Contents

December

GTU. Winter 2019.

Chap=1, chap.2 → 14 Marks
Chap 3, chap 4 → 14 Marks
Chap=5 → 18 Marks
Chap=6 → 17 Marks
Chap=7 → 7 Marks

70 Marks.

from:- D.G. BORAD

-: Shreenathji Engineering Zone :-

Defated

CHAPTER 4

Correlation and Regression

Chapter Outline

4.1 INTRODUCTION

Correlation and regression are the most commonly used techniques for investigating the relationship between two quantitative variables. *Correlation* refers to the relationship of two or more variables. It measures the closeness of the relationship between the variables. *Regression* establishes a functional relationship between the variables. In correlation, both the variables x and y are random variables, whereas in regression, x is a random variable and y is a fixed variable. The coefficient of correlation is a relative measure whereas the regression coefficient is an absolute figure.

4.2 CORRELATION

Correlation is the relationship that exists between two or more variables. Two variables are said to be correlated if a change in one variable affects a change in the other variable. Such a data connecting two variables is called *bivariate data*. Thus, correlation is a statistical analysis which measures and analyses the degree or extent to which two variables fluctuate with reference to each other. Some examples of such a relationship are as follows:

1. Relationship between heights and weights.
2. Relationship between price and demand of commodity.
3. Relationship between rainfall and yield of crops.
4. Relationship between age of husband and age of wife.

4.3 TYPES OF CORRELATIONS

Correlation is classified into four types:

1. Positive and negative correlations
2. Simple and multiple correlations
3. Partial and total correlations
4. Linear and nonlinear correlations

4.3.1 Positive and Negative Correlations

Depending on the variation in the variables, correlation may be positive or negative.

1. Positive Correlation If both the variables vary in the same direction, the correlation is said to be positive. In other words, if the value of one variable increases, the value of the other variable also increases, or, if value of one variable decreases, the value of the other variable decreases, e.g., the correlation between heights and weights of group of persons is a positive correlation.

Height (cm)	150	152	155	160	162	165
Weight (kg)	60	62	64	65	67	69

2. Negative Correlation If both the variables vary in the opposite direction, correlation is said to be negative. In other words, if the value of one variable increases, the value of the other variable decreases, or, if the value of one variable decreases, the value of the other variable increases, e.g., the correlation between the price and demand of a commodity is a negative correlation.

Price (₹ per unit)	10	8	6	5	4	1
Demand (units)	100	200	300	400	500	600

4.3.2 Simple and Multiple Correlations

Depending upon the study of the number of variables, correlation may be simple or multiple.

1. Simple Correlation When only two variables are studied, the relationship is described as simple correlation, e.g., the quantity of money and price level, demand and price, etc.

2. Multiple Correlation When more than two variables are studied, the relationship is described as multiple correlation, e.g., relationship of price, demand, and supply of a commodity.

4.3.3 Partial and Total Correlations

Multiple correlation may be either partial or total.

1. Partial Correlation When more than two variables are studied excluding some other variables, the relationship is termed as partial correlation.

2. Total Correlation When more than two variables are studied without excluding any variables, the relationship is termed total correlation.

4.3.4 Linear and Nonlinear Correlations

Depending upon the ratio of change between two variables, the correlation may be linear or nonlinear.

1. Linear Correlation If the ratio of change between two variables is constant, the correlation is said to be linear. If such variables are plotted on a graph paper, a straight line is obtained, e.g.,

Milk (l)	5	10	15	20	25	30
Curd (kg)	2	4	6	8	10	12

2. Nonlinear Correlation If the ratio of change between two variables is not constant, the correlation is said to nonlinear. The graph of a nonlinear or curvilinear relationship will be a curve, e.g.,

Advertising expenses (₹ in lacs)	3	6	9	12	15
Sales (₹ in lacs)	10	12	15	15	16

4.4 METHODS OF STUDYING CORRELATION

There are two different methods of studying correlation, (1) Graphic methods (2) Mathematical methods.

Graphic methods are (a) scatter diagram, and (b) simple graph.

Mathematical methods are (a) Karl Pearson's coefficient of correlation, and (b) Spearman's rank coefficient of correlation.

4.5 SCATTER DIAGRAM

The scatter diagram is a diagrammatic representation of bivariate data to find the correlation between two variables. There are various correlationships between two variables represented by the following scatter diagrams.

1. Perfect Positive Correlation If all the plotted points lie on a straight line rising from the lower left-hand corner to the upper right-hand corner, the correlation is said to be perfectly positive (Fig. 4.1).

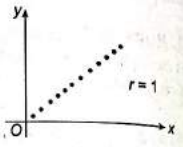

Fig. 4.1

2. Perfect Negative Correlation If all the plotted points lie on a straight line falling from the upper-left hand corner to the lower right-hand corner, the correlation is said to be perfectly negative (Fig. 4.2).

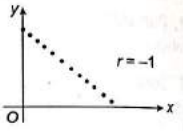

Fig. 4.2

3. High Degree of Positive Correlation If all the plotted points lie in the narrow strip, rising from the lower left-hand corner to the upper right-hand corner, it indicates a high degree of positive correlation (Fig. 4.3).

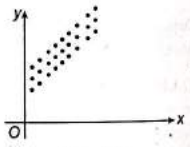

Fig. 4.3

4. High Degree of Negative Correlation If all the plotted points lie in a narrow strip, falling from the upper left-hand corner to the lower right-hand corner, it indicates the existence of a high degree of negative correlation (Fig. 4.4).

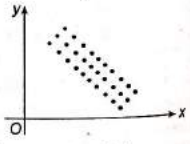

Fig. 4.4

5. No Correlation If all the plotted points lie on a straight line parallel to the x-axis or y-axis or in a haphazard manner, it indicates the absence of any relationship between the variables (Fig. 4.5).

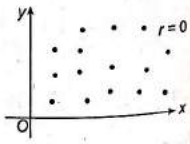

Fig. 4.5

Merits of a Scatter Diagram

1. It is simple and nonmathematical method to find out the correlation between the variables.

2. It gives an indication of the degree of linear correlation between the variables.
3. It is easy to understand.
4. It is not influenced by the size of extreme items.

4.6 SIMPLE GRAPH

A *simple graph* is a diagrammatic representation of bivariate data to find the correlation between two variables. The values of the two variables are plotted on a graph paper. Two curves are obtained, one for the variable x and the other for the variable y. If both the curves move in the same direction, the correlation is said to be positive. If both the curves move in the opposite direction, the correlation is said to be negative. This method is used in the case of a time series. It does not reveal the extent to which the variables are related.

4.7 KARL PEARSON'S COEFFICIENT OF CORRELATION

The coefficient of correlation is the measure of correlation between two random variables X and Y, and is denoted by r.

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

where cov (X, Y) is the covariance of variables X and Y,

σ_X is the standard deviation of variable X,

and σ_Y is the standard deviation of variable Y.

This expression is known as Karl Pearson's coefficient of correlation or Karl Pearson's product-moment coefficient of correlation.

$$\text{cov}(X,Y) = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y})$$

$$\sigma_X = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

$$\sigma_Y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

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

The above expression can be further modified.

Expanding the terms,

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

$$=\frac{\sum xy-\bar{y}\sum x-\bar{x}\sum y+\bar{x}\bar{y}\sum 1}{\sqrt{\sum x^2-2\bar{x}\sum x+\bar{x}^2\sum 1}\sqrt{\sum y^2-2\bar{y}\sum y+\bar{y}^2\sum 1}}$$

$$=\frac{\sum xy-\frac{\sum y}{n}\sum x-\frac{\sum x}{n}\sum y+\frac{\sum x}{n}\frac{\sum y}{n}\cdot n}{\sqrt{\sum x^2-2\frac{\sum x}{n}\sum x+\left(\frac{\sum x}{n}\right)^2 n}\sqrt{\sum y^2-2\frac{\sum y}{n}\sum y+\left(\frac{\sum y}{n}\right)^2 n}}$$

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

4.8 PROPERTIES OF COEFFICIENT OF CORRELATION

1. The coefficient of correlation lies between −1 and 1, i.e., $-1 \le r \le 1$.

Proof Let $\bar{x}$ and $\bar{y}$ be the mean of x and y series and σ_x and σ_y be their respective standard deviations.

Let $\sum\left(\frac{x-\bar{x}}{\sigma_x}\pm\frac{y-\bar{y}}{\sigma_y}\right)^2 \ge 0$ $\left[\because \begin{array}{l}\text{sum of squares of real quantities}\\ \text{cannot be negative}\end{array}\right]$

$$\frac{\sum(x-\bar{x})^2}{\sigma_x^2}+\frac{\sum(y-\bar{y})^2}{\sigma_y^2}\pm\frac{2\sum(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}\ge 0$$

$$n+n\pm 2nr\ge 0$$

$$2n\pm 2nr\ge 0$$

$$2n(1\pm r)\ge 0$$

$$1\pm r\ge 0$$

i.e., $1+r\ge 0$ or $1-r\ge 0$

$r\ge -1$ or $r\le 1$

Hence, the coefficient of correlation lies between −1 and 1, i.e., $-1 \le r \le 1$.

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

Proof Let $d_x=\frac{x-a}{h}$, $d_y=\frac{y-b}{k}$

$x=a+hd_x$, $y=b+kd_y$

where a, b, h (>0) and k(>0) are constants.

$$x=a+hd_x \Rightarrow \bar{x}=a+h\bar{d}_x \Rightarrow x-\bar{x}=h(d_x-\bar{d}_x)$$

$$y=b+kd_y \Rightarrow \bar{y}=b+h\bar{d}_y \Rightarrow y-\bar{y}=k(d_y-\bar{d}_y)$$

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

$$=\frac{\sum h(d_x-\bar{d}_x)\,k(d_y-\bar{d}_y)}{\sqrt{\sum h^2(d_x-\bar{d}_x)^2}\sqrt{\sum k^2(d_y-\bar{d}_y)^2}}$$

$$=\frac{\sum(d_x-\bar{d}_x)(d_y-\bar{d}_y)}{\sqrt{\sum(d_x-\bar{d}_x)^2}\sqrt{(d_y-\bar{d}_y)^2}}$$

$$=r_{d_xd_y}$$

Hence, the correlation coefficient is independent of change of origin and change of scale.

Note Since correlation coefficient is independent of change of origin and change of scale,

$$r=\frac{\sum d_xd_y-\frac{\sum d_x\sum d_y}{n}}{\sqrt{\sum d_x^2-\frac{(\sum d_x)^2}{n}}\sqrt{\sum d_y^2-\frac{(\sum d_y)^2}{n}}}$$

3. Two independent variables are uncorrelated.

Proof If random variables X and Y are independent,

$$\sum(x-\bar{x})(y-\bar{y})=0 \text{ or } \operatorname{cov}(X,Y)=0$$

$$\therefore \quad r=0$$

Thus, if X and Y are independent variables, they are uncorrelated.

Note The converse of the above property is not true, i.e., two uncorrelated variables may not be independent.

Example 1

Calculate the correlation coefficient between x and y using the following data:

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

Solution

$n = 6$

x	y	x^2	y^2	xy
2	18	4	324	36
4	12	16	144	48
5	10	25	100	50
6	8	36	64	48
8	7	64	49	56
11	5	121	25	55
$\Sigma x = 36$	$\Sigma y = 60$	$\Sigma x^2 = 266$	$\Sigma y^2 = 706$	$\Sigma xy = 293$

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

$$= \frac{293 - \frac{(36)(60)}{6}}{\sqrt{266 - \frac{(36)^2}{6}} \sqrt{706 - \frac{(60)^2}{6}}}$$

$$= -0.9203$$

Note Σx, Σy, Σx^2, Σy^2, Σxy can be directly obtained with the help of scientific calculator.

Example 2

Calculate the coefficient of correlation from the following data:

x	12	9	8	10	11	13	7
y	14	8	6	9	11	12	3

Solution

$n = 7$

x	y	x^2	y^2	xy
12	14	144.	196	168
9	8	81	64	72
8	6	64	36	48
10	9	100	81	90
11	11	121	121	121
13	12	169	144	156
7	3	49	9	21
$\Sigma x = 70$	$\Sigma y = 63$	$\Sigma x^2 = 728$	$\Sigma y^2 = 651$	$\Sigma xy = 676$

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

$$= \frac{676 - \frac{(70)(63)}{7}}{\sqrt{728 - \frac{(70)^2}{7}} \sqrt{651 - \frac{(63)^2}{7}}}$$

$$= 0.949$$

Example 3

Calculate the coefficient of correlation for the following data:

x	9	8	7	6	5	4	3	2	1
y	15	16	14	13	11	12	10	8	9

Solution

$n = 9$

x	y	x^2	y^2	xy
9	15	81	225	135
8	16	64	256	128
7	14	49	196	98
6	13	36	169	78
5	11	25	121	55
4	12	16	144	48
3	10	9	100	30
2	8	4	64	16
1	9	1	81	9
$\Sigma x = 45$	$\Sigma y = 108$	$\Sigma x^2 = 285$	$\Sigma y^2 = 1356$	$\Sigma xy = 597$

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

$$= \frac{597 - \frac{(45)(108)}{9}}{\sqrt{285 - \frac{(45)^2}{9}} \sqrt{1356 - \frac{(108)^2}{9}}}$$

$$= 0.95$$

Example 4

Calculate the correlation coefficient between the following data:

x	5	9	13	17	21
y	12	20	25	33	35

Solution

$n = 5$

$$\bar{x} = \frac{\sum x}{n} = \frac{65}{5} = 13$$

$$\bar{y} = \frac{\sum y}{n} = \frac{125}{5} = 25$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
5	12	−8	−13	64	169	104
9	20	−4	−5	16	25	20
13	25	0	0	0	0	0
17	33	4	8	16	64	32
21	35	8	10	64	100	80
$\Sigma x = 65$	$\Sigma y = 125$	$\Sigma(x-\bar{x}) = 0$	$\Sigma(y-\bar{y}) = 0$	$\Sigma(x-\bar{x})^2 = 160$	$\Sigma(y-\bar{y})^2 = 358$	$\Sigma(x-\bar{x})(y-\bar{y}) = 236$

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

$$= \frac{236}{\sqrt{160}\sqrt{358}}$$

$$= 0.986$$

Note Since Σx, Σy, Σx^2, Σy^2, Σxy can be directly obtained with the help of scientific calculator, correlation coefficient can be calculated without using mean.

Example 5

Calculate the correlation coefficient between for the following values of demand and the corresponding price of a commodity:

Demand in Quintals	65	66	67	67	68	69	70	72
Price in rupees per kg	67	68	65	68	72	72	69	71

Solution

Let the demand in quintal be denoted by x and the price in rupees per kg be denoted by y.

$n = 8$

$$\bar{x} = \frac{\sum x}{n} = \frac{544}{8} = 68$$

$$\bar{y} = \frac{\sum y}{n} = \frac{552}{8} = 69$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
65	67	−3	−2	9	4	6
66	68	−2	−1	4	1	2
67	65	−1	−4	1	16	4
67	68	−1	−1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
$\Sigma x = 544$	$\Sigma y = 552$	$\Sigma(x-\bar{x}) = 0$	$\Sigma(y-\bar{y}) = 0$	$\Sigma(x-\bar{x})^2 = 36$	$\Sigma(y-\bar{y})^2 = 44$	$\Sigma(x-\bar{x})(y-\bar{y}) = 24$

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

$$= \frac{24}{\sqrt{36}\sqrt{44}}$$

$$= 0.603$$

Example 6

Calculate the coefficient of correlation for the following pairs of x and y:

x	17	19	21	26	20	28	26	27
y	23	27	25	26	27	25	30	33

Solution

Let $a = 23$ and $b = 27$ be the assumed means of x and y series respectively.

$d_x = x - a = x - 23$

$d_y = y - b = y - 27$

$n = 8$

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
17	23	−6	−4	36	16	24
19	27	−4	0	16	0	0
21	25	−2	−2	4	4	4
26	26	3	−1	9	1	−3
20	27	−3	0	9	0	0
28	25	5	−2	25	4	−10
26	30	3	3	9	9	9
27	33	4	6	16	36	24
		$\Sigma d_x = 0$	$\Sigma d_y = 0$	$\Sigma d_x^2 = 124$	$\Sigma d_y^2 = 70$	$\Sigma d_x d_y = 48$

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum dx)^2}{n}}\sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

$$= \frac{48-0}{\sqrt{124-0}\sqrt{70-0}}$$

$$= 0.515$$

Note Since Σx, Σy, Σx^2, Σy^2, Σxy can be directly obtained with the help of scientific calculator, the correlation coefficient can be calculated without using assumed mean.

Example 7

Calculate the correlation coefficient from the following data:

x	23	27	28	29	30	31	33	35	36	39
y	18	22	23	24	25	26	28	29	30	32

Solution

Let $a = 30$ and $b = 25$ be the assumed means of x and y series respectively.

$d_x = x - a = x - 30$

$d_y = y - b = x - 25$

$n = 10$

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
23	18	–7	–7	49	49	49
27	22	–3	–3	9	9	9
28	23	–2	–2	4	4	4
29	24	–1	–1	1	1	1
30	25	0	0	0	0	0
31	26	1	1	1	1	1
33	28	3	3	9	9	9
35	29	5	4	25	16	20
36	30	6	5	36	25	30
39	32	9	7	81	49	63
		$\Sigma d_x = 11$	$\Sigma d_y = 7$	$\Sigma d_x^2 = 215$	$\Sigma d_y^2 = 163$	$\Sigma d_x d_y = 186$

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}} \sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

$$= \frac{186 - \frac{(11)(7)}{10}}{\sqrt{215 - \frac{(11)^2}{10}} \sqrt{163 - \frac{(7)^2}{10}}}$$

$= 0.996$

Example 8

Calculate the coefficient of correlation between the ages of cars and annual maintenance costs.

Age of cars (year)	2	4	6	7	8	10	12
Annual maintenance cost (₹)	1600	1500	1800	1900	1700	2100	2000

Solution

Let the ages of cars in years be denoted by x and annual maintenance costs in rupees be denoted by y.

Let $a = 7$ and $b = 1800$ be the assumed means of x and y series respectively.

Let $h = 1$, $k = 100$

$$d_x = \frac{x - a}{h} = \frac{x - 7}{1} = x - 7$$

$$d_y = \frac{y - b}{k} = \frac{y - 1800}{100}$$

$n = 7$

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
2	1600	–5	–2	25	4	10
4	1500	–3	3	9	9	9
6	1800	–1	0	1	0	0
7	1900	0	1	0	1	0
8	1700	1	–1	1	1	–1
10	2100	3	3	9	9	9
12	2000	5	2	25	4	10
		$\Sigma d_x = 0$	$\Sigma d_y = 0$	$\Sigma d_x^2 = 70$	$\Sigma d_y^2 = 28$	$\Sigma d_x d_y = 37$

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}} \sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

$$= \frac{37 - 0}{\sqrt{70 - 0}\sqrt{28 - 0}}$$

$= 0.836$

Example 9

Calculate Karl Pearson's coefficient of correlation for the data given below:

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

Solution

Let $a = 22$ and $b = 24$ be the assumed means of x and y series respectively.

Let $h = 4, k = 6$

$$d_x = \frac{x-a}{h} = \frac{x-22}{4}$$

$$d_y = \frac{y-b}{k} = \frac{y-24}{6}$$

$n = 6$

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
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
		$\Sigma d_x = -3$	$\Sigma d_y = -3$	$\Sigma d_x^2 = 19$	$\Sigma d_y^2 = 19$	$\Sigma d_x d_y = 12$

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}} \sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

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

$$= 0.6$$

Example 10

The coefficient of correlation between two variables X and Y is 0.48. The covariance is 36. The variance of X is 16. Find the standard deviation of Y.

Solution

$r = 0.48, \quad \text{cov}(X, Y) = 36, \quad \sigma_X^2 = 16$

$\therefore \quad \sigma_X = 4$

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

$$0.48 = \frac{36}{4\sigma_Y}$$

$$\therefore \quad \sigma_Y = 18.75$$

Example 11

Given $n = 10$, $\sigma_X = 5.4$, $\sigma_Y = 6.2$, and sum of the product of deviations from the mean of x and y is 66. Find the correlation coefficient.

Solution

$n = 10, \sigma_X = 5.4, \sigma_Y = 6.2$

$\sum (x-\bar{x})(y-\bar{y}) = 66$

$$\sigma_X = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

$$5.4 = \sqrt{\frac{\sum (x-\bar{x})^2}{10}}$$

$$\therefore \quad \sum (x-\bar{x})^2 = 291.6$$

$$\sigma_Y = \sqrt{\frac{\sum (y-\bar{y})^2}{n}}$$

$$6.2 = \sqrt{\frac{\sum (y-\bar{y})^2}{10}}$$

$$\therefore \quad \sum (y-\bar{y})^2 = 384.4$$

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

$$= \frac{66}{\sqrt{291.6}\sqrt{384.4}}$$

$$= 0.197$$

Example 12

From the following information, calculate the value of n.

$$\sum x = 4, \sum y = 4, \sum x^2 = 44, \sum y^2 = 44, \sum xy = -40, r = -1$$

Solution

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

$$-1=\frac{-40-\frac{(4)(4)}{n}}{\sqrt{44-\frac{(4)^2}{n}}\sqrt{44-\frac{(4)^2}{n}}}$$

$\therefore \quad n=8$

Example 13

From the following data, find the number of items n.

$$r=0.5,\ \sum(x-\bar{x})(y-\bar{y})=120,\ \sigma_Y=8,\ \sum(x-\bar{x})^2=90$$

Solution

$$\sigma_Y=\sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

$$8=\sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

$$\sum(y-\bar{y})^2=64n$$

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

$$0.5=\frac{120}{\sqrt{90}\sqrt{64n}}$$

$\therefore \quad n=10$

Example 14

Calculate the correlation coefficient between x and y from the following data:

$$n=10,\ \sum x=140,\ \sum y=150,\ \sum(x-10)^2=180$$
$$\sum(y-15)^2=215,\ \sum(x-10)(y-15)=60$$

Solution

$$\sum d_x^2=\sum(x-10)^2=180$$
$$\sum d_y^2=\sum(y-15)^2=215$$
$$\sum d_x d_y=\sum(x-10)(y-15)=60$$
$$a=10$$
$$b=15$$
$$n=10$$

$$\bar{x}=\frac{\sum x}{n}=\frac{140}{10}=14$$

$$\bar{y}=\frac{\sum y}{n}=\frac{150}{10}=15$$

$$\bar{x}=a+\frac{\sum d_x}{n}$$

$$14=10+\frac{\sum d_x}{10}$$

$\therefore \quad \sum d_x=40$

$$\bar{y}=b+\frac{\sum d_y}{n}$$

$$15=15+\frac{\sum d_y}{10}$$

$\therefore \quad \sum d_y=0$

$$r=\frac{\sum d_x d_y-\frac{\sum d_x\sum d_y}{n}}{\sqrt{\sum d_x^2-\frac{(\sum d_x)^2}{n}}\sqrt{\sum d_y^2-\frac{(\sum d_y)^2}{n}}}$$

$$=\frac{60-\frac{(40)(0)}{10}}{\sqrt{180-\frac{(40)^2}{10}}\sqrt{215-\frac{0}{10}}}$$

$$=0.915$$

Example 15

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

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

It was 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

$$n = 25$$

$$\text{Corrected } \sum x = \text{Incorrect } \sum x - (\text{Sum of incorrect } x) + (\text{Sum of correct } x)$$
$$= 125 - (6+8) + (8+6)$$
$$= 125$$

Similarly,

$$\text{Corrected } \sum y = 100 - (14+6) + (12+8) = 100$$
$$\text{Corrected } \sum x^2 = 650 - (6^2+8^2) + (8^2+6^2) = 650$$
$$\text{Corrected } \sum y^2 = 460 - (14^2+6^2) + (12^2+8^2) = 436$$
$$\text{Corrected } \sum xy = 508 - (84+48) + (96+48) = 520$$

Correct value of correlation coefficient

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}} \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$$
$$= \frac{520 - \frac{(125)(100)}{25}}{\sqrt{650 - \frac{(125)^2}{25}} \sqrt{436 - \frac{(100)^2}{25}}}$$
$$= 0.67$$

EXERCISE 4.1

1. Draw a scatter diagram to represent the following data:

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

Calculate the coefficient of correlation between x and y.

[Ans.: –0.92]

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

x	10	12	18	24	23	27
y	13	18	12	25	30	10

[Ans.: 0.223]

3. From the following information relating to the stock exchange quotations for two shares *A* and *B*, ascertain by using Pearson's coefficient of correlation how shares *A* and *B* are correlated in their prices?

Price share (*A*) ₹	160	164	172	182	166	170	178
Price share (*B*) ₹	292	280	260	234	266	254	230

[Ans.: –0.96]

4. Find the correlation coefficient between the income and expenditure of a wage earner.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul
Income	46	54	56	56	58	60	62
Expenditure	36	40	44	54	42	58	54

[Ans.: 0.769]

5. From the following data, examine whether the input of oil and output of electricity can be said to be correlated.

Input of oil	6.9	8.2	7.8	4.8	9.6	8.0	7.7
Output of Electricity	1.9	3.5	6.5	1.3	5.5	3.5	2.2

[Ans.: 0.696]

6. For the following data, show that cov $(x, x^2) = 0$.

x	–3	–2	–1	0	1	2	3
x^2	9	4	1	0	1	4	9

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

x	62	64	65	69	70	71	72	74
y	126	125	139	145	165	152	180	208

[Ans.: 0.9032]

8. The following data gave the growth of employment in lacs in the organized sector in India between 1988 and 1995:

Year	1988	1989	1990	1991	1992	1993	1994	1995
Public sector	98	101	104	107	113	120	125	128
Private sector	65	65	67	68	68	69	68	68

Find the correlation coefficient between the employment in public and private sectors.

[Ans.: 0.77]

9. Calculate Karl Pearson's coefficient of correlation from the following data, using 20 as the 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.: −0.954]

10. A sample of 25 pairs of values x and y lead to the following results:

$\sum x = 127, \sum y = 100, \sum x^2 = 760, \sum y^2 = 449, \sum xy = 500$

Later on, it was found that two pairs of values were taken as (8, 14) and (8, 6) instead of the correct values (8, 12) and (6, 8). Find the corrected coefficient between x and y.

[Ans.: −0.31]

4.9 RANK CORRELATION

Let a group of n individuals be arranged in order of merit with respect to some characteristics. The same group would give a different order (rank) for different characteristics. Considering the orders corresponding to two characteristics A and B, the correlation between these n pairs of ranks is called the *rank correlation* in the characteristics A and B for that group of individuals.

4.9.1 Spearman's Rank Correlation Coefficient

Let x, y be the ranks of the i^{th} individuals in two characteristics A and B respectively where $i = 1, 2, ..., n$. Assuming that no two individuals have the same rank either for x or y, each of the variables x and y take the values 1, 2, ..., n.

$$\bar{x} = \bar{y} = \frac{1+2+3+\cdots+n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

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

$$= \sum x^2 - 2\bar{x}\sum x + \bar{x}^2 \sum 1$$

$$= \sum x^2 - 2n\bar{x}^2 + n\bar{x}^2 \qquad \left[\because \sum x = n\bar{x} \text{ and } \sum 1 = n\right]$$

$$= \sum x^2 - n\bar{x}^2$$

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

$$= \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)^2}{4}$$

$$= \frac{1}{12}(n^3 - n)$$

Similarly, $\sum (y-\bar{y})^2 = \frac{1}{12}(n^3 - n)$

If d denotes the difference between the ranks of the i^{th} individuals in the two variables,

$$d = x - y = (x-\bar{x}) - (y-\bar{y}) \qquad [\because \bar{x} = \bar{y}]$$

Squaring and summing over i from 1 to n,

$$\sum d^2 = \sum [(x-\bar{x}) - (y-\bar{y})]^2$$

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

$$\sum (x-\bar{x})(y-\bar{y}) = \frac{1}{2}\left[\sum (x-\bar{x})^2 + \sum (y-\bar{y})^2 - \sum d^2\right]$$

$$= \frac{1}{12}(n^3 - n) - \frac{1}{2}\sum d^2$$

Hence, the coefficient of correlation between these variables is

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

$$= \frac{\frac{1}{12}(n^3 - n) - \frac{1}{2}\sum d^2}{\frac{1}{12}(n^3 - n)}$$

$$= 1 - \frac{6\sum d^2}{n^3 - n}$$

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

This is called Spearman's rank correlation coefficient and is denoted by ρ.

Note $\sum d = \sum (x-y) = \sum x - \sum y = n(\bar{x} - \bar{y}) = 0$

Example 1

Ten participants in a contest are ranked by two judges as follows:

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

Calculate the rank correlation coefficient.

Solution

$n = 10$

Rank by first Judge x	Rank by second Judge y	$d = x - y$	d^2
1	3	−2	4
3	1	2	4
7	4	3	9
5	5	0	0
4	6	−2	4
6	9	−3	9
2	7	−5	25
10	8	2	4
9	10	−1	1
8	2	6	36
		$\sum d = 0$	$\sum d^2 = 96$

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$
$$= 1 - \frac{6(96)}{10\left[(10)^2 - 1\right]}$$
$$= 0.418$$

Example 2

Ten competitors in a musical test were ranked by the three judges A, B, and C in the following order:

Rank by A	1	6	5	10	3	2	4	9	7	8
Rank by B	3	5	8	4	7	10	2	1	6	9
Rank by C	6	4	9	8	1	2	3	10	5	7

Using the rank correlation method, find which pair of judges has the nearest approach to common liking in music. **[Summer 2015]**

Solution

$n = 10$

Rank by A x	Rank by B y	Rank by C z	$d_1 = x - y$	$d_2 = y - z$	$d_3 = z - x$	d_1^2	d_2^2	d_3^2
1	3	6	−2	−3	5	4	9	25
6	5	4	1	1	−2	1	1	4
5	8	9	−3	−1	4	9	1	16
10	4	8	6	−4	−2	36	16	4
3	7	1	−4	6	−2	16	36	4
2	10	2	−8	8	0	64	64	0
4	2	3	2	−1	−1	4	1	1
9	1	10	8	−9	1	64	81	1
7	6	5	1	1	−2	1	1	4
8	9	7	−1	2	−1	1	4	1
			$\sum d_1 = 0$	$\sum d_2 = 0$	$\sum d_3 = 0$	$\sum d_1^2 = 200$	$\sum d_2^2 = 214$	$\sum d_3^2 = 60$

$$r(x, y) = 1 - \frac{6\sum d_1^2}{n(n^2-1)}$$
$$= 1 - \frac{6(200)}{10\left[(10)^2 - 1\right]}$$
$$= -0.21$$
$$r(y, z) = 1 - \frac{6\sum d_2^2}{n(n^2-1)}$$
$$= 1 - \frac{6(214)}{10\left[(10)^2 - 1\right]}$$
$$= -0.296$$
$$r(z, x) = 1 - \frac{6\sum d_3^2}{n(n^2-1)}$$
$$= 1 - \frac{6(60)}{10\left[(10)^2 - 1\right]}$$
$$= 0.64$$

Since $r(z, x)$ is maximum, the pair of judges A and C has the nearest common approach.

Example 3

Ten students got the following percentage of marks in mathematics and physics:

Mathematics (x)	8	36	98	25	75	82	92	62	65	35
Physics (y)	84	51	91	60	68	62	86	58	35	49

Find the rank correlation coefficient.

Solution

$n = 10$

x	y	Rank in mathematics x	Rank in Physics y	$d = x - y$	d^2
8	84	10	3	7	49
36	51	7	8	−1	1
98	91	1	1	0	0
25	60	9	6	3	9
75	68	4	4	0	0
82	62	3	5	−2	4
92	86	2	2	0	0
62	58	6	7	−1	1
65	35	5	10	−5	25
35	49	8	9	−1	1
				$\sum d = 0$	$\sum d^2 = 90$

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$
$$= 1 - \frac{6(90)}{10\left[(10)^2 - 1\right]}$$
$$= 0.455$$

Example 4

The coefficient of rank correlation of the marks obtained by 10 students in physics and chemistry was found to be 0.5. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the rank coefficient of the rank correlation.

Solution

$n = 10$

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$
$$0.5 = 1 - \frac{6\sum d^2}{10(100-1)}$$
$$\therefore \quad \sum d^2 = 82.5$$

Correct $\sum d^2$ = Incorrect $\sum d^2$ − (Incorrect rank difference)^2 + (Correct rank difference)^2

$$= 82.5 - (3)^2 + (7)^2$$
$$= 122.5$$

Correct coefficient of rank correlation $r = 1 - \frac{6(122.5)}{10(100-1)}$

$$= 0.26$$

4.9.2 Tied Ranks

If there is a tie between two or more individuals ranks, the rank is divided among equal individuals, e.g., if two items have fourth rank, the 4th and 5th rank is divided between them equally and is given as $\frac{4+5}{2} = 4.5^{th}$ rank to each of them. If three items have the same 4th rank, each of them is given $\frac{4+5+6}{3} = 5^{th}$ rank. As a result of this, the following adjustment or correction is made in the rank correlation formula. If m is the number of item having equal ranks then the factor $\frac{1}{12}(m^3 - m)$ is added to $\sum d^2$. If there are more than one cases of this type, this factor is added corresponding to each case.

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

Example 1

Obtain the rank correlation coefficient from the following data:

x	10	12	18	18	15	40
y	12	18	25	25	50	25

Solution

Here, $n = 6$

x	y	Rank x	Rank y	$d = x - y$	d^2
10	12	1	1	0	0
12	18	2	2	0	0
18	25	4.5	4	0.5	0.25
18	25	4.5	4	0.5	0.25
15	50	3	6	−3	9
40	25	6	4	2	4
					$\sum d^2 = 13.5$

There are two items in the x series having equal values at the rank 4. Each is given the rank 4.5. Similarly, there are three items in the y series at the rank 3. Each of them is given the rank 4.

$$m_1 = 2, m_2 = 3$$

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

$$= 1 - \frac{6\left[13.50 + \frac{1}{12}(8-2) + \frac{1}{12}(27-3)\right]}{6\left[(6)^2 - 1\right]}$$

$$= 0.5429$$

EXERCISE 4.2

1. Compute Spearman's rank correlation coefficient from the following data:

x	18	20	34	52	12
y	39	23	35	18	46

[Ans.: −0.9]

2. Two judges gave the following ranks to a series of eight one-act plays in a drama competition. Examine the relationship between their judgements.

Judge A	8	7	6	3	2	1	5	4
Judge B	7	5	4	1	3	2	6	8

[Ans.: 0.62]

3. From the following data, calculate Spearman's rank correlation between x and y.

x	36	56	20	42	33	44	50	15	60
y	50	35	70	58	75	60	45	80	38

[Ans.: 0.92]

4. Ten competitors in a voice test are ranked by three judges in the following order:

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

Use the method of rank correlation to gauge which pairs of judges has the nearest approach to common liking in voice.

[Ans.: The first and third judge**]**

5. The following table gives the scores obtained by 11 students in English and Tamil translation. Find the rank correlation coefficient.

Scores in English	40	46	54	60	70	80	82	85	85	90	95
Scores in Tamil	45	45	50	43	40	75	55	72	65	42	70

[Ans.: 0.36]

6. Calculate Spearman's coefficient of rank correlation for the following data:

x	53	98	95	81	75	71	59	55
y	47	25	32	37	30	40	39	45

[Ans.: −0.905]

7. Following are the scores of ten students in a class and their IQ:

Score	35	40	25	55	85	90	65	55	45	50
IQ	100	100	110	140	150	130	100	120	140	110

Calculate the rank correlation coefficient between the score IQ.

[Ans.: 0.47]

4.10 REGRESSION

Regression is defined as a method of estimating the value of one variable when that of the other is known and the variables are correlated. *Regression analysis* is used to predict or estimate one variable in terms of the other variable. It is a highly valuable tool for prediction purpose in economics and business. It is useful in statistical estimation of demand curves, supply curves, production function, cost function, consumption function, etc.

4.11 TYPES OF REGRESSION

Regression is classified into two types:

1. Simple and multiple regressions
2. Linear and nonlinear regressions

4.11.1 Simple and Multiple Regressions

Depending upon the study of the number of variables, regression may be simple or multiple.

1. Simple Regression The regression analysis for studying only two variables at a time is known as simple regression.

2. Multiple Regression The regression analysis for studying more than two variables at a time is known as multiple regression.

4.11.2 Linear and Nonlinear Regressions

Depending upon the regression curve, regression may be linear or nonlinear.

1. Linear Regression If the regression curve is a straight line, the regression is said to be linear.

2. Nonlinear Regression If the regression curve is not a straight line i.e., not a first-degree equation in the variables x and y, the regression is said to be nonlinear or curvilinear. In this case, the regression equation will have a functional relation between the variables x and y involving terms in x and y of the degree higher than one, i.e., involving terms of the type x^2, y^2, x^3, y^3, xy, etc.

4.12 METHODS OF STUDYING REGRESSION

There are two methods of studying correlation:

(i) Method of scatter diagram
(ii) Method of least squares

4.12.1 Method of Scatter Diagram

It is the simplest method of obtaining the lines of regression. The data are plotted on a graph paper by taking the independent variable on the x-axis and the dependent variable on the y-axis. Each of these points are generally scattered in a narrow strip. If the correlation is perfect, i.e., if r is equal to one, positive, or negative, the points will lie on a line which is the line of regression.

4.12.2 Method of Least Squares

This is a mathematical method which gives an objective treatment to find a line of regression. It is used for obtaining the equation of a curve which fits best to a given set of observations. It is based on the assumption that the sum of squares of differences between the estimated values and the actual observed values of the observations is minimum.

4.13 LINES OF REGRESSION

If the variables, which are highly correlated, are plotted on a graph then the points lie in a narrow strip. If all the points in the scatter diagram cluster around a straight line, the line is called the *line of regression*. The line of regression is the line of best fit and is obtained by the principle of least squares.

Line of Regression of y on x

It is the line which gives the best estimate for the values of y for any given values of x. The regression equation of y on x is given by

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

It is also written as

$$y=a+bx$$

Line of Regression of x on y

It is the line which gives the best estimate for the values of x for any given values of y. The regression equation for x on y is given by

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

It is also written as

$$x=a+by$$

where $\bar{x}$ and $\bar{y}$ are means of x series and y series respectively, σ_x and σ_y are standard deviations of x series and y series respectively, r is the correlation coefficient between x and y.

4.14 REGRESSION COEFFICIENTS

The slope b of the line of regression of y on x is also called the *coefficient of regression* of y on x. It represents the increment in the value of y corresponding to a unit change in the value of x.

b_{yx} = Regression coefficient of y on x

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

Similarly, the slope b of the line of regression of x on y is called the coefficient of regression of x on y. It represents the increment in the value of x corresponding to a unit change in the value of y.

b_{xy} = Regression coefficient of x on y

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

Expressions for Regression Coefficients

(i) We know that

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

$$\sigma_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

$$\sigma_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

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

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

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

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

(ii) We know that

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

$$\sigma_x = \sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$\sigma_y = \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}$$

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

$$= \frac{\sum xy - \frac{\sum x\sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

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

$$= \frac{\sum xy - \frac{\sum x\sum y}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}}$$

(iii) We know that

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}\sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

$$\sigma_x = \sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}$$

$$\sigma_y = \sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}$$

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

$$= \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}}$$

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

$$= \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

4.15 PROPERTIES OF REGRESSION COEFFICIENTS

1. The coefficient of correlation is the geometric mean of the coefficients of regression, i.e., $r = \sqrt{b_{yx} b_{xy}}$.

Proof We know that

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

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

$$b_{yx}\, b_{xy} = r\frac{\sigma_y}{\sigma_x} \cdot r\frac{\sigma_x}{\sigma_y}$$

$$= r^2$$

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

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

Proof Let $b_{yx} > 1$

We know that

$$r^2 \le 1 \text{ and } r^2 = b_{yx}\, b_{xy}$$

$$b_{yx}\, b_{xy} \le 1$$

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

Hence, if $b_{yx} < 1$, $b_{xy} > 1$

3. The arithmetic mean of regression coefficients is greater than or equal to the coefficient of correlation.

Proof We have to prove that

$$\frac{1}{2}(b_{yx} + b_{xy}) \ge r$$

i.e., $$\frac{1}{2}\left(r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y}\right) \ge r$$

i.e., $$\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \ge 2$$

i.e., $$\sigma_y^2 + \sigma_x^2 - 2\sigma_x\sigma_y \ge 0$$

i.e., $$(\sigma_y - \sigma_x)^2 \ge 0$$

which is always true, since the square of a real quantity is $1 \ge 0$.

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

Proof Let $$d_x = \frac{x-a}{h},\ d_y = \frac{y-b}{k}$$

$$x = a + hd_x,\ y = b + kd_y$$

where a, b, h (> 0) and $k (> 0)$ are constants.

$$r_{d_x d_y} = r_{xy},\ \sigma_{d_x}^2 = \frac{1}{h^2}\sigma_x^2,\ \sigma_{d_y}^2 = \frac{1}{k^2}\sigma_y^2$$

$$b_{d_x d_y} = r_{d_x d_y}\frac{\sigma_{d_x}}{\sigma_{d_y}}$$

$$= r_{xy}\frac{\sigma_x}{h}\frac{k}{\sigma_y}$$

$$= \frac{k}{h} r_{xy}\frac{\sigma_x}{\sigma_y}$$

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

Similarly, $$b_{d_y d_x} = \frac{h}{k} b_{yx}$$

5. Both regression coefficients will have the same sign i.e., either both are positive or both are negative.

6. The sign of correlation is same as that of the regression coefficients, i.e., $r > 0$ if $b_{xy} > 0$ and $b_{yx} > 0$; and $r < 0$ if $b_{xy} < 0$ and $b_{yx} < 0$.

4.16 PROPERTIES OF LINES OF REGRESSION (LINEAR REGRESSION)

1. The two regression lines x on y and y on x always intersect at their means $(\bar{x}, \bar{y})$.
2. Since $r^2 = b_{yx}\, b_{xy}$, i.e., $r = \sqrt{b_{yx}\, b_{xy}}$, therefore, r, b_{yx}, b_{xy} all have the same sign.
3. If $r = 0$, the regression coefficients are zero.
4. The regression lines become identical if $r = \pm 1$. It follows from the regression equations that $x = \bar{x}$ and $y = \bar{y}$. If $r = 0$, these lines are perpendicular to each other.

Example 1

The regression lines of a sample are $x + 6y = 6$ and $3x + 2y = 10$. Find
(i) sample means $\bar{x}$ and $\bar{y}$, and
(ii) the coefficient of correlation between x and y.
(iii) Also estimate y when $x = 12$.

Solution

(i) The regression lines pass through the point $(\bar{x}, \bar{y})$.

$$\bar{x} + 6\bar{y} = 6 \quad \ldots(1)$$

$$3\bar{x} + 2\bar{y} = 10 \quad \ldots(2)$$

Solving Eqs (1) and (2),

$$\bar{x} = 3, \quad \bar{y} = \frac{1}{2}$$

(ii) Let the line $x + 6y = 6$ be the line of regression of y on x.

$$6y = -x + 6$$

$$y = -\frac{1}{6}x + 1$$

$$\therefore \quad b_{yx} = -\frac{1}{6}$$

Let the line $3x + 2y = 10$ be the line of regression of x on y.

$$3x = -2y + 10$$

$$x = -\frac{2}{3}y + \frac{10}{3}$$

$$\therefore \quad b_{xy} = -\frac{2}{3}$$

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{\left(-\frac{1}{6}\right)\left(-\frac{2}{3}\right)} = \frac{1}{3}$$

Since b_{yx} and b_{xy} are negative, r is negative.

$$r = -\frac{1}{3}$$

Estimated value of y when $x = 12$ is

$$y = -\frac{1}{6}(12) + 1 = -1$$

Example 2

If the two lines of regression are $4x - 5y + 30 = 0$ and $20x - 9y - 107 = 0$, which of these are lines of regression of x on y and y on x? Find r_{xy} and σ_y when $\sigma_x = 3$.

Solution

For the line $4x - 5y + 30 = 0$,

$$-5y = -4x - 30$$

$$y = 0.8x + 6$$

$$\therefore \quad b_{yx} = 0.8$$

For the line $20x - 9y - 107 = 0$

$$20x = 9y + 107$$

$$x = 0.45y + 5.35$$

$$\therefore \quad b_{xy} = 0.45$$

Both b_{yx} and b_{xy} are positive.
Hence, line $4x - 5y + 30 = 0$ is the line of regression of y one x and line $20x - 9y - 107 = 0$ is the line of regression of x on y.

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(0.8)(0.45)} = 0.6$$

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

$$0.8 = 0.6\left(\frac{\sigma_y}{3}\right)$$

$$\therefore \quad \sigma_y = 4$$

Example 3

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

$$\sum x = 15000, \quad \sum x^2 = 2272500, \quad \sum y = 6800$$

$$\sum y^2 = 463025, \quad \sum xy = 1022250$$

Find the coefficient of correlation between height and weight and also the equation of regression of height and weight.

Solution

$$n = 100$$

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

$$= \frac{1022250 - \frac{(15000)(6800)}{100}}{2272500 - \frac{(15000)^2}{100}}$$

$$= 0.1$$

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

$$= \frac{1022250 - \frac{(15000)(6800)}{100}}{463025 - \frac{(6800)^2}{100}}$$

$$= 3.6$$

$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(0.1)(3.6)} = 0.6$$

$$\bar{x} = \frac{\sum x}{n} = \frac{15000}{100} = 150$$

$$\bar{y} = \frac{\sum y}{n} = \frac{6800}{100} = 68$$

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

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

$$y - 68 = 0.1(x - 150)$$

$$y = 0.1x + 53$$

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

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

$$x - 150 = 3.6(y - 68)$$

$$x = 3.6y - 94.8$$

Example 4

For a bivariate data, the mean value of x is 20 and the mean value of y is 45. The regression coefficient of y on x is 4 and that of x on y is $\frac{1}{9}$. Find

(i) the coefficient of correlation, and

(ii) the standard deviation of x if the standard deviation of y is 12.

(iii) Also write down the equations of regression lines..

Solution

$$\bar{x} = 20, \quad \bar{y} = 45, \quad b_{yx} = 4, \quad b_{xy} = \frac{1}{9}$$

(i) $$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(4)\left(\frac{1}{9}\right)} = \frac{2}{3} = 0.667$$

(ii) $$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$4 = \frac{2}{3}\left(\frac{12}{\sigma_x}\right)$$

$$\therefore \quad \sigma_x = 2$$

(iii) The equation of the regression line of y on x is

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

$$y - 45 = 4(x - 20)$$

$$y = 4x - 35$$

The equation of the regression line of x on y is

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

$$x - 20 = \frac{1}{9}(y - 45)$$

$$x = \frac{1}{9}y + 15$$

Example 5

From the following results, obtain the two regression equations and estimate the yield when the rainfall is 29 cm and the rainfall, when the yield is 600 kg:

	Yield in kg	Rainfall in cm
Mean	508.4	26.7
SD	36.8	4.6

The coefficient of correlation between yield and rainfall is 0.52.

Solution

Let rainfall in cm be denoted by x and yield in kg be denoted by y.

$\bar{x} = 26.7, \quad \bar{y} = 508.4, \quad \sigma_x = 4.6, \quad \sigma_y = 36.8, \quad r = 0.52$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$
$$= 0.52\left(\frac{36.8}{4.6}\right)$$
$$= 4.16$$
$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$
$$= 0.52\left(\frac{4.6}{36.8}\right)$$
$$= 0.065$$

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

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$y - 508.4 = 4.16(x - 26.7)$$
$$y = 4.16x + 397.328$$

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

$$x - \bar{x} = b_{xy}(y - \bar{y})$$
$$x - 26.7 = 0.065(y - 508.4)$$
$$x = 0.065y - 6.346$$

Estimated yield when the rainfall is 29 cm is

$y = 4.16\,(29) + 397.328 = 517.968$ kg

Estimated rainfall when the yield is 600 kg is

$x = 0.065\,(600) - 6.346 = 32.654$ cm

Example 6

Find the regression coefficients b_{yx} and b_{xy} and hence, find the correlation coefficient between x and y for the following data:

x	4	2	3	4	2
y	2	3	2	4	4

Solution

$n = 5$

x	y	x^2	y^2	xy
4	2	16	4	8
2	3	4	9	6
3	2	9	4	6
4	4	16	16	16
2	4	4	16	8
$\Sigma x = 15$	$\Sigma y = 15$	$\Sigma x^2 = 49$	$\Sigma y^2 = 49$	$\Sigma xy = 44$

$$b_{yx} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$
$$= \frac{44 - \frac{(15)(15)}{5}}{49 - \frac{(15)^2}{5}}$$
$$= -0.25$$
$$b_{xy} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}}$$
$$= \frac{44 - \frac{(15)(15)}{5}}{49 - \frac{(15)^2}{5}}$$
$$= -0.25$$
$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(-0.25)(-0.25)} = 0.25$$

Since b_{yx} and b_{xy} are negative, r is negative.

$r = -0.25$

Note Σx, Σy, Σx^2, Σy^2, Σxy can be directly obtained with the help of scientific calculator.

Example 7

The following data give the experience of machine operators and their performance rating as given by the number of good parts turned out per 100 pieces.

Operator	1	2	3	4	5	6
Performance rating (x)	23	43	53	63	73	83
Experience (y)	5	6	7	8	9	10

Calculate the regression line of performance rating on experience and also estimate the probable performance if an operator has 11 years of experience. **[Summer 2015]**

Solution

$n = 6$

x	y	y^2	xy
23	5	25	115
43	6	36	258
53	7	49	371
63	8	64	504
73	9	81	657
83	10	100	830
$\Sigma x = 338$	$\Sigma y = 45$	$\Sigma y^2 = 355$	$\Sigma xy = 2735$

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

$$= \frac{2735 - \frac{(338)(45)}{6}}{355 - \frac{(45)^2}{6}}$$

$$= 11.429$$

$$\bar{x} = \frac{\sum x}{n} = \frac{338}{6} = 56.33$$

$$\bar{y} = \frac{\sum y}{n} = \frac{45}{6} = 7.5$$

The equation of regression line of x on y is

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

$$x - 56.33 = 11.429(y - 7.5)$$

$$x = 11.429\,y - 29.3875$$

Estimated performance if $y = 11$ is

$$x = 11.429(11) - 29.3875 = 96.3315$$

Example 8

The number of bacterial cells (y) per unit volume in a culture at different hours (x) is given below:

x	0	1	2	3	4	5	6	7	8	9
y	43	46	82	98	123	167	199	213	245	272

Fit lines of regression of y on x and x on y. Also, estimate the number of bacterial cells after 15 hours.

Solution

$n = 10$

x	y	x^2	xy	y^2
0	43	0	0	1849
1	46	1	46	2116
2	82	4	164	6724
3	98	9	294	9604
4	123	16	492	15129
5	167	25	835	27889
6	199	36	1194	39601
7	213	49	1491	45369
8	245	64	1960	60025
9	272	81	2448	73984
$\Sigma x = 45$	$\Sigma y = 1488$	$\Sigma x^2 = 285$	$\Sigma xy = 8924$	$\Sigma y^2 = 282290$

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

$$= \frac{8924 - \frac{(45)(1488)}{10}}{285 - \frac{(45)^2}{10}}$$

$$= 27.0061$$

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

$$= \frac{8924 - \frac{(45)(1488)}{10}}{282290 - \frac{(1488)^2}{10}}$$

$$= 0.0366$$

$$\bar{x} = \frac{\sum x}{n} = \frac{45}{10} = 4.5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{1488}{10} = 148.8$$

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

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

$$y - 148.8 = 27.0061(x - 4.5)$$

$$y = 27.0061x + 27.2726$$

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

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

$$x - 4.5 = 0.0366(y - 148.8)$$

$$x = 0.366y - 0.9461$$

At $x = 15$ hours,

$y = 27.0061\ (15) + 27.2726 = 432.3641$

Example 9

Find the regression coefficient of y on x for the following data:

x	1	2	3	4	5
y	160	180	140	180	200

Solution

$n = 5$

$$\bar{x} = \frac{\sum x}{n} = \frac{15}{5} = 3$$

$$\bar{y} = \frac{\sum y}{n} = \frac{860}{5} = 172$$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	160	−2	−12	4	24
2	180	−1	8	1	−8
3	140	0	−32	0	0
4	180	1	8	1	8
5	200	2	28	4	56
$\sum x = 15$	$\sum y = 860$	$\sum(x - \bar{x}) = 0$	$\sum(y - \bar{y}) = 0$	$\sum(x - \bar{x})^2 = 10$	$\sum(x - \bar{x})(y - \bar{y}) = 80$

$$b_{yx} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$= \frac{80}{10}$$

$$= 8$$

Note Since $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$, $\sum xy$ can be directly obtained with the help of scientific calculator, the regression coefficient can be calculated without using mean.

Example 10

Calculate the two regression coefficients from the data and find correlation coefficient.

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

Solution

$n = 5$

$\bar{x} = \frac{\sum x}{n} = \frac{30}{5} = 6$

$\bar{y} = \frac{\sum y}{n} = \frac{30}{5} = 6$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
7	6	1	0	1	0	0
4	5	–2	–1	4	1	2
8	9	2	3	4	9	6
6	8	0	2	0	4	0
5	2	–1	–4	1	16	4
$\sum x = 30$	$\sum y = 30$	$\sum(x-\bar{x}) = 0$	$\sum(y-\bar{y}) = 0$	$\sum(x-\bar{x})^2 = 10$	$\sum(y-\bar{y})^2 = 30$	$\sum(x-\bar{x})(y-\bar{y}) = 12$

$b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$

$= \frac{12}{10}$

$= 1.2$

$b_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$

$= \frac{12}{30}$

$= 0.4$

$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(1.2)(0.4)} = 0.693$

Example 11

Obtain the two regression lines from the following data and hence, find the correlation coefficient.

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

[Summer 2015]

Solution

$n = 5$

$\bar{x} = \frac{\sum x}{n} = \frac{30}{5} = 6$

$\bar{y} = \frac{\sum y}{n} = \frac{40}{5} = 8$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
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
$\sum x = 30$	$\sum y = 40$	$\sum(x-\bar{x}) = 0$	$\sum(y-\bar{y}) = 0$	$\sum(x-\bar{x})^2 = 40$	$\sum(y-\bar{y})^2 = 20$	$\sum(x-\bar{x})(y-\bar{y}) = -26$

$b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$

$= \frac{-26}{40}$

$= -0.65$

$b_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$

$= \frac{-26}{20}$

$= -1.3$

The equation of regression line of y on x is

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

$y - 8 = -0.65(x-6)$

$y = -0.65x + 11.9$

The equation of regression line of x on y is

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

$x - 6 = -1.3(y-8)$

$x = -1.3y + 16.4$

$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(-0.65)(-1.3)} = 0.9192$

Since b_{yx} and b_{xy} are negative, r is negative.

$r = -0.9192$.

Example 12

Calculate the regression coefficients and find the two lines of regression from the following data:

x	57	58	59	59	60	61	62	64
y	67	68	65	68	72	72	69	71

Find the value of y when x = 66.

Solution

$n = 8$

$$\bar{x} = \frac{\sum x}{n} = \frac{480}{8} = 60$$

$$\bar{y} = \frac{\sum y}{n} = \frac{552}{8} = 69$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
57	67	−3	−2	9	4	6
58	68	−2	−1	4	1	2
59	65	−1	−4	1	16	4
59	68	−1	−1	1	1	1
60	72	0	3	0	9	0
61	72	1	3	1	9	3
62	69	2	0	4	0	0
64	71	4	2	16	4	8
$\Sigma x = 480$	$\Sigma y = 552$	$\Sigma(x-\bar{x}) = 0$	$\Sigma(y-\bar{y}) = 0$	$\Sigma(x-\bar{x})^2 = 36$	$\Sigma(y-\bar{y})^2 = 44$	$\Sigma(x-\bar{x})(y-\bar{y}) = 24$

$$b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

$$= \frac{24}{36}$$

$$= 0.667$$

$$b_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$$

$$= \frac{24}{44}$$

$$= 0.545$$

The equation of regression line of y on x is

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

$$y - 69 = 0.667(x - 60)$$

$$y = 0.667x + 28.98$$

The equation of regression line of x on y is

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

$$x - 60 = 0.545(y - 69)$$

$$x = 0.545y + 22.395$$

Value of y when x = 66 is

$$y = 0.667\ (66) + 28.98 = 73.002$$

Example 13

The following data represents rainfall (x) and yield of paddy per hectare (y) in a particular area. Find the linear regression of x on y.

x	113	102	95	120	140	130	125
y	1.8	1.5	1.3	1.9	1.1	2.0	1.7

Solution

Let $a = 120$ and $b = 1.8$ be the assumed means of x and y series respectively.

$$d_x = x - a = x - 120$$

$$d_y = y - b = y - 1.8$$

$$n = 7$$

x	y	d_x	d_y	d_y^2	d_xd_y
113	1.8	–7	0	0	0
102	1.5	–18	–0.3	0.09	5.4
95	1.3	–25	–0.5	0.25	12.5
120	1.9	0	0.1	0.01	0
140	1.1	20	–0.7	0.49	–14
130	2.0	10	0.2	0.04	2.0
125	1.7	5	–0.1	0.01	–0.5
$\Sigma x = 825$	$\Sigma y = 11.3$	$\Sigma d_x = -15$	$\Sigma d_y = -1.3$	$\Sigma d_y^2 = 0.89$	$\Sigma d_xd_y = 5.4$

$$b_{xy} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}$$

$$= \frac{5.4 - \frac{(-15)(-1.3)}{7}}{0.89 - \frac{(-1.3)^2}{7}}$$

$$= 4.03$$

$$\bar{x} = \frac{\sum x}{n} = \frac{825}{7} = 117.86$$

$$\bar{y} = \frac{\sum y}{n} = \frac{11.3}{7} = 1.614$$

The equation of the regression line of x on y is

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

$$x - 117.86 = 4.03(y - 1.614)$$

$$x = 4.03y + 111.36$$

Note Since Σx, Σy, Σx^2, Σy^2, Σxy can be directly obtained with the help of scientific calculator, the regression coefficient can be calculated without using assumed mean.

Example 14

Find the two lines of regression from the following data:

Age of husband (x)	25	22	28	26	35	20	22	40	20	18
Age of wife (y)	18	15	20	17	22	14	16	21	15	14

Hence, estimate (i) the age of the husband when the age of the wife is 19, and (ii) the age of the wife when the age of the husband is 30.

Solution

Let $a = 26$ and $b = 17$ be the assumed means of x and y series respectively.

$$d_x = x - a = x - 26$$

$$d_y = y - b = y - 17$$

$$n = 10$$

x	y	d_x	d_y	d_x^2	d_y^2	d_xd_y
25	18	–1	1	1	1	–1
22	15	–4	–2	16	4	8
28	20	2	3	4	9	6
26	17	0	0	0	0	0
35	22	9	5	81	25	45
20	14	–6	–3	36	9	18
22	16	–4	–1	16	1	4
40	21	14	4	196	16	56
20	15	–6	–2	36	4	12
18	14	–8	–3	64	9	24
$\Sigma x = 256$	$\Sigma y = 172$	$\Sigma d_x = -4$	$\Sigma d_y = 2$	$\Sigma d_x^2 = 450$	$\Sigma d_y^2 = 78$	$\Sigma d_xd_y = 172$

$$b_{yx} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}$$

$$= \frac{172 - \frac{(-4)(2)}{10}}{450 - \frac{(-4)^2}{10}}$$

$$= 0.385$$

$$b_{xy} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}$$

$$= \frac{172 - \frac{(-4)(2)}{10}}{78 - \frac{(2)^2}{10}}$$

$$= 2.227$$

$$\bar{x} = \frac{\sum x}{n} = \frac{256}{10} = 25.6$$

$$\bar{y} = \frac{\sum y}{n} = \frac{172}{10} = 17.2$$

The equation of the regression line of y on x is

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

$$y - 17.2 = 0.385(x - 25.6)$$

$$y = 0.385x + 7.344$$

The equation of the regression line of x on y is

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

$$x - 25.6 = 2.227(y - 17.2)$$

$$x = 2.227y - 12.704$$

Estimated age of the husband when the age of the wife is 19 is

$x = 2.227\ (19) - 12.704 = 29.601$ or 30 nearly

Age of the husband = 30 years

Estimated age of the wife when the age of the husband is 30 is

$y = 0.385\ (30) + 7.344 = 18.894$ or 19 nearly

Age of the wife = 19 years

Example 15

From the following data, obtain the two regression lines and correlation coefficient.

Sales (x)	100	98	78	85	110	93	80
Purchase (y)	85	90	70	72	95	81	74

Solution

Let $a = 93$ and $b = 81$ be the assumed means of x and y series respectively.

$d_x = x - a = x - 93$

$d_y = y - b = y - 91$

$n = 7$

x	y	d_x	d_y	d_x^2	d_y^2	$d_x d_y$
100	85	7	4	49	16	28
98	90	5	9	25	81	45
78	70	−15	−11	225	121	165
85	72	−8	−9	64	81	72
110	95	17	14	289	196	238
93	81	0	0	0	0	0
80	74	−13	−7	169	49	91
$\sum x = 644$	$\sum y = 567$	$\sum d_x = -7$	$\sum d_y = 0$	$\sum d_x^2 = 821$	$\sum d_y^2 = 544$	$\sum d_x d_y = 639$

$$b_{yx} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sum d_x^2 - \frac{\left(\sum d_x\right)^2}{n}}$$

$$= \frac{639 - \frac{(-7)(0)}{7}}{821 - \frac{(-7)^2}{7}}$$

$$= 0.785$$

$$b_{xy} = \frac{\sum d_x dy - \frac{\sum d_x \sum d_y}{n}}{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}$$

$$= \frac{639 - \frac{(-7)(0)}{7}}{544 - \frac{(0)^2}{7}}$$

$$= 1.1746$$

$$\bar{x} = \frac{\sum x}{n} = \frac{644}{7} = 92$$

$$\bar{y} = \frac{\sum y}{n} = \frac{567}{7} = 81$$

The equation of regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$y - 81 = 0.785(x - 92)$$
$$y = 0.785x + 8.78$$

The equation of regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$
$$x - 92 = 1.1746(y - 81)$$
$$x = 1.1746 - 3.1426$$
$$r = \sqrt{b_{yx}\, b_{xy}} = \sqrt{(0.785)(1.1746)} = 0.9602$$

EXERCISE 4.3

1. The following are the lines of regression $4y = x + 38$ and $9y = x + 288$. Estimate y when $x = 99$ and x when $y = 30$. Also, find the means of x and y.

   [Ans.: $y = 43$, $x = 82$, $\bar{x} = 162$, $\bar{y} = 50$]

2. The equations of the two lines of regression are $x = 19.13 - 0.87y$ and $y = 11.64 - 0.50x$. Find (i) the means of x and y, and (ii) the coefficient of correlation between x and y.

   [Ans.: $\bar{x} = 15.79$, $\bar{y} = 3.74$, (ii) $r = -0.66$, $b_{yx} = -0.5$, $b_{xy} = 0.87$]

3. Given var(x) = 25. The equations of the two lines of regression are $5x - y = 22$ and $64x - 45y = 24$. Find (i) $\bar{x}$ and $\bar{y}$, (ii) r, and (iii) σ_y.

   [Ans.: $\bar{x} = 6$, $\bar{y} = 8$, (ii) $r = 1.87$ (iii) $\sigma_y = 0.2$]

4. In a partially destroyed laboratory record of analysis of correlation data the following results are legible. Variance = 9, the equations of the lines of regression $4x - 5y + 33 = 0$, $20x - 9y - 107 = 0$. Find (i) the mean values of x and y, (ii) the standard deviation of y, and (iii) the coefficient of correlation between x and y

   [Ans.: (i) $\bar{x} = 13$, $\bar{y} = 17$, (ii) $\sigma_y = 4$, (iii) $r = 0.6$]

5. From a sample of 200 pairs of observation, the following quantities were calculated:

   $\sum x = 11.34$, $\sum y = 20.78$, $\sum x^2 = 12.16$, $\sum y^2 = 84.96$, $\sum xy = 22.13$

   From the above data, show how to compute the coefficients of the equation $y = a + bx$.

   [Ans.: $a = 0.0005$, $b = 1.82$]

6. In the estimation of regression equations of two variables x and y, the following results were obtained:

   $\bar{x} = 90$, $\bar{y} = 70$, $n = 10$, $\Sigma(x - \bar{x})^2 = 6360$, $\Sigma(y - \bar{y})^2 = 2860$

   $\Sigma(x - \bar{x})(y - \bar{y}) = 3900$

   Obtain the two lines of regression.

   [Ans.: $x = 1.361\, y - 5.27$, $y = 0.613\, x + 14.812$]

7. Find the likely production corresponding to a rainfall of 40 cm from the following data:

	Rainfall (in cm)	Output (in quintals)
mean	30	50
SD	5	10
$r = 0.8$		

   [Ans.: 66 quintals]

8. The following table gives the age of a car of a certain make and annual maintenance cost. Obtain the equation of the line of regression of cost on age.

Age of a car	2	4	6	8
Maintenance	1	2	2.5	3

   [Ans.: $x = 0.325\, y + 0.5$]

9. Obtain the equation of the line of regression of y on x from the following data and estimate y for $x = 73$.

x	70	72	74	76	78	80
y	163	170	179	188	196	220

   [Ans.: $y = 5.31\, x - 212.57$, $y = 175.37$]

10. The heights in cm of fathers (x) and of the eldest sons (y) are given below:

x	165	160	170	163	173	158	178	168	173	170	175	180
y	173	168	173	165	175	168	173	165	180	170	173	178

Estimate the height of the eldest son if the height of the father is 172 cm and the height of the father if the height of the eldest son is 173 cm. Also, find the coefficient of correlation between the heights of fathers and sons.

[Ans.: (i) $y = 1.016\,x - 5.123$ (ii) $x = 0.476\,y + 98.98$
(iii) 169.97, 173.45 (iv) $r = 0.696$]

11. Find (i) the lines of regression, and (ii) coefficient of correlation for the following data:

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

[Ans.: (i) $y = 19.64 + 0.72\,x$, $x = 33.29 + 0.5\,y$, (ii) $r = 0.604$]

12. Find the line of regression for the following data and estimate y corresponding to $x = 15.5$.

x	10	12	13	16	17	20	25
y	19	22	24	27	29	33	37

[Ans.: $y = 1.21x + 7.71$, $y = 26.465$]

13. The following data give the heights in inches (x) and weights in lbs (y) of a random sample of 10 students:

x	61	68	68	64	65	70	63	62	64	67
y	112	123	130	115	110	125	100	113	116	126

Estimate the weight of a student of height 59 inches.

[Ans.: 126.4 lbs]

14. Find the regression equations of y on x from the data given below taking deviations from actual mean of x and y.

Price in rupees (x)	10	12	13	12	16	15
Demand (y)	40	38	43	45	37	43

Estimate the demand when the price is ₹20.

[Ans.: $y = -0.25\,x + 44.25$, $y = 39.25$]