

Solution: Given that $E(X) = 65$; $E(Y) = 67$; $\sigma_X = 2.5$; $\sigma_Y = 3.5$; $r = 0.8$

$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$ $\quad$ $b_{XY} = r\frac{\sigma_X}{\sigma_Y}$

(i) Regression line of Y on X is

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$\Rightarrow Y - 67 = 0.8\left(\frac{3.5}{2.5}\right)(X - 65)$$

$$\Rightarrow Y = 1.12X - 5.8$$

Regression line of X on Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$\Rightarrow X - 65 = 0.8\left(\frac{2.5}{3.5}\right)(Y - 67)$$

$$\Rightarrow X = 0.571Y + 26.743$$

Regression line of Y on X is

$$\Rightarrow Y = 1.12X - 5.8$$

Regression line of X on Y is

$$\Rightarrow X = 0.571Y + 26.743$$

(ii) When Y = 70, then value of X is

$$X = 0.571(70) + 26.743$$

$$= 66.713$$

(iii) When X = 66.713, then Y is

$$Y = 1.12(66.713) - 5.8$$

$$= 68.92$$

14

Example: The correlation coefficient between X and Y variables is 0.60. If $\sigma_X = 1.5$, $\sigma_Y = 2.0, \bar{X} = 10, \bar{Y} = 20$, find the equations of the regression lines (i) Y on X (ii) X on Y.

Solution: Given that $r = 0.6$; $\sigma_X = 1.5$; $\sigma_Y = 2$; $\bar{X} = 10$, $\bar{Y} = 20$

$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$ $\quad$ $b_{XY} = r\frac{\sigma_X}{\sigma_Y}$

Regression line of Y on X is

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$\Rightarrow Y - 20 = 0.6\left(\frac{2}{1.5}\right)(X - 10)$$

$$\Rightarrow Y = \frac{4}{5}X + 12$$

Regression line of X on Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$\Rightarrow X - 10 = 0.6\left(\frac{1.5}{2}\right)(Y - 20)$$

$$\Rightarrow X = 0.45Y + 1$$

15

Example: Find out σ_Y and r from the following data: $3X = Y;\ 4Y = 3X;\ \sigma_X = 2$

Solution:

S.D

Y on X

X on Y

$Y = 3X$ $\quad b_{YX} = 3$

$X = \frac{4}{3}Y$ $\quad b_{XY} = \frac{4}{3}$

$r = \sqrt{3 \times \frac{4}{3}}$

$= 2 \notin [-1, 1]$

The regression lines are

$$3X = Y;\ 4Y = 3X$$

Regression line of Y on X

is $Y = \frac{3}{4}X$

$\Rightarrow b_{YX} = \frac{3}{4}$

Hence,

$$r = \sqrt{b_{XY} \times b_{YX}}$$

$$= \sqrt{\frac{1}{3} \times \frac{3}{4}}$$

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

Regression line of X on Y

is $X = \frac{1}{3}Y$

$\Rightarrow b_{XY} = \frac{1}{3}$

For σ_Y

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$$

$$\Rightarrow \frac{3}{4} = \frac{1}{2}\frac{\sigma_Y}{2}$$

$$\Rightarrow \sigma_Y = 3$$

16

Example: The following data about the sales and advertisement expenditure of a firm is given in Table. Coefficient of correlation between them is **0.9**

	Sales (in crores)	Advertisement Expenditure (in crores)
Mean	40	6
S.D.	10	1.5

(i) Estimate the likely sales from a proposed advertisement expenditure of 10 crores.

(ii) What should be the advertisement expenditure if the firm proposes a sales target of 60 crores?

Solution:

(i) Given that Y = 10; X = ?

Regression line of X on Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$\Rightarrow X - 40 = 0.9\left(\frac{10}{1.5}\right)(10 - 6)$$

$$\Rightarrow X = 64 \text{ crores}$$

$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$

(ii) Given that X = 60; Y = ?

Regression line of Y on X is

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$\Rightarrow Y - 6 = 0.9\left(\frac{1.5}{10}\right)(60 - 40)$$

$$\Rightarrow Y = 8.7 \text{crores}$$

$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$

17

Example: Given that the regression equation of Y on X and X on Y respectively $Y = X$ and $4X - Y = 3$. Find the correlation coefficient between X and Y.

Solution: Given that

Regression line of Y on X is

Y=X

$\Rightarrow b_{YX} = 1$

Regression line of X on Y is

$X = \frac{1}{4}Y + \frac{3}{4}$

$\Rightarrow b_{XY} = \frac{1}{4}$

Hence,

$$r = \sqrt{b_{XY} \times b_{YX}}$$

$$= \sqrt{1 \times \frac{1}{4}}$$

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

18

Example: From the following data, $X = 0.854\,Y$; $Y = 0.89X$; $\sigma_X = 3$

Calculate (i) coefficient of correlation (ii) S.D. of Y

Solution: From the given lines, we can get

$b_{YX} = 0.89$; $b_{XY} = 0.854$

$$r = \sqrt{b_{XY} \times b_{YX}}$$

$$= \sqrt{0.89 \times 0.854}$$

$$= 0.87$$

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

$$\Rightarrow 0.854 = 0.87\frac{3}{\sigma_Y}$$

$$\Rightarrow \sigma_Y = 3.06$$

19.

Example 5 *Given that $x = 4y + 5$ and $y = kx + 4$ are the regression lines of X on Y and of Y on X respectively, show that $0 \le k \le \frac{1}{4}$. If $k = \frac{1}{16}$, find the means of X and Y and r_{XY}.*

Solution From the given equations, we note that

$$b_{YX} = k \text{ and } b_{XY} = 4$$

$$r_{XY}^2 = b_{XY} \cdot b_{YX} = 4k$$

Since $0 \le r_{XY}^2 \le 1$, we get $0 \le 4k \le 1$

$$\therefore \quad 0 \le k \le \frac{1}{4}.$$

When $k = \frac{1}{16}$, $r_{XY}^2 = \frac{1}{4}$

$$\therefore \quad r_{XY} = \pm \frac{1}{2}$$

But both b_{YX} and b_{XY} are positive.

$$\therefore \quad r_{XY} = \frac{1}{2}$$

When $k = \frac{1}{16}$, the regression equations become

$$x = 4y + 5 \qquad (1)$$

and

$$y = \frac{1}{16}x + 4 \qquad (2)$$

Solving equations (1) and (2), we get

$$x = 28 \text{ and } y = 5.75$$

$$\therefore \quad x = 28 \text{ and } y = 5.75$$

Angle Between Two Lines of Regression

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

Slopes of these lines are $r \cdot \frac{\sigma_Y}{\sigma_X}$ and $\frac{\sigma_Y}{r\sigma_X}$ respectively. If θ is the angle between the two lines of regression then

$$\tan\theta = \frac{r\frac{\sigma_Y}{\sigma_X} \sim \frac{\sigma_Y}{r\sigma_X}}{1 + r\frac{\sigma_Y}{\sigma_X} \cdot \frac{\sigma_Y}{r\sigma_X}} = \frac{r^2 \sim 1}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)$$

$$= \frac{1 \sim r^2}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right) \qquad (\because r^2 \le 1)$$

$$\therefore \quad \theta = \tan^{-1}\left\{\frac{1 - r^2}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)\right\} \qquad \ldots(10\cdot19)$$

Case (i). ($r = 0$). If $r = 0$, $\tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$

Example 6 *Find the angle between the two lines of regression. Deduce the condition for the two lines to be (i) at right angles and (ii) coincident.*

Solution The equations of the regression lines

are $$y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(x - \bar{x}) \qquad (1)$$

and $$x - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(y - \bar{y}) \qquad (2)$$

Slope of line (1) $= r\frac{\sigma_Y}{\sigma_X} = m_1$, say.

Slope of line (2) $= \frac{\sigma_Y}{r\sigma_X}$, m_2, say.

If θ is the acute angle between the two lines, then $\tan\theta = \frac{|m_1 - m_2|}{1 + m_1 m_2}$

$$= \frac{\left| r\frac{\sigma_Y}{\sigma_X} - \frac{\sigma_Y}{r\sigma_X} \right|}{1 + \frac{\sigma_Y^2}{\sigma_X^2}}$$

$$= \frac{\left| r - \frac{1}{r} \right| \sigma_X \sigma_Y}{\sigma_X^2 + \sigma_Y^2}$$

$$= \frac{\left(1 - r^2\right)}{|r|} \frac{\sigma_X \sigma_Y}{\sigma_X^2 + \sigma_Y^2}$$

The two regression lines are at angles when $\theta = \frac{\pi}{2}$, i.e., $\tan\theta = \infty$

i.e., $r = 0$

$\therefore$ When the linear correlation between X and Y is zero, the two lines of regression will be at right angles.

The two regression lines are coincident, when $\theta = 0$, i.e., when $\tan\theta = 0$

i.e., when $r = \pm 1$.

$\therefore$ When the correlation between X and Y is perfect, the two regression lines will coincide.

Properties of Regression coefficients with proof

Properties

- Value of $r = \sqrt{b_{XY} \times b_{YX}} \in [-1,1]$
  - b_{XY}, b_{YX}, r all have same sign.
- If one of regression coefficient is greater than 1 then other MUST be less than one.

etc....

1

Property: The geometric mean between the two regression coefficients is equal to the correlation coefficient, i.e., $r = \sqrt{b_{XY} \times b_{YX}}$

Proof: The regression coefficients are

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} \quad ; \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

b_{XY}, b_{YX}

On multiplying, we get

$$b_{XY} \times b_{YX} = r^2$$

$$\Rightarrow \quad r = \sqrt{b_{XY} \times b_{YX}}$$

Hence result.

Property 2: If one regression coefficient is greater than unity, then others will be lesser than unity.

Proof: We know that $r = \sqrt{b_{XY} \times b_{YX}}$

r is correlation coefficient

$$\Rightarrow b_{XY} \times b_{YX} = r^2 \leq 1$$

$$\Rightarrow b_{XY} \leq \frac{1}{b_{YX}}$$

If $b_{YX} > 1$, then $b_{XY} < 1$ and vice versa too.

Hence, the result.

Property 3: Both of the regression coefficients must have the same sign.

i.e., If b_{YX} is positive, b_{XY} will also be positive and it is true for vice versa.

Proof: We know that

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y} \quad ; \quad b_{YX} = r\frac{\sigma_Y}{\sigma_X}$$

On Dividing, we get

$$\frac{b_{XY}}{b_{YX}} = \frac{\sigma_X^2}{\sigma_Y^2} > 0$$

Thus, both the regression coefficients have same sign.

Property 4: The regression coefficients b_{YX} , b_{XY} and correlation coefficient r have the same sign.

Proof: We know that $b_{XY} = r\frac{\sigma_X}{\sigma_Y} \quad ; \quad b_{YX} = r\frac{\sigma_Y}{\sigma_X}$

$$\Rightarrow \frac{b_{XY}}{r} = \frac{\sigma_X}{\sigma_Y} > 0 \quad ; \quad \frac{b_{YX}}{r} = \frac{\sigma_Y}{\sigma_X} > 0$$ Since $\sigma_X, \sigma_Y > 0$,

Thus, nature of b_{XY}, b_{YX} and r have always same sign.

Hence, the result.

Example: The regression coefficient of regression equation of X on Y is $-2/3$ and of Y on X is $-3/4$. Find the correlation coefficient.

Solution: $b_{XY} = \ominus\frac{2}{3}$; $b_{YX} = \ominus\frac{3}{4}$

We know that $r = \sqrt{b_{XY} \times b_{YX}}$

$$= \pm\sqrt{-\frac{2}{3} \times -\frac{3}{4}}$$

$$= -\frac{1}{2}$$

Property 5: The arithmetic means of both regression coefficients is equal to or greater than the coefficient of correlation, i.e., $\frac{b_{XY}+b_{YX}}{2} \geq r$

Proof: The regression coefficients are

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} \quad ; \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

Now,

$$\frac{b_{XY}+b_{YX}}{2} = \frac{r}{2}\left(\frac{\sigma_X}{\sigma_Y}+\frac{\sigma_Y}{\sigma_X}\right)$$

$$= r\left(\frac{\sigma_X^2+\sigma_Y^2}{2\sigma_X\sigma_Y}\right)$$

Since

$$\sigma_X^2+\sigma_Y^2-2\sigma_X\sigma_Y = (\sigma_X-\sigma_Y)^2 \geq 0$$

$$\Rightarrow \sigma_X^2+\sigma_Y^2 \geq 2\sigma_X\sigma_Y$$

$$\Rightarrow \frac{\sigma_X^2+\sigma_Y^2}{2\sigma_X\sigma_Y} \geq 1$$

$$\Rightarrow r\left(\frac{\sigma_X^2+\sigma_Y^2}{2\sigma_X\sigma_Y}\right) \geq r$$

$$\Rightarrow \frac{b_{XY}+b_{YX}}{2} \geq r$$

Property 6: The regression coefficients are independent of the change of the origin, i.e., if $U = X \pm a$; $V = Y \pm b$ then $b_{UV} = b_{XY}$; $b_{VU} = b_{YX}$

Proof: By definition, $b_{UV} = r\frac{\sigma_U}{\sigma_V}$ where $r = r(U,V)$

We know that $r(X \pm a, Y \pm b) = r(X,Y)$, i.e., $r(U,V) = r(X,Y) = r$

$$\sigma_U^2 = Var(U)$$
$$= Var(X \pm a)$$
$$= Var(X)$$
$$= \sigma_X^2$$

$$\sigma_V^2 = Var(V)$$
$$= Var(Y \pm b)$$
$$= Var(Y)$$
$$= \sigma_Y^2$$

Hence, $b_{UV} = r\frac{\sigma_X}{\sigma_Y}$

$$= b_{XY}$$

Similarly, $b_{VU} = b_{YX}$

Example: The regression coefficient of regression equation of X on Y is 0.4 and of Y on X is 1.6. Find the regression coefficients of $X+3$ on $Y-2$ and Y-2 on X+3.

Solution: Since $b_{XY} = 0.4 \quad ; \quad b_{YX} = 1.6$

Take $U = X + 3 \; ; \quad V = Y - 2$, we get

$$b_{UV} = b_{XY} = 0.4$$

$$b_{VU} = b_{YX} = 1.6$$

Property 7: The regression coefficients are not independent of the change of the scale. If $U = aX \; ; V = bY$ then $b_{UV} \neq b_{XY} \; ; \; b_{VU} \neq b_{YX}$

If x and y are multiplied by any constant, then the regression coefficient will change.

Proof: $b_{UV} = r\frac{\sigma_U}{\sigma_V}$ where $r = r(U,V)$

We know that $r(aX, bY) = \begin{cases} r(X,Y) & \text{if } ab > 0 \\ -r(X,Y) & \text{if } ab < 0 \end{cases}$

$$\sigma_U^2 = Var(U) = Var(aX) = a^2 Var(X) = a^2\sigma_X^2$$

$$\sigma_V^2 = Var(V) = Var(bY) = b^2 Var(Y) = b^2\sigma_Y^2$$

Hence, $b_{UV} = \begin{cases} r\frac{a\sigma_X}{b\sigma_Y} & ; \; if \; ab > 0 \\ -r\frac{a\sigma_X}{b\sigma_Y} & ; \; if \; ab < 0 \end{cases}$

$$= \begin{cases} \frac{a}{b}b_{XY} & ; \; if \; ab > 0 \\ -\frac{a}{b}b_{XY} & ; \; if \; ab < 0 \end{cases}$$

Similarly, $b_{VU} = \begin{cases} \frac{b}{a}b_{YX} & ; \; if \; ab > 0 \\ -\frac{b}{a}b_{YX} & ; \; if \; ab < 0 \end{cases}$

Example: The regression coefficient of regression equation of X on Y is 0.4 and of Y on X is 1.6. Find the regression coefficients of $3X$ on $2Y$ and 2Y on 3X.

Solution: Since $b_{XY} = 0.4 \quad ; \quad b_{YX} = 1.6$

Take $U = 3X \; ; \quad V = 2Y$, we get

b/a


$b_{UV} = \frac{3}{2} b_{XY}$
$= \frac{3}{2}(0.4)$
$= 0.6$
$b_{VU} = \frac{2}{3} b_{YX}$
$= \frac{2}{3}(1.6)$
$= 1.066$