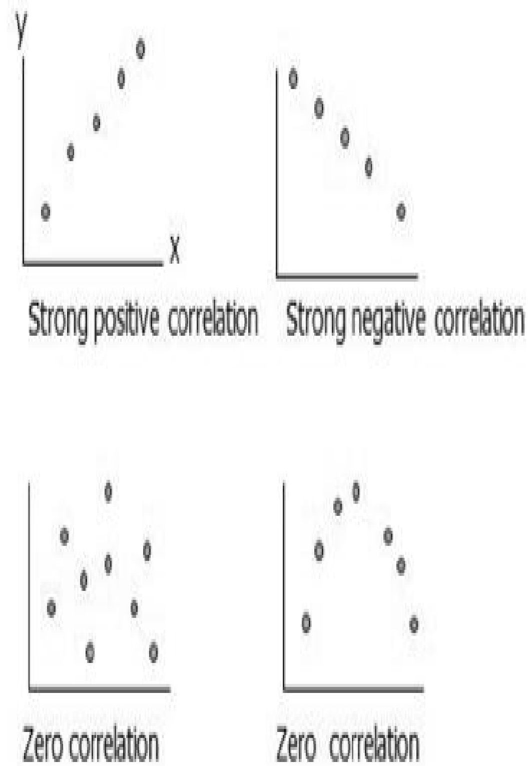
Week 6

Correlation

Correlation is a measure of association between two variables; which may be dependent or independent. Whenever two variables x and y are so related; that increase in one in accompanied by an increase or decrease in the other, then the variables are said to be correlated. Coefficient of correlation (r) lies between -1 and $+1$, i.e. $-1 \leq r \leq 1$.



If r is zero; no correlation between two variables, positive correlation ($0 < r \leq +1$); when both variables increase or decrease simultaneously, and negative correlation ($-1 \leq r < 0$); when increase in one is associated with decrease in other variable and vice-versa.

Correlation Analysis

Correlation analysis is used to measure the strength of the relationship between two variables. It is represented as a number. The correlation coefficient is a measure of how closely related two data series are. In particular, the correlation coefficient measures the direction and extent of **linear** association between two variables.

Characteristics of the correlation coefficient

A correlation coefficient has no units. The sample correlation coefficient is denoted by *r*.

- The value of *r* is always $-1 \le r \le 1$.
- A value of *r* greater than 0 indicates a positive linear association between the two variables.
- A value of *r* less than 0 indicates a negative linear association between the two variables.
- A value of *r* equal to 0 indicates no linear relation between the two variables.

Calculating and Interpreting the Correlation Coefficient

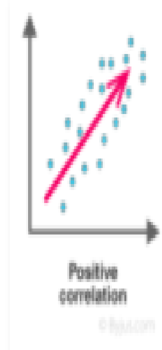
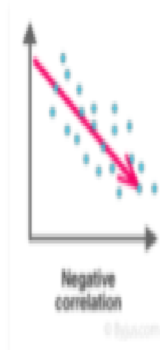
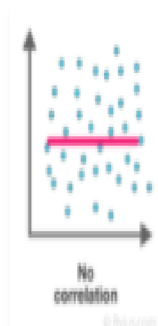
In order to calculate the correlation coefficient between two variables, X and Y, we need the following:

1. Covariance between X and Y, denoted by Cov (X,Y)
2. Standard deviation of X, denoted by σ_x
3. Standard deviation of Y, denoted by σ_y

How to find the Correlation Coefficient

Correlation is used almost everywhere in statistics. Correction illustrates the relationship between two or more variables. It is expressed in the form of a number that is known as correlation coefficient. There are mainly two types of correlations:

- **Positive Correlation**
- **Negative Correlation**

Positive Correlation	The value of one variable increases linearly with increase in another variable. This indicates a similar relation between both the variables. So its correlation coefficient would be positive or 1 in this case.	
Negative Correlation	When there is a decrease in values of one variable with decrease in values of other variable. In that case, correlation coefficient would be negative.	
Zero Correlation or No Correlation	There is one more situation when there is no specific relation between two variables.	

Properties Of Correlation Coefficient

Correlation coefficient **r** is all about establishing relationships between two variables. Some properties of correlation coefficient are as follows:

1. The value of r ranges from – 1.0 to 0.0 or from 0.0 to 1.0
2. A value of r = 1.0 indicates that there exists perfect positive correlation between the two variables.
3. A value of r = - 1.0 indicates that there exists perfect negative correlation between the two variables.
4. A value r = 0.0 indicates zero correlation i.e., it shows that there is no correlation at all between the two variables.
5. A positive value of r shows a positive correlation between the two variables.
6. A negative value of r shows a negative correlation between the two variables.
7. A value of r = 0.9 and above indicates a very high degree of positive correlation between the two variables.
8. A value of $-0.9 \geq r > -1.0$ shows a very high degree of negative correlation between the two variables.
9. For a reasonably high degree of positive correlation, we require r to be from 0.75 to 1.0.
10. A value of r from 0.6 to 0.75 may be taken as a moderate degree of positive correlation.

Karl Pearson Coefficient of Correlation

Coefficient of correlation (r) between two variables x and y is defined as

$$r = \frac{Covariance\,(x,y)}{\sqrt{Variance\,(x)}\sqrt{Variance\,(y)}} = \frac{\sum d_x d_y}{\sqrt{(\sum d_x^2)(\sum d_y^2)}} = \frac{\rho}{\sigma_x \sigma_y}$$

where $d_x = x - \bar{x}$, $d_y = y - \bar{y}$, $\bar{x}$, $\bar{y}$ are means of x and y data values.

$\rho = Cov\,(x, y) = \dfrac{\sum d_x d_y}{n}$ is the covariance between the variables x and y.

Also $\sigma_x = \sqrt{\dfrac{\sum d_x^2}{n}}$ and $\sigma_y = \sqrt{\dfrac{\sum d_y^2}{n}}$

Alternatively,

Calculating the Correlation Coefficient

Sample correlation coefficient:

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

or the algebraic equivalent:

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

where:

r = Sample correlation coefficient

n = Sample size

x = Value of the independent variable

y = Value of the dependent variable

Example:1

If $Cov\,(x,y) = 10$, $var(x) = 25$, $var(y) = 9$ find coefficient of correlation.

Solution: $r = \dfrac{Cov\,(x,y)}{\sqrt{Var(x)}\,\sqrt{Var(y)}} = \dfrac{10}{\sqrt{25}\,\sqrt{9}} = \dfrac{10}{5\times 3} = 0.67$

Example: 2

Calculation Example

Tree Height	Trunk Diameter			
y	x	xy	y^2	x^2
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
$\Sigma=321$	$\Sigma=73$	$\Sigma=3142$	$\Sigma=14111$	$\Sigma=713$

Calculation

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

$$= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}}$$

$$= 0.886$$

r = 0.886 → relatively strong positive linear association between x and y

Example:

A study is conducted involving 10 students to investigate the association between statistics and science tests. The question arises here; is there a relationship between the degrees gained by the 10 students in statistics and science tests?

Table: Student degree in Statistic and science

Students	1	2	3	4	5	6	7	8	9	10
Statistics	20	23	8	29	14	12	11	20	17	18
Science	20	25	11	24	23	16	12	21	22	26

Calculation:

Notes: the marks out of 30

Suppose that (x) denotes for statistics degrees and (y) for science degree

Calculating the mean $(\bar{x}\ ,\ \bar{y})$;

$$\bar{x} = \frac{\sum x}{n} = \frac{173}{10} = 17.3\ ,\ \bar{y} = \frac{\sum y}{n} = \frac{200}{10} = 20$$

Where the mean of statistics degrees $\bar{x} = 17.3$ and the mean of science degrees $\bar{y} = 20$

Statistics	Science					
x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
20	20	2.7	7.29	0	0	0
23	25	5.7	32.49	5	25	28
8	11	-9.3	86.49	-9	81	83
29	24	11.7	136.89	4	16	46
14	23	-3.3	10.89	3	9	-9.9
12	16	-5.3	28.09	-4	16	21.2
11	12	-6.3	39.69	-8	64	50.4
21	21	3.7	13.69	1	1	3.7
17	22	-0.3	0.09	2	4	-0.6
18	26	0.7	0.49	6	36	4.2
173	200	0	356.1	0	252	228

$$\sum(x-\bar{x})^2 = 356.1 , \sum(y-\bar{y})^2 = 252 ,$$

$$\sum(x-\bar{x})(y-\bar{y}) = 228$$

Calculating the Pearson correlation coefficient;

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{228}{\sqrt{356.1}\sqrt{252}}$$

$$= \frac{228}{(18.8706)(15.8745)} = \frac{228}{299.5614} = 0.761$$

Other solution

Also; the Pearson correlation coefficient is given by the following equation:

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

x	y	xy	x^2	y^2
20	20	400	400	400
23	25	575	529	625
8	11	88	64	121
29	24	696	841	576
14	23	322	196	529
12	16	192	144	256
11	12	132	121	144
21	21	441	441	441
17	22	374	289	484
18	26	468	324	676
173	200	3688	3349	4252

Required calculation

$\sum x = 173$, $\sum y = 200$

$\sum xy = 3688$

$\sum x^2 = 3349$

$\sum y^2 = 4252$

Calculating the Pearson correlation coefficient by substitute in the aforementioned equation;

$$r = \frac{3688 - \frac{(173)(200)}{10}}{\sqrt{\left(3349 - \frac{(173)^2}{n10}\right)\left(4252 - \frac{(200)^2}{10n}\right)}} = \frac{228}{\sqrt{(356.1)(252)}} = \frac{228}{299.5614} = 0.761$$

The calculation shows a strong positive correlation (0.761) between the student's statistics and science degrees. This means that as degrees of statistics increases the degrees of science increase also. Generally the student who has a high degree in statistics has high degree in science and vice versa.

Example: Calculate coefficient of correlation from the following data:

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

Solution:

x	$(x-\bar{x})$	$(x-\bar{x})^2$	y	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
9	4	16	15	3	9	12
8	3	9	16	4	16	12
7	2	4	14	2	4	4
6	1	1	13	1	1	1
5	0	0	11	-1	1	0
4	-1	1	12	0	0	0
3	-2	2	10	-2	4	4
2	-3	9	8	-4	16	12
1	-4	16	9	-3	9	12
$\sum x = 45$ $\bar{x} = 5$		$\sum(x-\bar{x})^2$ $=60$	$\sum y = 108$ $\bar{y} = 12$		$\sum(y-\bar{y})^2$ $=60$	$\sum(x-\bar{x})(y-\bar{y})$ $= 57$

$$\sum(x-\bar{x})^2 = 60 \quad , \sum(y-\bar{y})^2 = 60 \ ,$$

$$\sum(x-\bar{x})(y-\bar{y}) = 57$$

Calculating the Pearson correlation coefficient;

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{57}{\sqrt{60}\sqrt{60}}$$

$$= \frac{57}{60} \quad = 0.95$$

Example :3 Calculate coefficient of correlation from the following data:

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

Ans: $r = 0.9879$

Example:

Calculate the co-efficient of correlation between the values of X and Y given below:

X	78	89	97	69	59	79	68	61
Y	125	137	156	112	107	136	123	108

Let $u = X - 69$ and $v = Y - 112$. Then $r_{xy} = r_{uv}$. The calculations needed to find r are

X	Y	u	v	u^2	v^2	uv
78	125	9	13	81	169	117
89	137	20	25	400	625	500
97	156	28	44	784	1936	1232
69	112	0	0	0	0	0
59	107	−10	−5	100	25	50
79	136	10	24	100	576	540
68	123	−1	11	1	121	−11
61	108	−8	−4	64	16	32
600	1004	48	108	1530	3468	2160

Now $$r = \frac{\sum uv - (\sum u)(\sum v)/n}{\sqrt{\left[\sum u^2 - \frac{(\sum u)^2}{n}\right]\left[\sum v^2 - \frac{(\sum v)^2}{n}\right]}}$$

$$= \frac{2160 - \frac{48 \times 108}{8}}{\sqrt{\left[1530 - \frac{(48)^2}{8}\right]\left[3468 - \frac{(108)^2}{8}\right]}}$$

$$= \frac{\mathbf{1512}}{\mathbf{1578}}$$

$$= \mathbf{0.96}$$

Hence the correlation coefficient between X and Y is 0.96

Coefficient of Correlation by Rank differences

Rank correlation is used for attributes (like beauty, intelligence etc.) which cannot be measured quantitatively but can be provided with comparative ranks.

Spearman's Rank Correlation in given by: $r = 1 - \frac{6\sum D^2}{n(n^2-1)}$, where $D = R_1 - R_2$

Tied Ranks: If two or more observations in a data are equal, each observation is provided with an average rank and a correction factor is applied to correlation formula given as: Correction Factor (C.F.) = $\sum m(m^2 - 1)$, m is the number of times each observation is repeated.

Spearman's Rank Correlation for repeated ranks is given by:

$$r = 1 - \frac{6\left(\sum D^2 + \frac{1}{12}\text{C.F.}\right)}{n(n^2-1)}, \text{ where } D = R_1 - R_2$$

Example: Calculate the coefficient of correlation from the following data; given ranks of 10 students in English and Mathematics.

Rank in English	3	1	5	4	2	6	8	10	9	7
Rank in Mathematics	2	4	3	1	5	10	7	9	8	6

Solution: Since comparative ranks are given; instead of marks, using Spearman's Rank Correlation is given by: $r = 1 - \frac{6\sum D^2}{n(n^2-1)}$, where $D = R_1 - R_2$

Rank in English R_1	Rank in Mathematics R_2	$D = R_1 - R_2$	D^2
3	2	1	1
1	4	-3	9
5	3	2	4
4	1	3	9
2	5	-3	9
6	10	-4	16
8	7	1	1
10	9	1	1
9	8	1	1
7	6	1	1
			$\sum D^2 = 52$

$$\therefore r = 1 - \frac{6(52)}{10(10^2-1)} = 0.6848$$

Example 10.10 Find the co-efficient of rank correlation from the following rankings of 10 marks in Statistics and Mathematics.

Statistics (x):	1	2	3	4	5	6	7	8	9	10
Mathematics (y):	2	4	3	1	7	5	8	10	6	9

x_i	y_i	$d_i(=x_i-y_i)$	d_i^2
1	2	−1	1
2	4	−2	4
3	3	0	0
4	1	3	9
5	7	−2	4
6	5	1	1
7	8	−1	1
8	10	−2	4
9	6	3	9
10	9	1	1
---	---	0	34

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6\times 34}{10\times 99} = 1 - 0.2 = +0.8.$$

This indicates a high correlation between Statistics and Mathematics.

Example: Eight competitors in a beauty contest got marks (out of 10) by three judges as given:

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

Use rank correlation to discuss which pair of judges has the nearest approach to common tastes in beauty.

Solution: Since instead of ranks; marks are given by the three judges, converting the given data to comparative ranks for the eight competitors

Judge A		Judge B		Judge C		D_{AB}	D_{AB}^2	D_{BC}	D_{BC}^2	D_{AC}	D_{AC}^2
Marks	Rank	Marks	Rank	Marks	Rank						
9	2	3	6	6	4	-4	16	2	4	-2	4
6	3	5	4	4	5	-1	1	-1	1	-2	4
5	4	8	2	9	2	2	4	0	0	2	4
10	1	4	5	8	3	-4	16	2	4	-2	4
3	6	7	3	1	8	3	9	-5	25	-2	4
1	8	10	1	2	7	7	49	-6	36	1	1
4	5	2	7	3	6	-2	4	1	1	-1	1
2	7	1	8	10	1	-1	1	7	49	6	36

Here D_{AB} = Rank by Judge A – Rank by Judge B, also $\sum D_{AB}^2 = 100$

Similarly D_{BC} = Rank by Judge B – Rank by Judge C, also $\sum D_{BC}^2 = 120$

D_{AC} = Rank by Judge A – Rank by Judge C, also $\sum D_{AC}^2 = 58$

Rank Correlation between judges A and B is given by:

$$r_{AB} = 1 - \frac{6\sum D_{AB}^2}{n(n^2-1)} = 1 - \frac{6(100)}{8(8^2-1)} = -0.1905$$

Rank Correlation between judges B and C is given by:

$$r_{BC} = 1 - \frac{6\sum D_{BC}^2}{n(n^2-1)} = 1 - \frac{6(120)}{8(8^2-1)} = -0.4286$$

Rank Correlation between judges A and C is given by:

$$r_{AC} = 1 - \frac{6\sum D_{AC}^2}{n(n^2-1)} = 1 - \frac{6(58)}{8(8^2-1)} = 0.3095$$

Therefore Judges A and C have the nearest approach to common tastes in beauty, while Judges B and C have most different beauty tastes.

Example Obtain rank correlation coefficient for following marks in economics (x) and Mathematics (y) out of 25 for eight students.

x	20	24	12	20	10	12	24	20
y	18	19	16	22	14	16	19	12

Solution: Converting data into ranks: Ranks of x as R_x , Ranks of y as R_y

X	R_x	Y	R_y	$D = R_x - R_y$	D^2
20	4	18	4	0	0
24	1.5	19	2.5	-1	1
12	6.5	16	5.5	1	1
20	4	22	1	3	9
10	8	14	7	1	1
12	6.5	16	5.5	1	1
24	1.5	19	2.5	-1	1
20	4	12	8	-4	16
					$\sum D^2 = 30$

Correction Factor $= \sum m(m^2 - 1)$, m is the number of times each data value is repeated $\therefore$ $\text{C.F.} = 2(2^2 - 1) + 3(3^2 - 1) + 2(2^2 - 1) + 2(2^2 - 1) + 2(2^2 - 1)$

$$= 6 + 24 + 6 + 6 + 6 = 48$$

Spearman's Rank Correlation for repeated ranks is given by:

$$r = 1 - \frac{6\left(\sum D^2 + \frac{1}{12}\text{C.F.}\right)}{n(n^2-1)}, \text{ where } D = R_x - R_y$$

$$\therefore r = 1 - \frac{6\left(30+\frac{48}{12}\right)}{8(8^2-1)} = \frac{25}{42} = 0.595$$

Example Obtain rank correlation coefficient for following data

x	68	64	75	50	64	80	75	40	55	64
y	62	58	68	45	81	60	68	48	50	70

Solution: Converting data into ranks: Ranks of x as R_x , Ranks of y as R_y

x	R_x	Y	R_y	$D = R_x - R_y$	D^2
68	4	62	5	-1	1
64	6	58	7	-1	1
75	2.5	68	3.5	-1	1
50	9	45	10	-1	1
64	6	81	1	5	25
80	1	60	6	-5	25
75	2.5	68	3.5	-1	1
40	10	48	9	1	1
55	8	50	8	0	0
64	6	70	2	4	16
					$\sum D^2 =$ 72

Correction Factor (C.F.) $= \sum m(m^2-1)$, m is the number of times each data value is repeated $\therefore$ C.F. $= 2(2^2-1)+3(3^2-1)+2(2^2-1)=36$

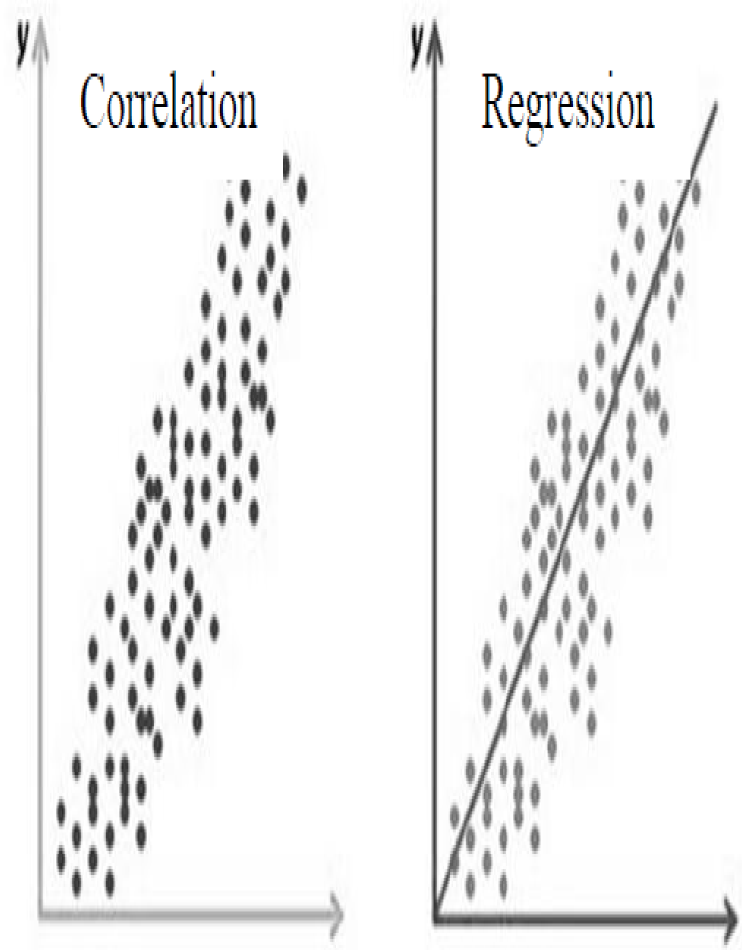
Spearman's Rank Correlation for repeated ranks is given by:

$$r = 1 - \frac{6\left(\sum D^2 + \frac{1}{12}\text{C.F.}\right)}{n(n^2-1)}, \text{ where } D = R_x - R_y$$

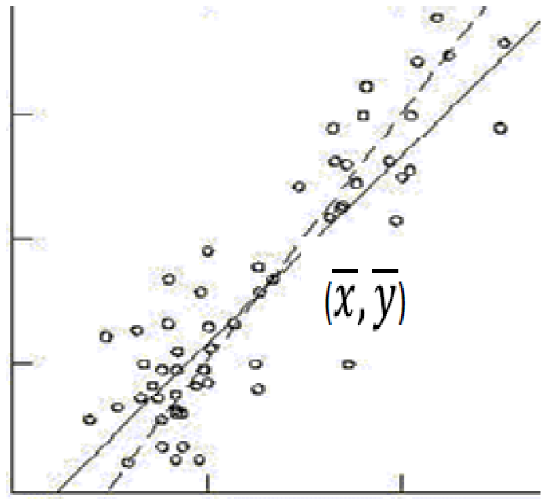
$$\therefore r = 1 - \frac{6\left(72+\frac{36}{12}\right)}{10(10^2-1)} = \frac{6}{11} = 0.545$$

Linear Regression

Regression describes the functional relationship between dependent and independent variables; which helps us to make estimates of one variable from the other. Correlation quantifies the association between the two variables; whereas linear regression finds the best line that predicts y from x and also x from y .The difference between correlation and regression is illustrated in the adjoining figure.



Lines of Regression: If we plot the observations of the linear regression between two variables, actually two straight lines can approximately be drawn through the scatter diagram. One line estimates values of y for specified values of x (known as line of regression of y on x); and other predicts values of x from given values of y (called line of regression of x on y).



Let line of regression of y on x be represented by $y = a + bx$ ...①

Normal equations as derived by the method of least Square are:

$$\sum y = an + b\sum x \qquad \text{...②}$$

and $\sum xy = a\sum x + b\sum x^2$ ...③

Dividing ② by n, we get

$$\frac{\sum y}{n} = a + b\frac{\sum x}{n} \Rightarrow \bar{y} = a + b\bar{x}$$

Where $\bar{x}$ and $\bar{y}$ are the means of x series and y series. This shows that $(\bar{x}, \bar{y})$ lies on the line of regression given by ①.

Again as $(\bar{x}, \bar{y})$ satisfies ① , shifting the origin to $(\bar{x}, \bar{y})$ in equation ③ , we get

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

$$\Rightarrow \sum(x-\bar{x})(y-\bar{y}) = b\sum(x-\bar{x})^2 \qquad \because \sum(x-\bar{x}) = 0$$

$$\Rightarrow b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum d_x d_y}{\sum d_x^2} \qquad \text{...④}$$

$$\text{Again } r = \frac{\sum d_x d_y}{\sqrt{(\sum d_x^2)(\sum d_y^2)}} = \frac{\sum d_x d_y}{n\sqrt{\frac{\sum d_x^2}{n}}\sqrt{\frac{\sum d_y^2}{n}}} = \frac{\sum d_x d_y}{n\sigma_x\sigma_y} \quad \because \sigma_x = \sqrt{\frac{\sum d_x^2}{n}},\ \sigma_y = \sqrt{\frac{\sum d_y^2}{n}}$$

Here σ_x , σ_y are standard deviations of x and y data points respectively

$$\Rightarrow \sum d_x d_y = nr\sigma_x\sigma_y \qquad \text{...⑤}$$

Using ⑤ in ④ , we get

$$b = \frac{n\,r\,\sigma_x\sigma_y}{\sum d_x^2} = \frac{r\,\sigma_x\sigma_y}{\sigma_x^2}$$

$$\Rightarrow b = \frac{r\,\sigma_y}{\sigma_x}$$ which is slope of line of regression line of y on x

$$\therefore\ b_{yx} = \frac{r\,\sigma_y}{\sigma_x}\ ,$$ b_{yx} denotes slope of line of regression line of y on x.

Thus line of regression of y on x given by ①, passes through $(\bar{x}, \bar{y})$ and is having slope $b_{yx} = \frac{r\,\sigma_y}{\sigma_x}$

$\therefore$ Equation of line of regression of y on x is given by $y - \bar{y} = b_{yx}(x - \bar{x})$

Similarly line of regression of x on y is given by: $x - \bar{x} = b_{xy}(y - \bar{y})$

where $b_{xy} = \frac{r\,\sigma_x}{\sigma_y}$ is slope of line of regression line of x on y

Here b_{xy} and b_{yx} are known coefficients of regression and are connected by the relation:

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

Example:

Compute the least squares regression equation of Y on X for the follo
What is the regression coefficient and what does it mean?

X	5	6	8	10	12	13	15	16	17
Y	16	19	23	28	36	41	44	45	50

The estimated regression line of Y on X is

$$\hat{Y} = a + bX,$$

and the two normal equations are

$$\sum Y = na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

	X	Y	XY	X^2
	5	16	80	25
	6	19	114	36
	8	23	184	64
	10	28	280	100
	12	36	432	144
	13	41	533	169
	15	44	660	225
	16	45	720	256
	17	50	850	289
Total	102	302	3853	1308

Now $\quad \overline{X} = \dfrac{\sum X}{n} = \dfrac{102}{9} = 11.33,\ \overline{Y} = \dfrac{\sum Y}{n} = \dfrac{302}{9} = 33.56,$

$$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{9(3853) - (102)(302)}{9(1308) - (102)^2}$$

$$= \frac{34677 - 30804}{11772 - 10404} = \frac{3873}{1368} = 2.831, \text{ and}$$

$$a = \overline{Y} - b\overline{X} = 33.56 - (2.831)(11.33) = 1.47.$$

…e the desired estimated regression line of Y on X is

$$\hat{Y} = 1.47 + 2.831X.$$

The estimated regression co-efficient, $b = 2.831$, which indicates that the values of Y increase by … units for a unit increase in X.

<u>Example:</u>

In an experiment to measure the stiffness of a spring, the length of the spring under …nt loads was measured as follows:

X=Loads (1b)	3	5	6	9	10	12	15	20	22	28
Y=length (in)	10	12	15	18	20	22	27	30	32	34

Find the regression equations appropriate for predicting

i) the length, given the weight on the spring;

ii) the weight, given the length of the spring.

The data come from a bivariate population, *i.e.* both X and Y are random, therefore there are two …sion lines. To find the regression equation for predicting length (Y), we take Y as dependent …e and treat X as independent variable (*i.e.* non-random). For the second regression, the choice of …ables is reversed.

The computations needed for the regression lines are given in the following table:

	X	Y	X^2	Y^2	XY
	3	10	9	100	30
	5	12	25	144	60
	6	15	36	225	90
	9	18	81	324	162
	10	20	100	400	200
	12	22	144	484	264
	15	27	225	729	405
	20	30	400	900	600
	22	32	484	1024	704
	28	34	784	1156	932
Total	130	220	2288	5486	3467

The estimated regression equation appropriate for predicting the length, Y, given the weight X, is

$$\hat{Y} = a_0 + b_{yx} X,$$

where $b_{YX} = \dfrac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \dfrac{(10)(3467) - (130)(220)}{(10)(2288) - (130)^2}$

$$= \frac{6070}{5980} = 1.02, \text{ and}$$

$$a_0 = \overline{Y} - b_{yx}\overline{X} = 22 - (1.02)(13) = 8.74$$

Hence the desired estimated regression equation is

$$\hat{Y} = 8.74 + 1.02\,X$$

ii) The estimated regression equation appropriate for predicting the weight, X, given the length

$$\hat{X} = a_1 + b_{xy} Y,$$

where $b_{XY} = \dfrac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2} = \dfrac{(10)(3467) - (130)(220)}{(10)(5486) - (220)^2}$

$$= \frac{6070}{6460} = 0.94, \text{ and}$$

$$a_1 = \overline{X} - b_{xy}\overline{Y} = 13 - (0.94)(22) = -7.68$$

Hence $\hat{X} = 0.94Y - 7.68$ is the estimated regression equation appropriate for predicting the weight given the length (Y).

Standard Error of Estimate

Definition: The **Standard Error of Estimate** is the measure of variation of an observation made around the computed regression line. Simply, it is used to check the accuracy of predictions made with the regression line.

The observed values of (X,Y) do not fall on the regression line but scatter away from it. The degree of dispersion of the observed values about the regression line is measured by the deviation of regression or the standard error of estimate of Y on X. For the population data, the standard deviation that measures the observations about the regression line is denoted by $\sigma_{Y.X}$ and is defined by

$$\sigma_{Y.X} = \sqrt{\frac{\sum[Y-(\alpha+\beta X)]^2}{N}}$$

where N is the population size.

For sample data, we estimate $\sigma_{Y.X}$ by $s_{y.x}$ which is defined as

$$s_{y.x} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-2}},$$

To find $\sum(Y-\hat{Y})^2$, we have to calculate $\hat{Y}$ from the estimated regression line for the observed of X, which is not an easy task. We therefore use an alternative form obtained as below:

$$s_{y.x} = \sqrt{\frac{\sum Y_i^2 - a\sum Y_i - b\sum X_i Y_i}{n-2}},$$

Example: Using the data given below

X	5	6	8	10	12	13	15	16	17
Y	16	19	23	28	36	41	44	45	50

a) Find the values of $\hat{Y}$ and show that $\sum(Y - \hat{Y}) = 0$

b) Compute the standard error of estimate $S_{y.x}$

X	Y	$\hat{Y}$ (=1.47+2.831X)	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	Y^2
5	16	15.625	0.375	0.140625	256
6	19	18.456	0.544	0.295936	361
8	23	24.118	–1.118	1.249924	529
10	28	29.780	–1.780	3.168400	784
12	36	35.442	0.558	0.311364	1296
13	41	38.273	2.727	7.436529	1681
15	44	43.935	0.065	0.004225	1936
16	45	46.766	–1.766	3.118756	2025
17	50	49.597	0.403	0.162409	2500
102	302	301.992	0.008	15.888168	110368

i) The estimated values $\hat{Y}$ appear in the third column of the table and $\Sigma(Y-\hat{Y})$ turns out to be 0.008. This small difference is due to rounding off.

ii) The standard error of estimate of Y on X is

$$s_{y.x} = \sqrt{\frac{\Sigma(Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{15.888168}{7}} = \sqrt{2.269738} = 1.51$$

Using the alternative form for the calculation of $s_{y.x}$, we get

$$s_{y.x} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n-2}}$$

$$= \sqrt{\frac{11368 - (1.47)(302) - (2.831)(3853)}{9-2}}$$

$$= \sqrt{\frac{16.217}{7}} = \sqrt{2.316714} = 1.52.$$

10.4.5 Co-efficient of Determination. The variability among the values of the dependen[t] Y, called the *total variation*, is given by $\Sigma(Y-\bar{Y})^2$. This is composed of two parts (i) that [which is] explained by (associated with) the regression line, *i.e.* $\Sigma(\hat{Y}-\bar{Y})^2$, (ii) that which the regression li[ne fails] to explain, *i.e.* $\Sigma(Y-\hat{Y})^2$ (see figure). In symbols

$$\Sigma(Y-\bar{Y})^2 = \Sigma(Y-\hat{Y})^2 + \Sigma(\hat{Y}-\bar{Y})^2.$$

Co-efficient of Determination:

The coefficient of determination is a statistical measurement that examines how differences in one variable can be explained by the difference in a second variable, when predicting the outcome of a given event.

The standard error of estimate gives some indication of how certain we can be about a particular prediction of Y using the regression equation; it still does not tell us how well the independent variable explains variation in the dependent variable. The coefficient of determination does exactly this: it measures the fraction of the total variation in the dependent variable that is explained by the independent variable.

Thus the sample co-efficient of determination is

$$r^2 = \frac{\text{Explained variation}}{\text{Total variatioin}} = \frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(Y - \bar{Y})^2}$$

$$= 1 - \frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \bar{Y})^2}.$$

alternative form for calculating the coefficient of determination is

$$r^2 = \frac{a\Sigma Y + b\Sigma XY - (\Sigma Y)^2 / n}{\Sigma Y^2 - (\Sigma Y)^2 / n}.$$

Properties of Regression Coefficients

- As $\sqrt{b_{xy}b_{yx}} = r$, the coefficient of correlation is the geometric mean between the two regression coefficients.
- Since $\frac{b_{xy}+b_{yx}}{2} \geq \sqrt{b_{xy}b_{yx}} = r$, $\therefore$ arithmetic mean of the two regression coefficients is greater than or equal to the correlation coefficient (r).

- If there is a perfect correlation between the two variables under consideration, then $b_{xy} = b_{yx} = r$; and the two lines of regression coincide. Converse is also true, i.e. if two lines of regression coincide, then there is a perfect correlation; $r = \pm 1$.
- Since $b_{xy}b_{yx} = r^2 > 0$, the signs of both regression coefficients b_{xy} and b_{yx} and coefficient of correlation (r) must be same; either all three negative or all positive.
- $\because b_{xy}b_{yx} = r^2 \leq 1$, if one of the regression coefficients is greater than unity, other must be less than unity.
- Point of intersection of two lines of regression is $(\bar{x}, \bar{y})$, Where $\bar{x}$ and $\bar{y}$ are the means of x series and y series.
- If both lines of regression cut each other at right angle, there is no correlation between the two variables; i.e. $r = 0$.

Example Prove that arithmetic mean of coefficients of regression is greater than the coefficient of correlation.

Solution: We know that $b_{xy} = \frac{r\,\sigma_x}{\sigma_y}$ and $b_{yx} = \frac{r\,\sigma_y}{\sigma_x}$

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

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

or $\frac{1}{2}\left[\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x\sigma_y}\right] > 1$

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

or $\sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y > 0$

or $[\sigma_x - \sigma_y]^2 > 0$

which is true

Note : A.M. $= r$ if $b_{xy} = b_{yx} = r = \pm 1$

Angle between the Lines of Regression

If θ be the acute angle between the two regression lines for two variables x and y, then $\tan\theta = \frac{1-r^2}{r}\,\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$

Proof: The two lines of regression are given by:

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

$$\text{and } x-\bar{x} = \frac{r\,\sigma_x}{\sigma_y}(y-\bar{y}) \qquad \ldots ②$$

If m_1 and m_2 are slopes of lines ① and ②, then

$\tan\theta = \frac{m_2-m_1}{1+m_1m_2}$, where $m_1 = \frac{r\,\sigma_y}{\sigma_x}$, $m_2 = \frac{\sigma_y}{r\,\sigma_x}$

$$\Rightarrow \tan\theta = \frac{\frac{\sigma_y}{r\,\sigma_x}-\frac{r\,\sigma_y}{\sigma_x}}{1+\frac{r\sigma_y}{\sigma_x}\,\frac{\sigma_y}{r\sigma_x}} = \frac{\left(\frac{1}{r}-r\right)\frac{\sigma_y}{\sigma_x}}{1+\frac{\sigma_y^2}{\sigma_x^2}} = \frac{1-r^2}{r}\,\frac{\sigma_x\,\sigma_y}{\sigma_x^2+\sigma_y^2} \qquad \ldots ③$$

- When $r=0$, $\tan\theta = \infty \;\Rightarrow \theta = \frac{\pi}{2}$ from ③

  $\therefore$ when $r=0$, the two lines of regression are perpendicular to each other.

- When $r=\pm 1$, $\tan\theta = 0 \;\Rightarrow \theta = 0$ from ③

  $\therefore$ when $r=\pm 1$, the two lines of regression are coincident

Properties of regression equation

1. If $r = 0$, the variables are uncorrelated, the lines of regression become perpendicular to each other.

2. If $r = 1$, the two lines of regression either coincide or parallel to each other.

3. Angle between the two regression lines is $\theta = \tan^{-1} (m_1 - m_2 \ / \ 1 + m_1 m_2)$ where m1 and m2 are the slopes of regression lines X on Y and Y on X respectively.

4. The angle between the regression lines indicates the degree of dependence between the variable.

5. Regression equations intersect at $(\overline{X} , \overline{Y})$

Example Find the correlation coefficient between x and y, when the two lines of regression are given by: $2x - 9y + 6 = 0$ and $x - 2y + 1 = 0$

Solution: Let the line of regression of x on y be $2x - 9y + 6 = 0$ ...①

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

Now ① $\Rightarrow x = \frac{9}{2}y - 3 \quad \therefore b_{xy} = \frac{9}{2}$

Also ② $\Rightarrow y = \frac{1}{2}x + \frac{1}{2} \quad \therefore b_{yx} = \frac{1}{2}$

$\therefore r = \sqrt{b_{xy}b_{yx}} = \sqrt{\frac{9}{2} \times \frac{1}{2}} = \frac{3}{2}$, which is not possible as $-1 \leq r \leq 1$

So our choice of regression lines is incorrect.

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

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

Also line of regression of y on x is $2x - 9y + 6 = 0$

$\Rightarrow y = \frac{2}{9}x + \frac{2}{3} \quad \therefore b_{yx} = \frac{2}{9}$

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

Hence coefficient of correlation between x and y is $\frac{2}{3}$

Example The regression equations calculated from a given set of observations for two random variables are: $x = -0.4y + 6.4$ and $y = -0.6x + 4.6$

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

Solution: The two equations of regression are:

$$x = -0.4y + 6.4 \quad \text{...①}$$

$$y = -0.6x + 4.6 \quad \text{...②}$$

$\Rightarrow b_{xy} = -0.4$ and $b_{yx} = -0.6$

$\therefore r^2 = b_{xy}b_{yx} = 0.24$

$\Rightarrow \quad r = \pm 0.49$

We know that the signs of b_{xy}, b_{yx} and r must be same

$\therefore r = -0.49$

Again we know that the point of intersection of two regression lines is $(\overline{x}, \overline{y})$

Therefore solving ① and ②, we get $\overline{x} = 6$, $\overline{y} = 1$

Example From a partially destroyed lab data, following results were retrieved:

Lines of regression are:

$x = 0.45y + 5.35$ and $y = 0.8x + 6.6$, $\sigma_x^2 = 9$

Find $\overline{x}, \overline{y}, \sigma_y$ and r for the existing data.

Solution: The two equations of regression are:

$$x = 0.45y + 5.35 \qquad \ldots ①$$

$$y = 0.8x + 6.6 \qquad \ldots ②$$

We know that the point of intersection of two regression lines is $(\bar{x}, \bar{y})$

Therefore solving ① and ②, we get $\bar{x} = 13$, $\bar{y} = 17$

Again ① ⇒ $b_{xy} = 0.45$ and $b_{yx} = 0.8$

$\therefore r^2 = b_{xy} b_{yx} = 0.36$

$\Rightarrow \quad r = \pm 0.6$

We know that the signs of b_{xy} , b_{yx} and r must be same

$\therefore \; r = 0.6$

Also $b_{yx} = \frac{r\,\sigma_y}{\sigma_x} \Rightarrow 0.8 = \frac{(0.6)\sigma_y}{3} \Rightarrow \sigma_y = \frac{0.8 \times 3}{0.6} = 4$

Example

Is there any mistake in the data provided about the two regression lines $Y = -1.5\ X + 7$, and $X = 0.6\ Y + 9$? Give reasons.

Solution:

The regression coefficient of Y on X is $b_{YX} = -1.5$

The regression coefficient of X on Y is $b_{XY} = 0.6$

Both the regression coefficients are of different sign, which is a contrary. So the given equations cannot be regression lines.

Example

If two regression coefficients are $b_{YX} = 5/6$ and $b_{XY} = 9/20$, what would be the value of r_{XY}?

Solution:

The correlation coefficient r_{XY} =

$$r_{XY} = \pm\sqrt{(b_{YX})(b_{XY})}$$

$$= \pm\sqrt{\frac{5}{6}\times\frac{9}{20}} = 0.375$$

Since both the signs in b_{YX} and b_{XY} are positive, correlation coefficient between X and Y is positive.

Example

Given that $b_{YX} = 18/7$ and $b_{XY} = -5/6$. Find r ?

Solution:

$$r_{XY} = \pm\sqrt{(b_{YX})(b_{XY})}$$

$$= \sqrt{-\frac{18}{7} \times -\frac{5}{6}} = \sqrt{\frac{15}{7}} = -0.553.$$

Since both the signs in b_{YX} and b_{XY} are negative, correlation coefficient between X and Y is negative.

Example:

	mean	S.D
Yield of wheat (kg. unit area)	10	8
Annual Rainfall (inches)	8	2

Correlation coefficient: 0.5

Estimate the yield when rainfall is 9 inches

Solution:

Let us denote the dependent variable yield by Y and the independent variable rainfall by X.

Regression equation of Y on X is given by

$$Y - \bar{y} = r_{XY} \frac{SD(Y)}{SD(X)} (x - \bar{x})$$

$\bar{x} = 8,\ SD(X) = 2, \bar{y} = 10,\ SD(Y) = 8, \quad r_{XY} = 0.5$

$$Y - 10 = 0.5 \times \frac{8}{2}(x - 8)$$
$$= 2\,(x - 8)$$

When $x = 9$,

$Y - 10 = 2\,(9 - 8)$

$Y = 2 + 10$

= 12 kg (per unit area)

Corresponding to the annual rain fall 9 inches the expected yield is 12 kg (per unit area).

Example Find the regression line of y on x from the following data:

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

Also estimate the value of y, when $x = 10$

Solution: Let line of regression of y on x be:

$$y = a + bx \qquad \text{...①}$$

Then normal equations are given by:

$$\sum y = an + b\sum x \qquad \text{...②}$$

and $\sum xy = a\sum x + b\sum x^2$ ...③

Calculating $\sum x$, $\sum y$, $\sum xy$ and $\sum x^2$

x	y	x^2	xy
1	1	1	1
3	2	9	6
4	4	16	16
6	4	36	24
8	5	64	40
9	7	81	63
11	8	121	88
14	9	196	126
$\sum x = 56$	$\sum y = 40$	$\sum x^2 = 524$	$\sum xy = 364$

Substituting values of $\sum x, \sum y, \sum xy$ and $\sum x^2$ in ② and ③

$$\Rightarrow 40 = 8a + 56b \qquad \text{...④}$$

$$\text{and} \quad 364 = 56a + 524b \qquad \text{...⑤}$$

Solving ④and ⑤, we get $a = \frac{6}{11}$ and $b = \frac{7}{11}$

Substituting in ①, line of regression of y on x is $y = \frac{6}{11} + \frac{7}{11}x$

$\Rightarrow 7x - 11y + 6 = 0$

Also at $x = 10$, $y = \frac{76}{11}$

Example Following data depicts the statistical values of rainfall and production of wheat in a region for a specified time period.

	Mean	Standard Deviation
Production of Wheat (kg. per unit area)	10	8
Rainfall (cm)	8	2

Estimate the production of wheat when rainfall is 9cm if correlation coefficient between production and rainfall is given to be 0.5.

Solution: Let the variables x and y denote production and rainfall respectively.

Given that $\overline{x} = 10$, $\overline{y} = 8$ also $\sigma_x = 8$, $\sigma_y = 2$

Now equation of regression of x on y is given by:

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

$$\Rightarrow x - 10 = \frac{(0.5)8}{2}(y - 8)$$

$$\Rightarrow x = 2y - 6$$

$\therefore$ When rainfall is 9cm, production of wheat is estimated to be $2(9) - 6 = 12$ kg. per unit area

Example Find the coefficient of correlation and the lines of regression for the data given below:

$$n = 18, \sum x = 12, \sum y = 18, \sum x^2 = 60, \sum y^2 = 96 \text{ and} \sum xy = 48$$

Solution: $\overline{x} = \frac{\sum x}{n} = \frac{12}{18} = 0.67$, $\overline{y} = \frac{\sum y}{n} = \frac{18}{18} = 1$

$$\sigma_x^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{60}{18} - \left(\frac{12}{18}\right)^2 = 2.89 \therefore \sigma_x = 1.7$$

$$\sigma_y^2 = \frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2 = \frac{96}{18} - \left(\frac{18}{18}\right)^2 = 4.33 \therefore \sigma_y = 2.08$$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{(\sum x^2) - \frac{1}{n}(\sum x)^2}\sqrt{(\sum y^2) - \frac{1}{n}(\sum y)^2}}$$

$$= \frac{48 - \frac{(12)(18)}{18}}{\sqrt{(60) - \frac{1}{18}(12)^2}\sqrt{(96) - \frac{1}{18}(18)^2}} = \frac{36}{(7.2)(8.83)} = 0.57$$

$$b_{xy} = \frac{r\,\sigma_x}{\sigma_y} = \frac{(0.57)(1.7)}{2.08} = 0.47 \quad , b_{yx} = \frac{r\,\sigma_y}{\sigma_x} = \frac{(0.57)(2.08)}{1.7} = 0.7$$

Equations of lines of regression are:

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

$$\Rightarrow y - 1 = 0.7(x - 0.67) \text{ and } x - 0.67 = 0.47(y - 1)$$

$$\Rightarrow y = 0.7x + 0.53 \qquad \text{and } x = 0.47y + 0.2$$

Example Marks obtained by 11 students in statistics papers are given below:

Paper I	60	65	68	70	75	85	80	45	55	56	58
Paper II	62	64	65	70	74	90	82	56	50	48	60

Calculate the coefficient of correlation for the above data. Also find the equations of lines of regression.

Solution: Let marks obtained in paper I be denoted by x and marks obtained in paper II be denoted by y.

$$\text{Let } A_x = 65,\ A_y = 70 \ \therefore\ d_x = x - 65\ ,\ d_y = y - 70$$

Calculating $\sum d_x, \sum d_y, \sum d_x^2, \sum d_y^2$ and $\sum d_x d_y$

x	d_x $(x-65)$	d_x^2	y	d_y $(y-70)$	d_y^2	$d_x d_y$
60	-5	25	62	-8	64	40
65	0	0	64	-6	36	0
68	3	9	65	-5	25	-15
70	5	25	70	0	0	0
75	10	100	74	4	16	40
85	20	400	90	20	400	400
80	15	225	82	12	144	180
45	-20	400	56	-14	196	280
55	-10	100	50	-20	400	200
56	-9	81	48	-22	484	198
58	-7	49	60	-10	100	70
	$\sum d_x = 2$	$\sum d_x^2 = 1414$		$\sum d_y = -49$	$\sum d_y^2 = 1865$	$\sum d_x d_y = 1393$

Karl Pearson coefficient of correlation (r) is given by:

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\left(\sum d_x^2\right) - \frac{1}{n}\left(\sum d_x\right)^2}\sqrt{\left(\sum d_y^2\right) - \frac{1}{n}\left(\sum d_y\right)^2}}$$

$$\therefore r = \frac{1393 - \frac{(2)(-49)}{11}}{\sqrt{(1414) - \frac{1}{11}(2)^2}\sqrt{(1865) - \frac{1}{11}(-49)^2}} = \frac{1401.9091}{(37.5984)(40.5799)} = 0.9188$$

Now $\bar{x} = A_x + \frac{\sum d_x}{n} = 65 + \frac{2}{11} = 65.1818$

$$\bar{y} = A_y + \frac{\sum d_y}{n} = 70 + \frac{-49}{11} = 65.5455$$

Also $\sigma_x = \sqrt{\frac{\sum d_x^2}{n} - \left(\frac{\sum d_x}{n}\right)^2} = \sqrt{\frac{1414}{11} - \left(\frac{2}{11}\right)^2} = 11.3363$

and $\sigma_y = \sqrt{\frac{\sum d_y^2}{n} - \left(\frac{\sum d_y}{n}\right)^2} = \sqrt{\frac{1865}{11} - \left(\frac{-49}{11}\right)^2} = 12.2353$

$$\therefore b_{xy} = \frac{r\,\sigma_x}{\sigma_y} = \frac{(0.9188)(11.3363)}{12.2353} = 0.8513$$

$$b_{yx} = \frac{r\,\sigma_y}{\sigma_x} = \frac{(0.9188)(12.2353)}{11.3363} = 0.9917$$

Equations of lines of regression are:

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

$$\Rightarrow y - 65.55 = 0.99(x - 65.18) \text{ and } x - 65.18 = 0.85(y - 65.55)$$

$$\Rightarrow y = 0.99x + 1.02 \qquad\qquad \text{and } x = 0.85y + 9.46$$

Example24 The regression equations calculated from a given set of observations for two variables x and y are: $x = 9y + 5$ and $y = kx + 9$

Show that $0 < k < \frac{1}{9}$. Also if $k = \frac{1}{10}$, find $\bar{x}, \bar{y}$ and r

Solution: The two equations of regression are:

$$x = 9y + 5 \qquad \ldots ①$$

$$y = kx + 9 \qquad \ldots ②$$

$\Rightarrow b_{xy} = 9$ and $b_{yx} = k$

$\therefore r^2 = b_{xy} b_{yx} = 9k$

$\Rightarrow r = 3\sqrt{k} \quad \because b_{xy} = 9$ is positive, therefore k and r are also positive

Now $0 < r < 1$ or $0 < 3\sqrt{k} < 1$

$\Rightarrow 0 < 9k < 1$ or $0 < k < \frac{1}{9}$

Now if $k = \frac{1}{10}$, equation ② becomes $10y = x + 90 \quad \ldots ③$

Solving ① and ③, the point of intersection of two regression lines is

$$\bar{x} = 860, \ \bar{y} = 95, \text{ also } r = 3\sqrt{k} = 3\sqrt{\frac{1}{10}} = 0.949$$

Exercise 4

1. Find the coefficient of correlation between x and y from the given data. Also find the two lines of regression.

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

2. Find the rank correlation for the following data:

x	56	42	72	36	63	47	55	49	38	42	68	60
y	147	125	160	118	149	128	150	145	115	140	152	155

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

4. If the coefficient of correlation between two variables x and y is 0.5 and the acute angle between their lines of regression is $\tan^{-1}\frac{3}{8}$, show that $\sigma_x = \frac{\sigma_y}{2+\sqrt{3}}$

5. From a partially destroyed lab data, following results were retrieved: Lines of regression are:

$$8x = 10y - 66 \text{ and } 18y = 40x - 214, \ \sigma_x = 3$$

Find $\overline{x}, \overline{y}, \sigma_y$ and r for the existing data.

6.

Calculate the co-efficient of correlation and obtain the lines of regression of the following data

Price (X)	3	4	5	6	7	8	9	10	11	12
Demand (Y)	25	24	20	20	19	17	16	13	10	6

7.

Find the correlation co-efficient between X and Y, given

X	5	12	4	16	18	21	22	23	25
Y	11	16	15	20	17	19	25	24	21

8.

Find the co-efficient of correlation between persons employed and cloth manufactured in a textile mill. Interpret the result

Persons employed	137	209	113	189	176	200	219
Cloth manufactured ('000 yds)	23	47	22	40	39	51	49

9.

Compute the co-efficient of rank correlation for the following ranks;

X	8	3	6.5	3	6.5	9	3	1	5
Y	8	9	6.5	2.5	4	5	6.5	1	2.5

Answers

1. $r = 0.96, x = 0.2y - 0.64, y = 4.69x + 4.9$
2. 0.932
3. $\bar{x} = 9.06, \bar{y} = 5.52, r = 0.46$
5. $\bar{x} = 13, \bar{y} = 17, \sigma_y = 4, r = 0.6$