CHAPTER TEN

Correlation

LEARNING OBJECTIVES. Upon completion of this chapter, you should be able to :

1. **Understand the meaning of the term correlation and significance of its study.**
2. **Compute and interpret Karl Pearson's correlation coefficient, *r*.**
3. **Derive important properties of correlation coefficient.**
4. **Understand the meaning and the significance of probable error of *r*.**
5. **Distinguish between Karl Pearson's correlation coefficient and Spearman's rank correlation coefficient.**

CHAPTER OUTLINE

10·1. INTRODUCTION

10·2. MEANING OF CORRELATION

10·3. SCATTER DIAGRAM

10·4. KARL PEARSON'S COEFFICIENT OF CORRELATION

- 10·4·1. Limits for Correlation Coefficient
- 10·4·2. Assumptions Underlying Karl Pearson's Correlation Coefficient

10·5. CALCULATION OF THE CORRELATION COEFFICIENT FOR A BIVARIATE FREQUENCY DISTRIBUTION

10·6. PROBABLE ERROR OF CORRELATION COEFFICIENT

10·7. RANK CORRELATION

- 10·7·1. Spearman's Rank Correlation Coefficient
- 10·7·2. Tied or Repeated Ranks
- 10·7·3. Repeated Ranks (continued)

CHAPTER CONCEPTS QUIZ / DISCUSSION & REVIEW QUESTIONS / ASSORTED REVIEW PROBLEMS FOR SELF-ASSESSMENT

10·1. INTRODUCTION

So far we have confined ourselves to univariate distributions, *i.e.*, the distributions involving only one variable. Often we come across situations in which our focus is simultaneously on two or more variables and invariably, we observe that movements in one variable are accompanied by movements in other variable. For example, husband's age and wife's age move together, scores on an I.Q. test move with scores in university examinations. Similarly, studies in income and expenditure on households or price and demand of commodities, exhibit accompaning movements of two variables. Notwithstanding the cases of spurious or nonsensical relations which we may enjoy through such funny combinations of variables as suggesting that the number of runs scored by a batsman increases with an increase in the consumption of fertilizer in the local market or the number of flights space is increasing with a decrease in the population of tigers, the study of variables indicating accompaning behaviour is of great interest in Statistics.

However, the above statements are not precise enough to be of use to decision makers. We are therefore on the look out for a *quantitative measure* of the relationship between the two variables, and also for an appropriate mathematical or statistical form of the relationship. While the second question will be discussed in the next chapter, we will take up the first question in the present chapter.

10·2. MEANING OF CORRELATION

In a bivariate distribution we may be interested to find out if there is any correlation or covariation 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 deviate in the same direction, *i.e.*, if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be *direct* or *positive*. But if they constantly deviate in the opposite directions, *i.e.*, if increase (or decrease) in one results in corresponding decrease (or increase) in the other, correlation is said to be *diverse* or *negative*. For example, the correlation between (*i*) the heights and weights of a group of persons, and (*ii*) the income and expenditure; is positive and the correlation between (*i*) price and demand of a commodity and (*ii*) the volume and pressure of a perfect gas; is negative. Correlation is said to be *perfect* if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

10·3. SCATTER DIAGRAM

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

10·4. KARL PEARSON'S COEFFICIENT OF CORRELATION

As a measure of itensity or degree of linear relationship between two variables, Karl Pearson (1867-1936), *a British Biometrician*, developed a formula called *Correlation Coefficient.*

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

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \qquad \ldots(10\cdot1)$$

If (x_i, y_i); $i = 1, 2, \ldots, n$ is the bivariate distribution, then

$$\left.\begin{aligned} \text{Cov}(X, Y) &= E[\{X - E(X)\}\{(Y - E(Y)\}] \\ &= \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \mu_{11} \\ \sigma_X^2 &= E\{X - E(X)\}^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2 \\ \sigma_Y^2 &= E\{Y - E(Y)\}^2 = \frac{1}{n}\Sigma(y_i - \bar{y})^2 \end{aligned}\right\} \qquad \ldots(10\cdot2)$$

the summation extending over i from 1 to n.

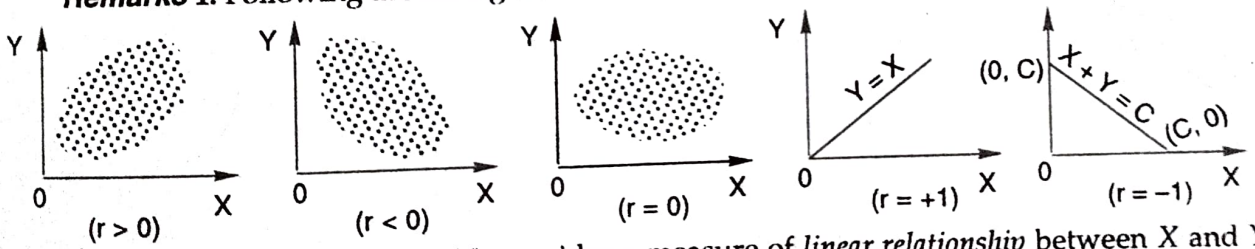
Another convenient form of the formula (10·2) for computational work is as follows :

$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})$$

$$= \frac{1}{n}\Sigma x_i y_i - \bar{y}\frac{1}{n}\Sigma x_i - \bar{x}\frac{1}{n}\Sigma y_i + \bar{x}\bar{y}$$

$$\therefore \quad \text{Cov}(X, Y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y}, \quad \sigma_X^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2 \text{ and } \sigma_Y^2 = \frac{1}{n}\Sigma y_i^2 - \bar{y}^2 \ldots (10\cdot2a)$$

Remarks **1.** Following are the figures of the standard data for $r > 0, < 0, = 0$, and $r = \pm 1$:



2. It may be noted that $r(X, Y)$ provides a measure of *linear relationship* between X and Y. For non-linear relationship, however, it is not very suitable.

3. Sometimes, we write : $\text{Cov}(X, Y) = \sigma_{XY}$.

4. Karl Pearson's correlation coefficient is also called *product-moment correlation coefficient*, since

$$\text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = \mu_{11}.$$

10·4·1. Limits for Correlation Coefficient. We have

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\left[\frac{1}{n}\Sigma(x_i - \bar{x})^2 \cdot \frac{1}{n}\Sigma(y_i - \bar{y})^2\right]^{1/2}},$$

$$\therefore \quad r^2(X, Y) = \frac{\left(\sum_i a_i b_i\right)^2}{\left(\sum_i a_i^2\right)\left(\sum_i b_i^2\right)}, \qquad \text{where} \begin{pmatrix} a_i = x_i - \bar{x} \\ b_i = y_i - \bar{y} \end{pmatrix} \qquad \ldots(*)$$

We have the Schwartz inequality which states that if $a_i, b_i, i = 1, 2, \ldots, n$ are real quantities then

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \le \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right),$$

the sign of equality holding if and only if $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \ldots = \frac{a_n}{b_n}$

Using Schwartz inequality, we get from (*) :

$$r^2(X, Y) \le 1 \text{ i.e., } |r(X, Y)| \le 1 \quad \Rightarrow \quad -1 \le r(X, Y) \le 1 \qquad \ldots(10\cdot3)$$

Hence correlation coefficient cannot exceed unity numerically. It always lies between –1 and +1. *If r = +1, the correlation is perfect and positive and if r = –1, correlation is perfect and negative.*

Aliter. If we write $E(X) = \mu_X$ and $E(Y) = \mu_Y$, then $E\left[\left(\frac{X-\mu_X}{\sigma_X}\right) \pm \left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]^2 \ge 0$

$$\Rightarrow \quad E\left(\frac{X-\mu_X}{\sigma_X}\right)^2 + E\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2 \pm 2\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \ge 0 \quad \Rightarrow \quad 1 + 1 \pm 2\, r(X, Y) \ge 0$$

$$\therefore \qquad -1 \le r(X, Y) \le 1.$$

Theorem 10·1. *Correlation coefficient is independent of change of origin and scale.*

Proof. Let $U = \frac{X-a}{h}$, $V = \frac{Y-b}{k}$, so that $X = a + hU$ and $Y = b + kV$, where a, b, h, k are constants; $h > 0, k > 0$.

We shall prove that $r(X, Y) = r(U, V)$.

Since $X = a + hU$ and $Y = b + kV$, on taking expectations, we get

$E(X) = a + hE(U)$ and $E(Y) = b + kE(V)$

$\therefore \quad X - E(X) = h[U - E(U)]$ and $Y - E(Y) = k[V - E(V)]$

$$\Rightarrow \quad \text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = E[h\{U - E(U)\}\{k\{V - E(V)\}]$$

$$= hk\, E[\{U - E(U)\}\{V - E(V)\}] = hk\, \text{Cov}(U, V) \qquad \ldots(10\cdot4)$$

$$\sigma_X^2 = E[\{X - E(X)\}^2] = E[h^2\{U - E(U)\}^2] = h^2\sigma_U^2$$

$$\Rightarrow \quad \sigma_X = h\sigma_U, (h > 0) \qquad \ldots(10\cdot4a)$$

and $\sigma_Y^2 = E[\{Y - E(Y)\}^2] = E[k^2\{V - E(V)\}^2] = k^2\sigma_V^2$

$$\Rightarrow \quad \sigma_Y = k\sigma_V, (k > 0) \qquad \ldots(10\cdot4b)$$

Substituting from (10·4), (10·4*a*) and (10·4*b*) in (10·1), we get

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{hk\, \text{Cov}(U, V)}{hk\, \sigma_U \sigma_V} = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = r(U, V).$$

Remark. This theorem is of fundamental importance in the numerical computation of the correlation coefficient.

Corollary. *If X and Y are random variables and a, b, c, d are any numbers provided only that* $a \ne 0, c \ne 0$, *then*

$$r(aX + b, cY + d) = \frac{ac}{|ac|} r(X, Y). \qquad \ldots(10\cdot4c)$$

Proof. With usual notations, we have

$$\text{Var}(aX + b) = a^2\sigma_X^2; \qquad \text{Var}(cY + d) = c^2\sigma_Y^2;$$

$$\text{Cov}(aX + b, cY + d) = ac\sigma_{XY}$$

$$\therefore \quad r(aX+b, cY+d) = \frac{\text{Cov}(aX+b, cY+d)}{[\text{Var}(aX+b)\,\text{Var}(cY+d)]^{1/2}} = \frac{ac\,\sigma_{XY}}{|a|\,|c|\,\sigma_X\sigma_Y} = \frac{ac}{|ac|}\,r(X, Y)$$

If $ac > 0$, *i.e.*, if a and c are of same signs, then $ac/|ac| = +1$.

If $ac < 0$, *i.e.*, if a and c are of opposite signs, then $ac/|ac| = -1$.

Remark. In particular, if we take $b = 0 = d$, then from (10·4*c*), we get

$$r(aX, cY) = \frac{ac}{|ac|}\,r(X, Y). \qquad \ldots (10\cdot4d)$$

Theorem 10·2. *Two independent variables are uncorrelated.*

Proof. If X and Y are independent variables, then

$$\text{Cov}(X, Y) = 0 \quad \Rightarrow \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} = 0$$

Hence two independent variables are uncorrelated.

But the *converse* of the theorem is not true, *i.e.*, two uncorrelated variables may not be independent as the following example illustrates :

X	-3	-2	-1	1	2	3	Total $\Sigma X = 0$
Y	9	4	1	1	4	9	$\Sigma Y = 28$
XY	-27	-8	-1	1	8	27	$\Sigma XY = 0$

$$\bar{X} = \frac{1}{n}\Sigma X = 0, \qquad \text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\,\bar{Y} = 0$$

$$\therefore \qquad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\,\sigma_Y} = 0$$

Thus in the above example, the variables X and Y are uncorrelated. But on careful examination we find that X and Y are not independent but they are connected by the relation $Y = X^2$. Hence two uncorrelated variables need not necessarily be independent. A simple reasoning for this strange conclusion is that $r(X, Y) = 0$, merely implies the absence of any linear relationship between the variables X and Y. There may, however, exist some other form of relationship between them, *e.g.*, quadratic, cubic or trigonometric.

Remarks **1.** Following are some more examples where two variables are uncorrelated but *not* independent.

(*i*) $X \sim N(0, 1)$ and $Y = X^2$. Since $X \sim N(0, 1)$, $E(X) = 0 = E(X^3)$

$$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)\,E(Y) = E(X^3) - E(X)\,E(Y) = 0 \qquad (\because Y = X^2)$$

$$\Rightarrow \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\,\sigma_Y} = 0.$$

Hence X and Y are uncorrelated but not independent ($\because Y = X^2$).

(*ii*) Let X be a *r.v.* with p.*d.f.* $f(x) = \frac{1}{2}$, $-1 \le x \le 1$, and let $Y = X^2$.

Here we shall get

$$E(X) = 0 \text{ and } E(XY) = E(X^3) = 0, \text{ so that } r(X, Y) = 0$$

2. However, the converse of the theorem holds in the following cases :

(*a*) If X and Y are jointly normally distributed with $\rho = \rho(X, Y) = 0$, then they are independent. If $\rho = 0$, then [*c.f.* § 11·3, Equation (11·7)]

$$f(x, y) = \frac{1}{\sigma_X \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{X-\mu_X}{\sigma_X}\right)^2\right] \times \frac{1}{\sigma_Y \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2\right] = f_1(x)\, f_2(y)$$

Hence X and Y are independent.

(*b*) *If each of the two variables X and Y takes two values, 0, 1 with positive probabilities, then* $r(X, Y) = 0 \Rightarrow X$ *and Y are independent.*

Proof. Let X take the values 1 and 0 with positive probabilities p_1 and q_1 respectively and let Y take the values 1 and 0 with positive probabilities p_2 and q_2 respectively. Then

$$r(X, Y) = 0 \quad \Rightarrow \quad \text{Cov}(X, Y) = 0$$

$$\Rightarrow \quad 0 = E(XY) - E(X)\,E(Y) = 1 \,.\, P(X = 1 \cap Y = 1) - [1 \,.\, P(X = 1) \times 1 \,.\, P(Y = 1)]$$

$$= P(X = 1 \cap Y = 1) - p_1 p_2$$

$$\Rightarrow \quad P(X = 1 \cap Y = 1) = p_1 p_2 = P(X = 1) \,.\, P(Y = 1).$$ Hence X and Y are independent.

10·4·2. Assumptions Underlying Karl Pearson's Correlation Coefficient. Pearsonian correlation coefficient *r* is based on the following assumptions :

(*i*) *The variables X and Y under study are linearly related.* In other words, the scatter diagram of the data will give a straight line curve.

(*ii*) *Each of the variables (series) is being affected by a large number of independent contributory causes of such a nature as to produce normal distribution.* For example, the variables (series) relating to ages, heights, weights, supply, price, etc., conform to this assumption. In the words of Karl Pearson :

"The sizes of the complex of organs (something measurable) are determined by a great variety of independent contributory causes, for example, climate, nourishment, physical training and innumerable other causes which cannot be individually observed or their effects measured." Karl Pearson further observes, "The variations in intensity of the contributory causes are small as compared with their absolute intensity and these variations follow the normal law of distribution."

(*iii*) *The forces so operating on each of the variable series are not independent of each other but are related in a causal fashion.* In other words, cause and effect relationship exists between different forces operating on the items of the two variable series. These forces must be common to both the series. If the operating forces are entirely independent of each other and not related in any fashion, then there cannot be any correlation between the variables under study.

For example, the correlation coefficient between,

(*a*) the series of heights and incomes of individuals over a period of time,

(*b*) the series of marriage rate and the rate of agricultural production in a country over a period of time, and

(*c*) the series relating to the size of the shoe and intelligence of a group of individuals,

should be zero, since the forces affecting the two variable series in each of the above cases are entirely independent of each other.

However, if in any of the above cases the value of *r* for a given set of data is not zero, then such correlation is termed as *chance correlation* or *spurious* or *nonsense correlation.*

Example 10·1. *Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y) :*

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

Solution.

CALCULATIONS FOR CORRELATION COEFFICIENT

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
Total 544	552	37028	38132	37560

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{544}{8} = 68, \qquad \bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{8} \times 552 = 69$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma XY - \bar{X}\bar{Y}}{\sqrt{\left(\frac{1}{n}\Sigma X^2 - \bar{X}^2\right)\left(\frac{1}{n}\Sigma Y^2 - \bar{Y}^2\right)}}$$

$$= \frac{\frac{1}{8} \times 37560 - 68 \times 69}{\sqrt{\left\{\frac{37028}{8} - (68)^2\right\}\left\{\frac{38132}{8} - (69)^2\right\}}}$$

$$= \frac{4695 - 4692}{\sqrt{(4628{\cdot}5 - 4624)\,(4766{\cdot}5 - 4761)}} = \frac{3}{\sqrt{4{\cdot}5 \times 5{\cdot}5}} = 0{\cdot}603.$$

Aliter. (SHORT-CUT METHOD)

X	Y	$U = X - 68$	$V = Y - 69$	U^2	V^2	UV
65	67	−3	−2	9	4	6
66	68	−2	−1	4	1	2
67	65	−1	−4	1	16	4
67	68	−1	−1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
Total		0	0	36	44	24

$$\bar{U} = \frac{1}{n}\Sigma U = 0, \quad \bar{V} = \frac{1}{n}\Sigma V = 0, \quad \text{Cov}(U, V) = \frac{1}{n}\Sigma UV - \bar{U}\,\bar{V} = \frac{1}{8} \times 24 = 3$$

$$\sigma_U^2 = \frac{1}{n}\Sigma U^2 - (\bar{U})^2 = \frac{1}{8} \times 36 = 4{\cdot}5, \quad \sigma_V^2 = \frac{1}{n}\Sigma V^2 - (\bar{V})^2 = \frac{1}{8} \times 44 = 5{\cdot}5$$

$$\therefore \quad r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{3}{\sqrt{4{\cdot}5 \times 5{\cdot}5}} = 0{\cdot}603 = r(X, Y)$$

Remark. The reader is advised to calculate the correlation coefficient by arbitrary origin method rather than by the direct method, since the latter leads to much simpler arithmetical calculations.

Example 10·2. *A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results :*

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

If was, however, later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values were

X	Y
8	12
6	8

. Obtain the correct value of correlation coefficient.

Solution.

$$\text{Corrected } \Sigma X = 125 - 6 - 8 + 8 + 6 = 125$$

$$\text{Corrected } \Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$$

$$\text{Corrected } \Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$$

$$\text{Corrected } \Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$$

$$\text{Corrected } \Sigma XY = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520$$

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{1}{25} \times 125 = 5, \qquad \bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{25} \times 100 = 4$$

$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\,\bar{Y} = \frac{1}{25} \times 520 - 5 \times 4 = \frac{4}{5}$$

$$\sigma_X^2 = \frac{1}{n}\Sigma X^2 - \bar{X}^2 = \frac{1}{25} \times 650 - (5)^2 = 1; \qquad \sigma_Y^2 = \frac{1}{n}\Sigma Y^2 - \bar{Y}^2 = \frac{1}{25} \times 436 - 16 = \frac{36}{25}$$

$$\therefore \quad \text{Corrected } r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{4/5}{\{1 \times (6/5)\}} = \frac{2}{3} = 0{\cdot}67$$

Example 10·3. *Show that if X′, Y′ are the deviations of the random variables X and Y from their respective means, then*

$$(i) \quad r = 1 - \frac{1}{2N}\sum_i \left(\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y}\right)^2 \qquad (ii) \quad r = -1 + \frac{1}{2N}\sum_i \left(\frac{X_i'}{\sigma_X} + \frac{Y_i'}{\sigma_Y}\right)^2.$$

Deduce that $-1 \le r \le +1$.

Solution. (*i*) Here $X_i' = (X_i - \bar{X})$ and $Y'_i = (Y_i - \bar{Y})$

$$\text{R.H.S.} = 1 - \frac{1}{2N}\sum_i \left(\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y}\right)^2 = 1 - \frac{1}{2N}\sum_i \left[\frac{X_i'^2}{\sigma_X^2} + \frac{Y_i'^2}{\sigma_Y^2} - \frac{2X_i'Y_i'}{\sigma_X \sigma_Y}\right]$$

$$= 1 - \frac{1}{2N}\left[\frac{1}{\sigma_X^2}\sum_i X_i'^2 + \frac{1}{\sigma_Y^2}\sum_i Y_i'^2 - \frac{2}{\sigma_X \sigma_Y}\sum_i X_i' Y_i\right]$$

$$= 1 - \frac{1}{2N}\left[\frac{1}{\sigma_X^2}\sum_i (X_i - \bar{X})^2 + \frac{1}{\sigma_Y^2}\sum_i (Y_i - \bar{Y})^2 - \frac{2}{\sigma_X \sigma_Y}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})\right]$$

$$= 1 - \frac{1}{2}\left(\frac{1}{\sigma_X^2}\cdot\sigma_X^2 + \frac{1}{\sigma_Y^2}\cdot\sigma_Y^2 - \frac{2}{\sigma_X\sigma_Y}\cdot r\sigma_X\sigma_Y\right) = 1 - \frac{1}{2}(1 + 1 - 2r) = r$$

(*ii*) Proceeding similarly, we will get R.H.S. $= -1 + \frac{1}{2}(1 + 1 + 2r) = r$

Deduction. Since $\left(\frac{X_i'}{\sigma_X} \pm \frac{Y_i'}{\sigma_Y}\right)^2$, being the square of a real quantity is always non-negative, $\sum_i \left(\frac{X_i'}{\sigma_X} \mp \frac{Y_i'}{\sigma_Y}\right)^2$ is also non-negative. From part (*i*), we get

$$r = 1 - (\text{some non-negative quantity}) \Rightarrow r \le 1 \qquad \ldots(*)$$

Also from part (*ii*), we get

$$r = -1 + (\text{some non-negative quantity}) \Rightarrow r \ge -1 \Rightarrow -1 \le r \qquad \ldots(**)$$

The sign of equality in (*) and (**) holds if and only if

$\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y} = 0$ and $\frac{X_i'}{\sigma_X} + \frac{Y_i'}{\sigma_Y} = 0$, $(\forall\, i = 1, 2, \ldots, n)$ respectively.

$\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y} = 0$, $\frac{X_i'}{\sigma_X} + \frac{Y_i'}{\sigma_Y} = 0$, $\forall\, i = 1, 2, \ldots, n$ respectively.

From (*) and (**), we conclude that $-1 \le r \le 1$.

Example 10·4. *The variables X and Y are connected by the equation $aX + bY + c = 0$. Show that the correlation between them is –1 if the signs of a and b are alike and + 1 if they are different.*

Solution. $aX + bY + c = 0 \Rightarrow aE(X) + bE(Y) + c = 0$

$\therefore$ $a[X - E(X)] + b[Y - E(Y)] = 0 \Rightarrow [X - E(X)] = -\frac{b}{a}[Y - E(Y)]$

$\therefore$ $\text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = -\frac{b}{a}E[\{Y - E(Y)\}^2] = -\frac{b}{a}\cdot\sigma_Y^2$

and $\sigma_X^2 = E\{X - E(X)\}^2 = \frac{b^2}{a^2}E[\{Y - E(Y)\}^2] = \frac{b^2}{a^2}\cdot\sigma_Y^2$

$\therefore$ $r = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} = \frac{-\frac{b}{a}\cdot\sigma_Y^2}{\sqrt{\sigma_Y^2}\sqrt{\frac{b^2}{a^2}\sigma_Y^2}} = \frac{-\frac{b}{a}\sigma_Y^2}{\left|\frac{b}{a}\right|\sigma_Y^2}$

$$= \begin{cases} +1, \text{ if } b \text{ and } a \text{ are of opposite signs.} \\ -1, \text{ if } b \text{ and } a \text{ are of same sign.} \end{cases}$$

Example 10·5. *(a) If $Z = aX + bY$ and r is the correlation coefficient between X and Y, show that $\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X\sigma_Y$.*

(b) Show that the correlation coefficient r between two random variables X and Y is given by

$$r = (\sigma_X^2 + \sigma_Y^2 - \sigma^2_{X-Y}) / 2\sigma_X\sigma_Y,$$

where σ_X, σ_Y and σ_{X-Y} are the standard deviations of X, Y and X – Y respectively.

Solution. (*a*) Taking expectation of both sides of $Z = aX + bY$, we get

$$E(Z) = aE(X) + bE(Y)$$

$$\therefore \quad Z - E(Z) = a\{X - E(X)\} + b\{Y - E(Y)\}$$

Squaring and taking expectation of both sides, we get

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab \text{ Cov } (X, Y) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X\sigma_Y$$

(*b*) Taking $a = 1, b = -1$ in the above case,

$$Z = X - Y \text{ and } \sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X\sigma_Y$$

$$\therefore \quad r = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2\sigma_X\sigma_Y}.$$

Remark. In the above example, we have obtained

$$V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2ab \text{ Cov } (X, Y)$$

Similarly, we could obtain the result $V(aX - bY) = a^2 V(X) + b^2 V(Y) - 2ab$ Cov (X, Y)

The above results are useful in solving theoretical problems.

Example 10·6. *X and Y are two random variables with variances σ^2_X and σ^2_Y respectively and r is the coefficient of correlation between them. If $U = X + kY$ and $V = X + (\sigma_X/\sigma_Y)Y$, find the value of k so that U and V are uncorrelated.*

Solution. Taking expectations of $U = X + kY$ and $V = X + \frac{\sigma_X}{\sigma_Y} Y$, we get

$$E(U) = E(X) + kE(Y) \text{ and } E(V) = E(X) + \frac{\sigma_X}{\sigma_Y} E(Y)$$

$$\therefore \quad U - E(U) = \{X - E(X)\} + k\{Y - E(Y)\}$$

$$\text{and} \quad V - E(V) = \{X - E(X)\} + \frac{\sigma_X}{\sigma_Y}\{Y - E(Y)\}$$

$$\text{Cov } (U, V) = E[\{U - E(U)\}\{V - E(V)\}]$$

$$= E[\{X - E(X)\} + k(Y - E(Y))] \times [\{X - E(X)\} + \frac{\sigma_X}{\sigma_Y}\{Y - E(Y)\}]$$

$$= \sigma_X^2 + \frac{\sigma_X}{\sigma_Y} \text{Cov } (X, Y) + k \text{ Cov } (X, Y) + k\frac{\sigma_X}{\sigma_Y} \cdot \sigma_Y^2$$

$$= (\sigma_X^2 + k\sigma_X\sigma_Y) + \left(\frac{\sigma_X}{\sigma_Y} + k\right) \text{Cov } (X, Y)$$

$$= \sigma_X(\sigma_X + k\sigma_Y) + \left(\frac{\sigma_X + k\sigma_Y}{\sigma_Y}\right) \text{Cov } (X, Y)$$

$$= (\sigma_X + k\sigma_Y)\left[\sigma_X + \frac{\text{Cov } (X, Y)}{\sigma_Y}\right] = (\sigma_X + k\sigma_Y)(1 + r)\sigma_X$$

U and V will be uncorrelated if $r(U, V) = 0 \quad \Rightarrow \quad \text{Cov } (U, V) = 0$

i.e., if $(\sigma_X + k\sigma_Y)(1 + r)\sigma_X = 0 \quad \Rightarrow \quad \sigma_X + k\sigma_Y = 0 \quad \Rightarrow \quad k = -\frac{\sigma_X}{\sigma_Y}$.

Example 10·7. *The random variables X and Y are jointly normally distributed and U and V are defined by $U = X \cos \alpha + Y \sin \alpha$, $V = Y \cos \alpha - X \sin \alpha$.*

Solution. We have

$$E(X) = \int_{-\frac{1}{2}}^{\frac{1}{2}} x f_1(x)\,dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x \cdot 1\,dx = \left|\frac{x^2}{2}\right|_{-\frac{1}{2}}^{\frac{1}{2}} = 0$$

If $f(x, y)$ is the joint *p.d.f.* of X and Y, then

$$f(x, y) = f(y \mid x) f_1(x) = f(y \mid x) \qquad [\because f_1(x) = 1] \qquad \ldots (**)$$

$$E(XY) = \int_{-\frac{1}{2}}^{0} \int_{x}^{x+1} xy f(x, y)\,dx\,dy + \int_{0}^{\frac{1}{2}} \int_{-x}^{1-x} xy f(x, y)\,dx\,dy$$

$$= \int_{-\frac{1}{2}}^{0} \left(x \int_{x}^{x+1} y\,dy \right) dx + \int_{0}^{\frac{1}{2}} \left(x \int_{-x}^{1-x} y\,dy \right) dx \qquad \text{[From (*) and (**)]}$$

$$= \frac{1}{2} \int_{-\frac{1}{2}}^{0} x(2x+1)\,dx + \frac{1}{2} \int_{0}^{\frac{1}{2}} x(1-2x)\,dx$$

$$= \frac{1}{2}\left[\frac{2}{3}x^3 + \frac{x^2}{2}\right]_{-1/2}^{0} + \frac{1}{2}\left[\frac{1}{2}x^2 - \frac{2}{3}x^3\right]_{0}^{1/2}$$

$$= \frac{1}{2}\left(\frac{1}{12} - \frac{1}{8} - \frac{1}{12} + \frac{1}{8}\right) = 0$$

$\therefore$ $\quad \text{Cov}(XY) = E(XY) - E(X)\,E(Y) = 0 \quad \Rightarrow \quad r(X, Y) = 0$

Hence the variables X and Y are uncorrelated.

10·5. CALCULATION OF THE CORRELATION COEFFICIENT FOR A BIVARIATE FREQUENCY DISTRIBUTION

When the data are considerably large, they may be summarised by using a two-way table. Here, for each variable a suitable number of classes are taken, keeping in view the same considerations as in the univariate case. If there are m classes for X and n classes for Y, there will be in all $m \times n$ cells in the two-way table. By going through the pairs of values of X and Y, we can find the frequency for each cell. The whole set of cell frequencies will then define a *bivariate frequency distribution.* The column totals and row totals will give us the marginal distributions of X and Y respectively. A particular column or row will be called the conditional distribution of Y for given X or of X for given Y respectively.

Suppose that the bivariate data on X and Y are presented in a two-way correlation table (shown on *page 10·18*) where there are n classes of Y placed along the horizontal lines and m classes of X along the vertical lines and f_{ij} is the frequency of individuals lying in the (i, j)th cell.

Here $\sum_x f(x, y) = g(y)$, is the sum of the frequencies along any row and

$\sum_y f(x, y) = f(x)$ is the sum of the frequencies along any column.

We observe that $\sum_x \sum_y f(x, y) = \sum_y \sum_x f(x, y) = \sum_x f(x) = \sum_y g(y) = N$

$$\therefore \quad \bar{x} = \frac{1}{N}\sum_x \sum_y x f(x, y) = \frac{1}{N}\left[\sum_x \left\{x \sum_y f(x, y)\right\}\right] = \frac{1}{N}\sum_x x f(x)$$

Similarly

$$\bar{y} = \frac{1}{N}\sum_x \sum_y y f(x, y) = \frac{1}{N}\sum_y y\, g(y)$$

$$\sigma_X^2 = \frac{1}{N}\sum_x \sum_y x^2 f(x, y) - \bar{x}^2 = \frac{1}{N}\sum_x x^2 f(x) - \bar{x}^2$$

$$\sigma_Y^2 = \frac{1}{N}\sum_x \sum_y y^2 f(x, y) - \bar{y}^2 = \frac{1}{N}\sum_y y^2 g(y) - \bar{y}^2$$

BIVARIATE FREQUENCY TABLE (CORRELATION TABLE)

Y Series ↓ / X Series →	Classes: Mid Points x_1 x_2 ... x_i ... x_m	Total of Frequencies of Y $g(y)$
y_1		
y_2		
⋮		
y_j	$f(x, y)$	$g(y) = \sum_x f(x, y)$
⋮		
⋮		
y_n		
Total of frequencies of X $f(x)$	$f(x) = \sum_y f(x, y)$	$N \to \sum_x \sum_y f(x, y)$ ↓ $\sum_y \sum_x f(x, y)$

Example 10·14. *The following table gives, according to age, the frequency of marks obtained by 100 students in an intelligence test :*

Ages in years → / Marks ↓	18	19	20	21	Total
10—20	4	2	2	—	8
20—30	5	4	6	4	19
30—40	6	8	10	11	35
40—50	4	4	6	8	22
50—60	—	2	4	4	10
60—70	—	2	3	1	6
Total	19	22	31	28	100

Example 10·15. *The joint probability distribution of X and Y is given below :*

Y \ X	−1	+1
0	$\frac{1}{8}$	$\frac{3}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$

Find the correlation coefficient between X and Y.

Solution.

COMPUTATION OF MARGINAL PROBABILITIES

Y \ X	−1	+1	$g(y)$
0	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{4}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{4}{8}$
$p(x)$	$\frac{3}{8}$	$\frac{5}{8}$	1

We have

$$E(X) = \Sigma x p(x) = (-1)\times\frac{3}{8} + 1\times\frac{5}{8} = \frac{1}{4}, \qquad E(X^2) = \Sigma x^2 p(x) = (-1)^2\times\frac{3}{8} + 1^2\times\frac{5}{8} = 1$$

$$\therefore \quad \text{Var}(X) = E(X^2) - [E(X)]^2 = 1 - \frac{1}{16} = \frac{15}{16}$$

$$E(Y) = \Sigma y\, g(y) = 0\times\frac{4}{8} + 1\times\frac{4}{8} = \frac{1}{2}, \qquad E(Y^2) = \Sigma y^2\, g(y) = 0^2\times\frac{4}{8} + 1^2\times\frac{4}{8} = \frac{1}{2}$$

$$\therefore \quad \text{Var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

$$E(XY) = 0\times(-1)\times\frac{1}{8} + 0\times 1\times\frac{3}{8} + 1\times(-1)\times\frac{2}{8} + 1\times 1\times\frac{2}{8} = -\frac{2}{8} + \frac{2}{8} = 0$$

$$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)\,E(Y) = 0 - \frac{1}{4}\times\frac{1}{2} = -\frac{1}{8}$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\,\sigma_Y} = \frac{-\frac{1}{8}}{\sqrt{\frac{15}{16}\times\frac{1}{4}}} = \frac{-1}{\sqrt{15}} = \frac{-1}{3{\cdot}873} = -0{\cdot}2582.$$

Example 10·16. *If X_1, X_2 and X_3 are uncorrelated variables with equal mean M and variances V_1^2, V_2^2 and V_3^2 respectively, prove that correlation coefficient ρ between $Z_1 = \frac{X_1}{X_3}$ and $Z_2 = \frac{X_2}{X_3}$ is given by ;* $\rho = \dfrac{V_3^2}{\sqrt{[(V_1^2 + V_3^2)(V_2^2 + V_3^2)}}$ ...

Solution. Let us make N observations on each of the variables X_1, X_2 and X_3. Then we are given :

$$\overline{X}_1 = \overline{X}_2 = \overline{X}_3 = M$$

Let x_i be the deviation of X_i, $(i = 1, 2, 3)$ from the mean M so that

$$x_{ij} = X_{ij} - M \quad \Rightarrow \quad X_{ij} = x_{ij} + M \ (i = 1, 2, 3);\ j = 1, 2, \ldots, N \qquad \ldots (i)$$

10·7. RANK CORRELATION

Let us suppose that a group of n individuals is arranged in order of merit or proficiency in possession of two characteristics A and B. These ranks in the two characteristics will, in general, be different. For example, if we consider the relation between intelligence and beauty, it is not necessary that a beautiful individual is intelligent also. Let (x_i, y_i); $i = 1, 2, \ldots, n$ be the ranks of the ith individual in two characteristics A and B respectively. Pearsonian coefficient of correlation between the ranks x_i's and y_i's is called the *rank correlation coefficient* between A and B for that group of individuals.

10·7·1. Spearman's Rank Correlation Coefficient. Assuming that no two individuals are bracketed equal in either classification, each of the variables X and Y takes the values $1, 2, \ldots, n$.

Hence $\quad \bar{x} = \bar{y} = \frac{1}{n}(1 + 2 + 3 + \ldots + n) = \frac{n+1}{2}$

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}(1^2 + 2^2 + \ldots + n^2) - \left(\frac{n+1}{2}\right)^2 = \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}$$

$$\therefore \qquad \sigma_X^2 = \frac{n^2-1}{12} = \sigma_Y^2 \qquad \ldots (*)$$

In general $x_i \neq y_i$. Let $d_i = x_i - y_i \quad \therefore \quad d_i = (x_i - \bar{x}) - (y_i - \bar{y}) \qquad (\because \bar{x} = \bar{y})$

Squaring and summing over i from 1 to n, we get

$$\sum_i d_i^2 = \sum_i [(x_i - \bar{x}) - (y_i - \bar{y})]^2 = \Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2 - 2\Sigma(x_i - \bar{x})(y_i - \bar{y})$$

Dividing both sides by n, we get

$$\frac{1}{n}\Sigma d_i^2 = \sigma_X^2 + \sigma_Y^2 - 2\,\mathrm{Cov}\,(X, Y) = \sigma_X^2 + \sigma_Y^2 - 2\rho\,\sigma_X\sigma_Y,$$

where ρ is the rank correlation coefficient between A and B.

$$\therefore \qquad \frac{1}{n}\Sigma d_i^2 = 2\sigma_X^2 - 2\rho\sigma_X^2 \quad \Rightarrow \quad 1 - \rho = \frac{\Sigma d_i^2}{2n\sigma_X^2}$$

$$\Rightarrow \qquad \rho = 1 - \frac{\sum_{i=1}^{n} d_i^2}{2n\sigma_X^2} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad \text{[From (*)]} \qquad \ldots(10\cdot7)$$

which is the *Spearman's formula for the rank correlation coefficient.*

Remark. We always have $\qquad \Sigma d_i = \Sigma(x_i - y_i) = \Sigma x_i - \Sigma y_i = n(\bar{x} - \bar{y}) = 0 \qquad (\because \bar{x} = \bar{y})$

This serves as a check on the calculations.

10·7·2. Tied Ranks. If some of the individuals receive the same rank in a ranking of merit, they are said to be tied. Let us suppose that m of the individuals, say, $(k+1)^{th}$, $(k+2)^{th}$, ..., $(k+m)^{th}$ are tied. Then each of these m individuals is assigned a common rank, which is arithmetic mean of the ranks $k+1, k+2, \ldots, k+m$.

Derivation of $\rho(X, Y)$. We have :

$$\rho(X, Y) = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{[\Sigma (X - \bar{X})^2 . \Sigma (Y - \bar{Y})^2]^{1/2}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}}, \quad \ldots(i)$$

where $x = X - \bar{X}$, $y = Y - \bar{Y}$.

If X and Y each takes the values 1, 2, ..., n, then $\bar{X} = (n + 1)/2 = \bar{Y}$

and $$n\sigma_X^2 = \Sigma x^2 = \frac{n(n^2 - 1)}{12} \text{ and } n\sigma_Y^2 = \Sigma y^2 = \frac{n(n^2 - 1)}{12} \quad \ldots(ii)$$

Also $$\Sigma d^2 = \Sigma (X - Y)^2 = \Sigma [(X - \bar{X}) - (Y - \bar{Y})]^2 = \Sigma (x - y)^2$$

$$\Rightarrow \quad \Sigma d^2 = \Sigma x^2 + \Sigma y^2 - 2\Sigma xy \quad \Rightarrow \quad \Sigma xy = \frac{1}{2}(\Sigma x^2 + \Sigma y^2 - \Sigma d^2) \quad \ldots(iii)$$

We shall now investigate the effect of common ranking (in case of ties), on the sum of squares of the ranks. Let S^2 and S_1^2 denote the sum of the squares of untied and tied ranks respectively. Then we have :

$$S^2 = (k + 1)^2 + (k + 2)^2 + \ldots + (k + m)^2$$

$$= mk^2 + (1^2 + 2^2 + \ldots + m^2) + 2k (1 + 2 + \ldots + m)$$

$$= mk^2 + \frac{m (m + 1) (2m + 1)}{6} + mk (m + 1)$$

$$S_1^2 = m \text{ (Average rank)}^2 = m \left[\frac{(k + 1) + (k + 2) + \ldots + (k + m)}{m} \right]^2$$

$$= m\left(k + \frac{m + 1}{2}\right)^2 = mk^2 + \frac{m (m + 1)^2}{4} + mk (m + 1)$$

$$\therefore \quad S^2 - S_1^2 = \frac{m(m + 1)}{12}[2 (2m + 1) - 3 (m + 1)] = \frac{m (m^2 - 1)}{12}$$

Thus the effect of tying m individuals (ranks) is to reduce the sum of the squares by $m (m^2 - 1)/12$, though the mean value of the ranks remains the same, *viz.*, $(n + 1)/2$.

Suppose that there are s such sets of ranks to be tied in the X-series so that the total sum of squares due to them is :

$$\frac{1}{12} \sum_{i=1}^{s} m_i (m_i^2 - 1) = \frac{1}{12} \sum_{i=1}^{s} (m_i^3 - m_i) = T_X, \text{ (say)} \quad \ldots (10{\cdot}7a)$$

Similarly suppose that there are t such sets of ranks to be tied with respect to the other series Y so that sum of squares due to them is :

$$\frac{1}{12} \sum_{j=1}^{t} m_j' (m_j'^2 - 1) = \frac{1}{12} \sum_{j=1}^{t} (m_j'^3 - m_j') = T_Y, \text{ (say)} \quad \ldots (10{\cdot}7b)$$

Thus, in the case of ties, the new sums of squares are given by :

$$n \text{ Var}'(X) = \Sigma x^2 - T_X = \frac{n(n^2 - 1)}{12} - T_X, \qquad n \text{ Var}'(Y) = \Sigma y^2 - T_Y = \frac{n(n^2 - 1)}{12} - T_Y$$

and $$n\,\text{Cov}'(X, Y) = \frac{1}{2}(\Sigma x^2 - T_X + \Sigma y^2 - T_Y - \Sigma d^2)$$ [From (*iii*)]

$$= \frac{1}{2}\left[\frac{n(n^2-1)}{12} - T_X + \frac{n(n^2-1)}{12} - T_Y - \Sigma d^2\right]$$

$$= \frac{n(n^2-1)}{12} - \frac{1}{2}[(T_X + T_Y) + \Sigma d^2]$$

$$\rho(X, Y) = \frac{\frac{n(n^2-1)}{12} - \frac{1}{2}(T_X + T_Y + \Sigma d^2)}{\left[\frac{n(n^2-1)}{12} - T_X\right]^{1/2}\left[\frac{n(n^2-1)}{12} - T_Y\right]^{1/2}}$$

$$= \frac{\{n(n^2-1)/6\} - (\Sigma d^2 + T_X + T_Y)}{\left[\frac{n(n^2-1)}{6} - 2T_X\right]^{1/2}\left[\frac{n(n^2-1)}{6} - 2T_Y\right]^{1/2}} \quad \ldots(10\cdot7c)$$

where T_X and T_Y are given by (10·7*a*) and (10·7*b*).

Remark. If we adjust only the covariance term, *i.e.*, Σxy and not the variances σ_X^2 (or Σx^2) and σ_Y^2 (or Σy^2) for ties, then the formula (10·7*c*) reduces to :

$$\rho(X, Y) = \frac{\{n(n^2-1)/6\} - (\Sigma d^2 + T_X + T_Y)}{n(n^2-1)/6} = 1 - \frac{6(\Sigma d^2 + T_X + T_Y)}{n(n^2-1)} \quad \ldots(10\cdot7d)$$

a formula which is commonly used in practice for numerical problems. For illustration, see *Example 10·19.*

Example 10·17. *The ranks of same 16 students in Mathematics and Physics are as follows. Two numbers within brackets denote the ranks of the students in Mathematics and Physics :*

(1, 1) (2, 10) (3, 3) (4, 4) (5, 5) (6, 7) (7, 2) (8, 6) (9, 8) (10, 11) (11, 15) (12, 9) (13, 14) (14, 12) (15, 16) (16, 13).

Calculate the rank correlation coefficient for proficiencies of this group in Mathematics and Physics.

Solution.

Ranks in Maths. (X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
Ranks in Physics (Y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13	
$d = X - Y$	0	−8	0	0	0	−1	5	2	1	−1	−4	3	−1	2	−1	3	0
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9	136

Rank correlation coefficient is given by :

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 136}{16 \times 255} = 1 - \frac{1}{5} = \frac{4}{5} = 0{\cdot}8.$$

Example 10·18. *Ten competitors in a musical test were ranked by the three judges A, B and C in the 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

Using rank correlation method, discuss which pair of judges has the nearest approach to common likings in music.

Solution. Here $n = 10$

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			$\Sigma d_1 = 0$	$\Sigma d_2 = 0$	$\Sigma d_3 = 0$	$\Sigma d_1^2 = 200$	$\Sigma d_2^2 = 60$	$\Sigma d_3^2 = 214$

$$\rho(X, Y) = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)} = 1 - \frac{6 \times 200}{10 \times 99} = 1 - \frac{40}{33} = -\frac{7}{33}$$

$$\rho(X, Z) = 1 - \frac{6\Sigma d_2^2}{n(n^2-1)} = 1 - \frac{6 \times 60}{10 \times 99} = 1 - \frac{4}{11} = \frac{7}{11}$$

$$\rho(Y, Z) = 1 - \frac{6\Sigma d_3^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = 1 - \frac{214}{165} = -\frac{49}{165}.$$

Since $\rho(X, Z)$ is maximum, we conclude that the pair of judges A and C has the nearest approach to common likings in music.

10·7·3. Repeated Ranks (Continued). If any two or more individuals are bracketed equal in any classification with respect to characteristics A and B, or if there is more than one item with the same value in the series, then the Spearman's formula (10·7) for calculating the rank correlation coefficient breaks down, since in this case each of the variables X and Y does not assume the values 1, 2, ..., n and consequently, $\bar{x} \neq \bar{y}$.

In this case, common ranks are given to the repeated items. This common rank is the average of the ranks which these items would have assumed if they were slightly different from each other and the next item will get the rank next to the ranks already assumed. As a result of this, following adjustment or correction is made in the rank correlation formula [see (10·7*d*)].

In the formula, we add the factor $\frac{m(m^2-1)}{12}$ to Σd^2, where m is 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.

Example 10·19. *Obtain the rank correlation coefficient for the following data :*

X :	68	64	75	50	64	80	75	40	55	64
Y :	62	58	68	45	81	60	68	48	50	70

Solution.

CALCULATIONS FOR RANK CORRELATION

X	Y	Rank X (x)	Rank Y (y)	$d = x - y$	d^2
68	62	4	5	–1	1
64	58	6	7	–1	1
75	68	2·5	3·5	–1	1
50	45	9	10	–1	1
64	81	6	1	5	25
80	60	1	6	–5	25
75	68	2·5	3·5	–1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
				$\Sigma d = 0$	$\Sigma d^2 = 72$

In the X-series we see that the value 75 occurs 2 times. The common rank given to these values is 2·5 which is the average of 2 and 3, the ranks which these values would have taken if they were different. The next value 68, then gets the next rank which is 4. Again we see that value 64 occurs thrice. The common rank given to it is 6 which is the average of 5, 6 and 7. Similarly in the Y-series, the value 68 occurs twice and its common rank is 3·5 which is the average of 3 and 4. As a result of these common rankings, the formula for 'ρ' has to be corrected. To Σd^2 we add $\frac{m(m^2-1)}{12}$ for each value repeated, where m is the number of times a value occurs. In the X-series the correction is to be applied twice, once for the value 75 which occurs twice ($m = 2$) and then for the value 64 which occurs thrice ($m = 3$). The total correction for the X-series is : $\frac{2(4-1)}{12} + \frac{3(9-1)}{12} = \frac{5}{2}$. Similarly, this correction for the Y-series is $\frac{2(4-1)}{12} = \frac{1}{2}$, as the value 68 occurs twice.

$$\therefore \quad \rho = 1 - \frac{6\left(\Sigma d^2 + \frac{5}{2} + \frac{1}{2}\right)}{n(n^2-1)} = 1 - \frac{6(72+3)}{10 \times 99} = 0{\cdot}545.$$

10·7·4. Limits for the Rank Correlation Coefficient. Spearman's Rank correlation coefficient is given by : $\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$.

'ρ' is maximum, if $\sum_{i=1}^{n} d_i^2$ is minimum, *i.e.*, if each of the deviations d_i is minimum. But the minimum value of d_i is zero in the particular case $x_i = y_i$, *i.e.*, if the ranks of the ith individual in the two characteristics are equal. Hence the maximum value of ρ is +1, *i.e.*, $\rho \leq 1$. 'ρ' is minimum, if $\sum_{i=1}^{n} d_i^2$ is maximum, *i.e.*, if each of the deviations d_i is

maximum which is so if the ranks of the n individuals in the two characteristics are in the opposite directions as given below :

x	1	2	3	...	...	$n-1$	n
y	n	$n-1$	$n-2$	...	...	2	1

... (*)

Case I. Suppose n is odd equal to $(2m + 1)$, then the values of d are :

$d : 2m, 2m-2, 2m-4, \ldots, 2, 0, -2, -4, \ldots -(2m-2), -2m.$

$$\therefore \quad \sum_{i=1}^{n} d_i^2 = 2[(2m)^2 + (2m-2)^2 + \ldots + 4^2 + 2^2]$$

$$= 8\left[m^2 + (m-1)^2 + \ldots + 2^2 + 1^2\right] = \frac{8m\,(m+1)\,(2m+1)}{6}$$

$$\text{Hence} \quad \rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{8m\,(m+1)(2m+1)}{(2m+1)\,[(2m+1)^2-1]} = 1 - \frac{8\,m(m+1)}{4m\,(m+1)} = -1.$$

Case II. Let n be even and equal to $2m$, (say).

Then the values of d are : $(2m-1), (2m-3), \ldots, 1, -1, -3, \ldots, -(2m-3), -(2m-1)$

$$\therefore \quad \Sigma d_i^2 = 2[(2m-1)^2 + (2m-3)^2 + \ldots + 1^2]$$

$$= 2[\{(2m)^2 + (2m-1)^2 + (2m-2)^2 + \ldots + 2^2 + 1^2\}$$

$$-\{(2m)^2 + (2m-2)^2 + \ldots + 4^2 + 2^2\}]$$

$$= 2\left[1^2 + 2^2 + \ldots + (2m)^2 - \{2^2m^2 + 2^2(m-1)^2 + \ldots + 2^2\}\right]$$

$$= 2\left[\frac{2m\,(2m+1)(4m+1)}{6} - 4\,\frac{m\,(m+1)\,(2m+1)}{6}\right]$$

$$= \frac{2m}{3}\left[(2m+1)\,(4m+1) - 2(m+1)\,(2m+1)\right]$$

$$= \frac{2m}{3}\left[(2m+1)\,(4m+1-2m-2)\right] = \frac{2m}{3}\,(2m+1)\,(2m-1) = \frac{2m\,(4m^2-1)}{3}$$

$$\therefore \quad \rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{4m\,(4m^2-1)}{2m\,(4m^2-1)} = -1$$

Thus the limits for rank correlation coefficient are given by $-1 \le \rho \le 1$.

Aliter. For an alternate and simpler proof for obtaining the minimum value of ρ, from (*) onward, proceed as in Hint to *Question Number 10·51* of Assorted Review Problems for Self-Assessment.

Remarks on Spearman's Rank Correlation Coefficient.

1. $\Sigma d = \Sigma x - \Sigma y = n(\bar{x} - \bar{y}) = 0$, which provides a check for numerical calculations.

2. Since Spearman's rank correlation coefficient ρ is nothing but Pearsonian correlation coefficient between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.

3. Karl Pearson's correlation coefficient assumes that the parent population from which sample observations are drawn is normal. If this assumption is violated, then we need a measure which is distribution free (or non-parametric). A distribution-free measure is one which does not make any assumptions about the parameters of the population. Spearman's ρ is

such a measure (*i.e.*, distribution-free), since no strict assumptions are made about the form of the population from which sample observations are drawn.

4. Spearman's formula is easy to understand and apply as compared with Karl Pearson's formula. The value obtained by the two formulae, *viz.*, Pearsonian r and Spearman's ρ, are generally different. The difference arises due to the fact that when ranking is used instead of full set of observations, there is always some loss of information. Unless many ties exist, the coefficient of rank correlation should be only slightly lower than the Pearsonian coefficient.

5. Spearman's formula is the only formula to be used for finding correlation coefficient if we are dealing with qualitative characteristics which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of extreme observations, Spearman's formula is preferred to Pearson's formula.

6. Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution (Correlation Table). For $n > 30$, this formula should not be used unless the ranks are given, since in the contrary case the calculations are quite time-consuming.

CHAPTER CONCEPTS QUIZ

1. State whether the following statements are TRUE or FALSE :

   (*i*) $r_{XY} = 0 \Rightarrow X$ and Y are independent.

   (*ii*) If $r_{XY} > 0$, then $r_{X,-Y} > 0$, $r_{-X,Y} > 0$ and $r_{-X,-Y} > 0$.

   (*iii*) $r_{XY} > 0 \Rightarrow E(XY) > E(X)\,E(Y)$.

   (*iv*) Pearson's coefficient of correlation is independent of origin but not of scale.

   (*v*) The numerical value of product moment correlation coefficient 'r' between two variables X and Y cannot exceed unity.

   (*vi*) If the correlation coefficient between the variables X and Y is zero then the correlation coefficient between X^2 and Y^2 is also zero.

   (*vii*) If $r > 0$, then as X increases, Y also increases.

   (*viii*) r measures every type of relationship between the two variables.

   (*ix*) "High positive coefficient of correlation between increase in the sale of newspapers and increase in the number of crimes leads to the conclusion that newspaper reading may be responsible for the increase in the number of crimes."

   (*x*) "A high positive value of r between the increase in cigarette smoking and increase in lung cancer establishes that cigarette smoking is responsible for lung cancer."

   (*xi*) Let (X, Y) be jointly discrete random variables such that each X and Y has at most two mass points. Then X and Y are independent if and only if they are uncorrelated.

2. Fill in the blanks :

   (*i*) If $r_{XY} = 1$, the line on the graph will extend from ... to ...

   (*ii*) If $r_{XY} = 0$, it depicts ... association.

   (*iii*) Correlation coefficient is a ... number.

   (*iv*) If r is the correlation coefficient, the $\sqrt{1 - r^2}$ is termed as ... .

   (*v*) The Karl Pearson coefficient of correlation between variables X and Y is .....

   (*vi*) Two independent variables are .....

   (*vii*) Limits for correlation coefficient are .....

   (*viii*) If r be the correlation coefficient between the random variables X and Y then the variance of $X + Y$ is .....

(*ix*) The absolute value of the product moment correlation coefficient is less than .....

(*x*) Correlation coefficient is invariant under changes of ..... and .....

3. Indicate the correct answer :

(*i*) The coefficient of correlation will have positive sign when (*a*) X is increasing and Y is decreasing, (*b*) both X and Y are increasing, (*c*) X is decreasing and Y is increasing, (*d*) there is no change in X and Y.

(*ii*) The coefficient of correlation (*a*) can take any value between – 1 and + 1, (*b*) is always less than – 1, (*c*) is always more than + 1, (*d*) cannot be zero.

(*iii*) The coefficient of correlation (*a*) cannot be positive, (*b*) cannot be negative, (*c*) is always positive, (*d*) can be both positive as well as negative.

(*iv*) Probable error of r is

(*a*) $0{\cdot}6745\,\frac{1-r^2}{\sqrt{n}}$, (*b*) $0{\cdot}6754\,\frac{1+r^2}{\sqrt{n}}$, (*c*) $0{\cdot}6547\,\frac{1-r^2}{n}$, (*d*) $0{\cdot}6754\,\frac{1-r^2}{n}$.

(*v*) The coefficient of correlation between X and Y is 0·6. Their covariance is 4·8. The variance of X is 9. Then the S.D. of Y is

(*a*) $\frac{4{\cdot}8}{3\times 0{\cdot}6}$, (*b*) $\frac{0{\cdot}6}{4{\cdot}8\times 3}$, (*c*) $\frac{3}{4{\cdot}8\times 0{\cdot}6}$, (*d*) $\frac{4{\cdot}8}{9\times 0{\cdot}6}$.

(*vi*) The coefficient of correlation is independent of (*a*) change of scale only, (*b*) change of origin only, (*c*) both change of scale and origin, (*d*) neither change of scale nor change of origin.

(*vii*) X, Y and Z are three uncorrelated variables having variances σ_X^2, σ_Y^2 and σ_Z^2 respectively, then the correlation coefficient between $X + Y$ and $Y + Z$ is :

(*a*) 0, (*b*) $\frac{1}{2}$, (*c*) 1, (*d*) None of the above.

(*viii*) Let X be normally distributed with mean 0 and variance σ_1^2 and Y be normally distributed with mean 0 and variance σ_2^2. Let ρ be the correlation coefficient between X and Y, then $(X/\sigma_1) + (Y/\sigma_2)$ and $(X/\sigma_1) - (Y/\sigma_2)$:

(*a*) have correlation coefficient = ρ (*b*) have correlation coefficient $\leq \rho$

(*c*) have correlation coefficient $\geq \rho$ (*d*) None of the above.

(*ix*) If $X_1, X_2, \ldots, X_n$ are n uncorrelated random variables following the same distribution, then the correlation coefficient between $\overline{X}$ and $X_i - \overline{X}$ is :

(*a*) 0, (*b*) 1, (*c*) $\frac{1}{2}$, (*d*) None of the above.

4. Prove or disprove :

(*a*) $r(X, Y) = 0 \Rightarrow r(X|, Y) = 0$ (*b*) $r(X, Y) = 0, r(Y, Z) = 0 \Rightarrow r(X, Z) = 0$

DISCUSSION AND REVIEW QUESTIONS

CHAPTER ELEVEN

Linear and Curvilinear Regression

LEARNING OBJECTIVES. Upon completion of this chapter, you should be able to :

1. **Understand the meaning of regression, and its role in statistical analysis.**
2. **Use regression analysis to develop equations for estimating the relationship between two variables.**
3. **Understand the concepts and meaning of regression coefficients, and their relationship with correlation coefficient.**
4. **Demonstrate the meaning of curvilinear regression and fitting of non-linear functions, using the Principle of Least squares .**
5. **Explain the concept and computation of regression curves for means.**
6. **Appreciate the use of regression analysis for estimation and prediction purposes**

CHAPTER OUTLINE

11·1. INTRODUCTION

11·2. LINEAR REGRESSION

- 11·2·1. Regression Coefficients
- 11·2·2. Properties of Regression Coefficients
- 11·2·3. Angle between Two Lines of Regression
- 11·2·4. Standard Error of Estimate or Residual Variance
- 11·2·5. Correlation Coefficient between Observed and Estimated Values

11·3. CURVILINEAR REGRESSION

11·4. REGRESSION CURVES

CHAPTER CONCEPTS QUIZ / DISCUSSION & REVIEW QUESTIONS / ASSORTED REVIEW PROBLEMS FOR SELF-ASSESSMENT

11·1. INTRODUCTION

The term *"regression"* literally means *"stepping back towards the average"*. It was first used by British biometrician Sir Francis Galton (1822 - 1911), in connection with the inheritance of stature. Galton found that the offsprings of abnormally tall or short parents tend to *"regress"* or *"step back"* to the average population height. But the term "regression" as now used in Statistics is only a convenient term without having any reference to biometry.

Definition. *Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.*

In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called *dependent variable* and the variable which influences the values or is used for prediction is called *independent variable*. In regression analysis independent variable is also known as *regressor or predictor* or *explanatory variable* while the dependent variable is also known as *regressed* or *explained variable*.

11·2 . LINEAR REGRESSION

If the variables in a bivariate distribution are related, we 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 line of regression is the line which gives the 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 in the bivariate distribution (x_i, y_i) ; $i = 1, 2, ..., n$; Y is dependent variable and X is independent variable. Let the line of regression of Y on X be

$$Y = a + bx \qquad \ldots (11\cdot1)$$

The above equation (11·1) represents a family of straight lines for different values of the arbitrary constants *'a'* and *' b'*. The problem is to determine *'a'* and *'b'* so that the line (11·1) is the line of *'best fit'*.

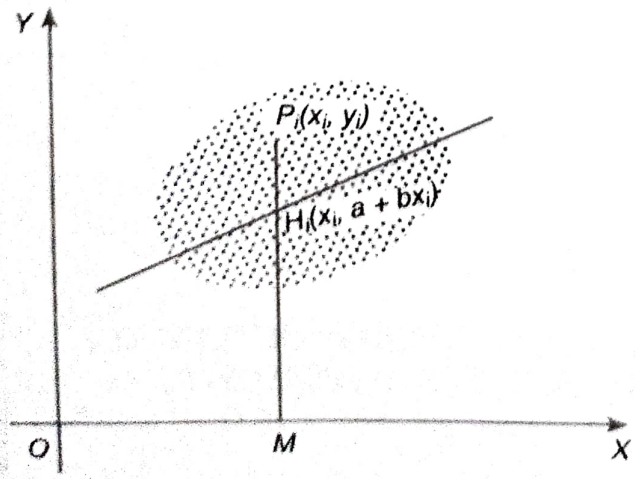


Fig. 11·1.

The term *'best fit'* is interpreted in accordance with Legender's principle of least squares which consists in minimising the sum of the squares of the deviations of the actual values of y from their estimated values as given by the line of best fit.

Let $P_i\,(x_i, y_i)$ be any general point in the scatter diagram (§ 10·2). Draw $P_i M$ $\perp$ to x-axis meeting the line, (11·1) in H_i. Abscissa of H_i is x_i and since H_i lies on (11·1), its ordinate is $a + bx_i$. Hence the co-ordinates of H_i are $(x_i, a + bx_i)$.

$$P_i H_i = P_i M - H_i M = y_i - (a + bx_i),$$

is called the *error of estimate* or the *residual* for y_i.

According to the principle of least squares, we have to determine a and b so that

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

is minimum. From the principle of maxima and minima, the partial derivatives of E, *w.r.* to a and b should vanish separately, *i.e.*,

$$\frac{\partial E}{\partial a} = 0 = -2 \sum_{i=1}^{n} (y_i - a - bx_i) \quad \text{and} \quad \frac{\partial E}{\partial b} = 0 = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i)$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \quad \ldots (11\cdot2) \quad \text{and} \quad \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \quad \ldots (11\cdot3)$$

Remark. Equations (11·2) and (11·3) are known as the *normal equations* for estimating a and b. All the quantities $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i y_i$, can be obtained from the given set of points (x_i, y_i) ; $i = 1, 2, \ldots, n$ and the equations (11·2) and (11·3) can be solved for a and b. With the values of a and b so obtained, equation (11·1) is the line of best fit to the given set of points $(x_i, y_i, i = 1, 2, \ldots, n$.

From (11·2) on dividing by n, we get $\bar{y} = a + b\bar{x}$ ... (11·4)

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

Now $\mu_{11} = \text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} \Rightarrow \frac{1}{n} \sum_{i=1}^{n} x_i y_i = \mu_{11} + \bar{x}\,\bar{y}$ ... (11·5)

Also $\sigma_X^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \Rightarrow \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \sigma_X^2 + \bar{x}^2$ ... (11·5*a*)

Dividing (11·3) by n and using (11·5) and (11·5*a*), we get

$$\mu_{11} + \bar{x}\,\bar{y} = a\bar{x} + b(\sigma_X^2 + \bar{x}^2) \quad \ldots (11\cdot6)$$

Multiplying (11·4) by $\bar{x}$ and then subtracting from (11·6), we get

$$\mu_{11} = b\,\sigma_X^2 \quad \Rightarrow \quad b = \frac{\mu_{11}}{\sigma_X^2} \quad \ldots (11\cdot7)$$

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 :

$$Y - \bar{y} = b(X - \bar{x}) = \frac{\mu_{11}}{\sigma_X^2}(X - \bar{x}) \quad \ldots (11\cdot8)$$

$$\Rightarrow \quad Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \quad \ldots (11\cdot8a)$$

---

* The general problem in curve fitting is to find, if possible, an analytic expression of the form $y = f(x)$, for the functional relationship suggested by the given data. Fitting of curves to a set of numerical data is of considerable importance, theoretical as well as practical. Theoretically it is useful in the study of correlation and regression, *e.g.*, lines of regression can be regarded as fitting of linear curves to the given bivariate distribution. In practical statistics it enables us to represent the relationship between two variables by simple algebraic expressions, *e.g.*, polynomials, exponential or logarithmic functions. Moreover, it may be used to estimate the values of one variable which would correspond to the specified values of the other variable.

Starting with the equation $X = A + BY$ and proceeding similarly or by simply interchanging the variables X and Y in (11·8) and (11·8a), the equation of the line of regression of X on Y becomes

$$X - \bar{x} = \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{y}) \qquad \ldots (11\cdot9)$$

$$\Rightarrow \qquad X - \bar{x} = r\,\frac{\sigma_X}{\sigma_Y}\,(Y - \bar{y}) \qquad \ldots (11\cdot9a)$$

Aliter. The straight line defined by $Y = a + bX$ ...(*i*)

and satisfying the residual (least square) condition $S = E\,[(Y - a - bX)^2] = \text{Min.}$ ... (*ii*)

for variations in a and b, is called the line of regression of Y on X.

The necessary and sufficient conditions for a minima of S, subject to variations in a and b are :

$$(i)\ \frac{\partial S}{\partial a} = 0, \frac{\partial S}{\partial b} = 0 \quad \ldots (*) \quad \text{and } (ii)\ \Delta = \begin{vmatrix} \frac{\partial S}{\partial a^2} & \frac{\partial^2 S}{\partial a \partial b} \\ \frac{\partial^2 S}{\partial b\,\partial a} & \frac{\partial^2 S}{\partial b^2} \end{vmatrix} > 0 \text{ and } \frac{\partial^2 S}{\partial a^2} > 0 \quad \ldots (**)$$

Using (*), we get

$$\frac{\partial S}{\partial a} = -2\,E\,[Y - a - bX] \quad = 0 \qquad \ldots (iii)$$

$$\frac{\partial S}{\partial b} = -2\,E\,[X\,(Y - a - bX)] \; = 0 \qquad \ldots (iv)$$

$$\Rightarrow \qquad E(Y) = a + bE(X) \quad \ldots (v) \quad \text{and} \quad E(XY) = aE(X) + bE(X^2) \qquad \ldots (vi)$$

Equation (*v*) implies that the line (*i*) of regression of Y on X passes through the mean value $[E(X), E(Y)]$.

Multiplying (*v*) by $E(X)$ and subtracting from (*vi*), we get

$$E\,(XY) - E(X)\,E\,(Y) = b\,[E(X^2) - \{E\,(X)\}^2]$$

$$\Rightarrow \qquad \text{Cov}\,(X, Y) = b\,\sigma_X^2 \quad \Rightarrow \quad b = \frac{\text{Cov}\,(X, Y)}{\sigma_X^2} = \frac{r\sigma_Y}{\sigma_X} \qquad \ldots (vii)$$

Subtracting (*v*) from (*i*) and using (*vii*), we obtain the equation of line of regression of Y on X as :

$$Y - E\,(Y) = \frac{\text{Cov}\,(X, Y)}{\sigma_X^2}\,[X - E\,(X)] \quad \Rightarrow \quad Y - E\,(Y) = \frac{r\sigma_Y}{\sigma_X}\,[X - E\,(X)]$$

Similarly, the straight line defined by $X = A + BY$ and satisfying the residual condition $E\,[X - A - BY]^2 = \text{Minimum}$, is called the line of regression of X on Y.

***Remarks* 1.** We note that

$$\frac{\partial^2 S}{\partial a^2} = 2 > 0, \quad \frac{\partial^2 S}{\partial b^2} = 2E\,(X^2) \quad \text{and} \quad \frac{\partial^2 S}{\partial a \partial b} = 2E(X)$$

Substituting in (**), we have

$$\Delta = \frac{\partial^2 S}{\partial a^2} \cdot \frac{\partial^2 S}{\partial b^2} - \left(\frac{\partial^2 S}{\partial a \partial b}\right)^2 = 4\,[E(X^2) - \{E\,(X)\}^2] = 4\sigma_X^2 > 0$$

Hence the solution of the least square equations (*iii*) and (*iv*), in fact provides a minima of S.

2. The regression equation (11·8*a*) implies that the line of regression of Y on X passes through the point $(\bar{x}, \bar{y})$. Similarly (11·9*a*) implies that the line of regression of X on Y also passes through the point $(\bar{x}, \bar{y})$. Hence both the lines of regression pass through the point $(\bar{x}, \bar{y})$. In other words, the mean values $(\bar{x}, \bar{y})$ can be obtained as the point of intersection of the two regression lines.

3. *Why two lines of Regression* ? There are always two lines of regression, one of Y on X and the other of X on Y. The line of regression of Y on X (11·8*a*) is used to estimate or predict the value of Y for any given value of X, *i.e.*, when Y is a dependent variable and X is an independent variable. The estimate so obtained will be best in the sense that it will have the minimum possible error as defined by the principle of least squares. We can also obtain an estimate of X for any given value of Y by using equation (11·8*a*) but the estimate so obtained will not be best since (11·8*a*) is obtained on minimising the sum of the squares of errors of estimates in Y and not in X. Hence to estimate or predict X for any given value of Y, we use the regression equation of X on Y (11·9*a*) which is derived on minimising the sum of the squares of errors of estimates in X. Here X is a dependent variable and Y is an independent variable. The two regression equations are not reversible or interchangeable because of the simple reason that the basis and assumptions for deriving these equations are quite different. The regression equation of Y on X is obtained on minimising the sum of the squares of the errors parallel to the Y-axis while the regression equation of X on Y is obtained on minimising the sum of squares of the errors parallel to the X-axis.

In a particular case of perfect correlation, positive or negative, *i.e.*, $r \pm 1$, the equation of line of regression of Y on X becomes :

$$Y - \bar{y} = \pm \frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \quad \Rightarrow \quad \frac{Y - \bar{y}}{\sigma_Y} = \pm\left(\frac{X - \bar{x}}{\sigma_X}\right) \qquad \ldots (11\cdot10)$$

Similarly, the equation of the line of regression of X on Y becomes :

$$Y - \bar{x} = \pm \frac{\sigma_X}{\sigma_Y}(Y - \bar{y}) \quad \Rightarrow \quad \frac{Y - \bar{y}}{\sigma_Y} = \pm\left(\frac{X - \bar{x}}{\sigma_X}\right)$$

which is same as (11·10).

Hence in case of perfect correlation, $(r = \pm 1)$, both the lines of regression coincide. Therefore, in general, we always have two lines of regression except in the particular case of perfect correlation when both the lines coincide and we get only one line.

11·2·1. Regression Coefficients. 'b', the slope of the line of regression of Y on X is also called the coefficient of regression of Y on X. It represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable X. More precisely, we write

$$b_{YX} = \text{Regression coefficient of } Y \text{ on } X = \frac{\mu_{11}}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X} \qquad \ldots (11\cdot11)$$

Similarly, the coefficient of regression of X on Y indicates the change in the value of variable X corresponding to a unit change in the value of variable Y and is given by :

$$b_{XY} = \text{Regression coefficient of } X \text{ on } Y = \frac{\mu_{11}}{\sigma_Y^2} = r\frac{\sigma_X}{\sigma_Y} \qquad \ldots (11\cdot11a)$$

11·2·2. Properties of Regression Coefficients.

(*a*) *Correlation coefficient is the geometric mean between the regression coefficients.*

Proof. Multiplying (11·11) and (11·11*a*), we get

$$b_{XY} \times b_{YX} = r\frac{\sigma_X}{\sigma_Y} \times r\frac{\sigma_Y}{\sigma_X} = r^2 \quad \Rightarrow \quad r = \pm\sqrt{b_{XY} \times b_{YX}} \qquad \ldots (11\cdot12)$$

Remark. We have

$$r = \frac{\mu_{11}}{\sigma_X \cdot \sigma_Y}, b_{YX} = \frac{\mu_{11}}{\sigma_X^2} \quad \text{and} \quad b_{XY} = \frac{\mu_{11}}{\sigma_Y^2}$$

It may be noted that the sign of correlation coefficient is the same as that of regression coefficients, since the sign of each depends upon the co-variance term μ_{11}. Thus, if the regression coefficients are positive, 'r' is positive and if the regression coefficients are negative 'r' is negative.

Hence, the sign to be taken before the square root in (11·12) is that of the regression coefficients.

(b) If one of the regression coefficients is greater than unity, the other must be less than unity.

Proof. Let one of the regression coefficients (say) b_{YX} be greater than unity, then we have to show that $b_{XY} < 1$.

Now $b_{YX} > 1 \Rightarrow \frac{1}{b_{YX}} < 1 \quad \ldots (*)$ Also $r^2 \le 1 \Rightarrow b_{YX} \cdot b_{XY} \le 1$

Hence $b_{XY} \le \frac{1}{b_{YX}} < 1$ [From (*)]

(c) The modulus value of the arithmetic mean of the regression coefficients is not less than the modulus value of the correlation coefficient r.

Proof. We have to prove that $\left|\frac{1}{2}(b_{YX} + b_{XY})\right| > |r|$.

$$\Rightarrow \left|\frac{1}{2}\left(r\frac{\sigma_Y}{\sigma_X} + r\frac{\sigma_X}{\sigma_Y}\right)\right| \ge |r| \Rightarrow \frac{\sigma_Y}{\sigma_X} + \frac{\sigma_X}{\sigma_Y} \ge 2 \qquad (\because |r| > 0)$$

$$\Rightarrow \sigma_Y^2 + \sigma_X^2 - 2\sigma_X\sigma_Y > 0 \Rightarrow (\sigma_Y - \sigma_X)^2 > 0,$$

which is always true, since the square of a real quantity is ≥ 0.

(d) Regression coefficients are independent of the change of origin but not of scale.

Proof. Let $U = \frac{X-a}{h}, V = \frac{Y-b}{k} \Rightarrow X = a + hU, Y = b + kV,$

where $a, b, h\ (>0)$ and $k\ (>0)$ are constants.

Then $\text{Cov}(X, Y) = hk\,\text{Cov}(U, V), \ \sigma_X^2 = h^2\sigma_U^2$ and $\sigma_Y^2 = k^2\sigma_V^2$

$$\therefore \quad b_{YX} = \frac{\text{Cov}(X, Y)}{\sigma_X^2} = \frac{hk\,\text{Cov}(U, V)}{h^2\sigma_U^2} = \frac{k}{h} \cdot \frac{\text{Cov}(U, V)}{\sigma_U^2} = \frac{k}{h} b_{VU}$$

Similarly, we can prove that $b_{XY} = (h/k)\, b_{UV}$.

11·2·3. Angle Between Two Lines of Regression. Equations of the lines of regression of Y on X, and X on Y are :

$$Y - \bar{y} = r \cdot \frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \quad \text{and} \quad X - \bar{x} = r \cdot \frac{\sigma_X}{\sigma_Y}(Y - \bar{y}) \Rightarrow Y - \bar{y} = \frac{\sigma_Y}{r\sigma_X}(X - \bar{x})$$

Slopes of these lines are $r \cdot \frac{\sigma_Y}{\sigma_X}$ and $\frac{\sigma_Y}{r\sigma_X}$ respectively. If θ is the acute angle between the two lines of regression, then

$$\tan\theta = \left| \frac{r\frac{\sigma_Y}{\sigma_X} - \frac{\sigma_Y}{r\sigma_X}}{1 + r\frac{\sigma_Y}{\sigma_X}\cdot\frac{\sigma_Y}{r\sigma_X}} \right| = \left|\frac{r^2 - 1}{r}\right| \left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)$$

$$= \frac{1 - r^2}{|r|}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right) \qquad (\because r^2 \le 1)$$

$$\therefore \qquad \theta = \tan^{-1}\left\{\frac{1 - r^2}{|r|}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)\right\} \qquad \ldots (11\cdot13)$$

Case (I). $(r = 0)$. If $r = 0$, $\tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$. Thus if the two variables are uncorrelated, the lines of regression become perpendicular to each other.

Case (II). $(r = \pm 1)$. If $r = \pm 1$, $\tan\theta = 0 \Rightarrow \theta = 0$ or π. In this case the two lines of regression either coincide or they are parallel to each other. But since both lines of regression pass through the point $(\bar{x}, \bar{y})$, they cannot be parallel. Hence in the case of perfect correlation, positive or negative, the two lines of regression coincide.

Remarks **1.** *Whenever* two lines intersect, there are two angles between them, one acute angle and other obtuse angle. Further $\tan\theta > 0$ if $0 < \theta < \pi/2$, *i.e.*, θ is an acute angle and $\tan\theta < 0$ if $\pi/2 < \theta < \pi$, *i.e.*, θ is an obtuse angle and since $0 < r^2 < 1$, the acute angle (θ_1) and obtuse angle θ_2 between the two lines of regression are given by :

$$\theta_1 = \text{Acute angle} = \tan^{-1}\left\{\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\cdot\frac{1 - r^2}{|r|}\right\}, \quad \text{and} \quad \theta_2 = \pi - \theta_1$$

2. When $r = 0$, *i.e.*, variables X and Y are uncorrelated, then the lines of regressions of Y on X and X on Y are given respectively by :

[From (11·8*a*) and (11·9*a*)] $Y = \bar{Y}$ and $X = \bar{X}$

as shown in the adjoining diagram. Hence, in this case $(r = 0)$, the lines of regression are perpendicular to each other and are parallel to X-axis and Y-axis respectively.

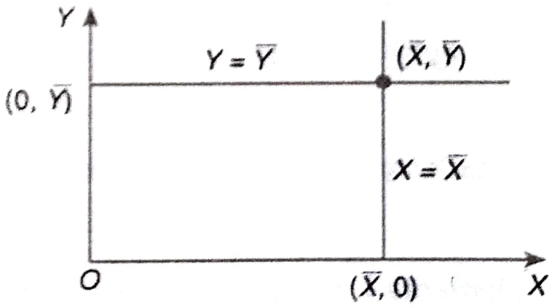


Fig. 11·2

3. The fact that if $r = 0$ (variables uncorrelated), the two lines of regression are perpendicular to each and if $r = \pm 1$, $\theta = 0$, *i.e.*, the two lines coincide, leads us to the conclusion that for higher degree of correlation between the variables, the angle between the lines is smaller, *i.e.*, the two lines of regression are nearer to each other. On the other hand, if the lines of regression make a larger angle, they indicate a poor degree of correlation between the variables and ultimately for $\theta = \pi/2$, $r = 0$, *i.e.*, the lines become perpendicular if no correlation exists between the variables. Thus by plotting the lines of regression on a graph paper, we can have an approximate idea about the degree of correlation between the two variables under study. Consider the following illustrations :

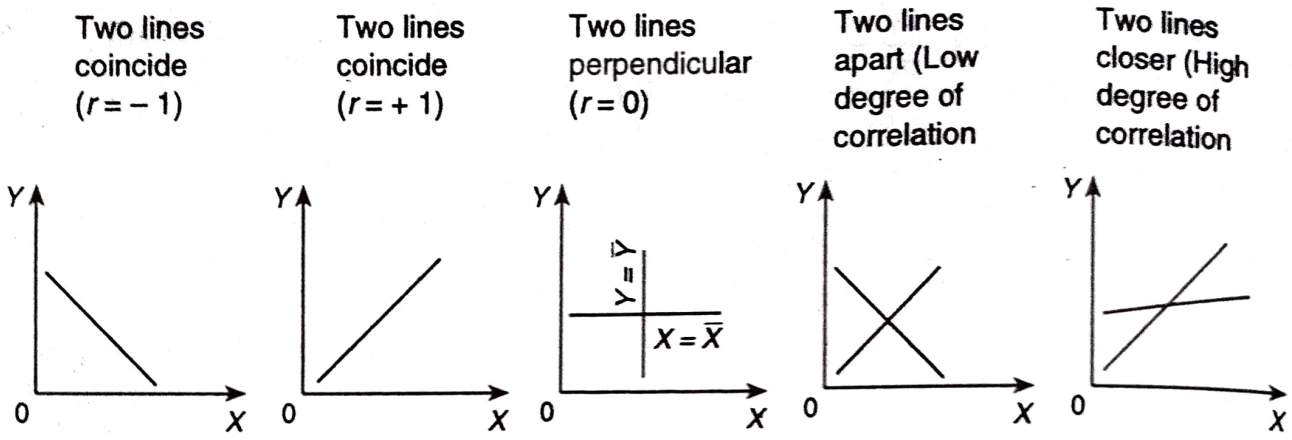


Fig. 11·3

11·2·4· Standard Error of Estimate or Residual Variance. The equation of the line of regression of Y on X is :

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \Rightarrow \frac{Y - \bar{Y}}{\sigma_Y} = r\frac{X - \bar{X}}{\sigma_X}$$

The residual variance S_Y^2 is the expected value of the squares of deviations of the observed values of Y from the expected values as given by the line of regression of Y on X. Thus

$$S_Y^2 = E[Y - \hat{Y}]^2 = E\left[Y - \{\bar{Y} + (r\sigma_Y (X - \bar{X})/\sigma_X\}\right]^2.$$

$$= \sigma_Y^2 E\left[\frac{Y - \bar{Y}}{\sigma_Y} - r\left(\frac{X - \bar{X}}{\sigma_X}\right)\right]^2 = \sigma_Y^2 E(Y^* - rX^*)^2,$$

where X^* and Y^* are standardised variates so that

$$E(X^{*2}) = 1 = E(Y^{*2}) \text{ and } E(X^* Y^*) = r.$$

$$\therefore \quad s_Y^2 = \sigma_Y^2 \left[E(Y^{*2}) + r^2 E(X^{*2}) - 2r E(X^* Y^*)\right] = \sigma_Y^2 (1 - r^2)$$

$$\Rightarrow \qquad s_Y = \sigma_Y (1 - r^2)^{1/2} \qquad \ldots (11\cdot13a)$$

Similarly, the standard error of estimate of X is given by :

$$s_X = \sigma_X (1 - r^2)^{1/2} \qquad \ldots (11\cdot13b)$$

Remarks **1.** Since s_X^2 or $s_Y^2 \geq 0$, it follows that $(1 - r^2) \geq 0 \Rightarrow |r| \leq 1$.

Hence $-1 \leq r(X, Y) \leq 1$.

2. If $r = \pm 1$, $s_X = s_Y = 0$ so that each deviation is zero, and the two lines of regression are coincident.

3. Since, as $r^2 \to 1$, s_X and $s_Y \to 0$, it follows that departure of the value r^2 from unity indicates the departure of the relationship between the variables X and Y from linearity.

4. From the definition of linear regression, the minima condition implies that s^2_Y or s^2_X is the minimum variance.

11·2·5. Correlation Coefficient Between Observed and Estimated Values. Here we will find the correlation between Y and $\hat{Y} = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$,

where $\hat{Y}$ is the estimated value of Y as given by the line of regression of Y on X.

We have : $$r(Y, \hat{Y}) = \frac{\text{Cov}(\hat{Y}, Y)}{\sigma_Y \sigma_{\hat{Y}}}$$

We have $E(\hat{Y}) = E\left[\bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})\right] = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}E(X - \bar{X}) = \bar{Y}$

$\therefore$ $\text{Var}(\hat{Y}) = E[\hat{Y} - E(\hat{Y})]^2 = E\left[r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})\right]^2 = r^2\sigma_Y^2$

$\Rightarrow$ $$\sigma_{\hat{Y}} = r\sigma_Y$$

Also $\text{Cov}(Y, \hat{Y}) = E[\{Y - E(Y)\}\{\hat{Y} - E(\hat{Y})\}]$

$$= E\left[b\{(X - E(X))\}\left\{r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})\right\}\right]$$

$$= b\, r\frac{\sigma_Y}{\sigma_X} E[\{X - E(X)\}^2] = \left(r\frac{\sigma_Y}{\sigma_X}\right)^2 \sigma_X^2 = r^2\sigma_Y^2 \left[\because b = b_{YX} = r\frac{\sigma_Y}{\sigma_X}\right]$$

$\therefore$ $$r(Y, \hat{Y}) = \frac{r^2\sigma_Y^2}{\sigma_Y r \sigma_Y} = r = r(X, Y)$$

Hence the correlation coefficient between observed and estimated value of Y is the same as the correlation coefficient between X and Y.

Example 11·1. *Obtain the equations of two lines of regression for the following data. Also obtain the estimate of X for $Y = 70$.*

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

Solution. Let $U = X - 68$ and $V = Y - 69$, then

$\bar{U} = 0, \bar{V} = 0, \sigma_U^2 = 4{\cdot}5, \sigma_V^2 = 5{\cdot}5$, $\text{Cov}(U, V) = 3$ and $r(U, V) = 0{\cdot}6$ (*c.f.* page 10·6)

Since correlation coefficient is independent of change of origin, we get

$$r = r(X, Y) = r(U, V) = 0{\cdot}6$$

We know that if $U = \frac{X - \alpha}{h}$, $V = \frac{Y - b}{k}$, then

$$\bar{X} = a + h\bar{U}, \quad \bar{Y} = b + k\bar{V}, \quad \sigma_X = h\sigma_U \quad \text{and} \quad \sigma_Y = k\sigma_Y$$

Here we are given : $h = k = 1, a = 68$ and $b = 69$.

Thus $\bar{X} = 68 + 0 = 68, \bar{Y} = 69 + 0 = 69$

$$\sigma_X = \sigma_U = \sqrt{4{\cdot}5} = 2{\cdot}12 \quad \text{and} \quad \sigma_Y = \sigma_V = \sqrt{5{\cdot}5} = 2{\cdot}35$$

Equation of line of regression of Y on X is : $Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$

i.e., $$Y = 69 + 0{\cdot}6 \times \frac{2{\cdot}35}{2{\cdot}12}(X - 68) \quad \Rightarrow \quad Y = 0{\cdot}665\,X + 23{\cdot}78$$

Equation of line of regression of X on Y is :

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

$$\Rightarrow \quad X = 68 + 0{\cdot}6 \times \frac{2{\cdot}12}{2{\cdot}35}(Y - 69) \quad \Rightarrow \quad X = 0{\cdot}54Y + 30{\cdot}74$$

To estimate X for given Y, we use the line of regression of X on Y. If $Y = 70$, estimated value of X is given by : $\hat{X} = 0{\cdot}54 \times 70 + 30{\cdot}74 = 68{\cdot}54$,

where $\hat{X}$ is estimate of X.

Example 11·2. *In a partially destroyed laboratory, record of an analysis of correlation data, the following results only are legible :*

Variance of $X = 9$. Regression equations : $8X - 10Y + 66 = 0$, $40X - 18Y = 214$.

What are : (i) the mean values X and Y, (ii) the correlation coefficient between X and Y, and (iii) the standard deviation of Y ?

Solution. (*i*) Since both lines of regression pass through the point $(\bar{X}, \bar{Y})$, we have : $8\bar{X} - 10\bar{Y} + 66 = 0$, and $40\bar{X} - 18\bar{Y} = 214$. Solving, we get $\bar{X} = 13$, $\bar{Y} = 17$.

(*ii*) Let $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$ be the lines of regression of Y on X and X on Y respectively. These equations can be put in the form :

$$Y = \frac{8}{10}X + \frac{66}{10} \quad \text{and} \quad X = \frac{18}{40} + \frac{214}{40}$$

$$\therefore \quad b_{YX} = \text{Regression coefficient of } Y \text{ on } X = \frac{8}{10} = \frac{4}{5}$$

$$\text{and} \quad b_{XY} = \text{Regression coefficient of } X \text{ on } Y = \frac{18}{40} = \frac{9}{20}$$

$$\text{Hence} \quad r^2 = b_{YX} \cdot b_{XY} = \frac{4}{5} \cdot \frac{9}{20} = \frac{9}{25} \qquad \therefore \; r = \pm\frac{3}{5} = \pm 0{\cdot}6$$

But since both the regression coefficients are positive, we take $r = +0{\cdot}6$.

(*iii*) We have $b_{YX} = r \cdot \dfrac{\sigma_Y}{\sigma_X} \Rightarrow \dfrac{4}{5} = \dfrac{3}{5} \times \dfrac{\sigma_Y}{3}$

Hence $\sigma_Y = 4$.

Remarks **1.** It can be verified that the values of $\bar{X} = 13$ and $\bar{Y} = 17$ as obtained in part (*i*) satisfy both the regression equations. In numerical problems of this type, this check should invariably be applied to ascertain the correctness of the answer.

2. If we had assumed that $8X - 10Y + 66 = 0$, is the equation of the line of regression of X on Y and $40X - 18Y = 214$ is the equation of line of regression of Y on X, then we get respectively :

$$8X = 10Y - 66 \quad \text{and} \quad 18Y = 40X - 214$$

$$\Rightarrow \quad X = \frac{10}{8}Y - \frac{66}{8} \quad \text{and} \quad Y = \frac{40}{18}X - \frac{214}{18}$$

$$\Rightarrow \quad b_{XY} = \frac{10}{8} \quad \text{and} \quad b_{YX} = \frac{40}{18}$$

$$\therefore \quad r^2 = b_{XY} \cdot b_{YX} = \frac{10}{8} \times \frac{40}{18} = 2{\cdot}78$$

But since r^2 always lies between 0 and 1, our supposition is wrong.

Example 11·3. *Find the most likely price in Mumbai corresponding to the price of Rs. 70 at Kolkata from the following :*

	Kolkata	*Mumbai*
Average price	65	67
Standard deviation	2·5	3·5

Correlation coefficient between the prices of commodities in the two cities is 0·8.

Solution. Let the prices (in Rupees) in Kalkata and Mumbai be denoted by X and Y respectively. Then we are given :

$\bar{X} = 65$, $\bar{Y} = 67$, $\sigma_X = 2{\cdot}5$, $\sigma_Y = 3{\cdot}5$ and $r = r(X, Y) = 0{\cdot}8$. We want Y for $X = 70$.

Line of regression of Y on X is :

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \quad \Rightarrow \quad Y = 67 + 0{\cdot}8 \times \frac{3{\cdot}5}{2{\cdot}5}(X - 65)$$

When $X = 70$, $\hat{Y} = 67 + 0{\cdot}8 \times \frac{3{\cdot}5}{2{\cdot}5}((70 - 65) = 72{\cdot}6$

Hence, the most likely price in Mumbai corresponding to the price of Rs. 70 at Kolkata is Rs. 72·60.

Example 11·4. *Can $Y = 5 + 2{\cdot}8X$ and $X = 3 - 0{\cdot}5Y$ be the estimated regression equations of Y on X and X on Y respectively ? Explain your answer with suitable theoretical arguments.*

Solution . Line of regression of Y on X is : $Y = 5 + 2{\cdot}8X \quad \Rightarrow \quad b_{YX} = 2{\cdot}8 \quad \ldots (*)$

Line of regression of X on Y is : $X = 3 - 0{\cdot}5Y \quad \Rightarrow \quad b_{XY} = -0{\cdot}5 \quad \ldots (**)$

This is not possible, since each of the regression coefficients b_{YX} and b_{XY} must have the same sign, which is same as that of Cov (X, Y). If Cov (X, Y) is positive, then both the regression coefficients are positive and if Cov (X, Y) is negative, then both the regression coefficients are negative. Hence (*) and (**) cannot be the estimated regression equations of Y on X and X on Y respectively.

Example 11·5. *By minimising $\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)^2$ for variations in α and p, show that there are two straight lines passing through the mean of the distribution for which the sum of squares of normal deviations has an extreme value. Prove also that their slopes are given by :*

$$\tan 2\alpha = \frac{2\mu_{11}}{\sigma_X^2 - \sigma_Y^2}.$$

Solution. We have to minimize : $S = \sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)^2 \quad \ldots (i)$

for variations in α and p.

Equating to zero, the partial derivatives of (*) *w.r. to* α and p, we have

$$\frac{\partial S}{\partial \alpha} = 0 = 2\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)(-x_i \sin\alpha + y_i \cos\alpha) \quad \ldots (ii)$$

$$\frac{\partial S}{\partial p} = 0 = -2\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p) \quad \ldots (iii)$$

Equation (*iii*) can be written as :

$$\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p) = 0 \quad \Rightarrow \quad \bar{x} \cos\alpha + \bar{y} \sin\alpha - p = 0 \qquad \ldots (iv)$$

From equation (*ii*), we get a quadratic equation which shows that there are two straight lines for extreme values of E.

From equation (*iv*), it becomes clear that both the straight lines pass through the point $(\bar{x}, \bar{y})$.

Again equation (*ii*) can be written as :

$$\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)(y_i \cos\alpha - x_i \sin\alpha) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} \left\{ f_i \left[\cos\alpha (x_i - \bar{x}) + \sin\alpha (y_i - \bar{y}) \right] (y_i \cos\alpha - x_i \sin\alpha) \right. = 0 \quad \text{[Using } (iv)\text{]}$$

$$\Rightarrow \quad \cos^2\alpha \sum_{i=1}^{n} f_i y_i (x_i - \bar{x}) - \sin\alpha \cos\alpha \sum_{i=1}^{n} f_i x_i (x_i - \bar{x})$$

$$+ \sin\alpha \cos\alpha \sum_{i=1}^{n} f_i y_i (y_i - \bar{y}) - \sin^2\alpha \sum_{i=1}^{n} f_i x_i (y_i - \bar{y}) = 0 \qquad \ldots (v)$$

We have $\mu_{11} = \frac{1}{N} \sum_i f_i (x_i - \bar{x})(y_i - \bar{y})$

$$= \frac{1}{N} \sum_i f_i x_i (y_i - \bar{y}) - \bar{x} \cdot \frac{1}{N} \sum_i f_i (y_i - \bar{y}) = \frac{1}{N} \sum f_i x_i (y_i - \bar{y})$$

Similarly, $\mu_{11} = \frac{1}{N} \sum_i f_i y_i (x_i - \bar{x})$

$$\sigma_X^2 = \frac{1}{N} \sum_i f_i x_i (x_i - \bar{x}) \quad \text{and} \quad \sigma_Y^2 = \frac{1}{N} \sum_i f_i y_i (y_i - \bar{y})$$

Substituting these values in (*v*), we get the required result.

11·3. CURVILINEAR REGRESSION

One of the criteria set forth previously in our regression studies has been that the variables X and Y are related linearly. In many cases, this assumption may not be valid. For example, in a scatter diagram showing the relationship between X, the number of overtime hours, versus Y, the resulting number of additional units produced for a manufacturing firm, it appears by inspection that a curvlinear regression may explain more of the variability of Y than does a linear line. For this reason, we shall fit the regression line using the following formula :

$$Y = a + b_1 X + b_2 X^2 \qquad \ldots (11\cdot14)$$

Using the principle of least squares, we have to determine the constants a, b_1 and b_2 so that : $E = \sum_{i=1}^{n} (y_i - a - b_1 x_i - b_2 x_i^2)^2$ is minimum.

Equating to zero the partial derivatives of E with respect to a, b and c separately, we get the normal equations for estimating a, B and C as

$$\left.\begin{aligned}\frac{\partial E}{\partial a} &= 0 = -2\sum_i (y_i - a - b_1 x_i - b_2 x_i^2)\\ \frac{\partial E}{\partial b_1} &= 0 = -2\sum_i x_i (y_i - a - b_1 x_i - b_2 x_i^2)\\ \frac{\partial E}{\partial b_2} &= 0 = -2\sum_i x_i^2 (y_i - a - b_1 x_i - b_2 x_i^2)\end{aligned}\right\} \quad \ldots (11\cdot15)$$

$$\Rightarrow \left\{\begin{aligned}\Sigma y_i &= na + b_1 \Sigma x_i + b_1 \Sigma x_i^2\\ \Sigma x_i y_i &= a\Sigma x_i + b_1 \Sigma x_i^2 + b_1 \Sigma x_i^3\\ \Sigma x_i^2 y_i &= a\Sigma x_i^2 + b_1 \Sigma x_i^3 + b_1 \Sigma x_i^4,\end{aligned}\right. \quad \ldots (11\cdot15a)$$

summation taken over i from 1 to n.

For given set of points (x_i, y_i) ; $i = 1, 2, \ldots, n$, equations (11·15a) can be solved for a, b_1 and b_2, and with these values of a, b_1 and b_2 (11·14) is the parabola of best fit.

***Remarks* 1 .** If $Y = a_0 + a_1 X + a_2 X^2 + \ldots + a_k X^k$ ... (11·16)

is the k^{th} degree polynomial of best fit to the set of points (x_i, y_i) ; $i = 1, 2, \ldots, n$ the constants $a_0, a_1, a_2, \ldots, a_k$ are to be obtained so that

$$E = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_k x_i^k)^2$$

is minimum. Thus, the normal equations for estimating $a_0, a_1, \ldots, a_k$ are obtained on equating to zero the partial derivatives of E *w.r.to.* $a_0, a_1, \ldots, a_k$ separately, *i.e.*,

$$\left.\begin{aligned}\frac{\partial E}{\partial a_0} &= 0 = -2\Sigma (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_k x_i^k)\\ \frac{\partial E}{\partial a_1} &= 0 = -2\Sigma x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_k x_i^k)\\ &\vdots \qquad\qquad \vdots\\ \frac{\partial E}{\partial a_k} &= 0 = -2\Sigma x_i^k (y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_k x_i^k)\end{aligned}\right\} \quad \ldots (11\cdot17)$$

$$\Rightarrow \left.\begin{aligned}\Sigma y_i &= na_0 + a_1 \Sigma x_i + a_2 \Sigma x_i^2 + \ldots + a_k \Sigma x_i^k\\ \Sigma x_i y_i &= a_0 \Sigma x_i + a_1 \Sigma x_i^2 + a_2 \Sigma x_i^3 + \ldots + a_k \Sigma x_i^{k+1}\\ &\vdots \qquad\qquad \vdots\\ \Sigma x_i^k y_i &= a_0 \Sigma x_i^k + a_1 \Sigma x_i^{k+1} + a_2 \Sigma x_i^{k+2} + \ldots + a_k \Sigma x_i^{2k}\end{aligned}\right\} \quad \ldots (11\cdot17a)$$

summation extended over i from 1 to n. These are $(k + 1)$ equations in $(k + 1)$ unknowns $a_0, a_1, a_2, \ldots, a_k$ and can be solved with the help of algebra.

2. Depending on the pattern suggested by the scatter diagram, there are many non-linear equations from which we can select a regression line. In such cases' the original data which is not in a linear form can be reduced to linear form by simple transformation of variables. This will be illustrated by considering the following curves.

(a)* Fitting of a Power Curve $Y = aX^b$ *to a set of n points. ... (11·18)

Taking logarithm of each side, we get

$$\log Y = \log a + b \log X \quad \Rightarrow \quad U = A + bV,$$

where $U = \log Y$, $A = \log a$ and $V = \log X$.

This is a linear equation in V and U.

Normal equations for estimating A and B are :

$$\Sigma U = nA + b\Sigma V \quad \text{and} \quad \Sigma UV = A\Sigma V + b\Sigma V^2 \qquad \ldots (11\cdot18a)$$

These equations can be solved for A and b and consequently, we get a = antilog (A)
With the values of a and b so obtained, (11·18) is the curve of best fit to the set of n points.

(b) Fitting of Exponential Curves. (i) $Y = ab^X$, (ii) $Y = ae^{bX}$ to a set of n points.

(i) $$Y = ab^X \qquad \ldots (11\cdot19)$$

Taking logarithm of each side, we get

$$\log Y = \log a + X \log b \quad \Rightarrow \quad U = A + BX,$$

where $U = \log Y$, $A = \log a$ and $B = \log b$.

This is linear equation in X and U.

The normal equations for estimating A and B are :

$$\Sigma U = nA + B\Sigma X \quad \text{and} \quad \Sigma XU = A\Sigma X + B\Sigma X^2 \qquad \ldots (11\cdot19a)$$

Solving these equations for A and B, we finally get

a = antilog (A) and B = antilog (B)

With these values of a and b, (11·19) is the curve of best fit to the given set of n points.

(ii) $$Y = ae^{bX} \qquad \ldots (11\cdot20)$$

$$\log Y = \log a + bX \log e = \log a + (b \log e)\, X \quad \Rightarrow \quad U = A + BX,$$

where $U = \log Y$, $A = \log a$ and $B = b \log e$.

This is linear equation in X and U, and the normal equations are :

$$\Sigma U = nA + B\Sigma X \quad \text{and} \quad \Sigma XU = A\Sigma X + B\Sigma X^2 \qquad \ldots (11\cdot20a)$$

From these we find A and B and consequently a = Antilog A and $b = \dfrac{B}{\log e}$.

Example 11·6. *For 10 randomly selected observations, the following data were recorded :*

Observation No.	:	1	2	3	4	5	6	7	8	9	10
Overtime hrs. (X)	:	1	1	2	2	3	3	4	5	6	7
Additional units (Y)	:	2	7	7	10	8	12	10	14	11	14

Determine the coefficients of regression and regression equation using the non-linear form : $Y = a + b_1 X + b_2 X^2$.

Solution.

S. No	X	Y	X^2	X^3	X^4	XY	X^2Y
1	1	2	1	1	1	2	2
2	1	7	1	1	1	7	7
3	2	7	4	8	16	14	28
4	2	10	4	8	16	20	40
5	3	8	9	27	81	24	72
6	3	12	9	27	81	36	108
7	4	10	16	64	256	40	160
8	5	14	25	125	625	70	350
9	6	11	36	216	1296	66	396
10	7	14	49	343	2401	98	686
Total	34	95	154	820	4774	377	1849

Using normal equations (11·15a), we get

$10a + 34\, b_1 + 154\, b_2 = 95$, $34a + 154\, b_1 + 820\, b_2 = 377$, and $154\, a + 820\, b_1 + 4774\, b_2 = 1849$.

The solutions to these three simultaneous equations are :

$$a = 1\cdot80, \quad b_1 = 3\cdot48 \quad \text{and} \quad b_2 = -0\cdot27$$

The regression equations, therefore, is :

$$Y = 1\cdot80 + 3\cdot48\, X - 0\cdot27X^2.$$

Example 11·7. *Fit an exponential curve of the form* $Y = ab^X$ *to the following data :*

X :	1	2	3	4	5	6	7	8
Y :	1·0	1·2	1·8	2·5	3·6	4·7	6·6	9·1

Solution.

S. No.	X	Y	$U = \log Y$	XU	X^2
1	1	1·0	0·0000	0·0000	1
2	2	1·2	0·0792	0·1584	4
3	3	1·8	0·2553	0·7659	9
4	4	2·5	0·3979	1·5916	16
5	5	3·6	0·5563	2·7815	25
6	6	4·7	0·6721	4·0326	36
7	7	6·6	0·8195	5·7365	49
8	8	9·1	0·9590	7·6720	64
Totals	36	30·5	3·7393	22·7385	204

(11·18*a*) gives the normal equation as :

$$3·7393 = 8A + 36B \quad \text{and} \quad 22·7385 = 36A + 204B$$

Solving, we get

$$B = 0·1408 \quad \text{and} \quad A = -0·1662 = \bar{1}·8338$$

$$\therefore \quad b = \text{Antilog } B = 1·383 \quad \text{and} \quad d = \text{Antilog } A = 0·6821$$

Hence the equation of the required curve is : $Y = 0·6821\ (1·38)^X$.

11·4. REGRESSION CURVES

In modern terminology, the conditional mean $E\ (Y \mid X = x)$ for a continuous distribution is called the regression function of Y on X and the graph of this function of x is known as the *regression curve of* Y *on* X or sometimes the regression curve for the mean of Y. Geometrically, the regression function represents the y co-ordinate of the centre of mass of the bivariate probability mass in the infinitesimal vertical strip bounded by x and $x + dx$.

Similarly, the regression function of X on Y is $E\ (X \mid Y = y)$ and the graph of this function of y is called the *regression curve (of the mean) of* X *on* Y.

In case a regression curve is a straight line, the corresponding regression is said to be *linear*. If one of the regressions is linear, it does not however follow that the other is also linear.

Theorem. *Let* (X, Y) *be a two-dimensional random variable with* $E(X) = \bar{X}$, $E(Y) = \bar{Y}$, $V(X) = \sigma_X^2$, $V(Y) = \sigma_Y^2$ *and let* $r = r\ (X,Y)$ *be the correlation coefficient between* X *and* Y. *If the regression of* Y *on* X *is linear, then*

$$E\ (Y \mid X) = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \qquad \ldots (11·21a)$$

Similarly, if the regression of X *on* Y *is linear, then*

$$E\ (X \mid Y) = \bar{X} + r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y}).$$