Module-4

Correlation and Regression

Correlation:

In a bivariate distribution we have to find out the if there is any correlation or covariance between the two variables under study. If the change in one variable affects a change in the other variable, the variables are said to be correlated. If the two variables are deviate in the same direction, that is, if the increase (or decrease) in one result in a corresponding increase (or decrease) in the other, correlation is said to be positive. But, if they are constantly deviate in the opposite directions, that is if increase (or decrease) in one result in corresponding decrease (or increase) in the other, correlation is said to be negative.

Type of Correlation:

(a) Positive and Negative Correlation
(b) Linear and Non-linear Correlation

Positive and Negative Correlation:

If the values of the two variables deviate in the same direction, that is, if the increase of one variable results, on an average, in a corresponding increase in the values of the other variable or if decrease in the values of one variable results, on an average, in a corresponding decrease in the values of the other variable, correlation is said to be positive or direct.

Examples:

- Heights and weights
- Price and supply of a commodity
- The family income and expenditure on luxury items, etc.

On the other hand, correlation is said to be negative or inverse if the variables deviate in the opposite direction that is, if the increase or decrease in the values of one variable results, on the average, in a corresponding decrease or increase in the values of the other variable.

Examples:

- Price and demand of a commodity
- Volume and pressure of a perfect gas, etc.

Linear and Non-linear Correlation:

The correlation between two variables is said to be linear if corresponding to a unit change in one variable, there is a constant change in the other variable over the entire range of the values.

Example:

Let us consider the following data:

x	1	2	3	4	5
y	5	7	9	11	13

Thus, for a c unit change in the variable of x, there is constant change in the corresponding values of y. Mathematically, the above data can be expressed by the relation

$$y = 2x + 3$$

In general, two variables $x\ and\ y$ are said to be linearly related, if there exists a relationship of the form

$$y = a + bx \qquad (1)$$

between them. From eq. (1) of straight line with slope b and which makes an intercept a on the $y - axis$. Hence, if the values of the two variables are plotted as points in the $xy - plane$. Then we get a straight line.

The relationship between two variables is said to be non-linear or curvilinear if corresponding to unit change in one variable, the other variable does not change at a constant rate but at fluctuating rate. In such cases if the data are plotted on the $xy - plane$, we do not get a straight-line curve.

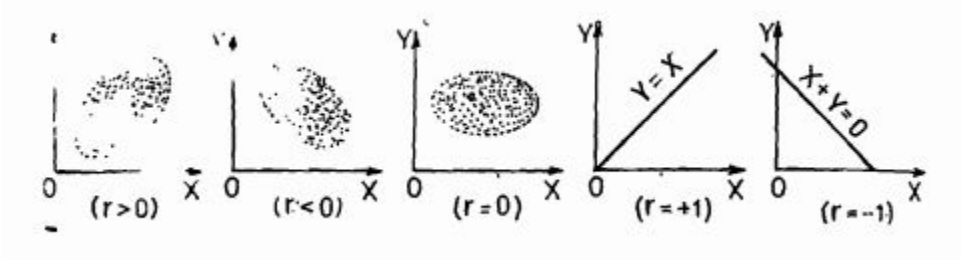
Methods of studying Correlation:

The methods of ascertaining only linear relationship between two variables. The commonly used methods for studying the correlation between two variables are:

(a) Scatter diagram or plot method
(b) Karl Pearson's coefficient of correlation (or Covariance method)
(c) Two-way frequency table (Bivariate correlation method)
(d) Rank correlation method

(a) Scatter diagram method:

If the simplest way of the diagrammatic representation of bivariate data. Thus, for the bivariate distribution $(x_i, y_i); i = 1,2,3 \ldots, n$, if the values of the variables $X\ and\ Y$ are plotted along $x - axis$ and $y - axis$ respectively in the $xy - plane$, diagram of dots so obtained is known as scatter diagram. From the scatter diagram, we can form a fairly good, whether the variables are correlated or not. For example, if the points are very dense, i.e., very close to each other, we should expect a fairly good amount of correlation is expected. This method, however, is not suitable if the number of observations is fairly large.



(b) Karl Pearson's coefficient of Correlation (Covariance method):

As a measure of intensity or degree of linear relationship between two variables, Karl Pearson's, a British Biometrician, developed a formula called Correlation coefficient.

Correlation coefficient between two variables $X\ and\ Y$, usually denoted by $r(X,Y)$ or simply r_{XY} or simply r, is a numerical measure of linear relationship between them and is defined as:

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \qquad (1)$$

Or

It is defined as the ratio of covariance between $X\ and\ Y$ say $Cov(X,Y)$ to the product of the standard deviations $X\ and\ Y$, say

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

If $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$ are n pairs of observations of the variables $X\ and\ Y$ in a bivariate distribution, then

$$Cov(x,y) = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y});\ \ \sigma_x = \sqrt{\frac{1}{n}\sum(x-\bar{x})^2},\ \ \sigma_y = \sqrt{\frac{1}{n}\sum(y-\bar{y})^2}$$

(2)

Summation being taken over n pairs of observations.

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

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \qquad (3)$$

Eq. (3) can also be written as

$$r = \frac{\sum dx\ dy}{\sqrt{\sum dx^2\ \sum dy^2}}$$

Where, $dx = x - \bar{x}\ \ and\ \ \ dy = y - \bar{y}$.

Problem 1:

Calculate Karl Pearson's coefficient of correlation between expenditure on advertising and sales from the data given below

Advertising expenses (thousands Rs.)	39	65	62	90	82	75	25	98	36	78
Sales (lakhs Rs.)	47	53	58	86	62	68	60	91	51	84

Solution:

Let the advertising expenses('000Rs.) be denoted by the variable x and the sales (in lakhs Rs.) be denoted by the variable y.

We have to find the Calculation for correlation coefficient

x	y	$dx = x - \bar{x} = x - 65$	$dy = y - \bar{y} = y - 66$	dx^2	dy^2	$dxdy$
39	47	-26	-19	676	361	494
65	53	0	-13	0	169	0
62	58	-3	-8	9	64	24
90	86	25	20	625	400	500
82	62	17	-4	289	16	-68
75	68	10	2	100	4	20
25	60	-40	-6	1600	36	240
98	91	33	25	1089	625	825
63	51	-29	-15	841	225	435
78	84	13	18	169	324	234

$\sum x$ =650	$\sum y$ = 660	$\sum dx$ = 0	$\sum dy$ = 0	$\sum dx^2$ = 5398	$\sum dy^2$ = 2224	$\sum dxdy$ = 2704

$$\sum \bar{x} = \frac{\sum x}{n} = \frac{650}{10} = 65$$

$$\sum \bar{y} = \frac{\sum y}{n} = \frac{660}{10} = 66$$

$$dx = x - \bar{x} = x - 65$$

$$dy = y - \bar{y} = y - 66$$

$$r = \frac{\sum dxdy}{\sqrt{\sum dx^2 \sum dy^2}} = \frac{2704}{\sqrt{5398 \times 2224}} = \frac{2704}{\sqrt{12005152}} = \frac{2704}{3464.8451} = 0.7804$$

Hence, there is a fairly high degree of positive correlation between expenditure on advertising sales. We may, therefore conclude that in general; sales have increased with an increase in the advertising expenditures.

Problem 2:

From the following table calculate the coefficient of correlation by Karl Pearson's method

X	6	2	10	4	8
Y	9	11	?	8	7

Arithmetic mean of X and Y series of 6 and 8 respectively.

Solution:

First of all, we shall find the missing value of . Let the missing value of Y series be a.Then the mean of $\bar{y}$ is given by:

$$\bar{y} = \frac{\sum y}{n} = \frac{9 + 11 + a + 8 + 7}{5} = \frac{35 + a}{5} = 8\ (given)$$

$$35 + a = 5 \times 8$$

$$a = 40 - 35 = 5$$

Now, we calculate the Correlation coefficient

X	Y	$X - \bar{X} = X - 6$	$Y - \bar{Y} = Y - 8$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$	$(X - \bar{X})\,(Y - \bar{Y})$
6	9	0	1	0	1	0
2	11	-4	3	16	9	-12
10	5	4	-3	16	9	-12
4	8	-2	0	4	0	0
8	7	2	-1	4	1	1
$\sum X$ =30	$\sum Y$ =40	0	0	$\sum(X - \bar{X})^2$ =40	$\sum(Y - \bar{Y})^2$ =20	$\sum(X - \bar{X})\,(Y - \bar{Y})$ =-26

$$\bar{X} = \frac{\sum x}{5} = \frac{30}{5} = 6$$

$$\bar{Y} = \frac{\sum y}{5} = \frac{40}{5} = 8$$

Karl Pearson's correlation coefficient is given by

$$r = \frac{COV(X,Y)}{\sigma_x \sigma_y} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}} = \frac{-26}{\sqrt{40 \times 20}} = \frac{-26}{\sqrt{800}} = \frac{-26}{28.2843} = -0.9192$$

$$r \approx -0.92$$

Problem 3:

Calculate the coefficient of correlation between $X$and Y series from the following data

	Series	
	X	Y
No. of series observations	15	15
Arithmetic mean	25	18
Standard deviation	3.01	3.03
Sum of squares of deviations from mean	136	138

Summation of product deviation of $X\ and\ Y$ series from their respective arithmetic mean=122.

Solution:

In the usual notations, we are given

$n = 15,\ \bar{x} = 25, \bar{y} = 18, \sigma_x = 3.01, \sigma_y = 3.03, \sum(x - \bar{x})^2 = 136,\ \sum(y - \bar{y})^2 = 138$ and $\sum(x - \bar{x})(y - \bar{y}) = 122.$

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{n\ \sigma_x \sigma_y} = \frac{122}{15 \times 3.01 \times 3.03} = 0.8917$$

Problem 4:

A computer while calculating correlation coefficient between two variables $X\ and\ Y$ from 25 pairs of observations obtained the following results:

$$n = 25, \sum X = 125, \sum X^2 = 650, \sum Y = 100, \sum Y^2 = 460,\ \sum XY = 508$$

It was, however, discovered at the time of checking that two pairs of observations were not correctly copied. They were taken as (6, 14) and (8, 6) while the correct values were (8, 12) and (6, 8). Prove that the correct value of the correlation coefficient should be 2/3.

Solution:

Corrected $\sum X = 125 - 6 - 8 + 8 + 6 = 125$

Corrected $\sum Y = 100 - 14 - 6 + 12 + 8 = 100$

Corrected $\sum X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Corrected $\sum Y^2 = 460 - 6^2 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

Corrected $\sum XY = 508 - (6 \times 14) - (8 \times 6) + (8 \times 12) + (6 \times 8) = 520$

Corrected value of r is given by

$$r = \frac{n\ \sum XY - \sum X\ \sum Y}{\sqrt{[n\ \sum X^2 - (\sum X)^2] \times [n\ \sum Y^2 - (\sum Y)^2]}}$$

$$r = \frac{(25 \times 520) - (125 \times 100)}{\sqrt{[25 \times 520 - 125^2] \times [25 \times 436 - 100^2]}} = \frac{2}{3}$$

Properties of correlation coefficient:

1. Pearson coefficient cannot exceed 1 numerically. In other words, it lies between -1 and +1 i.e., $-1 \leq r \leq 1$
2. Correlation coefficient is independent of the change of origin and scale. Mathematically, if X and y are the given variables and they are transformed to the new variables $u \; and \; v$ by the change of origin and scale
   $u = \frac{X-A}{h} \; and \; v = \frac{y-B}{k}, \; h > 0, k > 0.$

   Where, A, B, h and k are constants, $h > 0, k > 0$, then the correlation between $x \; and \; y$ is same the correlation coefficient between $u \; and \; v$ i.e., $r(x, y) = r(u, v)$

   $$r_{xy} = r_{uv}$$

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

   $$r_{uv} = \frac{n \sum uv - (\sum u)(\sum v)}{\sqrt{[n \sum u^2 - (\sum u)^2] \times [n \sum v^2 - (\sum v)^2]}}$$
3. Two independent variables are uncorrelated i.e., $r_{xy} = 0$.
4. $r(aX + b, cY + d) = \frac{a \times c}{|a \times c|} . r(X, Y)$

Example:

Calculate the coefficient of correlation for the ages of husbands and wives

Ages of husbands (years)	23	27	28	29	30	31	33	35	36	39
Ages of wives(years)	18	22	23	24	25	26	28	29	30	32

Solution:

x	y	u $= x - 31$	v $= y - 25$	u^2	v^2	uv
23	18	-8	-7	64	49	56
27	22	-4	-3	16	9	12
28	23	-3	-2	9	4	6
29	24	-2	-1	4	1	2
30	25	-1	0	1	0	0
31	26	0	1	0	1	0
33	28	2	3	4	9	6
35	29	4	4	16	16	16
36	30	5	5	28	25	25
39	32	8	7	64	49	56
$\sum$x =311	$\sum$y =257	$\sum u$ =1	$\sum v$ =7	$\sum$u^2 =203	$\sum$v^2 =163	$\sum$uv =179

Karl Pearson's correlation coefficient between u and v is given by

$$r_{uv} = \frac{n\sum uv - (\sum u)(\sum v)}{\sqrt{[n\sum u^2 - (\sum u)^2] \times [n\sum v^2 - (\sum v)^2]}}$$

$$= \frac{10 \times 179 - 1 \times 7}{\sqrt{[10 \times 203 - (1)^2] \times [10 \times 163 - (7)^2]}}$$

$$= \frac{1790 - 7}{\sqrt{[2030 - 1] \times [1630 - 49]}}$$

$$= \frac{1783}{\sqrt{2029 \times 1581}}$$

$$= \frac{1783}{45.04 \times 39.76}$$

$$= \frac{1783}{1790.79} = 0.9956$$

Since Karl Pearson's correlation coefficient (r) is independent of change of origin, we get

$$r_{xy} = r_{uv} = 0.9956$$

(c) **Rank Correlation method**:

Sometimes we come across statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This happens when we are dealing with qualitative characteristics (attributes) such as honesty, beauty,

character, morality, etc., which cannot be measured quantitatively but can be arranged serially. In such situations Karl Pearson's coefficient of correlation cannot be used as such. Charles Edward Spearman, a British psychologist, developed a formula in 1904 which consists in obtaining the correlation coefficient between the ranks of n individuals in the two attributes under study.

Suppose we want to find if two characteristics A, say, intelligence and B, say, beauty are related or not. Both the characteristics are incapable of quantitative measurements but we can arrange a group of n individuals in order of merit (ranks) w.r.t. proficiency in the two characteristics. Let the random variables X and Y denote the ranks of the individuals in the characteristics A and B respectively. If we assume that there is no tie, i.e., if no two individuals get the same rank in a characteristic then, obviously, X and Y assume numerical values ranging from 1 to n.

The Pearsonian correlation coefficient between the ranks X and Y is called the rank correlation coefficient between the characteristics A and B for that group of individuals.

Spearman's rank correlation coefficient, usually denoted by ρ (Rho) is given by the formula

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

Where d is the difference between the pair of ranks of the same individual in the two characteristics and n is the number of pairs.

Computation of rank correlation coefficient:

We shall discuss below the method of computing the Spearman's rank correlation coefficient ρ under the following situations:

I. When actual ranks are given
II. When ranks are not given

Case I: When actual ranks are given:

In this situation the following steps are involved:

i. Compute d, the difference of ranks.

ii. Compute d^2
iii. Obtain the sum $\sum d^2$
iv. Use formula (1) to get the value of ρ.

Example.

The ranks of the same 15 students in two subjects A and B are given below:

the two numbers within the brackets denoting the ranks of the same student in A and B respectively. (1,10), (2,7), (3,2), (4,6), (5,4), (6,8), (7,3), (8,1), (9,11), (10,15), (11,9), (12,5), (13,14), (14,12), (15,13).

Use Spearman's formula to find the rank correlation coefficient.

Solution:

Rank in A (x)	Rank in B (y)	$d = x - y$	d^2
1	10	-9	81
2	7	-5	25
3	2	1	1
4	6	-2	4
5	4	1	1
6	8	-2	4
7	3	4	16
8	1	7	49
9	11	-2	4
10	15	-5	25
11	9	2	4
12	5	7	49
13	14	-1	1
14	12	2	4
15	13	2	4
		$\sum d = 0$	$\sum d^2 = 272$

Spearman's rank correlation coefficient ρ (Rho) is given by

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

$$= 1 - \frac{6 \times 272}{15(225-1)} = 1 - \frac{17}{35} = \frac{18}{35} = 0.51$$

Example:

Calculate Spearman's rank correlation coefficient between advertisement cost and sales from the following data,

Advertising cost (thousands Rs.)	39	65	62	90	82	75	25	98	36	78
Sales (lakhs Rs.)	47	53	58	86	62	68	60	91	51	84

Solution:

Let X denotes the advertising cost('000Rs.) and Y denotes the Sales (lakhs Rs.).

X	Y	Rank of $X(x)$	Rank of $Y(y)$	$d = x - y$	d^2
39	47	8	10	-2	4
65	53	6	8	-2	4
62	58	7	7	0	0
90	86	2	2	0	0
82	62	3	5	-2	4
75	68	5	4	1	1
25	60	10	6	4	16
98	91	1	1	0	0
63	51	9	9	0	0
78	84	4	3	1	1
				$\sum d$ =0	$\sum d^2$ =30

Here $n = 10$

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 30}{10 \times 99} = 1 - \frac{2}{11} = \frac{9}{11} = 0.82.$$

Example:

Find the rank correlation coefficient from the following data

Ranks in X	1	2	3	4	5	6	7
Ranks in Y	4	3	1	2	6	5	7

Solution:

In this problem ranks are not repeated

x	y	$d_i = x_i - y_i$	${d_i}^2$
1	4	-3	9
2	3	-1	1
3	1	2	4
4	2	2	4
5	6	-1	1
6	5	1	1
7	7	0	0
			$\sum {d_i}^2 = 20$

In this problem ranks are not repeated, so the rank correlation coefficient is

$$r(x, y) = \rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6 \times 20}{7(7^2 - 1)} = 0.6429$$

Example:

Calculate the rank correlation coefficient from the following data, which give the ranks of 10 students in Mathematics and Computer Science

Mathematics (x)	1	5	3	4	7	6	10	2	9	8
Computer Science(y)	6	9	1	3	5	4	8	2	10	7

Solution:

x	y	$d_i = x_i - y_i$	${d_i}^2$
1	6	-5	25
5	9	-4	16
3	1	2	4

4	3	1	1
7	5	2	4
6	4	2	4
10	8	2	4
2	2	0	0
9	10	-1	1
8	7	1	1
			$\sum d_i^{\,2} = 60$

In this problem ranks are not repeated, so the rank correlation coefficient is

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

$$= 1 - \frac{6 \times 60}{10(10^2-1)} = 0.63636$$

Try yourself:

The ranks of same 16 students in mathematics and physics are as follows. Calculate rank correlation coefficients for proficiency in mathematics and physics

Mathematics (x)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Physics (y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13

Example:

Ten competitors in a beauty contest are ranked by three judges in the following order

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

Use the rank correlation coefficient to determine which pair of judges has the nearest approach in common tastes in beauty.

Solution:

Let R_1, R_2 and R_3 denote the ranks given by the first, second and third judges respectively and let ρ_{ij} be the rank correlation coefficient between the ranks given by ith and jth judges $i \neq j = 1,2,3$. Let $d_{ij} = R_i - R_j$, be the difference of ranks of an individual given by the ith and jth judge.

R_1	R_2	R_3	$d_{12} = R_1 - R_2$	$d_{13} = R_1 - R_3$	$d_{23} = R_2 - R_3$	${d_{12}}^2$	${d_{13}}^2$	${d_{23}}^2$
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	-9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	1	2	1	1	1
			$\sum d_{12} = 0$	$\sum d_{13} = 0$	$\sum d_{23} = 0$	$\sum {d_{12}}^2 = 200$	$\sum {d_{13}}^2 = 60$	$\sum {d_{23}}^2 = 214$

We have $n = 10$

Spearman's rank correlation coefficient ρ is given by

$$\rho_{12} = 1 - \frac{6\sum {d_{12}}^2}{n(n^2-1)} = 1 - \frac{6 \times 200}{10 \times 99} = -\frac{7}{33} = -0.2121$$

$$\rho_{13} = 1 - \frac{6\sum {d_{13}}^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{10 \times 99} = \frac{7}{11} = 0.6363$$

$$\rho_{23} = 1 - \frac{6\sum {d_{23}}^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = -\frac{49}{165} = -0.2970$$

Since ρ_{13} is maximum, the pair of first and third judges has the nearest approach to common tastes in beauty.

Remark, since ρ_{12} and ρ_{23} are negative, the pair of judges (1,2) and (2,3) have opposite (divergent) tastes for beauty.

Case II: When ranks are not given

Spearman's rank correlation formula can also be used even if we are dealing with variables which are measured quantitatively, i.e., when the actual data but not the ranks relating to two variables are given. In such a case we shall have to convert the data into ranks. The highest (smallest) observation is given the rank 1. The next highest (next lowest) observation is given rank 2 and so on. It is immaterial in which way (descending or ascending) the ranks are assigned. However, the same approach should be followed for the all the variables under consideration.

Repeated ranks:

In case of attributes if there is a tie i.e., if any two or more individuals are placed together in any classification with respect to an attribute or if in case of variable data there is more than one item with the same value in either or both the series, then Spearman's formula for calculating the rank correlation coefficient breaks down, since in this case the variables X [the ranks of individuals in characteristic A (1^{st} series)] and Y [the ranks of individuals in characteristic B (2^{nd} series) do not take the values from 1 to n and consequently $\bar{x} \neq \bar{y}$, while Spearman's formula proving we had assumed that $\bar{x} = \bar{y}$.

In this case, common ranks are assigned to the repeated items. These common ranks are the arithmetic mean of the ranks which these items should have got if they are different from each other and the next item will get the rank next to the rank used computing the common rank.

For example, suppose an item is repeated at rank 4. The common rank common rank to be assigned to each item is (4+5)/2 i.e., 4.5 which is the average of 4 and 5, the ranks which these observations would have assigned if they were different. The next item will be assigned the rank 6. If an item is repeated thrice at rank 7, then the common rank to be assigned to each value will e (7+8+9)/3 i.e., 8 which is arithmetic mean of 7, 8 and 9. The ranks these observations would have got if they were different from each other. The next rank to be assigned will be 10.

In the Spearman's formula add the factor $\frac{m(m^2-1)}{12}$ to $\sum d^2$, where m is the number of times is repeated. This correction factor is to be added for each repeated value in both the series.

Problem:

A psychologist wanted to compare two methods A and B of teaching. He selected a random sample of 22 students. He grouped them into 11 pairs so the students in a pair have approximately equal scores on an intelligence test. In each pair one student was taught by method A and the other by method B and examined after the course. The marks obtained by them are tabulated below:

Pair	1	2	3	4	5	6	7	8	9	10	11
A	24	29	19	14	30	19	27	30	20	28	11
B	37	35	16	26	23	27	19	20	16	11	21

Find the rank correlation coefficient.

Solution:

In the X-series, we seen that the value 30 occurs twice. The common rank assigned to each of these values is 1.5, the arithmetic mean of 1 and 2, the ranks which these which observations would have taken if they were different. The next value 29 gets the next i.e. rank 3. Again, the value 19 occurs twice. The common rank assigned to it as 8.5, the arithmetic mean of 8 and 9 and the next value, 14 gets the rank 10. Similarly, in the y-series the value 16 occurs twice and the common rank assigned to each is 9.5, the arithmetic mean of 9 and 10, the next value, 11 gets the rank 11.

X	Y	Rank of X (x)	Rank of Y (y)	d=x-y	d^2
24	37	6	1	5	25
29	35	3	2	1	1
19	16	8.5	9.5	-1	1
14	26	10	4	6	36
30	23	1.5	5	-3.5	12.25
19	27	8.5	3	5.5	30.25
27	19	5	8	-3	9

30	20	1.5	7	-5.5	30.25
20	16	7	9.5	-2.5	6.25
28	11	4	11	-7	49
11	21	11	6	5	25
				$\sum d = 0$	$\sum d^2 = 225$

Hence, we see that in the X-series the items 19 and 30 are repeated, each occurring twice and, in the Y-series in the item 16 is repeated. Thus, in each of the three cases $m = 2$. Hence on applying the correction factor $\frac{m(m^2-1)}{12}$ for each repeated item, we get

$$\rho = 1 - \frac{6\left[\sum d^2 + 2\left(\frac{4-1}{12}\right) + 2\left(\frac{4-1}{12}\right) + 2\left(\frac{4-1}{12}\right)\right]}{11(121-1)}, \quad \text{here n=11}$$

$$\rho = 1 - \frac{6 \times 226.5}{11 \times 120} = 1 - 1.0225 = -0.0225$$

Problem:

A sample of 12 fathers and their eldest sons have the following data about their heights in inches.

Fathers (x)	65	63	67	64	68	63	70	66	68	67	69	71
Sons (y)	68	66	68	65	69	66	68	65	71	67	68	70

Calculate the rank correlation coefficient.

Solution:

Fathers (x)	Sons (y)	Rank of x	Rank of y	$d = x - y$	d^2
65	68	9	5.5	3.5	12.25
63	66	11	9.5	1.5	2.25
67	68	6.5	5.5	1	1
64	65	10	11.5	-1.5	2.25

68	69	4.5	3	1.5	2.25
62	66	12	9.5	2.5	6.25
70	68	2	5.5	-3.5	12.25
66	65	8	11.5	-3.5	12.25
68	71	4.5	1	3.5	12.25
67	67	6.5	8	-1.5	2.25
69	68	3	5.5	2.5	6.25
71	70	1	2	-1	1
				$\sum d$ =0	$\sum d^2$ =72.5

Correlation factors

In x, 68 is repeated twice, then $\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{1}{2}$

In x, 67 is repeated twice, then $\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{1}{2}$

In y, 67 is repeated 4 times, then $\frac{m(m^2-1)}{12} = \frac{4(4^2-1)}{12} = 5$

In y, 66 is repeated twice, then $\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{1}{2}$

In y, 65 is repeated twice, then $\frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{1}{2}$

Rank correlation is

$$\rho = 1 - \frac{6\left[\sum d^2 + \frac{1}{2} + \frac{1}{2} + 5 + \frac{1}{2} + \frac{1}{2}\right]}{12(144-1)} = 0.722$$

Linear Regression:

If the variables in bivariate distribution are related, will find that the points in the scatter diagram will cluster round some curve called the ‘’curve of regression’’. If the curve is a straight

line, it is called the line of regression and there is said to be linear regression between the variables, otherwise regression is said to be curvilinear.

The lines of regression are the line which gives to be best estimate to the value of one variable for any specific value of the other variable. Thus, the line of regression is the line of 'best fit' and is obtained by the principle of least squares.

Let us suppose that the in the bivariate distribution $(x_i, y_i); i = 1,2,3, \dots, n;$ y is dependent variable and x is independent variable. Let the line of regression is the line of y on x be

$$y = a + bx \tag{1}$$

Eq. (1) represents the family of straight lines for different values of the arbitrary constants $'a' and\ 'b'$. The problem is to determine the $'a' and\ 'b$ so that the line Eq. (1) is the line of best fit.

According to the principle of the principle of least squares, we have to determine $'a' and\ 'b'$.

$$E = \sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2$$

Is minimum. From the principle of maxima and minima, the partial derivatives of E, with respect to $'a' and\ 'b'$ should vanish separately, i.e.,

$$\frac{\partial E}{\partial a} = 0 = -2\sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)$$

$$\sum_{i=1}^{n} y_i = an + \sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i \tag{2}$$

$$\frac{\partial E}{\partial b} = 0 = -2\sum_{i=1}^{n} x_i\left(x_i - (a + bx_i)\right)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} {x_i}^2 \tag{3}$$

Dividing on both sides to the Eq. (2) by n, we get

$$\bar{y} = a + b\bar{x} \tag{4}$$

Now, the line of regression of $Y\ on\ X$ passes through the point $(\bar{x}, \bar{y}\,)$.

$$\mu_{11} = Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i = \mu_{11} + \bar{x}\bar{y} \quad (5)$$

$$\sigma_x{}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i{}^2 - \bar{x}^2$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i{}^2 = \sigma_x{}^2 + \bar{x}^2 \quad (6)$$

Dividing Eq. (3) by n and using Eqs. (5) and (6), we get

$$\mu_{11} + \bar{x}\bar{y} = a\bar{x} + b(\sigma_x{}^2 + \bar{x}^2) \quad (7)$$

Eq. (7)- Eq. (4)$\times\ \bar{x}$, we get

$$\mu_{11} = b\sigma_x{}^2$$

$$b = \frac{\mu_{11}}{\sigma_x{}^2}$$

Since 'b' is the slope of the line of regression of Y on X and since the line of regression passes through the point $(\bar{x}, \bar{y})$ its equation is

$$Y - \bar{y} = b(x - \bar{x}) = \frac{\mu_{11}}{\sigma_x{}^2}(X - \bar{x})$$

$$Y - \bar{y} = r\,\frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

Starting the equation $X = A + BY$ and proceeding similarly, we get

$$X - \bar{x} = r\,\frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

Problem:

From the following data, obtain the two regression equations

Sales	91	97	108	121	67	124	51	73	111	57

Purchases	71	75	69	97	70	91	39	61	80	47

Solution:

Let us denote the sales by the variable x *and* y the purchases by the variable y

x	y	$dx = x - 90$	$dy = y - 70$	dx^2	dy^2	$dx\ dy$
91	71	1	1	1	1	1
97	75	7	5	49	25	35
108	69	18	-1	324	1	-18
121	97	31	27	961	729	837
67	70	-23	0	529	0	0
124	91	34	21	1156	441	714
51	39	-39	-31	1521	961	1209
73	61	-17	-9	289	81	153
111	80	21	10	441	100	210
57	47	-33	-23	1089	529	759
$\sum x = 900$	$\sum y = 700$	$\sum dx = 0$	$\sum dy = 0$	$\sum dx^2 = 6360$	$\sum dy^2 = 2868$	$\sum dx\ dy = 3900$

We have, $\bar{x} = \frac{\sum x}{n} = \frac{900}{10} = 90$

$$\bar{y} = \frac{\sum y}{n} = \frac{700}{10} = 70$$

$$b_{yx} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum dxdy}{\sum dx^2} = \frac{3900}{6360} = 0.6132$$

$$b_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2} = \frac{\sum dxdy}{\sum dy^2} = \frac{3900}{2868} = 1.361$$

Equation of regression of y *on* x is

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

$$y - 70 = 0.6132(x - 90)$$

$$y - 70 = 0.6132\ x - 0.613 \times 90$$

$$= 0.6132\ x - 55.188$$

$$y = 0.6132\ x - 55.188 + 70$$

$$y = 0.6132\ x + 14.812$$

Equation of regression of x on y is

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

$$x - 90 = 1.361(y - 70)$$

$$x - 90 = 1.361\ \ y - 1.361 \times 70$$

$$= 1.361\ y - 95.27$$

$$x = 1.361\ y - 95.27 + 90$$

$$x = 1.361\ y - 5.27$$

$$r^2 = b_{yx}\ .b_{xy}$$

$$r^2 = 0.6132 \times 1.361 = 0.8346$$

$$r = \mp 0.9135$$

But since, both the regression coefficients are positive, r must be positive.

$$r = 0.9135$$

Problem:

From the data given below find

(a) Two regression coefficients
(b) The two regression equations
(c) The coefficient of correlation between the marks in Economics and Statistics
(d) The most likely marks in Statistics when marks in Economics are 30.

Marks in Economics	25	28	35	32	31	36	29	38	34	32
Marks in Statistics	43	46	49	41	36	32	31	30	33	39

Solution:

x	y	$dx = x - 32$	$dy = y - 38$	dx^2	dy^2	$dx\ dy$
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
$\sum x = 320$	$\sum y = 380$	$\sum dx = 0$	$\sum dy = 0$	$\sum dx^2 = 140$	$\sum dy^2 = 398$	$\sum dx\,dy = -93$

$$\bar{x} = \frac{\sum x}{n} = \frac{320}{10} = 32$$

$$\bar{y} = \frac{\sum y}{n} = \frac{380}{10} = 38$$

(a) **Regression coefficients**:

Coefficient of regression y *on* x is given by

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

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

(b) Equations of the line of regression of x *on* y is

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

$$x - 32 = -0.2337(y - 38)$$

$$x - 32 = -0.2337\,y + 38 \times 0.2337$$

$$= -0.2337\,y + 8.8806$$

$$x = -0.2337\,y + 8.8806 + 32$$

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

Equation of line of regression of y *on* x is

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

$$y - 38 = -0.6643(x - 32)$$

$$y - 38 = -0.6643\,x + 0.6643 \times 32$$

$$= -0.6643\,x + 0.6643 \times 32 + 38$$

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

(c) Correlation coefficient:

$$r^2 = b_{yx} \, . b_{xy}$$

$$r^2 = (-0.6643)(-0.2337) = 0.1552$$

$$r = \mp 0.394$$

Since the both regression coefficients are negative. Hence the discarding plus sign, we get

$$r = -0.394$$

(d) In order to estimate the most likely marks in Statistics (y) when marks in Economics (x) are 30, we use the line of regression of $y \; on \; x$.

The equation is

$$y - 38 = -0.6643 \; (30) + 59.2576$$

$$y = 39.3286$$

Hence the most likely marks in Statistics when in Economics are 30, are 39.3286$\approx$ 39.

Problem:

The following is an estimated supply regression for sugar:

$y = 0.025 + 1.5 \; x$, where y is supply in kilos and x is price in rupees per kilo.

(a) Interpret the coefficient of variable x
(b) Predict the supply when supply when price is Rs. 20 per kilo
(c) Given that $r(x, y) = 1$, interpret the implied relationship between price and quality supplied.

Solution:

The regression equation of y (supply in kgs) on x (price in rupees per kg) is given to be

$$y = 0.025 + 1.5 \; x = a + bx \text{ (say)} \qquad (1)$$

(a) The coefficients of variation x
$b = 1.5$ is the coefficient of regression of $y \; on \; x$. It reflects the unit change in the value of y, for a unit change in the corresponding value of x. This means that if the price of sugar goes up by Re. 1 per kg, the estimated supply of sugar goes up by 1.5 kg.

(b) From eq. (1), the estimated supply of sugar when its price is Rs. 20 per kg is given by

$y = 0.025 + 1.5 \times 20 = 30.025$ kg

(c) $r(x, y) = 1$

The relationship between that $x\ and\ y$ is exactly linear. i.e., all the observed values (x, y) lies on straight line.

Problem:

Given that the regression equations of $y\ on\ x$ and of $x\ on\ y$ are respectively $y = x$ and $4x - y = 3$, and that the second moment of x about the origin is 2, find

(a) The correlation coefficient between $x\ and\ y$
(b) The standard deviation of y

Solution:

Regression equation of $y\ on\ x$ is $y = x$

$$b_{yx} = 1$$

Regression equation of $x\ on\ y$ is $4x - y = 3$

$$x = \frac{1}{4}y + \frac{3}{4}$$

$$b_{xy} = \frac{1}{4}$$

(a) The correlation coefficient between $x\ and\ y$ is

$$r^2 = b_{yx}\ .b_{xy}$$

$$r^2 = 1 \times \frac{1}{4} = \frac{1}{4}$$

$$r = \mp 0.5$$

Since the both the regression coefficients are positive $r = 0.5$.

(b) We are given that the second moment of x about origin is 2. i.e., $\frac{\sum x^2}{n} = 2$

Since $(\bar{x}, \bar{y})$ is the point of intersection of the two lines of regression

Solving $y = x$ and $4x - y = 3$, then $x = 1 = y$

$$\bar{x} = 1\ and\ \bar{y} = 1$$

$$\sigma_x{}^2 = \frac{\sum x^2}{n} - \bar{x}^2 = 2 - 1 = 1$$

Also, $b_{yx} = r\ \frac{\sigma_y}{\sigma_x}$

$$1 = \frac{1}{2}\left(\frac{\sigma_y}{1}\right)$$

$$\sigma_y = 2$$

Coefficient of Determination:

Coefficient of correlation between two variable series is a measure of linear relationship between them and indicates the amount of variation of one variable which is associated with or accounted for by another variable. A more useful and readily comprehensible measure for this purpose is the coefficient of determination which gives the percentage variation in the dependent variable that is accounted for by the independent variable.

In other words, the coefficient of determination gives the ratio of the explained variance to the total variance. The coefficient is given by the square of the correlation coefficient i.e.,

$r^2 = \frac{explained\ variance}{total\ variance}$.

Ex:

If the value of $r = 0.8$, we cannot conclude that 80% of the variation in the relative series (dependent variable) is due to the variation in the subject series (independent variable). But the coefficient of determination in this case $r^2 = 0.64$ which implies that only 64% of the variation in the relative series has been explained by the subject series and the remaining 36% of the variation is due to other factors.

Coefficient of Partial correlation:

Sometimes the correlation between **two variables $X_1 and\ X_2$ may be partly due to the correlation of third variable X_3 with both $X_1\ and\ X_2$**. In such a situation, one may want to know what the correlation between $X_1 and\ X_2$would be if the effect of X_3 an each of $X_1 and\ X_2$ were eliminated. **This correlation is called partial correlation and the correlation coefficient between $X_1 and\ X_2$ after the linear effect of X_3 on each of them has been eliminated is called the partial coefficient.**

The partial correlation coefficient between $X_1 and\ X_2$, usually denoted by $r_{12.3}$ is given by

$$r_{12.3} = \frac{r_{12} - r_{13}\ r_{23}}{\sqrt{(1 - r_{13}{}^2)(1 - r_{23}{}^2)}}$$

Similarly,

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{32}}{\sqrt{(1 - {r_{12}}^2)(1 - {r_{32}}^2)}}$$

and

$$r_{23.1} = \frac{r_{23} - r_{21}\, r_{31}}{\sqrt{(1 - {r_{21}}^2)(1 - {r_{31}}^2)}}$$

Multiple correlation in terms of total and partial correlations:

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

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

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

Note:

$$1 - {R_{1.23}}^2 = \frac{\omega}{\omega_{11}}$$

Where, $\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 1 - {r_{12}}^2 - {r_{13}}^2 - {r_{23}}^2 + 2r_{12}r_{13}r_{23}$

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

Problem:

From the data relating to the yield of dry bark (X_1), height (X_2) and girth (X_3) for 18 cinchona plants, the following correlation coefficients were obtained:

$r_{12} = 0.77, r_{13} = 0.72\ and\ r_{23} = 0.52$. Find the partial correlation coefficients $r_{12.3}$ and multiple correlation coefficient $R_{1.23}$.

Solution:

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

$$= \frac{0.77 - 0.72 \times 0.52}{\sqrt{(1 - 0.72^2)(1 - 0.52^2)}} = 0.62$$

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

$$= \frac{0.77^2 + 0.72^2 - 2 \times 0.77 \times 0.72 \times 0.52}{1 - 0.52^2} = 0.7334$$

$R_{1.23} = +0.8564$ (since multiple correlation is non-negative).

Problem:

In a trivariate distribution $\sigma_1 = 2, \sigma_2 = \sigma_3 = 3, r_{12} = 0.7, r_{23} = r_{31} = 0.5$.

Find (i) $r_{23.1}$ (ii) $R_{1.23}$ (iii) $b_{12.3}$, $b_{13.2}$ and (iv) $\sigma_{1.23}$.

Solution:

$$r_{23.1} = \frac{r_{23} - r_{21}\, r_{31}}{\sqrt{(1 - {r_{21}}^2)(1 - {r_{31}}^2)}}$$

$$= \frac{0.5 - 0.7 \times 0.5}{\sqrt{(1 - 0.7^2)(1 - 0.5^2)}} = 0.2425$$

(ii)

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

$${R_{1.23}}^2 = \frac{0.7^2 + 0.5^2 - 2 \times 0.7 \times 0.5 \times 0.5}{1 - 0.5^2} = 0.52$$

$$R_{1.23} = +0.7211$$

(iii)

$$b_{12.3} = r_{12.3}\left(\frac{\sigma_{1.3}}{\sigma_{2.3}}\right) \text{ and } b_{13.2} = r_{13.2}\left(\frac{\sigma_{1.2}}{\sigma_{3.2}}\right) \qquad (1)$$

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - {r_{13}}^2)(1 - {r_{23}}^2)}} = 0.6$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{32}}{\sqrt{(1 - {r_{12}}^2)(1 - {r_{32}}^2)}} = 0.2425$$

$$\sigma_{1.3=}\, \sigma_1 \sqrt{(1 - {r_{13}}^2)} = 2 \times \sqrt{(1 - 0.5^2)} = 1.7320$$

$$\sigma_{2.3=}\, \sigma_2 \sqrt{(1 - {r_{23}}^2)} = 3 \times \sqrt{(1 - 0.5^2)} = 2.5980$$

$$\sigma_{1.2=}\, \sigma_1 \sqrt{(1 - {r_{12}}^2)} = 2 \times \sqrt{(1 - 0.7^2)} = 1.4282$$

$$\sigma_{3.2=}\, \sigma_3 \sqrt{(1 - {r_{32}}^2)} = 2 \times \sqrt{(1 - 0.5^2)} = 2.5980$$

Eq. (1) gives $b_{12.3} = 0.6 \times \frac{1.7320}{2.5980} = 0.4$ and $b_{13.2} = 0.2425 \times \frac{1.4282}{2.5980} = 0.1333$

(iv) $$\sigma_{1.23} = \sigma_1 \left(\sqrt{\frac{\omega}{\omega_{11}}}\right)$$

$$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 1 - {r_{12}}^2 - {r_{13}}^2 - {r_{23}}^2 + 2 r_{12} r_{13} r_{23} = 0.36$$

and $\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - {r_{23}}^2 = 1 - 0.5^2 = 0.75$

$$\sigma_{1.23} = 2 \times \left(\sqrt{\frac{0.36}{0.75}}\right) = 1.3856$$

Multiple regression:

Bivariate Regression equation:

→ Here we try to study the linear relationship between two variables x and y.

$$\mathrm{Y} = \mathrm{a} + \mathrm{bX} = \beta_0 + \beta_1 X$$

We see that the $a = \beta_0$ is the Y intercept, $b = \beta_1$ is the slope of the linear relationship between the variable X and Y.

Multivariate regression equation

- $\mathrm{Y} = \mathrm{a} + \mathrm{b}_1\mathrm{X}_1 + \mathrm{b}_2\mathrm{X}_2 = \beta_0 + \beta_1\mathrm{X}_1 + \beta_2\mathrm{X}_2$

- $b_1 = \beta_1$ = partial slope of the linear relationship between the first independent variable and Y, indicates the change in Y for one unit change in X_1.
- $b_2 = \beta_2$ = partial slope of the linear relationship between the second independent variable and Y, indicates the change in Y for one unit change in X_2.

Formulas for finding partial slopes:

$$b_1 = \beta_1 = \frac{S_y}{S_1}\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)$$

$$b_2 = \beta_2 = \frac{S_y}{S_2}\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)$$

S_y= standard deviation of Y

S_1 = standard deviation of the first independent variable (X_1)

S_2 = standard deviation of the second independent variable (X_2)

r_{y1} = bivariate correlation between Y and X1

r_{y2} = bivariate correlation between Y and X2

r_{12} = bivariate correlation between X_1 and X_2

Example:

1) The salary of a person in an organisation has to be regressed in terms of experience (X_1) and mistakes (X_2). If it is given that the values

$$\bar{Y} = 3.3;\ \overline{X_1} = 2.7;\ \overline{X_2} = 13.7$$
$$S_y = 2.1; S_1 = 1.5;\ S_2 = 2.6$$

and the zero order correlations :

$$r_{y1} = 0.5;\ r_{y2} = -0.3;\ r_{12} = -0.47;$$

Find the linear regression and interpret the results.

So,

$$b_1 = \beta_1 = \frac{S_y}{S_1}\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)$$

$$b_1 = \beta_1 = \frac{2.1}{1.5}\left(\frac{0.50 - (-0.3)(-0.47)}{1 - (-0.47)^2}\right) = 0.65$$

Similarly,

$$b_2 = \beta_2 = \frac{S_y}{S_2}\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)$$

$$b_2 = \beta_2 = \frac{2.1}{2.6}\left(\frac{0.30 - (0.5)(-0.47)}{1 - (-0.47)^2}\right) = -0.07$$

Calculation of a:

$$a = \bar{Y} - b_1\overline{X_1} - b_2\overline{X_2}$$
$$\beta_0 = \bar{Y} - \beta_1\overline{X_1} - \beta_2\overline{X_2}$$
$$= 3.3 - (0.65)(2.7) - (-0.07)\ 13.7$$
$$= 2.5$$

Interpretation:

1) If a person has no experience and has not done any mistakes, he would get a salary of 2.5 units.
2) If the experience goes up by 1 unit, there would be an increment in the salary by 0.65 units.
3) If he/ she commits a mistake, then the salary would decrease by 0.07 units.