15. $\Sigma y = m\,a_0 + a_1 \Sigma x + a_2 \Sigma x^2 + a_3 \Sigma x^3$

$\Sigma xy = a_0 \Sigma x + a_1 \Sigma x^2 + a_2 \Sigma x^3 + a_3 \Sigma x^4$

$\Sigma x^2 y = a_0 \Sigma x^2 + a_1 \Sigma x^3 + a_2 \Sigma x^4 + a_3 \Sigma x^5$

$\Sigma x^3 y = a_0 \Sigma x^3 + a_1 \Sigma x^4 + a_2 \Sigma x^5 + a_3 \Sigma x^6$.

16. $y = \dfrac{3.774}{x(x-2)}$; $y(5) = 0.2516$

3.31 CORRELATION

In a bivariate distribution, 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 *i.e.*, 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.*

e.g., the correlation between income and expenditure is positive.

If the two variables deviate in opposite direction *i.e.*, if the increase (or decrease) in one results in a corresponding decrease (or increase) in the other, correlation is said to be *inverse or negative.*

e.g., the correlation between volume and the pressure of a perfect gas or the correlation between price and demand is negative.

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

3.32 REASONS RESPONSIBLE FOR THE EXISTENCE OF CORRELATION

1. Due to mere chance. The correlation between variables may be due to mere chance. Consider the data regarding six students selected at random from a college.

Students:	A	B	C	D	E	F
% of marks obtained in previous exam.	43%	47%	60%	80%	55%	40%
Height (in inches):	60	62	65	70	64	59

Here the variables are moving in the same direction and a high degree of correlation is expected between the variables. We cannot expect this degree of correlation to hold good for any other sample drawn from the concerned population. In this case, the correlation has occurred just due to chance.

2. Due to the effect of some common cause. The correlation between variables may be due to the effect of some common cause. For example, positive correlation between the number of girls seeking admission in colleges A and B of a city may be due to the effect of increasing interest of girls towards higher education.

3. Due to the presence of cause-effect relationship between variables. For example, a high degree correlation between 'temperature' and 'sale of coffee' is due to the fact that people like taking coffee in winter season.

4. Due to the presence of interdependent relationship between the variables. For example, the presence of correlation between amount spent on entertainment of family and total expenditure of family is due to fact that both variables affect each other.

3.33 SCATTER OR DOT DIAGRAMS

It is the simplest method of the diagrammatic representation of bivariate data. Let (x_i, y_i), $i = 1, 2, 3, \ldots\ldots, n$ be a bivariate distribution. Let the values of the variables x and y be plotted along the x-axis and y-axis on a suitable scale. Then corresponding to every ordered pair, there corresponds a point or dot in the xy-plane. The diagram of dots so obtained is called a *dot or scatter diagram.*

If the dots are very close to each other and the number of observations is not very large, a fairly good correlation is expected. If the dots are widely scattered, a poor correlation is expected.

3.34 KARL PEARSON'S CO-EFFICIENT OF CORRELATION (OR PRODUCT MOMENT CORRELATION CO-EFFICIENT)

[U.P.T.U. (C.O.) 2009; U.P.T.U. 2006, 2007, 2015]

Correlation co-efficient between two variables x and y, usually denoted by $r(x, y)$ or r_{xy} is a numerical measure of linear relationship between them and is defined as

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2\,\Sigma(y_i - \bar{y})^2}} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2 \cdot \frac{1}{n}\Sigma(y_i - \bar{y})^2}} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}.$$

$$\therefore \quad \boxed{r_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n\sigma_x \sigma_y}}$$

Alternate form of r(x, y):

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

Here n is the no. of pairs of values of x and y.

Note. Correlation co-efficient is independent of change of origin and scale.

Let us define two new variables u and v as

$$u = \frac{x - a}{h}, v = \frac{y - b}{k} \text{ where } a, b, h, k \text{ are constants, then } r_{xy} = r_{uv}.$$

Then,
$$\boxed{r(u, v) = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\,\sqrt{n\Sigma v^2 - (\Sigma v)^2}}}.$$

EXAMPLES

Example 1. *Find the coefficient of correlation between the values of x and y:*

[U.P.T.U. (C.O.) 2008]

x	1	3	5	7	8	10
y	8	12	15	17	18	20

Sol. Here, $n = 6$. The table is as follows:

x	y	x^2	y^2	xy
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
$\Sigma x = 34$	$\Sigma y = 90$	$\Sigma x^2 = 248$	$\Sigma y^2 = 1446$	$\Sigma xy = 582$

Karl Pearson's coefficient of correlation is given by

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

$$= \frac{(6 \times 582) - (34 \times 90)}{\sqrt{(6 \times 248) - (34)^2}\sqrt{(6 \times 1446) - (90)^2}} = 0.9879$$

Example 2. *The following data regarding the heights (y) and weights (x) of 100 college students are given:*

$\Sigma x = 15000$, $\Sigma x^2 = 2272500$, $\Sigma y = 6800$, $\Sigma y^2 = 463025$ and $\Sigma xy = 1022250$.

Find the correlation coefficient between height and weight.

Sol. Here, $n = 100$

Correlation co-efficient $r(x, y)$ is given by

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

$$= \frac{(100 \times 1022250) - (15000 \times 6800)}{\sqrt{(100 \times 2272500) - (15000)^2}\sqrt{(100 \times 463025) - (6800)^2}} = 0.6.$$

Example 3. *Find the co-efficient of correlation for the following table:* (U.K.T.U. 2011)

x:	*10*	*14*	*18*	*22*	*26*	*30*
y:	*18*	*12*	*24*	*6*	*30*	*36*

Sol. Let $u = \frac{x - 22}{4}$, $v = \frac{y - 24}{6}$

x	y	u	v	u^2	v^2	uv
10	18	-3	-1	9	1	3
14	12	-2	-2	4	4	4
18	24	-1	0	1	0	0
22	6	0	-3	0	9	0
26	30	1	1	1	1	1
30	36	2	2	4	4	4
Total		$\Sigma u = -3$	$\Sigma v = -3$	$\Sigma u^2 = 19$	$\Sigma v^2 = 19$	$\Sigma uv = 12$

Here, $n = 6$, $\bar{u} = \frac{1}{n}\Sigma u = \frac{1}{6}(-3) = -\frac{1}{2}$; $\bar{v} = \frac{1}{n}\Sigma v = \frac{1}{6}(-3) = -\frac{1}{2}$

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

$$= \frac{(6 \times 12) - (-3)(-3)}{\sqrt{(6 \times 19) - (-3)^2}\sqrt{(6 \times 19) - (-3)^2}} = \frac{63}{\sqrt{105}\sqrt{105}} = 0.6$$

Hence, $r_{xy} = r_{uv} = 0.6$.

Example 4. *Ten students got the following percentage of marks in Principles of Economics and Statistics:*

Roll Nos.:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Marks in Economics:	*78*	*36*	*98*	*25*	*75*	*82*	*90*	*62*	*65*	*39*
Marks in Statistics:	*84*	*51*	*91*	*60*	*68*	*62*	*86*	*58*	*53*	*47*

Calculate the co-efficient of correlation.

Sol. Let the marks in the two subjects be denoted by x and y respectively.

x	y	$u = x - 65$	$v = y - 66$	u^2	v^2	uv
78	84	13	18	169	324	234
36	51	-29	-15	841	225	435
98	91	33	25	1089	625	825
25	60	-40	-6	1600	36	240
75	68	10	2	100	4	20
82	62	17	-4	289	16	-68
90	86	25	20	625	400	500
62	58	-3	-8	9	64	24
65	53	0	-13	0	169	0
39	47	-26	-19	676	361	494
Total		$\Sigma u = 0$	$\Sigma v = 0$	$\Sigma u^2 = 5398$	$\Sigma v^2 = 2224$	$\Sigma uv = 2734$

Here, $n = 10$, $\bar{u} = \frac{1}{n}\Sigma u_i = 0$, $\bar{v} = \frac{1}{n}\Sigma v_i = 0$

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

$$= \frac{(10 \times 2734) - (0 \times 0)}{\sqrt{(10 \times 5398) - (0)^2}\sqrt{(10 \times 2224) - (0)^2}} = 0.789$$

Hence, $r_{xy} = r_{uv} = 0.789$.

Example 5. *A computer while calculating correlation co-efficient 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 co-efficient.

Sol. 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 $\Sigma XY = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520$
(Subtract the incorrect values and add the corresponding correct values)

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

Corrected $$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{(25 \times 520) - (125 \times 100)}{\sqrt{(25 \times 650) - (125)^2}\sqrt{(25 \times 436) - (100)^2}} = 0.67.$$

Example 6. *If $z = ax + by$ and r is the correlation co-efficient between x and y, show that*
$\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\,\sigma_x\,\sigma_y$.

Sol. $z = ax + by$

$\Rightarrow$ $\bar{z} = a\bar{x} + b\bar{y}$, $z_i = ax_i + by_i$

$z_i - \bar{z} = a(x_i - \bar{x}) + b(y_i - \bar{y})$

Now, $$\sigma_z^2 = \frac{1}{n}\Sigma(z_i - \bar{z})^2 = \frac{1}{n}\Sigma[a(x_i - \bar{x}) + b(y_i - \bar{y})]^2$$

$$= \frac{1}{n}\Sigma[a^2(x_i - \bar{x})^2 + b^2(y_i - \bar{y})^2 + 2ab(x_i - \bar{x})(y_i - \bar{y})]$$

$$= a^2 \cdot \frac{1}{n}\Sigma(x_i - \bar{x})^2 + b^2 \cdot \frac{1}{n}\Sigma(y_i - \bar{y})^2 + 2ab \cdot \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})$$

$$= a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\,\sigma_x\sigma_y \quad \left| \because\ r = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x\sigma_y} \right.$$

Example 7. *Establish the formula: $\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r_{xy}\sigma_x\sigma_y$*

where r_{xy} is the correlation coefficient between x and y. Using the above formula, calculate the correlation coefficient from the following data relating to the marks of 10 candidates in aptitude test (x) and Achievement rating (y).

	Marks									
Aptitude (x):	22	53	46	67	43	35	88	11	95	13
Achievement (y):	18	39	31	42	55	64	82	10	96	14

Sol. Let $z = x - y$

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

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

or, $(z - \bar{z})^2 = (x - \bar{x})^2 + (y - \bar{y})^2 - 2(x - \bar{x})(y - \bar{y})$

Summing up for n terms,

$$\Sigma(z - \bar{z})^2 = \Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2 - 2\Sigma(x - \bar{x})(y - \bar{y})$$

or, $$\frac{\Sigma(z - \bar{z})^2}{n} = \frac{\Sigma(x - \bar{x})^2}{n} + \frac{\Sigma(y - \bar{y})^2}{n} - \frac{2\Sigma(x - \bar{x})(y - \bar{y})}{n}$$

$\Rightarrow$ $$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y, \quad \text{where } r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y}$$

$\Rightarrow$ $$\boxed{\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y} \qquad ...(1)$$

Now, $n = 10$

$$\bar{x} = \frac{22 + 53 + 46 + 67 + 43 + 35 + 88 + 11 + 95 + 13}{10} = \frac{473}{10} = 47.3$$

$$\bar{y} = \frac{18 + 39 + 31 + 42 + 55 + 64 + 82 + 10 + 96 + 14}{10} = \frac{451}{10} = 45.1$$

Now we form the table as

x	y	$z = x - y$	$x - \bar{x}$	$y - \bar{y}$	$z - \bar{z}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(z - \bar{z})^2$
22	18	4	− 25.3	− 27.1	1.8	640.09	734.41	3.24
53	39	14	5.7	− 6.1	11.8	32.49	37.21	139.24
46	31	15	− 1.3	− 14.1	12.8	1.69	198.81	163.84
67	42	25	19.7	− 3.1	22.8	38.09	9.61	519.84
43	55	− 12	− 4.3	9.9	− 14.2	18.49	98.01	201.64
35	64	− 29	− 12.3	18.9	− 31.2	151.29	357.21	973.44
88	82	6	40.7	36.9	3.8	1656.49	1361.61	14.44
11	10	1	− 36.3	− 35.1	− 1.2	1317.69	1232.01	1.44
95	96	− 1	47.7	50.9	− 3.2	2275.29	2590.81	10.24
13	14	− 1	− 34.3	− 31.1	− 3.2	1176.49	967.21	10.24
						$\Sigma(x - \bar{x})^2$ = 7658.1	$\Sigma(y - \bar{y})^2$ = 7586.9	$\Sigma(z - \bar{z})^2$ = 2037.6

where, $$\bar{z} = \frac{\Sigma z}{n} = \frac{22}{10} = 2.2$$

Now, $$\sigma_x^2 = \frac{\Sigma(x - \bar{x})^2}{n} = \frac{7658.1}{10} = 765.81, \quad \sigma_y^2 = \frac{\Sigma(y - \bar{y})^2}{n} = \frac{7586.9}{10} = 758.69$$

$$\sigma_{x-y}^2 = \sigma_z^2 = \frac{\Sigma(z - \bar{z})^2}{n} = \frac{2037.6}{10} = 203.76.$$

Substituting the values in the formula (1),

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

$\Rightarrow$ $203.76 = 765.81 + 758.69 - 2r\,(27.67)(27.54)$

$$r = \frac{1524.5 - 203.76}{1524.06} = 0.866.$$

Example 8. *(i) Calculate coefficient of correlation from the following results:* $n = 10,\ \Sigma X = 100,\ \Sigma Y = 150,\ \Sigma(X-10)^2 = 180,\ \Sigma(Y-15)^2 = 215,\ \Sigma(X-10)(Y-15) = 60.$

(ii) Calculate Karl Pearson's coefficient of correlation between X and Y for the following information:

$n = 12,\ \Sigma X = 120,\ \Sigma Y = 130,\ \Sigma(X-8)^2 = 150,\ \Sigma(Y-10)^2 = 200$ *and* $\Sigma(X-8)(Y-10) = 50$

Sol. (*i*) Mean of first series, $\overline{X} = \frac{\Sigma X}{n} = \frac{100}{10} = 10$

Mean of second series, $\overline{Y} = \frac{\Sigma Y}{n} = \frac{150}{10} = 15$

Now, $$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2 \,.\, \Sigma(Y-\overline{Y})^2}}$$

$$= \frac{\Sigma(X-10)(Y-15)}{\sqrt{\Sigma(X-10)^2 \,.\, \Sigma(Y-15)^2}} = \frac{60}{\sqrt{180 \times 215}} = 0.305$$

(*ii*) $\Sigma x = \Sigma(X-8) = \Sigma X - \Sigma 8 = 120 - (8 \times 12) = 24$

$\Sigma y = \Sigma(Y-10) = \Sigma Y - \Sigma 10 = 130 - (10 \times 12) = 10$

$\Sigma xy = \Sigma(X-8)(Y-10) = 50$ (given)

$\Sigma x^2 = \Sigma(X-8)^2 = 150$

$\Sigma y^2 = \Sigma(Y-10)^2 = 200$

Now, $$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{(12 \times 50) - (24 \times 10)}{\sqrt{(12 \times 150) - (24)^2}\sqrt{(12 \times 200) - (10)^2}}$$

$$= \frac{360}{\sqrt{1224}\sqrt{2300}} = 0.2146.$$

3.35 CALCULATION OF CO-EFFICIENT OF CORRELATION FOR A BIVARIATE FREQUENCY DISTRIBUTION

If the bivariate data on x and y is presented on a two way correlation table and f is the frequency of a particular rectangle in the correlation table, then

$$r_{xy} = \frac{\Sigma fxy - \frac{1}{n}\Sigma fx\,\Sigma fy}{\sqrt{\left[\Sigma fx^2 - \frac{1}{n}(\Sigma fx)^2\right]\left[\Sigma fy^2 - \frac{1}{n}(\Sigma fy)^2\right]}}$$

Since change of origin and scale do not affect the co-efficient of correlation.

$\therefore$ $r_{xy} = r_{uv}$ where the new variables u, v are properly chosen.

Example 9. *The following table gives according to age the frequency of marks obtained by 100 students in an intelligence test:*

Age (in years) / Marks	18	19	20	21	Total
10–20	4	2	2		8
20–30	5	4	6	4	19
30–40	6	8	10	11	35
40–50	4	4	6	8	22
50–60		2	4	4	10
60–70		2	3	1	6
Total	19	22	31	28	100

Calculate the co-efficient of correlation between age and intelligence.

Sol. Let age and intelligence be denoted by x and y respectively.

Mid value	x / y	18	19	20	21	f	u	fu	fu^2	fuv
15	10–20	4	2	2		8	– 3	– 24	72	30
25	20–30	5	4	6	4	19	– 2	– 38	76	20
35	30–40	6	8	10	11	35	– 1	– 35	35	9
45	40–50	4	4	6	8	22	0	0	0	0
55	50–60		2	4	4	10	1	10	10	2
65	60–70		2	3	1	6	2	12	24	– 2
	f	19	22	31	28	100	Total	**– 75**	**217**	**59**
	v	– 2	– 1	0	1	Total				
	fv	– 38	– 22	0	28	**– 32**				
	fv^2	76	22	0	28	**126**				
	fuv	56	16	0	– 13	**59**				

Let us define two new variables u and v as $u = \frac{y-45}{10}$, $v = x - 20$

$$r_{xy} = r_{uv} = \frac{\Sigma fuv - \frac{1}{n}\Sigma fu\,\Sigma fv}{\sqrt{\left[\Sigma fu^2 - \frac{1}{n}(\Sigma fu)^2\right]\left[\Sigma fv^2 - \frac{1}{n}(\Sigma fv)^2\right]}}$$

$$= \frac{59 - \frac{1}{100}(-75)(-32)}{\sqrt{\left[217 - \frac{1}{100}(-75)^2\right]\left[126 - \frac{1}{100}(-32)^2\right]}} = \frac{59 - 24}{\sqrt{\frac{643}{4} \times \frac{2894}{25}}} = 0.25.$$

3.36 RANK CORRELATION

Sometimes we have to deal with problems in which data cannot be quantitatively measured but qualitative assessment is possible.

Let a group of n individuals be arranged in order of merit or proficiency in possession of two characteristics A and B. The ranks in the two characteristics are, in general, different. For example, if A stands for intelligence and B for beauty, it is not necessary that the most intelligent individual may be the most beautiful and *vica versa*. Thus an individual who is ranked at the top for the characteristic A *may be* ranked at the bottom for the characteristic B. Let (x_i, y_i), $i = 1, 2, \ldots\ldots n$ be the ranks of the n individuals in the group for the characteristics A and B respectively. Pearsonian co-efficient of correlation between the ranks x_i's and y_i's is called the *rank correlation co-efficient* between the characteristics A and B for that group of individuals.

Thus rank correlation co-efficient

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \qquad \ldots(1)$$

Now x_i's and y_i's are merely the permutations of n numbers from 1 to n. *Assuming that no two individuals are bracketed or tied in either classification i.e., $(x_i, y_i) \neq (x_j, y_j)$ for $i \neq j$, both x and y take all integral values from 1 to n.*

$$\therefore \quad \bar{x} = \bar{y} = \frac{1}{n}(1 + 2 + 3 + \ldots\ldots + n) = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$

$$\Sigma x_i = 1 + 2 + 3 + \ldots\ldots + n = \frac{n(n+1)}{2} = \Sigma y_i$$

$$\Sigma x_i^2 = 1^2 + 2^2 + \ldots\ldots + n^2 = \frac{n(n+1)(2n+1)}{6} = \Sigma y_i^2$$

If D_i denotes the difference in ranks of the i^{th} individual, then

$$D_i = x_i - y_i = (x_i - \bar{x}) - (y_i - \bar{y}) \qquad [\because \bar{x} = \bar{y}]$$

$$\frac{1}{n}\Sigma D_i^2 = \frac{1}{n}\Sigma[(x_i - \bar{x}) - (y_i - \bar{y})]^2$$

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

$$= \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y \qquad \ldots(2) \quad | \text{ using (1)}$$

But $$\sigma_x^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2 = \frac{1}{n}\Sigma y_i^2 - \bar{y}^2 = \sigma_y^2$$

$$\therefore \text{ From (2), } \frac{1}{n}\Sigma D_i^2 = 2\sigma_x^2 - 2r\sigma_x^2 = 2(1-r)\,\sigma_x^2 = 2(1-r)\left[\frac{1}{n}\Sigma x_i^2 - \bar{x}^2\right]$$

$$= 2(1-r)\left[\frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}\right]$$

$$= (1-r)(n+1)\left[\frac{4n+2-3n-3}{6}\right] = \frac{(1-r)(n^2-1)}{6} \quad \text{or} \quad 1 - r = \frac{6\Sigma D_i^2}{n(n^2-1)}$$

Hence, $$\boxed{r = 1 - \left[\frac{6\Sigma D_i^2}{n(n^2-1)}\right]}$$

Note. This is called *Spearman's Formula for* Rank Correlation.

$\Sigma d_i = \Sigma(x_i - y_i) = \Sigma x_i - \Sigma y_i = 0$ always.

This serves as a check on calculations.

EXAMPLES

Example 1. *Compute the rank correlation coefficient for the following data:*

Person:	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Rank in Maths:	*9*	*10*	*6*	*5*	*7*	*2*	*4*	*8*	*1*	*3*
Rank in Physics:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

Sol. Here the ranks are given and $n = 10$.

Person	R_1	R_2	$D = R_1 - R_2$	D^2
A	9	1	8	64
B	10	2	8	64
C	6	3	3	9
D	5	4	1	1
E	7	5	2	4
F	2	6	– 4	16
G	4	7	– 3	9
H	8	8	0	0
I	1	9	– 8	64
J	3	10	– 7	49
				$\Sigma D^2 = 280$

$$\therefore \quad r = 1 - \left\{\frac{6\Sigma D^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6 \times 280}{10(100-1)}\right\} = 1 - 1.697 = -0.697.$$

Example 2. *The marks secured by recruits in the selection test (X) and in the proficiency test (Y) are given below:*

Serial No.:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
X:	*10*	*15*	*12*	*17*	*13*	*16*	*24*	*14*	*22*
Y:	*30*	*42*	*45*	*46*	*33*	*34*	*40*	*35*	*39*

Calculate the rank correlation co-efficient.

Sol. Here the marks are given. Therefore, first of all, write down ranks. In each series, the item with the largest size is ranked 1, next largest 2 and so on. Here $n = 9$.

X	10	15	12	17	13	16	24	14	22	Total
Y	30	42	45	46	33	34	40	35	39	
Ranks in X (x)	9	5	8	3	7	4	1	6	2	
Ranks in Y (y)	9	3	2	1	8	7	4	6	5	
$D = x - y$	0	2	6	2	– 1	– 3	– 3	0	– 3	0
D^2	0	4	36	4	1	9	9	0	9	72

$$\therefore \quad r = 1 - \left\{\frac{6\Sigma D^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6\times 72}{9\times 80}\right\} = 1 - .6 = 0.4$$

Example 3. *Rank correlation co-efficient of marks obtained by 10 students in Mathematics and English was found to be 0.5. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct rank correlation co-efficient.*

Sol. Incorrect $r_k = 0.5$, $n = 10$

$$\therefore \text{ Incorrect } \Sigma D^2 = \frac{n(n^2-1)(1-r_k)}{6} = \frac{10\times 99\times 0.5}{6} = 82.5$$

Now, correct ΣD^2 = Incorrect $\Sigma D^2 - (3)^2 + (7)^2 = 82.5 - 9 + 49 = 122.5$

$$\therefore \quad \text{Correct } r_k = 1 - \left\{\frac{6\Sigma D^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6\times 122.5}{10(100-1)}\right\} = 0.2575.$$

Example 4. *Ten competitors in a beauty contest were ranked by three judges in the following orders:*

First Judge:	1	6	5	10	3	2	4	9	7	8
Second Judge:	3	5	8	4	7	10	2	1	6	9
Third Judge:	6	4	9	8	1	2	3	10	5	7

Use the method of rank correlation to determine which pair of judges has the nearest approach to common taste in beauty?

Sol. Let R_1, R_2, R_3 be the ranks given by three judges.

Calculation of rank correlation coefficient

Competitor	R_1	R_2	R_3	$D_{12} = R_1 - R_2$	$D_{13} = R_1 - R_3$	$D_{23} = R_2 - R_3$	D_{12}^2	D_{13}^2	D_{23}^2
A	1	3	6	−2	−5	−3	4	25	9
B	6	5	4	1	2	1	1	4	1
C	5	8	9	−3	−4	−1	9	16	1
D	10	4	8	6	2	−4	36	4	16
E	3	7	1	−4	2	6	16	4	36
F	2	10	2	−8	0	8	64	0	64
G	4	2	3	2	1	−1	4	1	1
H	9	1	10	8	−1	−9	64	1	81
I	7	6	5	1	2	1	1	4	1
J	8	9	7	−1	1	2	1	1	4
$n = 10$							$\Sigma D_{12}^2 = 200$	$\Sigma D_{13}^2 = 60$	$\Sigma D_{23}^2 = 214$

Rank correlation coefficient between first and second judges,

$$r_{k_{12}} = 1 - \left\{\frac{6\Sigma D_{12}^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6\times 200}{10(99)}\right\} = -0.212$$

Rank correlation coefficient between first and third judges,

$$r_{k_{13}} = 1 - \left\{\frac{6\Sigma D_{13}^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6\times 60}{10\times 99}\right\} = 0.636$$

Rank correlation coefficient between second and third judges,

$$r_{k_{23}} = 1 - \left\{\frac{6\Sigma D_{23}^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6\times 214}{10\times 99}\right\} = -0.297$$

Correlation between first and second judges is negative *i.e.*, their opinions regarding beauty test are opposite to each other. Similarly, opinions of second and third judges are opposite to each other, but the opinions of first and third judges are of similar type as their correlation is positive. It means that their likings and dislikings are very much common.

3.37 TIED RANKS

If any two or more individuals have same rank or the same value in the series of marks, then the above formula fails and requires an adjustment. In such cases, each individual is given an average rank. This common average rank is the average of the ranks which these individuals would have assumed if they were slightly different from each other. Thus, if two individual are ranked equal at the sixth place, they would have assumed the 6th and 7th ranks if they were ranked slightly different. Their common rank $= \frac{6+7}{2} = 6.5$. If three individuals are ranked equal at fourth place, they would have assumed the 4th, 5th and 6th ranks if they were ranked slightly different. Their common rank $= \frac{4+5+6}{3} = 5$.

Adjustment. Add $\frac{1}{12}m(m^2-1)$ to ΣD^2 where m stands for the number of times an item is repeated.

This adjustment factor is to be added for each repeated item.

Thus
$$r = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}m(m^2-1) + \frac{1}{12}m(m^2-1) + \ldots\right\}}{n(n^2-1)}.$$

Example 5. *Obtain the rank correlation co-efficient 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.

Sol. Here, marks are given, so write down the ranks.

X	68	64	75	50	64	80	75	40	55	64	Total
Y	62	58	68	45	81	60	68	48	50	70	
Ranks in X (x)	4	6	2.5	9	6	1	2.5	10	8	6	
Ranks in Y (y)	5	7	3.5	10	1	6	3.5	9	8	2	
$D = x - y$	−1	−1	−1	−1	5	−5	−1	1	0	4	0
D^2	1	1	1	1	25	25	1	1	0	16	72

In the X-series, the value 75 occurs twice. Had these values been slightly different, they would have been given the ranks 2 and 3. Therefore, the common rank given to them is $\frac{2+3}{2}$ = 2.5. The value 64 occurs thrice. Had these values been slightly different, they would have been given the ranks 5, 6, and 7. Therefore the common rank given to them is $\frac{5+6+7}{3} = 6$.

Similarly, in the Y-series, the value 68 occurs twice. Had these values been slightly different, they would have been given the ranks 3 and 4? Therefore, the common rank given to them is $\frac{3+4}{2} = 3.5$.

Thus, m has the values 2, 3, 2.

$$\therefore \quad r = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}m_1(m_1^2-1) + \frac{1}{12}m_2(m_2^2-1) + \frac{1}{12}m_3(m_3^2-1)\right\}}{n(n^2-1)}$$

$$= 1 - \frac{6\left[72 + \frac{1}{12}.2(2^2-1) + \frac{1}{12}.3(3^2-1) + \frac{1}{12}.2(2^2-1)\right]}{10(10^2-1)}$$

$$= 1 - \left\{\frac{6\times 75}{990}\right\} = \frac{6}{11} = 0.545.$$

ASSIGNMENT

1. Calculate the coefficient of correlation for the following data: *(U.P.T.U. 2006)*

Husband's age (in yrs.) x	23	27	28	28	29	30	31	33	35	36
Wife's age (in yrs.) y	18	20	22	27	21	29	27	29	28	29

2. Calculate the coefficient of correlation for the following data:

Height of father (in inches)	65	66	67	67	68	69	70	72
Height of son (in inches)	67	68	65	68	72	72	69	71

3. Define Karl Pearson's coefficient of correlation. How would you interpret the sign and magnitude of a correlation coefficient?

4. Calculate the coefficient of correlation between the marks obtained by 8 students in Mathematics and Statistics:

Students	A	B	C	D	E	F	G	H
Mathematics	25	30	32	35	37	40	42	45
Statistics	08	10	15	17	20	23	24	25

[U.P.T.U. (C.O.) 2009]

5. Find the correlation coefficient between x and y for the following data:

x	60	34	40	50	45	41	22	43
y	75	32	34	40	45	33	12	30

[U.K.T.U. 2010]

6. The marks of the same 15 students in two subjects A and B are analysed. The two numbers within the brackets denote the ranks of the same student in A and B respectively:
(1, 10) (2, 7) (3, 2) (4, 6) (5, 4) (6, 8) (7, 3) (8, 1) (9, 11) (10, 15) (11, 9) (12, 5) (13, 14) (14, 12) (15, 13)
Use Spearman's formula to find the rank correlation coefficient.

7. Ten students got the following percentage of marks in Chemistry and Physics:

Students:	1	2	3	4	5	6	7	8	9	10
Marks in Chemistry:	78	36	98	25	75	82	90	62	65	39
Marks in Physics:	84	51	91	60	68	62	86	58	63	47

Calculate the rank correlation co-efficient.

8. A firm not sure of the response to its product in ten different colour shades decides to produce them in those colour shades, if the ranking of these colour shades by two typical consumer judges is highly correlated.
The two judges rank the ten colours in the following order:

Colour :	*Red*	*Green*	*Blue*	*Yellow*	*White*	*Black*	*Pink*	*Purple*	*Orange*	*Ivory*
Ranking by I Judge :	6	4	3	1	2	7	9	8	10	5
Ranking by II Judge :	4	1	6	7	5	8	10	9	3	2

Is there any agreement between the two judges, to allow the introduction of the product by the firm in the market?

9. Two judges in a music competition rank the 12 entries as follows:

x	1	2	3	4	5	6	7	8	9	10	11	12
y	12	9	6	10	3	5	4	7	8	2	11	1

What degree of agreement is there between the judgement of the two judges?

10. Calculate the coefficient of correlation between the following ages of husband (x) and wife (y) by taking 30 and 28 as assumed mean incase of x and y respectively:

x :	24	27	28	28	29	30	32	33	35	35	40
y :	18	20	22	25	22	28	28	30	27	30	32

Answers

1. 0.82 **2.** 0.603 **4.** 0.9804 **5.** 0.9158
6. 0.51 **7.** 0.84 **8.** 0.22, no **9.** – 0.454
10. 0.8926.

3.38 REGRESSION ANALYSIS

(U.P.T.U. 2015)

The term 'regression' was first used by Sir Francis Galton (1822–1911), a British Biometrician in connection with the height of parents and their offsprings. He found that the offspring of tall or short parents tend to regress to the average height. In other words, though tall fathers do tend to have tall sons yet the average height of tall fathers is more than the average height of their sons and the average height of short fathers is less than the average height of their sons.

The term 'regression' stands for some sort of functional relationship between two or more related variables. The only fundamental difference, if any, between problems of curve-fitting and regression is that in regression, any of the variables may be considered as independent or dependent while in curve-fitting, one variable cannot be dependent.

Regression measures the nature and extent of correlation. Regression is the estimation or prediction of unknown values of one variable from known values of another variable.

3.39 CURVE OF REGRESSION AND REGRESSION EQUATION

If two variates x and y are correlated *i.e.*, there exists an association or relationship between them, then the scatter diagram will be more or less concentrated round a curve. This curve is called the *curve of regression* and the relationship is said to be expressed by means of *curvilinear regression.*

The mathematical equation of the regression curve is called regression equation.

3.40 LINEAR REGRESSION

When the points of the scatter diagram concentrate round a straight line, the regression is called linear and this straight line is known as the line of regression.

Regression will be called non-linear if there exists a relationship other than a straight line between the variables under consideration.

3.41 LINES OF REGRESSION

(U.P.T.U. 2006, 2007)

A line of regression is the straight line which gives the best fit in the least square sense to the given frequency.

In case of n pairs (x_i, y_i) ; $i = 1, 2, ..., n$ from a bivariate data, we have no reason or justification to assume y as dependent variable and x as independent variable. Either of the two may be estimated for the given values of the other. Thus if we wish to estimate y for given values of x, we shall have the regression equation of the form $y = a + bx$, called the regression line of y on x. If we wish to estimate x for given values of y, we shall have the regression line of the form $x = A + By$, called the regression line of x on y.

Thus it implies, in general, *we always have two lines of regression.*

If the line of regression is so chosen that the sum of squares of deviation parallel to the axis of y is minimised [*See Fig. (a)*], it is called *the line of regression of y on x* and it gives *the best estimate of y for any given value of x.*

If the line of regression is so chosen that the sum of squares of deviations parallel to the axis of x is minimised [*See Fig. (b)*], it is called *the line of regression of x on y* and it gives *the best estimate of x for any given value of y.*

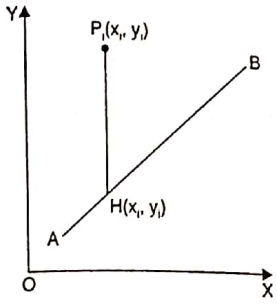

Fig. (a)

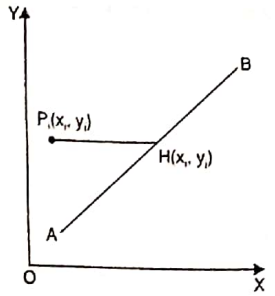

Fig. (b)

The independent variable is called *predictor* or regresser or explanator and the dependent variable is called the *predictant* or regressed or explained variable.

3.42 DERIVATION OF LINES OF REGRESSION

3.42.1. Line of Regression of y on x

To obtain the line of regression of y on x, we shall assume y as dependent variable and x as independent variable. Let $y = a + bx$ be the equation of regression line of y on x.

The residual for i^{th} point is $E_i = y_i - a - bx_i$.

Introduce a new quantity U such that

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad ...(1)$$

According to the *principle of Least squares*, the constants a and b are chosen in such a way that the sum of the squares of residuals is minimum.

Now, the condition for U to be maximum or minimum is

$$\frac{\partial U}{\partial a} = 0 \quad \text{and} \quad \frac{\partial U}{\partial b} = 0$$

From (1), $$\frac{\partial U}{\partial a} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1)$$

$$\frac{\partial U}{\partial a} = 0 \text{ gives } 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0$$

$\Rightarrow$ $$\boxed{\Sigma y = na + b\,\Sigma x} \qquad ..(2)$$

Also, $$\frac{\partial U}{\partial b} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i)$$

$$\frac{\partial U}{\partial b} = 0 \text{ gives } 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0$$

$\Rightarrow$ $$\boxed{\Sigma xy = a\,\Sigma x + b\,\Sigma x^2} \qquad ...(3)$$

Equations (2) and (3) are called *normal equations.*

Solving (2) and (3) for 'a' and 'b', we get

$$b = \frac{\Sigma xy - \frac{1}{n}\Sigma x\,\Sigma y}{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} \quad \ldots(4)$$

and

$$a = \frac{\Sigma y}{n} - b\frac{\Sigma x}{n} = \bar{y} - b\bar{x} \quad \ldots(5)$$

Eqn. (5) gives $\bar{y} = a + b\bar{x}$

Hence $y = a + bx$ line passes through point $(\bar{x}, \bar{y})$.

Putting $a = \bar{y} - b\bar{x}$ in equation of line $y = a + bx$, we get

$$\boxed{y - \bar{y} = b(x - \bar{x})} \quad \ldots(6)$$

Equation (6) is called regression line of y on x. 'b' is called the regression coefficient of y on x and is usually denoted by b_{yx}.

Hence eqn. (6) can be rewritten as

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

where $\bar{x}$ and $\bar{y}$ are mean values while

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2}$$

In equation (3), shifting the origin to $(\bar{x}, \bar{y})$, we get

$$\Sigma(x - \bar{x})(y - \bar{y}) = a\,\Sigma(x - \bar{x}) + b\,\Sigma(x - \bar{x})^2$$

$$\Rightarrow \quad nr\,\sigma_x\sigma_y = a(0) + bn\sigma_x^2$$

$$\Rightarrow \quad b = r\frac{\sigma_y}{\sigma_x}$$

$\because\ \Sigma(x - \bar{x}) = 0,$

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

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

Hence, regression coefficient b_{yx} can also be defined as

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

where r is the coefficient of correlation, σ_x and σ_y are the standard deviations of x and y series respectively.

3.42.2. Line of Regression of x on y

Proceeding in the same way as 3.42.1, we can derive the regression line of x on y as

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

where b_{xy} is the regression coefficient of x on y and is given by

$$\boxed{b_{xy} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma y^2 - (\Sigma y)^2}}$$

or

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

where the terms have their usual meanings.

Note. If $r = 0$, the two lines of regression become $y = \bar{y}$ and $x = \bar{x}$ which are two straight lines parallel to x and y axes respectively and passing through their means $\bar{y}$ and $\bar{x}$. They are mutually perpendicular. If $r = \pm 1$, the two lines of regression will coincide.

3.43 USE OF REGRESSION ANALYSIS

(U.P.T.U. 2008)

(*i*) In the field of Business, this tool of statistical analysis is widely used. Businessmen are interested in predicting future production, consumption, investment, prices, profits and sales etc.

(*ii*) In the field of economic planning and sociological studies, projections of population, birth rates, death rates and other similar variables are of great use.

3.44 COMPARISON OF CORRELATION AND REGRESSION ANALYSIS

[G.B.T.U. M.B.A. (C.O.) 2011]

Both the correlation and regression analysis helps us in studying the relationship between two variables yet they differ in their approach and objectives.

(*i*) Correlation studies are meant for studying the covariation of the two variables. They tell us whether the variables under study move in the same direction or in reverse directions. The degree of their covariation is also reflected in the correlation co-efficient but the correlation study does not provide the nature of relationship. It does not tell us about the relative movement in the variables and we cannot predict the value of one variable corresponding to the value of other variable. This is possible through regression analysis.

(*ii*) Regression presumes one variable as a cause and the other as its effect. The independent variable is supposed to be affecting the dependent variable and as such we can estimate the values of the dependent variable by projecting the relationship between them. However, correlation between two series is not necessarily a cause-effect relationship.

(*iii*) Coefficient of correlation cannot exceed unity but one of the regression coefficients can have a value higher than unity but the product of two regression coefficients can never exceed unity.

3.45 PROPERTIES OF REGRESSION CO-EFFICIENTS

Property I. *Correlation co-efficient is the geometric mean between the regression co-efficients.*

Proof. The co-efficients of regression are $\frac{r\sigma_y}{\sigma_x}$ and $\frac{r\sigma_x}{\sigma_y}$.

$$\text{G.M. between them} = \sqrt{\frac{r\sigma_y}{\sigma_x} \times \frac{r\sigma_x}{\sigma_y}} = \sqrt{r^2} = r = \text{co-efficient of correlation.}$$

Property II. *If one of the regression co-efficients is greater than unity, the other must be less than unity.*

Proof. The two regression co-efficients are $b_{yx} = \frac{r\sigma_y}{\sigma_x}$ and $b_{xy} = \frac{r\sigma_x}{\sigma_y}$.

Let $b_{yx} > 1$, then $\frac{1}{b_{yx}} < 1$ ...(1)

Since $b_{yx} \cdot b_{xy} = r^2 \le 1$ $(\because\ -1 \le r \le 1)$

$\therefore$ $b_{xy} \le \frac{1}{b_{yx}} < 1.$ | Using (1)

Similarly, if $b_{xy} > 1$, then $b_{yx} < 1$.

Property III. *Arithmetic mean of regression co-efficients is greater than the correlation co-efficient.*

Proof. We have to prove that

$$\frac{b_{yx}+b_{xy}}{2} > r$$

or $$r\frac{\sigma_y}{\sigma_x}+r\frac{\sigma_x}{\sigma_y} > 2r$$

or $$\sigma_x^2+\sigma_y^2 > 2\sigma_x\sigma_y$$

or $$(\sigma_x-\sigma_y)^2 > 0, \quad \text{which is true.}$$

Property IV. *Regression co-efficients are independent of the origin but not of scale.*

Proof. Let $u=\frac{x-a}{h}, v=\frac{y-b}{k}$, where a, b, h and k are constants

$$b_{yx}=\frac{r\sigma_y}{\sigma_x}=r\cdot\frac{k\sigma_v}{h\sigma_u}=\frac{k}{h}\left(\frac{r\sigma_v}{\sigma_u}\right)=\frac{k}{h}b_{vu}$$

Similarly, $$b_{xy}=\frac{h}{k}b_{uv}.$$

Thus, b_{yx} and b_{xy} are both independent of a and b but not of h and k.

Property V. *The correlation co-efficient and the two regression co-efficients have same sign.*

Proof. Regression co-efficient of y on $x = b_{yx} = r\frac{\sigma_y}{\sigma_x}$

Regression co-efficient of x on $y = b_{xy} = r\frac{\sigma_x}{\sigma_y}$

Since σ_x and σ_y are both positive; b_{yx}, b_{xy} and r have same sign.

3.46 ANGLE BETWEEN TWO LINES OF REGRESSION

If θ is the acute angle between the two regression lines in the case of two variables x and y, show that

$$\tan\theta=\frac{1-r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}, \quad \textit{where } r, \sigma_x, \sigma_y \textit{ have their usual meanings.}$$

Explain the significance of the formula when $r = 0$ and $r = \pm 1$.

[U.P.T.U. 2007, 2015; G.B.T.U. (C.O.) 2011]

Proof. Equations to the lines of regression of y on x and x on y are

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

Their slopes are $$m_1=\frac{r\sigma_y}{\sigma_x} \quad \text{and} \quad m_2=\frac{\sigma_y}{r\sigma_x}.$$

$$\therefore \quad \tan\theta=\pm\frac{m_2-m_1}{1+m_2m_1}=\pm\frac{\frac{\sigma_y}{r\sigma_x}-\frac{r\sigma_y}{\sigma_x}}{1+\frac{\sigma_y^2}{\sigma_x^2}}$$

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

Since $r^2 \le 1$ and σ_x, σ_y are positive.

$\therefore$ +ve sign gives the acute angle between the lines.

Hence $$\boxed{\tan\theta=\frac{1-r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}}$$

when $r = 0$, $\theta=\frac{\pi}{2}$ $\therefore$ The two lines of regression are perpendicular to each other.

Hence the estimated value of y is the same for all values of x and vice-versa.

When $r = \pm 1$, $\tan\theta = 0$ so that $\theta = 0$ or π

Hence the lines of regression coincide and there is perfect correlation between the two variates x and y.

EXAMPLES

Example 1. *If the regression coefficients are 0.8 and 0.2, what would be the value of coefficient of correlation?*

Sol. We know that,

$$r^2 = b_{yx}\cdot b_{xy} = 0.8\times 0.2 = 0.16$$

Since r has the same sign as both the regression coefficients b_{yx} and b_{xy}

Hence $$r=\sqrt{0.16}=0.4.$$

Example 2. *Calculate linear regression coefficients from the following:*

x	→	1	2	3	4	5	6	7	8
y	→	3	7	10	12	14	17	20	24

Sol. Linear regression coefficients are given by

$$b_{yx}=\frac{n\Sigma xy-\Sigma x\Sigma y}{n\Sigma x^2-(\Sigma x)^2} \quad \text{and} \quad b_{xy}=\frac{n\Sigma xy-\Sigma x\Sigma y}{n\Sigma y^2-(\Sigma y)^2}$$

Let us prepare the following table:

x	y	x^2	y^2	xy
1	3	1	9	3
2	7	4	49	14
3	10	9	100	30
4	12	16	144	48
5	14	25	196	70
6	17	36	289	102
7	20	49	400	140
8	24	64	576	192
$\Sigma x = 36$	$\Sigma y = 107$	$\Sigma x^2 = 204$	$\Sigma y^2 = 1763$	$\Sigma xy = 599$

Here, $n = 8$

$$\therefore \quad b_{yx} = \frac{(8 \times 599) - (36 \times 107)}{(8 \times 204) - (36)^2} = \frac{940}{336} = 2.7976$$

and

$$b_{xy} = \frac{(8 \times 599) - (36 \times 107)}{(8 \times 1763) - (107)^2} = \frac{940}{2655} = 0.3540$$

Example 3. *The following table gives age (x) in years of cars and annual maintenance cost (y) in hundred rupees:*

X:	*1*	*3*	*5*	*7*	*9*
Y:	*15*	*18*	*21*	*23*	*22*

Estimate the maintenance cost for a 4 year old car after finding the regression equation.

Sol.

x	y	xy	x^2
1	15	15	1
3	18	54	9
5	21	105	25
7	23	161	49
9	22	198	81
$\Sigma x = 25$	$\Sigma y = 99$	$\Sigma xy = 533$	$\Sigma x^2 = 165$

Here, $n = 5$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{25}{5} = 5, \ \bar{y} = \frac{\Sigma y}{n} = \frac{99}{5} = 19.8$$

$$\therefore \quad b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} = \frac{(5 \times 533) - (25 \times 99)}{(5 \times 165) - (25)^2} = 0.95$$

Regression line of y on x is given by

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

$$\Rightarrow \quad y - 19.8 = 0.95(x - 5)$$

$$\Rightarrow \quad y = 0.95x + 15.05$$

When $x = 4$ years, $y = (0.95 \times 4) + 15.05 = 18.85$ hundred rupees = ₹ 1885.

Example 4. *In a partially destroyed laboratory record of an analysis of a correlation data, the following results only are legible:*

Variance of $x = 9$

Regression equations: $8x - 10y + 66 = 0$, $40x - 18y = 214$.

What were (a) the mean values of x and y (b) the standard deviation of y and the co-efficient of correlation between x and y? [U.P.T.U. 2008, 2009; U.K.T.U. 2010]

Sol. (*a*) **Since both the lines of regression pass through the point** $(\bar{x}, \bar{y})$ therefore, we have

$$8\bar{x} - 10\bar{y} + 66 = 0 \qquad \ldots(1)$$

$$40\bar{x} - 18\bar{y} - 214 = 0 \qquad \ldots(2)$$

Multiplying (1) by 5, $40\bar{x} - 50\bar{y} + 330 = 0$ ...(3)

Subtracting (3) from (2), $32\bar{y} - 544 = 0$ $\therefore \quad \bar{y} = 17$

$\therefore$ From (1), $8\bar{x} - 170 + 66 = 0$ or $8\bar{x} = 104 \ \therefore \ \bar{x} = 13$

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

(*b*) Variance of $x = \sigma_x^2 = 9$ (given)

$\therefore \quad \sigma_x = 3$

The equations of lines of regression can be written as

$$y = 0.8x + 6.6 \quad \text{and} \quad x = 0.45y + 5.35$$

$\therefore$ The regression co-efficient of y on x is $\dfrac{r\sigma_y}{\sigma_x} = 0.8$ ...(4)

The regression co-efficient of x on y is $\dfrac{r\sigma_x}{\sigma_y} = 0.45$ ...(5)

Multiplying (4) and (5), $r^2 = 0.8 \times 0.45 = 0.36 \quad \therefore \ r = 0.6$

(+ve sign with square root is taken because regression co-efficients are +ve).

From (4), $\sigma_y = \dfrac{0.8\sigma_x}{r} = \dfrac{0.8 \times 3}{0.6} = 4.$

Example 5. *The regression lines of y on x and x on y are respectively $y = ax + b$, $x = cy + d$. Show that*

$$\frac{\sigma_y}{\sigma_x} = \sqrt{\frac{a}{c}}, \ \bar{x} = \frac{bc + d}{1 - ac} \ \text{and} \ \bar{y} = \frac{ad + b}{1 - ac}.$$

[U.P.T.U. (C.O.) 2009, U.P.T.U. (MCA) 2008]

Sol. The regression line of y on x is

$$y = ax + b \qquad \ldots(1)$$

$$\therefore \quad b_{yx} = a$$

The regression line of x on y is

$$x = cy + d \qquad \ldots(2)$$

$$\therefore \quad b_{xy} = c$$

We know that, $b_{yx} = r\dfrac{\sigma_y}{\sigma_x}$ ...(3)

and $b_{xy} = r\dfrac{\sigma_x}{\sigma_y}$ ...(4)

Dividing eqn. (3) by (4), we get

$$\frac{b_{yx}}{b_{xy}} = \frac{\sigma_y^2}{\sigma_x^2} \Rightarrow \frac{a}{c} = \frac{\sigma_y^2}{\sigma_x^2} \Rightarrow \frac{\sigma_y}{\sigma_x} = \sqrt{\frac{a}{c}}$$

Since both the regression lines pass through the point $(\bar{x}, \bar{y})$ therefore,

$$\bar{y} = a\bar{x} + b \quad \text{and} \quad \bar{x} = c\bar{y} + d$$

$$\Rightarrow \quad a\bar{x} - \bar{y} = -b \qquad \ldots(5)$$

$$\bar{x} - c\bar{y} = d \qquad \ldots(6)$$

Multiplying equation (6) by a and then subtracting from (5), we get

$$(ac - 1)\,\bar{y} = -ad - b \quad \Rightarrow \quad \bar{y} = \frac{ad + b}{1 - ac}$$

Similarly, we get $\bar{x} = \dfrac{bc + d}{1 - ac}$.

Example 6. *For two random variables, x and y with the same mean, the two regression equations are* $y = ax + b$ *and* $x = \alpha y + \beta$. *Show that* $\dfrac{b}{\beta} = \dfrac{1 - a}{1 - \alpha}$. *Find also the common mean.* (G.B.T.U. 2010)

Sol. Here, $b_{yx} = a,\ b_{xy} = \alpha$

Let the common mean be m, then regression lines are

$$y - m = a\,(x - m)$$

$$\Rightarrow \quad y = ax + m\,(1 - a) \qquad \ldots(1)$$

and

$$x - m = \alpha(y - m)$$

$$\Rightarrow \quad x = \alpha y + m\,(1 - \alpha) \qquad \ldots(2)$$

Comparing (1) and (2) with the given equations.

$$b = m\,(1 - a),\ \beta = m\,(1 - \alpha)$$

$$\therefore \quad \frac{b}{\beta} = \frac{1 - a}{1 - \alpha}$$

Since regression lines pass through $(\bar{x}, \bar{y})$

$$\therefore \quad \bar{x} = \alpha\bar{y} + \beta \quad \text{and} \quad \bar{y} = a\bar{x} + b \text{ will hold.}$$

$$\Rightarrow \quad m = am + b,\ m = \alpha m + \beta$$

$$\Rightarrow \quad am + b = \alpha m + \beta$$

$$\Rightarrow \quad m = \frac{\beta - b}{a - \alpha}.$$

Example 7. *(i) Obtain the line of regression of y on x for the data given below:*

x:	*1.53*	*1.78*	*2.60*	*2.95*	*3.42*
y:	*33.50*	*36.30*	*40.00*	*45.80*	*53.50.*

(ii) The following data regarding the heights (y) and weights (x) of 100 college students are given:

$\Sigma x = 15000,\quad \Sigma x^2 = 2272500,\quad \Sigma y = 6800,\quad \Sigma y^2 = 463025$ *and* $\Sigma xy = 1022250.$

Find the equation of regression line of height on weight.

Sol. *(i)* The line of regression of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x}) \qquad \ldots(1)$$

where b_{yx} is the coefficient of regression given by

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} \qquad \ldots(2)$$

Now we form the table as,

x	y	x^2	xy
1.53	33.50	2.3409	51.255
1.78	36.30	2.1684	64.614
2.60	40.00	6.76	104
2.95	45.80	8.7025	135.11
3.42	53.50	11.6964	182.97
$\Sigma x = 12.28$	$\Sigma y = 209.1$	$\Sigma x^2 = 32.6682$	$\Sigma xy = 537.949$

Here, $n = 5$

$$b_{yx} = \frac{(5 \times 537.949) - (12.28 \times 209.1)}{(5 \times 32.6682) - (12.28)^2} = 9.726$$

Also, mean $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{12.28}{5} = 2.456$ and $\bar{y} = \dfrac{\Sigma y}{n} = \dfrac{209.1}{5} = 41.82$

$\therefore$ From (1), we get

$$y - 41.82 = 9.726(x - 2.456) = 9.726x - 23.887$$

$$y = 17.932 + 9.726x$$

(ii) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{15000}{100} = 150,\quad \bar{y} = \dfrac{\Sigma y}{n} = \dfrac{6800}{100} = 68$

Regression coefficient of y on x,

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} = \frac{(100 \times 1022250) - (15000 \times 6800)}{(100 \times 2272500) - (15000)^2} = 0.1$$

Regression line of height (y) on weight (x) is given by

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

$$\Rightarrow \quad y - 68 = 0.1(x - 150)$$

$$\Rightarrow \quad y = 0.1x + 53.$$

Example 8. *For 10 observations on price (x) and supply (y), the following data were obtained (in appropriate units):*

$\Sigma x = 130,\quad \Sigma y = 220,\quad \Sigma x^2 = 2288,\quad \Sigma y^2 = 5506$ *and* $\Sigma xy = 3467$

Obtain the two lines of regression and estimate the supply when the price is 16 units.

Sol. Here, $n = 10,\ \bar{x} = \dfrac{\Sigma x}{n} = 13$ and $\bar{y} = \dfrac{\Sigma y}{n} = 22$

Regression coefficient of y on x is

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} = \frac{(10 \times 3467) - (130 \times 220)}{(10 \times 2288) - (130)^2} = 1.015$$

$\therefore$ Regression line of y on x is

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

$$y - 22 = 1.015(x - 13)$$

$$y = 1.015x + 8.805 \quad \ldots(1)$$

$\Rightarrow$ Regression coefficient of x on y is

$$b_{xy} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma y^2 - (\Sigma y)^2} = \frac{(10 \times 3467) - (130 \times 220)}{(10 \times 5506) - (220)^2} = 0.9114$$

Regression line of x on y is

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

$$x - 13 = 0.9114(y - 22)$$

$$x = 0.9114y - 7.0508 \quad \ldots(2)$$

Since we are to estimate supply (y) when price (x) is given therefore we are to use regression line of y on x here.

When $x = 16$ units,

$$y = 1.015(16) + 8.805 = 25.045 \text{ units.}$$

Example 9. *The following results were obtained from records of age (x) and systolic blood pressure (y) of a group of 10 men:*

	x	y
Mean	53	142
Variance	130	

and $\Sigma(x - \bar{x})(y - \bar{y}) = 1220$

Find the appropriate regression equation and use it to estimate the blood pressure of a man whose age is 45.

Sol. Given

Mean $\bar{x} = 53$ and $\bar{y} = 142$; Variance $\sigma_x^2 = 130$

$n = 10$ and $\Sigma(x - \bar{x})(y - \bar{y}) = 1220$

Since we are to estimate blood pressure (y) of a 45 years old man, we will find regression line of y on x.

Regression coefficient

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n\,\sigma_x\,\sigma_y}\left(\frac{\sigma_y}{\sigma_x}\right)$$

$$= \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n\,\sigma_x^2} = \frac{1220}{(10)(130)} = 0.93846$$

Regression line of y on x is given by

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

$\Rightarrow$ $$y - 142 = 0.93846(x - 53)$$

$\Rightarrow$ $$y = 0.93846x + 92.26162$$

When $x = 45$, $\therefore$ $y = 134.49$

Hence the required blood pressure = 134.49.

Example 10. *The following results were obtained from marks in Applied Mechanics and Engineering Mathematics in an examination:*

	Applied Mechanics (x)	*Engineering Mathematics (y)*
Mean	47.5	39.5
Standard Deviation	16.8	10.8

$r = 0.95$.

Find both the regression equations. Also estimate the value of y for $x = 30$.

Sol. $\bar{x} = 47.5$, $\bar{y} = 39.5$

$\sigma_x = 16.8$, $\sigma_y = 10.8$ and $r = 0.95$.

Regression coefficients are

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.95 \times \frac{10.8}{16.8} = 0.6107$$

and

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.95 \times \frac{16.8}{10.8} = 1.477.$$

Regression line of y on x is

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

$\Rightarrow$ $$y - 39.5 = 0.6107(x - 47.5) = 0.6107x - 29.008$$

$$y = 0.6107x + 10.49 \quad \ldots(1)$$

Regression line of x on y is

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

$\Rightarrow$ $$x - 47.5 = 1.477(y - 39.5)$$

$$x = 1.477y - 10.8415 \quad \ldots(2)$$

Putting $x = 30$ in equation (1), we get

$$y = (0.6107)(30) + 10.49 = 28.81.$$

Example 11. *The equations of two regression lines, obtained in a correlation analysis of 60 observations are:*

$$5x = 6y + 24 \text{ and } 1000y = 768x - 3608.$$

What is the correlation coefficient? Show that the ratio of coefficient of variability of x to that of y is $\frac{5}{24}$. What is the ratio of variances of x and y?

Sol. Regression line of x on y is

$$5x = 6y + 24$$

$\Rightarrow$ $$x = \frac{6}{5}y + \frac{24}{5}$$

$\therefore$ $$b_{xy} = \frac{6}{5} \quad \ldots(1)$$

Regression line of y on x is

$$1000y = 768x - 3608$$

$\Rightarrow$ $$y = 0.768x - 3.608$$

$\therefore$ $$b_{yx} = 0.768 \quad \ldots(2)$$

From (1), $r\frac{\sigma_x}{\sigma_y} = \frac{6}{5}$ ...(3)

From (2), $r\frac{\sigma_y}{\sigma_x} = 0.768$ ...(4)

Multiplying equations (3) and (4), we get

$r^2 = 0.9216 \Rightarrow r = 0.96$ ...(5)

Dividing (4) by (3), we get

$$\frac{\sigma_x^2}{\sigma_y^2} = \frac{6}{5 \times 0.768} = 1.5625.$$

Taking square root, we get

$$\frac{\sigma_x}{\sigma_y} = 1.25 = \frac{5}{4} \quad \text{...(6)}$$

Since the regression lines pass through the point $(\bar{x}, \bar{y})$, we have

$$5\bar{x} = 6\bar{y} + 24$$

$$1000\bar{y} = 768\bar{x} - 3608.$$

Solving the above equations for $\bar{x}$ and $\bar{y}$, we get $\bar{x} = 6$, $\bar{y} = 1$.

Co-efficient of variability of $x = \frac{\sigma_x}{\bar{x}}$

Co-efficient of variability of $y = \frac{\sigma_y}{\bar{y}}$.

$\therefore$ Required ratio $= \frac{\sigma_x}{\bar{x}} \times \frac{\bar{y}}{\sigma_y} = \frac{\bar{y}}{\bar{x}}\left(\frac{\sigma_x}{\sigma_y}\right) = \frac{1}{6} \times \frac{5}{4} = \frac{5}{24}$. | Using (6)

Example 12. *A panel of two judges, A and B, graded seven TV serial performances by awarding marks independently as shown in the following table:*

Performance	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Marks by A	*46*	*42*	*44*	*40*	*43*	*41*	*45*
Marks by B	*40*	*38*	*36*	*35*	*39*	*37*	*41*

The eighth TV performance which judge B could not attend, was awarded 37 marks by judge A. If the judge B had also been present, how many marks would be expected to have been awarded by him to the eighth TV performance?

Use regression analysis to answer this question.

Sol. Let the marks awarded by judge A be denoted by x and the marks awarded by judge B be denoted by y.

Here, $n = 7$; $\bar{x} = \frac{\Sigma x}{n} = \frac{46+42+44+40+43+41+45}{7} = 43$

$\bar{y} = \frac{\Sigma y}{n} = \frac{40+38+36+35+39+37+41}{7} = 38$

Let us form the table as

x	y	xy	x^2
46	40	1840	2116
42	38	1596	1764
44	36	1584	1936
40	35	1400	1600
43	39	1677	1849
41	37	1517	1681
45	41	1845	2025
$\Sigma x = 301$	$\Sigma y = 266$	$\Sigma xy = 11459$	$\Sigma x^2 = 12971$

Regression coefficient,

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{(7 \times 11459) - (301 \times 266)}{(7 \times 12971) - (301)^2} = 0.75$$

Regression line of y on x is given by

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

$$y - 38 = 0.75(x - 43)$$

$\Rightarrow$ $y = 0.75x + 5.75$

when $x = 37$, $y = 0.75(37) + 5.75 = 33.5$ marks

Hence, if judge B had also been present, 33.5 marks would be expected to have been awarded to the eighth TV performance.

Example 13. *Two variables x and y have zero means, the same variance σ^2 and zero correlation, show that:*

$$u = x\cos\alpha + y\sin\alpha \quad \text{and} \quad v = x\sin\alpha - y\cos\alpha$$

have the same variance σ^2 and zero correlation. (U.P.T.U. 2007)

Sol. We are given that

$$r(x, y) = 0 \Rightarrow \text{Cov}(x, y) = 0, \quad \sigma_x^2 = \sigma_y^2 = \sigma^2$$

We have, $\sigma_u^2 = V(x\cos\alpha + y\sin\alpha)$

$= \cos^2\alpha\, V(x) + \sin^2\alpha\, V(y) + 2\sin\alpha\cos\alpha\, \text{Cov}(x, y)$

$= (\cos^2\alpha + \sin^2\alpha)\,\sigma^2$ | $\because$ Cov $(x, y) = 0$

$= \sigma^2$

Similarly, $\sigma_v^2 = \sigma^2$

$\text{Cov}(u, v) = E[(u - \bar{u})(v - \bar{v})]$

$= E[(x\cos\alpha + y\sin\alpha - \bar{x}\cos\alpha - \bar{y}\sin\alpha)$

$(x\sin\alpha - y\cos\alpha - \bar{x}\sin\alpha + \bar{y}\cos\alpha)]$

$= E[(x\cos\alpha + y\sin\alpha)(x\sin\alpha - y\cos\alpha)]$ | $\because$ $\bar{x} = 0 = \bar{y}$

$= [E(x^2) - E(y^2)]\sin\alpha\cos\alpha + E(xy)(\sin^2\alpha - \cos^2\alpha)$

$= 0$ | $\because$ $\sigma_x^2 = \sigma_y^2 = \sigma^2$ and $E(xy) = 0$

$\therefore$ $r = \frac{\text{Cov}(u, v)}{\sigma_u\,\sigma_v} = 0.$

ASSIGNMENT

1. (*i*) Discuss regression and its importance. Given the following data:

x:	1	5	3	2	1	7	3
y:	6	1	0	0	1	2	1	5

Find a regression line of x on y. *(U.P.T.U. 2008)*

(*ii*) In a study between the amount of rainfall and the quantity of air pollution removed the following data were collected:

Daily rainfall: (*in .01 cm*)	4.3	4.5	5.9	5.6	6.1	5.2	3.8	2.1
Pollution removed: (mg/m^3)	12.6	12.1	11.6	11.8	11.4	11.8	13.2	14.1

Find the regression line of y on x.

(*iii*) Find the two lines of regression and coefficient of correlation for the data given below: *[U.P.T.U. (MCA) 2009]*

$n = 18, \Sigma x = 12, \Sigma y = 18, \Sigma x^2 = 60, \Sigma y^2 = 96, \Sigma xy = 48$

(*iv*) From the data given, find the equation of lines of regression of x on y and y on x. Also calculate the correlation co-efficient.

x:	2	4	6	8	10
y:	5	7	9	8	11

(U.P.T.U. 2011)

2. (*i*) Can $Y = 5 + 2.8X$ and $X = 3 - 0.5Y$ be the estimated regression equations of Y on X and X on Y respectively? Explain your answer with suitable theoretical arguments.

(*ii*) Find the co-efficient of correlation when the two regression equations are

$X = -0.2Y + 4.2, \quad Y = -0.8X + 8.4$

3. (*i*) If F is the pull required to lift a load W by means of a pulley block, fit a linear law of the form $F = mW + c$ connecting F and W, using the data

W:	50	70	100	120
F:	12	15	21	25

where F and W are in kg wt. Compute F when W = 150 kg wt. *(U.P.T.U. 2007)*

(*ii*) A simply supported beam carries a concentrated load P (kg) at its mid-point. The following table gives maximum deflection y (cm) corresponding to various values of P:

P:	100	120	140	160	180	200
y:	0.45	0.55	0.60	0.70	0.80	0.85

Find a law of the form $y = a + bP$. Also find the value of maximum deflection when P = 150 kg.

4. (*i*) Find both the lines of regression of following data:

x:	5.60	5.65	5.70	5.81	5.85
y:	5.80	5.70	5.80	5.79	6.01

(*ii*) Obtain regression line of x on y for the given data:

x:	1	2	3	4	5	6
y:	5.0	8.1	10.6	13.1	16.2	20.0

[U.P.T.U. (MCA) 2007]

(*iii*) Given that:

x:	1	3	5	7	8	10
y:	8	12	15	17	18	20

Find the equations of both lines of regression. *[U.P.T.U. (C.O.) 2008]*

5. (*i*) The two regression equations of the variables x and y are $x = 19.13 - 0.87y$ and $y = 11.64 - 0.50x$. Find (*a*) mean of x's (*b*) mean of y's and (*c*) correlation coefficient between x and y.

(*ii*) Two random variables have the regression lines with equations $3x + 2y = 26$ and $6x + y = 31$. Find the mean values and the correlation coefficient between x and y. *[G.B.T.U. (MBA) 2011]*

(*iii*) In a partially destroyed laboratory data, only the equations giving the two lines of regression of y on x and x on y are available and are respectively

$7x - 16y + 9 = 0, \quad 5y - 4x - 3 = 0.$

Calculate the coefficient of correlation, $\bar{x}$ and $\bar{y}$.

(*iv*) The regression equations calculated from a given set of observations for two random variables are

$x = -0.4y + 6.4$ and $y = -0.6x + 4.6$

Calculate (*i*) $\bar{x}$ (*ii*) $\bar{y}$ (*iii*) r.

(*v*) Two lines of regression are given by

$x + 2y - 5 = 0$ and $2x + 3y - 8 = 0$ and $\sigma_x^2 = 12$,

Calculate:

(*a*) the mean values of x and y (*b*) variance of y

(*c*) the coefficient of correlation between x and y. *[U.P.T.U. (MCA) 2008, G.B.T.U. (C.O.) 2011]*

6. An analyst for a company was studying travelling expenses (y) in ₹ and duration (x) of these trips for 102 sales trip. He has found relation between x and y linear and data as follows:

$\Sigma x = 510, \quad \Sigma y = 7140, \quad \Sigma x^2 = 4150, \quad \Sigma xy = 54900, \quad \Sigma y^2 = 740200$

Calculate (*i*) Two regression lines

(*ii*) A given trip has to take 7 days. How much money should be allowed so that they will not run short of money?

7. Assuming that we conduct an experiment with 8 fields planted with corn, four fields having no nitrogen fertiliser and four fields having 80 kgs of nitrogen fertilizer. The resulting corn yields are shown in table in bushels per acre :

Field:	1	2	3	4	5	6	7	8
Nitrogen (kgs) x:	0	0	0	0	80	80	80	80
Corn yield y: (*acre*)	120	360	60	180	1280	1120	1120	760

(*a*) Compute a linear regression equation of y on x.

(*b*) Predict corn yield for a field treated with 60 kgs of fertilizer.

8. If the coefficient of correlation between two variables x and y is 0.5 and the acute angle between their lines of regression is $\tan^{-1}\left(\frac{3}{5}\right)$, show that $\sigma_x = \frac{1}{2}\sigma_y$. *(U.P.T.U. 2009)*

9. Given N = 50, Mean of $y = 44$, Variance of x is $\frac{9}{16}$ of the variance of y.

Regression equation of x on y is $3y - 5x = -180$

Find (*i*) Mean of x (*ii*) Coeff. of correlation between x and y.

10. The means of a bivariate frequency distribution are at (3, 4) and $r = 0.4$. The line of regression of y on x is parallel to the line $y = x$. Find the two lines of regression and estimate value of x when $y = 1$.

11. The following results were obtained in the analysis of data on yield of dry bark in ounces (y) and age in years (x) of 200 cinchona plants:

	x	y
Average:	9.2	16.5
Standard deviation:	2.1	4.2

Correlation coefficient = 0.84

Construct the two lines of regression and estimate the yield of dry bark of a plant of age 8 years.

12. A panel of judges A and B graded 7 debators and independently awarded the following marks:

Debator:	1	2	3	4	5	6	7
Marks by A:	40	34	28	30	44	38	31
Marks by B:	32	39	26	30	38	34	28

An eighth debator was awarded 36 marks by judge A while judge B was not present. If judge B were also present, how many marks would you expect him to award to the eighth debator assuming that the same degree of relationship exists in their judgement?

Answers

1. (*i*) $72x = -20y + 247$ (*ii*) $y = -0.6842x + 15.5324$

(*iii*) $y = 0.6923x + 0.53846$; $x = 0.4615y + 0.2051$

(*iv*) $x = 1.3y - 4.4$, $y = .65x + 4.1$; $r = .9192$

2. (*i*) No (*ii*) $r = -0.4$

3. (*i*) F = 0.18793W + 2.27595; F = 30.4654 kg wt.

(*ii*) $y = 0.04765 + 0.004071$ P; $y = 0.6583$ cm

4. (*i*)Regression line of y on x: $y = 0.74306x + 1.56821$

Regression line of x on y: $x = 0.63602y + 2.0204$

(*ii*) $x = 0.34195y - 0.660355$ (*iii*) $y = 1.3012x + 7.6265$; $x = 0.75y - 5.5833$.

5. (*i*) (*a*) 15.935 (*b*) 3.67 (*c*) -0.659

(*ii*) $\bar{x} = 4$, $\bar{y} = 7$, $r = -0.5$ (*iii*) $r = 0.7395$, $\bar{x} = -0.1034$, $\bar{y} = 0.5172$

(*iv*) $\bar{x} = 6$, $\bar{y} = 1$, $r = -0.48989$

(*v*) (*a*) $\bar{x} = 1$, $\bar{y} = 2$ (*b*) 4 (*c*) $-\frac{\sqrt{3}}{2}$

6. (*i*) $y = 12x + 10$, $x = 0.07986y - 0.59068$ (*ii*) ₹ 94

7. (*a*) $y = 11.125x + 180$ (*b*) 847.5 acre

9. (*i*) 62.4 (*ii*) 0.8

10. $y = x + 1$; $x = 0.16y + 2.36$; $x = 2.52$

11. $y = 1.68x + 1.044$, $x = 0.42y + 2.27$; $y = 14.484$

12. 33 marks.

3.47 POLYNOMIAL FIT: NON-LINEAR REGRESSION

Let $$y = a + bx + cx^2 \quad \ldots(1)$$

be a second degree parabolic curve of regression of y on x to be fitted for the data (x_i, y_i), $i = 1, 2, \ldots\ldots, n$.