[Academic Script]

Correlation

Subject:	Business Economics
Course:	B. A. (Hons.), 1st Semester, Undergraduate
Paper No. & Title:	Paper – 102 Statistics for Business Economics
Unit No. & Title:	Unit – 3 Multivariate Analysis
Lecture No. & Title:	Lecture – 1 Correlation

Academic Script

1. Introduction: Multi-variate Data

When information regarding two different characteristics of an object is collected at same point of time then the group of information is called bi-variate data. For example, suppose during last fifteen days, information regarding the price and demand of a particular type of vegetable is collected then the group of these fifteen pair of observations is called bi-variate data.

Similarly, when information regarding three characteristics of an object is collected at same point of time then the group of information is called tri-variate data. For example, suppose during last fifteen days, information regarding quantity, price and demand of a particular type of vegetable is collected then group of these fifteen values is called tri-variate data.

In general, group of information regarding two or more characteristics of an object collected at same point of time is called multivariate-variate data.

2. Introduction: Correlation

While studying such characteristics, it is observed that as the value of one characteristic (variable) changes then the value of other characteristic (variable) also changes, which indicate the existence of some relationship between them and hence to know the strength of such relationship the study of correlation is must.

The Direct or indirect cause and effect relationship between two variables of a bi-variate data is called correlation. i.e. when the changes in the values of two variables is simultaneous and the value of one variable

changes due to the change in the value of other variable then two variables are called correlated.

As in the above example there may be change in demand due to change in price of that vegetable, it is called direct cause and effect relationship. When the correlation between two variables exists due to the existence of third variable then it is called indirect cause and effect relationship, e.g. let us consider the demand of raincoat and total yield of crop then there is no apparent relation between them but due to the third variable rainfall there exists relation.

3. Multiple and Partial Correlation

When the correlation between one variable and the linear combination of other variable is to be studied then it is called multiple correlation, e.g. in the above tri-variate example the correlation between demand and price by ignoring the third variable quantity then it is called partial correlation between demand and price by ignoring the effect of quantity.

4. Assumptions for the study

The followings are important assumptions for the study:

1. There must exist logical relation between two variables
2. There is linear relation between two variables i.e. the proportionate changes in the values of variable are same.

In general, it may happen that the relationship between the variables may not be linear then the correlation between them is called non-linear correlation. Here according to the syllabus, we discuss only linear correlation.

5. Types of Correlation

There are two types of correlation

I. Positive correlation: When the changes in the values of two variables are in same direction then the correlation between them is called positive correlation. For example unemployment and poverty in a region

II. Negative correlation: When the changes in the values of two variables are in opposite direction then the correlation between them is called negative correlation. For example population and per capita income of a country.

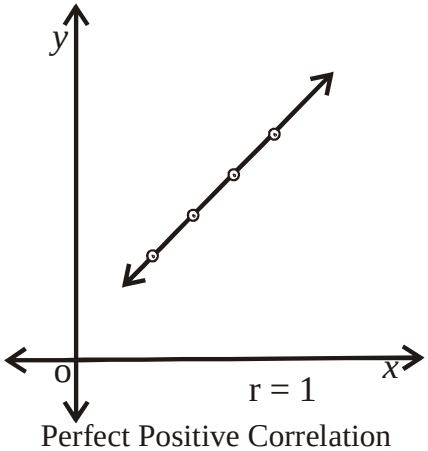
6. Methods for the Study of Correlation

There are many methods for the study of correlation, like scatter diagram method, Karl Pearson's product moment Method, Spearman's rank method, concurrent deviation method, intra-class correlation method, etc., among them within the boundary of the syllabus some methods are discussed here.

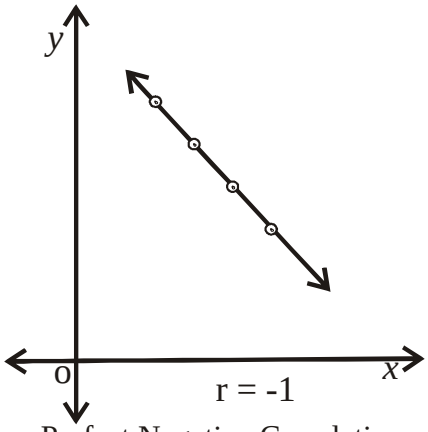
1. Scatter diagram method:

The ***scatter diagram*** is the graphical presentation of the data, by taking the values of variable x on $x-axis$ and values of variable y on $y-axis$ if we plot the given data values on graph paper then the figure obtained is called scatter diagram. If the data points are in increasing direction then it indicates positive correlation between the variables and if the points are in decreasing direction then it indicates negative correlation between the variables.

If the data points are on same line then it indicates the perfect linear correlation between the variables.

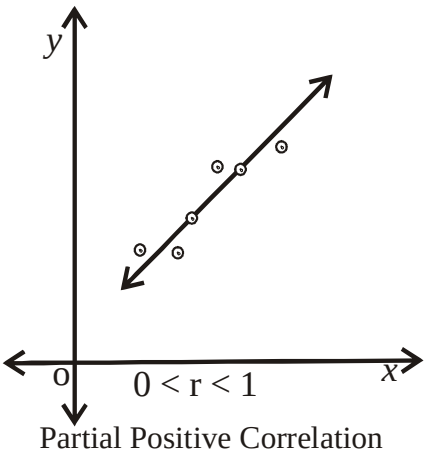


Perfect Positive Correlation

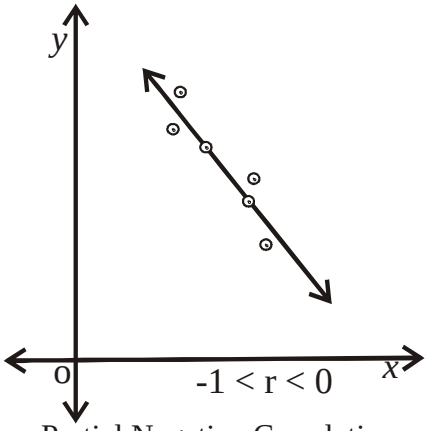


Perfect Negative Correlation

As the data point on graph are away from the line then it indicate the weak linear relationship between the variables. The figures show partial correlation between the variables.

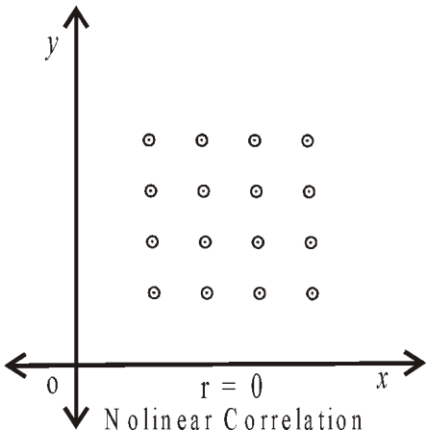


Partial Positive Correlation

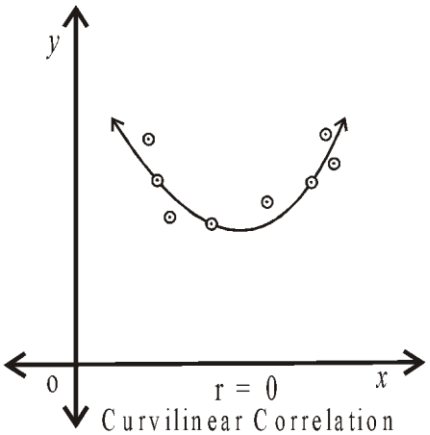


Partial Negative Correlation

If the points on a diagram are randomly distributed, then it indicates the absence of linear correlation. In such case there may be any other non linear correlation between the variables. In this case, the value of correlation coefficient is zero i.e., $r = 0$ and two variables are linearly independent. There may be curvilinear or periodical or cyclical correlation between the variables.



Nolinear Correlation



Curvilinear Correlation

If the point on scatter diagram lies on same line but the line is horizontal or vertical to the axis then there does not exists any kind of correlation between the variables. There is constant relation between two variables.

From the diagram the nature or the type or the direction of the correlation between variables is to be studied very easily but degree or numeric measure of the relationship is not known and hence this method is not used in practice.

Correlation coefficient:

A numerical measure which shows the degree and direction of the relationship between the correlated variable is called correlation coefficient and is usually denoted by 'r'.

The following are important properties of correlation coefficient.

1. The value of correlation coefficient lies from – 1 to +1, i.e. $-1 \leq r \leq 1$.
2. It is relative measure i.e. it is unit free measure.
3. It is independent of change of origin and scale but when the variables are multiplied by opposite sign then the value of correlation coefficient changes by its sign only.

2. Karl Pearson's product moment method:

The ***Karl Pearson's*** correlation coefficient is a ratio of covariance between two variables to the product of their standard deviations.

$$r = \frac{\mathrm{cov}(x,y)}{Sx \cdot Sy};$$

Where,

$$\mathrm{cov}(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}; \qquad Sx = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}} \quad Sy = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}$$

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

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

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}; \; Where \; u = \frac{x-A}{C_x}, \quad v = \frac{y-B}{C_y}$$

For bi-variate frequency distribution the following formula is used.

$$r = \frac{n\,\text{å}\,uvfxy - \left(\text{å}\,ufx\right)\left(\text{å}\,vfy\right)}{\sqrt{n\,\text{å}\,u^2 fx - \left(\text{å}\,ufx\right)^2}\sqrt{n\,\text{å}\,v^2 fy - \left(\text{å}\,vfy\right)^2}}$$

Problem 1: A manager of a company suggests that they should do aggressive marketing of newly design product and that the company has to invest more amount of money in advertisement. A survey is conducted for knowing the relationship between the expense in advertisement and total sales revenue of that product and the information regarding the advertise expense and sales revenue in rupees are given below. Your duty is to give advice to

the management that whether they should follow the manager's suggestion or not?

Amount spend in advertisement (in lakh rupees)	55	56	57	58	59	60	61
Amount of sales revenue (in crore rupees)	57	58	56	59	62	62	59

Solution:

Consider the amount spend in advertisement (in lakh rupees) as x and amount of sales revenue (in crore rupees) as y then it can be seen that the average amount spend in advertisement is $\underline{x} = 58$ lakhs rupees and the average amount of sales revenue is $\underline{y} = 59$ crore rupees.

					$(y-\bar{y})^2$	
55	57	−3	−2	9	4	6
56	58	−2	−1	4	1	2
57	56	−1	−3	1	9	3
58	59	0	0	0	0	0
59	62	1	3	1	9	3
60	62	2	3	4	9	6
61	59	3	0	9	0	0
0	**0**	**28**	**32**	**20**		

∴ The Karl Pearson's correlation coefficient is

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

$$= \frac{20}{\sqrt{28}\sqrt{32}}$$

$$= 0.67$$

Which shows positive correlation between the variables but the magnitude of the relationship is not that much high. It means that this study suggest that if you invest large amount of money in the advertisement of the product it doesn't gives surety about the equivalent high return of the sales revenue i.e., the sales revenue will also depends upon some other factors, hence for increasing the sales revenue, to increase in the advertisement expenditure is not only the solution but along with it some other precautionary steps are also to be taken along with the increase in the advertisement expense.

Limitations of Karl Pearson's correlation coefficient:

This method gives the most reliable value of correlation coefficient in case of linear relationship only. The main limitations for the method are as under

- The method is not used for non-linear relationship between the variables.
- Also when the nature of data is qualitative then this method in not used directly.

2. Spearman's Rank Method.

This method is an extension of Karl Pearson's method to the qualitative data. When the characteristic under study is not measured in terms of numbers then it is known as qualitative data or attributes. For example honesty, beauty, poverty, love, hungriness, angriness, etc. are not measured in terms of number so they are called qualitative or attribute data. It is also interested to study the association ship between beauty and intelligence, honesty and poverty, etc. All such variables are not measured in terms of number but can be assigned ranks

depending its priority or importance and then the coefficient of relationship can be determined among them. This coefficient is called Spearman's rank correlation coefficient and it also possesses all properties of correlation coefficient. This method is also used for the quantitative data. It is observed that when the variations in the data values are large then the Karl Pearson's method is not reliable, in such case, first of all the ranks are assigned to the data values either in ascending or descending order. For the sake of simplicity usually the ranks are assigned in ascending order, i.e., the first rank is given to the highest data value and the next highest has to be given second rank, and so on. If the data values are same, then one has to assign the average of the ranks assignable to them and for obtaining correlation coefficient, the following formulas can be used.

$$r = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} \qquad \left(\text{When ranks are not repeated.}\right)$$

$$r = 1 - \frac{6\left[\sum d^2 + \frac{m}{12}\left(m^2 - 1\right) + \frac{m}{12}\left(m^2 - 1\right) + ...\right]}{n\left(n^2 - 1\right)} \left(\text{When ranks are repeated.}\right)$$

Where $d = R_x - R_y$ = difference in ranks and $\frac{m}{12}\left(m^2 - 1\right)$ is called correction factor.

This method is less reliable as compared to Karl Pearson's method in general but in the case of qualitative data or the case when the relation between the variables is very weak linear then this method is used. The main disadvantage of this method is that it is not used for bi-variate frequency distribution.

Followings are some examples for illustrating the method.

Problem 2: A college has organized a talent hunt competition during festive season. For the final competition two judges from the related fields were invited. The two judges have given ranks to ten finalists for their performance as follows. The in-charge faculty of the college was not happy with the ranks assigned by the judges and hence he claim that the two judges doesn't have enough synchronized attitude toward the identification of the talent of the participants. Check whether the faculty's claim is correct or not?

Ranks by Judge – 1	9	6	1	2	10	7	4	8	5	3
Ranks by Judge – 2	6	4	9	8	1	2	3	10	5	7

Solution: Here the ranks assigned by the two judges to the ten competitors are given

Ranks by Judge – 1 *Rx*	9	6	1	2	10	7	4	8	5	3	
Ranks by Judge – 2 *Ry*	6	4	9	8	1	2	3	10 1	5	7	**Total**
Difference in ranks $d = Rx - Ry$	3	2	–8	–6	9	5	–1	–2	0	–4	**0**
d^2	9	4	64	36	81	25	1	4	0	16	**240**

Now

$$r = 1 - \frac{6\Sigma d^2}{n\left(n^2 - 1\right)}$$

$$= 1 - \frac{6(240)}{10(100 - 1)}$$

$$= 1 - 1.45$$

$$= - 0.45$$

Since *r* is negative so the two judges differ in their judgment i.e. the criterion for assigning the ranks are different and hence there

is no synchronized attitude toward the identification of the talent of the participants and hence the faculty's claim is correct.

Problem 3: In a stock market the assumption about the market volatility is made by two experts. The followings are the chances (in terms of percentage) about the instability of the market are predicted for ten consecutive working days. Can it be concluded that both the experts have similar approach in predicting the market volatility?

% of instability by expert 1	35	40	42	43	40	53	54	49	41	55
% of instability by expert 2	80	95	100	109	95	110	110	98	95	119

Solution: Here consider *x* as percentage of instability by expert 1 and *y* as percentage of instability by expert 2. Since the numeric values represented in the above table represent the belief of the experts about the volatility of the stock market so it is advisable to consider it as qualitative data and hence one has to assign ranks to the data values and then the Spearman's rank correlation coefficient is determined.

x	*Y*	R*x*	R*y*	d=R*x*-R*y*	d2
35	80	10	10	0.0	0
40	95	8.5	8	0.5	0.25
42	100	6	5	1.0	1.00
43	109	5	4	1.0	1.00
40	95	8.5	8	0.5	0.25
53	110	3	2.5	0.5	0.25
54	110	2	2.5	-0.5	0.25
49	98	4	6	-2.0	4.00
41	95	7	8	-1.0	1.00

55	119	1	1	0	0.00
0	**8.00**				

Here in the given data three data values 40, 95 and 110 are repeated so the ranks are repeated, so the correction factor is to be written three times and the respective formula can be written as under

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

In x, 40 is repeated 2 times so $m = 2$, in y, 95 is repeated 3 times so $m = 3$ and 110 is repeated 2 times so $m = 2$.

$$r = 1 - \frac{6\left[8 + \frac{2}{12}(4-1) + \frac{3}{12}(9-1) + \frac{2}{12}(4-1)\right]}{10(100-1)}$$

$$= 1 - \frac{6[8 + 0.5 + 2 + 0.5]}{990}$$

= 1 - 0.067

= 0.933

It indicates high degree of positive correlation between the attitudes of assigning percentages to the volatility of stock market. This suggests that both the experts have almost similar approach in predicting the instability of market.

7. Probable Error

A sample is drawn from a bi-variate population and a correlation coefficient is obtained for a sample with an object to make inference about the parent population. From the given population number of different samples can be drawn and from each of them correlation coefficient can be obtained which may differ

from population correlation coefficient. ***An average of the absolute differences between population correlation coefficient and all possible sample correlation coefficients is called probable error (PE).*** The value of probable error depends upon the value of sample size, as sample size increases, value of probable error decreases. If the sample size is n then its formula is given as under:

$$PE = \frac{0.6745(1-r^2)}{\sqrt{n}}$$

It is used to check whether there exists significant correlation between the population variables or not?

If $|r| > 6(PE)$ then there exists significant correlation between population variables

If $|r| < (PE)$ then there exists insignificant correlation between population variables

It is also used to obtain the limit for population correlation coefficient within which it expected to lie. This limit for population correlation coefficient is given by $(r - PE,\ r + PE)$

For example suppose for 10 pairs of data *r = 0.9* then

$$PE = \frac{0.6745(1-0.9^2)}{\sqrt{10}} = 0.0405$$

$|r| > 6(PE);\ since,\ 0.9 > 0.24$ this indicates that there exists significant correlation between population variables and also the value of population correlation coefficient may lie within the interval *(0.8595, 0.9405)*.

8. Summary

Let us summarize today's talk. Today we learn correlation, types of correlation and different methods like scatter diagram, Karl Pearson's product moment method, spearman's rank method for

the study of correlation. We also learn how to apply that methods in real life and probable error. Thank you.