CHAPTER

59

CORRELATION AND REGRESSION

59.1 INTRODUCTION

A relationship may be obtained in two series. **For example**; two series relating to the heights and weights of a group of persons are given. It may be observed that weights increase with increase in heights - so that tall people are heaviour than short sized people. We also know that the area A of circle of radius r is given by $A = \pi r^2$. It means larger radius will always have a larger area than a circle with smaller radius.

The intensity of light on the table decreases as the distance between source of light and table increases.

In this chapter we shall study the relationship of two series. Such a relationship is called statistical relationship.

59.2 TYPE OF DISTRIBUTION

There are two types of distributions.

(1) **Univariate Distribution.** A distribution in which there is only one variable, such as heights of students of a class.

(2) **Bivariate Distribution.** The distribution involving two variables such as heights and weights of the students of a class.

59.3 COVARIANCE

Let the corresponding values of two variables X and Y, given by ordered pairs

$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots\ldots\ldots (x_n, y_n)$

Then the covariance between X and Y is denoted by cov. (X, Y).

It is defined as

$$\text{cov}\,(X, Y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ldots\ldots\ldots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

$$= \frac{1}{n}\sum_{n=1}^{n} (x_n - \bar{x})(y_n - \bar{y})$$

$\text{cov}\,(X, Y) = E(XY) - E(X)\,E(Y)$

$E(XY)$, $E(X)$, $E(Y)$ are the corresponding means

Working Rule

Step I. Calculate the sums $\sum_{1}^{n} x_i$ and $\sum_{1}^{n} y_i$

Step II. Calculate the sum $\sum_{1}^{n} x_i y_i$ of the products of x_i and y_i.

Step III. Divide the values obtained in steps I, II by n to get $\frac{\Sigma x_i}{n}, \frac{\Sigma y_i}{n}, \frac{\Sigma x_i y_i}{n}$

Step IV. Obtain the difference $\sum_1^n \frac{x_i y_i}{n} - \left(\frac{\sum_i^n x_i}{n}\right) \cdot \left(\frac{\sum_i^n y_i}{n}\right)$ to get cov (X, Y).

Example 1. *Calculate the covariance of the following pairs of observations of two variates.*

(1, 4), (2, 2), (3, 4), (4, 8), (5, 9), (6, 12)

Solution. $\Sigma x_i = 1 + 2 + 3 + 4 + 5 + 6 = 21$

$\Sigma y_i = 4 + 2 + 4 + 8 + 9 + 12 = 39$

$\Sigma x_i y_i = (1 \times 4) + (2 \times 2) + (3 \times 4) + (4 \times 8) + (5 \times 9) + (6 \times 12)$

$= 4 + 4 + 12 + 32 + 45 + 72 = 169$

$$\text{Cov } (X, Y) = \frac{\Sigma x_i y_i}{n} - \frac{\Sigma x_i}{n}\frac{\Sigma y_i}{n} = \left[\frac{169}{6} - \frac{21}{6} \times \frac{39}{6}\right] = \frac{169}{6} - \frac{91}{4} = \frac{65}{12}$$ **Ans.**

Example 2. *Find the covariance of the following pairs of observations of two variates :*

(10, 35) (15, 20) (20, 30) (25, 30) (30, 35)

(35, 38) (40, 42) (45, 30) (50, 40) (55, 70)

Solution. $\sum_{i=1}^n x_i = 10 + 15 + 20 + 25 + 30 + 35 + 40 + 45 + 50 + 55 = 325$

$\sum_{i=1}^n y_i = 35 + 20 + 30 + 30 + 35 + 38 + 42 + 30 + 40 + 70 = 370$

$\sum_{i=1}^n x_i y_i = (10 \times 35) + (15 \times 20) + (20 \times 30) + (25 \times 30) + (30 \times 35) + (35 \times 38)$

$+ (40 \times 42) + (45 \times 30) + (50 \times 40) + (55 \times 70)$

$= 350 + 300 + 600 + 750 + 1050 + 1330 + 1680 + 1350 + 2000 + 3850$

$= 13260$

$$\text{Cov. } (X, Y) = \left[\frac{\Sigma x_i y_i}{n} - \frac{\Sigma x_i}{n} \cdot \frac{\Sigma y_i}{n}\right] = \left(\frac{13260}{10} - \frac{325}{10} \cdot \frac{370}{10}\right) = 1326 - 1202.5$$

$= 123.5$ **Ans.**

59.4 CORRELATION

Whenever two variables x and y are so related that an increase in the one is accompanied by an increase or decrease in the other, then the variables are said to be correlated.

For example, the yield of crop varies with the amount of rainfall.

59.5 TYPES OF CORRELATIONS

(1) Positive correlation

If an increase in the value of one variable X results in a corresponding increase in value of other variable Y on an average.

OR

If a decrease in the value of one variable X results in a corresponding decrease in value of other variable Y on an average.

The correlation is said to be positive.

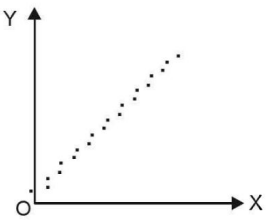
(2) Negative correlation

If the increase in the values of one variable X results in a corresponding decrease in the values of other variable Y.

OR

If the decrease in the values of one variable X results in the increase to a corresponding values of Y.

The correlation between X and Y is said to be negative.



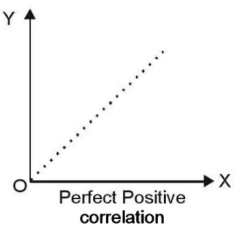
(3) Linear correlation

When all the ploted points lie approximately on a straight line, then the correlation is said to be linear correlation.

(4) Perfect correlation

If the deviation of one variable X is proportional to the deviation in other variable Y, then the correlation is said to be perfect.

In this case the plotted points on a graph lie exactly on a straight line.

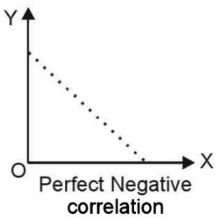


4 (a) Positive perfect correlation

If increase in one variable X is proportional to the increase in the other variable Y. The graph will be exactly straight line.

4 (*b*) Negative perfect correlation

If increase in one variable is proportional to the decrease in the other variable. The graph will be exactly a straight line.

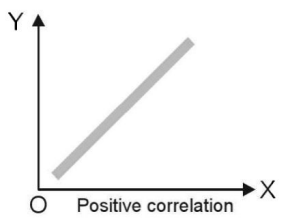


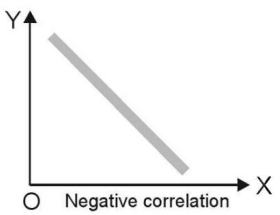
Perfect Correlation: If two variables vary in such a way that their ratio is always constant, then the correlation is said to be perfect.

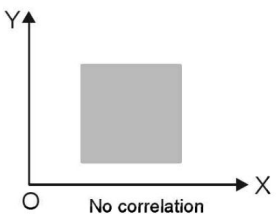
59.6 SCATTER OR DOT-DIAGRAM

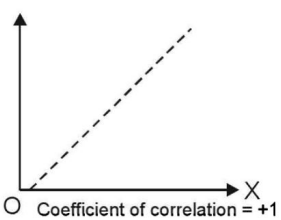
When we plot the corresponding values of two variables, taking one on x-axis and the other along y-axis, it shows a collection of dots.

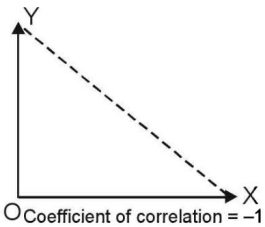
This collection of dots is called a dot diagram or a scatter diagram.



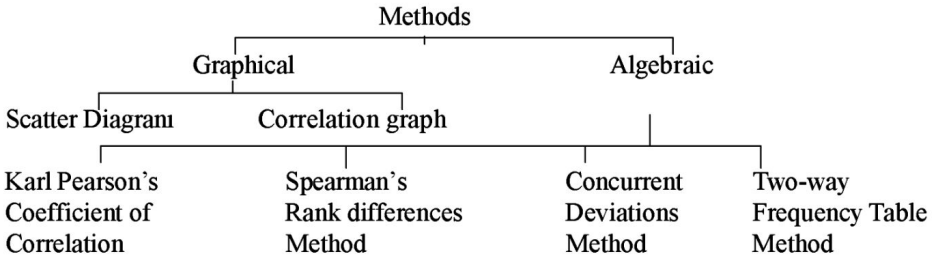








Methods of Determining Simple Correlation



59.7 KARL PEARSON'S COEFFICIENT OF CORRELATION

r between two variables x and y is defined by the relation

$$r = \frac{\sum XY}{\sqrt{(\sum X^2)(\sum Y^2)}} = \frac{P}{\sigma_x \sigma_y} = \frac{Covariance(x, y)}{\sqrt{variance\ x}\sqrt{variance\ y}},$$

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

i.e. X, Y are the deviations measured from their respective means,

$$P = \left(\frac{\sum XY}{n}\right) = \text{co variance}$$

and σ_x, σ_y being the standard deviations of these series.

Example 3. *Calculate the coefficient of correlation between x and y series from the following data:*

$$\Sigma(x-\bar{x})^2 = 136, \qquad \Sigma(y-\bar{y})^2 = 138$$
$$\Sigma(x-\bar{x})\,(y-\bar{y}) = 122$$

Solution. Here, we have

$$\Sigma X^2 = \Sigma(x-\bar{x})^2 = 136$$
$$\Sigma Y^2 = \Sigma(y-\bar{y})^2 = 138$$
$$\Sigma XY = \Sigma\,(x-\bar{x})\,(y-\bar{y}) = 122$$
$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2}.\sqrt{\Sigma Y^2}} \qquad \ldots (1)$$

Putting the values of ΣXY, ΣX^2 & ΣY^2 in (1), we get

$$r = \frac{122}{\sqrt{136}\,\sqrt{138}} = \frac{122}{11.66 \times 11.75} = \frac{122}{137.005} = 0.89 \qquad \textbf{Ans.}$$

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

Roll No.	*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 coefficient of correlation.

Solution. Let the marks of two subjects be denoted by x and y respectively.

Then the mean for x marks $= \frac{650}{10} = 65$ and the mean of y marks $= \frac{660}{10} = 66$

If X and Y are deviations of x's and y's from their respective means, then the data may be arranged in the following form :

x	y	$X = x - 65$	$Y = y - 66$	X^2	Y^2	XY
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
650	660	0	0	5398	2224	2704

Here $\sum X^2 = 5398$, $\sum Y^2 = 2224$, $\sum XY = 2704$

$$\therefore \quad r = \frac{\sum XY}{\sqrt{(\sum X^2)(\sum Y^2)}} = \frac{2704}{\sqrt{5398 \times 2224}} = \frac{2704}{73.4 \times 47.1} = \frac{2704}{3457} = 0.78$$ **Ans.**

Example 5. *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. III Semester, 2009-2010)

Solution. Let the marks of two subjects be denoted by x and y respectively.

Let the assemed mean for x marks be 35 and that of for y be 17.

x	y	$X' = x - 35$	$Y' = y - 17$	X'^2	Y'^2	$X'Y'$
25	08	−10	−9	100	81	90
30	10	−5	−7	25	49	35
32	15	−3	−2	9	4	6
35	17	0	0	0	0	0
37	20	2	3	4	9	6
40	23	5	6	25	36	30
42	24	7	7	49	49	49
45	25	10	8	100	64	80
N = 8		$\Sigma X' = 6$	$\Sigma Y' = 6$	$\Sigma X'^2 = 312$	$\Sigma Y'^2 = 292$	$\Sigma X'Y' = 296$

We know that, $$r = \frac{\frac{\Sigma X'Y'}{N} - \left(\frac{\Sigma X'}{N}\right)\left(\frac{\Sigma Y'}{N}\right)}{\sqrt{\left\{\frac{\Sigma X'^2}{N} - \left(\frac{\Sigma X'}{N}\right)^2\right\}\left\{\frac{\Sigma Y'^2}{N} - \left(\frac{\Sigma Y'}{N}\right)^2\right\}}} = \frac{\frac{296}{8} - \left(\frac{6}{8}\right)\left(\frac{6}{8}\right)}{\sqrt{\left[\left\{\frac{312}{8} - \left(\frac{6}{8}\right)^2\right\}\left\{\frac{292}{8} - \left(\frac{6}{8}\right)^2\right\}\right]}}$$

$$= \frac{\frac{583}{16}}{\sqrt{\left(\frac{615}{16}\right)\left(\frac{575}{16}\right)}} = \frac{583}{\sqrt{353625}} = \frac{583}{594.66} = 0.98039$$ **Ans.**

59.8 COEFFICIENT OF CORRELATION OF GROUPED DATA

$$r = \frac{\frac{\Sigma f X'Y'}{N} - \left(\frac{\Sigma f X'}{N}\right)\left(\frac{\Sigma f Y'}{N}\right)}{\sqrt{\left\{\frac{\Sigma f X'^2}{N} - \left(\frac{\Sigma f X'}{N}\right)^2\right\}}\sqrt{\left\{\frac{\Sigma f Y'^2}{N} - \left(\frac{\Sigma f Y'}{N}\right)^2\right\}}}$$

where r is the coefficient of correlation.

X' = Deviation from *assumed* mean of $x = x - a$

Y' = Deviation from *assumed* mean of $y = y - b$

N = Total number of items.

Example 6. *Find the coefficient of correlation between the age and the sum assured from the following table.*

Sum assured in ₹.

Age – group	10,000	20,000	30,000	40,000	50,000	*No. of persons*
20 – 30	4	6	3	7	1	21
30 – 40	2	8	15	7	1	33
40 – 50	3	9	12	6	2	32
50 – 60	8	4	2	–	–	14
	17	27	32	20	4	100

Solution. Let the sum assured denoted by x and the age group by y.

$$x' = \frac{x - 30,000}{10,000}, \qquad y' = \frac{y - 45}{10}$$

$$r = \frac{\frac{\Sigma f X'Y'}{N} - \left(\frac{\Sigma f X'}{N}\right)\left(\frac{\Sigma f Y'}{N}\right)}{\sqrt{\left\{\frac{\Sigma f X'^2}{N} - \left(\frac{\Sigma f X'}{N}\right)^2\right\}}\sqrt{\left[\frac{\Sigma f Y'^2}{N} - \left(\frac{\Sigma f Y'}{N}\right)^2\right]}} \quad \text{... (1)} \qquad [N = \Sigma f]$$

Putting the values in (1), we get

x			10,000		20,000		30,000		40,000		50,000		Row-wise			
		x'	– 2		– 1		0		1		2					
y		y'	f	$fX'Y'$	f	$fX'Y'$	f	$fX'Y'$	f	$fX'Y'$	f	$fX'Y'$	Σf	$f.Y'$	fY'^2	$fX'Y'$
20–30	25	– 2	4	16	6	12	3	0	7	–14	1	–4	21	–42	84	+10
30–40	35	–1	2	4	8	8	15	0	7	–7	1	–2	33	–33	33	+3
40–50	45	0	3	0	9	0	12	0	6	0	2	0	32	0	0	0
50–60	55	1	8	–16	4	–4	2	0	–	0	–	0	14	14	14	–20
		Σf	17		27		32		20		4		N = 100	$\Sigma fY'$ = –61	$\Sigma fY'^2$ – 131	$\Sigma fX'Y'$ = – 7
		fX'	–34		–27		0		20		8		$\Sigma fX'$ = – 33			
		fX'^2	68		27		0		20		16		$\Sigma fX'^2$ = 131			
		$fX'Y'$		4		16		0		–21		–6	$\Sigma fX'Y'$ = –7			

$$r = \frac{\frac{-7}{100} - \left(\frac{-33}{100}\right)\left(\frac{-61}{100}\right)}{\sqrt{\left\{\frac{131}{100} - \left(\frac{-33}{100}\right)^2\right\}}\sqrt{\left\{\frac{131}{100} - \left(\frac{-61}{100}\right)^2\right\}}}$$

Multiplying numerator and denominator by 10,000, we get

$$= \frac{100(-7) - (-33)(-61)}{\sqrt{100(131) - (-33)^2}\sqrt{100(131) - (-61)^2}} = \frac{-700 - 2013}{\sqrt{13100 - 1089}\sqrt{13100 - 3721}}$$

$$= \frac{-2713}{\sqrt{12011}\sqrt{9379}} = \frac{-2713}{109.59 \times 96.85} = \frac{-2713}{10613.7915} = -0.2556$$

Hence, the age and sum assured are negatively correlated, *i.e.*, as age goes up the sum assured comes down. **Ans.**

Example 7. *Calculate the coefficient of correlation for the following table :*

x – *age* / *marks*	0 – 4	4 – 8	8 – 12	12 – 16	Total
0 – 5	7	—	—	—	7
5 – 10	6	8	—	—	14
10 – 15	—	5	3	—	8
15 – 20	—	7	2	—	9
20 – 25	—	—	—	9	9
Total	13	20	5	9	47

Solution.

	x	*x*	2		6		10		14		Row-wise			
		X′	– 2		– 1		0		1					
	y	*Y′*	*f*	*fX′Y′*	*f*	*fX′Y′*	*f*	*fX′Y′*	*f*	*fX′Y′*	Σ*f*	*fY′*	*fY′*²	Σ*fX′Y′*
0–5	2.5	– 2	7	28							7	– 14	28	28
5– 10	7.5	– 1	6	12	8	8					14	– 14	14	20
10–15	12.5	0			5	0	3	0			8	0	0	0
15–20	17.5	1			7	– 7	2	0			9	9	9	– 7
20–25	22.5	2							9	18	9	18	36	18
	Σ*f*		13		20		5		9		47	Σ*fY′* = – 1	Σ*fY′*² = 87	Σ*fX′Y′* = 59
	fX′		– 26		– 20		0		9		Σ*fX′* = – 37			
	*fX′*²		52		20		0		9		Σ*fX′*² = 81			
	fX′Y′			40		1		0		18	Σ*fX′Y′* = 59			

Here, $\Sigma f X' = -37,\ \Sigma f X'^2 = 81,\ \Sigma f Y' = -1,\ \Sigma f Y'^2 = 87,\ \Sigma f X' Y' = 59$

$$r = \frac{\frac{\Sigma f X'Y'}{N} - \left(\frac{\Sigma f X'}{N}\right)\left(\frac{\Sigma f Y'}{N}\right)}{\sqrt{\frac{\Sigma f X'^2}{N} - \left(\frac{\Sigma f X'}{N}\right)^2}\sqrt{\frac{\Sigma f Y'^2}{N} - \left(\frac{\Sigma f Y'}{N}\right)^2}}$$

$$= \frac{\frac{59}{47} - \left(\frac{-37}{47}\right)\left(\frac{-1}{47}\right)}{\sqrt{\left\{\frac{81}{47} - \left(\frac{-37}{47}\right)^2\right\}}\sqrt{\left\{\frac{87}{47} - \left(\frac{-1}{47}\right)^2\right\}}} = \frac{1.255 - 0.017}{\sqrt{1.723 - 0.620}\sqrt{1.851 - 0.0005}}$$

$$= \frac{1.238}{\sqrt{1.103}\sqrt{1.8505}} = \frac{1.238}{1.05 \times 1.36} = \frac{1.238}{1.428} = 0.87$$ **Ans.**

Example 8. *A computer operator while calculating the coefficient between two variates x and y for 25 pairs of observations obtained the following constants :*

$n = 25,\ \Sigma x = 125,\ \Sigma x^2 = 650,\quad \Sigma y = 100,\ \Sigma y^2 = 460,\ \Sigma xy = 508$

It was however later discovered at the time of checking that he had copied down two pairs as (6, 14) and (8, 6) while the correct pairs were (8, 12) and (6, 8). Obtain the correct value of the correlation coefficient.

Solution. Here, corrected Σx = Incorrect $\Sigma x - (6 + 8) + (8 + 6) = 125 - 14 + 14 = 125$

Corrected Σy = Incorrect $\Sigma y - (14 + 6) + (12 + 8) = 100 - 20 + 20 = 100$

Corrected $\Sigma x^2 = 650 - (6^2 + 8^2) + (8^2 + 6^2) = 650 - 100 + 100 = 650$

Corrected $\Sigma y^2 = 460 - (14^2 + 6^2) + (12^2 + 8^2) = 460 - 232 + 208 = 436$

Corrected $\Sigma xy = 508 - [(6)(14) + (8)(6)] + (8)(12) + (6)(8)$

$= 508 - (84 + 48) + (96 + 48) = 508 - 132 + 144 = 520$

Corrected value of correlation coefficient is

$$r_{xy} = \frac{520 - \frac{125 \times 100}{25}}{\sqrt{\left[650 - \frac{(125)^2}{25}\right]\left[436 - \frac{(100)^2}{25}\right]}} = \frac{520 - 500}{\sqrt{(650 - 625)(436 - 400)}} = \frac{20}{\sqrt{25 \times 36}} = \frac{2}{3} = 0.67$$

Ans.

59. 9 SPEARMAN'S RANK CORRELATION

The coefficient of rank correlation is applied to the problems in which data cannot be measured quantitatively but qualitative assessment is possible such as beauty, honesty etc. In this case the best individual is given the rank no. 1 next rank no. 2 and so on.

59. 10 SPEARMAN'S RANK CORRELATION COEFFICIENT

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

Solution. Let $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$ be the ranks of n individuals corresponding to two characteristics.

Assuming nor two individuals are equal in either classification, each individual takes the values 1, 2, 3, ... n and hence their arithmetic means are, each

$$= \frac{\Sigma n}{n} = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$$

Let $x_1, x_2, x_3, \ldots x_n$ be the values of variable X and $y_1, y_2, y_3, \ldots y_n$ those of Y.

Then $$d = X - Y = \left(x - \frac{n+1}{2}\right) - \left(y - \frac{n+1}{2}\right) = x - y$$

where X and Y are deviations from the mean.

$$\sum X^2 = \sum\left(x - \frac{n+1}{2}\right)^2 = \sum x^2 - (n+1)\sum x + \sum\left(\frac{n+1}{2}\right)^2$$

$$= \frac{n(n+1)(2n+1)}{6} - \frac{(n+1)n(n+1)}{2} + n\left(\frac{n+1}{2}\right)^2 = \frac{n(n^2-1)}{12}$$

Clearly, $\sum X = \sum Y$ and $\sum X^2 = \sum Y^2$ $\therefore$ $\sum Y^2 = \frac{n(n^2-1)}{12}$

Hence $$\sum d^2 = \sum (x-y)^2 = \sum x^2 + \sum y^2 - 2\sum xy$$

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

Putting these values in $r = \dfrac{\sum XY}{\sqrt{\sum X^2}\sqrt{\sum Y^2}} = \dfrac{\frac{1}{12}n(n^2-1) - \frac{1}{2}\sum d^2}{\frac{n(n^2-1)}{12}} = 1 - \dfrac{6\sum d^2}{n(n^2-1)}$ **Ans.**

Working Rule

Step I. Assign ranks to each item of both series, if they are not given.

Step II. Calculate the difference D of ranks of X from the rank of Y and write it in a separate column.

Step III. Square the difference D and write D^2 in a separate column.

Step IV. Apply the formula to get the Rank correlation.

$$r = 1 - \frac{6\Sigma D^2}{n(n^2-1)}$$

where n is the total number of pairs of observations.

Example 9. *Compute Spearman's rank correlation coefficient r for the following data:*

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

Solution.

Person	*Rank in statistics*	*Rank in income*	$d = R_1 - R_2$	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
				$\Sigma d^2 = 280$

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}, \quad r = 1 - \frac{6\times 280}{10\,(100-1)} = 1 - 1.697 = -0.697$$ **Ans.**

59.11 EQUAL RANKS

If there are more than one item with the same rank. The rank to the euqal items is assigned by average rank to each of these individuals.

For example; Suppose an item is repeated at the rank 5th (*i.e.*, the 5th and 6th items are having the same values then the common rank is assigned to 5th and 6th item is $\frac{5+6}{2} = 5.5$, which is the average of 5 and 6. The next rank assigned will be seven.

If an item is repeated thrice at rank 2, then the common rank assigned to each value will be $\frac{2+3+4}{3} = 3$ which is the arithmetic mean of 2, 3 and 4. Then next rank to be assigned would be 5.

To find the rank of correlation coefficient of repeated ranks, correlation factor is added to the Spearman's rank correlation formula.

59.12 CORRELATION FACTOR

In the formula of rank correlation coefficient, add the factor $\frac{m(m^2-1)}{12}$ to Σd^2, where m is the number of times an item (say a_1) is repeated. This factor is added for each repeated value in both the series.

The total number of observations is denoted by n.

The modified formula for the rank correlation coefficient is given below

$$r = 1 - \frac{6\left[\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3) + \ldots\ldots\right]}{n(n^2-1)}$$

Example 10. *Obtain the ranks 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

(Nagpur University, Summer 2002, Winter 2002)

Solution.

x	y	Rank in $x = x'$	Rank in $y = y'$	$d = x' - y'$	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
				Total	$\Sigma d^2 = 72$

Repeated Rank of x column	No. of times	Repeated Rank of y column	No. of times
2.5	$2 = m_1$	3.5	$2 = m_2$
6	$3 = m_3$		

Rank correlation coefficient

$$r = 1 - \frac{6\left[\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3)\right]}{n(n^2-1)}$$

$$r = 1 - \frac{6\left[72 + \frac{1}{12}(2^3-2) + \frac{1}{12}(2^3-2) + \frac{1}{12}(3^3-3)\right]}{10(100-1)} = 1 - \frac{6(72+0.5+0.5+2)}{10(99)} = 0.545$$ **Ans.**

Example 11. *Find rank correlation coefficient to the following data* :

x	65	63	67	64	68	62	70	66	68	67	69	71
y	68	66	68	65	69	66	68	65	71	67	68	70

(*Nagpur University, Summer 2005*)

Solution. Here we assign rank to the values of x and y and we have a table of the following form:

x	y	Rank in $x = x'$	Rank in $y = y'$	$d = x' - y'$	d^2
65	68	9	5.5	3.5	12.25
63	66	11	9.5	1.5	2.25
67	68	6.5	5.5	1	1
64	65	10	11.5	− 1.5	2.25
68	69	4.5	3	1.5	2.25
62	66	12	9.5	2.5	6.25
70	68	2	5.5	− 3.5	12.25
66	65	8	11.5	− 3.5	12.25
68	71	4.5	1	3.5	12.25
67	67	6.5	8	− 1.5	2.25
69	68	3	5.5	− 2.5	6.25
71	70	1	2	− 1	1
				Total	$\Sigma d^2 = 72.5$

Repeated Rank of x column	No. of times	Repeated Rank of y column	No. of times
4.5	$2 = m_1$	11.5	$2 = m_3$
6.5	$2 = m_2$	9.5	$2 = m_4$
		5.5	$4 = m_5$

The rank correlation coefficient r is given by

$$r = 1 - 6\left[\frac{\Sigma d^2 + \frac{1}{12}m_1(m_1^2-1) + \frac{1}{12}m_2(m_2^2-1) + \frac{1}{12}m_3(m_3^2-1) + \frac{1}{12}m_4(m_4^2-1) + \frac{1}{12}m_5(m_5^2-1)}{n(n^2-1)}\right]$$

Here $n = 12$, two x values are repeated twice so it is of the same rank. Two y values are repeated twice and one y values is repeated four times.

$$r = 1 - 6\frac{\left[72.5 + \frac{1}{12}2(2^2-1) + \frac{1}{12}2(2^2-1) + \frac{1}{12}2(2^2-1) + \frac{1}{12}2(2^2-1) + \frac{1}{12}4(4^2-1)\right]}{12(144-1)}$$

$$= 1 - \frac{6(72.5+2+5)}{12 \times 143} = 1 - 0.27797 = 0.722$$ **Ans.**

Example 12. *Establish the formula* $\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$ *where r is the correlation coefficient between x and y.*

Solution. We know that $\sigma_x^2 = \dfrac{\Sigma(x-\bar{x})^2}{n}$

$\therefore \quad \sigma_{x-y}^2 = \dfrac{\Sigma[(x-y)-\overline{(x-y)}]^2}{n}$

$\overline{x-y}$ = mean of $(x-y)$ series = mean of x– mean of $y = \bar{x} - \bar{y}$

$$\sigma_{x-y}^2 = \frac{\Sigma[(x-y)-(\bar{x}-\bar{y})]^2}{n} = \frac{\Sigma[(x-\bar{x})-(y-\bar{y})]^2}{n}$$

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

$$= \frac{\Sigma(x-\bar{x})^2}{n} + \frac{\Sigma(y-\bar{y})^2}{n} - \frac{2\Sigma(x-\bar{x})(y-\bar{y})}{n} = \sigma_x^2 + \sigma_y^2 - \frac{2\Sigma(x-\bar{x})(y-\bar{y})}{n} \quad ...(1)$$

We know that $r = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}$ or $\dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n} = r\sigma_x\sigma_y$

Putting this value in (1), we get

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

Example 13. *If X and Y are uncorrelated random variables, find the coefficient of correlation between X + Y and X – Y.*

Solution.

Let $u = X + Y$ and $v = X - Y$

Then $r = \dfrac{\Sigma(u-\bar{u})(v-\bar{v})}{n\sigma_u\sigma_v}$

Now $u = X + Y, \bar{u} = \bar{X} + \bar{Y}$

Similarly $\bar{v} = \bar{X} - \bar{Y}$

Now $\Sigma(u-\bar{u})(v-\bar{v}) = \Sigma(X-\bar{X}+Y-\bar{Y})[(X-\bar{X})-(Y-\bar{Y})]$

$= \Sigma(x+y)(x-y) = \Sigma x^2 - \Sigma y^2 = n\sigma_x^2 - n\sigma_y^2$

Also $\sigma_u^2 = \dfrac{\Sigma(u-\bar{u})^2}{n} = \dfrac{1}{n}\Sigma[(X-\bar{X})+(Y-\bar{Y})]^2 = \dfrac{1}{n}\Sigma(x+y)^2 = \dfrac{1}{n}(\Sigma x^2 + \Sigma y^2 + 2\Sigma xy)$

$= \sigma_x^2 + \sigma_y^2$ (As X and Y are not correlated, we have $\Sigma xy = 0$)

Similarly $\sigma_v^2 = \sigma_x^2 + \sigma_y^2$

$\therefore \quad r = \dfrac{\Sigma(u-\bar{u})(v-\bar{v})}{n\sigma_u\sigma_v} = \dfrac{n(\sigma_x^2 - \sigma_y^2)}{\sqrt{n(\sigma_x^2+\sigma_y^2)}\sqrt{n(\sigma_x^2+\sigma_y^2)}} = \dfrac{\sigma_x^2 - \sigma_y^2}{\sigma_x^2 + \sigma_y^2}$ **Ans.**

EXERCISE 59.1

1. Calculate the coefficient of correlation from the data given below :

x	4	6	8	10	12
y	2	3	4	6	10

Ans. 0.95

3. Find the coefficient of corrrelation of the following data taking new origin of x at 70 and for y at 67:

x	67	68	64	68	72	70	69	70
y	65	66	67	67	68	69	71	73

(*A.M.I.E., Winter 2002*) **Ans.** 0.472

3. Calculate Karl Pearson's coefficient of corrrelation from the following data, using 20 as working mean for price and 70 as working mean for demand.

Price	14	16	17	18	19	20	21	22	23
Demand	84	78	70	75	66	67	62	58	60

Ans. 1.044

4. The ranks of the same 16 students in two subjects A and B where as follows. Two numbers within brackets denote the ranks of the students in A and B respectively :

(1, 1), (2, 10), (3, 3), (4, 4), (5, 5), (6, 7), (7, 2), (8, 6), (9, 8), (10, 11), (11, 15), (12, 9), (13, 14), (14, 12), (15, 16), (16, 13).

Calculate the rank correlation for proficiencies of this group in subjects A and B. **Ans.** 0.8

5. Show that $E(x) = 0$, $E(x, y) = 0$ and hence deduce that the correlation between x and y is zero.

6. x and y are two random variables with the same standard deviation and correlation coefficient r. Show that the coefficient of correlation between x and $x + y$ is $\sqrt{\frac{1+r}{2}}$

7. Calculate the coefficients of correlation between x (Marks in Mathematics) and y (marks in Physics) given in this following data :

y \ x	10 – 40	40 – 70	70 – 100	Total
0 – 30	5	20	—	25
30 – 60	—	28	2	30
60 – 90	—	32	13	45
	5	80	15	100

Ans. 0.4517

8. Calculate from the data reproduced pertaining to 66 selected villages in Meerut district, the value of r, between 'total cultivated area' and 'the area under wheat'.

Area under wheat (in Bighas)	0 – 500	500 – 1000	1000 – 1500	1500 – 2000	2000 – 2500	Total
0 – 200	12	6	–	–	–	18
200 – 400	2	18	4	2	1	27
400 – 600	–	4	7	3	–	14
600 – 800	–	1	–	2	1	4
800 – 1000	–	–	–	1	2	3
Total	14	29	11	8	4	66

Ans. 0.749

9. Find the coefficient of correlation for the following data :

y \ x	16 – 18	18 – 20	20 – 22	22 – 24	Total
10 – 20	2	1	1	–	4
20 – 30	3	2	3	2	10
30 – 40	3	4	5	6	18
40 – 50	2	2	3	4	11
50 – 60	–	1	2	2	5
60 – 70	–	1	2	1	4
	10	11	16	15	52

Ans. 0.28

10. Two judges in a beauty contest rank the ten competitors in the following order :

6	4	3	1	2	7	9	8	10	5
4	1	6	7	5	8	10	9	3	2

Do the two judges appear to agree in their standard ? **Ans.** 0.224

59.13 REGRESSION

If the scatter diagram indicates some relationship between two variables x and y, then the dots of the scatter diagram will be concentrated round a curve. This curve is called the *curve of regression.*

Regression analysis is the method used for estimating the unknown values of one variable corresponding to the known value of another variable.

59.14 LINE OF REGRESSION

When the curve is a straight line, it is called a line of regression. A line of regression is the straight line which gives the best fit in the least square sense to the given frequency.

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

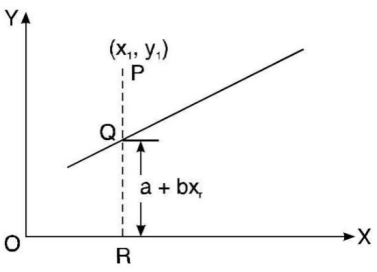
59.15 EQUATIONS TO THE LINES OF REGRESSION

Let $y = a + bx$ ... (1)

be the equation of the line of regression of y on x.

Let (x_r, y_r) be any point of dot.

From the figure



$$PR = y_r$$

$$QR = a + bx_r$$

$$PQ = PR - QR = y_r - a - bx_r$$

Let S be the sum of the squares of such distances, then

$$S = \sum (y - a - bx)^2$$

According to the principle of least squares, we have to choose a and b so that S is minimum. The method of least square gives the condition for minimum value of S.

$$\frac{\partial S}{\partial a} = -2\sum (y - a - bx), \qquad \frac{\partial S}{\partial b} = -2\sum (y - a - bx)x$$

$$\frac{\partial S}{\partial a} = 0, \qquad \frac{\partial S}{\partial b} = 0, \quad \text{for } S \text{ minimum}$$

i.e. $$\sum (y - a - bx) = 0 \Rightarrow \sum y - na - b\sum x = 0 \Rightarrow \sum y = na + b\sum x \qquad \text{... (2)}$$

and $$\sum (xy - ax - bx^2) = 0 \Rightarrow \sum xy - a\sum x - b\sum x^2 = 0$$

$$\Rightarrow \qquad \sum xy = a\sum x + b\sum x^2 \qquad \text{...(3)}$$

Dividing (2) by n, we get

$$\frac{\Sigma y}{n} = a + b\frac{\Sigma x}{n} \qquad \left(\bar{y} = \frac{\sum y}{n}, \bar{x} = \frac{\sum x}{n}\right)$$

$$\bar{y} = a + b\bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the means of x series and y series.

This shows that $(\bar{x}, \bar{y})$ lie on the line of regression (1), shifting the origin to $(\bar{x}, \bar{y})$, the equation (3) becomes

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

But $$\sum (x-\bar{x}) = 0$$

$\Rightarrow$ $$\sum (x-\bar{x})(y-\bar{y}) = b\sum (x-\bar{x})^2$$

or $$b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{\sum XY}{\sum X^2} \qquad \ldots(4)$$

We know that $$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2}\sqrt{\Sigma Y^2}} = \frac{\Sigma XY}{n\sqrt{\frac{\Sigma X^2}{n}}\sqrt{\frac{\Sigma Y^2}{n}}} = \frac{\Sigma XY}{n\sigma_x \sigma_y}$$

or $$\sum XY = nr\sigma_x\sigma_y$$

Putting the value of $\sum XY$ in (4), we get $$b = \frac{nr\sigma_x\sigma_y}{\sum X^2} = \frac{r\sigma_x\sigma_y}{\frac{\sum X^2}{n}} = \frac{r\sigma_x\sigma_y}{\sigma_x^2} = \frac{r\sigma_y}{\sigma_x}$$

i.e. slope of the line of regression $= b = r\frac{\sigma_y}{\sigma_x}$

The line of regression passes through $(\bar{x}, \bar{y})$.

Hence the equation to the line of regression is $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$

Similarly the regression line of x on y is $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$.

Note. $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ are known as the coefficients of regression.

$$b_{yx}.b_{xy} = \left(r\frac{\sigma_y}{\sigma_x}\right)\left(r\frac{\sigma_x}{\sigma_y}\right) = r^2$$

Example 14. *If* θ *be the acute angle between the two regression lines in the case of two variables x and y, show that*

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

where r, σ_x, σ_y *have their usual meanings. Explain the significance where* $r = 0$ *and* $r = \pm 1$. *(Nagpur University, Winter 2004, A.M.I.E., Winter 2001)*

Solution. Lines of regression are

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \qquad \ldots(1) \qquad \therefore m_1 = r\frac{\sigma_y}{\sigma_x}$$

and $$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y-\bar{y}) \qquad \ldots(2) \qquad \therefore m_2 = \frac{1}{r}\frac{\sigma_y}{\sigma_x}$$

$$\tan\theta = \frac{m_2 - m_1}{1+m_1m_2} = \frac{\frac{1}{r}\frac{\sigma_y}{\sigma_x} - r\frac{\sigma_y}{\sigma_x}}{1 + r\frac{\sigma_y}{\sigma_x}\times\frac{1}{r}\frac{\sigma_y}{\sigma_x}} = \frac{\left(\frac{1}{r}-r\right)\frac{\sigma_y}{\sigma_x}}{1+\frac{\sigma_y^2}{\sigma_x^2}}$$

$$\tan\theta = \frac{1-r^2}{r}\cdot\frac{\left(\frac{\sigma_y}{\sigma_x}\right)\sigma_x^2}{\sigma_x^2+\sigma_y^2} = \frac{1-r^2}{r}\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2} \qquad ...(3) \qquad \textbf{Proved.}$$

(*a*) If $r = 0$, then there is no relationship between the two variables and they are independent.

On putting the value of $r = 0$ in (3) we get $\tan\theta = \infty, \theta = \frac{\pi}{2}$. So the lines (1) and (2) are perpendicular.

(*b*) If $r = 1$ or -1

On putting these values of r in (3) we get, $\tan\theta = 0$ or $\theta = 0$

i.e. lines (1) and (2) coincide.

The correlation between the variables is perfect. **Ans.**

Example 15. *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. III Semester, June 2009*)

Solution. Here, we have

$$r = 0.5$$

$$\theta = \tan^{-1}\left(\frac{3}{5}\right) \quad \Rightarrow \quad \tan\theta = \frac{3}{5}$$

$$\tan\theta = \frac{1-r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^{\,2}+\sigma_y^{\,2}} \qquad ... (1) \quad \text{(From Example 14)}$$

Putting the values of r and $\tan\theta$ in (1), we get

$$\frac{3}{5} = \frac{1-\frac{1}{4}}{\frac{1}{2}}\frac{\sigma_x\sigma_y}{\sigma_x^{\,2}+\sigma_y^{\,2}}$$

$$\frac{3}{5} = \frac{3}{2}\frac{\sigma_x\sigma_y}{\sigma_x^{\,2}+\sigma_y^{\,2}}$$

$$2\sigma_x^{\,2} + 2\sigma_y^{\,2} - 5\sigma_x\sigma_y = 0$$

$$\Rightarrow \quad 2\sigma_x^{\,2} - 5\sigma_x\sigma_y + 2\sigma_y^{\,2} = 0$$

$$\Rightarrow \quad 2\sigma_x^{\,2} - 4\sigma_x\sigma_y - \sigma_x\sigma_y + 2\sigma_y^{\,2} = 0$$

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

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

$\Rightarrow$ Either $\sigma_x - 2\sigma_y = 0 \Rightarrow \sigma_x = 2\sigma_y$ (*Not desired*)

or $2\sigma_x - \sigma_y = 0 \Rightarrow \sigma_x = \frac{1}{2}\sigma_y$ **Proved.**

Example 16. *Find the correlation coefficient between x and y, when the lines of regression are:*

$$2x - 9y + 6 = 0 \text{ and } x - 2y + 1 = 0$$

Solution. Let the line of regression of x on y be $2x-9y+6=0$

Then, the line of regression of y on x is $x-2y+1=0$

$\therefore$ $2x-9y+6=0 \Rightarrow x=\frac{9}{2}y-3 \Rightarrow b_{xy}=\frac{9}{2}$

and $x-2y+1=0 \Rightarrow y=\frac{1}{2}x+\frac{1}{2} \Rightarrow b_{yx}=\frac{1}{2}$

$$r=\sqrt{b_{xy}\cdot b_{yx}}=\sqrt{\frac{9}{2}\times\frac{1}{2}}=\frac{3}{2}>1 \text{ which is not possible.}$$

So our choice of regression line is incorrect.

$\therefore$ The regression line of x on y is $x-2y+1=0$

And, the regression line of y on x is $2x-9y+6=0$

$\therefore$ $x-2y+1=0 \Rightarrow x=2y-1 \Rightarrow b_{xy}=2$

And $2x-9y+6=0 \Rightarrow y=\frac{2}{9}x+\frac{2}{3} \Rightarrow b_{yx}=\frac{2}{9}$

$$r=\sqrt{b_{xy}\cdot b_{yx}}=\sqrt{2\times\frac{2}{9}}=\frac{2}{3}$$

Hence, the correlation coefficient between x and y is $\frac{2}{3}$. **Ans.**

Example 17. *Two lines of regression are given by*

$5y - 8x + 17 = 0$ and $2y - 5x + 14 = 0$

If $\sigma_y^2 = 16$, find (i) the mean values of x and y (ii) σ_x^2 (iii) the coefficient of correlation between x and y.

Solution. We have, $5y - 8x + 17 = 0$ ... (1)

$2y - 5x + 14 = 0$

Since $(\bar{x}, \bar{y})$ is a common point of the two lines of regression, we have

$$5\bar{y} - 8\bar{x} + 17 = 0 \quad \text{... (2)}$$

$$2\bar{y} - 5\bar{x} + 14 = 0 \quad \text{... (3)}$$

On solving (2) and (3) for $\bar{x}$ and $\bar{y}$, we have

$$\frac{\bar{x}}{17\times2-14\times5}=\frac{\bar{y}}{-8\times14-(-5\times17)}=\frac{1}{5\times(-5)-(-8)\times2}$$

$$\Rightarrow \frac{\bar{x}}{34-70}=\frac{\bar{y}}{-112+85}=\frac{1}{-25+16} \Rightarrow \frac{\bar{x}}{-36}=\frac{\bar{y}}{-27}=\frac{1}{-9} \Rightarrow \bar{x}=\frac{36}{9}=4 \text{ and } \bar{y}=\frac{27}{9}=3$$

The equations of line of regression can be written as

$$y=\frac{8}{5}x-\frac{17}{5} \text{ and } x=\frac{2}{5}y+\frac{14}{5}$$

$$r\frac{\sigma_y}{\sigma_x}=\frac{8}{5} \text{ and } r\frac{\sigma_x}{\sigma_y}=\frac{2}{5}$$

On multiplication of two equations, we get

$$\left(r\frac{\sigma_y}{\sigma_x}\right)\left(\frac{r\sigma_x}{\sigma_y}\right)=\frac{8}{5}\times\frac{2}{5}=\frac{16}{25} \Rightarrow r^2=\frac{16}{25} \Rightarrow r=\pm\frac{4}{5} \quad \text{... (4)}$$

Now we have to determine the sign of r *i.e.* + or –

as σ_x, σ_y are always +ve, so r is also +ve from (4). $r = \frac{4}{5}$

We are given $\sigma_y^2 = 16 \qquad \therefore \ \sigma_y = 4$

And $r\frac{\sigma_y}{\sigma_x} = \frac{8}{5}$ ... (5)

On putting the values of r and σ_y in (5), we get

$$\left(\frac{4}{5}\right)\frac{4}{\sigma_x} = \frac{8}{5} \quad \Rightarrow \quad \frac{16}{5\sigma_x} = \frac{8}{5} \quad \Rightarrow \quad \sigma_x = 2$$

$$\Rightarrow \quad \sigma_x^2 = 4$$

Hence (*i*) $\bar{x} = 4, \ \bar{y} = 3,$ (*ii*) $\sigma_x^2 = 4,$ (*iii*) $r = \frac{4}{5}$ **Ans.**

Example 18. *In a partially destroyed laboratory record of an analysis of correlation data, the following results only are eligible:*

$\sigma_x^2 = 9$ *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, (c) coefficient of correlation between x and y?

(U.P., III Semester, Dec. 2009, Nagpur University, Summer 2002)

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

$$8\bar{x} - 10\bar{y} + 66 = 0 \quad \text{... (1)}$$

$$40\bar{x} - 18\bar{y} - 214 = 0 \quad \text{... (2)}$$

On solving (1) and (2), by cross multiplication method, we have

$$\frac{\bar{x}}{(-10)(-214) - (66)(-18)} = \frac{\bar{y}}{(66)(40) - (8)(-214)} = \frac{1}{8(-18) - (-10)(40)}$$

$$\frac{\bar{x}}{2140 + 1188} = \frac{\bar{y}}{2640 + 1712} = \frac{1}{-144 + 400}$$

$$\Rightarrow \quad \frac{\bar{x}}{3328} = \frac{\bar{y}}{4352} = \frac{1}{256} \quad \Rightarrow \quad \bar{x} = \frac{3328}{256} = 13, \quad \bar{y} = \frac{4352}{256} = 17$$

Also given lines of regression can be written as $y = 0.8x + 6.6 : x = 0.45\,y + 5.35$

We get $r\frac{\sigma_y}{\sigma_x} = 0.8; \qquad r\frac{\sigma_x}{\sigma_y} = 0.45$

$$\left(r\frac{\sigma_y}{\sigma_x}\right)\left(r \cdot \frac{\sigma_x}{\sigma_y}\right) = (0.8)\ (0.45)$$

$$\Rightarrow \quad r^2 = 0.36 \quad \Rightarrow \quad r = 0.6$$

Ans. $r\frac{\sigma_y}{\sigma_x} = 0.8$ ... (3)

On putting the values of r and σ_x in (3), we get

$$(0.6)\frac{\sigma_y}{3} = 0.8 \quad \Rightarrow \quad \sigma_y = \frac{3(0.8)}{0.6} = 4$$

Hence (*a*) $\bar{x} = 13, \ \bar{y} = 17,$ (*b*) $\sigma_y = 4$ (*c*) r = 0.6 **Ans.**

Example 19. *The two regression equations of the variables x and y are*

$$x = 19.13 - 0.87y \text{ and } y = 11.64 - 0.50x.$$

Find (i) Mean of x's; (ii) Mean of y's; (iii) The correlation coefficient between x and y.

Solution. $$x = 19.13 - 0.87y \quad ...(1)$$

$$y = 11.64 - 0.50x \quad ...(2)$$

As (1) and (2) pass through $(\bar{x}, \bar{y})$:

$$\bar{x} = 19.13 - 0.87\bar{y} \quad ...(3)$$

$$\bar{y} = 11.64 - 0.50\bar{x} \quad ...(4)$$

On solving (3) and (4) we get

$$\bar{x} = 15.937, \ \bar{y} = 3.67$$

From (1) $$r\frac{\sigma_x}{\sigma_y} = -0.87 \quad ...(5)$$

From (2) $$r\frac{\sigma_y}{\sigma_x} = -0.50 \quad ...(6)$$

As σ_x and σ_y are always positive, so r is negative.
Multiplying (5) and (6), we get

$$r\frac{\sigma_x}{\sigma_y} \cdot r\frac{\sigma_y}{\sigma_x} = -0.87 \times (-0.50)$$

$$r^2 = 0.435 \quad \Rightarrow \quad r = \pm 0.66 \quad \textbf{Ans.}$$

Example 20. *The regression equations calculated from a given set of observations for two random variables are*

$$x = -0.4y + 6.4 \text{ and } y = -0.6x + 4.6$$

Calculate $\bar{x}, \bar{y}$ and r.

Solution. The regression equations are

$$x = -0.4y + 6.4 \quad ... (1)$$

$$y = -0.6x + 4.6 \quad ... (2)$$

From (1) coefficient of regression of x on $y = r\frac{\sigma_x}{\sigma_y} = -0.4 \quad ...(3)$

From (2) coefficient of regression of y on $x = r\frac{\sigma_y}{\sigma_x} = -0.6 \quad ...(4)$

From (3) and (4), we have

$$\left(r\frac{\sigma_x}{\sigma_y}\right)\left(r\frac{\sigma_y}{\sigma_x}\right) = (-0.4)(-0.6)$$

$\Rightarrow$ $$r^2 = 0.24 \quad \Rightarrow \quad r = \pm 0.49$$

In (3) and (4), σ_x and σ_y are (always) negative so r is negative

$$r = -0.49$$

To find $\bar{x}$ and $\bar{y}$ we solve the equations (1) and (2) simultaneously. Their point of intersection is $(\bar{x}, \bar{y})$.

$$\bar{x} = 6, \quad \bar{y} = 1 \quad \textbf{Ans.}$$

Example 21. *Show that the geometric mean of the coefficients of regression is the coefficient of correlation.* (*AMIE, Summer 2001*)

Solution. The coefficients of regressions are $r\frac{\sigma_y}{\sigma_x}$ and $r\frac{\sigma_x}{\sigma_y}$

i.e. $$G.M = \sqrt{r\frac{\sigma_y}{\sigma_x}.r\frac{\sigma_x}{\sigma_y}} = r$$

= coefficient of correlation. **Proved.**

Example 22. *The regression lines of y on x and of x on y are respectively* $y = ax + b$ *and* $x = cy + d$. *Show that the means are* $\bar{x} = (bc + d)(1 - ac)$ *and* $\bar{y} = (ad + b)(1 - ac)$ *and correlation coefficient between x and y is* $\sqrt{ac}$. *Also, show that the ratio of the standard deviations of y and x is* $\sqrt{\frac{a}{c}}$.

Solution. Here, we have

The regression line of y on x is $y = ax + b$ ... (1)

The regression line of x on y is $x = cy + d$ ... (2)

As (1) and (2) pass through $(\bar{x}, \bar{y})$, so

$$\bar{y} = a\bar{x} + b \quad \text{... (3)}$$

$$\bar{x} = c\bar{y} + d \quad \text{... (4)}$$

Solving (3) and (4), we get

$$\bar{x} = \frac{bc + d}{1 - ac} \quad \Rightarrow \quad \bar{y} = \frac{ad + b}{1 - ac}$$

We know that $$r\frac{\sigma_y}{\sigma_x} = a = \text{slope of (1)} \quad \text{... (5)}$$

and $$r\frac{\sigma_x}{\sigma_y} = c \quad \text{... (6)}$$

Multiplying (5) and (6), we get

$$r\frac{\sigma_y}{\sigma_x}.r\frac{\sigma_x}{\sigma_y} = a.c \quad \Rightarrow \quad r^2 = ac \quad \Rightarrow \quad r = \sqrt{ac}$$ **Proved.**

Dividing (5) by (6), we get

$$\frac{r\frac{\sigma_y}{\sigma_x}}{r\frac{\sigma_x}{\sigma_y}} = \frac{a}{c} \quad \Rightarrow \quad \left(\frac{\sigma_y}{\sigma_x}\right)^2 = \frac{a}{c} \quad \Rightarrow \quad \frac{\sigma_y}{\sigma_x} = \sqrt{\frac{a}{c}}$$ **Proved.**

Example 23. *Prove that arithmetic mean of the coefficients of regression is greater than the coefficient of correlation.* (*A.M.I.E.T.E., Summer 2000*)

Solution. Coefficients of regression are $r\frac{\sigma_y}{\sigma_x}$ and $r\frac{\sigma_x}{\sigma_y}$

We have to prove that *A.M.* > r

$$\Rightarrow \quad \frac{1}{2}\left[r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y}\right] > r \quad \Rightarrow \quad \frac{1}{2}\left[\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y}\right] > 1$$

$$\Rightarrow \quad \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} - 2 > 0 \quad \Rightarrow \quad \frac{1}{\sigma_x\sigma_y}\left[\sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y\right] > 0$$

$$\Rightarrow \quad \frac{1}{\sigma_x\sigma_y}\left[\sigma_x - \sigma_y\right]^2 > 0$$ which is true. **Proved.**

Example 24. *In a study between the amount of rainfall and the quantity of air pollution removed the following data were collected.*

Daily Rainfall in 0.01 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. (*A.M.I.E., Summer 2000*)

Solution.

S.N	*x (metre)*	*y*	*xy*	x^2
1	4.3	12.6	54.18	18.49
2	4.5	12.1	54.45	20.25
3	5.9	11.6	68.44	34.81
4	5.6	11.8	66.08	31.36
5	6.1	11.4	69.54	37.21
6	5.2	11.8	61.36	27.04
7	3.8	13.2	50.16	14.44
8	2.1	14.1	29.61	4.41
	37.5	98.6	453.82	188.01

Let $y - a + bx$ be the equation of the line of regression of y on x, where a and b are given by the following equations.

$$\Sigma y = na + b\,\Sigma x \quad \Rightarrow \quad 98.6 = 8a + 37.5b \qquad ...(1)$$

$$\Sigma xy = a\,\Sigma x + b\,\Sigma x^2 \quad \Rightarrow \quad 453.82 = 37.5a + 188.01b \qquad ...(2)$$

On solving (1) and (2), we get $a = 15.49$ and $b = -0.675$.

The equation of the line of regression is $y = 15.49 - 0.675x$ **Ans.**

Example 25. *Find the correlation coefficient and regression lines for the data :*

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

(*Nagpur University, Summer 2000*)

Solution. We have,

x	1	2	3	4	5	$\Sigma x = 15$
y	2	5	3	8	7	$\Sigma y = 25$

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{1}{5} \times 15 = 3 \qquad \bar{y} = \frac{1}{n}\Sigma y = \frac{1}{5} \times 25 = 5$$

x	y	$X = x - 3$	$Y = y - 5$	X^2	Y^2	XY
1	2	– 2	– 3	4	9	6
2	5	–1	0	1	0	0
3	3	0	– 2	0	4	0
4	8	1	3	1	9	3
5	7	2	2	4	4	4
		Total		10	26	13

Correlation coefficient 'r' $$= \frac{\Sigma XY}{\sqrt{\Sigma X^2}\sqrt{\Sigma Y^2}} = \frac{13}{\sqrt{10 \times 26}} = 0.8062$$

and $$r = \frac{\Sigma XY}{n\sigma_x \sigma_y} \quad \Rightarrow \quad \Sigma XY = n r \sigma_x \sigma_y$$

Slope of regression line of y on $x = \dfrac{\Sigma XY}{\Sigma X^2} = \dfrac{13}{10}$

Slope of regression line of x on $y = \dfrac{\Sigma XY}{\Sigma Y^2} = \dfrac{13}{26} = \dfrac{1}{2}$

Equation of regression line of y on x is

$$y - \bar{y} = \frac{13}{10}(x - \bar{x}) \Rightarrow y - 5 = \frac{13}{10}(x - 3) \Rightarrow y = 1.3x + 1.1$$

Equation of regression line of x on y is

$$x - \bar{x} = \frac{1}{2}(y - \bar{y}) \Rightarrow x - 3 = 0.5\,(y - 5) \Rightarrow x = 0.5y + 0.5$$ **Ans.**

Example 26. *Find the coefficient of correlation and regression lines to the following data:*

x	5	7	8	10	11	13	16
y	33	30	28	20	18	16	9

(*Nagpur University, Winter 2003*)

Solution. Here $n = 7$, $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{70}{7} = 10$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{154}{7} = 22$$

$\therefore$ $X = x - \bar{x} = x - 10$

$Y = y - \bar{y} = y - 22$

The various calculations are shown in the following table :

x	y	$X = x - 10$	$Y = y - 22$	XY	X^2	Y^2
5	33	– 5	11	– 55	25	121
7	30	– 3	8	– 24	9	64
8	28	– 2	6	– 12	4	36
10	20	0	– 2	0	0	4
11	18	1	– 4	– 4	1	16
13	16	3	– 6	– 18	9	36
16	9	6	– 13	– 78	36	169
$\Sigma x = 70$	$\Sigma y = 154$			$\Sigma XY = -191$	$\Sigma X^2 = 84$	$\Sigma Y^2 = 446$

Coefficient of correlation

Now, $$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 . \Sigma Y^2}} = \frac{-191}{\sqrt{84 \times 446}} = -0.9868$$

$$r\frac{\sigma_y}{\sigma_x} = r\sqrt{\frac{\Sigma Y^2}{\Sigma X^2}} = -0.9868\sqrt{\frac{446}{84}} = -2.2738$$

and $$r\frac{\sigma_x}{\sigma_y} = r\sqrt{\frac{\Sigma X^2}{\Sigma Y^2}} = -0.9868\sqrt{\frac{84}{446}} = -0.4283$$

$\therefore$ Equation of line of regression y on x is

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}), \qquad y - 22 = -2.2738\,(x - 10)$$

$$y - 22 = -2.2738x + 22.738, \qquad y = -2.2738x + 44.738$$

and equation of line of regression x on y is

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

$$\Rightarrow \quad x - 10 = -0.4283\,(y - 22)$$

$$\Rightarrow \quad x - 10 = -0.4283\,y + 9.4226$$

$$\Rightarrow \quad x = -0.4283\,y + 19.4226$$

$$x + 0.4283\,y = 19.4226$$

Hence the equation of the line of regression of y on x is $y + 2.2738\,x = 44.738$

The equation of the line of regression of x on y is $x + 0.4283\,y = 19.4226$ **Ans.**

Example 27. *Find the coefficient of correlation and obtain the equation of the lines of regression for the data.*

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

(*Nagpur University, Winter 2000*)

Solution. Here, we have

x	6	2	10	4	8	$\Sigma x = 30$
y	9	11	5	8	7	$\Sigma y = 40$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{30}{5} = 6, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{40}{5} = 8$$

x	y	$X = x - 6$	$Y = y - 8$	X^2	Y^2	XY
6	9	0	1	0	1	0
2	11	− 4	3	16	9	− 12
10	5	4	− 3	16	9	− 12
4	8	− 2	0	4	0	0
8	7	2	− 1	4	1	− 2
				$\Sigma X^2 = 40$	$\Sigma Y^2 = 20$	$\Sigma XY = -26$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 . \Sigma Y^2}} = \frac{-26}{\sqrt{40 \times 20}} = \frac{-26}{28.2842} = -0.919$$

The regression coefficient of y on x is

$$\frac{\Sigma XY}{\Sigma X^2} = -\frac{26}{40} = -0.65 \qquad \left(\frac{\Sigma XY}{\Sigma X^2} = r\frac{\sigma_y}{\sigma_x}\right)$$

The equation of line of regression of y on x is

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

$$\Rightarrow \quad y - 8 = -0.65\,(x - 6) \qquad \Rightarrow \; y = -0.65\,x + 11.9$$

The regression coefficient of x on y is

$$\frac{\Sigma XY}{\Sigma Y^2} = -\frac{26}{20} = -1.3$$

The equation of line of regression of x on y is

$$x - \bar{x} = \frac{\Sigma XY}{\Sigma Y^2}(y - \bar{y}) \qquad \left(\frac{\Sigma XY}{\Sigma Y^2} = r\frac{\sigma_x}{\sigma_y}\right)$$

$\Rightarrow \qquad x - 6 = -1.3\,(y - 8) \quad \Rightarrow \quad x = -1.3y + 16.4$

Hence, the equation of the line of regression of y on x is

$$y = -0.65x + 11.9$$

The equation of the line of regression of x on y is

$$x = -1.3y + 16.4$$ **Ans.**

59.16 MULTIPLE REGRESSION

We know that the production of wheat depends not only on the amount of rain fall x_1 but also on the fertilizer x_2, pestisides x_3, quality of seeds x_4, quality of soil x_5 etc. In a multiple regression the dependent variable is a function of more than one independent variable.

Linear regression is a linear relationship between y and x_1, x_2, x_3.....

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + ...$$

In multiple non-linear regression equation is not linear; *for example*

$$y = a_0 + a_1 x^{\alpha} + a_2 x^{\beta} + a_3 x^{\gamma} + ...$$

NON LINEAR RELATIONSHIP

Example 28. *Fit a non linear relationship between the following data :*

x	1	2	3	4
y	1.7	1.8	2.3	3.2

(U.P. , III Semester, June 2009)

Solution. Here, we have

	x	y	x^2	xy	x^3	x^4	x^2y
	1	1.7	1.	1.7	1	1	1.7
	2	1.8	4	3.6	8	16	7.2
	3	2.3	9	6.9	27	81	20.7
	4	3.2	16	12.8	64	256	51.2
Total	10	9.0	30	25.0	100	354	80.8

Let the non linear relationship is $y = a_0 + a_1 x + a_2 x^2$

Normal equations are

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

Substituting the values of Σy, n, Σx, Σx^2 etc. in these equations, we get

$$9 = 4a_0 + 10a_1 + 30a_2$$
$$25 = 10a_0 + 30a_1 + 100a_2$$
$$80.8 = 30a_0 + 100a_1 + 354a_2$$

Solving these equations, we get $a_0 = 2$, $a_1 = -0.5$ and $a_2 = 0.2$

Then the non-linear relationship is $y = 2 - 0.5x + 0.2x^2$ **Ans.**

59.17 ERROR OF PREDICTION

The deviation of the predicted value from the observed value is known as the standard error of prediction. It is given by

$$E_{yx} = \sqrt{\frac{\sum (y - y_r)^2}{n}}$$

where y is the actual value and y_r the predicted value.

Example 29. *Prove that* (*i*) $E_{yx} = \sigma_y.\sqrt{1-r^2}$ (*ii*) $E_{xy} = \sigma_x\sqrt{1-r^2}$

Solution. The equation of the line of regression of y on x is

$$y-\bar{y} = r\frac{\sigma_y}{\sigma_x}(x-\bar{x}), \qquad y_r = \bar{y} + r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

So, $$E_{yx} = \sqrt{\frac{\sum(y-y_r)^2}{n}} = \left[\frac{1}{n}\sum\left\{y-\bar{y}-r\frac{\sigma_y}{\sigma_x}(x-\bar{x})\right\}^2\right]^{1/2}$$

$$= \left[\frac{1}{n}\cdot\Sigma\left\{(y-\bar{y})^2 + \frac{r^2\sigma_y^2}{\sigma_x^2}(x-\bar{x})^2 - \frac{2r\sigma_y}{\sigma_x}(x-\bar{x})(y-\bar{y})\right\}\right]^{1/2}$$

$$= \left[\sum\frac{(y-\bar{y})^2}{n} + r^2\frac{\sigma_y^2}{\sigma_x^2}\sum\frac{(x-\bar{x})^2}{n} - 2r\frac{\sigma_y}{\sigma_x}\sum\frac{(x-\bar{x})(y-\bar{y})}{n}\right]^{1/2}$$

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

$$= \sigma_y\sqrt{1-r^2}$$ **Proved.**

(*ii*) Similarly (*ii*) may be proved.

Example 30. *Find the standard error of estimate of y on x for the data given below:*

x	*1*	*3*	*4*	*6*	*8*	*9*	*11*	*14*
y	*1*	*2*	*4*	*4*	*5*	*7*	*8*	*9*

Solution. The equation of the line of regression of y on x is

$$y = \frac{7}{11}x + \frac{6}{11}. \quad \text{So} \quad y_r = \frac{7x}{11} + \frac{6}{11}$$

S.No	x	y	y_r	$(y-y_r)$	$(y-y_r)^2$
1	1	1	$\frac{13}{11}$	$-\frac{2}{11}$	$\frac{4}{121}$
2	3	2	$\frac{27}{11}$	$-\frac{5}{11}$	$\frac{25}{121}$
3	4	4	$\frac{34}{11}$	$\frac{10}{11}$	$\frac{100}{121}$
4	6	4	$\frac{48}{11}$	$-\frac{4}{11}$	$\frac{16}{121}$
5	8	5	$\frac{62}{11}$	$-\frac{7}{11}$	$\frac{49}{121}$
6	9	7	$\frac{69}{11}$	$\frac{8}{11}$	$\frac{64}{121}$
7	11	8	$\frac{83}{11}$	$\frac{5}{11}$	$\frac{25}{121}$
8	14	9	$\frac{104}{11}$	$-\frac{5}{11}$	$\frac{25}{121}$
					$\Sigma(y-y_r)^2 = \frac{308}{121}$

$$E_{yx} = \sqrt{\frac{\sum(y-y_r)^2}{n}} = \sqrt{\frac{308}{121\times 8}} = \sqrt{\frac{7}{22}} = 0.564$$ **Ans.**

59.18 RELATION BETWEEN REGRESSION ANALYSIS AND CORRELATION ANALYSIS

Sr. No.	Correlation Analysis	Regression Analysis
1.	The relationship between two variables is given by the coefficient of correlation.	1. In this case some points are stepped up and some are stepped down for making an average value.
2.	It is a measure of direction and degree of relationship between x and y.	2. b_{xy} and b_{yx} are mathematical measure of average relationship between the two variables.
3.	Here, $r_{xy} = r_{yx}$	3. $b_{xy} \neq b_{yx}$
4.	It does not reflect upon the nature of variable (dependent or independent variable).	4. It indicates which is dependent variable and which is independent variable
5.	It does not imply cause and effect relationship between the variables.	5. It indicates the cause and effect relationship between the variables.
6.	It is a relative measure and have no units.	6. It is an absolute measure.
7.	It indicates the degree of association.	7. It is used to forecast the nature of the dependent variable when the value of independent variable is given.
8.	It is confined to the study of linear relationship.	8. It has not only application of linear relationship but non-linear relationship also.

EXERCISE 59.2

1. Find the regression line of y on x for the data :

x	1	4	2	3	5
y	3	1	2	5	4

Ans. $y = 2.7 + 0.1x$

2. Compute the regression line of x on y for the following data :

x	2	4	6	8	10
y	12	10	8	6	4

Ans. $x = 1 + 3yy$

3. Compute the regression line of y on x for the following data :

x	1	2	3	4	5	6
y	2	2	2	2	2	2

Ans. $y = 6 - x$

4. Find the regression lines of from the given data :

x	1	2	3	4	5	6	7	8	9	10
y	10	12	16	28	25	36	41	49	40	50

Ans. $x = 0.2\,y - 0.64$, $y = 4.69x + 4.9$

5. Find the equations to the lines of regression and the coefficient of correlation for the following data:

x	2	4	5	6	8	11
y	18	12	10	8	7	5

Ans. $y - 10 = -1.34\,(x - 6)$, $x - 6 = -0.632\,(y - 10)$, $r = -0.92$

6. The following marks have been obtained by a class of students in statistics.

Paper I	80	45	55	56	58	60	65	68	70	75	85
Paper II	81	56	50	48	60	62	64	65	70	74	90

Compute the coefficient of correlation for the above data. Find the lines of regression.

Ans. $r = 0.918$, $y - 65.45 = 0.981\,(x - 65.18)$, $x - 65.18 = 0.859\,(y - 65.45)$

7. 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	165

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

Find the appropriate regression eqaution and use it to estimate the blood pressure of a man whose age is 45. **Ans.** $y = 0.94\,x + 92.26$, Blood pressure = 134.56

8. The following results were obtained from lineups in Applied Mechanics and Engineering Mathematics in an examination :

	Applied Mechanics(x)	*Engg. Maths.*(y)
Mean	47.5	39.5
Standard deviation	16.8	10.8

$r = 0.95$

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

Ans. $y = 0.611x + 10.5$, $x = 1.478\,y - 1.143$, $y = 28.83$

9. If two regression coefficients are 0.8 and 0.2, what would be the value of coefficient of correlation?

Ans. $r = 0.4$

10. The regression equation are : $7x - 16y + 9 = 0$, $5y - 4x - 3 = 0$, find $\bar{x}, \bar{y}$ and r.

(*AMIE, Winter 2003*) **Ans.** $\bar{x} = -\frac{3}{29}$, $\bar{y} = \frac{15}{29}$, $r = \frac{3}{4}$

11. The following regression equations and variances are obtained from a correlation table :

$20x - 9y - 107 = 0$, $4x - 5y + 33 = 0$, variance of $x = 9$.

Find (*i*) the mean values of x and y, (*ii*) the standard deviation of y. (*A.M.I.E., Winter 2000*)

Ans. $\bar{x} = 13, \bar{y} = 17, \sigma_y = 4.$

12. Two random variables have the least square regression lines with equation $3x + 2y = 26$ and $6x + y = 31$. Find mean values and correlation coefficient between x and y.

Ans. $\bar{x} = 4, \bar{y} = 7, r = 0.5$

13. The regression equations of two variables x and y are $x = 0.7y + 5.2$, $y = 0.3x + 2.8$. Find the means of the variables and the coefficient of correlation between them.

Ans. $r = 0.7395$, $\bar{x} = -0.1034$, $\bar{y} = 0.5172$

14. Two lines of regression are given by $x + 2y = 5$ & $2x + 3y = 8$.

Calculate: (*i*) mean values of x and y

(*ii*) the coefficient of correlation

(*iii*) the ratio of the regression coefficients. **Ans.** $\bar{x} = 4$, $\bar{y} = 7$, $r = -0.5$

15. Fill in the blanks :

(*a*) Arithmetic mean of the coefficients of regression is .....than the coefficient of correlation.

(*A.M.I.E., Summer 2000*) **Ans.** greater

(*b*) If two regression lines concide then the coefficient of correlation is .....

(*A.M.I.E., Winter 2000*) **Ans.** ±1

CHAPTER

60

CORRELATION AND MULTIPLE REGRESSION ANALYSIS

60.1 INTRODUCTION

So for we have considered correlation between two variables only. But often it is necessary to find correlation between three or more variables.

For example; Crops are influenced not only by rainfall but of different fertilizers used.

Correlation between crops rainfall and fertilizer is the multiple correlation.

60.2 MULTIPLE CORRELATION

In multiple correlation we study three or more variables at a time.

In multiple correlation the effect of all the independent variables on a dependent variables is studied.

Let three variables be x_1, x_2 and x_3.

$R_{1.23}$ = Multiple correlation coefficient with x_1 as dependent variable and x_2 and x_3 as independent variables.

$R_{2.13}$ = Multiple correlation coefficient with x_2 as dependent variable and x_1 and x_3 as independent variables.

$R_{3.12}$ = Multiple correlation coefficient with x_3 as dependent variable and x_1 and x_2 as independent variables.

60.3 FORMULAE FOR THE CALCULATION OF MULTIPLE CORRELATION COEFFICIENT

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}\;; \quad R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{13}^2}}\;; \quad R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{12}^2}}$$

60.4 PROPERTIES OF MULTIPLE CORRELATION

(1) Its value lies between 0 and 1. (2) If $R_{1.23} = 0$, then $r_{12} = 0$ and $r_{13} = 0$

(3) $R_{1.23} \geq r_{12}$ and $R_{1.23} \geq r_{13}$ Q (4) $R_{1.23} = R_{1.32}$

Coefficient of Multiple correlation between four Variables :

$$R_{1.234} = \sqrt{1 - \left(1 - r_{14}^2\right)\left(1 - r_{12.3}^2\right)\left(1 - r_{12.34}^2\right)}$$

Example 1. *A simple correlation coefficient between quantity of production of wheat x_1, fertilizer (x_2) and rainfall (x_3) are given*

$r_{12} = 0.4, \quad r_{13} = 0.5$ and $r_{23} = 0.6$

Find the coefficient of multiple correlation $R_{1.23}$.

Solution. Here, we have

$$r_{12} = 0.4, \quad r_{13} = 0.5 \quad \text{and} \quad r_{23} = 0.6$$

We know that

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.4)^2 + (0.5)^2 - 2(0.4)\,(0.5)\,(0.6)}{1 - (0.6)^2}}$$

$$= \sqrt{\frac{0.16 + 0.25 - 0.24}{1 - 0.36}} = \sqrt{\frac{0.17}{0.64}} = \sqrt{0.2656} = 0.515 \qquad \textbf{Ans.}$$

Example 2. *If* $r_{12} = 0.6$. $r_{23} = 0.35$ *and* $r_{31} = 0.4$ *then find* $R_{3.12}$.

Solution. We have

$$r_{12} = 0.6, \qquad r_{23} = 0.35, \qquad \text{and} \quad r_{31} = 0.4$$

We know that

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{12}^2}} = \sqrt{\frac{0.4^2 + (0.35)^2 - 2(0.6)\,(0.4)\,(0.35)}{1 - (0.6)^2}}$$

$$= \sqrt{\frac{0.16 + 0.1225 - 0.168}{1 - 0.36}} = \sqrt{\frac{0.1145}{0.64}} = \sqrt{0.1789} = 0.423 \qquad \textbf{Ans.}$$

Example 3. *If* $r_{12} = 0.25$, $r_{13} = 0.35$ *and* $r_{23} = 0.45$ *then find* $R_{2.13}$.

Solution. Here, we have

$$r_{12} = 0.25, \qquad r_{13} = 0.35, \qquad \text{and} \quad r_{23} = 0.45$$

We know that

$$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{13}^2}} = \sqrt{\frac{(0.25)^2 + (0.45)^2 - 2(0.25)\,(0.35)\,(0.45)}{1 - (0.35)^2}}$$

$$= \sqrt{\frac{0.0625 + 0.2025 - 0.07875}{1 - 0.1225}} = \sqrt{\frac{0.18625}{0.8775}} = \sqrt{0.2123} = 0.461 \qquad \textbf{Ans.}$$

Example 4. *Given the following data :*

x_1	3	5	6	8	12	14
x_2	16	10	7	4	3	2
x_3	90	72	54	42	30	12

Compute the coefficient of linear multiple correlation of x_3 *on* x_1 *and* x_2.

Solution. Here $N = 6$; $\bar{x}_1 = \frac{\Sigma x_1}{6} = \frac{48}{6} = 8, \quad \bar{x}_2 = \frac{\Sigma x_2}{6} = \frac{42}{6} = 7, \quad \bar{x}_3 = \frac{\Sigma x_3}{6} = \frac{300}{6} = 50$

We have to compute the values of r_{13}, r_{23} and r_{12}.

$X_1 = x_1 - \bar{x}_1$			$X_2 = x_2 - \bar{x}_2$			$X_3 = x_3 - \bar{x}_3$					
x_1	X_1	X_1^2	x_2	X_2	X_2^2	x_3	X_3	X_3^2	X_1X_2	X_1X_3	X_2X_3
3	–5	25	16	+9	81	90	+40	1600	–45	–200	+360
5	–3	9	10	+3	9	72	+22	484	–9	–66	+66
6	–2	4	7	0	0	54	+4	16	0	–8	0
8	0	0	4	–3	9	42	–8	64	0	0	+24
12	+4	16	3	–4	16	30	–20	400	–16	–80	+80
14	+6	36	2	–5	25	12	–38	1444	–30	–228	+190
		ΣX_1^2 = 90			ΣX_2^2 = 140			ΣX_3^2 = 4008	ΣX_1X_2 = –100	ΣX_1X_3 = – 582	ΣX_2X_3 = 720

Now, $r_{12} = \dfrac{\Sigma X_1X_2}{\sqrt{\Sigma X_1^2 \times \Sigma X_2^2}} = \dfrac{-100}{\sqrt{90 \times 140}} = \dfrac{-100}{\sqrt{12600}} = -0.89$

Also, $r_{13} = \dfrac{\Sigma X_1X_3}{\sqrt{\Sigma X_1^2 \times \Sigma X_3^2}} = \dfrac{-582}{\sqrt{90 \times 4008}} = -0.97$

Again $r_{23} = \dfrac{\Sigma X_2X_3}{\sqrt{\Sigma X_2^2 \times \Sigma X_3^2}} = \dfrac{720}{\sqrt{140 \times 4008}} = 0.96$

We know that

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}\, r_{13}\, r_{23}}{1 - r_{12}^2}} \qquad \text{... (1)}$$

Substituting the values of r_{12}, r_{13} and r_{23} in (1), we get

$$R_{3.12} = \sqrt{\frac{(-0.97)^2 + (0.96)^2 - 2(-0.89)\,(-0.97)\,(0.96)}{1 - (-0.89)^2}}$$

$$= \sqrt{\frac{0.9409 + 0.9216 - 1.66}{1 - 0.7921}} = \sqrt{\frac{0.2025}{0.2079}} = \sqrt{0.9740} = 0.987 \qquad \textbf{Ans.}$$

EXERCISE 60.1

1. If $r_{12} = 0.59$, $r_{13} = 0.46$, and $r_{23} = 0.77$ then find $R_{1.23}$. **Ans.** 0.416
2. If $r_{12} = 0.5$, $r_{13} = 0.6$, and $r_{23} = 0.7$ then find $R_{2.13}$. **Ans.** 0.707
3. If $r_{12} = 0.6$, $r_{13} = 0.7$, and $r_{23} = 0.65$ then find $R_{3.12}$. **Ans.** 0.757
4. If $r_{12} = 0.8$, $r_{13} = -0.5$, and $r_{23} = 0.40$ then prove that $R_{1.23} = R_{1.32}$
5. If $r_{12} = 0.45$, $r_{13} = 0.32$, and $r_{23} = 0.61$ then find $R_{1.23}$. **Ans.** 0.339

60.5 MULTIPLE REGRESSION ANALYSIS

We have considered two types of regression equations one of x on y and the other of y on x.

Multiple regression analysis represents an extension of two variables to three or more variables.

We take x_1 as dependent variable and x_2 and x_3 as independent variable.

60.6 PURPOSE OF MULTIPLE REGRESSION

(1) From the regression equation to find out the estimate of dependent variable from two or more independent variables.

(2) To find out the error in the estimate.

(3) To derive a measure of proportion of variance in the dependent variable from the independent variables.

60.7 REGRESSION EQUATION OF THREE VARIABLES

$$X_1 = a + b_{12.3}\, x_2 + b_{13.2}\, x_3 \qquad \text{... (A)}$$

Normal equation of multiple regression equation is

$$S = \Sigma\,(x_1 - X_1)^2 \qquad \text{... (B)}$$

Putting the value of X_1 from (A) in (B), we get

$$S = \Sigma\,(x_1 - a - b_{12.3}\, x_2 - b_{13.2}\, x_3)^2$$

Differentiating partially above equation w.r.t. a, $b_{12.3}$ and $b_{13.2}$, we get

$$\frac{\partial S}{\partial a} = \Sigma\,(x_1 - 1 - b_{12.3}\, x_2 - b_{13.2}\, x_3) = 0 \qquad \text{... (1)}$$

$$\frac{\partial S}{\partial b_{12.3}} = \Sigma\, x_2\,(x_1 - a - b_{12.3}\, x_2 - b_{13.2}\, x_3) = 0 \qquad \text{... (2)}$$

$$\frac{\partial S}{\partial b_{13.2}} = \Sigma\, x_3\,(x_1 - a - b_{12.3}\, x_2 - b_{13.2}\, x_3) = 0 \qquad \text{... (3)}$$

Equation (1) can be rewritten as

$$\Sigma x_1 - \Sigma a - b_{12.3}\, \Sigma x_2 - b_{13.2}\, \Sigma x_3 = 0$$

since $\Sigma x_1 = \Sigma(X_1 - \bar{X}_1) = 0$, $\Sigma x_2 = \Sigma(X_2 - \bar{X}_2) = 0$, $\Sigma x_3 = \Sigma(X_3 - \bar{X}_3) = 0$

[Sum of the deviations from the mean = 0]

Therefore from (1), $a = 0$

Putting the value of $a = 0$ in (2) and (3), we get

$$\Sigma x_1 x_2 - b_{12.3}\, \Sigma x_2^2 - b_{13.2}\, \Sigma x_2 x_3 = 0 \qquad \text{... (4)}$$

$$\Sigma x_1 x_3 - b_{12.3}\, \Sigma x_2 x_3 - b_{13.2}\, \Sigma x_3^2 = 0 \qquad \text{... (5)}$$

On solving (4) and (5), we get the values of $b_{12.3}$, and $b_{13.2}$.

On putting the values of a, $b_{12.3}$ and $b_{13.2}$ in (A), we get the required regression equation.

Similarly

$$x_2 = b_{21.3}\, x_1 + b_{23.1}\, x_3$$
$$x_3 = b_{31.2}\, x_1 + b_{32.1}\, x_2$$

Second Method :

On putting the values of $\Sigma x_1 x_2$ etc. in (4) and (5), we get

$$r_{12} = \frac{\Sigma x_1 x_2}{\sigma_1 \sigma_2} \quad \Rightarrow \Sigma x_1 x_2 = r_{12}\, \sigma_1 \sigma_2 \text{ etc.}$$

$$r_{12}\, \sigma_1 \sigma_2 = b_{12.3}\, r_{23}\, \sigma_2 \sigma_3 + b_{13.2}\, \sigma_3^2 \qquad \text{... (6)}$$

$$r_{13}\, \sigma_1 \sigma_3 = b_{12.3}\, r_{23}\, \sigma_2 \sigma_3 + b_{13.2}\, \sigma_3^2 \qquad \text{... (7)}$$

where r_{ij} = coefficient of correlation between x_i and x_j.

Solving (6) and (7), we get

$$b_{12.3} = \begin{vmatrix} r_{12}\,\sigma_1 & r_{23}\,\sigma_3 \\ r_{13}\,\sigma_1 & \sigma_3 \end{vmatrix} \div \begin{vmatrix} \sigma_2 & r_{23}\,\sigma_3 \\ r_{23}\,\sigma_2 & \sigma_3 \end{vmatrix} = \frac{\dfrac{-\sigma_1}{\sigma_2}\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} = -\frac{\sigma_1}{\sigma_2}\frac{\Delta_{12}}{\Delta_{11}} \qquad \text{... (8)}$$

$$\text{and} \quad b_{13.2} = \frac{\dfrac{-\sigma_1}{\sigma_3}\begin{vmatrix} 1 & r_{12} \\ r_{23} & r_{13} \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} = -\frac{\sigma_1}{\sigma_3}\frac{\Delta_{13}}{\Delta_{11}} \qquad \ldots (9)$$

Where Δ_{ij} is the co-factor of the element in the ith row and jth column in the determinant.

$$\Delta = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

Hence, on substituting the values of $b_{12.3}$ and $b_{13.2}$ the equation to the regression plane of x_1 on x_2 and x_3 is

$$x_1 = \left[-\frac{\sigma_1}{\sigma_2}\frac{\Delta_{12}}{\Delta_{11}}\right]x_2 + \left[-\frac{\sigma_1}{\sigma_3}\frac{\Delta_{13}}{\Delta_{11}}\right]x_3 \qquad \ldots (10)$$

The above equation can also be written as :

$$x_1 = \frac{\sigma_1}{\sigma_2}\left[\frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\right]x_2 + \frac{\sigma_1}{\sigma_3}\left[\frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}\right]x_3$$

Similarly,

$$x_2 = \frac{\sigma_2}{\sigma_3}\left[\frac{r_{23} - r_{12}\, r_{13}}{1 - (r_{13})^2}\right]x_3 + \frac{\sigma_2}{\sigma_1}\left[\frac{r_{12} - r_{23}.r_{13}}{1 - (r_{31})^2}\right]x_1, \quad x_3 = \frac{\sigma_3}{\sigma_2}\left[\frac{r_{23} - r_{12}\, r_{13}}{1 - (r_{12})^2}\right]x_2 + \frac{\sigma_3}{\sigma_1}\left[\frac{r_{13} - r_{23}\, r_{12}}{1 - (r_{12})^2}\right]x_1$$

Standard Error of the estimate for multiple regression and multiple correlation

The standard error of the estimate X_1, X_2 and X_3 is

$$S_{1.23} = \sqrt{\frac{\Sigma(X_1 - Y_1)^2}{N - 3}}$$

where $S_{1.23}$ is standard error of the estimate of X_1 on X_2 and X_3.

X_1 is the original value of X and Y_1 is the estimated value on the basis of the regression equation.

Standard error in terms of multiple correlation

$$S_{1.23} = \sigma_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2 r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

Example 5. *If* $r_{12} = 0.6$, $r_{13} = 0.8$, $r_{23} = 0.3$
$\sigma_1 = 8$, $\sigma_2 = 9$, $\sigma_3 = 5$
Determine regression equation of x_1 *on* x_2 *and* x_3.

Solution. Let x_1, x_2 and x_3 be the respective deviations from the means of X_1, X_2 and X_3 series.

Regression equation of X_1 on X_2 and X_3 is

$$x_1 = b_{12.3}\, x_2 + b_{13.2}\, x_3 \qquad \ldots (1)$$

Now $$b_{12.3} = \frac{\sigma_1}{\sigma_2} \times \left[\frac{r_{12} - r_{13}\, r_{23}}{1 - r_{23}^2}\right]$$

$$= \frac{8}{9}\left[\frac{0.6 - 0.8 \times 0.3}{1 - (0.3)^2}\right] = \frac{8}{9}\left[\frac{0.6 - 0.24}{1 - 0.09}\right] = \frac{8}{9}\left[\frac{0.36}{0.91}\right] = \frac{2.88}{8.19} = 0.35$$

$$b_{13.2} = \frac{\sigma_1}{\sigma_3} \times \left[\frac{r_{13} - r_{12}\, r_{23}}{1 - r_{23}^2}\right] = \frac{8}{5}\left[\frac{0.8 - 0.6 \times 0.3}{1 - (0.3)^2}\right]$$

$$= \frac{8}{5}\left[\frac{0.8-0.18}{1-0.09}\right] = \frac{8}{5}\left[\frac{0.62}{0.91}\right] = \frac{4.96}{4.55} = 1.09$$

Substituting the values of $b_{12.3}$ and $b_{13.2}$ in the equation (1), we get

$$x_1 = 0.35\, x_2 + 1.09\, x_3$$ **Ans.**

Example 6. *If* $r_{12} = 0.75,\ r_{13} = 0.65,\ r_{23} = 0.55,\ \sigma_1 = 9,\ \sigma_2 = 7,\ \sigma_3 = 4$
Determine the regression equation of X_2 *on* X_1 *and* X_3.

Solution. The regression equation of X_2 on X_1 and X_3 is given by

$$x_2 = b_{21.3}\, x_1 + b_{23.1}\, x_3 \quad \ldots (1)$$

We know that

$$b_{21.3} = \frac{\sigma_2}{\sigma_1}\left[\frac{r_{12} - r_{23}\, r_{13}}{1-r_{31}^2}\right] = \frac{7}{9}\left[\frac{0.75-0.55\times 0.65}{1-(0.65)^2}\right] = \frac{7}{9}\left[\frac{0.75-0.3575}{1-0.4225}\right]$$

$$= \frac{7}{9}\left[\frac{0.3925}{0.5775}\right] = \frac{2.7475}{5.1975} = 0.5286$$

and

$$b_{23.1} = \frac{\sigma_2}{\sigma_3}\left[\frac{r_{23} - r_{12}\, r_{13}}{1-r_{13}^2}\right] = \frac{7}{4}\left[\frac{0.55-0.75\times 0.65}{1-(0.65)^2}\right] = \frac{7}{4}\left[\frac{0.55-0.4875}{1-0.4225}\right]$$

$$= \frac{7}{4}\left[\frac{0.0625}{0.5775}\right] = \frac{0.4375}{2.3100} = 0.1894$$

Substituting the values of $b_{12.3}$ and $b_{23.1}$ in (1), we get

$$x_2 = 0.5286\, x_1 + 0.1894\, x_3$$ **Ans.**

Example 7. *If* $\sigma_1 = 3,\ \sigma_2 = 2.5,\ \sigma_3 = 3.5$
$r_{12} = 0.3,\ r_{13} = 0.5,\ r_{23} = 0.4$
Find the regression equation of x_3 *on* x_1 *and* x_2.

Solution. Here, we have

$$\sigma_1 = 3,\quad \sigma_2 = 2.5,\quad \sigma_3 = 3.5$$
$$r_{12} = 0.3,\quad r_{13} = 0.5,\quad r_{23} = 0.4$$

The regression equation of x_3 on x_1 and x_2 is

$$x_3 = b_{31.2}\, x_1 + b_{32.1}\, x_2 \quad \ldots (1)$$

We know that

$$b_{31.2} = \frac{\sigma_3}{\sigma_1}\left[\frac{r_{13} - r_{23}\, r_{12}}{1-r_{12}^2}\right] = \frac{3.5}{3}\left[\frac{0.5-0.4\times 0.3}{1-(0.3)^2}\right]$$

$$= \frac{3.5}{3}\left[\frac{0.5-0.12}{1-0.09}\right] = \frac{3.5}{3}\left[\frac{0.38}{0.91}\right] = \frac{1.33}{2.73} = 0.487$$

$$b_{32.1} = \frac{\sigma_3}{\sigma_2}\left[\frac{r_{23} - r_{12}\, r_{13}}{1-r_{12}^2}\right] = \frac{3.5}{2.5}\left[\frac{0.4-0.3\times 0.5}{1-(0.3)^2}\right]$$

$$= \frac{3.5}{2.5}\left[\frac{0.4-0.15}{1-0.09}\right] = \frac{3.5}{2.5}\left[\frac{0.25}{0.91}\right] = \frac{0.875}{2.275} = 0.385$$

Substituting the values of $b_{31.2}$ and $b_{32.1}$ in (1), we get

$$x_3 = 0.487\, x_1 + 0.385\, x_2$$ **Ans.**

Example 8. *Find the multiple regression equation of x_1 on x_2 and x_3 from the data given below :*

X_1	3	5	6	8	12	10
X_2	10	10	5	7	5	2
X_3	20	25	15	16	15	2

Solution. The regression equation of X_1 on X_2 and X_3 is given by

$$X_1 = a_{1.23} + b_{12.3}\ X_2 + b_{13.2} X_3 \qquad \text{... (A)}$$

The three normal equations for getting the values of $a_{1.23}$, $b_{12.3}$ and $b_{13.2}$ are

$$\Sigma X_1 = N\, a_{1.23} + b_{12.3}\, \Sigma X_2 + b_{13.2}\, \Sigma X_3$$

$$\Sigma X_1 X_2 = a_{1.23}\, \Sigma X_2 + b_{12.3}\ \Sigma X_2^2 + b_{13.2}\, \Sigma X_2 X_3$$

$$\Sigma X_1 X_3 = a_{1.23}\, \Sigma X_3 + b_{12.3}\, \Sigma X_2 X_3 + b_{13.2}\ \Sigma X_3^2$$

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_2^2	X_3^2
3	10	20	30	60	200	100	400
5	10	25	50	125	250	100	625
6	5	15	30	90	75	25	225
8	7	16	56	128	112	49	256
12	5	15	60	180	75	25	225
10	2	2	20	20	4	4	4
ΣX_1 = 44	ΣX_2 = 39	ΣX_3 = 93	$\Sigma X_1 X_2$ = 246	$\Sigma X_1 X_3$ = 603	$\Sigma X_2 X_3$ = 716	ΣX_2^2 = 303	ΣX_3^2 = 1735

Substituting the values in normal equations, we get

$$6\ a_{1.23} + 39\ b_{12.3} + 93\ b_{13.2} = 44 \qquad \text{... (1)}$$

$$39\ a_{1.23} + 303\ b_{12.3} + 716\ b_{13.2} = 246 \qquad \text{... (2)}$$

$$93\ a_{1.23} + 716\ b_{12.3} + 1735\ b_{13.2} = 603 \qquad \text{... (3)}$$

Multiplying (1) by 13 and (2) by 2, we get

$$78\ a_{1.23} + 507\ b_{12.3} + 1209\ b_{13.2} = 572 \qquad \text{... (4)}$$

$$78\ a_{1.23} + 606\ b_{12.3} + 1432\ b_{13.2} = 492 \qquad \text{... (5)}$$

Subtracting (4) from (5), we get

$$99\ b_{12.3} + 223\ b_{13.2} = -80 \qquad \text{... (6)}$$

Multiplying (2) by 31 and (3) by 13, we get

$$1209\ a_{1.23} + 9393\ b_{12.3} + 22196\ b_{13.2} = 7626 \qquad \text{... (7)}$$

$$1209\ a_{1.23} + 9308\ b_{12.3} + 22555\ b_{13.2} = 7839 \qquad \text{... (8)}$$

Subtracting (8) from (7), we get

$$85\ b_{12.3} - 359\ b_{13.2} = -213 \qquad \text{... (9)}$$

Multiplying (6) by 85 and (9) by 99, we get

$$8415\ b_{12.3} + 18955\ b_{13.2} = -6800 \qquad \text{... (10)}$$

$$8415\ b_{12.3} - 35541\ b_{13.2} = -21087 \qquad \text{... (11)}$$

Subtracting (11) from (10), we get

$$54496\, b_{13.2} = 14287 \quad \Rightarrow \quad b_{13.2} = 0.2621$$

Putting the value of $b_{13.2}$ in (6), we get

$$99\, b_{12.3} + 223\ (0.2621) = -80 \quad \Rightarrow \quad 99\, b_{12.3} = -80 - 58.4483$$

$$\Rightarrow \qquad b_{12.3} = \frac{-138.4483}{99} = -1.3984$$

Putting the values of $b_{12.3}$ and $b_{13.2}$ in (1), we get

$$6\, a_{1.23} + 39\ (-1.3984) + 93\ (0.2621) = 44$$

$$\Rightarrow \qquad 6\, a_{1.23} - 54.5376 + 24.3753 = 44$$

$$\Rightarrow \qquad 6\, a_{1.23} = 44 + 54.5493 - 24.3846 \quad \Rightarrow \quad a_{1.23} = \frac{74.1647}{6} = 12.3608$$

Substituting the values $a_{1.23} = 12.3608$, $b_{12.3} = -1.3984$ and $b_{13.2} = 0.2621$ in equation (A), we get

$$X_1 = 12.3608 - 1.3984\, X_2 + 0.2621\, X_3$$

which is the required regression equation of X_1 on X_2 and X_3. **Ans.**

Example 9. *If $r_{12} = 0.60$, $r_{13} = 0.70$, $r_{23} = 0.65$, and $\sigma_1 = 1.0$, find $S_{1.23}$.*

Solution. Here, we have

$$r_{12} = 0.60, \quad r_{13} = 0.70, \quad r_{23} = 0.65, \quad \text{and} \quad S_1 = 1.0$$

We know that

$$S_{1.23} = \sigma_1 \times \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2 r_{12}\, r_{13}\, r_{23}}{1 - r_{23}^2}}$$

$$= 1.0 \times \sqrt{\frac{1 - (0.60)^2 - (0.70)^2 - (0.65)^2 + 2\,(0.60)\,(0.70)\,(0.65)}{1 - (0.65)^2}}$$

$$= 1.0 \times \sqrt{\frac{1 - 0.36 - 0.49 - 0.4225 + 0.546}{1 - 0.4225}}$$

$$= \sqrt{\frac{0.2735}{0.5775}} = \sqrt{0.4736} = 0.6882$$ **Ans.**

EXERCISE 60.2

1. If $r_{12} = 0.8$, $r_{13} = 0.7$, $r_{23} = 0.6$, $\sigma_1 = 10$, $\sigma_2 = 8$, $\sigma_3 = 5$, then find the regression eqaution of x_1 on x_2 and x_3. **Ans.** $x_1 = 0.742\, x_2 + 0.6875\, x_3$

2. If $\sigma_1 = 3$, $\sigma_2 = 4$, $\sigma_3 = 5$, $r_{12} = 0.7$, $r_{23} = 0.4$, $r_{31} = 0.6$, then determine the regression equation of x_1 on x_2 and x_3. **Ans.** $x_1 = 0.41\, x_2 + 0.229\, x_3$

3. If $r_{12} = 0.28$, $r_{23} = 0.49$, $r_{31} = 0.51$, $\sigma_1 = 2.7$, $\sigma_2 = 2.4$, $\sigma_3 = 2.7$, then find the regression eqaution of x_3 on x_1 and x_2. **Ans.** $x_3 = 0.405\, x_1 + 0.424\, x_2$

4. Find the multiple linear regression equation of x_1 on x_2 and x_3 from the data given below:

x_1	2	4	6	8
x_2	3	5	7	9
x_3	4	6	8	10

Ans. $x_1 = 2x_2 - x_3$

CHAPTER

61

PROBABILITY

61.1 PROBABILITY

Probability is a concept which numerically measure the degree of uncertainity and therefore, of certainity of the occurrence of events.

If an event A can happen in m ways, and fail in n ways, all these ways being equally likely to occur, then the probability of the happening of A is

$$= \frac{\text{Number of favourable cases}}{\text{Total number of mutually exclusive and equally likely cases}} = \frac{m}{m+n}$$

and that of its failing is defined as $\frac{n}{m+n}$

If the probability of the happening $= p$
and the probability of not happening $= q$

then $$p+q = \frac{m}{m+n} + \frac{n}{m+n} = \frac{m+n}{m+n} = 1 \text{ or } p+q = 1$$

For instance, on tossing a coin, the probability of getting a head is $\frac{1}{2}$.

61.2 DEFINITIONS

1. **Die :** It is a small cube. Dots are . .. ... :: ::. ::: marked on its faces. Plural of the die is dice. On throwing a die, the outcome is the number of dots on its upper face.
2. **Cards :** A pack of cards consists of four suits *i.e.* Spades, Hearts, Diamonds and Clubs. Each suit consists of 13 cards, nine cards numbered 2, 3, 4, ..., 10, and Ace, a King, a Queen and a Jack or Knave. Colour of Spades and Clubs is black and that of Hearts and Diamonds is red. Kings, Queens, and Jacks are known as *face* cards.
3. **Exhaustive Events or Sample Space :** The set of all possible outcomes of a single performance of an experiment is exhaustive events or sample space. Each outcome is called a sample point. In case of tossing a coin once, $S = (H,T)$ is the *sample space*. Two outcomes Head and Tail constitute an exhaustive event because no other outcome is possible.
4. **Random Experiment :** There are experiments, in which results may be altogether different, even though they are performed under identical conditions. They are known as random experiments. Tossing a coin or throwing a die is random experiment.
5. **Trail and Event :** Performing a random experiment is called a trial and outcome is termed as event. Tossing of a coin is a trial and the turning up of head or tail is an event.
6. **Equally likely events:** Two events are said to be '*equally likely*', if one of them cannot be expected in preference to the other. For instance, if we draw a card from well-shuffled pack, we may get any card, then the 52 different cases are equally likely.
7. **Independent event :** Two events may be *independent*, when the actual happening of one does not influence in any way the probability of the happening of the other.
   Example. The event of getting head on first coin and the event of getting tail on the second