Correlation Analysis

When two variables vary simultaneously in the same or opposite direction and change in the value of one is accompanied by a change in the variable of the other, then the two variables are said to be correlated.

Types :

① Positive and negative correlation

Two variables deviate in same direction ← Positive

Two variables deviate in opposite direction ← Negative

② Simple, partial or multiple correlation

Simple : when two variables are studied

Partial : Measure of correlation bet^n two variables when all the variables are kept constant

Multiple : when three or more variables are studied simultaneously.

③ Linear and Non-linear correlation

Amount of change in one variable tends to be a constant ratio to the change in the other. ← Linear

Amount of change in one variable doesn't bear a constant ratio to the change in the other ← Non-Linear

Coefficient of Correlation

The study of the extent or the degree of correlation that exists between two variables is called the coefficient of correlation.

It is denoted by r_{xy} or $r(x,y)$ or $\rho(x,y)$ or ρ or r.

Note

① $-1 \leq \rho \leq 1$

② for $\rho = 1$ ← correlation is perfect and positive

$\rho = -1$ ← correlation is perfect and negative

$\rho = 0$ ← no correlation i.e. independent variables

Covariance

Assume n-pairs of observations between two variables X and Y

$$(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$$

Then covariance is given by

$$\text{Cov}(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n} \left[\sum_i x_i y_i - \frac{1}{n} \left(\sum_i x_i \right) \left(\sum_i y_i \right) \right]$$

Q. Calculate the covariance of the following

① (1,6), (2,9), (3,6), (4,7), (5,8), (6,5), (7,12), (8,3), (9,17), (10,1).

Soln:

$$\Sigma x_i = 1+2+3+4+5+6+7+8+9+10 = 55$$

$$\Sigma y_i = 6+9+6+7+8+5+12+3+17+1 = 74$$

$$\Sigma x_i y_i = (1\times6)+(2\times9)+(3\times6)+(4\times7)+(5\times8)+(6\times5)+(7\times12)+(8\times3)+(9\times17)+(10\times1)$$

$$= 411$$

Now,

$$cov(x,y) = \frac{1}{n}\left[\Sigma x_i y_i - \frac{1}{n}\Sigma x_i \Sigma y_i\right]$$

$$= \frac{1}{10}\left[411 - \frac{1}{10}(55\times74)\right]$$

$$= \frac{1}{10}(411-407)$$

$$= 0.4$$

(2) (15,44), (20,43), (25,45), (30,37), (40,34), (50,37)

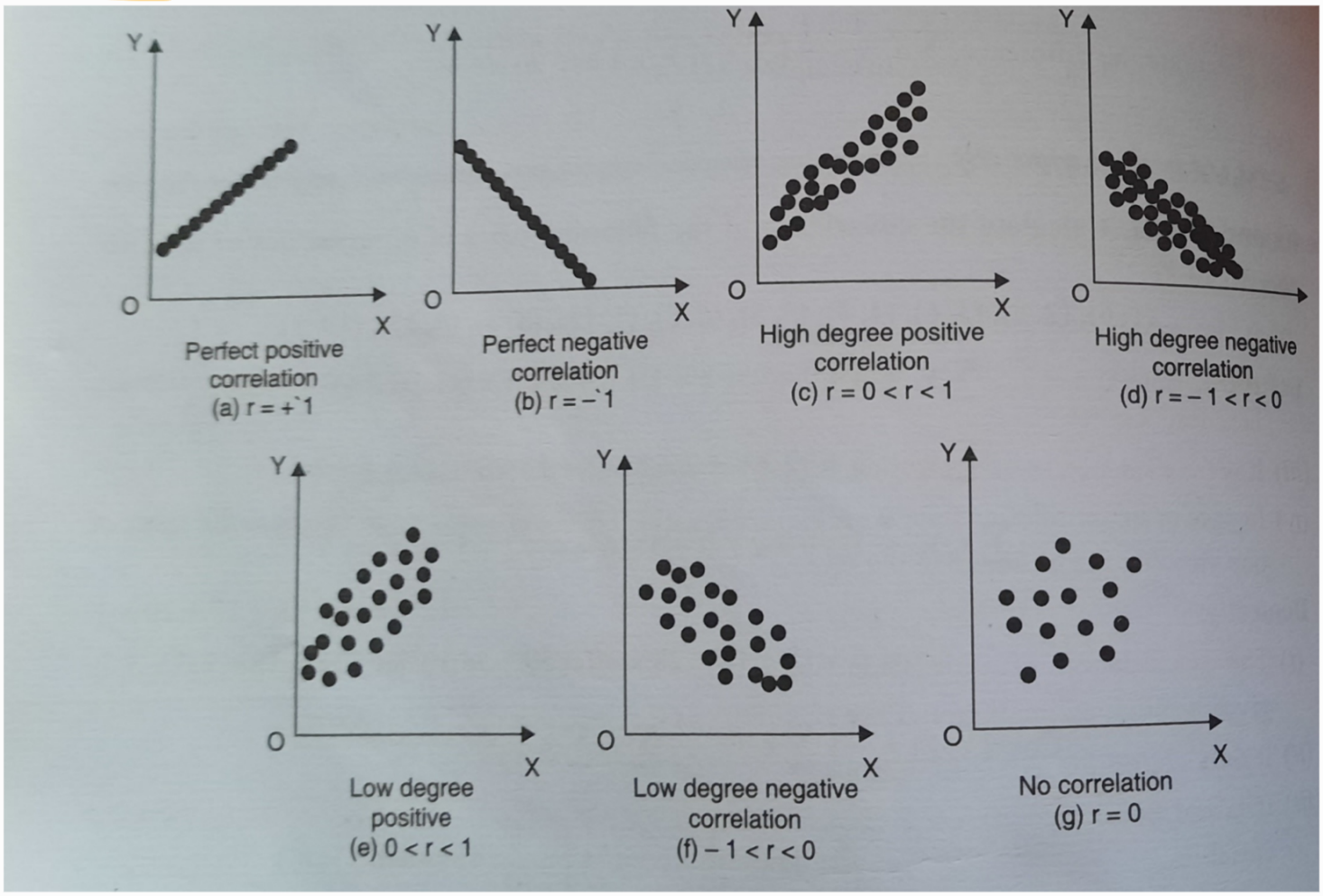
Methods of studying correlation

(i) Scatter Method

(ii) Karl Pearson's coefficient of correlation method

(iii) Spearman's Rank correlation coefficient method.

Scatter Method



Karl Pearson's Coefficient of Correlation

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

$$= \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\ \sqrt{\text{var}(Y)}}$$

Q. Find the coefficient of correlation from the following points of observation

$(1,3), (2,2), (3,5), (4,4), (5,6)$

x_i	y_i	$x_i - \bar{x} = x_i - 3$	$y_i - \bar{y} = y_i - 4$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	3	−2	−1	4	1	2
2	2	−1	−2	1	4	2
3	5	0	1	0	1	0
4	4	1	0	1	0	0
5	6	2	2	4	4	4
$\Sigma x_i = 15$	$\Sigma y_i = 20$	$\Sigma(x_i - \bar{x}) = 0$	$\Sigma(y_i - \bar{y}) = 0$	$\Sigma(x_i - \bar{x})^2 = 10$	$\Sigma(y_i - \bar{y})^2 = 10$	$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 8$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{15}{5} = 3 \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{20}{5} = 4$$

$$Cov(x,y) = \frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{5}(8) = 8/5$$

$$Var(x) = \frac{1}{n}\sum(x_i-\bar{x})^2 = \frac{1}{5}(10)$$

$$Var(y) = \frac{1}{n}\sum(y_i-\bar{y})^2 = \frac{1}{5}(10)$$

Now,

$$\rho(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}$$

$$= \frac{8/5}{\frac{\sqrt{10}}{\sqrt{5}} \times \frac{\sqrt{10}}{\sqrt{5}}}$$

$$= \frac{8}{10}$$

$$= 0.8$$

Q. Calculate the correlation coefficient betⁿ x & y

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

Spearman's Rank Correlation co-efficient method

There are three types of problems in this method

① When ranks are given

where actual ranks are given to us the steps required for computing rank correlation are

Step 1: Take the difference of two ranks i.e. $R_1 - R_2$ and denote these differences by D.

Step 2: Square these differences and obtain ΣD^2

Step 3: Apply the formula: $$R = 1 - \frac{6\Sigma D^2}{n(n^2-1)}$$

where n is the total no. of pairs of obs.

Q. Two ladies were asked to rank 7 different types of lipsticks. The ranks given by them are given below

Lipsticks	A	B	C	D	E	F	G
Anita	2	1	4	3	5	7	6
Sunita	1	3	2	4	5	6	7

Calculate Spearman's rank correlation coefficient.

Sol^n: Let the ranks given by Anita be R_1 & Sunita be R_2

R_1	R_2	$D = R_1 - R_2$	D^2
2	1	1	1
1	3	−2	4
4	2	2	4
3	4	−1	1
5	5	0	0
7	6	1	1
6	7	−1	1
			$\Sigma D^2 = 12$

Now,

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

Here, $\Sigma D^2 = 12$, $n = 7$

$$\therefore R = 1 - \frac{6(12)}{7(49-1)} = 1 - \frac{6(12)}{7(48)}$$

$$= 1 - \frac{12}{56}$$

$$= 0.786$$

Q. Ten competitors in a beauty contest are ranked by three judges in the following orders

1st Judge	1	6	5	10	3	2	4
2nd Judge	3	5	8	4	7	10	2
3rd Judge	6	4	9	8	1	2	3

1st Judge	9	7	8
2nd Judge	1	6	9
3rd Judge	10	5	7

Use the correlation coefficient to determine which pair of judges has the nearest approach to common taste in beauty.

Let R_1, R_2, R_3 respectively be the ranks given by 1st, 2nd and 3rd judge.

R_1	R_2	R_3	$D_{12} = R_1 - R_2$	$D_{13} = R_1 - R_3$	$D_{23} = R_2 - R_3$	D_{12}^2	D_{13}^2	D_{23}^2
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	9	64	1	1
7	6	8	1	2	1	1	4	81
8	9	7	-1	1	2	1	1	4

$\Sigma D_{12} = 0$ $\Sigma D_{13} = 0$ $\Sigma D_{23} = 0$ $\Sigma D_{12}^2 = 200$ $\Sigma D_{13}^2 = 60$ $\Sigma D_{23}^2 = 214$

Here, $n = 10$

$$R_{12} = 1 - \frac{6\Sigma D_{12}^2}{n(n^2-1)} = 1 - \frac{6(200)}{10(100-1)} = -0.21$$

$$R_{13} = 1 - \frac{6\Sigma D_{13}^2}{n(n^2-1)} = 1 - \frac{6(60)}{10(99)} = 0.64$$

$$R_{23} = 1 - \frac{6\Sigma D_{23}^2}{n(n^2-1)} = 1 - \frac{6(214)}{10(99)} = -0.29$$

∴ R_3 has max^m value, so 1st and 3rd Judge have common taste.

1st method to Q.1

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
1	6	-4.5	-1.4	6.3
2	9	-3.5	1.6	-5.6
3	6	-2.5	-1.4	3.5
4	7	-1.5	-0.4	0.6
5	8	-0.5	0.6	-0.3
6	5	0.5	-2.4	-1.2
7	12	1.5	4.6	6.9
8	3	2.5	-4.4	-11
9	17	3.5	9.6	33.6
10	1	4.5	-6.4	-28.8

$\Sigma(x-\bar{x})(y-\bar{y}) = 4$

$$\bar{x} = \frac{1+2+3+4+5+6+7+8+9+10}{10} = \frac{55}{10} = 5.5$$

$$\bar{y} = \frac{6+9+6+7+8+5+12+3+17+1}{10} = \frac{74}{10} = 7.4$$

$$Cov(x,y) = \frac{1}{n} \sum (x-\bar{x})(y-\bar{y})$$

$$= \frac{1}{10}(4)$$

$$= 0.4$$

II. When ranks are not given

In this case, we assign ranks to both the series x and y by giving the rank 1 to the highest values in both the series (or to the lowest values in both the series) and so on.

Q. From the data given below, calculate the coefficient of rank correlation bet^n x and y

X	78	89	97	69	59	79	68	57
Y	125	137	156	112	107	136	123	108

Soln: Here,

$n = 8$

X	Y	Rank in X (R_1)	Rank in Y (R_2)	$D = R_1 - R_2$	D^2
78	125	4	4	0	0
89	137	2	2	0	0
97	156	1	1	0	0
69	112	5	6	-1	1
59	107	7	8	-1	1
79	136	3	3	0	0
68	123	6	5	1	1
57	108	8	7	1	1
				$\Sigma D = 0$	$\Sigma D^2 = 4$

Coefficient of Rank Coefficient,

$$R = 1 - \frac{6\Sigma D^2}{n(n^2-1)} = 1 - \frac{6(4)}{8(64-1)}$$

$$= 0.95$$

Q. A random sample of 5 college students is selected and their grades in Mathematics and Statistics are found to be

Mathematics	85	60	73	40	90
Statistics	93	75	65	50	80

(iii) **When equal ranks are given for more than two attributes**

Assign average rank to each of the individuals having same rank. eg: Assume 6th and 7th items have same value then assign a common rank $\left(\frac{6+7}{2}\right) = 6.5$ to them and use the formula given below for soln purpose.

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

where $m_1, m_2, m_3, \ldots$ are the no. of times a value is repeated

9. From the following data of the marks obtained by 8 students in Mathematics and Physics paper compute rank of the coefficient of correlation

Mathematics	15	20	28	12	40	60	20	80
Physics	40	30	50	30	20	10	30	60

Soln: Here,

$n = 8$

Marks in Mathematics (X)	Marks in Physics (Y)	R_1	R_2	$D = R_1 - R_2$	D^2
15	40	2	6	−4	16
20	30	3.5	4	−0.5	0.25
28	50	5	7	−2	4
12	30	1	4	−3	9
40	20	6	2	4	16
60	10	7	1	6	36
20	30	3.5	4	−0.5	0.25
80	60	8	8	0	0
				$\Sigma D = 0$	$\Sigma D^2 = 81.5$

we know that,

$$R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right]}{n(n^2-1)} \quad \text{— (1)}$$

Now,

20 repeats itself two times in X

$\Rightarrow m_1 = 2$

and 40 repeats itself three times in Y

$\Rightarrow m_2 = 3$

$$\therefore R = 1 - \frac{6\left[81.5 + \frac{1}{12}\{(2)^3 - 2\} + \frac{1}{12}\{(3)^3 - 3\}\right]}{8(8^2 - 1)}$$

$$= 1 - \frac{504}{504}$$

$$= 0$$

Q. Calculate rank correlation coefficient

X	48	33	40	9	16	16	65	24
Y	13	13	24	6	15	4	20	9

16	57
6	19

Regression Analysis

Regression Lines

For a simple regression analysis, there are two regression lines — X on Y & Y on X. where X are Y are two variables.

* X on Y denotes that X is dependent & Y is independent
* Y on X denotes that Y is independent & X is dependent.

Assume $\bar{x}$ and $\bar{y}$ be the means for series X & Y respectively and σ_x and σ_y be their respective S.D.

let, r be the correlation coefficient

Then the regression line

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

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

Note

① $-1 \le r \le 1$

② $r \frac{\sigma_y}{\sigma_x}$ & $r \frac{\sigma_x}{\sigma_y}$ are called regression coefficients

It is denoted by b_{yx} & b_{xy}

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

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

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

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

Regression Lines of y on x

$y \leftarrow$ dependent

$x \leftarrow$ independent

Eqn of regression line : $y = a + bx$

Residual for i^{th} point is $\varepsilon_i = y_i - a - bx_i$

* **Normal equations**

$$\Sigma y = na + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

* Regression line of y on x

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

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

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

* Regression line of x on y

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

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

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

Q. Obtain the line of regression of y on x for the following data

x	1.53	1.78	2.60	2.45	3.42
y	33.50	36.30	40.00	45.80	53.50

Sol^n: Regression line of y on x is

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

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

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

Here, $n = 5$

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

$$\bar{x} = \frac{\Sigma x}{n} =$$

$$\bar{y} = \frac{\Sigma y}{n} =$$

Regression line of y on x is

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

Properties of Regression coefficients

① Correlation coefficient is the G.M. betⁿ the regression coefficients

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

③ A.M. of regression coefficients is greater than the correlation coefficients

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

⑤ The correlation coefficient and the two regression coefficients have same sign.

Angle betⁿ two lines of Regression

If θ be the acute angle betⁿ two regression lines in the case of two variables x and y then

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

when $r = 0$, $\theta = \pi/2$ then two lines of regression are perpendicular to each other.

When $r = \pm 1$, $\tan\theta$ so that $\theta = 0$ or π then lines of regression coincide and there is perfect correlation betⁿ the two variates x and y.

Q. The following data regarding the heights (y) and weights (x) of 100 college student are given

$$\Sigma x = 15000$$

$$\Sigma x^2 = 22,72,500$$

$$\Sigma y = 6,800$$

$$\Sigma y^2 = 4,63,025$$

$$\Sigma xy = 10,222,50$$

Find the eqn of regression line of height on weight

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

we know that

$$r^2 = b_{yx} \cdot b_{xy}$$

$$= 0.8 \times 0.2$$

$$= 0.16$$

$$\Rightarrow r = \sqrt{0.16} = 0.4$$

Q. Calculate linear regression co-efficients from the following

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

linear regression coefficients are given by

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

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

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

$\Sigma x = 36$ $\Sigma y = 107$ $\Sigma x^2 = 204$ $\Sigma y^2 = 1763$ $\Sigma xy = 599$

Here, $n = 8$

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{(8 \times 599) - (36 \times 107)}{(8 \times 204) - (36)^2} =$$

$$b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2} = \frac{(8 \times 599) - (36 \times 107)}{(8 \times 1763) - (107)^2} =$$

9. From the given data obtain the two regression equations using the method of least squares

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

Soln: We know that,

For Y on X, normal eqns are

$$\Sigma Y = na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad \text{—— (1)}$$

For X on Y, normal eqns are

$$\Sigma X = na + b\Sigma Y$$
$$\Sigma XY = a\Sigma Y + b\Sigma Y^2 \quad \text{—— (2)}$$

x	y	xy	x^2	y^2
2	5	10	4	25
4	7	28	16	49
6	9	54	36	81
8	8	64	64	64
10	11	110	100	121
$\Sigma x = 30$	$\Sigma y = 40$	$\Sigma xy = 266$	$\Sigma x^2 = 220$	$\Sigma y^2 = 340$

From (i),

$$\Sigma y = na + b\Sigma x$$

$$\Rightarrow 40 = 5a + b(30) \quad \text{---(i)} \quad (\text{Here } n = 5)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

$$\Rightarrow 266 = a(30) + b(220) \quad \text{---(ii)}$$

Solving (i) & (ii) we get

$$a = 4.1, \quad b = 0.65$$

∴ the reqd eqn is

$$y = a + bX$$

$$\Rightarrow y = 4.1 + 0.65X$$

Again,

From (2)

$$\Sigma X = na + b\Sigma y$$

$$\Rightarrow 30 = 5a + b(40) \quad \text{---(iii)}$$

and $$\Sigma xy = a\Sigma y + b\Sigma y^2$$

$$\Rightarrow 266 = a(40) + b(340) \quad \text{---(iv)}$$

Solving (iii) & (iv) we get

$$a = -4.4 \quad , \quad b = 1.3$$

∴ Reqd eqn is

$$x = a + by$$

$$\Rightarrow x = -4.4 + 1.3y$$

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

x	1	3	5	7	9
y	15	18	21	23	22

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