18

CORRELATION AND REGRESSION

Correlation analysis is a statistical procedure by which we determine the degree of association or relationship between two or more variables. But correlation does not predict anything about the cause and effect relationship. Even a high degree of correlation does not necessarily imply that a cause and effect relationship exists between the two variables. The study of correlation becomes important only if it is studied with regression analysis. The combined study of both correlation and regression analysis is specially interesting while studying problems in social science, educational research, policy making and arriving at decisions etc.

The correlation coefficient is denoted by r. Karl Pearson (1967 – 1936), a British Biometrician, developed the formula for Correlation Coefficient. The correlation coefficient between two variables X and Y are denoted by r(X,Y) or $r_{x,y.}$

Let $x_1, x_2, \ldots, x_n$ be a set of observations of the variate X and let $y_1, y_2, \ldots, y_n$ be the corresponding values of Y. Then we have

$$r_{xy} = \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2}} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

Where, Cov(X,Y) is called as the covariance between X and Y with σ_x and σ_y be the standard deviations of X and Y respectively.

Note

The formula of correlation coefficient can be expressed in terms of mathematical expectation as well. This is specially used when the correlation is to be found out between two random variables.

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

In terms of expectation $Cov(X,Y) = E(XY) - E(X)E(Y)$,
$Var(X) = E(X^2) - [E(X)]^2$ and $Var(Y) = E(Y^2) - [E(Y)]^2$

A Simplified Formula for Correlation Coefficient

We know that correlation coefficient is given by

$$r_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} \quad (1)$$

$$\text{Now, } \sum (x_i - \bar{x})^2 = \sum (x_i^2 + \bar{x} - 2x_i\bar{x}) = \sum x_i^2 + \sum \bar{x} - 2\bar{x}\sum x_i$$

$$= \sum x_i^2 + n\bar{x}^2 - 2\bar{x}n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2 \quad (2)$$

Similarly,

$$\sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 \quad (3)$$

and

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x}\,\bar{y})$$

$$= \sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + \sum \bar{x}\bar{y} = \sum x_i y_i - \bar{x} n\bar{y} - \bar{y} n\bar{x} + n\bar{x}\,\bar{y} \quad (4)$$

Replacing the values of (2), (3) and (4) in (1), we have

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

This formula is extensively used for calculation purpose.

Theorem: The correlation coefficient between X and Y lies between +1 and –1.

Proof: Let X and Y be two variates where X takes the values $x_1, x_2, \ldots, x_n$ and let Y takes the values $y_1, y_2, \ldots, y_n$. Also, let σ_x and σ_y be the standard deviations of X and Y respectively.
Now, we transform the variates x_i and y_i to u_i and v_i. Now, by the transformation we have:

$$u_i = \frac{x_i - \bar{x}}{\sigma_x} \text{ and } v_i = \frac{y_i - \bar{y}}{\sigma_y}$$

Now, $\sum u_i^2 = \sum \frac{(x_i - \bar{x})^2}{\sigma_x^2} = \frac{n\sigma_x^2}{\sigma_x^2} = n$

Similarly,

$$\sum v_i^2 = \sum \frac{(y_i - \bar{y})^2}{\sigma_y^2} = \frac{n\sigma_y^2}{\sigma_y^2} = n$$

Now,

$$\sum u_i v_i = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} = n \times \frac{1}{n} \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} = n \frac{Cov(X,Y)}{\sigma_x \sigma_y} = n \times r_{xy}$$

Now, $(u_i + v_i)^2 \geq 0$ because it is a perfect square. Hence, the sum of squares cannot be negative.
i.e. $\sum (u_i + v_i)^2 \geq 0$
$\Rightarrow \sum u_i^2 + \sum v_i^2 + 2\sum u_i v_i \geq 0$
$\Rightarrow n + n + 2n\, r_{xy} \geq 0$
$\Rightarrow 2n + 2n\, r_{xy} \geq 0$
$\Rightarrow 2n\,(1 + r_{xy}) \geq 0$
$\Rightarrow (1 + r_{xy}) \geq 0$ [Since, $n > 0$
$\Rightarrow r_{xy} \geq -1$ (1)
Simillarly,
i.e. $\sum (u_i - v_i)^2 \geq 0$
$\Rightarrow \sum u_i^2 + \sum v_i^2 - 2\sum u_i v_i \geq 0$
$\Rightarrow n + n - 2n\, r_{xy} \geq 0$

$\Rightarrow 2n - 2n r_{xy} \geq 0$
$\Rightarrow 2n(1 - r_{xy}) \geq 0$
$\Rightarrow (1 - r_{xy}) \geq 0$ [Since, $n > 0$
$\Rightarrow r_{xy} \leq 1$ (2)

Combining the results of (1) and (2) we have,

$-1 \leq r_{xy} \leq 1$

This proves that the correlation coefficient always lies between –1 and +1.

Uses of Correlation

1. The correlation analysis helps us to measure the degree of relationship that exists between the variables.
2. In business, forecasting is an important phenomenon and correlation helps us to make relatively more dependable forecast.
3. Correlation is used to determine the regression coefficient if standard deviation of two variables is known.
4. Correlation is used in problems of reliability and validity of tests.

Effect of Change of Origin and Scale on Correlation

Let $x_1, x_2, \ldots, x_n$ be a set of observations. Let us change the origin of x to a and scale by h, where a and h are arbitrary numbers (and h > 0). Then we have-

$$u_i = \frac{x_i - a}{h}$$
$$\Rightarrow x_i = a + h \times u_i \quad (1)$$

Multiplying and dividing both sides by Σf_i we have

$$\frac{\sum f_i x_i}{\sum f_i} = a \times \frac{\sum f_i}{\sum f_i} + h \frac{\sum f_i u_i}{\sum f_i}$$
$$\Rightarrow \bar{x} = a + h \times \bar{u} \quad (2)$$

Equation (1) – (2) gives

$$x_i - \bar{x} = h \times (u_i - \bar{u}) \quad (3)$$

Similarly, $y_1, y_2, \ldots, y_n$ be a set of observations. Let us change the origin of y to b and scale by k, where b and k are arbitrary numbers where h is positive. Then we have

$$v_i = \frac{y_i - b}{k}$$
$$\Rightarrow y_i = b + k \times v_i \quad (4)$$

Multiplying and dividing both sides by Σf_i we have

$$\frac{\sum f_i y_i}{\sum f_i} = b \times \frac{\sum f_i}{\sum f_i} + k \frac{\sum f_i v_i}{\sum f_i}$$
$$\Rightarrow \bar{y} = b + k \times \bar{v} \quad (5)$$

Equation (4) – (5) gives

$$y_i - \bar{y} = k \times (v_i - \bar{v}) \quad (6)$$

Thus, we have,

$$r_{xy} = \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2 \frac{1}{n}\sum (y_i - \bar{y})^2}} = \frac{\frac{1}{n}\sum h \times (u_i - \bar{u})k \times (v_i - \bar{v})}{\sqrt{\frac{1}{n}\sum h^2 \times (u_i - \bar{u})^2 \frac{1}{n}\sum k^2 \times (v_i - \bar{v})^2}}$$

$$= \frac{hk \times \frac{1}{n}\sum (u_i - \bar{u})(v_i - \bar{v})}{hk\sqrt{\frac{1}{n}\sum (u_i - \bar{u})^2 \frac{1}{n}\sum (v_i - \bar{v})^2}} = r_{uv}$$

Thus, we have $r_{xy} = r_{uv}$ so, correlation coefficient is independent of the effect of change of origin and scale. However, both the scales should be of same sign.

Relation between Correlation and Independence

Theorem: If two variates are independent then they are uncorrelated but not vice versa.

Proof: Let us consider two independent variates X and Y . Since the variates are independent so from the theorem of expectation we have,

$$E(XY) = E(X)\,E(Y)$$

Now we know that,

$$Corr(X,Y) = \frac{Cov\,(X,Y)}{S.D(X)\,S.D(Y)} = \frac{E(XY) - E(X)\,E(Y)}{S.D(X)\,S.D(Y)}$$

$$= \frac{E(X)E(Y) - E(X)\,E(Y)}{S.D(X)\,S.D(Y)} = 0$$

Thus, if two variates are independent then they are uncorrelated.

Now, let us consider the relation $y = x^2$. Let X takes the values *–3, -2, -1, 0, 1, 2, 3* thus using the above relation the corresponding values of y are *9, 4, 1, 0, 1,4, 9.* Thus we have

x	**$y = x^2$**	**xy**	**x^2**	**y^2**
-3	9	-27	9	81
-2	4	-8	4	16
-1	1	-1	1	1
0	0	0	0	0
1	1	1	1	1
2	4	8	4	16
3	9	27	9	81
$\Sigma x = 0$	$\Sigma y = 28$	$\Sigma xy = 0$	$\Sigma x^2 = 28$	$\Sigma y^2 = 196$

$$\bar{x} = \frac{\sum x}{n} = \frac{0}{n} = 0$$

$$r_{xy} = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}} = \frac{0 - n \times 0 \times \bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2}\sqrt{\sum y^2 - n\bar{y}^2}} = 0$$

Thus,

Thus, we see that the two variables are uncorrelated though practically they are associated by the relation $y = x^2$.

So, independence implies that the variables are uncorrelated but not vice versa.

Properties of Correlation Coefficient

1. It is rigidly defined.
2. It is based on all the observations.
3. The correlation coefficient is a pure number and has no unit of measurement.
4. It is independent of the change of origin and scale.
5. It lies between – 1 and + 1.
6. The sign of correlation coefficient depends on the sign of covariance.
7. If the two variables are independent, the correlation coefficient between them is zero but the converse is not true.

Interpretation of Correlation Coefficient

1. r = +1 indicates Perfect Positive Correlation, i.e, there is an equal proportional change in both the variables and in the same direction.
2. r = – 1 indicates Perfect Negative Correlation i.e. there is an equal proportional change in both the variables and in the opposite direction.
3. r = 0 implies that the variables are uncorrelated.
4. A value of r very near to 0 means very little correlation between X and Y i.e. X and Y are practically independent variates.
5. A value of r near to + 1 or –1 means Y is highly dependent on X or X is highly dependent on Y.

Note

If two variables are independent then the correlation coefficient between the two variables is equal to zero. But correlation coefficient equal to zero does not imply that the two variables are independent. At best we can say that there is no linear relation ship between the two variates where other relations like quadratic, exponential may exist.

Limitations of Correlation Coefficient

1. Correlation coefficient is a measure of the extent of linear relationship between two variables. However, a small value of r indicates only a poor linear relationship between the variables. This does not rule out the possibility that the variables may be related in some other way.
2. A high value of *r* does not always imply that there is a direct cause- and-effect relationship between the variables. The high correlation may be due to the fact that both the variables depend on a third variable.
3. Sometimes, it may happen that two series of observations show a high correlation coefficient even though there may not be any logical basis for relationship between them. Such correlation between two variables is known as '*Spurious correlation'* or '*Non-sense correlation'*. It is advisable that before computing correlation coefficient one should satisfy oneself about the logical relationship between them.
4. If the data are not reasonably homogeneous, the coefficient of correlation may give a misleading picture of the extent of association. If some reasonable basis can be found for separating the data into groups, only then it is desirable to compute values of *r* for each

group.

5. It is sometimes difficult to interpret the significance of correlation coefficient and often it is misinterpreted.

Types of Correlation

1. Positive or Negative

Correlation is said to be positive when the variables move in the same direction. It means that when the value of one variable increases, the value of the other variable will also increase and vice versa. If the variables move in opposite directions, correlation is negative, *i.e.* an increase in the value of one variable is associated with the decrease in the value of the other and vice versa.

Examples of positive correlation :

(i) Price and supply of a commodity.

(ii) Advertisement expenses and sales of a product.

Examples of negative correlation :

(i) Price and demand of a commodity.

(ii) Interest rate on loans and loans taken by the public.

2. Simple, Partial and Multiple Correlation

The simple correlation studies the relationship between two variables. When relationship is measured between three or more variables then it is a case of multiple or partial correlation. Multiple correlation studies the extent in which a variable is effected by the combined influence of a group of other variables. The study of relationship between two variables, keeping the effects of other variables constant is known as partial correlation.

Example of simple correlation :

(i) Price and supply of a commodity.

(ii) Price and demand of a commodity.

Example of multiple correlation :

(i) Yield of crop with the combined effect of rainfall, temperature, use of fertilizer, humidity etc.

(ii) Demand of a commodity with the combined effect of its price, income of the consumer, price of substitute products etc.

Example of partial correlation :

(i) Yield of crop and rainfall keeping the effect of temperature, use of fertilizer, humidity etc. as constant.

(ii) Demand of a commodity and its price keeping the effect of income of the consumer, price of substitute products etc as constant.

3. Linear and Non-linear Correlation

If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. If the amount of change in one variable does not bear a constant ratio to the amount of' change in the other variable, then the correlation is said to be non-linear or curvilinear.

Example of linear correlation:

(i) Amount of rainfall and reading in the rain gauge.

(ii) Amount of deposits in a savings bank account and the interest income of the depositor

Example of non–linear correlation:

(i) Amount of rainfall and yield of crop.
(ii) Income and expenditure of a person.

4. Spurious or Non-sense Correlation

It is sometimes seen that though there is no cause and effect relationship between two variables, but still the value of the correlation coefficient calculated on numerical basis may prove to be significant. Such type of correlation is called as spurious correlation or non-sense correlation. Examples of spurious correlation:

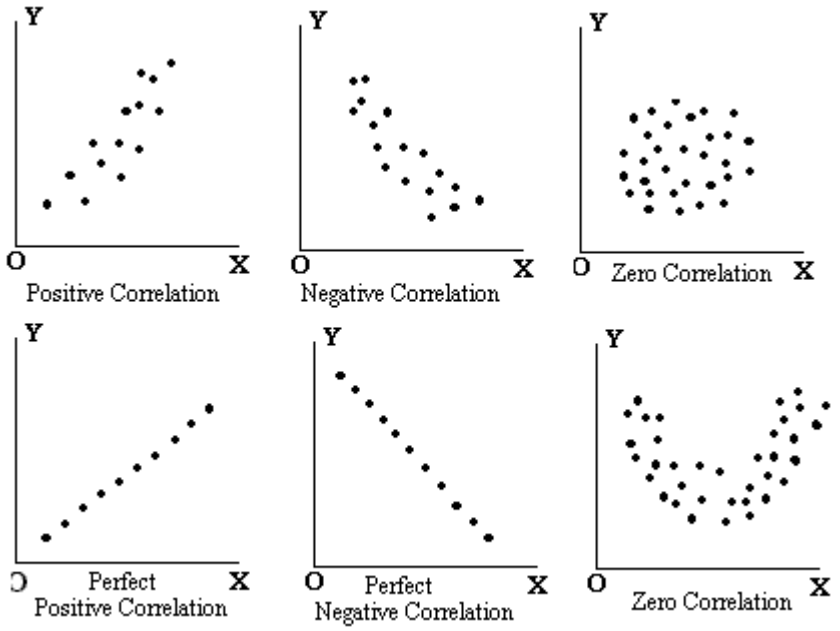
(i) The correlation between GDP of India and rainfall in Japan for the last 10 years.
(ii) Number of deaths in India due to snake bite and number of fresh graduates.

Scatter Diagram

The scatter diagram is the simplest method of studying relationship between two variables. The simplest device for ascertaining whether variables are related is to prepare a dot chart, where the horizontal axis represents one variable and vertical axis representing the other. The diagram so obtained is known as scatter diagram or dot diagram. From the scatter diagram one can have a fairly good idea about the relationship between variables. The following points may be borne in mind in interpretation of a scatter diagram regarding the correlation between two variables:

1. If the points are very close to each other, a fair amount of correlation may be expected between the two variables. On the other hand, if points are widely scattered, a poor correlation is expected between them.
2. If the points on the scatter diagram reveal any trend (upward or downward), the variables are supposed to be correlated.
3. If there is an upward trend rising from lower left hand corner to upper right hand corner, the correlation is positive. On the other hand, if there is a downward trend from upper left hand corner to lower right hand corner, the correlation is negative.
5. If all the points lie on a straight line starting from left bottom and going upwards to the right top, then the correlation is perfectly positive. On the other hand, if all the points lie on a straight line starting from left top and going downwards to the right bottom, the correlation is perfectly negative.

The figures below show the scatter diagram for various types of correlation:



Regression

Literally, the word 'regression' means to return or passing back to the original position. Sir Fransis Galton used the word 'regression' while he was dealing with inheritance of stature, in the later half of the Nineteenth century. Galton found a relation between height of the off springs with that of their parents who were either abnormally tall or short.

However, the word 'regression' is now used in a much wider sense. Regression analysis is now a powerful tool of statistics and is used for estimating or predicting the unknown value of a variable from the known value of another variable. More precisely, if X and Y are two related variables then regression analysis helps us to estimate the value of Y for a given value of X or vice versa.

According to Prof. Blair, "Regression is a mathematical measure of the average relationship between two or more variables in terms of original units of the data."

Regression analysis is now-a-days extensively used in natural, social and physical sciences. It is also used in business and economics to study the relationship between two or more variables for the estimation of demand and supply curves, cost functions, production and consumption functions.

Dependent and Independent Variables

The above definitions make it clear that regression analysis is used for estimating or predicting the unknown value of one variable from the known value of the other variable. The variable whose value is to be predicted is called as the dependent variable or 'explained variable'. On the other hand, the variable with the help of which prediction is done is called the independent variable or 'explanatory variable'. The independent variable is generally denoted by X and the dependent variable by Y.

Types of Regression

If regression analysis is confined to the study of two variables only, then it is known as simple regression. The regression analysis that studies more than two variables at the same time is called multiple regression.

In case of simple regression if the relation between the dependent and independent variable follows a straight line pattern then the regression is called as linear regression. However, if the relation is expressed in the form of a curve then the type of regression is called as curvilinear regression. Regression whether linear of curvilinear can be understood from the scatter diagram for the two variables. If the dots of a scatter diagram concentrate around a certain curve then the regression is curvilinear regression and the corresponding relation is called as the regression curve. If the points of a scatter diagram lie on a straight line then the line is called as the line of regression and the type of regression is linear regression.

Simple Linear Regression

If we take the case of two variables X and Y. In such a case we shall have two regression lines i.e. regression of X on Y and regression of Y on X. The regression line of Y on X gives the most probable values for Y for given values of X. Similarly, the regression line of X on Y gives the most probable value of X for given value of Y. However, when there is a perfect correlation between X and Y ($r = \pm 1$), then the regression lines will coincide i.e. we will have only one line. Other wise, the regression lines intersect each other at the point $(\bar{x}, \bar{y})$.

The regression equation of X on Y is given by,

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

Here, $\bar{x}$ and $\bar{y}$ are the arithmetic means of x and y respectively. In the equation, the coefficient of y, i.e., b_{xy} is called as the regression co-efficient of X on Y. The value of the regression coefficient $b_{xy} = r \times \sigma_x / \sigma_y$

Where, σ_x and σ_y are the standard deviations of X and Y respectively and r the correlation coefficient between X and Y.

Similarly, the regression equation of Y on X is given by,

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

In this equation the coefficient of X, i.e., b_{yx} is called as the regression co-efficient of Y on X. The value of the regression coefficient $b_{yx} = r \times \sigma_y / \sigma_x$

Note

1. Both the equations i.e. Y on X and X on Y are derived using the principle of least squares. It must be noted that the line of regression of Y on X should be used to predict Y for a given value of X and not vice versa because this equation will give the best estimate of Y for given X as defined by the principle of least squares.
2. Both the regression lines pass through the point $(\bar{x}, \bar{y})$ which, is eventually the solution of the lines of regression.

Relation Between Correlation and Regression Coefficient

We know that the simple regression coefficient for the line X on Y is given by $b_{xy} = r \times \sigma_x / \sigma_y$

Similarly, the simple regression coefficient for the line Y on X is given by

$b_{yx} = r \times \sigma_y / \sigma_x$

Now,

$$b_{xy} \times b_{yx} = r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x} = r^2$$

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

Thus, correlation coefficient is the geometric mean between the two regression coefficients.

Note

This formula can be used to calculate the correlation coefficient if the values of the regression coefficient are known. But one has to be careful in determining the sign of the correlation coefficient, because the calculated result will always be positive. Both the regression coefficients are either positive or negative. If the regression coefficients are positive then the correlation coefficient is positive and vice versa.

Effect of Change of Origin and Scale on Regression Coefficient

Let $x_1, x_2, \ldots, x_n$ be a set of observations and $y_1, y_2, \ldots, y_n$ be a set of observations. So, we have the regression coefficient of X on Y as,

$$b_{xy} = r \times \frac{\sigma_x}{\sigma_y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \times \frac{\sigma_x}{\sigma_y} = \frac{Cov(x,y)}{\sigma_y^2}$$

$$= \frac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum (y_i - \bar{y})^2}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2} \qquad (1)$$

Let us change the origin of x to a and scale by h, where a and h are arbitrary numbers $(h > 0)$. Then we have

$$u_i = \frac{x_i - a}{h}$$

$$\Rightarrow x_i = a + h \times u_i \qquad (2)$$

Multiplying and dividing both sides by Σf_i we have

$$\frac{\sum f_i x_i}{\sum f_i} = a \times \frac{\sum f_i}{\sum f_i} + h\frac{\sum f_i u_i}{\sum f_i}$$

$$\Rightarrow \bar{x} = a + h \times \bar{u} \qquad (3)$$

Equation (2) – (3) gives

$$x_i - \bar{x} = h \times (u_i - \bar{u}) \qquad (4)$$

Similarly, $y_1, y_2, \ldots, y_n$ be a set of observations. Let us change the origin of y to b and scale by k, where b and k are arbitrary numbers $(k > 0)$. Then we have

$$v_i = \frac{y_i - b}{k} \qquad (5)$$

$$\Rightarrow y_i = b + k \times v_i$$

Multiplying and dividing both sides by Σf_i we have

$$\frac{\sum fi\, yi}{\sum fi} = b \times \frac{\sum fi}{\sum fi} + k\,\frac{\sum fi\, vi}{\sum fi}$$

$$\Rightarrow \bar{y} = b + k \times \bar{v} \qquad (6)$$

Equation (5) – (6) gives

$$y_i - \bar{y} = k \times (v_i - \bar{v}) \qquad (7)$$

Replacing the values of (4) and (7) in (1) we have

$$b_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2} = \frac{\sum h \times (u_i - \bar{u}) k \times (v_i - \bar{v})}{\sum k^2 \times (v_i - \bar{v})^2}$$

$$= \frac{h\sum (u_i - \bar{u})(v_i - \bar{v})}{k\sum (v_i - \bar{v})^2} = \frac{h}{k} \times b_{uv}$$

Thus, the regression coefficient is independent of the change of origin but not of scale.

Properties of Regression Coefficient

1. The range of the regression coefficients is from $-\infty$ to $+\infty$.
2. The correlation coefficient between two variables is equal to the geometric mean of the regression coefficients.
   That is, $r = \sqrt{(b_{xy} \times b_{yx})}$
3. The signs of regression coefficients and correlation coefficient are always the same.
4. The arithmetic mean of the two regression coefficients is greater than or equal to the values of the correlation coefficient.
   That is, $(b_{xy} + b_{yx})/2 \geq r$
5. If one of the regression coefficient in a simple linear regression is greater than unity, the other will be less than unity.
   That is, if $b_{xy} > 1$ then $b_{yx} < 1$ and if $b_{yx} > 1$ then $b_{xy} < 1$.
6. If the variables X and *Y* are independent, then the regression coefficients are zero.
7. Both the regression coefficients are either positive or negative.

Difference between Correlation and Regression

Both correlation and regression analysis help us in studying relationship between two variables. However, they differ in their application and objectives:

1. The correlation coefficient measures the degree of association between the variables whereas the objective of regression analysis is to establish a mathematical relationship between the variables.
2. Correlation cannot be used for the purpose of prediction where as regression analysis is basically used for prediction purpose.
3. The coefficient of correlation varies between ± 1 while the regression coefficient ranges from - ∞ to + ∞.
4. In correlation analysis there is no concept of independent or dependent variable but in regression analysis one has to decide which variable is to be taken as the dependent and which one as the independent variable.

5. There may be nonsense correlation between two variables which is purely due to chance and has no practical relevance. However, there is nothing like nonsense regression.

5. Correlation is independent of change of scale and origin but regression coefficients are independent of change of origin but not of scale.

Formulae

1. $r = \dfrac{\frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2 \frac{1}{n}\sum (y_i - \bar{y})^2}} = \dfrac{\sum xy - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}}$ 2. $r_{xy} = r_{uv}$

3. $r = \dfrac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$

4. Regression line of X on Y $(x - \bar{x}) = b_{xy}(y - \bar{y})$ and X on Y is $(y - \bar{y}) = b_{yx}(x - \bar{x})$

5. $b_{xy} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})}$ 6. $r = \sqrt{b_{xy} \times b_{yx}}$ 7. $b_{xy} = \dfrac{h}{k}\, b_{uv}$

Solved Illustrations

Illustration 1 : Find the coefficient of correlation between x and y from the following data
$\Sigma x = 142, \Sigma y = 166, \Sigma xy = 2434, \Sigma x^2 = 2085, \Sigma y^2 = 2897, n = 10$

Solution : Correlation coefficient between two variables x and y is given by,

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

Here, $\bar{x} = \dfrac{\sum x}{n} = \dfrac{142}{10} = 14$ and $\bar{y} = \dfrac{\sum y}{n} = \dfrac{166}{10} = 16.6$

$\therefore$ Required correlation coefficient

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

$$= \frac{2434 - 10 \times 14.2 \times 16.6}{\sqrt{2085 - 10 \times (14.2)^2}\sqrt{2897 - 10 \times (16.6)^2}}$$

$$= \frac{2434 - 2357.2}{\sqrt{2085 - 2016.4}\sqrt{2897 - 2755.6}}$$

$$= \frac{76.8}{\sqrt{68.6}\sqrt{141.4}} = \frac{76.8}{8.28 \times 11.89} = \frac{76.8}{98.45} = 0.78$$

***Illustration** 2.* Compute the correlation coefficient between x and y from the following data

x : 65 66 67 68 69 70 71 67

y : 67 68 64 72 70 67 70 68

Solution: The correlation coefficient between two variables X and Y is given by

$$r = \frac{\sum uv - n\bar{u}\bar{v}}{\sqrt{(\sum u^2 - n\bar{u}^2)(\sum v^2 - n\bar{v}^2)}}$$

where, $u = \frac{x-a}{h}$ *and* $v = \frac{y-b}{k}$

Here, $u = \frac{x-68}{1}$ *and* $v = \frac{y-68}{1}$

To calculate the correlation coefficient, let us construct the following table:

x	y	u	v	uv	u^2	v^2
65	67	-3	-1	3	9	1
66	68	-2	0	0	4	0
67	64	-1	-4	4	1	16
68	72	0	4	0	0	16
69	70	1	2	2	1	4
70	67	2	-1	-2	4	1
71	70	3	2	6	9	4
67	68	-1	0	0	1	0
		Σu=-1	Σv =2	Σuv=13	Σu^2=29	Σv^2=42

Here, $\bar{u} = \frac{\sum u}{n}$ = -1/8 = - 0.13 and $\bar{v} = \frac{\sum v}{n}$ = 2/8 = 0.25

The required correlation coefficient ,

$$r_{xy} = r_{uv} = \frac{\sum uv - n\,\bar{u}\,\bar{v}}{\sqrt{(\sum u^2 - n\,\bar{u}^2)}\sqrt{(\sum v^2 - n\,\bar{v}^2)}}$$

$$= \frac{13 - 8(-0.13)(0.25)}{\sqrt{29 - 8(-0.13)^2}\sqrt{42 - 8(0.25)^2}}$$

$$= \frac{13 + 0.26}{\sqrt{29 - 0.1352}\sqrt{42 - 0.5}}$$

$$= \frac{13.26}{5.37 \times 6.44} = \frac{13.26}{34.58} = 0.39$$

Illustration 3 Find the correlation coefficient and the equations of regression

x : 1 2 3 4 5

y : 2 5 3 8 7

Estimate the value of *x* when *y=6* and the value of *y* when *x= 10.*

Solution: Correlation coefficient between two variables x and y is given by,

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

To calculate the correlation coefficient, let us construct the following table:

x	y	xy	x^2	y^2
1	2	2	1	4
2	5	10	4	25
3	3	9	9	9
4	8	32	16	64
5	7	35	25	49

$\Sigma x=15 \quad \Sigma y=25 \quad \Sigma xy=88 \quad \Sigma x^2=55 \quad \Sigma y^2=151$

Here, $\bar{x} = \frac{\sum x}{n} = \frac{15}{5} = 3$ and $\bar{y} = \frac{\sum y}{n} = \frac{25}{5} = 5$

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

$$= \frac{88 - 5\times 3\times 5}{\sqrt{55 - 5\times 3^2}\sqrt{151 - 5\times 5^2}} = \frac{88-75}{\sqrt{55-45}\sqrt{151-125}}$$

$$= \frac{13}{\sqrt{10}\sqrt{26}} = \frac{13}{3.16\times 5.1} = \frac{13}{16.12} = +0.81$$

The equation of the lines of regression of y on x and x on y are given by,

$$y - \bar{y} = r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \qquad \text{[y on x]}$$

$$x - \bar{x} = r_{xy}\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \text{[x on y]}$$

Here,

$$\sigma_x = \sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2} = \sqrt{\frac{1}{5}\times 55 - 3^2}$$

$$= \sqrt{11-9} = \sqrt{2} = 1.41$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum y^2 - \bar{y}^2} = \sqrt{\frac{1}{5}\times 151 - 5^2}$$

$$= \sqrt{30.2 - 25} = \sqrt{5.2} = 2.28$$

The required lines of regressions are

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

$$\Rightarrow y - 5 = 0.81\times\frac{2.28}{1.41}(x-3)$$

$$\Rightarrow y = 5 + 1.31(x-3)$$

$$\Rightarrow y = 5 + 1.31x - 3.93$$

$$\Rightarrow y = 1.31x + 1.07$$

and

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

$$\Rightarrow x - 3 = 0.81 \times \frac{1.41}{2.28}(y - 5)$$

$$\Rightarrow x = 3 + 0.5 \times (y - 5)$$

$$\Rightarrow x = 3 + 0.5y - 2.5$$

$$\Rightarrow x = 0.5y + 0.5$$

To estimate the value of x when y = 6 we use the regression line of x on y.

$x = 0.5y + 0.5 = 0.5 \times 6 + 0.5 = 3 + 0.5 = 3.5$

Similarly, to estimate the value of y when x = 10 we use the regression line of y on x.

$y = 1.31x + 1.07 = 1.31 \times 10 + 1.07 = 13.1 + 1.07 = 14.17$

Illustration 4: Obtain the lines of regression

x : -10 -5 0 5 10

y : 5 9 7 11 13

Solution : The equation of the lines of regression of y on x and x on y are given by,

$$y - \bar{y} = r_{xy} \frac{\sigma_y}{\sigma_x}(x - \bar{x}) \qquad \text{[y on x]}$$

$$x - \bar{x} = r_{xy} \frac{\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \text{[x on y]}$$

To calculate the regression equations, let us construct the following table:

x	y	xy	x^2	y^2
-10	5	-50	100	25
-5	9	-45	25	81
0	7	0	0	49
5	11	55	25	121
10	13	130	100	169
$\Sigma x = 0$	$\Sigma y = 45$	$\Sigma xy = 90$	$\Sigma x^2 = 250$	$\Sigma y^2 = 445$

Here, $\bar{x} = \frac{\sum x}{n} = \frac{0}{5} = 0$ and $\bar{y} = \frac{\sum y}{n} = \frac{45}{5} = 9$

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

$$= \frac{90 - 5 \times 0 \times 9}{\sqrt{250 - 5 \times 0^2}\sqrt{445 - 5 \times 9^2}} = \frac{90}{\sqrt{250}\sqrt{445 - 405}}$$

$$= \frac{90}{\sqrt{250}\sqrt{40}} = \frac{90}{15.81 \times 6.32} = \frac{90}{99.91} = +0.9$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2} = \sqrt{\frac{1}{5} \times 250 - 0^2}$$

$$= \sqrt{50} = 7.07$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum y^2 - \bar{y}^2} = \sqrt{\frac{1}{5} \times 445 - 9^2}$$

$$= \sqrt{89 - 81} = \sqrt{8} = 2.83$$

The required lines of regression are

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

$$\Rightarrow y - 9 = 0.9 \times \frac{2.83}{1.41}(x - 0)$$

$$\Rightarrow y - 9 = 0.36x$$

$$\Rightarrow y = 9 + 1036x$$

and

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

$$\Rightarrow x - 0 = 0.9 \times \frac{7.07}{2.83}(y - 9)$$

$$\Rightarrow x = 2.24 \times (y - 9)$$

$$\Rightarrow x = 2.24y - 20.16$$

Illustration 5. Compute correlation coefficient

n = 25, Σx = 125 Σy =100 Σx² = 650 Σy² = 460 Σxy = 508

Solution: Correlation coefficient between two variables *x* and *y* is given by,

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

Here, $\bar{x} = \frac{\sum x}{n} = \frac{125}{25} = 5$

$$\bar{y} = \frac{\sum y}{n} = \frac{100}{25} = 4$$

$\therefore$ Required correlation coefficient

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

$$= \frac{508 - 25 \times 5 \times 4}{\sqrt{650 - 25 \times (5)^2}\sqrt{460 - 25 \times (4)^2}}$$

$$= \frac{508 - 500}{\sqrt{650 - 625}\sqrt{460 - 400}}$$

$$= \frac{8}{\sqrt{25}\sqrt{60}} = \frac{8}{5 \times 7.75} = \frac{8}{38.75} = 0.21$$

Illustration 6. You are given the following data:

$\bar{x} = 36,\ \bar{y} = 85,\ \sigma_x = 11,\ \sigma_y = 8,\ r_{xy} = 0.66$

Obtain the lines of regression and estimate the value of *x* when *y = 75.*

Solution: The equation of the line of regression of y on x is given by,

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

$$\Rightarrow y - 85 = 0.66\frac{8}{11}(x - 36)$$

$$\Rightarrow y = 0.48(x - 36) + 85$$

$$\Rightarrow y = 0.48x - 17.28 + 85$$

$$\Rightarrow y = 0.48x + 67.72$$

The equation of the line of regression of x on y is given by

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

$$\Rightarrow x - 36 = 0.66\frac{11}{8}(y - 85)$$

$$\Rightarrow x = 0.91(y - 85) + 36$$

$$\Rightarrow x = 0.91y - 77.35 + 36$$

$$\Rightarrow x = 0.91y - 41.35$$

Estimation of x when y = 75

$x = 0.91y - 41.35 = 0.91 \times 75 - 41.35 = 68.25 - 41.35 = 26.90$

Illustration 7. A panel of two judges, A and B graded the acting performances by independently awarding marks as follows:

Actor	: 1	2	3	4	5	6	7
Marks by Judge A:	46	42	44	40	43	41	45
Marks by Judge B:	40	38	36	35	39	37	41

The performance of actor number 8 was awarded 37 marks by judge A while judge B forgot to award him any mark. If judge B has also awarded marks to actor number 8, what marks could be expected from him?

Solution: Here, we will first find the regression line between the marks awarded by judge B (y) on the marks provided by judge A(x). Then putting x = 37 we find the corresponding value of y . In order to solve the problem, let us first construct the following table

x	y	Xy	x^2	y^2

46	40	1840	2116	1600
42	38	1596	1764	1444
44	36	1584	1936	1296
40	35	1400	1600	1225
43	39	1677	1849	1521
41	37	1517	1681	1369
45	41	1845	2025	1681

$\Sigma x=301$ $\Sigma y=266$ $\Sigma xy=11459$ $\Sigma x^2=12971$ $\Sigma y^2=10136$

Here, $\bar{x} = \frac{\sum x}{n} = \frac{301}{7} = 43$ and $\bar{y} = \frac{\sum y}{n} = \frac{266}{7} = 38$

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

$$= \frac{11459 - 7 \times 43 \times 38}{\sqrt{12971 - 7 \times 43^2}\sqrt{10136 - 7 \times 38^2}} = \frac{11459 - 11438}{\sqrt{12971 - 12943}\sqrt{10136 - 10108}}$$

$$= \frac{21}{\sqrt{28}\sqrt{28}} = \frac{21}{28} = 0.75$$

$$\sigma x = \sqrt{\frac{1}{n}\sum x_i^2 - \bar{x}^2} = \sqrt{\frac{1}{7} \times 12971 - 43^2} = \sqrt{1853 - 1849} = \sqrt{4} = 2$$

$$\sigma y = \sqrt{\frac{1}{n}\sum y_i^2 - \bar{y}^2} = \sqrt{\frac{1}{7} \times 10136 - 38^2} = \sqrt{1448 - 1444} = \sqrt{4} = 2$$

Now,

So, $b_{yx} = r \times \frac{\sigma_y}{\sigma x} = 0.75 \times \frac{2}{2} = 0.75$

Now, regression line of y on x is given by

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

$\Rightarrow y - 38 = 0.75\,(x - 43)$

$\Rightarrow y = 38 + 0.75\,x - 32.25$

$\Rightarrow y = 5.75 + 0.75\,x$

Now when $x = 37$ we have

$y = 5.75 + 0.75 \times 37 = 33.5$

Thus, for the actor number 8 when the first judge gives 37marks, the second judge is expected to give *33.5* marks.

Illustration 8: For the variables X and Y the two lines of regression are given by $3x + 2y - 25 = 0$ and $6x + y - 30 = 0$.

(i) Identify the lines of regression of X on Y and Y on X.
(ii) Find the means of X and Y.
(iii) Find the correlation coefficient between X and Y.

Solution: Let

$3x + 2y - 25 = 0$ *(1)*

$6x + y - 30 = 0$ (2)

Let us assume that equation (1) is the line of X on Y and (2) is the line of Y on X
So from (1) we have,

$x = -2y/3 + 25/3$ Thus, $b_{xy} = -2/3$

Also, from (2) we have

$y = -6x + 30$ Thus, $b_{yx} = -6$

We know that

$r = \sqrt{(b_{xy} \times b_{yx})} = \sqrt{(-2/3 \times -6)} = \sqrt{4} = -2$

(as regression coefficient are both negative.)

This is not possible, as the correlation coefficient cannot be greater than 1 or less than –1. Thus our earlier assumptions regurding the regression lines should be reversed. Thus $3x + 2y - 25 = 0$ is the line of Y on X and $6x + y - 30 = 0$ is the line of X on Y.

So, $3x + 2y - 25 = 0$

$\Rightarrow y = -3x/2 + 25/3$ so $b_{yx} = -3/2$

and $6x + y - 30 = 0$

$\Rightarrow x = -y/6 + 5$ so $b_{xy} = -1/6$

Thus, correlation coefficient = $r = \sqrt{(b_{xy} \times b_{yx})}$

$= \sqrt{(-3/2 \times -1/6)} = -1/2 = -0.5$

Since, the two regression lines are different, so the solution of the regression lines is the point of intersection of the two lines and the point is $(\bar{x}, \bar{y})$. So, to get the values of the means we solve the two regression equations.

Multiplying (1) by 2 and subtracting from (2) we have

$$6x + y - 30 = 0$$
$$\underline{(-)\ 6x + 4y - 50 = 0}$$
$$-3y + 20 = 0$$

Thus, $\bar{y} = 20/3 = 6.67$

Now, replacing the value of $\bar{y} = 6.67$ in place of y in (1) we have,

$\bar{x} = 35/9 = 3.89$

Thus, $\bar{x} = 3.89$, $\bar{y} = 6.67$, $r = -0.5$,

Regression line of Y on X is $3x + 2y - 25 = 0$

Regression line of X on Y is $6x + y - 30 = 0$

Exercise

Theoretical

1. Prove that the correlation coefficient r lies between - 1 and + 1
2 If two variates are independent, their correlation coefficient is zero. Is the converse true? Explain by means of an example.
1. Prove that the coefficient of correlation is the geometric mean of the coefficients of regression.
4. Define correlation coefficient and mention its properties.
5. Write a short note on Regression
6. Define the coefficient of correlation (r) between the variables X and Y**.** What inference can be drawn if (i) r = 0, (ii) r =+ 1 and (iii) r = -1? Write a short note on scatter diagram.
7. Define the term 'regression'. Derive the equation of the line of regression of Y on X.
8.What do you mean by scatter diagram? Write a note on it.
9. What is correlation coefficient? Show that correlation coefficient ranges from -1 to 1.
10. Write a note on lines of regression.

1. 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

[Ans: $r = 0.94$]

2. Find the coefficient of correlation from the following data.

X:	78	36	98	25	75	82	90	62	65	39
Y:	84	51	91	60	68	62	86	58	53	47

[Ans: $r = 0.78$]

3. Data on heights and weights of a batch of students are given in the following series:

Height (inches):	48	49	50	51	52	53	54	*55*	*56*	57
Weight (lbs)	100	105	104	107	111	115	125	130	132	135

Find the coefficient of correlation. [Ans: $r = 0.977$]

4. Find the correlation coefficient from the following observations.

X:	2.52	2.49	2.47	2.42	1.69	3.43	4.72
Y:	*550*	610	730	870	880	930	400

[Ans: r = -0.54]

5. The following table gives the relative values of two variables

X:	42	44	58	55	89	98	66
Y:	56	49	53	58	65	76	58

Determine the regressions equations and hence find the correlation coefficient.

[Ans. x = 2.195 y – 65.549, y = 372 x +35.26, r =0.91]

6. From the following data calculate the regression equations:

ΣX=Rs. 580, ΣY= Rs. *370*, ΣXY=Rs. 11494

ΣX^2 = Rs. 41658 and ΣY^2 =Rs. 17206, N=12

Also compute the value of correlation co-efficient.

[Ans. $x = -1.1 y + 82.24$, $y = -0.47 x + 53.55$, $r = -0.72$]

7. In a partially destroyed record of an analysis of correlation data the following results only legible:

Variance of X = 9

Regression equations:

$8X-10Y+66 = 0$

$40X-18Y = 214$

Find *(i)* The mean values of X and Y

(ii) The co-efficient of correlation

(iii) The standard deviation of Y.

[Ans: $\bar{x} = 17$ and $\bar{y} = 13$ (ii) $r = 0.6$ (iii) $\sigma_y = 4$]

8. The following statistical co-efficients were deduced in the course of an examination of the relationship between the yield of wheat and the amount of rainfall.

	Yield in Lb. per acre	*Annual Rainfall in inches*
Mean	985.0	12.8
S.D.	70.1	1.6

$r = \pm 52$

From the above data, calculate (i) the most likely yield of wheat per acre when the annual rainfall

is 9.2 inches and (ii) the probable annual rainfall for yield of 1. 00 lb. per acre.

[Ans: (i) 902.99 lb/acre (ii) 17.74 inches]

9. Find the coefficient of correlation between x and y from the following data:

$\Sigma X = 50$, $\Sigma Y = -30$, $\Sigma XY = -115$, $\Sigma X^2 = 290$ and $\Sigma Y^2 = 300$, N=10

[Ans: –0.5864]

10. Calculate the correlation coefficient between

X:	1	2	5	8	9	10	36
Y:	3	4	8	10	12	11	48

[Answer: 0.97]

11. From the following data find the lines of regression equations, and calculate coefficient of correlation:

Sales :	91	97	103	121	67	124	51	73	111	57
Purchases:	97	75	69	97	70	91	39	61	83	47

[Answer*: $x = 8.333 + 1.118\,y$, $y = 0.636\,x + 15.678$, $r = 0.843$]*

12. The following calculations have been made for closing prices of 12 stocks (Y) on a stock exchange in a particular day, along with the volume of sales in thousands of shares (X). From this data find the regression line of Y on X.

$\Sigma X = 580$, $\Sigma Y = 370$, $\Sigma XY = 11494$, $\Sigma X^2 = 41658$ and $\Sigma Y^2 = 17206$

[Answer : $Y = 82.25 - 1.1\,X$]

13. Given the following data, estimate the value of Y when X = 60 and also find out the correlation coefficient.

$\bar{x} = 53.2$, $\bar{y} = 27.9$, $b_{xy} = -0.2$, $b_{yx} = -1.5$

[Answer : $y = 17.7$, $r = -0.548$]

14. The following data is about the sales and advertisement of a firm as given below:

	Sale (in Crores)	Advertisement Expenditure (in Crores)
Mean	40	6
S.D	10	1.5

Correlation coefficient $r = 0.9$

(i) Estimate the likely sale when advertisement expenditure is 10 crores.

(ii) What should the advertisement expenditure when the sales target of the firm is 60 crores.

[Answer : (i) 64 crores (ii) 8.7 crores]

@@@@@@@@@@@@