Correlation Analysis

Correlation is a statistical tool to measure the relationship between two or more variables. A group of statistical techniques to measure the strength of the association between variables is called correlation analysis. The examples of some correlated variables are:

(i) The family income and the expenditure on luxury items
(ii) The amount of rainfall up to a point and the production of rice
(iii) The price of a commodity and the quantity demanded
(iv) The supply of a commodity in the market and the price of the commodity.

Coefficient of correlation: The measure of correlation is called the coefficient of correlation which gives us a quantitative measure of the direction and strength of linear relationship between two numerically measured variables. It describe the strength of relationship between two sets of interval-scaled or ratio-scaled variables denoted by r and is originated by Karl Pearson. Coefficient of correlation is often referred to as Pearson's r.

The correlation coefficient r can take any value from -1 to +1 inclusive. The value of r equals -1 and +1 indicate the perfect positive and perfect negative correlation between the variables. The coefficient of correlation r close to (say, r=0.05) shows that the relationship between variables is poor or weak. The same conclusion can be drawn for r = -0.05. The correlation coefficient r = -0.9 and r = +0.9 have equal strength and both indicate very strong relationship between two variables. Thus we can say that the strength of the relationship does not depend on the direction of the variables. If there is absolutely no linear relation between variables, Pearson's r is zero.

Assumptions of correlation coefficient (r):

1. Both variables are measured on an interval or ratio scales
2. Both variables follow bivariate normal distribution
3. The relationship between the two variables is linear
4. The sample is of adequate size to assume normal

Types of correlation:

Correlation between variables may be of the following three types:

i) Positive and negative correlation.
ii) Linear and non linear correlation.
iii) Simple, multiple and partial correlation.

(i) Positive and negative correlation: The positive and negative correlation are also known as direct correlation and inverse correlation respectively. The positive and negative correlation depend upon the direction in the changes of the variables. If two variables vary in the same direction i.e. if the increase (or decrease) in the value of one variable results the increase (or decrease) in the value of other variable, then the two variables are said to have positive correlation.

For example:

(a) x:	10	20	25	50	(b) x:	100	50	30	10
y:	5	8	10	20	Y:	8	5	3	2

One the other hand, two variables are said to have negative correlation if two variables move in the opposite direction i.e. if the increase (or decrease) in the value of one variable results the decrease (or increase) in the value of other variable, then the two variables are said to have negative correlation.

For example:

(a) x:	10	20	25	50	(b) x:	100	50	30	10
y:	50	20	10	8	y:	8	15	23	28

(ii) Linear and non linear correlation:

The correlation between two variables is said to be linear when an unit change in one variable results a constant change in the other variable over the entire range of the values.

For example:

x:	1	2	3	4
y:	7	9	11	13

If corresponding to an unit change in one variable, there is no constant change in other variable, then correlation is said to be non linear.

As example :

x:	1	2	3	4
y:	7	10	11	20

(iii) Simple, multiple and partial correlation:

The correlation between two variables is known as simple correlation. When three or more variables are considered, then the correlation may be multiple or partial. In a multiple correlation, three or more variables are studied simultaneously. In a partial correlation, three will be three or more variables but we consider only two variables influencing each other and other variables being kept constant.

The following are the example of simple multiple and partial correlation.

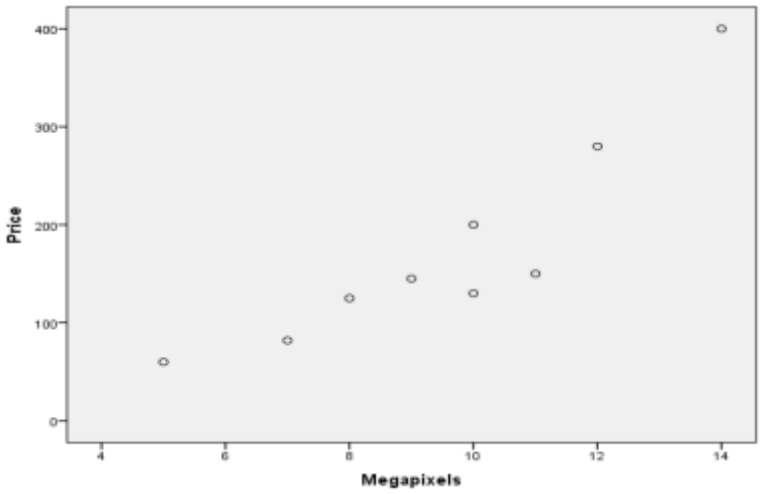
(a) The amount of fertilizer used and the yield of wheat per hectare.
(b) The amount of rainfall, quantity of fertilizer used and the yield of wheat per hectare.
(c) The amount of rainfall, and the yield of wheat per hectare, keeping the quantity of fertilizer used constant.

Methods of studying correlation:

The following methods can be used to study the correlation between two variables:

(a) Scatter diagram
(b) Karl Pearson correlation coefficient
(c) Spearmen's rank correlation

Scatter diagram:



Scatter diagram is a graphical method of studying correlation. It is the simplest method of ascertaining the correlation between two variables. Let X and Y be two variables, each consisting the same number of values. If we plot the x values along x-axis and corresponding y values along y-axis, we shall get a number of dots on the graph paper. The diagram so obtained consisting all the dots is said to be scatter diagram.

In the scatter diagram when all the dots shows an upward trend rising form lower left hand corner to the upper right hand corner, then the correlation is said to be positive and when all the dots lie in a straight line the correlation is said to be perfect positive (r=+1.0).

When all the dots shows an downward trend falling form upper left hand corner to the lower right hand corner, then the correlation is said to be negative and when all the dots lie in a straight line the correlation is said to be perfect negative (r=-1.0).

If the dots are widely scattered and they do not show any trend (rising or falling) then the variables are said to be uncorrelated.

Dependent and Independent Variables: The variable that is being predicted or estimated is called dependent variable or predicted variable. The variable that provides the basis for estimation is called independent variable or predictor variable. To develop a scatter diagram, it is common practice to scale the dependent variable on the vertical or y-axis and the independent variable on the horizontal or x-axis.

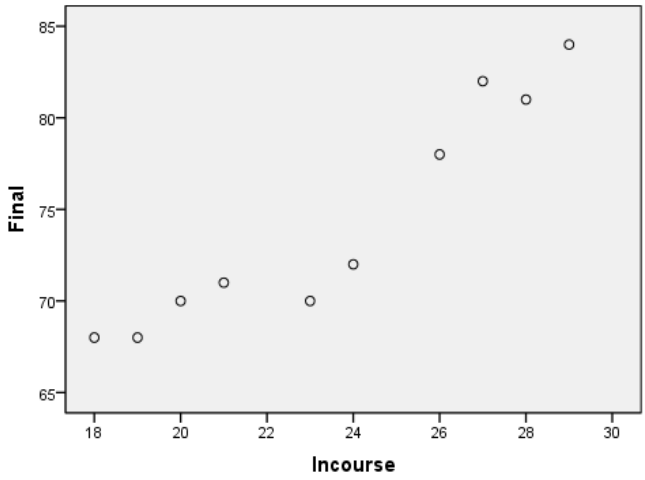
Now let us see the relationship between the variables observing the previous scatter diagrams:

Example 0.1: A teacher is trying to show his students the importance of incourse exam marks on final exam marks. He believes that the higher the incourse exam marks, the higher the final exam marks. A random sample of 10 students from his class are selected with the following marks they obtained in both exams.

Incourse exam marks	Final exam marks
21	71
18	68
27	82
29	84
24	72
20	70
28	81
26	78
19	68
23	70

Draw a scatter diagram for the data and comment on the relationship between incourse exam marks and final exam marks.

Solution: In a graph paper, plotting the incourse exam marks along x-axis and corresponding final exam marks along y-axis we get the following scatter diagram.



From the diagram we see, the dots show an upward trend from the lower left corner to upper right corner and which implies that there is a positive correlation between incourse exam marks and final exam marks i.e. a student who will get higher marks in incourse exam, it is very likely that he will get higher marks in final exam and vice versa.

Karl Pearson's Correlation coefficient:

One of the widely used mathematical methods of studying the correlation coefficient between two variables is Karl Pearson's correlation coefficient. It is also known as product moment correlation coefficient. Let (x_1, y_1), (x_2, y_2), ... (x_n, y_n) be n pairs of values of two variables x and y. The correlation coefficient between the variables x and y is denoted by r and is defined as:

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

$$= \frac{SP(x,y)}{\sqrt{SS(x) \times SS(y)}} = \frac{Sum\ of\ \ product\ of\ \ x\ and\ \ y}{\sqrt{Sum\ of\ squares of\ \ x \times sum\ of\ \ squares of\ \ y}}$$

Example 2: Calculate the coefficient of correlation (r) of the data given in **example 1.**

Solution: The coefficient of correlation**,** $r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$

Table for necessary calculation:

Incourse marks (x)	Final marks (y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})\ (y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
21	71	-2.5	-3.5	8.75	6.25	12.25
18	68	-5.5	-6.5	35.75	30.25	42.25
27	82	3.5	7.5	26.25	12.25	56.25
29	84	5.5	9.5	52.25	30.25	90.25
24	72	0.5	-2.5	-1.25	0.25	6.25
20	70	-3.5	-4.5	15.75	12.25	20.25
28	81	4.5	6.5	29.25	20.25	42.25
26	78	2.5	3.5	8.75	6.25	12.25
19	68	-4.5	-6.5	29.25	20.25	42.25
23	70	-0.5	-4.5	2.25	0.25	20.25
235	**745**			**207.00**	**138.5**	**344.5**

$$\bar{x} = \frac{\sum x}{n} = \frac{235}{10} = 23.5 \qquad \bar{y} = \frac{\sum y}{n} = \frac{745}{10} = 74.5$$

Now, **,** $r = \frac{207.00}{\sqrt{138.5 \times 344.5}} = \frac{207.00}{218.43} = 0.95$

Interpretation: The two variables, incourse marks and final marks are strongly positively correlated. The value of r=0.95 implies that if any student gets higher marks in incourse exam it is more likely that he will get higher marks in final exam and vice versa.

The coefficient of correlation can also be computed from the following formula based on actual values of variables x and y. This is a simplified version of previous formula.

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

Where,

$\sum x$ is the sum of the values of variable x

$\sum y$ is the sum of the values of variable y

$\sum x^2$ is the sum of squares the values of variable x

$\sum y^2$ is the sum of squares the values of variable y

$\sum xy$ is the sum of the product of the values of variable x and y

n is the number of pairs of values

Coefficient of determination

Coefficient of determination is the proportion of the total variation in the dependent variable y that is explained, or accounted for, by the variation in the independent variable x. It is a convenient way to evaluate the strength of regression equation. Coefficient of determination is computed by taking the square of the coefficient of correlation and is denoted by r^2. In previous example, the coefficient of correlation between the variables incourse marks and final marks is 0.95. The coefficient of determination, $r^2 = (0.95)^2 = 0.9025$, which implies that about 90 percent variation in the final marks is explained or accounted for, by the variation in the incourse marks.

Example: The owner of a motor shop wants to study the relationship between the age of a car and its selling price. Listed below is random sample of 12 used cars sold at his shop during last one year.

Age of cars (in Years)	Selling Price (in $000)	Age of cars (in Years)	Selling Price (in $000)
9	8.1	8	7.6
7	6.0	11	8.0
11	3.6	10	8.0
12	4.0	12	6.0
8	5.0	6	8.6
7	10.0	6	8.0

Requirements:

(a) If the owner of the shop wants to estimate selling price based on the age of the car, which variable is the dependent variable and which is the independent variable?

(b) Draw a scatter diagram and predict on the relationship between age and selling price of cars.

(c) Determine the coefficient of correlation.

(d) Determine the coefficient of determination.

(e) Interpret these statistical measures. Does it surprise you that the relationship is inverse?

Solution: (a) Since the owner of the shop wants to estimate selling price based on the age of the car, the selling price is the dependent variable (Y) and age is the independent variable (X).

(b) Plotting the independent variable (X) along x-axis and Dependent variable (Y) along y-axis we get the following scatter diagram.

DIAGRAM

(c) The coefficient of correlation, $r = \dfrac{\sum XY - \dfrac{\sum X \sum Y}{n}}{\sqrt{\{\sum X^2 - \dfrac{(\sum X)^2}{n}\}\{\sum Y^2 - \dfrac{(\sum Y)^2}{n}\}}}$

Table for necessary calculation:

Age (year) X	**Selling price ($000) Y**	**X**	**Y**	**XY**
7	8.1	81	65.61	72.9
9	6.0	49	36.00	42.0
11	3.6	121	12.96	39.6
12	4.0	144	16.00	48.0
8	5.0	64	25.00	40.0
7	10.0	49	100.00	70.0
8	7.6	64	57.76	60.8
11	8.0	121	64.00	88.0
10	8.0	100	64.00	80.0
12	6.0	144	36.00	72.0
6	8.6	36	73.96	51.6
6	8.0	36	64.00	48.0
107	**82.9**	**1009**	**615.29**	**712.9**

$$r = \frac{712.9 - \dfrac{107 \times 82.9}{12}}{\sqrt{(1009 - \dfrac{(107)^2}{12})\,(615.29 - \dfrac{(82.9)^2}{12})}} = \frac{712.9 - 739.19}{\sqrt{(1009 - 954.08)\,(615.29 - 572.70)}}$$

$= -0.54$

The value of the coefficient of correlation r=-0.54 implies that the variables age of car and selling price of car are negatively (inversely) correlated. That means if the age of a car goes up, the selling price of that car will go down and vice versa.

(d) the coefficient of determination $r^2 = (-0.54)^2 = 0.2916$ which indicates that 29 percent of the variation in the price of cars is explained or accounted for, by the variation in age of cars.

(e) It does not surprise me that the relationship between the variables is inverse. Because, when the age of a car increases it is quite natural that the price of the car will fall down i.e. low age of a car will produce high selling price.

Properties of correlation coefficient:

1. The coefficient of correlation is a symmetrical measure i.e. $r_{xy} = r_{yx}$

2. The coefficient of correlation is independent of the changes of origin and scale i.e. $r_{xy} = r_{uv}$

where $u = \frac{x-a}{h}$, $v = \frac{y-b}{k}$ where a, b are assumed means and h, k common factors which are called origin and scale of measurement.

3. The coefficient of correlation lies between -1 and +1
4. The coefficient of correlation is the geometric mean of two regression coefficients.

Rank Correlation: The correlation between the ranks of two variables is known as rank correlation. Rank correlation method is applied when the rank order data are available or when each variable can be ranked in some order. The measure based on this method is known as rank correlation coefficient.

The rank correlation method is recommended when

1. The values of the variables are available in rank order form.
2. The data are qualitative in nature and can be ranked is some order.
3. The data were originally quantitative in nature but because of smallness of the sample size or for convenience in fitting the requirements of analytical techniques were converted into ranks.

Computing rank correlation:

The Spearman rank correlation coefficient r_s is just the ordinary sample correlation coefficient r applied to the rank order data. The method calls for computing the sum of the squared differences between each pair of ranks, after each of the two variables to be correlated is arranged in order of ranks. Then if no tie in ranks exists we can apply the following formula for computing r_s:

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

where d_i is the difference between ranks of the i^{th} pair and n is the number of pairs included.

Example: Suppose we wish to determine whether the marks given by two independent examiners to 10 students in an examination are correlated. Let x and y respectively be the ranks of the marks given by the first examiner and the second examiner. Table below shows these marks, the rank ordering and the squares of the difference between the paired values on the basis of the marks.

First examiner			Second examiner			
Student	Marks	x_i	Marks	y_i	$d_i = x_i - y_i$	$D_i^2 = (x_i - y_i)^2$
1	65	10	30	9	+1	1
2	70	9	25	10	-1	1
3	76	7	35	8	-1	1
4	75	8	40	6	+2	4
5	80	5	38	7	-2	4

6	78	6	42	5	+1	1
7	83	4	48	3	+1	1
8	84	3	50	2	+1	1
9	85	2	55	1	+1	1
10	90	1	45	4	-3	9

Now $r_s = 1 - \frac{6 \times 24}{10(10^2 - 1)}$ = 1.0 – 0.15 = 0.85

Regression Analysis

The regression analysis is a technique of studying the dependence of one variable (called dependent variable) on one or more variables (called independent variables), with a view to estimate or predict the average value of the dependent variable in terms of the known or fixed values of the independent variables. The dependent and the independent variables are also called the explained and the explanatory variables respectively.

The regression technique is primarily used to:

(i) Estimate the relationship that exists, on the average, between the dependent variable and the explanatory variables.

(ii) Determine the effect of each of the explanatory variables on the dependent variable, controlling the effects of all other explanatory variables.

(iii) Predict the value of the dependent variable for a given value of the explanatory variable.

Regression Model:

A model is simply a mathematical equations that describes the relationship between a dependent variable and a set of independent variables.

A mathematical model in its simplest form involving two variables may be of the type

$$Y = \alpha + \beta X + \varepsilon.$$

This is the so called linear first order model, which says that for a given X, a corresponding observation Y consists of the value $\alpha + \beta X$ plus an amount ε, the increment by which any individual Y may fall off the regression line $\alpha + \beta X$.

The parameter α is the average value of Y for X= 0 and is called the Y intercept. The parameter β is the slope of the population regression line, also known as the population regression coefficient. It represents the amount of increases in Y for each unit increase in X.

Although we can not find the above parameters exactly without examining all possible occurrences of Y and X, we can use the information provided by the actual sample observations to provide us with the estimates a and b of α and β respectively. Thus we can write

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

where $\hat{Y}$ denotes the predicted value of Y for a given X when a and b are determined. The above Equation could then be used as a predictive equation. Substitution for a value of X would provide a prediction of true mean value of Y for that X.

The least squares method:

One of the important objectives of regression analysis is to find the estimates for α and β in the regression line $\mu_{y/x} = \alpha + \beta X$. from observed data, we shall designate these estimate by a and b respectively. The parameters β is called the regression coefficient of Y on X and measures the average increase in Y for a unit increase in X. β may be zero, positive or negative depending on the strength of relationship between X and Y. α is the intercept of the unknown regression equation on the Y axis. The estimates a and b will be called least squares estimates of α and β respectively.

The least squares method is thus a technique for minimizing the sum of suqares of the differences between the observed values and the estimated values of the dependent variable.

The estimating line $\hat{Y}_i$ = a+ bX_i is completely defined if the statistics a (the Y intercept) and b (the slope of the line) are known. As it appears from the diagram, Y_i is the ith observation of the variable Y associated with X_i, the ith observation on X. Then the least squares line is the line that minimizes:

$$\sum e_i^2 = \sum \left(y_i - \hat{y}_i\right)^2 = \sum \left(y_i - a - bX_i\right)^2$$

where

e_i = deviation of y_i from $\hat{y}_i$

$= y_i - \hat{y}_i$

$= y_i - a - bX_i$

The difference (Y_i - $\hat{y}_i$) between the observed and the estimated value of Y at X = X_i is called the residual or deviation corresponding to Y_i. The term $\sum e_i^2$ is known as the sum of squares of residuals.

One problem now is to compute the values of a and b that make the sum of squares e_i^2 as small as possible i.e. the values of a and b are to be so chosen that $\sum e_i^2$ is the minimum. One method of doing this is to set the partial derivatives of $\sum e_i^2$ with respect to both a and b equal to zero and solve the resulting equations. Thus differentiating first with respect to a and equating to zero

$$\frac{\partial}{\partial a} \sum e_i^2 = -2\sum \left(y_i - a - bX_i\right) = 0$$

So that, $\sum y_i = na + b\sum X_i$ ..... (1)

Again differentiating the same function with respect to b and equating to zero

$$\frac{\partial}{\partial a}\sum e_i^2 = -2\sum\left(y_i - a - bX_i\right) = 0$$

$$\sum X_i Y_i = a\sum X_i + b\sum X_i^2 \quad \text{..... (2)}$$

The equations (1) and (2) are known as the normal equations or least squares equations and the resulting estimates a and b are known as the least squares estimates of α and β respectively.

To solve these equations for a and b we multiply equation (1) by $\sum X_i$ and (2) by n, which yield

$$\sum X_i \sum y_i = na\sum X_i + b\left(\sum X_i\right)^2 \quad \text{..... (3)}$$

$$n\sum X_i Y_i = na\sum X_i + nb\sum X_i^2 \quad \text{..... (4)}$$

Subtracting (4) from (3) we get,

$$b\{n\sum x_i^2 - (\sum x_i)^2\} = n\sum x_i\, y_i - \sum x_i \sum y_i$$

$$\text{b} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} = \frac{\sum X_i Y_i - \frac{\sum x_i \sum Y_i}{n}}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$$

After the value of b has been obtained, we can compute the value of a by substituting the value of b into either equations. Thus from (1) we get

$$a = \frac{\sum Y_i}{n} - b\frac{\sum X_i}{n}$$

$$= \bar{Y} - b\bar{X}$$

Thus the fitted regression line is, $\hat{Y}_i = a + bX_i$

Example: A departmental store has the following statistics of sales (y) for a period of last one year of 10 salesman, who have varying years of experience (x).

Salesperson	Years of experience	Annual sales (in '000 Tk.)
1	1	80
2	3	97
3	4	92

4	4	102
5	6	103
6	8	111
7	10	119
8	10	119
9	10	123
10	13	136

(i) Find the regression line of y on x

(ii) Predict the annual sales volume of persons who have 12 and 15 years of sales experience.

Solution: Let the sales be Y and the experience be X. And the fitted regression line is $\hat{Y}_i = a + bX_i$

Where, b = $\dfrac{\sum X_i Y_i - \dfrac{\sum x_i \sum Y_i}{n}}{\sum X_i^{\,2} - \dfrac{(\sum X_i)^2}{n}}$ and $a = \bar{Y} - b\bar{X} = \dfrac{\sum Y_i}{n} - b\dfrac{\sum X_i}{n}$

The table for necessary calculations:

Salesperson	X_i	Y_i	$X_i^{\,2}$	$X_i Y_i$
1	1	80	1	80
2	3	97	9	291
3	4	92	16	368
4	4	102	16	408
5	6	103	36	618
6	8	111	64	888
7	10	119	100	1190
8	10	123	100	1230
9	11	117	121	1287
10	13	136	169	1768
Totals	70 $\sum X_i$	1080 $\sum Y_i$	632 $\sum X_i^{\,2}$	8128 $\sum X_i Y_i$

Now $b = \dfrac{8128 - \dfrac{70 \times 1080}{10}}{632 - \dfrac{(70)^2}{10}} = \dfrac{8128 - 7560}{632 - 490} = 4.0$

a = $\dfrac{1080}{10} - 4.0 \times \dfrac{70}{10}$ = 108 – 28 = 80.0

The fitted regression line of Y on X is, $\hat{Y}_i = 80 + 4X_i$ ..... (A)

Putting X=12 and X=15 in (A) we shall get the sales of the workers with experience 12 and 15 years respectively. Thus,

$\hat{Y}_{12} = 80 + 4 \times 12 = 128$ (in '000 Tk.) and $\hat{Y}_{15} = 80 + 4 \times 15 = 140$ (in '000 Tk.)

Coefficient of determination

One convenient way to evaluate the strength of regression equation is to compute coefficient of determination, which shows the proportion of the total variation in the dependent variable Y explained by the explanatory variable X. Coefficient of determination is computed by taking the square of the correlation coefficient and is denoted by r^2.

If r = 0.803 then $r^2 = 0.6448$ ➔ 64% of the variation in Y can be explained by X.

The larger the value of r^2, the better the fitted regression model in explaining the variability in the observed values of Y.

Difference between correlation and regression:

Basis for Comparison	Correlation	Regression
Meaning	Correlation is a statistical measure that determines the association between two variables.	Regression describes how to numerically relate an independent variable to the dependent variable.
Usage	To represent a linear relationship between variables.	To fit the best line and to estimate one variable based on another.
Dependent and independent variable	No difference	Both variables are different
Indicate	Correlation coefficient indicates the extent to which two variables move together	Regression indicates the impact of a change of unit on the estimated variable (y) in the known variable (x).
Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variables on the basis of the values of a fixed variables.