Module-3
Correlation and Regression

In this Module, we study the relationship between the variables. Also, the interest lies in establishing the actual relationship between two or more variables. This problem is dealt with regression. On the other hand, we are often not interested to know the actual relationship but are only interested in knowing the degree of relationship between two or more variables. This problem is dealt with correlation analysis.

Linear relationship between two variables is represented by a straight line which is known as regression line. In the study of linear relationship between two variables X and Y, suppose the variable Y is such that it depends on X, then we call it as the regression line of Y on X. If X depends on Y, then it is called as the regression line of X on Y.

To find out the regression line, the observations (x_i, y_i) on the variable X and Y are necessarily taken in pairs. For example, a chemical engineer may run a chemical process several times in order to study the relationship between the concentration of a certain catalyst and the yield of the process. Each time the process is run, the concentration X and the yield Y are recorded. Generally, the studies are based on samples of size 'n' and hence 'n' pairs of sample observations can be written as $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Correlation

In a bivariate distribution, we are interested to find out whether there is any relationship between two variables. The correlation is a statistical technique which studies the relationship between two or more variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of relationship between the two variables. When two variables are related in such a way that a change in the value of one is accompanied either by a direct change or by an inverse change in the values of the other, the two variables are said to be correlated. In the correlated variables an increase in one variable is accompanied by an increase or decrease in the other

variable. For instance, relationship exists between the price and demand of a commodity because keeping other things equal, an increase in the price of a commodity shall cause a decrease in the demand for that commodity. Relationship might exist between the heights and weights of the students and between amount of rainfall in a city and the sales of raincoats in that city.

Utility of Correlation

The study of correlation is very useful in practical life as revealed by these points.

1. With the help of correlation analysis, we can measure in one figure, the degree of relationship existing between variables like price, demand, supply, income, expenditure etc. Once we know that two variables are correlated then we can easily estimate the value of one variable, given the value of other.
2. Correlation analysis is of great use to economists and businessmen; it reveals to the economists the disturbing factors and suggests to him the stabilizing forces. In business, it enables the executive to estimate costs, sales etc. and plan accordingly.
3. Correlation analysis is helpful to scientists. Nature has been found to be a multiplicity of interrelated forces.

Types of Correlation

Correlation can be categorized as one of the following:

(i) Positive and Negative,

(ii) Simple and Multiple.

(iii) Partial and Total.

(iv) Linear and Non-Linear (Curvilinear)

(i) Positive and Negative Correlation : Positive or direct Correlation refers to the movement of variables in the same direction. The correlation is said to be positive when the increase (decrease) in the value of one variable is

accompanied by an increase (decrease) in the value of other variable also. Negative or inverse correlation refers to the movement of the variables in opposite direction. Correlation is said to be negative, if an increase (decrease) in the value of one variable is accompanied by a decrease (increase) in the value of other.

(ii) Simple and Multiple Correlation : Under simple correlation, we study the relationship between two variables only i.e., between the yield of wheat and the amount of rainfall or between demand and supply of a commodity. In case of multiple correlation, the relationship is studied among three or more variables. For example, the relationship of yield of wheat may be studied with both chemical fertilizers and the pesticides.

(iii) Partial and Total Correlation : There are two categories of multiple correlation analysis. Under partial correlation, the relationship of two or more variables is studied in such a way that only one dependent variable and one independent variable is considered and all others are kept constant. For example, coefficient of correlation between yield of wheat and chemical fertilizers excluding the effects of pesticides and manures is called partial correlation. Total correlation is based upon all the variables.

(iv) Linear and Non-Linear Correlation: When the amount of change in one variable tends to keep a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. But 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. The distinction between linear and non-linear is based upon the consistency of the ratio of change between the variables.

Methods of Studying Correlation

There are different methods which helps us to find out whether the variables are related or not.

1. Scatter Diagram Method.
2. Graphic Method.
3. Karl Pearson's Coefficient of correlation.
4. Rank Method.

Karl Pearson's Co-efficient of Correlation.

Karl Pearson's method, popularly known as Pearsonian co-efficient of correlation, is most widely applied in practice to measure correlation. The Pearsonian co-efficient of correlation is represented by the symbol r. Degree of correlation varies between + 1 and –1; the result will be + 1 in case of perfect positive correlation and – 1 in case of perfect negative correlation. Computation of correlation coefficient can be simplified by dividing the given data by a common factor. In such a case, the final result is not multiplied by the common factor because coefficient of correlation is independent of change of scale and origin.

$$r(X,Y) = \rho(X,Y) = \frac{Cov(x,Y)}{\sigma_X . \sigma_Y}$$

$$Cov(X,Y) = \frac{1}{n}\sum XY - \overline{X}\overline{Y}$$

$$\sigma_X = \sqrt{\frac{1}{n}\sum X^2 - \overline{X}^2} \quad , \quad \sigma_Y = \sqrt{\frac{1}{n}\sum Y^2 - \overline{Y}^2}$$

n - number of items in the given data

Standard Error

The standard error is the approximate standard deviation of a statistical sample population. The standard error is a statistical term that measures the accuracy with which a sample represents a population.

In statistics, a sample means deviates from the actual mean of a population; this deviation is the standard error.

$$S.E(r) = \frac{1 - r^2}{\sqrt{n}}$$

Probable Error= $P.E\ (r) = 0.675 \times S.E(r)$

Range:

$$r - S.E(r) \leq Population\ r \leq r + S.E(r)$$

Note: Two independent variables are uncorrelated when Cov(X,Y) = 0

Problems:

1. Find the correlation coefficient between annual advertising expenditures and annual sales revenue for the following data:

Year (i)	1	2	3	4	5	6	7	8	9	10
Annual advertising expenditure (X_i)	10	12	14	16	18	20	22	24	26	28
Annual sales (Y_i)	20	30	37	50	56	78	89	100	120	110

Solution: Now, $\bar{X} = \frac{\sum X}{n} = \frac{190}{10} = 19$, $\bar{Y} = \frac{\sum Y}{n} = \frac{690}{10} = 69$

i	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	10	20	-9	-49	81	2401	441
2	12	30	-7	-39	49	1521	273
3	14	37	-5	-32	25	1024	160
4	16	50	-3	-19	9	364	57
5	18	56	-1	-13	1	169	13
6	20	78	1	9	1	81	9
7	22	89	3	20	9	400	60
8	24	100	5	31	25	961	155
9	26	120	7	51	49	2601	357
10	28	110	9	41	81	1681	369
	190	**690**	**0**	**0**	**330**	**11200**	**1894**

Correlation coefficient is $r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{1894}{\sqrt{330}\sqrt{11200}} = 0.985$

The correlation coefficient between annual expenditure and annual sales revenue is 0.985.

2. Let X, Y and Z be uncorrelated random variables with zero means and standard deviations 5, 12 and 9 respectively. If U = X + Y and V = Y + Z, find the correlation coefficient between U and V.

Solution:

Given that all the three random variables have zero mean.

Hence, E(X) = E(Y) = E(Z) = 0.

Now, Var(X) = $E(X^2) - [E(X)]^2$

$\Rightarrow \quad E(X^2)$ = Var(X) { since, E(X) = 0}

$= 5^2 = 25$

Similarly, $E(Y^2) = 12^2 = 144$ and $E(Z^2) = 9^2 = 81$

Since X and Y are uncorrelated we have Cov(X,Y) = 0

$\Rightarrow$ E(XY) = E(X).E(Y) = 0

Similarly, E(YZ) = 0 and E(ZX) = 0.

To find $\rho(U,V)$:

Now, $\rho(U,V) = \frac{E(UV)-E(U).E(V)}{\sigma_U.\sigma_V}$

E(U) = E [X + Y] = E[X] + E[Y] = 0

E(V) = E [Y + Z] = E[Y] + E[Z] = 0

$E(U^2) = E[(X+Y)^2] = E[X^2] + E\ [Y^2] + 2E\ [XY]$

$= 25 + 144 + 0$

$= 169$

Similarly, $E(V^2) = 225$

Now, $Var(U) = E(U^2) - [E(U)]^2 = 169$

$\Rightarrow \sigma_U = \sqrt{169} = 13$

Similarly, $Var(V) = E(V^2) - [E(V)]^2 = 225$

$\Rightarrow \sigma_V = \sqrt{225} = 15$

E(UV) = E[(X+Y) (Y+Z)]
= E(XY) + $E(Y^2)$ + E(XZ) + E(YZ)
= 144

Therefore, $\rho(U,V) = \frac{E(UV)-E(U).E(V)}{\sigma_U.\sigma_V} = \frac{144}{195} = \frac{48}{65}$

3. If the joint pdf of (X,Y) is given by $f(x,y) = x + y, \quad 0 \leq x, y \leq 1$. Find ρ_{XY}.
Solution:
We know that, $\rho(X,Y) = \frac{E(XY)-E(X).E(Y)}{\sigma_X.\sigma_Y}$
Now, $E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy$
$= \int_0^1\int_0^1 xy(x+y)dxdy$
$= \int_0^1 \left[\frac{x^3y}{3} + \frac{x^2y^2}{2}\right]_0^1 dy$
$= \int_0^1 \left[\frac{y}{3} + \frac{y^2}{2}\right] dy$
$= \left[\frac{y^2}{6} + \frac{y^3}{6}\right]_0^1$
$= \frac{1}{3}$
The pdf of X and Y is given by
$f(x) = \int_0^1 f(x,y)dy = \int_0^1 (x+y)dy = \left[xy + \frac{y^2}{2}\right]_0^1 = x + \frac{1}{2}$
$f(y) = \int_0^1 f(x,y)dx = \int_0^1 (x+y)dx = \left[\frac{x^2}{2} + xy\right]_0^1 = y + \frac{1}{2}$
$E(X) = \int_0^1 xf(x)dx = \int_0^1 x\left(x+\frac{1}{2}\right)dx = \left[\frac{x^3}{3} + \frac{x^2}{4}\right]_0^1 = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$
$E(Y) = \int_0^1 yf(y)dy = \int_0^1 y\left(y+\frac{1}{2}\right)dy = \left[\frac{y^3}{3} + \frac{y^2}{4}\right]_0^1 = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$
$E(X^2) = \int_0^1 x^2f(x)dx = \int_0^1 x^2\left(x+\frac{1}{2}\right)dx = \left[\frac{x^4}{4} + \frac{x^3}{6}\right]_0^1 = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$

$E(Y^2) = \int_0^1 y^2 f(y)dy = \int_0^1 y^2 \left(y + \frac{1}{2}\right) dy = \left[\frac{y^4}{4} + \frac{y^3}{6}\right]_0^1 = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{5}{12} + \left(\frac{7}{12}\right)^2 = \frac{11}{144}$$

$$\Rightarrow \sigma_X = \frac{\sqrt{11}}{12}$$

$$Var(Y) = E(Y^2) - [E(Y)]^2 = \frac{5}{12} + \left(\frac{7}{12}\right)^2 = \frac{11}{144}$$

$$\Rightarrow \sigma_Y = \frac{\sqrt{11}}{12}$$

Therefore, $\rho(X,Y) = \frac{E(XY)-E(X).E(Y)}{\sigma_X.\sigma_Y} = \frac{\frac{1}{3} - \frac{7}{12}.\frac{7}{12}}{\frac{\sqrt{11}}{12}.\frac{\sqrt{11}}{12}} = \frac{-1}{11}$

4. The independent random variables X and Y have the pdf given by $f_X(x) = \begin{cases} 4ax & ,0 \le x \le 1 \\ 0 & ,otherwise \end{cases}$, $f_Y(y) = \begin{cases} 4by & ,0 \le y \le 1 \\ 0 & ,otherwise \end{cases}$
Find the correlation coefficient.

Solution:

$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 x4axdx = 4a\int_0^1 x^2dx = 4a\left[\frac{x^3}{3}\right]_0^1 = \frac{4a}{3}$

$E(Y) = \int_{-\infty}^{\infty} yf(y)dy = \int_0^1 y4bydy = 4b\int_0^1 y^2dy = 4b\left[\frac{y^3}{3}\right]_0^1 = \frac{4b}{3}$

Since X and Y are independent, the joint pdf of X and Y is given by $f(x,y) = f(x).f(y)$

$= (4ax)(4by)$

$= 16abxy, \quad 0 \le x \le 1, \quad 0 \le y \le 1$

Now, $E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy$

$=\int_0^1\int_0^1 xy(16abxy)dxdy \quad = \frac{16ab}{9}$

Therefore we get, Cov(X,Y) = E(XY) – E(X)E(Y)

$= \frac{16ab}{9} - \left(\frac{4a}{3}\right)\left(\frac{4b}{3}\right) = 0$

Which implies that the cor(X,Y)=0

That is, the variables X and Y are independent and there is no relationship between them.

SPEARMAN'S RANK CORRELATION COEFFICIENT

Rank correlation coefficient is useful for finding correlation between any two qualitative characteristics such as Beauty, Honesty, and Intelligence etc., which cannot be measured quantitatively but can be arranged serially in order of merit or proficiency possessing the two characteristics.

Suppose we associate the ranks to individuals or items in two series based on order of merit, the Spearman's Rank correlation coefficient r is given by

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

Where,

∑d2 = Sum of squares of differences of ranks between paired items in two series

n = Number of paired items

Remarks

Spearman's rank correlation coefficient can be used to find the correlation between two quantitative characteristics or variables. In this case, we associate the ranks to the observations based on their magnitudes for X and Y series separately. Let RX and Ry be the ranks of observations on two variables X and

Y respectively for a pair. Then the Spearman's rank correlation coefficient is given by

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

Where, $\Sigma d^2 = (R_x - R_y)^2$ = sum of squares of differences between the ranks of variables X and Y

n = number of pairs of observations

SPEARMAN'S RANK CORRELATION COFFICIENT FOR A DATA WITH TIED OBSERVATIONS

In any series, if two or more observations are having same values then the observations are said to be tied observations. If tie occurs for two or more observations in a series, then common ranks have to be given to the tied observations in that series; these common ranks are the average of the ranks, which these observations would have assumed if they were slightly different from each other and the next observation will get the rank next to the rank already assumed.

In the case of data with tied observations, the Spearman's rank correlation coefficient is given by

$$\rho = 1 - \left[\frac{6(Adj \sum d^2)}{n(n^2 - 1)} \right]$$

Where,

$$Adj \sum d^2 = \sum d^2 + \left[\frac{S_1^3 - S_1}{12} \right] + \left[\frac{S_2^3 - S_2}{12} \right] + \left[\frac{S_3^3 - S_3}{12} \right] + ...$$

Here,

S_1 is the number of times first tied observation is repeated

S_2 is the number of times second tied observation is repeated

S_3 is the number of times third observation is repeated etc.

Problem: In a quantitative aptitude test, two judges rank the ten competitors in the following order.

Competitor	1	2	3	4	5	6	7	8	9	10
Ranking of judge I	4	5	2	7	8	1	6	9	3	10
Ranking of judge II	8	3	9	10	6	7	2	5	1	4

Is there any concordance between the two judges ?

Solution: Let Rx: Ranking by Judge I and Ry: Ranking by Judge II The Spearman's rank correlation coefficient is given by

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

Where, $\Sigma d^2 = (R_x - R_y)^2$ and n = number of competitors.

R_x	R_y	$d = R_x - R_y$	d^2
4	8	-4	16
5	3	2	4
2	9	-7	79
7	10	-3	9
8	6	2	4
1	7	-6	36
6	2	4	16
9	5	4	16
3	1	2	4
10	4	6	36
		TOT	190

$$\rho = 1 - \left[\frac{6(190)}{10(100 - 1)} \right]$$

$$= 1 - 1.1515$$

$$= -0.1515$$

We say that there is low degree of negative rank correlation between the two judges.

Problem : Twelve recruits were subjected to selection test to ascertain their suitability for a certain course of training. At the end of training they were given a proficiency test. The marks scored by the recruits are recorded below:

Recruit	1	2	3	4	5	6	7	8	9	10	11	12
Selection Test Score	44	49	52	54	47	76	65	60	63	58	50	67
Proficiency Test Scrore	48	55	45	60	43	80	58	50	77	46	47	65

calculate rank correlation coefficient and comment on your result

Solution: Let selection test score be a variable X and proficiency test score be a variable Y. We associate the ranks to the scores based on their magnitudes. The spearman's rank correlation coefficient is given by

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

Where, $\Sigma d^2 = (R_x - R_y)^2$ = sum of squares of differences between the ranks of observations X and Y

n = number of recruits.

Given,

X	Y	R_x	R_y	$d = R_x - R_y$	d^2
44	48	12	8	4	16
49	55	10	6	4	16
52	45	8	11	-3	9
54	60	7	4	3	9

47	43	11	12	-1	1
76	80	1	1	0	0
65	58	3	5	-2	4
60	50	5	7	-2	4
63	77	4	2	2	4
58	46	6	10	-4	16
50	47	9	9	0	0
67	65	2	3	-1	1

From the table, we have,

$$\sum d^2 = 80,\ n = 12$$

$$\rho = 1 - \left[\frac{6(80)}{12(144-1)}\right]$$

$$= 1 - 0.2797$$

$$= 0.7203$$

We say that there is high degree of positive rank correlation between the scores of selection and proficiency tests.

Example:

Following is the data on heights and weights of ten students in a class:

Heights (in cm)	140	142	140	160	150	155	160	157	140	170
Weights (in cm)	43	45	42	50	45	52	57	48	49	53

Calculate rank correlation coefficient between heights and weights of students.

Solution:

Let height be a variable X and weight be a variable Y. Since, the data contains tied observations, we associate average ranks to the tied observations. The spearman's rank correlation coefficient is given by

$$\rho = 1 - \left[\frac{6(Adj \sum d^2)}{n(n^2 - 1)} \right]$$

Where,

$$Adj \sum d^2 = \sum d^2 + \left[\frac{S_1^3 - S_1}{12} \right] + \left[\frac{S_2^3 - S_2}{12} \right] + \left[\frac{S_3^3 - S_3}{12} \right] + ...$$

N= No. of students

X	Y	R_x	R_y	d= R_x - R_y	d^2
140	43	9	9	0	0
142	45	7	7.5	-0.5	0.25
140	42	9	10	-1	1
160	50	2.5	4	-1.5	2.25
150	45	6	7.5	-1.5	2.25
155	52	5	3	2	4
160	57	2.5	1	1.5	2.25
157	48	4	6	-2	4
140	49	9	5	4	16
170	53	1	2	-1	1
				TOT	33

From the table, we have,

$$n = 10, \sum d_2 = 33, S_1 = 3, S_2 = 2, S_3 = 33$$

Thus,

$$Adj \sum d^2 = \sum d^2 + \left[\frac{S_1^3 - S_1}{12} \right] + \left[\frac{S_2^3 - S_2}{12} \right] + \left[\frac{S_3^3 - S_3}{12} \right] + ...$$

$$Adj \sum d^2 = 33 + \left[\frac{3^3 - 3}{12} \right] + \left[\frac{2^3 - 2}{12} \right] + \left[\frac{2^3 - 2}{12} \right] + ...$$

$$= 33+2+0.5+0.5$$

$$= 36$$

$$\rho = 1 - \left[\frac{6(36)}{10(100-1)} \right]$$

$$\rho = 1 - 0.2182$$

$$= 0.7818$$

We say that there is high degree of positive rank correlation between heights and weights of students.

Partial and Multiple Correlation

Let us consider the example of yield of rice in a firm. It may be affected by the type of soil, temperature, amount of rainfall, usage of fertilizers etc. It will be useful to determine how yield of rice is influenced by one factor or how yield of rice is affected by several other factors. This is done with the help of partial and multiple correlation analysis.

The basic distinction between multiple and partial correlation analysis is that in the former, the degree of relationship between the variable Y and all the other variables $X_1, X_2, ..., X_n$ taken together is measured, whereas, in the later, the degree of relationship between Y and one of the variables $X_1, X_2, ..., X_n$ is measured by removing the effect of all the other variables.

Partial correlation

Partial correlation coefficient provides a measure of the relationship between the dependent variable and other variable, with the effect of the rest of the variables eliminated. If there are three variables X_1, X_2 *and* X_3, there will be three coefficients of partial correlation, each studying the relationship between two variables when the third is held constant. If we denote by $r_{12.3,}$ that is, the coefficient of partial correlation X_1 *and* X_2 keeping X_3 constant, it is calculated as

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}, \quad r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}},$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}}$$

1. In a trivariate distribution, it is found that $r_{12} = 0.7$, $r_{13} = 0.61$ and $r_{23} = 0.4$. Find the partial correlation coefficients.

Solution:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \frac{0.7 - (0.61 \times 0.4)}{\sqrt{1-(0.61)^2}\sqrt{1-(0.4)^2}} = 0.628$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}} = \frac{0.61 - (0.7 \times 0.4)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.4)^2}} = 0.504$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}} = \frac{0.4 - (0.7 \times 0.61)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.61)^2}} = -0.048$$

Multiple Correlation

In multiple correlation, we are trying to make estimates of the value of one of the variable based on the values of all the others. The variable whose value we are trying to estimate is called the dependent variable and the other variables on which our estimates are based are known as independent variables.

The coefficient of multiple correlation with three variables X_1, X_2 *and* X_3 are $R_{1.23,}$ $R_{2.13}$ and $R_{3.21}$. $R_{1.23,}$ is the coefficient of multiple correlation related to X_1 as a dependent variable and X_2, X_3 as two independent variables and it can be expressed in terms of $r_{12,}$ r_{23} and r_{13} as

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}},$$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{13}^2}},$$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{12}^2}}$$

PROPERTIES OF MULTIPLE CORRELATION COEFFICIENT

The following are some of the properties of multiple correlation coefficients:

1. Multiple correlation coefficient is the degree of association between observed value of the dependent variable and its estimate obtained by multiple regression,
2. Multiple Correlation coefficient lies between 0 and 1.
3. If multiple correlation coefficient is 1, then association is perfect and multiple regression equation may said to be perfect prediction formula.
4. If multiple correlation coefficient is 0, dependent variable is uncorrelated with other independent variables. From this, it can be concluded that multiple regression equation fails to predict the value of dependent variable when values of independent variables are known.
5. Multiple correlation coefficient is always greater or equal than any total correlation coefficient. If $R_{1.23}$ is the multiple correlation coefficient than $R_{1.23} \geq r_{12}$ or r_{13} or r_{23} and
6. Multiple correlation coefficient obtained by method of least squares would always be greater than the multiple correlation coefficient obtained by any other method.

Example:

1. The following zero-order correlation coefficients are given: $r_{12} = 0.98, r_{13} = 0.44$ and $r_{23} = 0.54$. Calculate multiple correlation coefficient treating first variable as dependent and second and third variables as independent.

Solution:

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}}$$

$$= \sqrt{\frac{(0.98)^2 + (0.44)^2 - 2(0.98)(0.54)(0.44)}{1 - (0.54)^2}} = 0.986$$

2. From the following data, obtain $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$

X_1	2	5	7	11
X_2	3	6	10	12
X_3	1	3	6	10

Solution:

We need r_{12}, r_{13} and r_{23} which are obtained from the following table:

S. No	X_1	X_2	X_3	$(X_1)^2$	$(X_2)^2$	$(X_3)^2$	$X_1 X_2$	$X_1 X_3$	$X_2 X_3$
1	2	3	1	4	9	1	6	2	3
2	5	6	3	25	36	9	30	15	18
3	7	10	6	49	100	36	70	42	60
4	11	12	10	121	144	100	132	110	120
TOT	25	31	20	199	289	146	238	169	201

Now we get the total correlation coefficient r_{12}, r_{13} and r_{23}

$$r_{12} = \frac{N(\sum X_1X_2) - (\sum X_1)(\sum X_2)}{\sqrt{\{N(\sum X_1{}^2) - (\sum X_1)^2\}}\sqrt{\{N(\sum X_2{}^2) - (\sum X_2)^2\}}}$$

$r_{12} = 0.97$

$$r_{13} = \frac{N(\sum X_1X_3) - (\sum X_1)(\sum X_3)}{\sqrt{\{N(\sum X_1{}^2) - (\sum X_1)^2\}}\sqrt{\{N(\sum X_3{}^2) - (\sum X_3)^2\}}}$$

$r_{13} = 0.99$

$$r_{23} = \frac{N(\sum X_2X_3) - (\sum X_2)(\sum X_3)}{\sqrt{\{N(\sum X_2{}^2) - (\sum X_2)^2\}}\sqrt{\{N(\sum X_3{}^2) - (\sum X_3)^2\}}}$$

$r_{23} = 0.97$

Now, we calculate $R_{1.23}$

We have, $r_{12} = 0.97$, $r_{13} = 0.99$ and $r_{23} = 0.97$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}}$$

$R_{1.23} = 0.99$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{13}^2}}$$

$R_{2.13} = 0.97$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{12}^2}}$$

$R_{3.12} = 0.99$

Regression:

Regression is a mathematical measure of the average relationship between two or more variables in terms of the original limits of the data.

- **Lines of regression:**

1. The line of regression of Y on X is given by $y-\bar{y}=r.\frac{\sigma_Y}{\sigma_X}(x-\bar{x})$
2. The line of regression of X on Y is given by $x-\bar{x}=r.\frac{\sigma_X}{\sigma_Y}(y-\bar{y})$

- **Regression Coefficients:**

1. Regression coefficient of Y on X : $r.\frac{\sigma_Y}{\sigma_X}=b_{YX}$

   Where $b_{YX}=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$

2. Regression coefficient of X on Y : $r.\frac{\sigma_X}{\sigma_Y}=b_{XY}$

   Where $b_{XY}=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$

3. Correlation coefficient $r=\pm\sqrt{b_{XY}\times b_{YX}}$

Remarks:

1. The lines of regression of Y on X and X on Y passes through the mean value of x and y. In other words, the mean value of x and y can be obtained as the point of intersection of the two regression lines.

2. In case of perfect correlation. *(r* = ± 1), both the lines of regression coincide. Therefore. in general. we always have two lines pf regression except in the particular case of perfect correlation when both the lines coincide and we get only one line.

3. The sign of correlation coefficient is the same as that of regression coefficients, since the sign of each depends upon the co-variance term. Thus, if the regression coefficients are positive, *'r'* is positive and if the

regression coefficients are negative $'r'$ is negative.

4. If one of the regression coefficients is greater than unity. the other must be less than unity.

5. If the two variables are uncorrelated, the lines of regression become perpendicular to each other.

Problems:

1. From the following data find (i) two regression equations (ii) the coefficient of correlation (iii) Find Y when $X = 30$

X	25	28	35	32	31	36	29	38	34	32
Y	43	46	49	41	36	32	31	30	33	39

Solution:

X	Y	$X-\bar{X}$ $= X-32$	$Y-\bar{Y}$ $= Y-38$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})$ $(Y-\bar{Y})$
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
320	**380**	**0**	**0**	**140**	**398**	**-93**

Here, $\bar{X} = \frac{\sum X}{n} = \frac{320}{10} = 32$, $\bar{Y} = \frac{\sum Y}{n} = \frac{380}{10} = 38$

The line of regression of X on Y is given by $x - \bar{x} = b_{XY}(y - \bar{y})$

$$b_{XY} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2} = \frac{-93}{398} = -0.2337$$

$$\Rightarrow (x-32) = -0.2337 \quad (y-38)$$

$$= -0.2337y + 0.2337 \times 38$$

$$\Rightarrow x = -0.2337y + 40.8806$$

The line of regression of Y on X is given by $y - \bar{y} = b_{YX}(x - \bar{x})$

$$b_{YX} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{-93}{140} = -0.6643$$

$$\Rightarrow (y-38) = -0.6643 \quad (x-32)$$

$$= -0.6643x + 0.6643 \times 32$$

$$\Rightarrow y = -0.6643x + 59.2576$$

Coefficient of correlation $r^2 = b_{YX} \times b_{XY}$

$$= (-0.6643)\ (-0.2337) = 0.1552$$

$$r = \pm\sqrt{0.1552} = \pm 0.394$$

When X = 30, Y = (- 0.6643) (30) + 59.2576

Y = 39.3286

2. The two lines of regression are $8x - 10y + 66 = 0$, $40x - 18y - 214 = 0$. The variance of X is 9. Find the mean values of X and Y.

Solution:

Since both the lines of regression passes through the mean values $\bar{x}$ and $\bar{y}$, the point $(\bar{x}, \bar{y})$ must satisfy the two given regression lines .

$8\bar{x} - 10\bar{y} = -66$ .....(1)

$40\bar{x} - 18\bar{y} = 214$ .....(2)

Solving these (1) and (2) we get, $\bar{x} = 13$, $\bar{y} = 17$

3. Estimate the regression line from the given information:

Solution:

$$\sum_{i=1}^{33} x_i = 1104,\ \sum_{i=1}^{33} y_i = 1124,\ \sum_{i=1}^{33} x_i y_i = 41{,}355,\ \sum_{i=1}^{33} x_i^2 = 41{,}086$$

Therefore,

$$b_1 = \frac{(33)(41{,}355) - (1104)(1124)}{(33)(41{,}086) - (1104)^2} = 0.903643 \text{ and}$$

$$b_0 = \frac{1124 - (0.903643)(1104)}{33} = 3.829633.$$

Thus, the estimated regression line is given by

$$\hat{y} = 3.8296 + 0.9036x.$$

4. The two regression lines are given as x+2y-5=0 and 2x+3y-8=0. Which one is the regression line of x on y?

Suppose $x + 2y - 5 = 0$ ***is the equn. of the reg. line of x on y &*** $2x + 3y - 8 = 0$ ***is the equn. of the reg. line of y on x ,***

then the 2 equns can be written as $x = -2y + 5$

& $y = -\frac{2}{3}x + \frac{8}{3}$ ***Hence*** $b_{yx} = -\frac{2}{3}$ & $b_{xy} = -2$

Now $r^2 = \frac{4}{3} > 1$

This is impossible. Hence our assumption is wrong

$\therefore$ $2x + 3y - 8 = 0$ ***is the equn. Of reg. line of x on y***

5. The Two Lines of Regressions Are X + 2y – 5 = 0 and 2x + 3y – 8 = 0 and the Variance of X is 12. Find the Variance of Y and the Coefficient of Correlation.

Let $y = -\frac{1}{2}x + \frac{5}{2}$ be the regression line of y on x

and $x = -\frac{3}{2}y + \frac{8}{2}$ be the regression line of x on y

Now, $b_{yx} = -\frac{1}{2}$ $b_{yx} = -\frac{3}{2}$

$$\sqrt{b_{yx} \cdot b_{xy}} = \sqrt{\frac{-1}{2} \cdot \frac{-3}{2}}$$

$$= \sqrt{\frac{3}{4}} = \frac{-\sqrt{3}}{2} < 1$$

$r = \frac{-\sqrt{3}}{2}$

Now, $\sigma_x = \sqrt{12} = 2\sqrt{3}$

We have: $b_{yx} = r \frac{\sigma_y}{\sigma_x}$

$$-\frac{1}{2} = -\frac{\sqrt{3}}{2} \cdot \frac{\sigma_y}{2\sqrt{3}}$$

$\Rightarrow \sigma_y = 2$

$\therefore$ Variance of y =4

coefficient of correlation = $\frac{-\sqrt{3}}{2}$ ...(same sign as b_{yx} and b_{yx})

Advanced types of linear regression (not in Syllabus)

Linear models are the oldest type of regression. It was designed so that statisticians can do the calculations by hand. However, OLS (Ordinary Least squares)has several weaknesses, including a sensitivity to both outliers and multicollinearity, and it is prone to to overfitting. To address these problems, statisticians have developed several advanced variants:

- **Lasso regression** (least absolute shrinkage and selection operator) performs variable selection that aims to increase prediction accuracy by identifying a simpler model. It is similar to Ridge regression but with variable selection.

- **Ridge regression** allows you to analyse data even when severe multicollinearity is present and helps prevent overfitting. This type of model reduces the large, problematic variance that multicollinearity causes

by introducing a slight bias in the estimates. The procedure trades away much of the variance in exchange for a little bias, which produces more useful coefficient estimates when multicollinearity is present.

- **Partial least squares (PLS) regression** is useful when you have very few observations compared to the number of independent variables or when your independent variables are highly correlated. PLS decreases the independent variables down to a smaller number of uncorrelated components, similar to Principal Components Analysis. Then, the procedure performs linear regression on these components rather the original data. PLS emphasizes developing predictive models and is not used for screening variables. Unlike OLS, you can include multiple continuous *dependent* variables. PLS uses the correlation structure to identify smaller effects and model multivariate patterns in the dependent variables.

Practice Problem:

1. Find the regression equations for the following data:

X	1	3	5	7	9
Y	15	18	21	23	22

Solution: $x = 0.887y - 12.562$, $y = 0.95x + 15.05$

Multiple Regression

If the number of independent variables in a regression model is more than one, then the model is called as multiple regression. In fact, many of the real-world applications demand the use of multiple regression models.

Assumptions of multiple linear regression

Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.

Independence of observations: the observations in the dataset were collected using statistically valid methods, and there are no hidden relationships among variables.

In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated (r2 > ~0.6), then only one of them should be used in the regression model.

Normality: The data follows a normal distribution.

Linearity: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

Multiple linear Regression formula

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

- **y** = the predicted value of the dependent variable
- $\mathbf{b_0}$ = the y-intercept (value of y when all other parameters are set to 0)
- $\mathbf{b_1X_1}$= the regression coefficient (B_1) of the first independent variable ($\mathbf{X_1}$) (a.k.a. the effect that increasing the value of the independent variable has on the predicted **y** value)
- **…** = do the same for however many independent variables you are testing
- $\mathbf{b_nX_n}$ = the regression coefficient of the last independent variable

Application:

where Y represents the economic growth rate of a country, X_1 represents the time period, X_2 represents the size of the populations of the country, X_3 represents the level of employment in percentage, X_4 represents the percentage of literacy, b_0 is the intercept and b_1, b_2, b_3 and b_4 are the slopes of the variables X_1, X_2, X_3 and X_4 respectively. In this regression model, X_1, X_2, X_3 and X_4 are the independent variables and Y is the dependent variable.

Regression model with two independent variables using normal equations:

If the regression equation with two independent variables is

$$Y = b_o + b_1X_1 + b_2X_2$$

Then, the normal equations are

$$\Sigma Y = nb_o + b_1 \Sigma X_1 + b_2 \Sigma X_2$$

$$\Sigma Y X_1 = b_o \Sigma X_1 + b_1 \Sigma X_1^{\,2} + b_2 \Sigma X_1 X_2$$

$$\Sigma Y X_2 = b_o \Sigma X_2 + b_1 \Sigma X_1 X_2 + b_2 \Sigma X_2^{\,2}$$

Problems:

1. The annual sales revenue (in crores of rupees) of a product as a function of sales force (number of salesmen) and annual advertising expenditure (in lakhs of rupees) for the past 10 years are summarized in the following table.

Annual sales revenue Y	**20**	**23**	**25**	**27**	**21**	**29**	**22**	**24**	**27**	**35**
Sales force X1	**8**	**13**	**8**	**18**	**23**	**16**	**10**	**12**	**14**	**20**
Annual advertising expenditures X2	**28**	**23**	**38**	**16**	**20**	**28**	**23**	**30**	**26**	**32**

Let the regression model be

$$Y = b_o + b_1 X_1 + b_2 X_2$$

Y	**X_1**	**X_2**	**X^2_1**	**X^2_2**	**X_1X_2**	**YX_1**	**YX_2**
20	8	28	64	784	224	160	560
23	13	23	169	529	299	299	529
25	8	38	64	1444	304	200	950
27	18	16	324	256	288	486	432
21	23	20	529	400	460	483	420
29	16	28	256	784	448	464	812
22	10	23	100	529	230	220	506

24	12	30	144	900	360	288	720	
27	14	26	196	676	364	378	702	
35	20	32	400	1024	640	700	1120	
253	142	264	2246	7326	3617	3678	6751	**Total**

Substituting the required values in the normal equations, we get the following simultaneous equations

$$253 = 10b_o + 142b_1 + 264b_2$$

$$3678 = 142b_o + 2246b_1 + 3617b_2$$

$$6751 = 264b_o + 3617b_1 + 7326b_2$$

The solution to the above set of simultaneous equation is

$$b_o = 5.1483, \quad b_1 = 0.6190 \quad and \quad b_2 = 0.4304$$

Therefore, the regression model is $Y = 5.1483 + 0.6190X_1 + 0.4304X_2$

If mean, standard deviation and partial correlation of the trivariate distribution are known, then the multiple regression of X_1 on X_2 and X_3 is given by

$$(X_1 - \bar{X}_1)\frac{\omega_{11}}{\sigma_1} + (X_2 - \bar{X}_2)\frac{\omega_{12}}{\sigma_2} + (X_3 - \bar{X}_3)\frac{\omega_{13}}{\sigma_3} = 0$$

where $$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2$$

$$\omega_{12} = -\begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = r_{13}\, r_{23} - r_{21}$$

$$\omega_{13} = r_{23}\, r_{12} - r_{13}$$

Example :

Find the regression equation of X_1 on X_2 and X_3 given the following results :—

Trait	*Mean*	*Standard deviation*	r_{12}	r_{23}	r_{31}
X_1	28·02	4·42	+ 0·80	—	—
X_2	4·91	1·10	—	−0·56	—
X_3	594	85	—	—	− 0·40

where X_1 = *Seed per acre;* X_2 = *Rainfall in inches*

X_3 = *Accumulated temperature above 42°F.*

Solution:

Regression equation of X_1 on X_2 and X_3 is given by

$$(X_1 - \bar{X}_1)\frac{\omega_{11}}{\sigma_1} + (X_2 - \bar{X}_2)\frac{\omega_{12}}{\sigma_2} + (X_3 - \bar{X}_3)\frac{\omega_{13}}{\sigma_3} = 0$$

where $$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix}$$

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 = 1 - (-0{\cdot}56)^2 = 0{\cdot}686$$

$$\omega_{12} = -\begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = r_{13}\, r_{23} - r_{21} = -0{\cdot}576$$

$$\omega_{13} = r_{23}\, r_{12} - r_{13} = (-0{\cdot}56)\,(0{\cdot}80) - (-0{\cdot}40) = -0{\cdot}048$$

$\therefore$ Required equation of plane of regression of X_1 on X_2 and X_3 is given by

$$\frac{0{\cdot}686}{4{\cdot}42}(X_1 - 28{\cdot}02) + \frac{(-0{\cdot}576)}{1{\cdot}10}(X_2 - 4{\cdot}91) + \frac{(-0{\cdot}048)}{85{\cdot}00}(X_3 - 594) = 0$$

Practice Problems :

1.

Data Collected From Random Sample of 5 General Motors Salespeople

Independent Variable 1 (X1)	Independent Variable 2 (X2)	Dependent Variable (Y)
Highest Year of School Completed	Motivation as Measured by Higgins Motivation Scale	Annual Sales in Dollars
12	32	$350,000
14	35	$399,765
15	45	$429,000
16	50	$435,000
18	65	$433,000

Solution : r = 0.9360 .