Correlation

In a bivariate distribution, if the change in one variable affects a change in other variables, then the variables are said to be correlated. For example, the two variables could be price and demand of commodities, income and expenditure on household items, rainfall and yield of crops etc. If the two variables move in the same direction i.e., if the decrease (or increase) in one variable results in the corresponding decrease (or increase) in other variables, then the variable are said to be positively correlated. Like correlation between income and expenditure is a positive correlation on the other hand if the variables move in the opposite direction i.e. the decrease (or increase) in one variables affects a increase (or decrease) in other variables, then the two variables are said to be negatively correlated. For example, the correlation between price and demand of a commodity is a negative. In what follows we shall concentrate on the relationship between two variables which is linear. The correlation coefficient defined later gives the degree of linear relationship between the two variables.

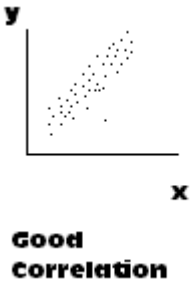
 # Scatter Diagram

The simplest way of getting an idea of relationship between two variables is through a graphical representation of the data. If (x_i, y_i), i= 1,2,….n denote the set of n observations on the two variables x and y , we first plot a graph by taking x- variable on the x-axis and y- variable on the y-axis and plot all the n points in the xy-palne. The diagram so obtained will be the scatter of n points and is called **scatter diagram**. By looking at the scatteredness of the plotted points, we can form an approximate idea about the degree of relationship between the two variables. If the points are widely scattered, then there exists a poor correlation between the two variables and if the points are very closely scattered around a straight line then the variables are said to have good correlation.

Fig-1 and Fig-2 show the scattered diagram of the different set of data. Fig-1 is an illustration of poor correlation whereas the Fig-2 indicates that the two variables have high degree correlation.

Fig-1

Fig-2



 # Karl Pearson's Coefficient of Correlation

Karl Pearson, a British Statistician, developed a formula called as Karl Pearson's coefficient of Correlation for studying the degree of linear relationship between the two variables. This is also known as product moment correlation coefficient. The coefficient of correlation $r(x,y)$ between the two variables x and y is defined as

$$r(x,y) = \frac{Cov(x,y)}{\sigma_x \, \sigma_y}$$

where σ_x^2, σ_y^2, $Cov(x,y)$ are respectively the variances of x and y and covariance between two variables x and y.

If (x_i, y_i); $i = 1,2\ldots n$ are the set of n observations of the pair of variables x and y, then we have

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

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

and so

$$r(x,y) = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}\ \sqrt{\frac{1}{n}\sum(y_i - \bar{y})^2}} \qquad (5.16)$$

where all the summation written as $\sum$ are from i=1 to n.

For computing the values of $r(x,y)$ we can make use of another simpler and convenient form of (5.16). We have

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

$$= \frac{1}{n}\sum x_iy_i - \bar{y}\frac{1}{n}\sum x_i - \bar{x}\frac{1}{n}\sum y_i + \bar{x}\,\bar{y}.\frac{1}{n}\sum 1$$

$$= \frac{1}{n}\sum x_iy_i - \bar{y}\,\bar{x} - \bar{x}\,\bar{y} + \bar{x}\,\bar{y} \quad = \frac{1}{n}\sum x_iy_i - \bar{x}\,\bar{y}$$

and $$\frac{1}{n}\sum(x_i-\bar{x})^2 = \frac{1}{n}\sum(x_i^2 + \bar{x}^2 - 2x_i\bar{x})$$

$$= \frac{1}{n}\sum x_i^2 + \bar{x}^2\frac{1}{n}\sum 1 - 2\bar{x}\sum x_i \quad = \frac{1}{n}\sum x_i^2 + x^{-2} - 2x^2 \quad = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

Similarly $$\frac{1}{n}\sum(y_i-y)^2 = \frac{1}{n}\sum y_i^{\,2} - \bar{y}^{\,2}$$

and so $$r(x,y) = \frac{\frac{1}{n}\sum x_iy_i - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\sum x_i^2 - \bar{x}^2}\;\sqrt{\frac{1}{n}\sum y_i^{\,2} - \bar{y}^2}} \qquad (5.17)$$

 # Limits of Correlation coefficient

From (5.17), we have

$$r(x,y) = \frac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2}\;\sqrt{\frac{1}{n}\sum(y_i-\bar{y})^2}} \quad \text{or} \quad r^2(x,y) = \frac{\left(\sum(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\sum(x_i-\bar{x})^2(y_i-\bar{y})^2} = \frac{\left(\sum a_i\,b_i\right)^2}{\sum a_i^{\,2}\sum b_i^{\,2}}$$

(5. 18)

where $x_i - \bar{x} = a_i$, $\quad y_i - \bar{y} = b_i$

Now, making use of Cauchy-Schwartz inequality, which states that, if a_i, b_i, i = 1,2,….,n, are real numbers, then

$$\left(\sum a_ib_i\right)^2 \le \left(\sum a_i^2\right)\left(\sum b_i^2\right) \quad \text{or} \quad \frac{\left(\sum a_ib_i\right)^2}{\left(\sum a_i^2\right)\left(\sum b_i^2\right)} \le 1 \qquad (5.\ 19)$$

Using (5. 19) in (5.18), we have

$r^2(x,y) \le 1$

or $\quad |\,r(x,y)\,| \le 1$

or $\quad -1 \le r\,(x,y) \le 1$

Hence the coefficient of correlation lies in the closed interval [-1, 1]. If r = +1, then the correlation is called perfect and positive and if r = -1, then it is perfect and negative.

 # Effect of change of origin and scale on the coefficient of correlation

Let $u = \frac{x-a}{h}$, $v = \frac{y-b}{k}$, *where* a, b, h > 0, k > 0 are constants

So, $x = a + uh$ *and* $y = b + vk$. Also $\bar{x} = a + \bar{u}h$ *and* $\bar{y} = b + \bar{v}k$

$\therefore$ $Cov\,(x,y) = \frac{1}{n}\sum(x_i-\bar{x})\,(y_i-\bar{y}) = \frac{hk}{n}\sum(u_i-\bar{u})\,(v-\bar{v})$

$$\sigma_x^{\,2} = \frac{1}{n}\sum(x_i-\bar{x})^2 = \frac{h^2}{n}\sum(u_i-\bar{u})^2$$

and $\sigma_y^2 = \frac{1}{n}\sum(y_i - \bar{y})^2 = \frac{k^2}{n}\sum(v_i - \bar{v})^2$

$$\therefore \quad r(x, y) = \frac{Cov(x, y)}{\sigma_x \, \sigma_y} = \frac{hk.\frac{1}{n}\sum(u_i - \bar{u})(v_i - \bar{v})}{hk\sqrt{\frac{1}{n}\sum(u_i - \bar{u})^2}\sqrt{\frac{1}{n}\sum(v_i - \bar{v})^2}} = \frac{Cov(u, v)}{\sigma_x \, \sigma_y} = r(u, v)$$

Thus, coefficient or correlation is independent of change of origin and scale.

Remark: If *r(x, y) = 0*, then the two variables *x* and *y* are said to be uncorrelated. It is not necessary that the two uncorrelated variable *x* and *y* are independent in the sense that there is no relationship between the two variables *x* and *y*. We give below an example where the variables are perfectly related but *r = 0*.

x :	-10	-9	-8	8	9	10	$\sum x = 0$
y:	100	81	64	64	81	100	$\sum y = 490$
xy:	-100	-729	-512	512	729	1000	$\sum xy = 0$

$$\therefore \quad \text{cov}(x, y) = \frac{1}{n}\sum xy - \bar{x}\,\bar{y} = 0 \qquad \text{and hence } r(x, y) = \frac{Cov(x, y)}{\sigma_x \, \sigma_y} = 0$$

Thus the variables x and y are correlated. You can see that the two variables are connected by the exact relationship $y = x^2$

Example 5.5 Find the coefficient of correlation between the two variables *x* and *y* as given in the following table:

x:	10	14	18	22	26	30
y:	18	12	24	6	30	36

Solution: Let u = $\frac{x - 22}{4}$, *and* $v = \frac{y - 24}{6}$

x	**y**	**u**	**v**	**u^2**	**v^2**	**uv**
10	18	-3	-1	9	1	3
14	12	-2	-2	4	4	4
18	24	0	0	1	0	-1
22	6	-3	-3	0	9	-3
26	30	1	1	1	1	1
30	36	2	2	4	4	4
Total		-3	-3	19	19	12

$$\bar{u} = \frac{1}{n}\sum u_i = \frac{1}{6}(-3) = -\frac{1}{2} \quad \textbf{and} \quad \bar{v} = \frac{1}{n}\sum v_i = \frac{1}{6}(-3) = -\frac{1}{2}$$

$$r(x, y) = r(u, v) = \frac{\frac{1}{n}\sum u_i v_i - \bar{u}\,\bar{v}}{\sqrt{(\frac{1}{n}\sum u_i^2 - \bar{u}^2)(\frac{1}{n}\sum v_i^2 - \bar{v}^2)}} = \frac{\frac{1}{6}(12) - \frac{1}{4}}{\sqrt{\frac{1}{6}(19) - \frac{1}{4}}\sqrt{\frac{1}{6}(19) - \frac{1}{4}}} = 0.6$$

Example 5.6: From the following data of 6 cities calculate the coefficient of correlation between the density of population and death rate.

Cities	***Area in sq. miles***	***Population in '000***	***No. of deaths***
A	150	30	300
B	180	90	1440
C	100	40	560
D	60	42	840
E	120	72	1224
F	80	24	312

Solution: First, we will find the density of population and death rate and denote them by *x* and *y*.

$$\text{Density} = \frac{Population}{Area} \quad \text{and Death rate} = \frac{No.\ of\ deaths}{Population} \times 1000$$

Cities	*Density* x	$(x-450)/100$ u	u^2	*Death rate* y	$(y-15)$ v	v^2	uv
A	200	-2.5	6.25	10	-5	25	+12.5
B	500	+0.5	0.25	16	+1	1	-0.5
C	400	-0.5	0.25	14	-1	1	+0.5
D	700	+2.5	6.25	20	+5	25	+12.5
E	600	+1.5	2.25	17	+2	4	+3.0
F	300	-1.5	2.25	13	-2	4	+3.0
		$\sum u = 0$	$\sum u^2 = 17.5$	$\sum y = 90$	$\sum v = 0$	$\sum v^2 = 60$	$\sum uv = 31$

$$r(x,y) = r(u,v) = \frac{\frac{1}{n}\sum uv - \bar{u}\,\bar{v}}{\sqrt{\frac{1}{n}\sum u^2 - \bar{u}^2}\ \sqrt{\frac{1}{n}\sum v^2 - \bar{v}^2}} = \frac{\sum uv}{\sqrt{\sum u^2 \sqrt{\sum v^2}}} = \frac{31}{\sqrt{17.5 \times 60}} = \frac{31}{32.404} = +0.957$$

Thus there is a high degree of positive correlation between density of population and death rate.

Rank Correlation

Sometimes we may have to handle the problems in which data cannot be measured quantitatively but can be measured qualitatively. In such situation we make use of rank correlation.

Let a group of n individuals be arranged in order of their merit or proficiency in possession of a certain characteristics. The same group would in general have different order for different characteristics. For example if we consider the relation between mathematical and musical ability of a group of individuals, it is not necessary that an individual who is best in mathematics will also be best in music. Let x_i and y_i, i = 1,2,3,….n be the ranks of the i^{th} individual in respect of two characteristics A and B respectively. The coefficient of correlation between the ranks x_i and y_i is called the rank correlation coefficient between A and B for that group of individuals.

Assume that ranks of no two individuals are same in either of the characteristic. (no ties). Therefore, each of the individual will get one of the ranks, 1,2,3,….n for the characteristics A and B

$$\therefore\ \bar{x} = \bar{y} = \frac{1+2+3+.....n}{n} = \frac{n+1}{2}$$

and
$$\sigma_x^{\,2} = \frac{1}{n}\sum x_i^2 - (\bar{x})^2 = \frac{1}{n}(1^2 + 2^2 + ...n^2) - \left(\frac{n+1}{2}\right)^2 = \frac{n(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n^2-1}{12} \qquad (5.20)$$

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

Defining, $d_i = x_i - y_i$, we have

$$d_i = (x_i - \bar{x}) - (y_i - \bar{y}) \qquad (\text{since } \bar{x} = \bar{y})$$

and
$$\sum_{i=1}^{n} d_i^{\,2} = \sum_{i=1}^{n}\left[(x_i - \bar{x})^2 - (y_i - \bar{y})\right]^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

or
$$\frac{1}{n}\sum_{i=1}^{n} d_i^{\,2} = \sigma_x^{\,2} + \sigma_y^{\,2} - 2\,Cov(x,y) \qquad (5.21)$$

we know that $r(x, y) = \frac{Cov\,(x, y)}{\sigma_x\,\sigma_y}$

$\Rightarrow$ $Cov\,(x, y) = r\,\sigma_x\,\sigma_y$, where r is the rank correlation between x and y.

So, (5.21) can be written as

$$\frac{1}{n}\sum_{i=1}^{n} d_i^{\,2} = 2\sigma_x^{\,2} - 2\,r\sigma_x^{\,2}, \text{ since } \sigma_x^{\,2} = \sigma_y^{\,2} \text{ in this case}$$

$= 2(1-r)\,\sigma_x^{\,2}$ or $1-r = \dfrac{\sum_{i=1}^{n} d_i^{\,2}}{2n\sigma_x^{\,2}}$ or $r = 1 - \dfrac{\sum_{i=1}^{n} d_i^{\,2}}{2n\sigma_x^{\,2}}$ or $r = 1 - \dfrac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)}$ (using 5.20)

This formula is called **Spearman's formula for the rank correlation**.

Remark: In order to rank a series the item with largest size is ranked 1, next largest 2 and so on. If there occurs a tie in the items of the series then they may be assigned the average of the ranks they would have got if they had differed slightly. For example if the two items are tied for 3rd rank, then each of the item be ranked as $\dfrac{3+4}{2} = 3.5$ and similarly if three items are tied for 5th rank, then each may be ranked as $\dfrac{5+6+7}{3} = 6$ and so on. As a result of these common rankings the formula for rank correlation has to be corrected. We add a correction factor $\dfrac{m(m^2-1)}{12}$ to $\sum d^2$ where m denotes the number of times an item is repeated. This correction factor is to be added for each repeated value in both the x-series and y-series. So, the formula for rank correlation in case of repeated ranks becomes

$$r = 1 - \frac{6\left(\sum d_i^{\,2} + \dfrac{m_1(m_1^{\,2}-1)}{12} + \dfrac{m_2(m_2^{\,2}-1)}{12} + \ldots\ldots\right)}{n(n^2-1)}$$

Example 5.7 Calculate the rank coefficient of correlation of the following data:

x:	80	78	75	75	68	67	60	59
y:	12	13	14	14	14	16	15	17

Solution:

x	Rank	*y*	Rank	Rank-differences	Square of rank differences
80	1	12	8	-7	49
78	2	13	7	-5	25
75	3.5	14	5	-1.5	2.25
75	3.5	14	5	-1.5	2.25
68	5	14	5	0	0
67	6	16	2	4	16
60	7	15	3	4	16
59	8	17	1	7	49
n = 8				$\sum d = 0$	$\sum d^2 = 159.50$

Here we see that the number 75 occurs twice in x-series so we give the rank 3.5 (mean of 3 and 4) to each of the two. Whereas the number 14 occurs three times in y-series so we give the rank 5 (mean of 4,5and 6) to each of the three. So we have $m_1 = 2$ and $m_2 = 3$

$$\therefore \text{ Coefficient of rank correlation } (r(x,y)) = 1 - \frac{6\left(\sum d^2 + \dfrac{2(2^2-1)}{12} + \dfrac{3(3^2-1)}{12}\right)}{n(n^2-1)} = 1 - \frac{6\left(159.5 + \dfrac{5}{2}\right)}{8(64-1)}$$

$$= -0.92$$

thus there is a high degree negative correlation between the two variables.

Example 5.8 Ten competitors in a beauty contest got marks by three judges in the following orders:

Judge A	1	6	5	10	3	2	4	9	7	8
Judge B	3	5	8	4	7	10	2	1	6	9
Judge C	6	4	9	8	1	2	3	10	5	7

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

Solution: we shall calculate the three sets of rank correlation coefficient

Ranks by A (X)	Ranks by B (Y)	Ranks by C (Z)	d_1 = X-Y	d_2 = X-Z	d_3 = Y-Z	d_1^2	d_2^2	d_3^2
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
Total			**0**	**0**	**0**	**200**	**60**	**214**

(i) For the first and second judge's opinion, we have

$$r(X,Y) = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 200}{100(100-1)} = -\frac{7}{33}$$

(ii) Similarly for the first and third Judge's opinion, we have

$$r\,(X, Z) = \frac{7}{11}$$

(iii) For second and third judge's opinion, we have

$$r\,(Y, Z) = -\frac{49}{165}$$

Thus we conclude that the judges A and C have nearest approach for beauty.

Regression

In regression analysis we deal with the problem of estimating (or predicting) the value of one variable from the given values of the other variable. To make such a prediction it is necessary to have a mathematical relationship between the variables. The variable whose value is to be predicted is called dependent (or regressed) variable and the variable which is used for prediction is called independent (or regressor) variable. This mathematical relationship could be any function of the form $y = f(x)$.

Let we be given the set $(x_i\, y_i)$, i = 1,2, …..n of n observations of the pair of variables x and y. To get an idea of the form of the function $f(x)$, we plot all the values of the pair x and y on xy-plane. Such a diagram is called scatter diagram.

Linear Regression

If the variables in a bivariate distribution are related, then the scatter diagram of that distribution will concentrate around some curve called as curve of regression. If this curve is a straight line then the line is called line of regression and there is said to be linear regression between the variables.

The line of regression is the line of best fit and is obtained by the principle of least squares.

Let us assume that in a bivariate distribution (x_i, y_i), i= 1,2,….n of a pair of variables x and y, x is independent variable and y is dependent variable. Let the line of regression of y and x be

$$y = a + bx \qquad (5.22)$$

We want to find a and b so that the line (5.22) be the line of best fit in the sense of principle of least squares i.e.,

We have to find a and b so that

$$\mathrm{E} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

is minimum. For E to be minimum we must have

$$\frac{\partial E}{\partial a} = 0, \quad \frac{\partial E}{\partial b} = 0$$

i.e., $-2\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 0$ and $-2\sum_{i=1}^{n}(y_i - a - bx_i)^2(x_i) = 0$

or $\sum_{i=1}^{n}(y_i - a - bx_i) = 0$ and $\sum_{i=1}^{n}(x_i y_i - ax_i - bx_i^2) = 0$

or $$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \qquad (5.23)$$

and $$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \qquad (5.24)$$

Since x_i and y_i are known, equation (5.23) and (5.24) known as the normal equations and can be solved for estimating a and b.

From (5.23), we have $\frac{1}{n}\sum_{i=1}^{n} y_i = a + \frac{b}{n}\sum_{i=1}^{n} x_i$

or $$\bar{y} = a + b\bar{x} \qquad (5.25)$$

Thus the line of regression of y on x passes through the point $(\bar{x}, \bar{y})$

Now, $\text{Cov}(x, y) = \frac{1}{n}x_i y_i - \bar{x}\,\bar{y}$

$\Rightarrow$ $$\frac{1}{n}\sum_{i=1}^{n} x_i y_i = Cov(x, y) + \bar{x}\,\bar{y} \qquad (5.26)$$

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

$\Rightarrow$ $$\frac{1}{n}\sum x_i^2 = \sigma_x^2 + \bar{x}^2 \qquad (5.27)$$

Dividing both sides of (5.24) by n and making use of (5.26) and (5.27), we get

$$Cov(x, y) + \bar{x}\,\bar{y} = a\bar{x} + b(\sigma_x^2 + \bar{x}^2) \qquad (5.28)$$

Multiplying (5. 25) by $\bar{x}$ and then subtracting from (5. 28), we get

$Cov(x, y) = b\,\sigma_x^2$ or $b = \frac{Cov(x, y)}{\sigma_x^2}$

Since b is the slope of the line of regression of y on x passing through $(\bar{x}, \bar{y})$, so its equation is

$$y - \bar{y} = b(x - \bar{x})$$

or $$y - \bar{y} = \frac{Cov(x, y)}{\sigma_x^2}(x - \bar{x})$$

But we know that $r(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y}$

$\therefore$ The equation of line of regression of y on x is:

$$y - \bar{y} = \frac{r\sigma_x\sigma_y}{\sigma_x^2}(x - \bar{x}) \quad \text{or} \quad y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})$$

or $$y - \bar{y} = b_{yx}(x - \bar{x}), \text{ where } b_{yx} = \frac{r\sigma_y}{\sigma_x} \qquad (5.29)$$

Proceeding in the similar way, the line

$$x = a + by$$

called as the line of regression of x and y is given by

$$x - \bar{x} = \frac{r\sigma_y}{\sigma_x}(y - \bar{y})$$

$$x - \bar{x} = b_{xy}(y - \bar{y}) \qquad \text{where } b_{xy} = \frac{r\sigma_x}{\sigma_y} \qquad (5.30)$$

 # Regression Coefficient

The slope b of the line of regression of y and x is called the regression coefficient of *y* on *x* and is denoted by b_{yx} and is given by

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} \qquad (5.31)$$

Similarly b_{xy} called as the regression coefficient of x and y is given by

$$b_{xy} = \frac{r\sigma_x}{\sigma_y} \qquad (5.32)$$

 # Properties of Regression Coefficients

(i) Coefficient of correlation is the geometric mean of the regression coefficients.

From (5. 31) and (5. 32), we have

$$b_{yx}\, b_{xy} = r\frac{\sigma_y}{\sigma_x}\; r\frac{\sigma_x}{\sigma y} = r^2$$

$$\Rightarrow \qquad r = \pm\sqrt{b_{yx}\, b_{xy}} \qquad (5.33)$$

Remark The following points must be noted about the regression coefficients:

(a) Both regression coefficients will have the same sign i.e. either both will be positive or both negative other wise $r = \sqrt{b_{yx}\, b_{xy}}$ will become imaginary.

(b) The coefficient of correlation has the same sign as that of regression coefficients. Since, we have

$$r = \frac{Cov(x, y)}{\sigma_x \sigma_y}\ ,\ b_{yx} = \frac{Cov(x, y)}{\sigma_x^{\,2}}\ \ and\ \ b_{xy} = \frac{Cov(x, y)}{\sigma_y^{\,2}} \qquad (5.34)$$

from (5.34) it is clear that the sign of coefficient of correlation depends upon the sign of covariance only (since σ_x and σ_y are always non-negative) and the sign of covariance depends upon the sign of regression coefficient. Hence the sign of coefficient of correlation depends on the sign of regression coefficients. Thus in (5.33), the sign of coefficient of correlation is taken positive if regression coefficients are positive and is taken negative if regression coefficients are negative.

(ii) *If one of the regression coefficient is greater than one then the other must be less than one.*

Let $b_{yx} > 1 \Rightarrow \frac{1}{b_{yx}} < 1$. Also $r^2 \le 1 \Rightarrow b_{xy}\, b_{yx} \le 1$. Hence $bxy \le \frac{1}{byx} < 1$

(iii) *Regression coefficients are independent of change of origin but not of scale.*

Let $u = \frac{x-a}{h},\ v = \frac{y-b}{k}$, where *a, b, h > 0, k > 0* are constants

then from article (5.8.4), we have

Cov *(x, y) = hk Cov (u, v)* $\qquad \sigma_x^{\,2} = h^2\, \sigma_u^{\,2} \qquad and \qquad \sigma_y^{\,2} = k^2 \sigma_v^{\,2}$

$$\therefore \qquad b_{yx} = \frac{Cov(x, y)}{\sigma_x^{\,2}} = \frac{k}{h}\,\frac{Cov(u, v)}{\sigma_u^{\,2}} = \frac{k}{h}\, b_{vu}$$

Similarly, we can prove that $b_{xy} = \frac{h}{k}\ b_{uv}$

 # Angle between two liens of regression

We know that the lines of regression *y* on *x* and that of *x* on *y* are

$$y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x}) \qquad (5.35)$$

and $\qquad x - \bar{x} = \frac{r\sigma_y}{\sigma_x}(y - \bar{y}) \quad \Rightarrow y - \bar{y} = \frac{\sigma_y}{r\sigma_x}(x - \bar{x}) \qquad (5.36)$

If θ is the angle between the two lines, then

$$tan\ \theta = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right|, \quad \text{where } m_1 \text{ and } m_2 \text{ are the slopes of the two lines.}$$

From (5. 35) and (5. 36), the slopes m_1 and m_2 of two lines are:

$$m_1 = \frac{r\sigma_y}{\sigma_x} \text{ and } \quad m_2 = \frac{\sigma_y}{r\sigma_x}$$

$$\therefore \qquad tan\ \theta = \left| \frac{\frac{r\sigma_y}{\sigma_x} - \frac{\sigma_y}{r\sigma_x}}{1 + \frac{r\sigma_y}{\sigma_x}\frac{\sigma_y}{r\sigma_x}} \right| = \left| \frac{r^2 - 1}{r} \right| \left(\frac{\sigma_x\, \sigma_y}{\sigma_x^{\,2} + \sigma_y^{\,2}} \right) = \frac{1 - r^2}{|r|} \frac{\sigma_x\, \sigma_y}{\sigma_x^{\,2} + \sigma_y^{\,2}}\ , \qquad (as\ r^2 \le 1)$$

$$\therefore \qquad \theta = \tan^{-1}\left\{ \frac{1 - r^2}{|r|} \frac{\sigma_x \sigma_y}{\sigma_x^{\,2}\, \sigma_y^{\,2}} \right\}$$

Corollary (i): If r = 0, then tan $\theta = \infty$ i.e. $\theta = \pi/2$
Thus if the two variables are uncorrelated, then the two lines of regression are perpendicular to each other.
Corrolar (ii) : If r = ±1, then tan θ = 0, i.e. θ = 0 or π. Thus in this case the two lines of regression are either parallel or coincide with each other. But since the two lines of regression intersect at the point $(\bar{x}, \bar{y})$, so they can not be parallel. Hence in case of perfect correlation i.e. when r = ±1, the two liens of regression coincide with each other.

Example 5.9 Find the equations of lines of regression of the following data:

Age of husband:	18	19	20	21	22	24	24	25	26	27
Age of wife :	17	17	18	18	17	17	19	20	21	22

Solution: Let age of husbands be denoted by *x* and age of wives by *y* and let

$u = x - 22, \quad v = y - 19$,

then from the following table, we have

Age of husband x	$u = (x-22)$	u^2	Age of wives y	$v = (y-19)$	v^2	uv
18	-4	16	17	-2	4	8
19	-3	9	17	-2	4	6
20	-2	4	18	-1	1	2
21	-1	1	18	-1	1	1
22	0	0	19	0	0	0
23	1	1	19	0	0	0
24	2	4	19	0	0	0
25	3	9	20	1	1	3
26	4	16	21	2	4	8
27	5	25	22	3	9	15
225	**-5**	**84**	**190**	**0**	**24**	**43**

$$\bar{u} = \frac{1}{10}(5) = 0.5, \qquad \bar{v} = 0, \qquad \sigma_u^{\,2} = \frac{1}{n}\sum u^2 - (\bar{u})^2 = \frac{1}{10}(84) - \left(\frac{1}{2}\right)^2 = 8.4 - 0.25 = 8.15$$

$$\sigma_v^{\,2} = \sum v^2 - (\bar{v})^2 = \frac{1}{10}(24) = 2.4, \qquad Cov\,(u,v) = \frac{1}{n}\sum uv - \bar{u}\,\bar{v} = \frac{1}{10} \times 43 = 4.3$$

and $\quad r(u,v) = \dfrac{Cov(u,v)}{\sigma_u \sigma_v} = \dfrac{4.3}{\sqrt{8.15 \times 2.4}} = 0.972$

Since Coefficient of correlation is independent of change of origin

$\therefore \qquad r(x, y) = r(u,v) = 0.972$

Moreover we know that if $\quad u = \dfrac{x-a}{h}, \quad v = \dfrac{y-b}{k}$, *then*

$$\bar{x} = a + h\,\bar{u}, \qquad \bar{y} = b + k\bar{v}$$

Here $\quad h = k = 1$, $a = 22, b = 19 \qquad \therefore\ \bar{x} = 22 + 0.5 = 22.5, \qquad \bar{y} = 19 + 0 = 19$

Also, $\quad \sigma_x = h\sigma_u = 2.85, \qquad \sigma_y = k\sigma_v = 1.54$

Now the equation of line of regression of Y on X is

$$y-\bar{y} = \frac{r\sigma_y}{\sigma_x}(x-\bar{x}), \text{ i.e., } y-19 = \frac{(0.972)(1.54)}{2.85}(x-22.5)$$

or $y - 19 = 0.53\,(x - 22.5)$ or $\boldsymbol{y = 0.53x + 7.1}$

Similarly the equation of line of regression of X on Y is given by

$$x-\bar{x} = \frac{r\sigma_x}{\sigma_y}(y-\bar{y}) \quad \text{i.e.} \quad x - 22.5 = \frac{(0.972)(2.85)}{1.54}(y-19)$$ **or** $\boldsymbol{x = 1.79y - 11.5}$

Example 5.10 Two lines of regression are given as under:

$6x + 10y - 119 = 0$

$-30x + 45y + 180 = 0$

The variance of y is known to be 4.

Find (i) The mean values of x and y

(ii) The coefficient of correlation between x and y

(iii) The variance of x.

Solution: (i) We know that the mean values of x and y satisfy the regression equations

$\therefore$ $6\bar{x} + 10\bar{y} = 119$ (5.37)

$-30\bar{x} + 45\bar{y} = -180$ (5.38)

Multiplying equation (5.37) by 5, we get on adding $95\bar{y} = 415$; or $\bar{y} = 4.37$

Putting the values of $\bar{y}$ in equation (5.37), we get

$6\bar{x} + 10(4.37) = 119$

$6\bar{x} = 119 - 43.7 = 75.3$

$\bar{x} = 12.55$

(ii) The Correlation coefficient:

For calculating the correlation coefficient, we have to find out the regression coefficients. Treating equation (5.37) as the line of regression of x on y, we get

$$6x = 119 - 10y, \text{ i.e., } x = \frac{119}{6} - \frac{10}{6}y; \; b_{xy} = \frac{-10}{6}$$

Treating equation (5.38) as the line of regression of y on x, we find that

$$-45y = -30x - 180, \text{ i.e., } y = \frac{30}{45}x + \frac{180}{45}; \; and \; so \; b_{yx} = \frac{30}{45}$$

Since it is not possible that one of the regression coefficients is positive and another is negative, so, our assumption while assuming the line of regression was wrong, hence we consider (5.37) as the equation of y on x and that of (5.38) as the equation of x on y.

So, from equation (5.37), we have $10y = 119 - 6x$

or $$y = \frac{119}{10} - \frac{6}{10}x; \; and \; so \; b_{yx} = \frac{-6}{10}$$

and from equation (5.38), we have $-30x = -180 + 45y$

or $$x = \frac{180}{30} - \frac{45}{30}y \;\; or \;\; b_{xy} = -\frac{45}{30}$$

thus on reversing the line of regression, both the regression coefficients are of the same sign.

$$\therefore \quad r = \sqrt{\frac{-6}{10}\times\frac{-45}{30}} = -0.949$$

here we have taken negative sign of r as the regression coefficients have negative sign.

(iii) Variance of x: We know that $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ $\therefore$ $-0.6 = -0.949\;\frac{2}{\sigma_x}$

or $-0.6\sigma_x = -1.898 \;\; or \;\; \sigma_x = 3.163$ or variance $x = 10.004$