UNIT I

CORRELATION

Definition:

The term correlation refers to the relationship between two variables.

Types of Correlation

Various types of correlation are considered under the following three heads. They are

(i) Positive or negative correlation

(ii) Simple or Partial or Multiple correlation

(iii) Linear or Non-linear or No correlation

Methods of studying Correlation

(i) Scatter Diagram

(ii) Karl Pearson's correlation coefficient(r)

(iii)Spearman's rank correlation coefficient(ρ)

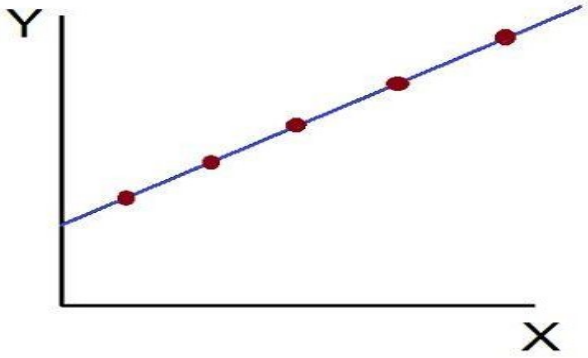
Scatter Diagram Method

Definition: The **Scatter Diagram Method** is the simplest method to study the correlation between two variables wherein the values for each pair of a variable is plotted on a graph in the form of dots thereby obtaining as many points as the number of observations. Then by looking at the scatter of several points, the degree of correlation is ascertained.

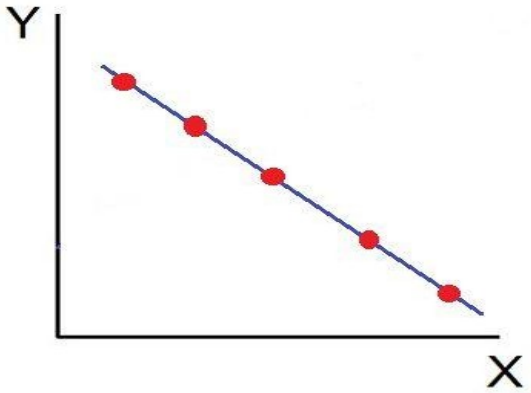
The degree to which the variables are related to each other depends on the manner in which the points are scattered over the chart. The more the points plotted are scattered over the chart, the lesser is the degree of correlation between the variables. The more the points plotted are closer to the line, the higher is the degree of correlation. The degree of correlation is denoted by **"r".**

The following types of scatter diagrams tell about the degree of correlation between variable X and variable Y.

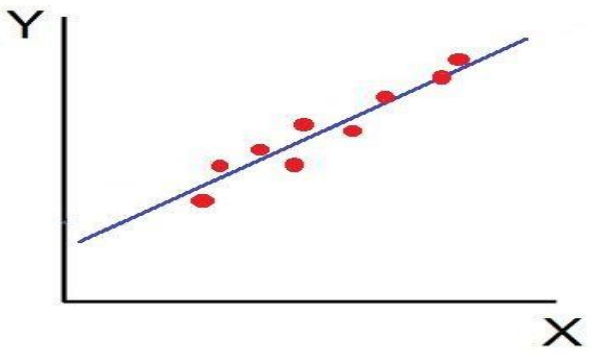
1. **Perfect Positive Correlation (r=+1):** The correlation is said to be perfectly positive when all the points lie on the straight line rising from the lower left-hand corner to the upper right-hand corner.



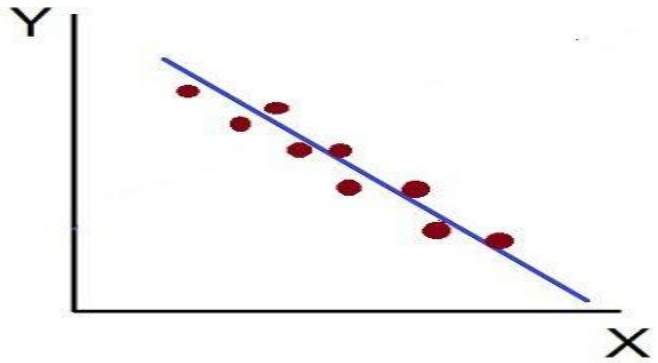
2. **Perfect Negative Correlation (r=-1):** When all the points lie on a straight line falling from the upper left-hand corner to the lower right-hand corner, the variables are said to be negatively correlated.



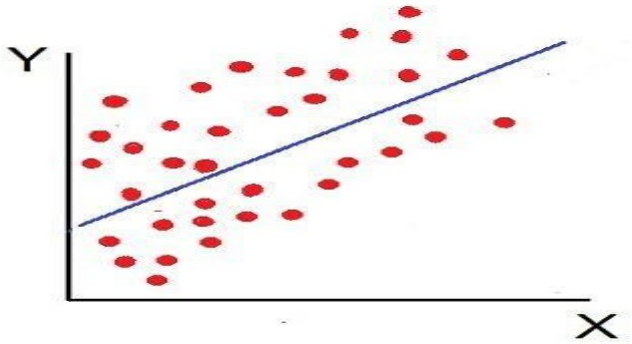
3. **High Degree of +Ve Correlation (r= + High):** The degree of correlation is high when the points plotted fall under the narrow band and is said to be positive when these show the rising tendency from the lower left-hand corner to the upper right-hand corner.


Y
X

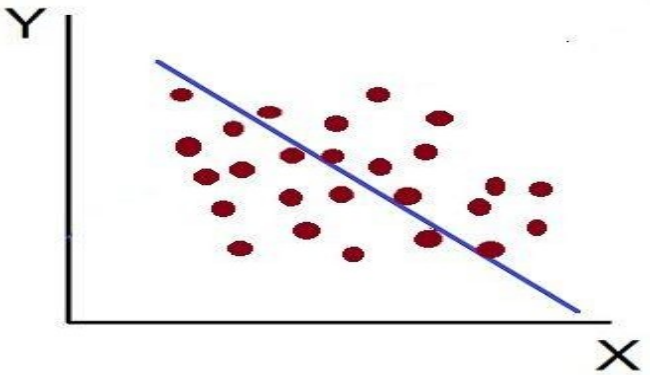
4. **High Degree of –Ve Correlation (r= – High):** The degree of negative correlation is high when the point plotted fall in the narrow band and show the declining tendency from the upper left-hand corner to the lower right-hand corner.



5. **Low degree of +Ve Correlation (r= + Low):** The correlation between the variables is said to be low but positive when the points are highly scattered over the graph and show a rising tendency left-hand corner to the upper right-hand corner.

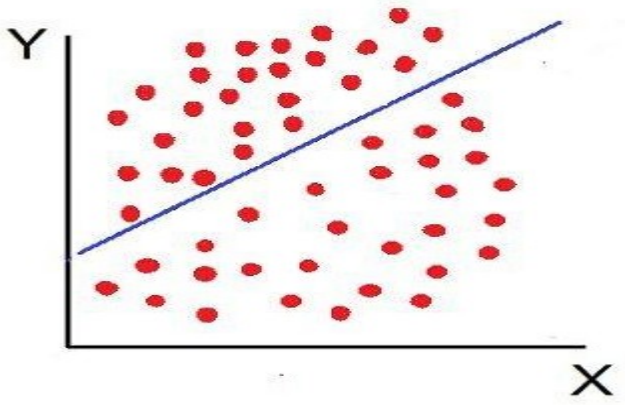


6. **Low Degree of –Ve Correlation (r= + Low):** The degree of correlation is low and negative when the points are scattered over the graph and the show the falling tendency from the upper



left-hand corner to the lower right-hand corner.

7. **No Correlation (r= 0):** The variable is said to be unrelated when the points are haphazardly scattered over the graph and do not show any specific pattern. Here the correlation is absent and



hence **r = 0**.

Thus, the scatter diagram method is the simplest device to study the degree of relationship between the variables by plotting the dots for each pair of variable values given. The chart on which the dots are plotted is also called as a **Dotogram**.

Karl Pearson's coefficient of correlation (r)

This is also called product moment correlation coefficient. This is denoted by r. This is covariance between the two variables divided by the product of their standard deviations.

Formula

$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}, \text{ Where } \sum x = 0, \sum y = 0$$

1. Compute the coefficient of correlation between X – Advertisement Expenditure and Y – Sales.

X	Y	X^2	Y^2	XY
10	88	100	7744	880
12	90	144	8100	1080

18	94	324	8836	1692
8	86	64	7396	688
13	87	169	7569	1131
20	92	400	8464	1840
22	96	484	9216	2112
15	94	225	8836	1410
5	88	25	7744	440
17	85	289	7225	1445
$\sum X = 140$	$\sum Y = 900$	$\sum X^2 = 2224$	$\sum Y^2 = 81130$	$\sum XY = 12718$

$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

$$r = \frac{10(12718) - (140)(900)}{\sqrt{10(2224) - (140)^2}\sqrt{10(81130) - (900)^2}}$$

$$r = \frac{127180 - 126000}{\sqrt{22240 - 19600}\sqrt{811300 - 810000}}$$

$$r = \frac{1180}{\sqrt{2640}\sqrt{1300}}$$

$$r = \frac{1180}{51.3809X36.0555}$$

$$r = \frac{1180}{1852.56404}$$

r = 0.6370

2.The following table gives aptitude test scores and productivity indices of 8 randomly selected workers. Calculate the correlation coefficient between the aptitude score and productivity index.

Let X = Aptitude Score, Y = Productivity Index

Aptitude Score X	Productivity Index Y	X^2	Y^2	XY
57	67	3249	4489	3819
58	68	3364	4624	3944
59	65	3481	4225	3835
59	68	3481	4624	4012

60	72	3600	5184	4320
61	72	3721	5184	4392
62	69	3844	4761	4278
64	71	4096	5041	4544
$\sum X = 480$	$\sum Y = 552$	$\sum X^2 = 28836$	$\sum Y^2 = 38132$	$\sum XY = 33144$

$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

$$r = \frac{8(33144) - (480)(552)}{\sqrt{8(28836) - (480)^2}\sqrt{8(38132) - (552)^2}}$$

$$r = \frac{265152 - 264960}{\sqrt{230688 - 230400}\sqrt{305056 - 304704}}$$

$$r = \frac{192}{\sqrt{288}\sqrt{352}}$$

$$r = \frac{192}{16.9706X18.7617}$$

$$r = \frac{192}{318.3974}$$

r = 0.6030

3. Calculate the coefficient of correlation between Expenditure on Advertisement in Rs.'000 (X) and Sales in Rs. Lakhs (Y) after allowing the time lag of two months.

Months	X	Y	X	Y	X^2	Y^2	XY
Jan	40	75	40	65	1600	4225	2600
Feb	45	69	45	64	2025	4096	2880
Mar	47	65	47	70	2209	4900	3290
Apr	50	64	50	71	2500	5041	3550
May	53	70	53	75	2809	5625	3975
June	60	71	60	83	3600	6889	4980
July	57	75	57	90	3249	8100	5130
Aug	51	83	51	92	2601	8464	4692
Sep	48	90					
Oct	45	92					

			$\sum X =$ 403	$\sum Y =$ 610	$\sum X^2 =$ 20593	$\sum Y^2 =$ 47340	$\sum XY =$ 31097

$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

$$r = \frac{8(31097) - (403)(610)}{\sqrt{8(20593) - (403)^2}\sqrt{8(47340) - (610)^2}}$$

$$r = \frac{248776 - 245830}{\sqrt{164744 - 162409}\sqrt{378720 - 372100}}$$

$$r = \frac{2946}{\sqrt{2335}\sqrt{6620}}$$

$$r = \frac{2946}{48.3218X81.3634}$$

$$r = \frac{2946}{3931.6259}$$

r = 0.7493

4. From the following data, compute the coefficient of correlation between X and Y

	X	Y
Sum of squares of deviations from the arithmetic mean	8250	724
Sum of products of deviations of X and Y from respective means	2350	
No of pairs of observations	10	

Solution:

Given

$\sum x^2 = \sum\left(X - \overline{X}\right)^2 = 8250$

$\sum y^2 = \sum\left(Y - \overline{Y}\right)^2 = 724$

$\sum xy = \sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right) = 2350$

N = 10

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = r = \frac{2350}{\sqrt{8250}\sqrt{724}} = r = \frac{2350}{90.8295X26.9072}$$

$$r = \frac{2350}{2443.9675} = 0.9615$$

Note:

1. Correlation coefficient (r) lies between $-1 \leq r \leq +1$
2. If r = 0, absence of linear correlation
3. If r = +1, perfect positive correlation
4. If r = -1, perfect negative correlation
5. r = Coefficient of correlation
6. r^2 = Coefficient of Determination
7. $K^2 = 1-r^2$ = Coefficient of Non – Determination
8. $K = \pm\sqrt{1-r^2}$ = Coefficient of Alienation
9. $S.E(r) = \frac{1-r^2}{\sqrt{N}}$

SPEARMAN'S RANK CORRELATION (ρ)

Formula

(i) When there is no tie

$$\rho = 1 - \left[\frac{6\sum d^2}{N(N^2-1)}\right]$$, Where d = difference between X and Y ranks

(ii) When one value occurs m times

$$\rho = 1 - \left[\frac{6\{\sum d^2 + \frac{m(m^2-1)}{12}\}}{N(N^2-1)}\right]$$

(iii) When more than one value is repeated

$$\rho = 1 - \left[\frac{6\{\sum d^2 + \frac{m(m^2-1)}{12}\} + \frac{m(m^2-1)}{12} + ...\}}{N(N^2-1)}\right]$$

5. Rankings of 10 trainees at the beginning (X) and at the end (Y) of a certain course are given below:

Trainees	X	Y	d X-Y	d^2
A	1	6	-5	25
B	6	8	-2	4
C	3	3	0	0
D	9	7	2	4
E	5	2	3	9
F	2	1	1	1
G	7	5	2	4
H	10	9	1	1
I	8	4	4	16
J	4	10	-6	36

			$\sum d = 0$	$\sum d^2 = 100$

$$\rho = 1 - \left[\frac{6\sum d^2}{N(N^2-1)}\right]$$

$$\rho = 1 - \left[\frac{6X100}{10(10^2-1)}\right] = 1 - \left[\frac{6X100}{10(100-1)}\right] = 1 - \left[\frac{6X100}{10X99}\right] = 1 - \left[\frac{600}{990}\right] = 1\text{-}0.6061$$

$$\rho = 0.3939$$

6. For the data given below, calculate the rank correlation coefficient.

		Ranks			
X	Y	X	Y	d X-Y	d^2
21	47	5	1	4	16
36	40	3	4	-1	1
42	37	1	5	-4	16
37	42	2	3	-1	1
25	43	4	2	2	4
				$\sum d = 0$	$\sum d^2 = 38$

$$\rho = 1 - \left[\frac{6\sum d^2}{N(N^2-1)}\right]$$

$$\rho = 1 - \left[\frac{6X38}{5(5^2-1)}\right] = 1 - \left[\frac{6X38}{5(25-1)}\right] = 1 - \left[\frac{6X38}{5x24}\right] = 1 - \left[\frac{228}{120}\right] = 1\text{-}1.9$$

$$\rho = -0.9$$

7. Find the rank correlation coefficient for the percentage of marks secured by a group of 8 students in Economics and Statistics.

		Ranks			
X	Y	X	Y	d X-Y	d^2
50	80	7	3	4	16
60	71	6	5	1	1
65	60	5	7	-2	4
70	75	3.5	4	-0.5	0.25
75	90	2	1	1	1
40	82	8	2	6	36
70	70	3.5	6	-2.5	6.25
80	50	1	8	-7	49
				$\sum d = 0$	$\sum d^2 = 113.5$

$$\rho = 1 - \left[\frac{6\{\sum d^2 + \frac{m(m^2-1)}{12}\}}{N(N^2-1)} \right];$$

when m = 2, $\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{2(4-1)}{12} = \frac{2X3}{12} = \frac{6}{12} = \frac{1}{2} = 0.5$

$$1 - \left[\frac{6\{113.5 + 0.5\}}{8(8^2-1)} \right] = 1 - \left[\frac{6X114}{8X63} \right] = 1 - \left[\frac{684}{504} \right] = 1\text{-}1.3571 = -0.3571$$

8. Marks obtained by 8 students in Accountancy (X) and Statistics (Y) are given below. Compute rank correlation coefficient.

		Ranks			
X	Y	X	Y	d	d^2
15	40	7	3	4	16
20	30	5.5	5	0.5	0.25
28	50	4	2	2	4
12	30	8	5	3	9
40	20	3	7	-4	16
60	10	2	8	-6	36
20	30	5.5	5	0.5	0.25
80	60	1	1	0	0
				$\sum d = 0$	$\sum d^2 = 81.5$

$$\rho = 1 - \left[\frac{6\{\sum d^2 + \frac{m(m^2-1)}{12}\} + \frac{m(m^2-1)}{12}\}}{N(N^2-1)} \right]$$

when m = 2, $\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{2(4-1)}{12} = \frac{2X3}{12} = \frac{6}{12} = \frac{1}{2} = 0.5$

when m = 3, $\frac{m(m^2-1)}{12} = \frac{3(3^2-1)}{12} = \frac{3(9-1)}{12} = \frac{3X8}{12} = \frac{24}{12} = 2$

$$\rho = 1 - \left[\frac{6\{81.5 + 0.5 + 2\}}{8(8^2-1)} \right] = 1 - \left[\frac{6X84}{8X63} \right] = 1 - \left[\frac{504}{504} \right] = 1\text{-}1 = 0$$

9. Ten competitors in a musical test were ranked by three judges A,B and C in the following order: Using rank correlation method, discuss which pair of judges has the nearest approach to common likings in music.

Ranks							
A	B	C	d_{AB}	d^2_{AB}	d_{AC}	d^2_{AC}	d_{BC}	d^2_{BC}

1	3	6	-2	4	-5	25	-3	9
6	5	4	1	1	2	4	1	1
5	8	9	-3	9	-4	16	-1	1
10	4	8	6	36	2	4	-4	16
3	7	1	-4	16	2	4	6	36
2	10	2	-8	64	0	0	8	64
4	2	3	2	4	1	1	-1	1
9	1	10	8	64	-1	1	-9	81
7	6	5	1	1	2	4	1	1
8	9	7	-1	1	1	1	2	4
			$\sum d_{AB}=$ 0	$\sum d^2_{AB}=$ 200	$\sum d_{AC}=$ 0	$\sum d^2_{AC}=$ 60	$\sum d_{BC}=$ 0	$\sum d^2_{BC}=$ 214

$$\rho_{AB} = 1 - \left[\frac{6\sum d^2}{N(N^2-1)}\right] = 1 - \left[\frac{6x200}{10(10^2-1)}\right] = 1 - \left[\frac{6X200}{10(100-1)}\right] = 1 - \left[\frac{1200}{990}\right] = 1 - 1.2121 = -0.2121$$

$$\rho_{AC} = 1 - \left[\frac{6\sum d^2}{N(N^2-1)}\right] = 1 - \left[\frac{6X60}{10(10^2-1)}\right] = 1 - \left[\frac{6X60}{10X99}\right] = 1 - \left[\frac{360}{990}\right] = 1 - 0.3636 = 0.6364$$

$$\rho_{BC} = 1 - \left[\frac{6\sum d^2}{N(N^2-1)}\right] = 1 - \left[\frac{6X214}{10(10^2-1)}\right] = 1 - \left[\frac{6X214}{10X99}\right] = 1 - \left[\frac{1284}{990}\right] = 1 - 1.2970 = -0.2970$$

ρ_{AC} is nearest to +1, so the pair A and C of judges has the nearest approach to common likings in music.

REGRESSION

The relationship between the two variables is called Regression, one variable is independent variable and the other variable is dependent variable.

The meaning of the word regression is returning or going back. The line which gives the average relationship between the variables is known as the regression line. The corresponding equation is the regression equation. The value of the dependent variable is estimated corresponding to any value of the independent variable by using the regression equation.

Methods of forming the Regression Equations

1. Regression Equations on the basis of Normal Equations
2. Regression Equations on the basis of $\overline{X}$, $\overline{Y}$, b_{XY}, b_{YX}

Regression Equations on the basis of $\overline{X}$, $\overline{Y}$, b_{XY}, b_{YX}

Regression equation of Y on X $Y - \overline{Y} = b_{YX}(X - \overline{X})$	Regression equation of X on Y $X - \overline{X} = b_{XY}(Y - \overline{Y})$

b_{YX} is called the regression coefficient of Y on X	b_{XY} is called the regression coefficient of X on Y
$b_{YX} = \frac{r\sigma_Y}{\sigma_X}$	$b_{XY} = \frac{r\sigma_X}{\sigma_Y}$
$b_{YX} = \frac{\sum xy}{\sum x^2}$	$b_{XY} = \frac{\sum xy}{\sum y^2}$
$b_{YX} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$	$b_{XY} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2}$

1. You are given the following data:

		X	Y
Arithmetic mean		36	85
Standard deviation		11	8
Correlation coefficient between X and Y	0.66		

(a) Find the two Regression equations

(b) Estimate the value of X when Y = 75

Solution:

Given; $\overline{X} = 36$, $\overline{Y} = 85$, $\sigma_X = 11$, $\sigma_Y = 8$, r = 0.66

$$b_{YX} = \frac{r\sigma_Y}{\sigma_X} = \frac{0.66X8}{11} = \frac{5.28}{11} = 0.48$$

$$b_{XY} = \frac{r\sigma_X}{\sigma_Y} = \frac{0.66X11}{8} = \frac{7.26}{8} = 0.9075$$

(a) Regression equation of Y on X

$Y - \overline{Y} = b_{YX}(X - \overline{X})$

Y – 85 = 0.48(X – 36)

Y – 85 = 0.48X – 17.28

Y = 0.48X – 17.28 + 85

Y = 0.48X + 67.72

Regression equation of X on Y

$X - \overline{X} = b_{XY}(Y - \overline{Y})$

X – 36 = 0.9075(Y- 85)

X – 36 = 0.9075Y – 77.1375

X = 0.9075Y – 77.1375 +36

X = 0.9075Y – 41.1375

X = 0.9075Y – 41.14

(b) When Y = 75, X = ?

Sub Y = 75 in the Regression equation of X on Y

X = 0.9075Y – 41.14

X = 0.9075(75) – 41.14
X = 68.0625 – 41.14
X = 26.9225
X = 26.92

2. From the following information on values of two variables X and Y find the two regression lines and the correlation coefficient:

N = 10, $\sum X = 20$, $\sum Y = 40$, $\sum X^2 = 240$, $\sum Y^2 = 410$, $\sum XY = 200$

Solution:

$$\overline{X} = \frac{\sum X}{N} = \frac{20}{10} = 2$$

$$\overline{Y} = \frac{\sum Y}{N} = \frac{40}{10} = 4$$

$$b_{YX} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum X^2 - \left(\sum X\right)^2} = \frac{10(200) - (20)(40)}{10(240) - (20)^2} = \frac{2000 - 800}{2400 - 400} = \frac{1200}{2000} = 0.6$$

$$b_{XY} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum Y^2 - \left(\sum Y\right)^2} = \frac{10(200) - (20)(40)}{10(410) - (40)^2} = \frac{2000 - 800}{4100 - 1600} = \frac{1200}{2500} = 0.48$$

(a) Regression equation of Y on X

$$Y - \overline{Y} = b_{YX}(X - \overline{X})$$

Y – 4 = 0.6(X-2)
Y -4 = 0.6X – 0.6X2
Y -4 = 0.6X – 1.2
Y = 0.6X – 1.2 + 4
Y = 0.6X 2.8

Regression equation of X on Y

$$X - \overline{X} = b_{XY}(Y - \overline{Y})$$

X – 2 = 0.48(Y- 4)
X -2 = 0.48Y – 0.48X4
X – 2 = 0.48Y – 1.92
X = 0.48Y – 1.92 +2
X = 0.48Y + 0.08

The correlation coefficient

$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

(or)

$$r = \pm\sqrt{b_{XY}Xb_{YX}}$$

$$r = +\sqrt{0.48X0.6}$$

$$r = +\sqrt{0.288}$$

r = 0.5367

3. Calculate the two regression equations from the following data:

Also estimate Y when X = 20.

X	Y	X^2	Y^2	XY
10	40	100	1600	400
12	38	144	1444	456
13	43	169	1849	559
12	45	144	2025	540
16	37	256	1369	592
15	43	225	1849	645
$\sum X = 78$	$\sum Y = 246$	$\sum X^2 = 1038$	$\sum Y^2 = 10136$	$\sum XY = 3192$

Solution:

$$\overline{X} = \frac{\sum X}{N} = \frac{78}{6} = 13, \quad \overline{Y} = \frac{\sum Y}{N} = \frac{246}{6} = 41$$

$$b_{YX} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum X^2 - \left(\sum X\right)^2} = \frac{6(3192) - (78)(246)}{6(1038) - (78)^2} = \frac{19152 - 19188}{6228 - 6084} = \frac{-36}{144} = -0.25$$

$$b_{XY} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum Y^2 - \left(\sum Y\right)^2} = \frac{6(3192) - (78)(246)}{6(10136) - (246)^2} = \frac{19152 - 19188}{60816 - 60516} = \frac{-36}{300} = -0.12$$

Regression equation of Y on X

$Y - \overline{Y} = b_{YX}(X - \overline{X})$

Y – 41 = -0.25(X – 13)

Y – 41 = -0.25X + 3.25

Y = -0.25X +3.25 + 41

Y = -0.25X + 44.25

Regression equation of X on Y

$X - \overline{X} = b_{XY}(Y - \overline{Y})$

X – 13 = -0.12 (Y – 41)

X – 13 = -0.12Y + 4.92

X = -0.12Y + 4.92 +13

X = -0.12Y + 17.92

When X = 20, Y = ?

Sub X = 20 in the Regression equation of Y on X

Y = -0.25X + 44.25

Y = -0.25(20) + 44.25

Y = -5 + 44.25

Y = 39.25

The correlation coefficient

$r = \pm\sqrt{b_{XY} X b_{YX}}$

$r = -\sqrt{-0.12X - 0.25}$

$r = -\sqrt{-0.03}$

r = -0.1732

4. From the data given below, Find

(a) The two regression equations

(b)The coefficient of correlation between the marks in Mathematics and Statistics

(c)The most likely marks in Statistics when the marks in Mathematics is 30

Solution:

Let the marks in Mathematics be X

Let the marks in Statistics be Y

X	Y	X^2	Y^2	XY
25	43	625	1849	1075
28	46	784	2116	1288
35	49	1225	2401	1715
32	41	1024	1681	1312
31	36	961	1296	1116
36	32	1296	1024	1152
29	31	841	961	899
38	30	1444	900	1140
34	33	1156	1089	1122
32	39	1024	1521	1248
$\sum X = 320$	$\sum Y = 380$	$\sum X^2 = 10380$	$\sum Y^2 = 14838$	$\sum XY = 12067$

$\overline{X} = \frac{\sum X}{N} = \frac{320}{10} = 32,\ \overline{Y} = \frac{\sum Y}{N} = \frac{380}{10} = 38$

$b_{YX} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum X^2 - \left(\sum X\right)^2} = \frac{10(12067) - (320)(380)}{10(10380) - (320)^2} = \frac{120670 - 121600}{103800 - 102400} = \frac{-930}{1400} =$ -0.6643

$b_{XY} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum Y^2 - \left(\sum Y\right)^2} = \frac{10(12067) - (320)(380)}{10(14838) - (380)^2} = \frac{120670 - 121600}{148380 - 144400} = \frac{-930}{3980} =$ -0.2337

(a) Regression equation of Y on X

$Y - \overline{Y} = b_{YX}(X - \overline{X})$

Y – 38 = -0.6643(X – 32)

Y – 38 = -0.6643X + 21.2576

Y = -0.6643X +21.2576 +38

Y = -0.6643X + 59.26

Regression equation of X on Y

$X - \overline{X} = b_{XY}(Y - \overline{Y})$

X – 32 = -0.2337(Y – 38)

X – 32 = -0.2337Y + 8.8806

X = -0.2337Y + 8.8806 + 32

X = -0.2337Y + 40.88

For finding the value of marks in Statistics (Y), when the marks in Mathematics is (X) = 30,

Substitute X = 30 in the Regression equation of Y on X

Y = -0.6643X + 59.26

Y = -0.6643(30) + 59.26

Y = -19.929 +59.26

Y = 39.33

The correlation coefficient

$r = \pm\sqrt{b_{XY} X b_{YX}}$

$r = -\sqrt{-0.2337X - 0.6643}$

$r = -\sqrt{0.15524}$

r = -0.3940

5. Height of the father and son are given below. Find the height of the son when the height of the father is 70 inches.

Father(inches) X	Son(inches) Y	X^2	Y^2	XY
71	69	5041	4761	4899
68	64	4624	4096	4352
66	65	4356	4225	4290
67	63	4489	3969	4221
70	65	4900	4225	4550
71	62	5041	3844	4402
70	65	4900	4225	4550
73	64	5329	4096	4672
72	66	5184	4356	4752
65	59	4225	3481	3835
66	62	4356	3844	4092
$\sum X = 759$	$\sum Y = 704$	$\sum X^2 = 52445$	$\sum Y^2 = 45122$	$\sum XY = 48615$

$$\overline{X} = \frac{\sum X}{N} = \frac{759}{11} = 69$$

$$\overline{Y} = \frac{\sum Y}{N} = \frac{704}{11} = 64$$

$$b_{YX} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum X^2 - \left(\sum X\right)^2} = \frac{11(48615) - (759)(704)}{11(52445) - (759)^2} = \frac{534765 - 534336}{576895 - 576081} = \frac{429}{814} = 0.5270$$

$$b_{XY} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum Y^2 - \left(\sum Y\right)^2} = \frac{11(48615) - (759)(704)}{11(45122) - (704)^2} = \frac{534765 - 534336}{496342 - 495616} = \frac{429}{726} = 0.5909$$

(a) Regression equation of Y on X

$Y - \overline{Y} = b_{YX}(X - \overline{X})$

Y - 64 = 0.5270(X – 69)

Y – 64 = 0.5270X – 69X0.5270

Y – 64 = 0.5270X – 36.363

Y = 0.5270X – 36.363 + 64

Y = 0.5270X + 27.637

Regression equation of X on Y

$X - \overline{X} = b_{XY}(Y - \overline{Y})$

X – 69 = 0.5909(Y-64)

X – 69 = 0.5909Y – 37.8176

X = 0.5909Y – 37.8176 + 69

X = 0.5909Y + 31.1824

For finding the value of Y when X = 70(father's height),

Substitute X = 70 in the Regression equation of Y on X

Y = 0.5270X + 27.637

Y = 0.5270(70) + 27.637

Y = 36.89 + 27.637

Y = 64.527

Y = 65

The correlation coefficient

$r = \pm\sqrt{b_{XY} X b_{YX}}$

$r = +\sqrt{0.5270 X 0.5909}$

$r = +\sqrt{0.3114043}$

r = 0.5580

Properties of Regression Lines and Coefficients

1. The two regression equations are generally different and are not to be interchanged in their usage.
2. The two regression lines intersect at $\left(\overline{X}, \overline{Y}\right)$.
3. Correlation coefficient is the geometric mean of the two regression coefficients. $r = \pm\sqrt{b_{XY} X b_{YX}}$
4. The two regression coefficients and the correlation coefficient have the same sign.
5. Both the regression coefficients cannot be greater than 1 numerically simultaneously.
6. Regression coefficients are independent of change of origin but are affected by change of scale
7. Each regression coefficient is in the unit of the measurement of the dependent variable
8. Each regression coefficient indicates the quantum of change in the dependent variable corresponding to unit increase in the independent variable.

Difference between Correlation and Regression:

	Correlation	Regression
1	Correlation is the relationship between variables. It is expressed numerically.	Regression means going back. The average relation between the variables is given as an equation.
2	Between two variables, non is identified as independent or dependent.	One of the variables is independent variable and the other is dependent variable in any particular context.
3	Correlation does not mean causation. One variable need not be the cause and the other effect.	Independent variable may be the 'cause' and dependent variable be the'effect'.
4.	Correlation stipulates the degree to which both of the variables can move together.	However, regression specifies the effect of the change in the unit, in the known variable on the evaluated variable