MBA Sem I

Business Statistics & Analysis for Decision Making

Module II : Correlation & Regression

Introduction:

In today's business world we come across many activities, which are dependent on each other. In businesses we see large number of problems involving the use of two or more variables. Identifying these variables and its dependency helps us in resolving the many problems. Many times there are problems or situations where two variables seem to move in the same direction such as both are increasing or decreasing. At times an increase in one variable is accompanied by a decline in another. For example, family income and expenditure, price of a product and its demand, advertisement expenditure and sales volume etc. If two quantities vary in such a way that movements in one are accompanied by movements in the other, then these quantities are said to be correlated.

Meaning:

Correlation is a statistical technique to ascertain the association or relationship between two or more variables. Correlation analysis is a statistical technique to study the degree and direction of relationship between two or more variables.

A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other.

Uses of correlations:

1. Correlation analysis helps inn deriving precisely the degree and the direction of such relationship.
2. The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality.
3. Correlation analysis contributes to the understanding of economic behavior,

aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing farces may become effective

4. Economic theory and business studies show relationships between variables like price and quantity demanded advertising expenditure and sales promotion measures etc.
5. The measure of coefficient of correlation is a relative measure of change.

Types of Correlation:

Correlation is described or classified in several different ways. Three of the most important are:

I. Positive and Negative
II. Simple, Partial and Multiple
III. Linear and non-linear

I. **Positive, Negative and Zero Correlation:**

Whether correlation is positive (direct) or negative (in-versa) would depend upon the direction of change of the variable.

Positive Correlation: If both the variables vary in the same direction, correlation is said to be positive. It means if one variable is increasing, the other on an average is also increasing or if one variable is decreasing, the other on an average is also deceasing, then the correlation is said to be positive correlation. For example, the correlation between heights and weights of a group of persons is a positive correlation.

Height (cm) : X	158	160	163	166	168	171	174	176
Weight (kg) : Y	60	62	64	65	67	69	71	72

Negative Correlation: If both the variables vary in opposite direction, the correlation is said to be negative. If means if one variable increases, but the other variable decreases or if one variable decreases, but the other variable increases, then the correlation is said to be negative correlation. For example, the correlation between the price of a product and its demand is a negative correlation.

Price of Product (Rs. Per Unit) : X	6	5	4	3	2	1
Demand (In Units) : Y	75	120	175	250	215	400

Zero Correlation: Actually it is not a type of correlation but still it is called as zero or no correlation. When we don't find any relationship between the variables then, it is said to be zero correlation. It means a change in value of one variable doesn't influence or change the value of other variable. For example, the correlation between weight of person and intelligence is a zero or no correlation.

II. **Simple, Partial and Multiple Correlation:**

The distinction between simple, partial and multiple correlation is based upon the number of variables studied.

Simple Correlation: When only two variables are studied, it is a case of simple correlation. For example, when one studies relationship between the marks secured by student and the attendance of student in class, it is a problem of simple correlation.

Partial Correlation: In case of partial correlation one studies three or more variables but considers only two variables to be influencing each other and the effect of other influencing variables being held constant. For example, in above example of relationship between student marks and attendance, the other variable influencing such as effective teaching of teacher, use of teaching aid like computer, smart board etc are assumed to be constant.

Multiple Correlation: When three or more variables are studied, it is a case of multiple correlation. For example, in above example if study covers the relationship between student marks, attendance of students, effectiveness of teacher, use of teaching aids etc, it is a case of multiple correlation.

III. **Linear and Non-linear Correlation:**

Depending upon the constancy of the ratio of change between the variables, the correlation may be Linear or Non-linear Correlation.

Linear Correlation: If the amount of change in one variable bears a constant ratio to the amount of change in the other variable, then correlation is said to be linear. If such

variables are plotted on a graph paper all the plotted points would fall on a straight line. For example: If it is assumed that, to produce one unit of finished product we need 10 units of raw materials, then subsequently to produce 2 units of finished product we need double of the one unit.

Raw material : X	10	20	30	40	50	60
Finished Product : Y	2	4	6	8	10	12

Non-linear Correlation: If the amount of change in one variable does not bear a constant ratio to the amount of change to the other variable, then correlation is said to be non-linear. If such variables are plotted on a graph, the points would fall on a curve and not on a straight line. For example, if we double the amount of advertisement expenditure, then sales volume would not necessarily be doubled.

Advertisement Expenses : X	10	20	30	40	50	60
Sales Volume : Y	2	4	6	8	10	12

Illustration 01:

State in each case whether there is

(a) Positive Correlation

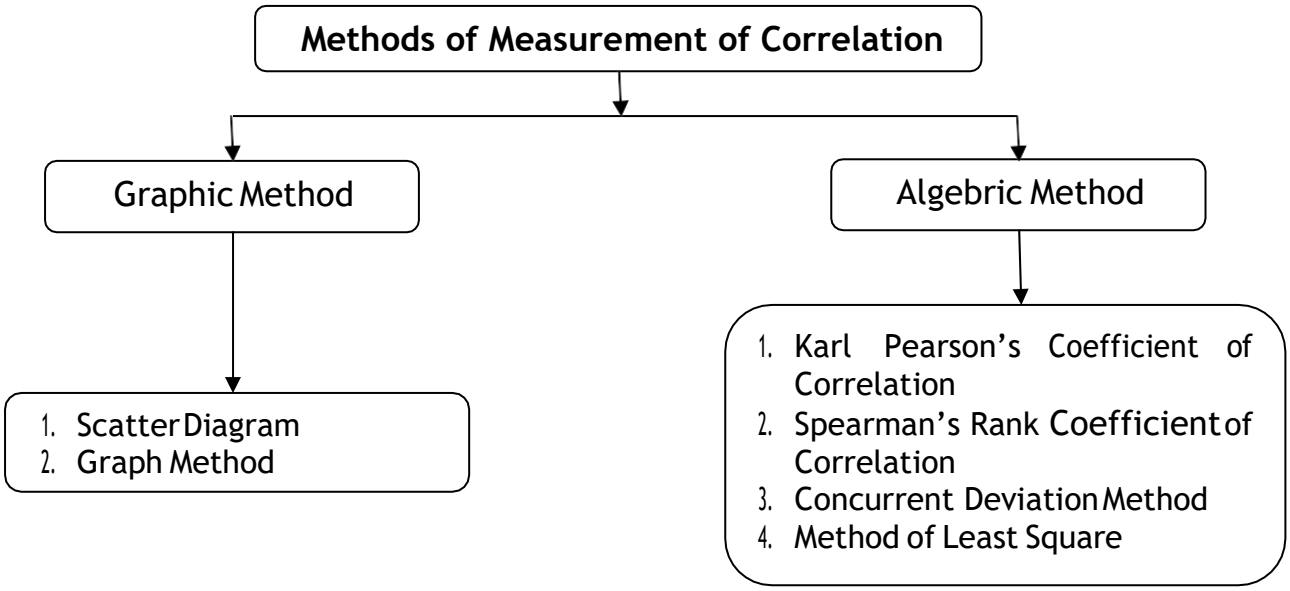
(b) Negative Correlation

(c) No Correlation

Sl No	Particulars	Solution
1	Price of commodity and its demand	Negative
2	Yield of crop and amount of rainfall	Positive
3	No of fruits eaten and hungry of a person	Negative
4	No of units produced and fixed cost per unit	Negative
5	No of girls in the class and marks of boys	No Correlation
6	Ages of Husbands and wife	Positive
7	Temperature and sale of woollen garments	Negative
8	Number of cows and milk produced	Positive
9	Weight of person and intelligence	No Correlation
10	Advertisement expenditure and sales volume	Positive

Methods of measurement of correlation:

Quantification of the relationship between variables is very essential to take the benefit of study of correlation. For this, we find there are various methods of measurement of correlation, which can be represented as given below:



Among these methods we will discuss only the following methods:

1. Scatter Diagram
2. Karl Pearson's Coefficient of Correlation
3. Spearman's Rank Coefficient of Correlation

Scatter Diagram:

This is graphic method of measurement of correlation. It is a diagrammatic representation of bivariate data to ascertain the relationship between two variables. Under this method the given data are plotted on a graph paper in the form of dot. i.e. for each pair of X and Y values we put dots and thus obtain as many points as the number of observations. Usually an independent variable is shown on the X-axis whereas the dependent variable is shown on the Y-axis. Once the values are plotted on the graph it reveals the type of the correlation between variable X and Y. A scatter diagram reveals whether the movements in one series are associated with those in the other series.

- Perfect Positive Correlation: In this case, the points will form on a straight line falling from the lower left hand corner to the upper right hand corner.
- Perfect Negative Correlation: In this case, the points will form on a straight line rising from the upper left hand corner to the lower right hand corner.
- High Degree of Positive Correlation: In this case, the plotted points fall in a narrow band, wherein points show a rising tendency from the lower left hand corner to the upper right hand corner.
- High Degree of Negative Correlation: In this case, the plotted points fall in a narrow band, wherein points show a declining tendency from upper left hand corner to the lower right hand corner.
- Low Degree of Positive Correlation: If the points are widely scattered over the

diagrams, wherein points are rising from the left hand corner to the upper right hand corner.

- Low Degree of Negative Correlation: If the points are widely scattered over the diagrams, wherein points are declining from the upper left hand corner to the lower right hand corner.
- Zero (No) Correlation: When plotted points are scattered over the graph haphazardly, then it indicate that there is no correlation or zero correlation between two variables.

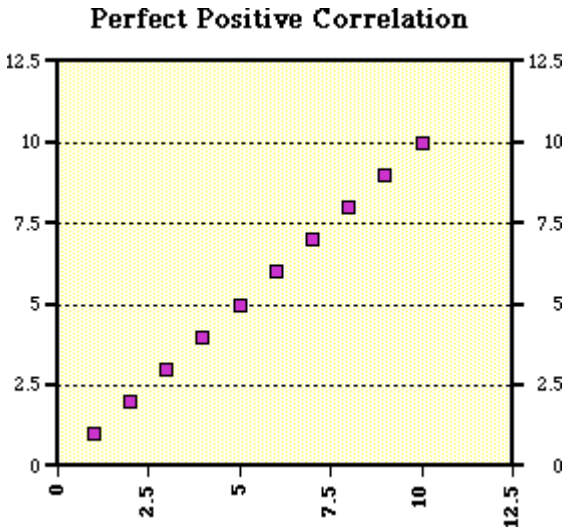


Diagram - I

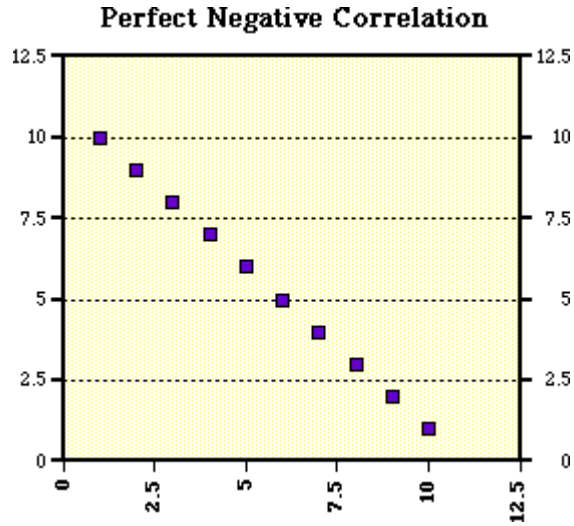


Diagram - II

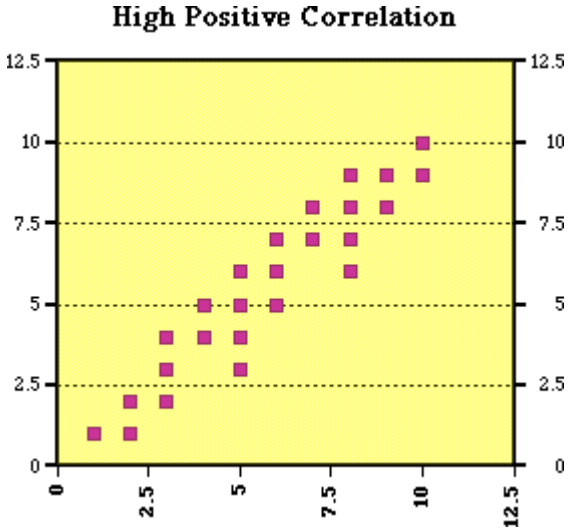


Diagram - III

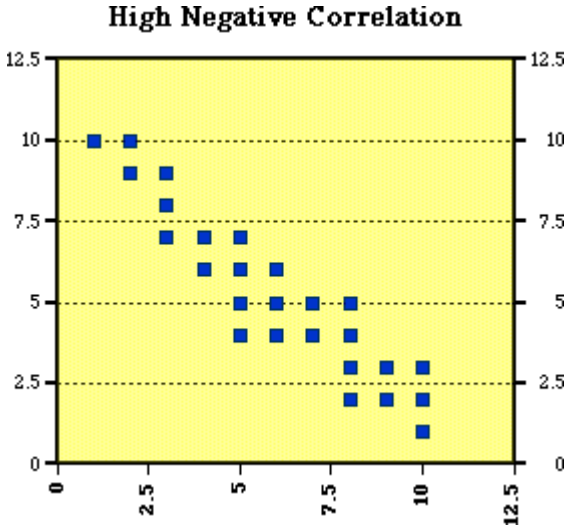


Diagram - IV

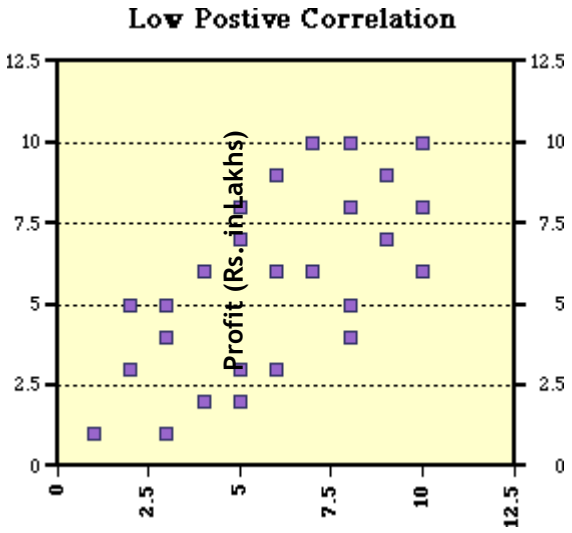


Diagram – V

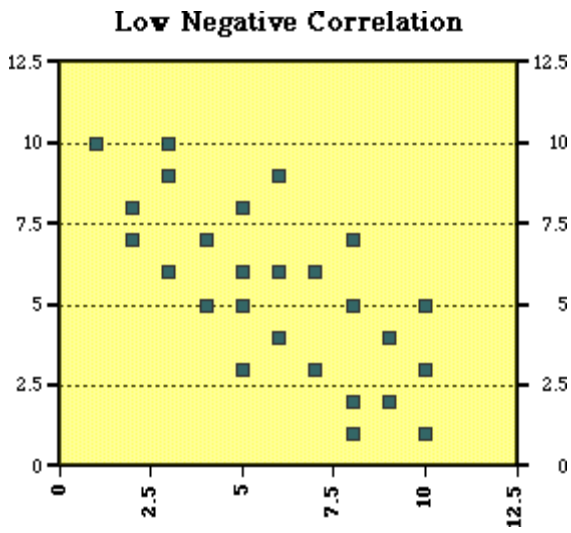


Diagram – VI

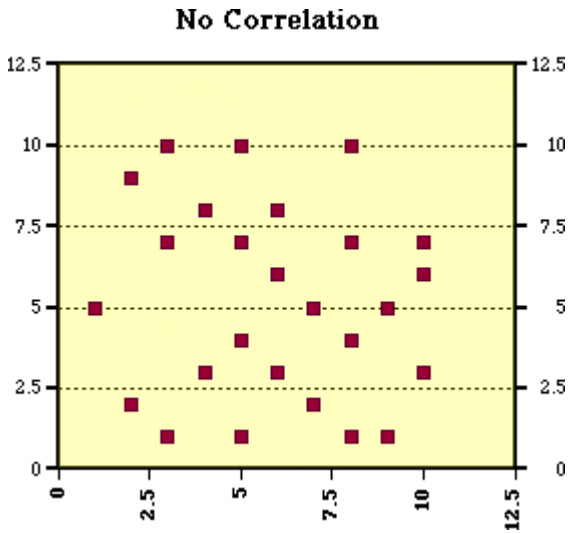


Diagram – VII

Illustration 02:

Given the following pairs of values:

Capital Employed (Rs. In Crore)	1	2	3	4	5	7	8	9	11	12
Profit (Rs. In Lakhs)	3	5	4	7	9	8	10	11	12	14

(a) Draw a scatter diagram

(b) Do you think that there is any correlation between profits and capital employed? Is it positive or negative? Is it high or low?

Solution:

From the observation of scatter diagram we can say that the variables are positively correlated. In the diagram the points trend toward upward rising from the lower left hand corner to the upper right hand corner, hence it is positive correlation. Plotted points are in narrow band which indicates that it is a case of high degree of positive correlation.

Karl Pearson's Coefficient of Correlation:

Karl Pearson's method of calculating coefficient of correlation is based on the covariance of the two variables in a series. This method is widely used in practice and the coefficient of correlation is denoted by the symbol "***r***". If the two variables under study are X and Y, the following formula suggested by Karl Pearson can be used for measuring the degree of relationship of correlation.

$$r = \frac{Covariance\ (x,y)}{S.D.(x)S.D.(y)}$$

$$r = \frac{Cov(X, Y)}{\sigma_x\ \sigma_y}$$

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \quad \text{where} \quad X = x - \bar{x},\ Y = y - \bar{y}$$

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

$$r = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2}\sqrt{(Y-\bar{Y})^2}}$$

Where, $\bar{X}$ = mean of X variable

$\bar{Y}$ = mean of Y variable

$$r = \frac{\sum f(dx)(dy) - \frac{\sum fdx)(\sum fdy)}{N}}{\sqrt{\sum (fdx) \wedge 2 - \frac{(\sum fdx) \wedge 2}{N}}\sqrt{\sum (dy) \wedge 2 - \frac{(\sum fdy) \wedge 2}{N}}}$$

$$d_x = X - A$$

$$d_y = Y - A$$

Above different formula's can be used in different situation depending upon the information given in the problem.

Illustration 03:

From following information find the correlation coefficient between advertisement expenses and sales volume using Karl Pearson's coefficient of correlation method.

Firm	1	2	3	4	5	6	7	8	9	10
Advertisement Exp. (Rs. In Lakhs)	11	13	14	16	16	15	15	14	13	13
Sales Volume (Rs. In Lakhs)	50	50	55	60	65	65	65	60	60	50

Solution:

Let us assume that advertisement expenses are variable X and sales volume are variable Y.

Calculation of Karl Pearson's coefficient of correlation

Firm	X	Y	$x = X - \dot{X}$	x^2	$y = Y - \dot{Y}$	y^2	xy
1	11	50	-3	9	-8	64	24
2	13	50	-1	1	-8	64	8
3	14	55	0	0	-3	9	0
4	16	60	2	4	2	4	4
5	16	65	2	4	7	49	14
6	15	65	1	1	7	49	7
7	15	65	1	1	7	49	7
8	14	60	0	0	2	4	0
9	13	60	-1	1	2	4	-2
10	13	50	-1	1	-8	64	8
	140	**580**		**22**		**360**	**70**
	ΣX	**ΣY**		**Σx2**		**Σy2**	**Σxy**

$$\dot{X} = \frac{\Sigma X}{n} = \frac{140}{10} = 14 \qquad \dot{Y} = \frac{\Sigma Y}{n} = \frac{580}{10} = 58$$

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \quad \text{where} \quad X = x - \bar{x}, \; Y = y - \bar{y}$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x2 \; \Sigma y2}} = \frac{70}{\sqrt{22 * 360}} = \frac{70}{88.9944} = \underline{\underline{0.7866}}$$

Interpretation: From the above calculation it is very clear that there is high degree of **positive correlation** i.e. ***r* = 0.7866,** between the two variables. i.e. Increase in advertisement expenses leads to increased sales volume.

Illustration 04:

Find the correlation coefficient between age and playing habits of the following students using Karl Pearson's coefficient of correlation method.

Age	15	16	17	18	19	20
Number of students	250	200	150	120	100	80
Regular Players	200	150	90	48	30	12

Solution:

To find the correlation between age and playing habits of the students, we need to compute the percentages of students who are having the playing habit.

Percentage of playing habits = No. of Regular Players / Total No. of Students * 100

Now, let us assume that ages of the students are variable X and percentages of playing habits are variable Y.

Calculation of Karl Pearson's coefficient of correlation

Age (X)	No of Students	Regular Players	Percentage of Playing Habits (Y)	$X - \dot{X}$	$(X - \dot{X})^2$	$Y - \dot{Y}$	$(Y - \dot{Y})^2$	$(X - \dot{X})(Y - \dot{Y})$
15	250	200	80	-2.5	6.25	30	900	-75
16	200	150	75	-1.5	2.25	25	625	-37.5
17	150	90	60	-0.5	0.25	10	100	-5
18	120	48	40	0.5	0.25	-10	100	-5
19	100	30	30	1.5	2.25	-20	400	-30
20	80	12	15	2.5	6.25	-35	1225	-87.5
105			**300**		**17.5**		**3350**	**-240**
ΣX			**ΣY**		**Σx2**		**Σy2**	**Σxy**

$$\dot{X} = \frac{\Sigma X}{n} = \frac{105}{6} = 17.5 \qquad \dot{Y} = \frac{\Sigma Y}{n} = \frac{300}{6} = 50$$

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{(Y-\bar{Y})^2}}$$

$$r = \frac{\Sigma(X-\dot{X})(Y-\dot{Y})}{\sqrt{\Sigma(X-\dot{X})2\ \Sigma(Y-\dot{Y})2}} = \frac{-240}{\sqrt{17.5 * 3350}} = \frac{-240}{242.126} = \underline{\mathbf{-0.9912}}$$

Interpretation: From the above calculation it is very clear that there is high degree of **negative correlation** i.e. ***r* = -0.9912,** between the two variables of age and playing habits. i.e. Playing habits among students decreases when their age increases.

Illustration 05:

Find Karl Pearson's coefficient of correlation between capital employed and profit obtained from the following data.

Capital Employed (Rs. In Crore)	10	20	30	40	50	60	70	80	90	100
Profit (Rs. In Crore)	2	4	8	5	10	15	14	20	22	50

Solution:

Let us assume that capital employed is variable X and profit is variable Y.

Calculation of Karl Pearson's coefficient of correlation

X	Y	X^2	Y^2	XY
10	2	100	4	20
20	4	400	16	80
30	8	900	64	240
40	5	1600	25	200
50	10	2500	100	500
60	15	3600	225	900
70	14	4900	196	980
80	20	6400	400	1600
90	22	8100	484	1980
100	50	10000	2500	5000
550	**150**	**38500**	**4014**	**11500**
ΣX	**ΣY**	**ΣX^2**	**ΣY^2**	**ΣXY**

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

$$r = \frac{n\Sigma XY - \Sigma X\ \Sigma Y}{\sqrt{[n(\Sigma X2) - (\Sigma X)2][n(\Sigma Y2) - (\Sigma Y)2]}}$$

$$r = \frac{(10 * 11500) - (550 * 150)}{\sqrt{[(10*38500) - (550^2)]\,[(10*4014) - (150^2)]}}$$

$$r = \frac{(1{,}15{,}000) - (82{,}500)}{\sqrt{[(3{,}85{,}000) - (3{,}02{,}500)]\,[(40{,}140) - (22{,}500)]}}$$

$$r = \frac{32{,}500}{\sqrt{(82{,}500)\,(17{,}640)}} = \frac{32{,}500}{\sqrt{1455300000}}$$

$$r = \frac{32{,}500}{38148.3945} = \mathbf{0.8519}$$

Illustration 06:

A computer while calculating the correlation coefficient between the variable X and Y obtained the following results:

N = 30; ΣX = 120 ΣX^2 = 600 ΣY = 90 ΣY^2 = 250 ΣXY = 335

It was, however, later discovered at the time of checking that it had copied down two pairs of observations as: (X, Y) : (8, 10) (12, 7)

While the correct values were: (X, Y) : (8, 12) (10, 8)

Obtain the correct value of the correlation coefficient between X and Y.

Solution:

Correct ΣX = 120 – 8 – 12 + 8 + 10 = 118

Correct ΣX^2 = $600 - 8^2 - 12^2 + 8^2 + 10^2$

= 600 – 64 – 144 + 64 + 100 = 556

Correct ΣY = 90 – 10 – 7 + 12 + 8 = 93

Correct ΣY^2 = $250 - 10^2 - 7^2 + 12^2 + 8^2$

= 250 – 100 – 49 + 144 + 64 = 309

Correct ΣXY = 335 – (8*10) – (12*7) + (8*12) + (10*8)

= 335 – 80 – 84 + 96 + 80 = 347

$$r = \frac{n\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{[n(\Sigma X2) - (\Sigma X)2][n(\Sigma Y2) - (\Sigma Y)2]}}$$

$$r = \frac{(30 * 347) - (118 * 93)}{\sqrt{[(30*556) - (118^2)]\,[(30*309) - (93^2)]}}$$

$$r = \frac{(10{,}410) - (10{,}974)}{\sqrt{[(16{,}680) - (13{,}924)]\,[(9270) - (8649)]}}$$

$$r = \frac{564}{\sqrt{(2{,}756)\,(621)}} = \frac{-564}{\sqrt{1711476}}$$

$$r = \frac{-564}{1308.2339} = \underline{\underline{\mathbf{-0.4311}}}$$

Therefore, the correct value of correlation coefficient between X and Y is moderately negative correlation of -0.4311.

Illustration 07:

Coefficient of correlation between X and Y is 0.3. Their covariance is 9. The variance of X is 16. Find the standard devotion of Y series.

Solution:

Given information:

$r = 0.3$ Cov (X, Y) = 9 Var (X) = 16

$$r = \frac{Cov(X,F)}{\sqrt{Var(X) * Var(F)}} \qquad 0.3 = \frac{9}{\sqrt{16 * Var(F)}} \qquad 0.3 = \frac{9}{4 * \sqrt{Var(F)}}$$

$$0.3 * 4 = \frac{9}{SD(F)} \qquad 1.2 = \frac{9}{SD(F)} \qquad SD(Y) = \frac{9}{1.2} = 7.5$$

Therefore the standard deviation of Y series = σ(Y) = 7.5

Illustration 08:

Calculate correlation coefficient from the following two-way table, with X representing the average salary of families selected at random in a given area and Y representing the average expenditure on entertainment.

Expenditure on	**Average Salary (Rs. '000)**				
Entertainment (Rs. '000)	100-150	150-200	200-250	250-300	300-350
0 – 10	5	4	5	2	4
10 – 20	2	7	3	7	1
20 – 30	-	6	-	4	5
30 – 40	8	-	4	-	8
40 – 50	-	7	3	5	10

Solution:

Let us assume that Average Salary is variable X and Expenditure on Entertainment is variable Y.

In case of grouped data, we need to follow the assumed mean method to calculate Karl Pearson's Coefficient of Correlation. Following steps are followed to compute correlation.

1. Identify the mid-point of the class intervals for variable X and Y.
2. Chose an assumed mean from the mid-point identified above for both X and Y.
3. To simplify further, deviation from assumed mean is computed by dividing deviation by a common factor.
4. Add the values in cell, row-wise and column-wise, to compute frequencies (f). Sum of either row-wise or column-wise represent the value of N.
5. Obtain the product of d_x and d_y and the corresponding frequencies (f) in each cell. Write the figure thus obtained in the right corner of each cell which represent the value of fd_xd_y.

Calculation of Karl Pearson's coefficient of correlation

X / Y	Mid Point	100 - 150	150 - 200	200 - 250	250 - 300	300 - 350	f	d_y	fd_y	fd_y^2	fd_xd_y
0 - 10	5	20 / 5	8 / 4	0 / 5	-4 / 2	-16 / 4	20	-2	-40	80	8
10 - 20	15	4 / 2	7 / 7	0 / 3	-7 / 7	-2 / 1	20	-1	-20	20	2
20 - 30	25	- / -	0 / 6	- / -	0 / 4	0 / 5	15	0	0	0	0
30 - 40	35	-16 / 8	- / -	0 / 4	- / -	16 / 8	20	1	20	20	0
40 - 50	45	- / -	-14 / 7	0 / 3	10 / 5	40 / 10	25	2	50	100	36
f		15	24	15	18	28	100 = N		**10**	**220**	**46**
d_x		-2	-1	0	1	2			Σfd_y	Σfd_y^2	Σfd_xd_y
fd_x		-30	-24	0	18	56	20	Σfd_x			
fd_x^2		60	24	0	18	112	214	Σfd_x^2			
fd_xd_y		8	1	0	-1	38	46	Σfd_xd_y			

d_x = Mid Point of Series X - Assumed Mean of Series X = MP(X) - 225
d_y = Mid Point of Series Y - Assumed Mean of Series Y = MP(Y) - 25

$$r = \frac{n\Sigma dxdy - \Sigma fdx\ \Sigma fdy}{\sqrt{[n(\Sigma dx^2) - (\Sigma dx)^2][n(\Sigma dy^2) - (\Sigma dy)^2]}} = \frac{(100 * 46) - (20 * 10)}{\sqrt{[(100 * 214) - (20)^2][(100 * 220) - (10)^2]}}$$

$$r = \frac{(4{,}600) - (200)}{\sqrt{[21{,}400 - 400][22{,}000 - 100]}} = \frac{4{,}400}{\sqrt{[21{,}000] * [21{,}900]}} = \frac{4{,}400}{21{,}445.2792} = \underline{\mathbf{0.2052}}$$

Interpretation: From the above calculation it is very clear that there is low degree of **positive correlation** i.e. r **= 0.2052,** between the two variables of salary and expenditure. It means average salary of income have slightly or low influence over entertainment expenditure.

Spearman's Rank Coefficient of Correlation:

When quantification of variables becomes difficult such beauty of female, leadership ability, knowledge of person etc, then this method of rank correlation is useful which was developed by British psychologist Charles Edward Spearman in 1904. In this method ranks are allotted to each element either in ascending or descending order. The correlation coefficient between these allotted two series of ranks is popularly called as "Spearman's Rank Correlation" and denoted by "***R***".

To find out correlation under this method, the following formula is used.

$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$ where, D =Difference of the ranks between paired items in two series.

N = Number of pairs of ranks

In case of tie in ranks or equal ranks:

In some cases it may be possible that it becomes necessary to assign same rank to two or more elements or individual or entries. In such situation, it is customary to give each individual or entry an average rank. For example, if two individuals are ranked equal to 5th place, then both of them are allotted with common rank (5+6)/2 = 5.5 and if three are ranked in 5th place, then they are given the rank of (5+6+7)/3 = 6. It means where two or more individuals are to be ranked equal, the rank assigned for the purpose of calculating coefficient of correlation is the average of the ranks with these individual or items or entries would have got had they differed slightly with each other.

Where equal ranks are assigned to some entries, an adjustment factor is to be added to the value of $6\Sigma D^2$ in the above formula for calculating the rank coefficient correlation. This adjustment factor is to be added for every repetition of rank.

Adjustment factor = $\frac{1}{12}(m_1^3 - m_1)$ where, m = number of items whose rank are common

For example, if a particular rank repeated two times then m=2 and if it repeats three times then m= 3 and so on.

Hence the above formula can be re-written as follows:

$$R = 1 - \frac{6 * [\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots]}{N^3 - N}$$

Illustration 09:

Find out spearman's coefficient of correlation between the two kinds of assessment of graduate students' performance in a college.

Name of students	A	B	C	D	E	F	G	H	I
Internal Exam	51	68	73	46	50	65	47	38	60
External Exam	49	72	74	44	58	66	50	30	35

Solution:

Calculation of Spearman's Rank Coefficient of Correlation

Name	Internal Exam	Ranks (R_1)	External Exam	Ranks (R_2)	$D = R_1 - R_2$	D^2
A	51	5	49	6	-1	1
B	68	2	72	2	0	0
C	73	1	74	1	0	0
D	46	8	44	7	1	1
E	50	6	58	4	2	4
F	65	3	66	3	0	0
G	47	7	50	5	2	4
H	36	9	30	9	0	0
I	60	4	35	8	-4	16
					ΣD^2 =	**26**

$$\mathbf{R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6*26}{9^3 - 9} = 1 - \frac{156}{729 - 9} = 1 - \frac{156}{720} = 1 - 0.2167 = \underline{\underline{0.7833}}}$$

Interpretation: From the above calculation it is very clear that there is high degree of **positive correlation** i.e. **R = 0.7833,** between two exams. It means there is a high degree of positive correlation between the internal exam and external exam of the students.

Illustration 10:

The coefficient of rank correlation of the marks obtained by 10 students in statistics and accountancy was found to be 0.8. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 7 instead of 9. Find the correct coefficient of rank correlation.

Solution:

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} \Rightarrow \quad 0.8 = 1 - \frac{6\Sigma D^2}{10^3 - 10} \Rightarrow 0.8 = 1 - \frac{6\Sigma D^2}{990} \Rightarrow \quad \frac{6\Sigma D^2}{990} = 1 - 0.8 \Rightarrow$$

$$\frac{6\Sigma D^2}{990} = 0.2 \Rightarrow \quad 6\Sigma D^2 = 0.2 * 990 \Rightarrow \quad \Sigma D^2 = 198/6 \Rightarrow \quad \Sigma D^2 = 33$$

But this is not correct ΣD^2 therefore we need to compute correct value

Correct $\Sigma D^2 = 33 - 7^2 + 9^2 = 65$

Hence, correct value of rank coefficient of correlation is:

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6*65}{990} = 1 - \frac{390}{990} = 1 - 0.394 = \underline{\underline{\mathbf{0.606}}}$$

Illustration 11:

Ten competitors in a beauty contest are ranked by three judges in the following order:

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

Use the rank correlation coefficient to determine which pairs of judges has the nearest approach to common tastes in beauty.

Solution:

In order to find out which pair of judges has the nearest approach to common tastes in beauty, we compare rank correlation between the judgements of

1. 1st Judge and 2nd Judge
2. 2nd Judge and 3rd Judge
3. 1st Judge and 3rd Judge

Calculation of Spearman's Rank Coefficient of Correlation

Rank by 1st Judge (R_1)	Rank by 2nd Judge (R_2)	Rank by 3rd Judge (R_3)	$D^2 = (R_1-R_2)^2$	$D^2 = (R_2-R_3)^2$	$D^2 = (R_1-R_3)^2$
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
N = 10	**N = 10**	**N = 10**	**$\sum D^2 = 200$**	**$\sum D^2 = 214$**	**$\sum D^2 = 60$**

1. 1st Judge and 2nd Judge: $R = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6*200}{10^3 - 10} = 1 - \frac{1200}{990} = 1 - 1.2121 = \mathbf{-0.2121}$
2. 2nd Judge and 3rd Judge: $R = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6*214}{10^3 - 10} = 1 - \frac{1284}{990} = 1 - 1.297 = \mathbf{-0.297}$
3. 1st Judge and 3rd Judge: $R = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6*60}{10^3 - 10} = 1 - \frac{360}{990} = 1 - 0.3636 = \mathbf{0.6364}$

Interpretation: From the above calculation it can be observed that coefficient of correlation is positive in the judgement of the first and third judges. Therefore, it can be concluded that first and third judges have the nearest approach to common tastes in beauty.

Illustration 12:

From the following data, compute the rank correlation.

X	82	68	75	61	68	73	85	68
Y	81	71	71	68	62	69	80	70

Solution:

In the problem we find there are repetitions of ranks. Value of X = 68 repeated 3 times and Value of Y = 71 repeated 2 times. Therefore we need to compute adjustment factor to be added to the value of $\sum D^2$.

Calculation of Spearman's Rank Coefficient of Correlation

X	Y	R_1	R_2	$D= R_1 - R_2$	D^2
82	81	2	1	1	1
68	71	6	3.5	2.5	6.25
75	71	3	3.5	-0.5	0.25
61	68	8	7	1	1
68	62	6	8	-2	4
73	69	4	6	-2	4
85	80	1	2	-1	1
68	70	6	5	1	1
				ΣD^2	**18.5**

$$R = 1 - \frac{6 * [\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)]}{N^3 - N}$$

When value X repeated three times, m=3,

Adjustment factor (1) = $\frac{1}{12}(3^3-3) = \frac{1}{12} * (27-3) = \frac{1}{12} * 24 = 2$

When value Y repeated two times, m=2,

Adjustment factor (2) = $\frac{1}{12}(2^3-2) = \frac{1}{12} * (8-2) = \frac{1}{12} * 6 = 0.5$

$$R = 1 - \frac{6 * [18.5 + 2 + 0.5]}{8^3 - 8} = 1 - \frac{6 * 21}{512 - 8} = 1 - \frac{126}{504} = 1 - 0.25 = \mathbf{0.75}$$

Spearman's Rank Coefficient of Correlation = 0.75, which indicates there is high degree of positive correlation.

Properties of Coefficient of Correlation:

1. The coefficient of correlation always lies between – 1 to +1, symbolically it can written as $-1 \le r \le 1$.
2. The coefficient of correlation is independent of change of origin and scale.
3. The coefficient of correlation is a pure number and is independent of the units of measurement. It means if X represent say height in inches and Y represent say weights in kgs, then the correlation coefficient will be neither in inches nor in kgs but only a pure number.
4. The coefficient of correlation is the geometric mean of two regression coefficient, symbolically $r = \pm\sqrt{bxy * byx}$
5. If X and Y are independent variables then coefficient of correlation is zero.

Probable Error:

The **Probable Error of Correlation Coefficient** helps in determining the accuracy and reliability of the value of the coefficient that in so far depends on the random sampling.

In other words, the probable error (P.E.) is the value which is added or subtracted from the coefficient of correlation **(r)** to get the upper limit and the lower limit respectively, within which the value of the correlation expectedly lies.

The probable error of correlation coefficient can be obtained by applying the following formula:

$$\mathbf{P.E.r = 0.6745\frac{1-r^2}{\sqrt{N}}}$$

r = coefficient of correlation
N = number of observations

- There is **no correlation** between the variables if the value of **'r' is less than P.E.** This shows that the coefficient of correlation is not at all significant.
- The correlation is said to be **certain** when the value of **'r' is six times more than the probable error**; this shows that the value of 'r' is significant.
- By **adding and subtracting the value of P.E** from the value of **'r,'** we get the upper limit and the lower limit, respectively within which the correlation of coefficient is expected to lie. Symbolically, it can be expressed

$$\mathbf{\rho(rho) = r \pm P.E.r}$$

where rho denotes the correlation in a population

The probable Error can be used only when the following three conditions are fulfilled:

1. The data must approximate to the bell-shaped curve, i.e. a **normal frequency curve.**
2. The Probable error computed from the statistical measure must have been **taken from the sample**.
3. The sample items must be selected in an **unbiased manner** and must be **independent of each other**.

Thus, the probable error is calculated to check the reliability of the value of coefficient calculated from the random sampling.

Coefficient of Determination:

The coefficient of Determination gives the percentage variation in the dependent variable that is accounted for by the independent variable.
In other words, the coefficient of determination gives the ratio of the explained variance to the total variance.
The coefficient of Determination is given by the square of the correlation coefficient, i.e., r^2. Thus,

$$Coefficient of Determination = r^2 = \frac{Explained\ variance}{Total\ variance}$$

Standard Error:

The Standard Error of a Correlation Coefficient The SE of a correlation coefficient r is computed by normalizing the fraction of the unexplained variations with respect to n – 2 degrees of freedom; i.e.

$$SEr = \text{sqrt}\ (1-r^2)\ /\ \text{sqrt}\ (n-2)$$

r^2 is the Coefficient of Determination which expresses the fraction of the explained variations; e.g., variations in y as the result of variations in x. To illustrate, if r^2 is 0.90, the independent variable y is said to explain 90% of the variance in the dependent variable x, but does not explain $(1-r^2)$ or 10% of the variance in the dependent variable.

REGRESSION

Meaning:

A study of measuring the relationship between associated variables, wherein one variable is dependent on another independent variable, called as Regression. It is developed by Sir Francis Galton in 1877 to measure the relationship of height between parents and their children.

Regression analysis is a statistical tool to study the nature and extent of functional relationship between two or more variables and to estimate (or predict) the unknown values of dependent variable from the known values of independent variable.

The variable that forms the basis for predicting another variable is known as the Independent Variable and the variable that is predicted is known as dependent variable. For example, if we know that two variables price (X) and demand (Y) are closely related we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X. Similarly, if we know that the amount of tax and the rise in the price of a commodity are closely related, we can find out the expected price for a certain amount of tax levy.

Uses of Regression Analysis:

1. It provides estimates of values of the dependent variables from values of independent variables.
2. It is used to obtain a measure of the error involved in using the regression line as a basis for estimation.
3. With the help of regression analysis, we can obtain a measure of degree of association or correlation that exists between the two variables.
4. It is highly valuable tool in economies and business research, since most of the problems of the economic analysis are based on cause and effect relationship.

Distinction between Correlation and Regression

Sl No	Correlation	Regression
1	It measures the degree and direction of relationship between the variables.	It measures the nature and extent of average relationship between two or more variables in terms of the original units of the data
2	It is a relative measure showing association between the variables.	It is an absolute measure of relationship.
3	Correlation Coefficient is independent of change of both origin and scale.	Regression Coefficient is independent of change of origin but not scale.
4	Correlation Coefficient is independent of units of measurement.	Regression Coefficient is not independent of units of measurement.
5	Expression of the relationship between the variables ranges from –1 to +1.	Expression o the relationship between the variables may be in any of the forms like: $Y = a + bX$ $Y = a + bX + cX^2$
6	It is not a forecasting device.	It is a forecasting device which can be used to predict the value of dependent variable from the given value of independent variable.
7	There may be zero correlation such as weight of wife and income of husband.	There is nothing like zero regression.

Regression Lines and Regression Equation:

Regression lines and regression equations are used synonymously. Regression equations are algebraic expression of the regression lines. Let us consider two variables: X & Y. If y depends on x, then the result comes in the form of simple regression. If we take the case of two variable X and Y, we shall have two regression lines as the regression line of X on Y and regression line of Y on X. The regression line of Y on X gives the most probable value of Y for given value of X and the regression line of X on Y given the most probable value of X for given value of Y. Thus, we have two regression lines. However, when there is either perfect positive or perfect negative correlation between the two variables, the two regression line will coincide, i.e. we will have one line. If the variables are independent, r is zero and the lines of regression are at right angles i.e. parallel to X axis and Y axis.

Therefore, with the help of simple linear regression model we have the following two regression lines

1. Regression line of Y on X: This line gives the probable value of Y (Dependent variable) for any given value of X (Independent variable).
   Regression line of Y on X : $Y - \dot{Y} = b_{yx} (X - \dot{X})$
   OR : $Y = a + bX$

2. Regression line of X on Y: This line gives the probable value of X (Dependent variable) for any given value of Y (Independent variable).
   Regression line of X on Y : $X - \dot{X} = b_{xy} (Y - \dot{Y})$
   OR : $X = a + bY$

In the above two regression lines or regression equations, there are two regression parameters, which are "a" and "b". Here "a" is unknown constant and "b" which is also denoted as "b_{yx}" or "b_{xy}", is also another unknown constant popularly called as regression coefficient. Hence, these "a" and "b" are two unknown constants (fixed numerical values) which determine the position of the line completely. If the value of either or both of them is changed, another line is determined. The parameter "a" determines the level of the fitted line (i.e. the distance of the line directly above or below the origin). The parameter "b" determines the slope of the line (i.e. the change in Y for unit change in X).

If the values of constants "a" and "b" are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the method of least squares. With the little algebra and differential calculus, it can be shown that the following two ***normal equations***, if solved simultaneously, will yield the values of the parameters "a" and "b".

Two normal equations:

X on Y	Y on X
$\Sigma X = Na + b\Sigma Y$ $\Sigma XY = a\Sigma Y + b\Sigma Y^2$	$\Sigma Y = Na + b\Sigma X$ $\Sigma XY = a\Sigma X + b\Sigma X^2$

This above method is popularly known as direct method, which becomes quite cumbersome when the values of X and Y are large. This work can be simplified if instead of dealing with actual values of X and Y, we take the deviations of X and Y series from their respective means. In that case:

Regression equation Y on X:

$Y = a + bX$ will change to $(Y - \dot{Y}) = b_{yx} (X - \dot{X})$

Regression equation X on Y:

$X = a + bY$ will change to $(X - \dot{X}) = b_{xy} (Y - \dot{Y})$

In this new form of regression equation, we need to compute only one parameter i.e. "b". This "b" which is also denoted either "b_{yx}" or "b_{xy}" which is called as regression coefficient.

Regression Coefficient:

The quantity "b" in the regression equation is called as the regression coefficient or slope coefficient. Since there are two regression equations, therefore, we have two regression coefficients.

1. Regression Coefficient X on Y, symbolically written as "b_{xy}"
2. Regression Coefficient Y on X, symbolically written as "b_{yx}" Different formula's used to compute regression coefficients:

Method	Regression Coefficient X on Y	Regression Coefficient Y on X
Using the correlation coefficient (r) and standard deviation (σ)	$b_{xy} = r \frac{\sigma_s}{\sigma_y}$	$b_{yx} = r \frac{\sigma_y}{\sigma_s}$
Direct Method: Using sum of X and Y	$b_{xy} = \frac{N\Sigma XY - \Sigma X\Sigma Y}{N\Sigma Y^2 - (\Sigma Y)^2}$	$b_{yx} = \frac{N\Sigma XY - \Sigma X\Sigma Y}{N\Sigma X^2 - (\Sigma X)^2}$
When deviations are taken from arithmetic mean	$b_{xy} = \frac{\Sigma sy}{\Sigma y^2}$ where $x = X - \dot{X}$ and $y = Y - \dot{Y}$	$b_{yx} = \frac{\Sigma sy}{\Sigma s^2}$ where $x = X - \dot{X}$ and $y = Y - \dot{Y}$

Properties of Regression Coefficients:

1. The coefficient of correlation is the geometric mean of the two regression coefficients. Symbolically $r = \pm\sqrt{bxy * byx}$
2. If one of the regression coefficients is greater than unity, the other must be less than unity, since the value of the coefficient of correlation cannot exceed unity. For example if bxy = 1.2 and byx = 1.4 "r" would be = $\sqrt{1.2 * 1.4} = 1.29$, which is not possible.
3. Both the regression coefficient will have the same sign. i.e. they will be either positive or negative. In other words, it is not possible that one of the regression coefficients is having minus sign and the other plus sign.
4. The coefficient of correlation will have the same sign as that of regression coefficient, i.e. if regression coefficient have a negative sign, "r" will also have negative sign and if the regression coefficient have a positive sign, "r" would also be positive. For example, if $b_{xy} = -0.2$ and $b_{yx} = -0.8$ then $r = -\sqrt{0.2 * 0.8} = \mathbf{-0.4}$
5. The average value of the two regression coefficient would be greater than the value of coefficient of correlation. In symbol (bxy + byx) / 2 > r. For example, if bxy = 0.8 and byx = 0.4 then average of the two values = (0.8 + 0.4) / 2 = 0.6 and the value of r = r = $\sqrt{0.8 * 0.4} = 0.566$ which less than 0.6
6. Regression coefficients are independent of change of origin but not scale.

Illustration 01:

Find the two regression equation of X on Y and Y on X from the following data:

X	:	10	12	16	11	15	14	20	22
Y	:	15	18	23	14	20	17	25	28

Solution:

Calculation of Regression Equation

X	Y	X^2	Y^2	XY
10	15	100	225	150
12	18	144	324	216
16	23	256	529	368
11	14	121	196	154
15	20	225	400	300
14	17	196	289	238
20	25	400	625	500
22	28	484	784	616
120	**160**	**1,926**	**3,372**	**2,542**
∑X	**ΣY**	$\sum X^2$	$\sum Y^2$	∑XY

Here N = Number of elements in either series X or series Y = 8

Now we will proceed to compute regression equations using normal equations.

Regression equation of X on Y: X = a + bY

The two normal equations are:

$$\sum X = Na + b\sum Y$$

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

Substituting the values in above normal equations, we get

120 = 8a + 160b ..... (i)

2542 = 160a + 3372b ..... (ii)

Let us solve these equations (i) and (ii) by simultaneous equation method

Multiply equation (i) by 20 we get 2400 = 160a + 3200b

Now rewriting these equations:

2400 = 160a + 3200b

2542 = 160a + 3372b

(-) (-) (-)

-142 = -172b

Therefore now we have -142 = -172b, this can rewritten as 172b = 142

Now, b = $\frac{142}{172}$ = 0.8256 (rounded off)

Substituting the value of b in equation (i), we get

120 = 8a + (160 *0.8256)

120 = 8a + 132 (rounded off)

8a = 120 - 132

8a = -12

a = -12/8

a = -1.5

Thus we got the values of a = -1.5 and b = 0.8256

Hence the required regression equation of X on Y:

X = a + bY => **X = -1.5 + 0.8256Y**

Regression equation of Y on X: Y = a + bX

The two normal equations are:

$\Sigma Y = Na + b\Sigma X$

$\Sigma XY = a\Sigma X + b\Sigma X^2$

Substituting the values in above normal equations, we get

160 = 8a + 120b ..... (iii)

2542 = 120a + 1926b ..... (iv)

Let us solve these equations (iii) and (iv) by simultaneous equation method

Multiply equation (iii) by 15 we get 2400 = 120a + 1800b

Now rewriting these equations:

2400 = 120a + 1800b

2542 = 120a + 1926b

(-) (-) (-)

-142 = -126b

Therefore now we have -142 = -126b, this can rewritten as 126b = 142

Now, b = $\frac{142}{126}$ = 1.127 (rounded off)

Substituting the value of b in equation (iii), we get

160 = 8a + (120 * 1.127)

160 = 8a + 135.24

8a = 160 - 135.24
8a = 24.76
a = 24.76/8
a = 3.095

Thus we got the values of a = 3.095 and b = 1.127

Hence the required regression equation of Y on X:

Y = a + bX => **Y = 3.095 + 1.127X**

Illustration 02:

After investigation it has been found the demand for automobiles in a city depends mainly, if not entirely, upon the number of families residing in that city. Below are the given figures for the sales of automobiles in the five cities for the year 2019 and the number of families residing in those cities.

City	No. of Families (in lakhs): X	Sale of automobiles (in '000): Y
Belagavi	70	25.2
Bangalore	75	28.6
Hubli	80	30.2
Kalaburagi	60	22.3
Mangalore	90	35.4

Fit a linear regression equation of Y on X by the least square method and estimate the sales for the year 2020 for the city Belagavi which is estimated to have 100 lakh families assuming that the same relationship holds true.

Solution:

Calculation of Regression Equation

City	X	Y	X^2	XY
Belagavi	70	25.2	4900	1764
Bangalore	75	28.6	5625	2145
Hubli	80	30.2	6400	2416
Kalaburagi	60	22.3	3600	1338
Mangalore	90	35.4	8100	3186
	375	**141.7**	**28,625**	**10,849**
	$\sum X$	**$\sum Y$**	$\sum X^2$	$\sum XY$

Regression equation of Y on X: Y = a + bX

The two normal equations are:

$\sum Y = Na + b\sum X$

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

Substituting the values in above normal equations, we get

141.7 = 5a + 375b ..... (i)

10849= 375a + 28625b ..... (ii)

Let us solve these equations (i) and (ii) by simultaneous equation method

Multiply equation (i) by 75 we get 10627.5 = 375a + 28125b

Now rewriting these equations:

10627.5	=	375a	+	28125b	
10849	=	375a	+	28625b	
(-)		(-)		(-)	.
-221.5	=			-500b	

Therefore now we have -221.5 = -500b, this can rewritten as 500b = 221.5

Now, $b = \frac{221.5}{500} = 0.443$

Substituting the value of b in equation (i), we get

141.7 =	5a	+	(375 * 0.443)	
141.7 =	5a	+	166.125	
5a	=	141.7 -	166.125	
5a	=	-24.425		
a	=	-24.425/5		
a	=	-4.885		

Thus we got the values of a = -4.885 and b = 0.443

Hence, the required regression equation of Y on X:

Y = a + bX => **Y = -4.885 + 0.443X**

Estimated sales of automobiles (Y) in city Belagavi for the year 2020, where number of families (X) are 100(in lakhs):

Y = -4.885 + 0.443X

Y = -4.885 + (0.443 * 100)

Y = -4.885 + 44.3

Y = 39.415 ('000)

Means sales of automobiles would be 39,415 when number of families are 100,00,000

Illustration 03:

From the following data obtain the two regression lines:

Capital Employed (Rs. in lakh):	7	8	5	9	12	9	10	15
Sales Volume (Rs. in lakh):	4	5	2	6	9	5	7	12

Solution:

Calculation of Regression Equation

X	Y	X^2	Y^2	XY
7	4	49	16	28
8	5	64	25	40
5	2	25	4	10
9	6	81	36	54
12	9	144	81	108
9	5	81	25	45
10	7	100	49	70
15	12	225	144	180
75	**50**	**769**	**380**	**535**
$\sum X$	**ΣY**	$\sum X^2$	$\sum Y^2$	$\sum XY$

Regression line / equation of X on Y:

$(X - \dot{X}) = b_{xy} (Y - \dot{Y})$

$\dot{X} = \frac{\Sigma X}{n} = \frac{75}{8} = 9.375$

$\dot{Y} = \frac{\Sigma Y}{n} = \frac{50}{8} = 6.25$

Regression coefficient of X on Y:

$b_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma Y^2 - (\Sigma Y)^2}$

$b_{xy} = \frac{(8 * 535) - (75 * 50)}{(8 * 380) - (50)^2}$

$= \frac{4280 - 3750}{3040 - 2500}$

$= \frac{530}{540} = \underline{0.9815}$

$(X - \dot{X}) = b_{xy} (Y - \dot{Y})$

$X - 9.375 = 0.9815 (Y - 6.25)$

$X - 9.375 = 0.9815Y - 6.1344$

$X = 9.375 - 6.1344 + 0.9815Y$

$X = 3.2406 + 0.9815Y$

Regression line / equation of Y on X:

$(Y - \dot{Y}) = b_{yx} (X - \dot{X})$

$\dot{X} = \frac{\Sigma X}{n} = \frac{75}{8} = 9.375$

$\dot{Y} = \frac{\Sigma Y}{n} = \frac{50}{8} = 6.25$

Regression coefficient of Y on X:

$b_{yx} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$

$b_{yx} = \frac{(8 * 535) - (75 * 50)}{(8 * 769) - (75)^2}$

$= \frac{4280 - 3750}{6152 - 5625}$

$= \frac{530}{527} = \underline{1.0057}$

$(Y - \dot{Y}) = b_{yx} (X - \dot{X})$

$Y - 6.25 = 1.0057 (X - 9.375)$

$Y - 6.25 = 1.0057X - 9.4284)$

$Y = 6.25 - 9.4284 + 1.0057X$

$Y = -3.1784 + 1.0057X$

Illustration 04:

From the following information find regression equations and estimate the production when the capacity utilisation is 70%.

	Average (Mean)	Standard Deviation
Production (in lakh units)	42	12.5
Capacity Utilisation (%)	88	8.5
Correlation Coefficient (*r*)	0.72	

Solution:

Let production be variable X and capacity utilisation be variable Y. Regression equation of production based on based on capacity utilisation shall be given by X on Y and regression equation of capacity utilisation of production shall be given by Y on X, which can be computed as given below:

Given Information: $\dot{X} = 42$ $\dot{Y} = 88$ $\sigma_x = 12.5$ $\sigma_y = 8.5$ $r = 0.72$

Regression coefficient of X on Y:

$b_{xy} = r\frac{a_s}{a_y} = 0.72 * \frac{12.5}{8.5} = 1.0588$

Regression Equation of X on Y:

$(X - \dot{X}) = b_{xy} (Y - \dot{Y})$

$X - 42 = 1.0588 (Y - 88)$

$X = 42 - 93.1744 + 1.0588Y$

$X = -51.1744 + 1.0588Y$

Regression coefficient of Y on X:

$b_{yx} = r\frac{a_y}{a_s} = 0.72 * \frac{8.5}{12.5} = 0.4896$

Regression Equation of Y on X:

$(Y - \dot{Y}) = b_{yx} (X - \dot{X})$

$Y - 88 = 0.4896 (X - 42)$

$Y = 88 - 20.5632 + 0.4896X$

$Y = 67.4368 + 0.4896X$

Estimation of the production when the capacity utilisation is 70% is regression equation X on Y, where Y = 70

Regression Equation of X on Y:

$(X - \dot{X}) = b_{xy} (Y - \dot{Y})$

$X = -51.1744 + 1.0588Y$

$= -51.1744 + (1.0588 * 70)$

$= -51.1744 + 74.116$

$= \mathbf{22.9416}$

Therefore, the estimated production would be **22,94,160** units when there is a capacity utilisation of 70%.

Illustration 05:

The following data gives the age and blood pressure (BP) of 10 sports persons.

Name	:	A	B	C	D	E	F	G	H	I	J
Age (X)	:	42	36	55	58	35	65	60	50	48	51
BP (Y)	:	98	93	110	85	105	108	82	102	118	99

i. Find regression equation of Y on X and X on Y (Use the method of deviation from arithmetic mean)
ii. Find the correlation coefficient (r) using the regression coefficients.
iii. Estimate the blood pressure of a sports person whose age is 45.

Solution:

Calculation of Regression Equation

Name	Age (X)	BP (Y)	x=X-$\dot{X}$ x=X-50	y=Y-$\dot{Y}$ y=Y-100	x^2	y^2	xy
A	42	98	-8	-2	64	4	16
B	36	93	-14	-7	196	49	98
C	55	110	5	10	25	100	50
D	58	85	8	-15	64	225	-120
E	35	105	-15	5	225	25	-75
F	65	108	15	8	225	64	120
G	60	82	10	-18	100	324	-180
H	50	102	0	2	0	4	0
I	48	118	-2	18	4	324	-36
J	51	99	1	-1	1	1	-1
	500 $\sum X$	**1,000** ΣY	**0** $\sum x$	**0** Σy	**904** $\sum x^2$	**1,120** $\sum y^2$	**-128** $\sum xy$

$\dot{X} = \frac{\Sigma X}{n} = \frac{500}{10} = 50$ $\quad \dot{Y} = \frac{\Sigma Y}{n} = \frac{1000}{10} = 100$

Regression coefficients can be computed using the following formula:

$b_{xy} = \frac{\Sigma sy}{\Sigma y^2}$ $\quad b_{yx} = \frac{\Sigma sy}{\Sigma s^2}$ where $x = X - \dot{X}$ and $y = Y - \dot{Y}$

Regression coefficient of X on Y:

$$b_{xy} = \frac{\Sigma sy}{\Sigma y^2} = \frac{-128}{1120} = -0.1143$$

Regression coefficient of Y on X:

$$b_{yx} = \frac{\Sigma sy}{\Sigma s^2} = \frac{-128}{904} = -0.1416$$

Regression equation of X on Y:

$$(X - \dot{X}) = b_{xy}\,(Y - \dot{Y})$$

- $X - 50 = -0.1143\,(Y - 100)$
- $X - 50 = -0.1143Y + 11.43$
- $X = 50 + 11.43 - 0.1143Y$
- **$X = 61.43 - 0.1143Y$**

Regression equation of Y on X:

$$(Y - \dot{Y}) = b_{yx}\,(X - \dot{X})$$

- $Y - 100 = -0.1416\,(X - 50)$
- $Y - 100 = -0.1416X + 7.08$
- $Y = 100 + 7.08 - 0.1416X$
- **$Y = 107.08 - 0.1416X$**

Computation of coefficient of correlation using regression coefficient:

$$r = -\sqrt{b_{xy} * b_{yx}} = -\sqrt{0.1143 * 0.1416} = -\sqrt{0.01618488} = -0.1272$$

Therefore, we have low degree of negative correlation between age and blood pressure of sports person.

Estimation of the blood pressure (Y) of a sports person whose age is X=45 can be calculated using regression equation Y on X:

Regression equation of Y on X:

$$(Y - \dot{Y}) = b_{yx}\,(X - \dot{X})$$

- $Y = 107.08 - 0.1416X = 107.08 - (0.1416 * 45) = 107.08 - 6.372 = \underline{\underline{\mathbf{100.708}}}$

It means estimated blood pressure of a sports person is 101 (rounded off) whose age is 45.

Illustration 06:

There are two series of index numbers, ***P*** for price index and ***S*** for stock of commodity. The mean and standard deviation of *P* are 100 and 8 and *S* are 103 and 4 respectively. The correlation coefficient between the two series is 0.4. With these data, work out a linear equation to read off values of *P* for various values of *S*. Can the same equation be used to read off values of *S* for various values of *P*?

Solution:

Let us assume that *P*=Price Index be variable X an *S*=Stock of Commodity be variable Y. Linear equation to read off values of P for various values of S would be regression equation of X on Y. Regression coefficient is to be computed using mean and standard deviation.

From the problem we can list out the given information:

$\dot{X} = 100$ $\quad \dot{Y} = 103$ $\quad \sigma_x = 8$ $\quad \sigma_y = 4$ $\quad r = 0.4$

Regression equation of X on Y:

$$(X - \dot{X}) = b_{xy}\,(Y - \dot{Y})$$

- $(X - \dot{X}) = \frac{r\sigma_x}{\sigma_y}(Y - \dot{Y})$

- $(X - 100) = (0.4 * 8)(Y - 103)$
- $(X - 100) = 0.8\,(Y - 103)$
- $(X - 100) = 0.8Y - 82.4$
- $X = 100 - 82.4 + 0.8Y$
- **$X = 17.6 + 0.8Y$**

Linear equation to read off values of *P* for various values of *S* is X = 17.6 + 0.8Y
To read off values of S for various values of P we need regression equation of Y on X and therefore above linear equation cannot be used. Hence, the following regression equation of Y on X be computed:

$(Y - \dot{Y}) = b_{yx}\,(X - \dot{X})$

- $(Y - \dot{Y}) = \dfrac{r\sigma_y}{\sigma_x}(X - \dot{X})$
- $(Y - 103) = 0.4\,(X - 100)$ as $\dfrac{4}{*} = \dfrac{8}{}$
- $(Y - 103) = 0.2\,(X - 100)$
- $Y - 103 = 0.2X - 20$
- $Y = 103 - 20 + 0.2X$
- **$Y = 83 + 0.2X$**

Hence, the linear equation to read off values of S for various values of *P* is Y = 83 + 0.2X

Review of Correlation and Regression Analysis:

In correlation analysis, when we are keen to know whether two variables under study are associated or correlated and if correlated what is the strength of correlation. The best measure of correlation is proved by Karl Pearson's Coefficient of Correlation. However, one severe limitation of this method is that it is applicable only in case of a linear relationship between two variables. If two variables say X and Y are independent or not correlated then the result of correlation coefficient is zero.

Correlation coefficient measuring a linear relationship between the two variables indicates the amount of variation one variable accounted for by the other variable. A better measure for this purpose is provided by the square of the correlation coefficient, known as "coefficient of determination". This can be interpreted as the ratio between the explained variance to total variance:

$r^2 = \dfrac{\text{Explained variance}}{\text{Total Variance}}$ Similarly, Coefficient of non-determination = $(1 - r^2)$.

Regression analysis is concerned with establishing a functional relationship between two variables and using this relationship for making future projection. This can be applied, unlike correlation for any type of relationship linear as well as curvilinear. The two lines of regression coincide i.e. become identical when r= -1 or +1

in other words, there is a perfect negative or positive correlation between the two variables under discussion if r = 0, then regression lines are perpendicular to each other.

* * * * *