Unit 3 Correlation and Regression

Introduction

- Correlation and regression are statistical methods that are commonly used to compare two or more variables. For example, Comparison between income and expenditure, price and demand, etc.
- Correlation measures the association between two or more variables and quantitates the strength of their relationship. It evaluates only the existing data.
- Regression means average relationship between two or more variables and this relationship is used to estimate the most likely values of one variable for specified values of the other variables.

Correlation

- It is exist between two variables
- It is use to represent linear relationship between two variables.

- Two variables are said to be correlation if a change in a one variable affects a change in a other variable .
- Such data connecting two variables is called bivariate data.

- Thus ,Correlation is a statistical analysis which measures and analysis the degree to which two variables fluctuate with reference to other.
- Examples :
- Relationship between heights and weights
- Relationship between price and demand of commodity.
- Relationship between rainfall and yields of crops

TYPES OF CORRELATIONS

Correlation is classified into four types:

I. Positive and negative correlations

2. Linear and nonlinear correlations

3. Partial and total correlations

4. Simple and multiple correlations

TYPES OF CORRELATIONS

POSITIVE CORRELATION

- If both the variables vary in the same direction, the correlation is said to be positive.
- In other words, if the value of one variable increases, the value of the other variable also increases, or, if value of one variable decreases, the value of the other variable decreases,
- e.g., the correlation between heights(cm) and weights(kg) of group of persons is a positive correlation

Height	150	157	163	170	178
Weight	58	62	68	73	80

NEGATIVE CORRELATION

- If both the variables vary in the opposite direction, correlation is said to be negative.
- In other words, if the value of one variable increases, the value of the other variable decreases, or, if the value of one variable decreases the value of the other variable increases,
- e.g. the correlation between the price($ per unit) and demand(unit) of a commodity is a negative correlation.

Price	10	8	6	5	4
Demand	100	200	300	400	500

TYPES OF CORRELATIONS

LINEAR CORRELATION

If the ratio of change between two variables is constant, the correlation is said to be linear. If such variables are plotted on a graph paper, a straight line is obtained, e.g.,

Milk(litter)	5	10	15	20	25
Paneer (kg)	2	4	6	8	10

NONLINEAR CORRELATION

If the ratio of change between two variables is not constant, the correlation is said to nonlinear. The graph of a nonlinear or curvilinear relationship will be a curve, e.g.,

Expenses	3	6	9	12	15
Sales	8	12	15	15	16

TYPES OF CORRELATIONS

Simple Correlation

When only two variables are studied, the relationship is described as simple correlation. e.g., the quantity of money and price level, demand and price, etc.

Multiple Correlation

When more than two variables are studied, the relationship is described as multiple correlation, e.g., relationship of price, demand, and supply of a commodity

Partial Correlation

When more than two variables are studied excluding some other variables, the relationship is termed as partial correlation.

Total Correlation

When more than two variables are studied without excluding any variables, the relationship is termed as total correlation.

METHOD OF STUDYING CORRELATION

There are two different methods of studying correlation,

(1) Graphical methods

(2) Mathematical methods.

Graphical methods are

(a) Scatter diagram

(b) Simple graph

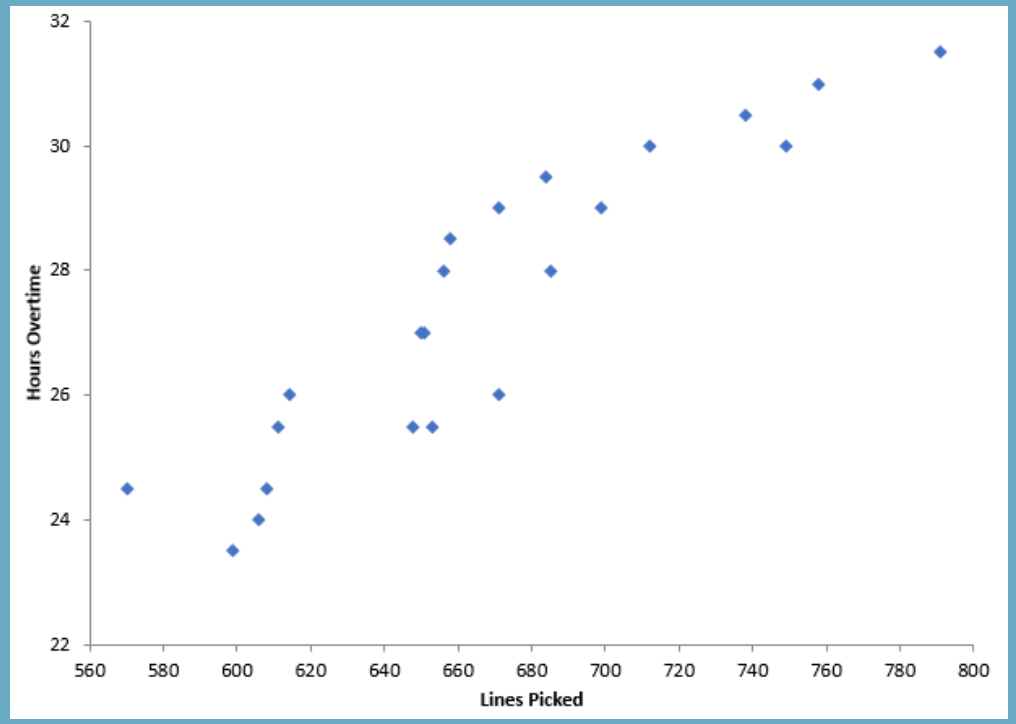
Mathematical methods are

(a) Karl Pearson’s coefficient

(b) Sperman’s rank coefficient of correlation

SCATTER DIAGRAM

- A scatter diagram is a graphical representation of the relation between two or more variables. In the scatter plot of two variables x and y, each point on the plot is an x-y pair.

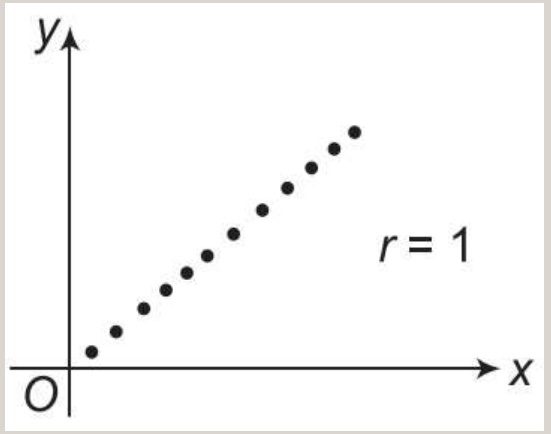


- There are various correlationships between two variables represented by the following scatter diagrams.

SCATTER DIAGRAM

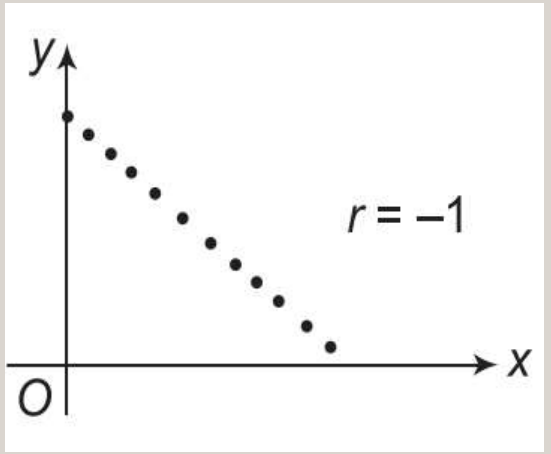
Perfect Positive Correlation

If all the plotted points lie on a straight line rising from the lower left-hand corner to the upper right-hand corner, the correlation is said to be perfectly positive.



Perfect Negative Correlation

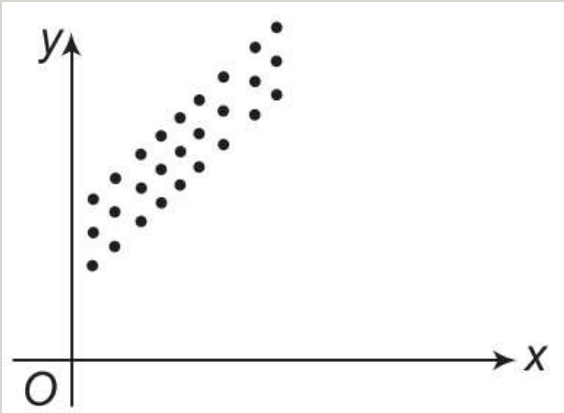
If all the plotted points lie on a straight line falling from the upper-left hand corner to the lower right-hand corner, the correlation is said to be perfectly negative.



SCATTER DIAGRAM

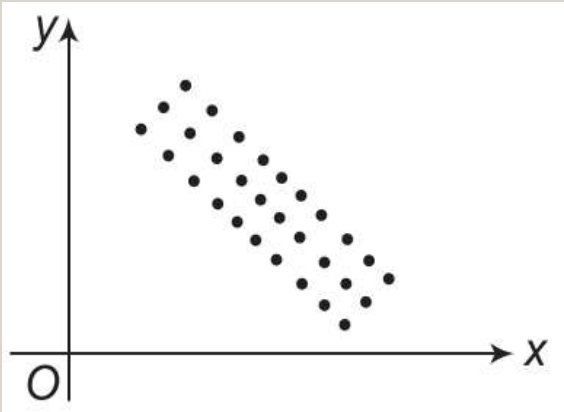
High Degree of Positive Correlation

If all the plotted points lie in the narrow strip, rising from the lower left-hand corner to the upper right-hand corner, it indicates a high degree of positive correlation.



High Degree of Negative Correlation

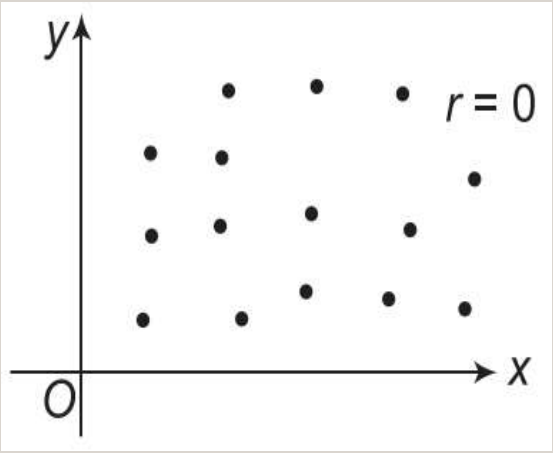
If all the plotted points lie in a narrow strip, falling from the upper left-hand corner to the lower right-hand corner, it indicates the existence of a high degree of negative correlation.



SCATTER DIAGRAM

No Correlation

If all the plotted points lie on a straight line parallel to the x-axis or y-axis or in a haphazard manner, it indicates the absence of any relationship between the variables.



SIMPLE GRAPH

- A simple graph is a diagrammatic representation of bivariate data to find the correlation between two variables.
- The values of the two variables are plotted on a graph paper. Two curves are obtained, one for the variable x and the other for the variable y.
- If both the curves move in the same direction, the correlation is said to be positive. If both the curves move in the opposite direction, the correlation is said to be negative.
- This method is used in the case of a time series. It does not reveal the extent to which the variables are related.

- Thus a scatter diagram is simple and nonmathematical method to find out the correlation between the variables .
- It gives an indication of the degree of linear correlation between the variables .
- It is easy to understand.

Mathematical method :

- Karl Pearson's coefficient of correlation

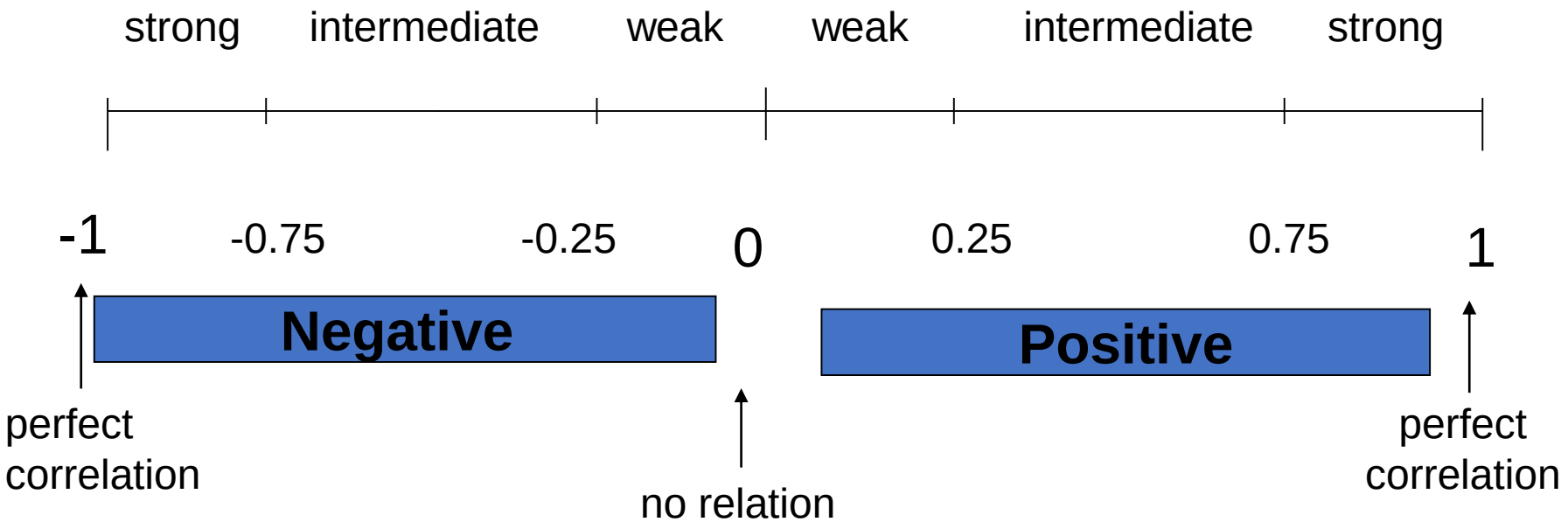
The coefficient of correlation is the measure of correlation between two random variables x and y , it denoted by r

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

➢ Note:

- The value of r lie between $-1 \leq r \leq 1$
- The correlation coefficient either positive or negative.
- The sign of correlation coefficient indicates the sign of linear relation ship.
- The magnitude (value) of correlation indicates the strength of correlation.

- The value of r ranges between (-1) and (+1)
- The value of r denotes the strength of the association as illustrated by the following diagram.



- If r = Zero this means no association or correlation between the two variables.
- If $0 < r < 0.25$ = weak correlation.
- If $0.25 \leq r < 0.75$ = intermediate correlation.
- If $0.75 \leq r < 1$ = strong correlation.
- If r = 1 = perfect correlation.

- For ex. Correlation r = 0.9 suggests a strong positive linear relation ship between two variables and if r = -0.2 suggest a weak negative correlation between two variables
- If two random variables are independent then r = 0

Example 1:

The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

a.) Display the scatter plot.

b.) Calculate the correlation coefficient *r*.

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50

Hours, x	0	1	2	3	3	5	5	5	6	7	7	10
Test score, y	96	85	82	74	95	68	76	84	58	65	75	50
xy												
x^2												
y^2												

Ex-2 Calculate the Karl Pearson's coefficient of correlation between x and y using the following data .

X	2	4	5	6	8	11
Y	18	12	10	8	7	5

X	Y	X^2	Y^2	XY
2	18			
4	12			
5	10			
6	8			
8	7			
11	5			
36	60			

Ex-3 Calculate the coefficient of correlation from the following data

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	12	3

X	Y	X^2	Y^2	XY
12	14			
9	8			
8	6			
10	9			
11	11			
13	12			
7	3			
70	63			

Ex-4 Calculate the coefficient of correlation following data

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

X	Y	X^2	Y^2	XY
9	15			
8	16			
7	14			
6	13			
5	11			
4	12			
3	10			
2	8			
1	9			
45	108			

Ex-5 Calculate the correlation coefficient between the following data

X	5	9	13	17	21
Y	12	20	25	33	35

X	Y	X^2	Y^2	XY

Rank correlation

- Let a group of n individuals be arranged in order to merit with respect to some characteristics . The same group would give a different order (rank) for different characteristics . Considering the orders corresponding to two characteristics x and y ,the correlation between these n pairs of ranks is called the rank correlation in the characteristics x and y for that group of individuals

- In statistics a **rank correlation** is any of several statistics that measure an **ordinal association**—the relationship between rankings of different ordinal variables or different rankings of the same variable, where a "ranking" is the assignment of the ordering labels "first", "second", "third", etc. to different observations of a particular variable.
- A **rank correlation coefficient** measures the degree of similarity between two rankings, and can be used to assess the significance of the relation between them

- When we are given the actual data and not a rank ,it will be necessary to assign a rank.
- Rank can be assigned by taking either the highest value as 1 or the lowest value as 1

Spearman’s Rank Correlation Coefficient

- Spearman's correlation coefficient, measures the strength and direction of association between two ranked variables
- The Spearman’ s Rank coefficient of two different characteristic x and y is denoted by r_s
- d= x-y

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Step in calculating Spearman's rank coefficient

- Convert the observes value to rank
- Find the difference between the rank ,square them and sum of the squared difference
- Write the formula and solve it and conclude based on finding value of r
- The rank correlation lie in between [-1 ,1]

- If r = +1 indicates a perfect positive association of ranks
- r = -1 indicates a perfect negative association of ranks
- r = 0 indicates a no association between the ranks
- If r is closer to zero , then it is weaker association between the ranks

Ex 6: The following table provides data about the percentage of students who have free university meals and their CGPA scores. Calculate the Spearman's Rank correlation between the two and interpret result

State University	% of students having free meals	% of student scoring above 8.5 CGPA
Pune	14.4	54
Chennai	7.2	64
Delhi	27.5	44
Kanpur	33.8	32
Ahmedabad	38	37
Indore	15.9	68
Guwahati	4.9	62

Let x =rank of students having free meals

y= rank of students scoring above 8.5 CGPA

State university	X	Y	d= x-y	d^2
Pune				
Chennai				
Delhi				
Kanpur				
Ahmedabad				
Indore				
Guwahati				

Ex-7 The ICC ranking for ODI and test matches for nine team as shown bellows
Check whether there is correlation between ranks

Team	test ranking	ODI ranking
India	1	1
Australia	2	3
South Africa	3	2
Srilanka	4	7
Pakistan	5	6
England	6	4
Newzealand	7	5
bangladesh	8	8
Westindies	9	9

Team	test ranking	ODI ranking	d	d^2
India				
Australia				
South Africa				
Srilanka				
Pakistan				
England				
Newzealad				
bangladesh				
Westindies				

Ex-8 Ten participants in a contest are ranked by two judges are as follows ,Calculate the rank correlation coefficient

X	1	3	7	5	4	6	2	10	9	8
Y	3	1	4	5	6	9	7	8	10	2

x	y	d=x-y	d^2

Ex-9 Ten students got the following percentage of marks in mathematics and physics ,Find rank correlation coefficient .

Mathematics (x)	8	36	98	25	75	82	92	62	5	35
Physics(y)	84	51	91	60	68	62	86	58	35	49

Ex-10 Ten competitors in a musical test were ranked by three judges a, b and c in the following order. Using the rank correlations method ,find which pair of judges has nearest approach to common liking in music.

Rank by a	1	6	5	10	3	2	4	9	7	8
Rank by b	3	5	8	4	7	10	2	1	6	9
Rank by c	6	4	9	8	1	2	3	10	5	7

Rank by a x	Rank by b y	Rank by c z	d_1=x-y	d_2=y-z	d_3=z-x	d_1^2	d_2^2	d_3^2

Example 11: The competitors in a beauty contest are ranked by three judges in the following order.
Use rank correlation coefficient to discuss which pair of judges has nearest approach to beauty.

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

REGRESSION

- Regression is defined as a method of estimating the value of one variable when that of the other is known and the variables are correlated.
- Regression analysis is used to predict or estimate one variable in terms of the other variable.
- It is a highly valuable tool for prediction purpose in economics and business.
- It is useful in statistical estimation of demand curves, supply curves, production function, cost function, consumption function, etc.

TYPES OF REGRESSION

Regression is classified into two types

1. Simple and multiple regressions
2. Linear and nonlinear regressions

TYPES OF REGRESSION

Simple and Multiple Regressions

Depending upon the study of the number of variables., regression may be simple or multiple.

- **Simple Regression**

The regression analysis for studying only two variables at a time is known as simple regression.

- **Multiple Regression**

The regression analysis for studying more than two variables at a time is known as multiple regression.

Linear and Nonlinear Regressions

Depending upon the regression curve, regression may be linear or nonlinear.

- **Linear Regression**

If the regression curve is a straight line, the regression is said to be linear.

- **Nonlinear Regression**

If the regression curve is not a straight line i.e., not a first-degree equation in the variables x and y, the regression is said to be nonlinear or curvilinear

LINES OF REGRESSION

Line of Regression of y on x

- It is the line which gives the best estimate for the values of y for any given values of x.
- The regression equation of y on x is given by

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

- It is also written as $y = a + bx$

Line of Regression of x on y

- It is the line which gives the best estimate for the values of x for any given values of y.
- The regression equation for x on y is given by

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

- It is also written as $x = a + by$

Note: $\bar{x}$ and $\bar{y}$ are means of x series and y series respectively. σ_x and σ_y are standard deviations of x series and y series respectively, r is the correlation coefficient between x and y.

REGRESSION COEFFICIENTS

- The slope b of the line of regression of y on x is also called the coefficient of regression of y on x.
- It represents the increment in the value of y corresponding to a unit change in the value of x.

b_{yx}= Regression coefficient of y on x

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

Similarly, b_{xy}= Regression coefficient of x on y

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

EXPRESSION FOR REGRESSION COEFFICIENT

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

$$\sigma_x = \sqrt{\sum x^2 - \frac{(\sum x)^2}{n}} \qquad \sigma_y = \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}$$

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

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

PROPERTIES OF REGRESSION COEFFICIENTS

1. The coefficient of correlation is the geometric mean of the coefficients of regression, i.e., $r = \sqrt{b_{yx} b_{xy}}$

2. The arithmetic mean of regression coefficients is greater than or equal to the coefficient of correlation i.e. $\frac{1}{2}\left(b_{yx} + b_{xy}\right) \geq r$

3. Both regression coefficients will have the same sign i.e., either both are positive or both are negative.

4. The sign of correlation is same as that of the regression coefficients, i.e., $r > 0$ **if** $b_{xy} > 0$ **and** $b_{yx} > 0$**; and** $r < 0$ **if** $b_{xy} < 0$ **and** $b_{yx} < 0$**.**

PROPERTIES OF LINES OF REGRESSION

1. The two regression lines x on y and y on x always intersect at their means $(\overline{x}, \overline{y})$.

2. Since $r^2 = b_{yx}b_{xy}$, i.e., $r = \sqrt{b_{yx}b_{xy}}$, therefore r, b_{yx}, b_{xy} all have the same sign.

3. If $r = 0$, the regression coefficients are zero.

4. The regression lines become identical if $r = \pm 1$. It follows from the regression equation that $x = \overline{x}$ and $y = \overline{y}$. If $r = 0$, these lines are perpendicular to each other.

EXAMPLES

Example 13: The regression lines of a sample are $x + 6y = 6$ and $3x + 2y = 10$. Find
(1) sample means $\overline{x}$ and $\overline{y}$, and
(2) the coefficient of correlation between x and y.
(3) Also estimate y when $x = 12$.

EXAMPLES

Example 14: The following data regarding the height (y) and weight (x) of 100 college students are given:

$$\sum x = 15000, \sum x^2 = 2272500, \sum y = 6800, \sum y^2 = 463025, \sum xy = 1022250$$

Find the coefficient of correlation between height and weight and also the equation of regression of height and weight.

EXAMPLES

Example 15: Find the regression coefficients b_{yx} and b_{xy} and hence, find the correlation coefficient between x and y for the following data:

x	4	2	3	4	2
y	2	3	2	4	4

x	y	x^2	y^2	xy

EXAMPLES

Example-16: The number of bacterial cells (y) per unit volume in a culture at different hours (x) is given below:

x	0	1	2	3	4	5	6	7	8	9
y	43	46	82	98	123	167	199	213	245	272

Fit lines of regression of y on x and x on y. Also, estimate the number of bacterial cells after 15 hours.

Solution:

EXAMPLES

Example 17: The regression lines of a sample are $4x - 5y + 30 = 0$ and $20x - 9y + 107 = 0$. Find

(1) Find the both regression coefficient

(2) Find r and σ_y when $\sigma_x = 3$