① (MU08)

Class: 08 Date: 11.10.23

Types of Correlation: Used to understand the reln betn variables.

1) Positive or negative correlation.
2) Simple or multiple
3) Linear or non-Linear

Correlation is the analysis of relationship betn two or more variable.

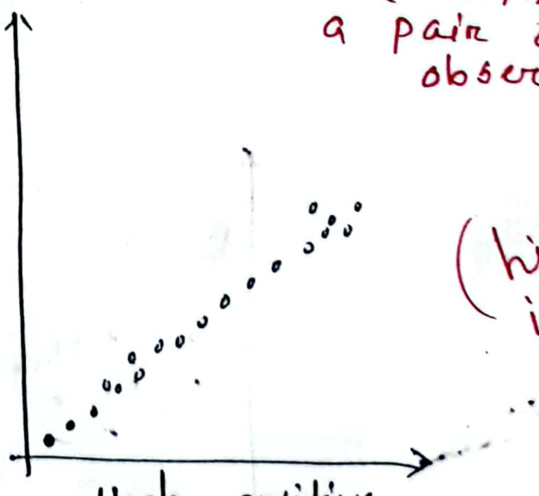
Positive Correlation: / Negative correlation

If the two variables deviate in the same direction, then its +ve correlation.

Scatter diagram => Suppose, $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ a pair of observation

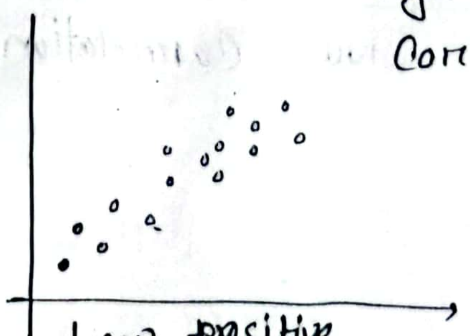
Not drastically change, Gradually slowly change

Perfect positive correlation



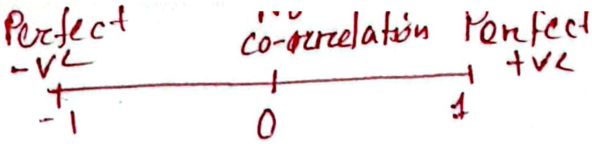
(hights increase)

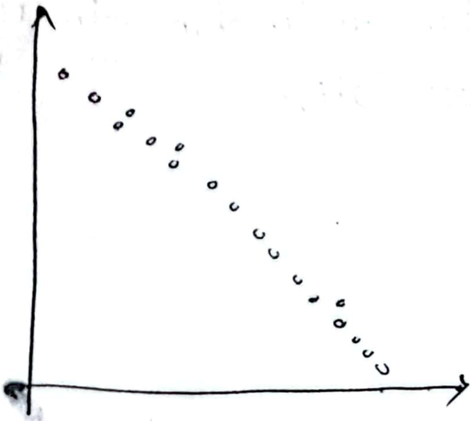
High positive Correlation



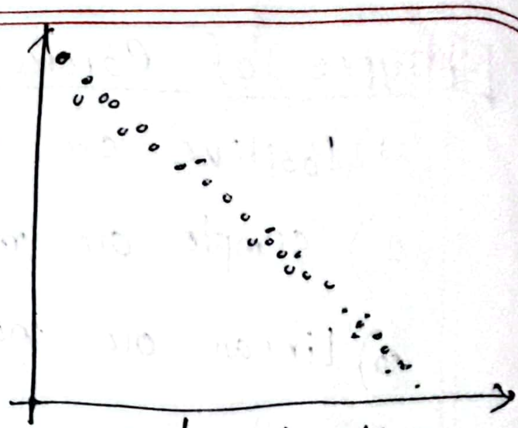
Some increases some decreases

Low positive correlation.





Perfect Negative correlation



High Negative correlation

Low Negative Correlation

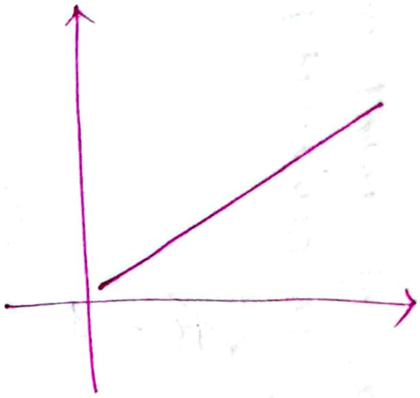
NO Correlation

"curved pattern" (weak correlation coefficient exists)

Linear and Non-Linear Correlation:

The graph of the variables having a linear relⁿ will form a straight line.

If not then Non-Linear.



X : 1, 2, 3, 4, 5, 6, 7, 8

Y : 5, 7, 9, 11, 13, 15, 17, 19

$(Y = 3 + 2x)$

$\boxed{\text{Correlation coefficient:}}$ [For Quantitative variable] (Numerical data) / (number based)

Measurement of strength of $\boxed{\text{Linear relationship}}$ between two or more variables.

Let, $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$ be the pairs of n observations. It is denoted by r_{xy},

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

This eqⁿ is called $\boxed{\text{Karl pearson's coefficient}}$ of correlation.

or,

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

(r is symmetric ; $r_{xy} = r_{yx}$)

$r = 1$ = Perfect positive correlation

$0.7 \le r < 1$ = strong/high positive correlation

$0.4 \le r < 0.7$ = Fairly positive correlation

$0 < c < 0.4$ = Weak positive correlation

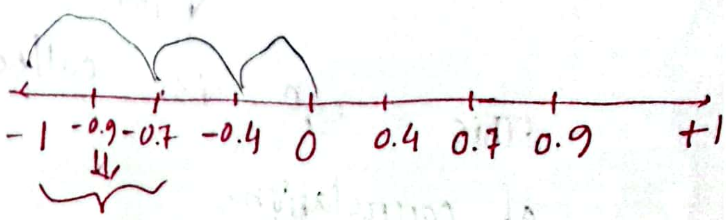
0 = No correlation (No Linear relationship betn x and y)

$-0.4 < c < 0$ = Weak neg cor.

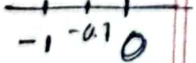
$-0.7 < c \le -0.4$ = fairly neg cor.

$-1 < c \le -0.7$ = Strong/high neg cor.

-1 = Perfect neg. cor.



($r = 0.30$; $r^2 = 0.09 \Rightarrow 9\%$ of total variation is explained)
(coefficient of determination)



2

Problem: Observe the following dataset and calculate the correlation co-efficient.

x_i : 68 65 70 62 60 55 58 65 69 63

y_i : 2 5 1 10 9 13 10 3 4 6

Soln:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
68	2	4624	4	136
65	5	4225	25	325
70	1	4900	1	70
62	10	3844	100	620
60	9	3600	81	540
55	13	3025	169	715
58	10	3364	100	580
65	3	4225	9	195
69	4	4761	16	276
63	6	3969	36	378

We know,

$$r_{xy} = \frac{(10 \times 3835) - (635 \times 63)}{\sqrt{(10 \times 40537) - (635)^2} \sqrt{(10 \times 541) - (63}}$$

$= -0.94$; $[-1 < r \leq -0.7]$

As $r = -0.94$; highly / strongly nega

correlation.

Problem: Compute r_{xy} for the following paired set of values.

(x, y) : $(1, 2)$, $(2, 3)$, $(3, 5)$, $(4, 4)$, $(5, 7$

Solⁿ:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	1	4	2
2	3	4	9	6
3	5	9	25	15
4	4	16	16	16
5	7	25	49	35

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

$$= \frac{(5 \times 74) - (15 \times 21)}{\sqrt{(5 \times 55) - (15)^2}\ \sqrt{(5 \times 103) - (21)^2}}$$

$$= 0.90$$

strongly positive correlation.

[Ans.]

(Abstract interpret ation nas

Rank Correlation: [Finding Rel^n between two qualita variables]

(Example: Gender, colour, ta smel

The Formula for computing the spearman's rank Correlation coefficient

$$R = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (d_i = x_i - y_i)$$

R = rank correlation coefficient

n = number of pairs of observations being ranked.

d = difference between rank of x and rank of y.

*Careful about two things:
1) Actual observations are given
2) Actual Ranks are given.

Problem: obtain the rank correlation coeffi[cient] for the following data:

A :	80	75	90	70	65	60
B :	65	70	60	75	85	80

Soln:

A	B	Rank of A (x)	Rank of B (y)	$d = x - y$	d^2
80	65	2	5	-3	9
75	70	3	4	-1	1
90	60	1	6	-5	25
70	75	4	3	1	1
65	85	5	1	4	16
60	80	6	2	4	16
					$\Sigma d^2 = 6[8]$

$\therefore$ Rank Correlation, $R = 1 - \frac{6 \Sigma d^2}{n(n^2-1)}$

$= 1 - \frac{6 \times 68}{6 \times (6^2 - 1)}$

$= -0.94$

Strongly negative correlation between A and B. [Ans.]

Repeated Ranks Correlation:

When ranks are repeated the following formula is used:

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots\right\}}{n(n^2-1)}$$

Problem:

Observe the following data:

Marks in math : 20 80 40 12 28 20 15 60

Marks in stat : 30 60 20 30 50 30 40 10.

Find Rank Correlation and comment

Soln:

$\frac{5+6}{2} = \frac{11}{2} = 5.5$

A	B		Rank of A (x)		Rank of B (y)		d	d^2
20	30	5.5	3.5	5	4	-0.5	+0.5	0.25
80	60	1	8	1	8	0	0	0
40	20	3	6	7	2	4	-4	16
12	30	8	1	5	4	-3	+3	9
28	50	4	5	2	7	-2	+2	4
20	30	5.5	3.5	5	4	-0.5	+0.5	0.25
15	40	7	2	3	6	-4	+4	16
60	10	2	7	8	1	6	6	36

(High ranking) (Low Ranking) Low Ranking High Rank.

$$\therefore R = 1 - \frac{6\{81.5 + \frac{1}{2}(2^3-2) + \frac{1}{2}(3^3-3)\}}{8(8^2-1)}$$

$$= 0$$

No correlation

[Ans]

Problem: The following figures relate to advertisement expenditure and profit:

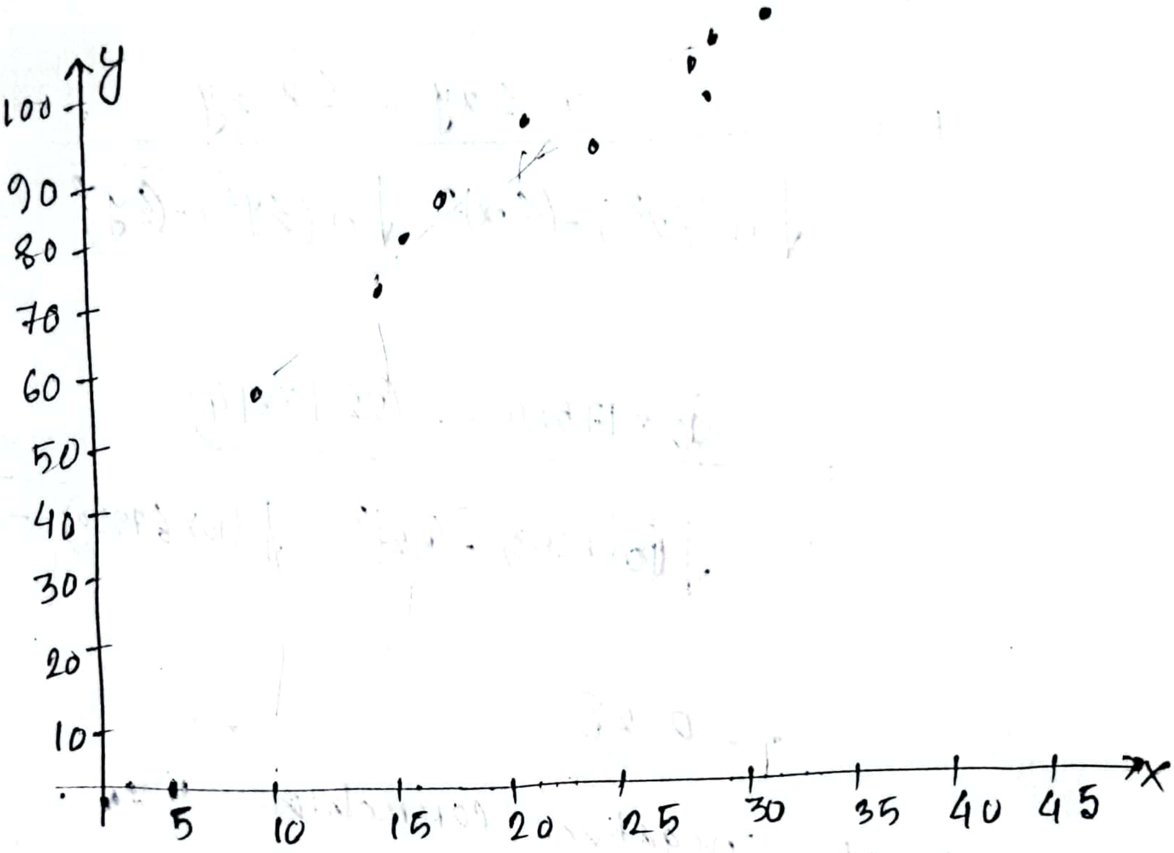
Profit (x):	25	28	27	33	31	10	16	16	18	23
Expenditure (y):	87	91	92	95	93	52	68	72	78	86

a) Draw a scatter diagram and comment.

b) Calculate Karl pearson's and Spearman rank correlation coefficients and comment.

Soln:

a)



Perfectly positive correlation.

(b)

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
25	87	2175	625	7569
28	91	2548	784	8281
27	92	2484	729	8464
33	95	3135	1089	9025
31	93	2883	961	8649
10	52	520	100	2704
16	68	1088	256	4624
16	72	1152	256	5184
18	78	1404	324	6084
23	86	1978	529	7396
$\Sigma x_i = 227$	$\Sigma y_i = 814$	17586	5653	67980

Karl Pearson's coefficient:

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

$$= \frac{(10 \times 17586) - (227 \times 814)}{\sqrt{(10 \times 5653) - (227)^2}\ \sqrt{(10 \times 67980) - (814)^2}}$$

$$= -0.28$$

Weak negative correlation.

Spearman Rank Correlation:

x_i	y_i	Rank(x)		Rank(y)		d	d^2
25	87	5	6	5	6	0	0
28	91	3	8	4	7	-1	+1
27	92	4	7	3	8	+1	1
33	95	1	10	1	10	0	0
31	93	2	9	2	9	0	0
40	52	10	1	10	1	0	0
16	68	8.5	2.5	9	2	-0.5	0.25
16	72	8.5	2.5	8	3	+0.5	0.25
18	78	7	4	7	4	0	0
23	86	6	5	6	5	0	0

United International University

Name (Optional)

ID No.

Section

Invigilator's Signature with date

Course Code

Trimester / Semester : Spring / Summer / Fall, 20........

Name of Exam : Class Test / Mid-term / Final

Date: ....

Qualitative
Quantitatio.

* **Correlation:** (Describe strength of variables)

=> Defines relation / connection between two or more variables

=> In correlation, we analyze if there is a consistent pattern between two variables.

=> There are no independent / dependent variables in correlation. We don't look out for cause and effects

Correlation
- Positive / negative correlation
- (Simple / Multiple)
- Linear / Non Linear correlation

(Correlation) simple $\longrightarrow$ Two var.

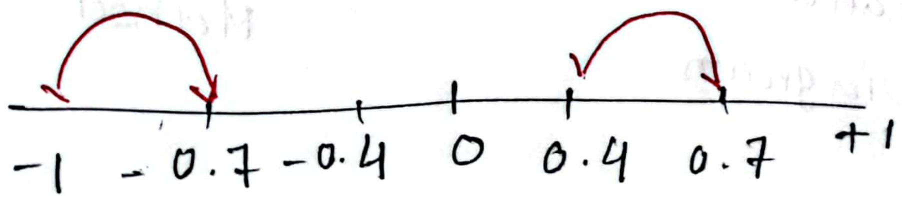
(Correlation) Multiple $\longrightarrow$ More than two var.

Problem:* Describe the strength of relation by using correlation coefficient $r = -0.72$ and $r = 0.41$?

Soln:

$r = -0.72$; it defines a strongly negative correlation between x and y.

$r = 0.41$; it defines fairly +ve correlation betn x and y.



United International University

Name (Optional)

ID No.

Section

Invigilator's Signature with date

Course Code

Trimester / Semester : Spring / Summer / Fall, 20........

Name of Exam : Class Test / Mid-term / Final

Date: ....

---

* Regression Line:

If we consider two variables x, Y.

So, here we have two regression lines.

→ Regression Line of Y on X (gives most probable values of Y for given value of x)

→ Regression Line of x on Y.

vice-versa

* Regression equation of Y on x:

$y = a + bx$

y: Dep. var.

b: (Regression co-efficient)

x: Ind. var.

$\therefore a = \bar{y} - b\bar{x}$

$= \frac{\Sigma Y}{n} - b\frac{\Sigma x}{n}$

* Regression Line of x on y :

$$x = a + by$$

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

$$= \frac{\Sigma x}{n} - b\frac{\Sigma y}{n}$$

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

m_1 = Repetition = 2

$$= 1 - \frac{6\left\{2.5 + \frac{1}{12}(2^3 - 2)\right\}}{10(10^2-1)}$$

$$= 1 - \frac{1}{55} = 0.98$$

Strongly positive correlation.

[Ans.]

(prediction of dependent variables)

average reln between two /more variables.

we look out here (cause and effect)

Regression Analysis:

b_{yx}: Dependent variable, Ind. var.

b_{xy}: Dep var., Ind var.

Regression equation:

The regression equation of y on x is expressed as follows: $\boxed{\bar{y} = a + b\bar{x}}$ where,

$$b_{yx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and } a = \bar{y} - b\bar{x}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

The regression equation of x on y is expressed $\boxed{x = a+by}$

$$b_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

and $a = \bar{x} - b\bar{y}$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}$$

Regression Co-efficient Properties:

1) It lies between $-\infty$ to $+\infty$

that is $-\infty < b_{yx} < \infty$

2) Reg. coeff. is not symmetric.

$b_{xy} \neq b_{yx}$.

3) G.M. of Reg.coeff is equal to Correlatn coeffici

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

4) A.M. of two reg.coeff is greater that corr. coeff.

$$\frac{b_{yx} + b_{xy}}{2} \geq r_{xy}$$

(5) If one reg. coeff is greater than unity the other must be less than unity that is

$b_{yx} \geq 1$ and $b_{xy} < 1$.

Problem: Observe the following sample:

x:	39	25	29	35	32	27	37
y:	37	18	20	25	25	20	30

a) compute Regression Line of y on x / using x ind. variable

b) Compute the value of correlation co-efficient r_{xy} with the help of Reg. coefficient.

c) Predict y when $x = 45$.

Soln: The equation of regression line of y on x is $y = a + bx$

Here,

$$b_{yx} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

x_i	y_i	x_i^2	y_i^2	x_iy_i
39	37	1521	1369	1443
25	18	625	324	450
29	20	841	400	580
35	25	1225	625	875
32	25	1024	625	800
27	20	729	400	540
37	30	1369	900	1110

$$b_{yx} = \frac{5798 - \frac{(224 \times 175)}{7}}{7334 - \frac{(224)^2}{7}} = 1.193$$

And $a = \bar{y} - b\bar{x}$

$$= \frac{\Sigma y}{n} - b\frac{\Sigma x}{n}$$

$$= \frac{175}{7} - \left(1.193 \times \frac{224}{7}\right)$$

$$= -13.176$$

$\therefore$ Regression line : $\hat{y} = a + bx = -13.176 + 1.193x$

(c) $\hat{y} = -13.176 + (1.193 \times 45) = 40.51$ years

Correlation coefficient with the help of Reg. coefficient:

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

$$= \sqrt{1.193 \times 0.739}$$

$$= 0.938$$

[Ans.]

[$b_{yx} = 1.193$;
$b_{xy} = 0.739$]

Just Do the calculation part of b_{xy}

Problem:

Temperature of water:

68 65 70 62 60 55 58 65 69 63

Reduction in pulse rate:

2 5 1 10 9 13 10 3 4 6

a) Calculate the correlation coefficient and regre
co-efficient

b) Also show that $\frac{b_{yx} + b_{xy}}{2} \geq r_{xy}$

Soln: a)

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
68	2	4624	4	136
65	5	4225	25	325
70	1	4900	1	70
62	10	3844	100	620
60	9	3600	81	540
55	13	3025	169	715
58	10	3364	100	580
65	3	4225	9	195
69	4	4761	16	276
63	6	3696	36	378

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

$$= -0.94$$

The Reg. coeff. of y on x is,

$$b_{yx} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

$$= \frac{3835 - \frac{635 \times 63}{10}}{40537 - \frac{(635)^2}{10}} = -0.77$$

and $$b_{xy} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}$$

$$= -1.1$$

(b) $$\frac{b_{yx} + b_{xy}}{2} = \frac{-0.77 - 1.1}{2}$$

$$= -0.935 \approx -0.94$$

which is equal to r_{xy}. [Showed]

Problem: Consider the following data set:

x : 1 2 3 4 5 6

y : 6 4 3 5 4 2

a) Find the equation of the regression Line y on x.

b) Graph the Line on a scatter diagram.

c) Estimate the value of y when $x = 4.5$.

d) Predict the value of y when $x = 8$.

Soln: a)

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	6	1	36	6
2	4	4	16	8
3	3	9	9	9
4	5	16	25	20
5	4	25	16	20
6	2	36	4	12
$\sum x_i = 21$	$\sum y_i = 24$	$\sum x_i^2 = 91$	$\sum y_i^2 = 106$	$\sum x_i y_i =$

$$b_{yx} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum x_i)^2}{n}} \qquad \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$= \frac{75 - \frac{(21 \times 24)}{6}}{91 - \frac{(21)^2}{6}}$$

$$= \cancel{0.00698} \quad -0.514$$

$$\hat{y} = a + bx$$

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

$$= \frac{\sum y}{n} - b\frac{\sum x}{n}$$

$$= \frac{24}{6} + (\overset{0.514}{\cancel{0.00698}}) \times \frac{21}{6}$$

$$= 4 + \cancel{0.02443}$$

$$= \cancel{3.97557} \quad 5.799$$

$\therefore$ Regression Line : $\hat{y} = a + bx$

$$\cancel{= 3.975 + 0.00698x}$$

$$= 5.799 - 0.514x \quad \text{[Ans]}$$

(c) From (a), $y = a + bx$

$= \cancel{3.975 + (0.00698 \times 4.5)}$

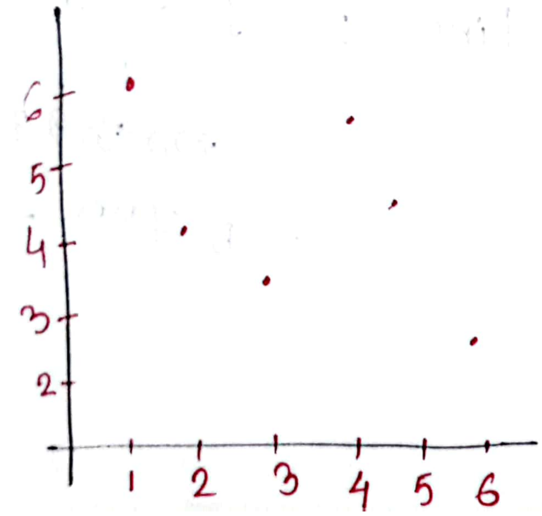
$= \cancel{4.00641} = 3.486$

(d) Again, $Y = \cancel{3.975 + (0.00698 \times}$

$= \cancel{4.03084}$ 1.687

[Ans.]

b)



Problem:

Annual sales (x): 90 75 78 86 95 110 130 145

Year of Experience (y): 7 4 5 6 11 12 13 17

(i) Fit two regression lines.

(ii) Estimate sales for Experience is 10

(iii) Estimate year of experience for sale is 100000.

Soln:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
90	7	8100	49	630
75	4	5625	16	300
78	5	6084	25	390
86	6	7396	36	516
95	11	9025	121	1045
110	12	12100	144	1320
130	13	16900	169	1690
145	17	21025	289	2465
$\Sigma x = 809$	$\Sigma y = 75$	$\Sigma x^2 = 86255$	$\Sigma y^2 = 849$	$\Sigma xy = 8356$

$$b_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}$$

$$= \frac{8356 - \frac{809 \times 75}{8}}{879 - \frac{(75)^2}{8}}$$

$$= 4.38$$

$$b_{yx} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

$$= \frac{8356 - \frac{809 \times 75}{8}}{86255 - \frac{(809)^2}{8}}$$

$$= 0.1735$$

For b_{yx}, $\quad y = a + bx$

$$\therefore a = \bar{y} - b\bar{x} = \frac{\sum y}{n} - b\frac{\sum x}{n}$$

$$= \frac{75}{8} - \left(0.1735 \times \frac{809}{8}\right)$$

$$= -8.17$$

$\therefore \hat{y} = -8.17 + 0.1735x$

Similarly, $\hat{x} = a + by$

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

$= \frac{\Sigma x}{n} - b\frac{\Sigma y}{n}$

$= 60.06$

$\therefore \hat{x} = 60.06 + 4.38y$

(ii) $\hat{x} = 60.06 + (4.38 \times 10)$

$= 103.86$

(iii) $\hat{y} = -8.17 + (0.1735 \times 100000)$

$= 17341.83$

[Ans.]

Problem: (34) If the correlation co-efficient of two variables is 0.72 and Reg. coefficient of x on y is 1.08. If $\bar{x} = 29.2$ and $\bar{y} = 37.5$. Find the regression line of y on x. Also, find the value of y when $x = 44$.

Soln: Given, $r_{xy} = 0.72$

$b_{xy} = 1.08$

$\bar{x} = 29.2$; $\bar{y} = 37.5$

$\hat{y} = a + bx$

~~Then, $a = \bar{y} - b\bar{x}$~~

~~$= 37.5 -$~~

~~Also, $\hat{x} = a + by$~~

~~Now, $a = \bar{x} - b\bar{y}$~~

~~$= 29.2 - (1.08 \times 37.5)$~~

~~$= -375.8$~~

Step 3:

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

$\Rightarrow 0.72 = \sqrt{1.08 \times b_{yx}}$

$\Rightarrow (0.72)^2 = 1.08 \times b_{yx}$

$\therefore b_{yx} = 0.48$

Step 2:

Now, $a = \bar{y} - b\bar{x}$

$= 37.5 - (0.48 \times 29.2)$

$= 23.48$

Step 1:

Reg. Line of y on x is

$\hat{y} = a + bx$

$= 23.48 + 0.48x$

Now, $y = 23.48 + (0.48 \times 44)$

$= 44.6$

[Ans.]