CORRELATION AND REGRESSION

2.1. Introduction:

Let us consider two characteristics X and Y which are numerically measurable. Assume X denotes the height and Y denote corresponding weight of college students. For a set of students, when we are recording their height (X) and weight (Y), we get two values for an individual. One value corresponds to the height and the other value corresponds to the weight of that individual. We record the data for that individual as an ordered pair. The procedure repeats for all the students and finally we get a set of ordered pairs on X and Y. We call such a data on two characteristics as bivariate data. To analyze whether there is any relation between these characteristics, there are two distinct aspects for the study. One is correlation analysis and the other is regression analysis. Correlation analysis is to determine the degree of linear relationship between the characteristics X and Y. Regression analysis is to establish the nature of linear relationship between the characteristics. A simple method to get a rough idea on correlation and regression of the two characteristics considered is scatter method.

2.2. Scatter Diagram:

Let $(x_1 , y_1) , (x_2 , y_2) ,\ldots, (x_n , y_n)$ be the set of observations obtained on the two characteristics X and Y. A diagram obtained by plotting these values $(x_1 , y_1) , (x_2 , y_2) ,\ldots, (x_n , y_n) ,$ on a graph is called *scatter diagram.* It consists of points scattered over the graph.
Consider the following scatter diagrams obtained by plotting the observations regarding to some X and Y.

In the first scatter diagram *(figure 1)*, we can observe that all the points are almost scattered around a straight line. Also the line is of the form, as X increases Y also increases. Then we can roughly say, there exist a positive linear relation between X and Y. Since the points are closely clustered around the straight line, there is a high degree of linear relation.

In the second diagram *(figure 2)*, also we observe that all the points are almost scattered around a straight line. But the line is of the form, as X increases Y decreases. Then we can suspect there exist a negative linear relation between X and Y. Here also, the points are closely clustered around the straight line. Hence the degree of linear relation is high.

In the next scatter diagram *(figure 3)*, no specific relation between X and Y is observed. Then one can infer that there is no correlation between X and Y.

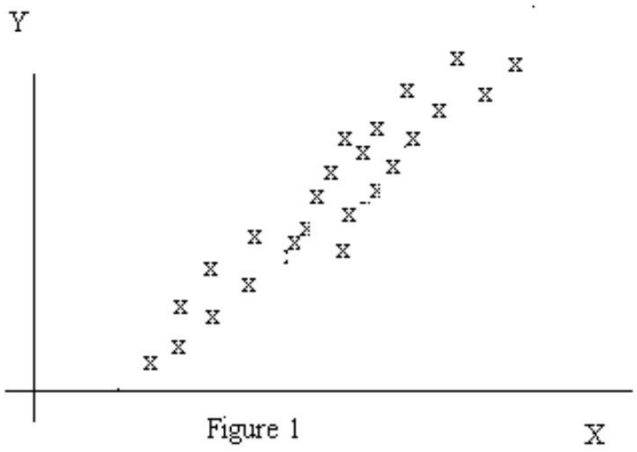

Figure 1

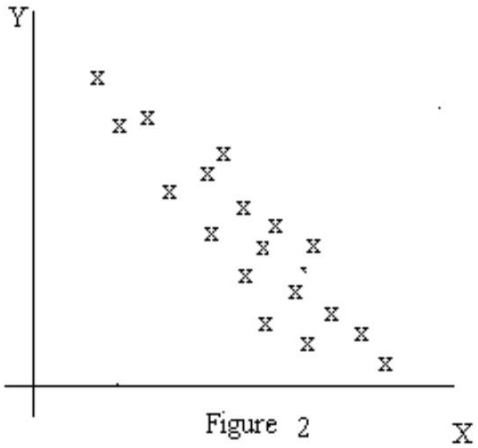

Figure 2

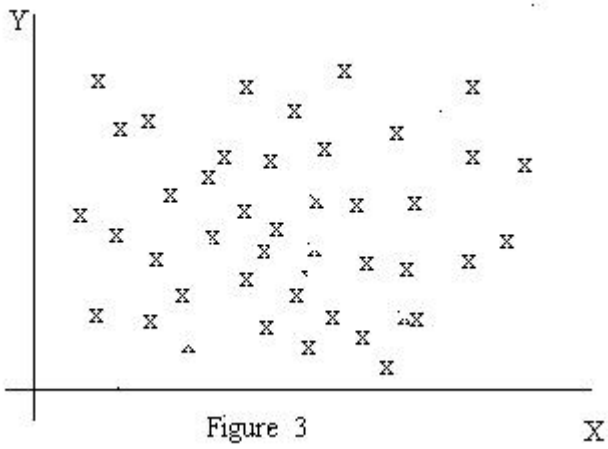

Figure 3

2.3. Curve Fitting:

We have a set of observations $(x_1 , y_1) , (x_2 , y_2) ,\ldots, (x_n , y_n)$ on two variables (characteristics) X and Y. If we feel that there is some relation between these two variables, let it be of the form, $y_i = f (x_i , a_1 , a_2 ,..., a_n) + \varepsilon_i$. $a_1 , a_2 ,..., a_n$ are the constants involved known as *parameters* and ε is the error term known as *residual error.* For example, if we feel a linear relation of the form $y = ax + b$, we can express it as $y_i = f (x_i , a , b) + \varepsilon_i$ for the point (x_i , y_i) . The relation $y = ax + b$ is only our assumption regarding the relation between X and Y. Hence all the (x_i , y_i) points may not strictly obey the relation. Using the assumed relation $y = ax + b$, between X and Y, we can calculate the value of y_i corresponds to given values of x_i . The difference between the given y_i values for a x_i value and the calculated y_i values for a x_i value using the proposed relation is the residual error. That is why we express the y_i value as $y_i = f (x_i , a , b) + \varepsilon_i$. Hence the error involved in the y_i value is $\varepsilon_i = y_i - f (x_i , a , b)$.

In general consider the relation between X and Y of the form, $y_i = f(x_i, a_1, a_2, ..., a_n) + \varepsilon_i$. Then the residual error on y_i value $\varepsilon_i = y_i - f(x_i, a_1, a_2, ..., a_n)$. To identify the relation between X and Y in terms of the parameters $a_1, a_2, ..., a_n$, it is to estimate the values of these parameters. The best values of $a_1, a_2, ..., a_n$ are those values of $a_1, a_2, ..., a_n$ which makes the residual errors minimum. The process of determining the best values of the parameters $a_1, a_2, ..., a_n$, statistically known as *curve fitting*. The values of the parameters are estimated using the *Principle of least squares*.

The Principle of least squares states that the best estimates of $a_1, a_2, ..., a_n$ are those values of $a_1, a_2, ..., a_n$ which minimize the sum of squares of the residual errors for all y_i values. Then it is to find the values of $a_1, a_2, ..., a_n$ which minimizing $E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - f(x_i, a_1, a_2, ..., a_n)]^2$.

The values of $a_1, a_2, ..., a_n$ which minimizes E can be obtained by solving the following equations, $\frac{\partial E}{\partial a_1} = 0$, $\frac{\partial E}{\partial a_2} = 0, \ldots, \frac{\partial E}{\partial a_n} = 0$.

These equations are known as *normal equations.*

- **Fitting of a straight line** $y = ax + b$

Consider $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the observations taken. It is to fit a maximum suitable straight line for the given data. That is to estimate the best values of the parameters involved *a* and *b*. By the Principle of least squares, the best values of *a* and *b* are those values of *a* and *b* which minimizes E, where,

$$E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - f(x_i, a, b)]^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

The normal equations are $\frac{\partial E}{\partial a} = 0$ and $\frac{\partial E}{\partial b} = 0$

$$\frac{\partial E}{\partial a} = 0 \Rightarrow \frac{\partial}{\partial a} \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 = 0$$

$$\Rightarrow -2 \sum_{i=1}^{n} [y_i - (ax_i + b)] x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i \quad ----(1)$$

$$\frac{\partial E}{\partial b} = 0 \Rightarrow \frac{\partial}{\partial b} \sum_{i=1}^{n} [\, y_i - (ax_i + b)\,]^2 = 0$$

$$\Rightarrow -1 \sum_{i=1}^{n} [\, y_i - (ax_i + b)] = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} x_i + n \times b \quad ----(2)$$

Solving (1) and (2) using the given data, the best estimates of *a* and *b* can be obtained.

If the given X, Y values are big values, to make the calculations easy, transform X, Y values to U, V values in the form, make a transformation on X, Y values to reduce them $u = \frac{x-a}{b}$ and $v = \frac{y-c}{d}$. Then fit a line of the form $v = a \cdot u + b$ and hence re substitute u and v to get the required relation in terms of X and Y.

Problem: *Fit a straight line to the following data*

x :	*3*	*4*	*5*	*6*	*7*
y :	*4*	*5*	*6*	*8*	*10*

Solution:

Consider the straight line of the form y = ax +b. To find the bst values of *a* and *b* by using the normal equations,

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

And $$\sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} x_i + nb$$

The calculations are as follows

x	y	x^2	xy
3	4	9	12
4	5	16	20
5	6	25	30
6	8	36	48
7	10	49	70
$\sum x$ =25	$\sum y$ =33	$\sum x_2$ =135	$\sum xy$ =180

The normal equation corresponds to the given data are,

$$180 = 135\,a + 25\,b \quad \text{-----(1)}$$

$$33 = 25\,a + 5b \quad \text{---- (2)}$$

From (2) we get, $165 = 125\,a + 25\,b$ ---(3)

(1) – (3) gives, $15 = 10a \quad \Rightarrow a = \frac{3}{2} = 1.5$

$a = \frac{3}{2}$ in (2) $\Rightarrow 33 = \frac{3}{2} \times 25 + 5b = -0.9$

Hence the required straight line fitted is , $y = 1.5x - 0.9$

- **Fitting of a curve** $y = ax^2 + bx + c$

Let (x_1, y_1) , (x_2, y_2) ,…, (x_n, y_n) are the given data. To fit the given curve, it is to estimate the values of *a, b and c*. By the principle of least squares the best estimates are those values of *a, b and c* which minimizing E. Here $E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - (ax_i^2 + bx_i + c))^2$

The normal equations are $\frac{\partial E}{\partial a} = 0$, $\frac{\partial E}{\partial b} = 0$ and $\frac{\partial E}{\partial c} = 0$. On differentiation the normal equations becomes;

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 \quad ----(1)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i \quad ----(2)$$

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + n \times c \quad ----(3)$$

Solve these normal equations using the given data to get the values of *a, b and c*.

While solving problems, appropriate transformations, if required to reduce calculations, can be done as illustrated in the case of fitting of a straight line.

Problem: *Fit a parabola of the form* $y = ax^2 + bx + c$ *to the following data:*

x :	*1960*	*1962*	*1964*	*1966*	*1968*
y :	*125*	*140*	*165*	*195*	*230*

Solution:

Let the equation of the parabola is in the form $y = a + bx + cx^2$.

To identify the best values of *a,b* and *c,* we use the following normal equations

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \quad ----(1)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \quad ----(2)$$

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \quad ----(3)$$

But here the given values of the variables are huge numbers. So first we transform x and y to some new variable u and v then, fit a parabola for u and v. Using this, derive the parabola for x and y. The working methods are shown below

x	y	u= $\frac{x-1964}{2}$	v= $\frac{y-165}{5}$	u^2	u^3	u^4	uv	u^2 v
1960	125	-2	-8	4	-8	16	16	-32
1962	140	-1	-5	1	-1	1	5	-1
1964	165	0	0	0	0	0	0	0
1966	195	1	6	1	1	1	6	6
1968	230	2	13	4	8	16	26	52
		0	6	10	0	34	53	25

The normal equations in terms of u and v are,

$$\sum_{i=1}^{n} u_i^2 v_i = a\sum_{i=1}^{n} u_i^2 + b\sum_{i=1}^{n} u_i^3 + c\sum_{i=1}^{n} u_i^4$$

$$\sum_{i=1}^{n} u_i v_i = a\sum_{i=1}^{n} u_i + b\sum_{i=1}^{n} u_i^2 + c\sum_{i=1}^{n} u_i^3$$

$$\sum_{i=1}^{n} v_i = na + b\sum_{i=1}^{n} u_i + c\sum_{i=1}^{n} u_i^2$$

Corresponds to the given data, these normal equation are,

$$25 = 10\,a + 0b + 34c \quad ---(1)$$

$$53 = 0\,a + 10b + 0c \quad ---(2)$$

$$6 = 5\,a + 0b + 10c \quad ---(3)$$

$$(2) \Rightarrow b = \frac{53}{10} = 5.3$$

$$(1) - 2(3) \Rightarrow 13 = 14\,c \Rightarrow c = \frac{13}{14} = 0.929$$

Then, $5a = 6 - 10(0.929) \Rightarrow a = -0.658$

Now the parabola is, $v = -0.658 + 5.3u + 0.929u^2$

Substitute u and v as $\frac{x-1964}{2}$ and $\frac{y-165}{5}$ respectively,

We get, $\frac{y-165}{5} = -0.658 + 5.3\left(\frac{x-1964}{2}\right) + 0.929\left(\frac{x-1964}{2}\right)^2$

$$y - 165 = -3.29 + 26.5\left(\frac{x-1964}{2}\right) + 4.645\left(\frac{x^2 - 3928x + 3857296}{4}\right)$$

$$\Rightarrow \quad y = 1.161x^2 - 4547.15x + 4410689.85$$

- **Fitting of a curve** $y = ab^x$

Taking logarithm to the base 10 on both sides, the curve $y = ab^x$ becomes, $\log y = \log a + x \log b$. Let $Y = \log y$, $A = \log a$ and $B = \log b$. Now the required curve is of the form, $Y = A + Bx$ or $Y = Bx + A$. If we are given x and Y values, it is easy to estimate the parameters A and B, using the method of fitting a straight line. Hence we can obtain a and b as the antilogarithm of A and B respectively.

To fit a curve of the form $y = ab^x$ for the given set of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, get Y_i values by taking the logarithm of the given y_i values. Using the x_i and Y_i values, solve the following normal equations for estimating A and B,

$$\sum_{i=1}^{n} x_i Y_i = B\sum_{i=1}^{n} x_i^2 + A\sum_{i=1}^{n} x_i \quad ----(1) \text{ and}$$

$$\sum_{i=1}^{n} Y_i = B\sum_{i=1}^{n} x_i + nA \quad ----(2)$$

Solve (1) and (2) to obtain A and B, then by taking antilogarithm of A and B we get a and b. Hence the curve $y = ab^x$ is fitted.

- **Fitting of a curve** $y = ax^b$

After taking logarithm on both sides the curve $y = ax^b$ also can be converted in the form of a straight line. That is, the curve becomes, $\log y = \log a + b \log x$. Let $Y = \log y$; $X = \log x$ and $A = \log a$; then the curve becomes, $Y = A + bX$ or $Y = bX + A$. Using the given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ values, obtain X_i *and* Y_i values taking logarithm on x_i *and* y_i values. Then solving the following normal equations A and b can be solved.

$$\sum_{i=1}^{n} X_i Y_i = b \sum_{i=1}^{n} X_i^2 + A \sum_{i=1}^{n} X_i \quad ----(1) \text{ and}$$

$$\sum_{i=1}^{n} Y_i = b \sum_{i=1}^{n} X_i + nA \quad ----(2)$$

The value of a is obtained by taking the antilogarithm of A. Hence the required curve is fitted.

- **Fitting of a curve** $y = ae^{bx}$

The method illustrated above can be used in the case of fitting of $y = ae^{bx}$ also. Taking logarithm on both sides, the curve becomes, $\log y = \log a + x \times b \log e$. Let $Y = \log y$; $A = \log a$ and $B = b \log e$, we get, $Y = A + Bx$ or $Y = Bx + A$. From the given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ values, taking the logarithm of y_i values Y_i values are obtained. Then use the following normal equations to obtain A and B.

$$\sum_{i=1}^{n} x_i Y_i = B \sum_{i=1}^{n} x_i^2 + A \sum_{i=1}^{n} x_i \quad ----(1) \quad \text{and}$$

$$\sum_{i=1}^{n} Y_i = B \sum_{i=1}^{n} x_i + nA \quad ----(2)$$

Now, a is the antilogarithm of A and $b = \dfrac{B}{\log e}$.

Problem: *for the data given below, find the equation to he fitting exponential curve of the form* $y = ae^{bx}$

x :	1	2	3	4	5	6
y :	1.6	4.5	13.8	40.2	125	300

Solution:

Taking logarithm to base 10 on both sides, the given curve $y = ae^{bx}$ is in the form, $\log y = \log a + xb \log c$

This is in the form, $Y = A + Bx$

where, $Y = \log y$, $A = \log a$, $B = b \log c$

Using the values of Y and x, we can fit the line, $Y = A + Bx$, that is we can find best values for A and B. Using these values of A and B, we can get the values of a and b.

For an easiness in calculation we transform x to u, where $u = x - 3$

Now using u and Y, fit a line of the form $Y = A' + B'u$, using the normal equations,

$$\sum_{i=1}^{n} u_i Y_i = B' \sum_{i=1}^{n} u_{i\,2} + A' \sum_{i=1}^{n} u_i \qquad \text{and}$$

$$\sum_{i=1}^{n} Y_i = B' \sum_{i=1}^{n} U_i + nA'$$

The calculations are as follows,

x	y	u=x-3	$Y = \log_{10} y$	uY	u^2
1	1.6	-2	0.204	-0.408	4
2	4.5	-1	0.653	-0.653	1
3	13.8	0	1.140	0	0
4	40.2	1	1.604	1.604	1
5	125	2	2.097	4.194	4
6	300	3	2.477	7.431	9
		3	8.175	12.168	19

Here the normal equations for u and Y are

$12.168 = 3A' + 19B'$ and $8.175 = 6A' + 3B'$

Solving these two equations we get, $A' = 1.13$ and $B' = 0.46$

Hence the line connecting u and Y is $Y = 1.13 + 0.46u$

That is $Y = 1.13 + 0.46(x - 3) \Rightarrow Y = -0.25 + 0.46x$

This implies, A = -0.25 and B = 0.46

That is $\log_{10} a = -0.25$ and $b \log_{10} c = 0.46$

From here we get a = 0.557 and b = 1.06

Hence the required curve is,

$$y = 0.557e^{1.06x}.$$

Problem: *Fit a curve of the form* $y = ax^b$ *for the following data*

x :	*66*	*64*	*55*	*51*	*42*	*32*	*24*
y :	*2.5*	*7.5*	*12.5*	*17.5*	*25*	*40*	*75*

Solution:

Taking logarithm on both sides of the required curve, $y = ax^b$, we get,

$\log y = \log a + b \log x$. This is in the form $Y = A + bX$, where $Y = \log y$, $A = \log a$, and $X = \log x$.

The calculations are:

x	y	$X = \log x$	$Y = \log y$	XY	X^2
66	2.5	1.8195	0.3979	0.7239	3.3106
64	7.5	1.8061	0.8751	1.5805	3.2619
55	12.5	1.7403	1.0969	1.9089	3.0286
51	17.5	1.7075	1.2430	2.1224	2.9156
42	25	1.6232	1.3979	2.2690	2.6347
32	40	1.5051	1.6021	2.4113	2.2653
24	75	1.3802	1.8750	2.5879	1.9049
		$\sum X$ =11.5819	$\sum Y$ =8.4879	$\sum XY$ =13.6036	$\sum X_2$ =19.3216

The normal equations for $Y = A + bX$ are,

$$\sum_{i=1}^{n} X_i Y_i = b \sum_{i=1}^{n} X_i^2 + A \sum_{i=1}^{n} X_i \quad \text{, and}$$

$$\sum_{i=1}^{n} Y_i = b \sum_{i=1}^{n} X_i + nA$$

Here the normal equations are,

13.6036 = 19.3216 b + 11.5819 A ---- (1)

8.4879 = 11.5819 b + 7 A ----- (2)

Solving these normal equations, we get, b = -2.773 and A = 5.8008

From A = 0.48, we get a = Anti log (A) = Anti log(5.8008) = 632120.68

Hence the required curve is ,

$$y = 632120.68\, x^{-2.773}\ .$$

Problem: *Fit a curve of the form* $y = ab^x$ *for the following data*

x :	*2*	*3*	*4*	*5*	*6*
y :	*144*	*172.8*	*207.4*	*248.8*	*298.6*

Solution:

Taking logarithm on both sides of the required curve, $y = ab^x$, we get $\log y = \log a + x \log b$. This is in the form $Y = A + B\,x$, where $Y = \log y$, $A = \log a$, and $B = \log b$.

Using the values of $Y = \log y$ and x , we can find the best values of A and B using the normal equations for fitting the line $Y = A + B\,x$. From this the value of a and b can be solved.

The calculations are:

x	y	$Y = \log y$	xY	x^2
2	144	2.16	4.32	4
3	172.8	2.24	6.72	9
4	207.4	2.32	9.28	16
5	248.8	2.40	12	25
6	298.6	2.47	14.82	36
$\sum x = 20$		$\sum Y = 11.59$	$\sum xY = 47.14$	$\sum x^2 = 90$

The normal equations for $Y = A + bX$ are,

$$\sum_{i=1}^{n} X_i Y_i = B\sum_{i=1}^{n} x_i^2 + A\sum_{i=1}^{n} x_i \quad \text{, and} \quad \sum_{i=1}^{n} Y_i = B\sum_{i=1}^{n} x_i + nA$$

Here the normal equations are, $47.14 = 90\ B + 20\ A$ --- (1)

$$11.59 = 20\ B + 5\ A \quad \text{--- (2)}$$

$$(1) - 4 \times (2) \Rightarrow 10\ B = 0.78 \quad \Rightarrow B = 0.078\ .$$

Solving (2) using $B = 0.078$, get $A = 2.006$.

Then, $a = Anti\log(2.006) = 101.3$, and $b = Anti\log(0.078) = 1.196$

Hence the required curve is,

$$y = (101.3) \times (1.196)^{x}$$

2.4. Regression lines:

Let (x_1, y_1) , (x_2, y_2) ,…, (x_n, y_n) be the given set of observations on two variables X and Y. A scatter plot of these points reveals an idea on the linear relation between X and Y. If a linear relation exists between X and Y, the line about which the points in the scatter diagram cluster is called the regression line and the equation representing this line is called the regression equation. There are two approaches for finding the regression line.
One is fitting a straight line of the form $y = ax + b$ to the given data (x_1, y_1) , (x_2, y_2) ,…, (x_n, y_n) , by minimizing the sum of squares of possible errors in y values. The other is fitting a straight line of the form $x = cy + d$ to the data, by minimizing the sum of squares of possible errors in x values. If all the given (x_i, y_i) values are perfectly obeys a linear relation, then the straight line fitted by the above two approaches will be same. But in general the (x_i, y_i) values may not perfectly obey a linear relation, and hence the above approaches may give two different straight lines for the given data. The straight line fitted to the data in the form $y = ax + b$ by minimizing the sum of squares of possible errors in y values is known as the *regression line y on x* and the straight line fitted to the data in the form $x = cy + d$ by minimizing the sum of squares of possible errors in x values is known as the *regression line x on y*.

To obtain the regression line Y on X of the form $y = ax + b$ for the given data (x_1, y_1) , (x_2, y_2) ,…, (x_n, y_n) , the following normal equations for fitting $y = ax + b$ are to be solved.

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i \quad \text{----(1) and}$$

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + n \times b \text{----(2)}$$

Let us transform x and y to X and Y as, $X = x - \bar{x}$ and $Y = y - \bar{y}$, where $\bar{x}$ and $\bar{y}$ are the means of x and y respectively. Now the normal equations for fitting a straight line = connecting X and Y in the form $Y\ aX + b$ are:

$$\sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_{i\,2} + b\sum_{i=1}^{n} X_i \quad ----(3) \text{ and}$$

$$\sum_{i=1}^{n} Y_i = a\sum_{i=1}^{n} X_i + n \times b ----(4)$$

But here, $\sum_{i=1}^{n} X_i = \sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n}(y_i - \bar{y}) = 0$

Hence,

$$(3) \Rightarrow \sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_{i2} + b \times 0$$

$$\Rightarrow \quad a = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^{\,2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})_2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})_2}$$

That is $a = \frac{Cov(x, y)}{\text{var}(x)}$

$$(4) \Rightarrow 0 = a \times 0 + n \times b \Rightarrow b = 0 .$$

Then, the straight line is, $Y = \frac{Cov(x, y)}{\text{var}(x)} X + 0$.

Hence the *regression line y on x* is, $(y - \bar{y}) = \frac{Cov(x, y)}{\text{var}(x)}(x - \bar{x})$.

In as similar way, the *regression line x on y* is derived as,

$$(x - \bar{x}) = \frac{Cov(x, y)}{\text{var}(y)}(y - \bar{y})$$

In the regression line *y on x*, the coefficient of *x*, $\frac{Cov(x, y)}{\text{var}(x)} = \frac{\rho_{xy}}{\sigma_x^2}$ is known as the *regression coefficient of y on x*, denoted by b_{yx} and in the regression line *x on y*, the coefficient of *y*, $\frac{Cov(x, y)}{\text{var}(y)} = \frac{\rho_{xy}}{\sigma_y^2}$ is known as the *regression coefficient of x on y*, denoted by b_{xy} .

The regression line *y on x* help us to predict the value of *y* for a given value of *x*, and the regression line *x on y* helps to predict the value of *x* for a given value of *y*.

Problem: *Obtain the line of regression of 'y on x' for the following data.*

Age x :	66	38	56	42	72	36	63	47	55	45
BP :	145	124	147	125	160	118	149	128	150	124

Estimate the blood pressure of a man whose age is 55.

Solution:

The regression line y on x is defined as,

$(y-\bar{y}) = \frac{P_{x,y}}{\sigma_x^2}(x-\bar{x})$, where $P_{x,y}$ = cov(X,Y), σ_x^2 = V(X).

Using the given data to find mean of x, mean of y, cov(X,Y) and V(X).

The calculations are as follows:

x	y	x^2	xy
66	145	4356	9570
38	124	1444	4712
56	147	3136	8232
42	125	1764	5250
72	160	5184	11520
36	118	1296	4248
63	149	3969	9387
47	128	2209	6016
55	150	3025	8250
45	124	2025	5580
520	1370	28408	72765

Mean of X = $\frac{520}{10} = 52$, Mean of Y = $\frac{1370}{10} = 137$

Cov (X,Y) = $\frac{1}{n}\sum xy - \bar{x}\times\bar{y} = \frac{72765}{10} - 52\times 137 = 152.5$

$$V(X) = \frac{1}{n}\sum x^2 - \bar{x}^2 = \frac{28408}{10} - 52^2 = 136.8$$

Hence the regression line of y on x is,

$$(y - 137) = \frac{152.5}{136.8}(x - 52) \quad \Rightarrow y = 1.1148\,(\quad)\,x + 79.03$$

Then the blood pressure of a man whose age x = 55 can be get by substituting x = 55 in the derived regression equation y on x, This implies, the blood pressure,

$$y = (1.1148) \times 55 + 79.03 \quad = 140.34\,.$$

Problem: *For 10 observations on X and Y, the following data were observed*

$$\sum x = 130\,,\ \sum y = 200\,,\ \sum x^2 = 2288\,,\ \sum y^2 = 5506\,,\ \sum xy = 3467$$

Obtain regression line of Y on X. Find y when x = 16.

Solution:

The regression line y on x is, $(y - \bar{y}) = \frac{P_{x,y}}{\sigma_x^2}(x - \bar{x})$, where $P_{x,y}$ = cov(X,Y), σ_x^2 = V(X)

$$Cov(X, Y) \equiv \frac{1}{n}\sum xy - \bar{x}\,\bar{y}$$

$$= \frac{1}{10}\left(\quad\right) - \left(\frac{130}{10}\right)\left(\frac{200}{10}\right) = 86.7 \quad 3467$$

$$V(X) = \frac{1}{n}\sum x^2 - \bar{x}^2 = \frac{1}{10}(2288) - \left(\frac{130}{10}\right)^2 = 59.8$$

The regression line Y on X is, $\left(y - \frac{200}{10}\right) = \frac{86.7}{59.8}\left(x - \frac{130}{10}\right)$

$$\Rightarrow y = 1.4498\,x + 1.1526\,.$$

When x = 16, we get,

$$y = 1.4498 \times 16 + 1.1526 = 24.3494\,.$$

2.5. Pearson's Coefficient of correlation:

If there is a linear relation between the variables x and y, the degree of linear relation is measured by the coefficient of correlation. If all they given (x_i , y_i) points are almost satisfying a linear relation, then we are saying that there is a high degree of linear relation between the variables. If the linear relation fitted for the variables is in such a

way that the increment in one variable results in the increment of the other also, then there is a direct (or positive) correlation existing between the variables. On the other hand, if the linear relation fitted for the variables is in such a way that the increment in one variable results in the decrease of the other, and then there is an inverse (or negative) correlation existing between the variables. If there is no linear relation existing between the variables, the correlation is zero.

A famous British Statistician, Karl Pearson suggested a coefficient measure of the degree of correlation between two variables *x and y,* known as Pearson's coefficient of correlation is denoted by r_{xy} , where,

$$r_{xy} = \frac{\mu_{xy}}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2 - (\bar{y})^2}}$$

Theorem: For two variable x and y, $-1 \le r_{xy} \le +1$, where r_{xy} is the Pearson's coefficient of correlation.

Proof:

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the observations on x and y. Consider $\frac{(x_i - \bar{x})}{\sigma_x}$ and $\frac{(y_i - \bar{y})}{\sigma_y}$, where $\bar{x}$ and $\bar{y}$ are the means and σ_x and σ_y are the standard deviations of x and y respectively.

We have, $\left[\frac{(x_i - \bar{x})}{\sigma_x} \pm \frac{(y_i - \bar{y})}{\sigma_y}\right]^2 \ge 0$, because it is the square of a real number.

Adding all such terms for i=1,2,…,n and dividing by n,

$$\Rightarrow \frac{1}{n}\sum_i \left[\frac{(x_i - \bar{x})}{\sigma_x} \pm \frac{(y_i - \bar{y})}{\sigma_y}\right]^2 \ge 0$$

On expansion, $\Rightarrow \frac{1}{n}\sum_i \left[\frac{(x_i - \bar{x})^2}{\sigma_x^2}\right] + \frac{1}{n}\sum_i \left[\frac{(y_i - \bar{y})^2}{\sigma_y^2}\right] \pm 2\frac{1}{n}\sum_i \frac{(x_i - \bar{x})}{\sigma_x}\frac{(y_i - \bar{y})}{\sigma_y} \ge 0$

$$\Rightarrow \frac{1}{\sigma_x^2}\frac{1}{n}\sum_i \left[(x_i - \bar{x})^2\right] + \frac{1}{\sigma_y^2}\frac{1}{n}\sum_i \left[(y_i - \bar{y})^2\right] \pm \frac{2}{\sigma_x \sigma_y}\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) \ge 0$$

$$\Rightarrow \quad \frac{\sigma_x^2}{\sigma_x^2} + \frac{\sigma_y^2}{\sigma_y^2} \pm 2\frac{Cov(x,y)}{\sigma_x \sigma_y} \geq 0 \text{ . That is, } 1 + 1 \pm 2\frac{P_{xy}}{\sigma_x \sigma_y} \geq 0$$

$$\Rightarrow \quad 2 \pm 2\, r_{xy} \geq 0 \text{ .} \qquad \text{That is, } 1 \pm r_{xy} \geq 0$$

This gives, $1 + r_{xy} \geq 0$ *or* $1 - r_{xy} \geq 0$

That is, $r_{xy} \geq -1$ *or* $r_{xy} \leq 1$

$$\Rightarrow \quad -1 \leq r_{xy} \leq +1$$

Remark: We have the regression coefficients y on x, $b_{yx} = \frac{P_{xy}}{\sigma_x^2}$ and the regression coefficients x on y, $b_{xy} = \frac{P_{xy}}{\sigma_y^2}$. The geometric mean of these regression coefficients gives the magnitude of the coefficient of correlation r_{xy} . The sign of correlation is determined by the sign of covariance between x and y, P_{xy} . If P_{xy} is positive r_{xy} is positive in sign and if P_{xy} is negative r_{xy} is negative in sign.

Theorem: **(Invariance of correlation coefficient under linear transformation):** A $\frac{-}{c}$ $\frac{-}{d}$ is making no change in the coefficient of correlation between the variables. That is, $r_{xy} = r_{uv}$.

Proof:

Let (x_1, y_1) , (x_2, y_2) ,…, (x_n, y_n) are the observations on x and y.

Then, $$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Let, $u = \frac{x - A}{c}$ and $v = \frac{y - B}{d}$;

Then, Pearson's coefficient of correlation between u and v,

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

$$\Rightarrow r_{uv} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - A}{c} - \left(\frac{\bar{x} - A}{c}\right)\right]\left[\frac{y_i - B}{d} - \left(\frac{\bar{y} - B}{d}\right)\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - A}{c} - \left(\frac{\bar{x} - A}{c}\right)\right]^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - B}{d} - \left(\frac{\bar{y} - B}{d}\right)\right]^2}}$$

$$\Rightarrow r_{uv} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x - \bar{x}}{c}\right]\left[\frac{y - \bar{y}}{d}\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x - \bar{x}}{c}\right]^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{y - \bar{y}}{d}\right]^2}}$$

$$\Rightarrow r_{uv} = \frac{\frac{1}{cd}\times\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}][y_i - \bar{y}]}{\frac{1}{cd}\times\sqrt{\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - \bar{y}]^2}}$$

$$\Rightarrow r_{uv} = \frac{\frac{1}{cd}\times P_{xy}}{\frac{1}{cd}\times\sigma_x\sigma_y} = \frac{P_{xy}}{\sigma_x\sigma_y}$$

$$\Rightarrow r_{uv} = r_{xy}.$$

Problem: Find the coefficient of correlation for the following data on X and Y.

X:	65	66	67	67	68	69	70	72
Y:	67	68	65	68	72	72	69	71

Solution:

$$\text{Coefficient of correlation, } r_{xy} = \frac{P_{xy}}{\sigma_x\sigma_y}$$

To find $\bar{x}$, $\bar{y}$, P_{xy} , σ_x^2 and σ_y^2

$$P_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \;;\; \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 \text{ and } \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y})^2$$

The calculations are as follows:

x	y	x^2	y^2	xy
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
544	552	37028	38132	37560

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{8} \times 544 = 68 ; \qquad \bar{y} = \frac{1}{n}\sum y_i = \frac{1}{8} \times 552 = 69$$

$$P_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy} = \frac{1}{8} \times 37560 - 68 \times 69 = 3$$

$$\sigma_{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \frac{1}{8} \times 37028 - (68)^2 = 4.5$$

$$\sigma_{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y})^2 = \frac{1}{8} \times 38132 - (69)^2 = 5.5$$

Coefficient of correlation, $r_{xy} = \frac{P_{xy}}{\sigma_x \sigma_y} = \frac{3}{\sqrt{4.5}\sqrt{5.5}} = 0.603$.

Problem: *Calculate Karl Pearson's coefficient of correlation for the following data;*

x:	*10*	*12*	*13*	*16*	*17*	*20*	*25*
y:	*19*	*22*	*26*	*27*	*29*	*33*	*37*

Solution:

$$\text{Coefficient of correlation } r = \frac{Cov(X, Y)}{S.D.(X) \times S.D.(Y)}$$

The problem can be solved by simply following the steps shown in above example. But for some computational easiness the problem can also be solved as in the following illustration.

We have the result that correlation coefficient is independent of change of origin and scale. Hence we can calculate the correlation between X and Y by altering

X and Y by some linear transformation. Here, consider U = X – 16 and V = Y – 27.

The correlation between U and V is same to correlation between X and Y.

Correlation between U and V, $r = \dfrac{Cov(U, V)}{S.D.(U) \times S.D.(V)}$

The calculations are:

x	y	U = X – 16	V = Y – 27	U^2	V^2	UV
10	*19*	*-6*	*-8*	*36*	*64*	*48*
12	*22*	*-4*	*-5*	*16*	*25*	*20*
13	*26*	*-3*	*-1*	*9*	*1*	*3*
16	*27*	*0*	*0*	*0*	*0*	*0*
17	*29*	*1*	*2*	*1*	*4*	*2*
20	*33*	*4*	*6*	*16*	*36*	*24*
25	*37*	*9*	*10*	*81*	*100*	*90*
		1	*4*	*159*	*230*	*187*

$$Cov(U, V) = \frac{1}{n}\sum uv - \bar{u}\bar{v} = \frac{1}{7}(187) - \left(\frac{1}{7} \times \frac{4}{7}\right) = 26.71 - .082 = 26.628$$

$$V(U) = \frac{1}{n}\sum u^2 - \bar{u}^2 = \frac{1}{7}(159) - \left(\frac{1}{7}\right)^2 = 22.71 - 0.02 = 22.69$$

$$V(V) = \frac{1}{n}\sum v^2 - \bar{v}^2 = \frac{1}{7}(230) - \left(\frac{4}{7}\right)^2 = 32.86 - 0.327 = 32.533$$

Now, Correlation between U and V, $r = \dfrac{26.628}{\sqrt{22.69} \times \sqrt{32.533}} = 0.98$

That is the correlation coefficient of X and Y =

2.8. Rank correlation coefficient

When we are considering two characteristics which are qualitative in nature, they are not possible to measure numerically. For example consider the characteristics of the ability in drawing (let it be X) and the ability in music (let it be Y). It is not possible to measure numerically the values of X and Y, for an individual. But if there are *n* individuals, it is possible to rank these *n* individuals according to the ability in drawing

(X) and according to their ability in music (Y). If these two characteristics are having high positive correlation, then ranks obtained for the individuals based of X and Y will be in same order. If these two characteristics are having high negative correlation, then ranks obtained for the individuals based of X and Y will be in reverse order. Using the ranks obtained for the *n* individuals based on the characteristics X and Y, a method of finding the coefficient of correlation is derived by C.Spearman in 1904. The coefficient of correlation for two characteristics which are calculated based on the ranks is known as Spearman's Rank Correlation Coefficient.

Let there be n individuals ranked according to two qualitative characteristics considered. Let (x_i , y_i) denote the rank of the i_{th} individual when ranked according to the characteristics. So the x_i , y_i values are the numbers from 1 to n.

Since x_i values are the numbers from 1 to n, the mean of x values,

$$\bar{x} = \frac{\text{sum of first n natural numbers}}{n} = \frac{1}{n} \times \frac{n(n+1)}{2} = \frac{(n+1)}{2}$$

Similarly,

$$\bar{y} = \frac{\text{sum of first n natural numbers}}{n} = \frac{1}{n} \times \frac{n(n+1)}{2} = \frac{(n+1)}{2}$$

Variance of x_i values,

$$\sigma_x^2 = \frac{\text{sum of squares of first n natural numbers}}{n} - \left[\frac{(n+1)}{2}\right]^2$$

$$\Rightarrow \quad \sigma_x^2 = \frac{1}{n} \times \frac{n(n+1)(2n+1)}{6} - \left[\frac{(n+1)}{2}\right]^2 ;$$

$$= \frac{n^2 - 1}{12}$$

Similarly, $\sigma_{y^2} = \frac{n_2 - 1}{12}$.

Let $d_i = (x_i - y_i)$. This gives, $\bar{d} = \bar{x} - \bar{y} = 0$

Variance of 'd' values,

$$\sigma_d^2 = \frac{1}{n}\sum_{i=1}^{n} d^2 - |\bar{d}|^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 - [0]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} d_i^2$$

Since $\bar{x} = \bar{y}$, we can re write $\frac{1}{n}\sum_{i=1}^{n} d_i^2$ as, $\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x} + \bar{y} - y_i)^2$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) - (y_i - \bar{y})^2$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} d_i^2 = \sigma_x^2 + \sigma_y^2 - 2\,\mathrm{cov}(x, y)$$

But, we have, $\mathrm{cov}(x, y) = r\sigma_x\sigma_y$, where r is the coefficient of correlation. Hence,

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$$

Since, $\sigma_x^2 = \sigma_y^2 = \frac{n^2 - 1}{12}$,

we get, $\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{n^2 - 1}{12} + \frac{n^2 - 1}{12} - 2r\sqrt{\frac{n^2 - 1}{12}}\sqrt{\frac{n^2 - 1}{12}}$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} d_i^2 = 2\times\frac{n^2 - 1}{12} - 2\times r\frac{n^2 - 1}{12}$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} d_i^2 = (1 - r)\frac{n^2 - 1}{6}$$

$$\Rightarrow 1 - r = \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad or$$

the coefficient of correlation $r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$.

Problem: *The following are the ranks obtained by 10 students in Statistics and Mathematics*

Statistics:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Mathematics:	*1*	*4*	*2*	*5*	*3*	*9*	*7*	*10*	*6*	*8*

To what extent is the knowledge of students in the two subjects related?

Solution:

Here to find the rank correlation coefficient of the ranks in Statistics and Mathematics. Rank correlation coefficient is defined as,

$r = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, d_i is the difference in ranks.

The calculations are:

Rank in Stat. x_i	Rank in Maths y_i	$d_i = x_i - y_i$	d_i^2
1	1	0	0
2	4	-2	4
3	2	1	1
4	5	1	1
5	3	2	4
6	9	3	9
7	7	0	0
8	10	-2	4
9	6	3	9
10	8	2	4
			36

Hence, $r = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = = 1 - \frac{6 \times 36}{10(10^2-1)} = 1 - 0.2189 = 0.7819$

Problem: *10 competitors in a music test were ranked by three judges A, B, and C in following order.*

Ranks by A:	*1*	*6*	*5*	*10*	*3*	*2*	*4*	*9*	*7*	*8*
Ranks by B:	*3*	*5*	*8*	*4*	*7*	*10*	*2*	*1*	*6*	*9*
Ranks by C:	*6*	*4*	*9*	*8*	*1*	*2*	*3*	*10*	*5*	*7*

Discuss which pair of judges has the nearest approaches to common likings in music.

Solution:

Here to find the rank correlation coefficient between each pair of the judges considering the ranks they given. Identify the pair of judges with high correlation coefficient. They are considered having nearest approaches to common likings in music.

The calculations follow:

Ranks by A x_i	Ranks by B y_i	Ranks by C z_i	$x_i - y_i$	$x_i - z_i$	$y_i - z_i$	$(x_i - y_i)^2$	$(x_i - z_i)^2$	$(y_i - z_i)^2$
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	4
						200	60	214

Rank correlation between A and B, $r = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 200}{10(10^2-1)} = -0.212$

Rank correlation between A and C, $r = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{10(10^2-1)} = 0.6364$

Rank correlation between B and C, $r = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10(10^2-1)} = -0.297$

It can be observed that the judges A and C are having nearest approaches to common likings in music.

Problem: *Find the rank correlation coefficient for the following data:*

X:	*92*	*89*	*87*	*86*	*84*	*77*	*71*	*63*	*53*	*50*
Y:	*86*	*83*	*91*	*77*	*68*	*85*	*52*	*82*	*37*	*57*

Solution:

First, the given values of X and Y should be ranked. If an observation repeats, then the sum of the ranks is equally divided among the observations. (For eg., when we are ranking the observations in order, and let a number, say *a*, coming in the *6th* and 7th position then the first and second *a* values are assigned with the rank 6.5).

Here the observations are ranked in descending order. Then find the rank correlation coefficient.

x	y	Rank of X, x_i	Rank of Y, y_i	x_i - y_i	$(x_i - y_i)^2$
92	86	1	2	-1	1
89	83	2	4	-2	4
87	91	3	1	2	4
86	77	4	6	-2	4
84	68	5	7	-2	4
77	85	6	3	3	9
71	52	7	9	-2	4
63	82	8	5	3	9
53	37	9	10	-1	1
50	57	10	8	2	4
					44

$$\text{Rank correlation coefficient, } r = 1 - \frac{6\sum_i d_i^{\,2}}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 44}{10(10^{\,2}-1)} = 0.733$$

Rank correlation coefficient when equal ranks (Tied ranks):

It may be noted that the Spearman's rank correlation formula is derived on the assumption that all the ranks are different. But in practice, there are many situations, where more than one individual are getting the same rank. In a competition consider, three individuals received 3rd rank. They would have given the 3rd ,4th, and 5th rank, if there were slight difference in the evaluation. Then we add 3,4 and 5, which is 12. Then 12 is equally divided for these three individuals. Hence we assign the rank 4 to each of these three individual. In such situations it is more accurate to calculate the Pearson's coefficient of correlation between the ranks directly after assigning the average rank to those with the same rank. But there is also a modified formula of Spearman's rank correlation coefficient, which is as follows:

$$r = 1 - \frac{6\left[\sum_{i=1}^{n} d_i^2 + \frac{1}{12}\sum_i m_i\left(m_i^2 - 1\right) + \frac{1}{12}\sum_j m_j\left(m_j^2 - 1\right)\right]}{\left(\quad\right)}$$

, where, m_i stands for the number of times the i_{th} rank repeats in the x series of ranks and m_j is the number of times the j_{th} rank repeats in the y series of ranks when the average ranks are assigned. The method is illustrated below:

Obtain the rank correlation coefficient for the following data:

X:	15	20	28	12	40	60	20	80
Y:	40	30	50	30	20	10	30	60

Illustration:

At first we assign ranks for X and Y values. Here we have 8 sets of data. That is $n=8$.

The ranks are:

X:	7	5.5	4	8	3	2	5.5	1
Y:	3	5	2	5	7	8	5	1

Here in X values, 20 repeats twice, with the possible ranks, 5 and 6. Hence its average 5.5 is supplied for the value 20. Similarly in Y values, 30 repeat thrice, with possible ranks 4, 5 and 6. Hence their average 5 is assigned as the ranks of the values 30.
Now the difference in ranks, $d_i = X_i - Y_i$ values are:

d_i :	4	0.5	2	3	-4	-6	0.5	0
d_i^2:	16	0.25	4	9	16	36	0.25	0

This gives, $\sum_i d_i^2 = 81.50$.

$m_i = 2$ (Because on X values, only the value 20 repeats twice) and $m_j = 3$ (because on Y values, only the value 30 repeats thrice).

Hence, $r = 1 - \dfrac{6\left[\sum_{i=1}^{n} d_i^2 + \frac{1}{12}\sum_i m_i(m_i^2 - 1) + \frac{1}{12}\sum_j m_j(m_j^2 - 1)\right]}{n(n^2 - 1)}$

$$= 1 - \frac{6\left[81.50 + \frac{1}{12} \times 2(2^2 - 1) + \frac{1}{12} \times 3(3^2 - 1)\right]}{8(8^2 - 1)}$$

$$= 1 - \frac{6\left[81.50 + 0.5 + 2\right]}{8\left(63\right)} \quad = 0.$$