230

About the Author

Prof. T. Veerarajan, Professor of Mathematics is currently heading the Department of Science and Humanities, Sree Sowdambika College of Engineering, Aruppukottai, Tamil Nadu. A Gold Medalist from Madras University, he has had a brilliant academic career all through. He has more than 42 years of teaching experience at undergraduate and postgraduate levels in various established engineering colleges in Tamil Nadu including Anna University, Chennai.

Other books by Prof. T Veerarajan in ASCENT Series

Engineering Mathematics, 2e (For First Year)
[ISBN: 0-07-053483-7]
Engineering Mathematics, 2e (For Third Semester)
[ISBN: 0-07-049501-7]
Trigonometry, Algebra and Calculus
[ISBN: 0-07-053507-8]
Analytical Geometry, Real and Complex Analysis
[ISBN: 0-07-053489-6]
Numerical Methods: 2E
[ISBN: 0-07-060161-5]

PROBABILITY, STATISTICS AND RANDOM PROCESSES

Second Edition

T Veerarajan
Professor of Mathematics and Head
Science and Humanities Department
Sree Sowdambika College of Engineering
Aruppukottai, Tamil Nadu

Tata McGraw-Hill Publishing Company Limited
NEW DELHI

McGraw-Hill Offices
New Delhi New York St Louis San Francisco Auckland Bogotá Caracas
Kuala Lumpur Lisbon London Madrid Mexico City Milan Montreal
San Juan Santiago Singapore Sydney Tokyo Toronto

Further if the points in the scatter diagram appear to lie near a straight line, we assume that the R.V.'s have *linear correlation*. If they cluster round a well defined curve other than a straight line, the R.V.'s are assumed to be *non-linear*. In this section we will assume linear correlation between the concerned R.V.'s and discuss how to measure the degree of linear correlation.

CORRELATION COEFFICIENT

As the variance $E\{X - E(X)\}^2$ measures the variations of the R.V. X from its mean value $E(X)$, the quantity $E\{[X - E(X)]\,[Y - E(Y)]\}$ measures the simultaneous variation of two R.V.'s X and Y from their respective means and hence it is called *the covariance of* X, Y and denoted as Cov (X, Y).

Cov $(X, Y) = E\{[X - E(X)]\,[Y - E(Y)]\}$ is also called the *product moment* of X and Y and is also denoted as $p(X, Y)$.

Though $p(X, Y)$ is a useful measure of the degree of correlation between X and Y, it is to be expressed in mixed units of X and Y. To avoid this difficulty and to express the degree of correlation in absolute units, we divide $p(X, Y)$ by $\sigma_x \cdot \sigma_y$, so that $\dfrac{p(x, y)}{\sigma_x \sigma_y}$ is a mere number, free from the units of X and Y.

$\dfrac{p(x, y)}{\sigma_x \sigma_y}$ is a measure of intensity of linear relationship between X and Y and is called *Karl Pearson's Product Moment Correlation Coefficient* or simply *correlation coefficient* between X and Y. It is denoted by $r(X, Y)$ or r_{XY} or simply r.

Thus
$$r_{XY} = \frac{E\{[X - E(X)][Y - E(Y)]\}}{\sqrt{E\{X - E(X)\}^2 E\{Y - E(Y)\}^2}} \tag{1}$$

since σ_x, the standard deviation of X is the positive square root of the variance of X.

Now
$$\begin{aligned} &E[\{X - E(X)\}\,\{Y - E(Y)\}] \\ &= E[XY - E(Y) \cdot X - E(X) \cdot Y + E(X) \cdot E(Y)] \\ &= E(XY) - E(Y) \cdot E(X) - E(X) \cdot E(Y) + E(X) \cdot E(Y) \\ &\qquad [\Theta\ E(X) \text{ and } E(Y) \text{ are non-random constants}] \\ &= E\{XY\} - E(X) \cdot E(Y) \end{aligned} \tag{2}$$

Also we know that
$$E\{X - E(X)\}^2 = E(X^2) - \{E(X)\}^2 \tag{3}$$

and
$$E\{Y - E(Y)\}^2 = E\{Y^2\} - \{E(Y)\}^2 \tag{4}$$

Using (2), (3) and (4) in (1), we get

$$r_{XY} = \frac{E(XY) - E(X) \cdot E(Y)}{\sqrt{\left\{E\left(X^2\right) - E^2(X)\right\}\left\{E\left(Y^2\right) - E^2(Y)\right\}}} \tag{5}$$

where $E^2(X)$ means $\{E(X)\}^2$.

We will mainly deal with linear correlation of discrete R.V.'s X and Y. X will take the values $x_1, x_2, \cdots, x_n$ with frequency 1 each and Y will simultaneously take the values $y_1, y_2, \cdots, y_n$ with frequency 1 each. Hence $E(X) = \frac{1}{n}\sum x_i$; $E(X^2) = \frac{1}{n}\sum x_i^2$, $E(XY) = \frac{1}{n}\sum x_i y_i$ etc. Using these values in (5), the working formula for the compution of r_{XY} is got as

$$r_{XY} = \frac{\frac{1}{n}\Sigma x_i\, y_i - \frac{1}{n}\Sigma x_i \cdot \frac{1}{n}\Sigma y_i}{\sqrt{\left\{\frac{1}{n}\Sigma x_i^2 - \left(\frac{1}{n}\Sigma x_i\right)^2\right\}\left\{\frac{1}{n}\Sigma y_i^2 - \left(\frac{1}{n}\Sigma y_i\right)^2\right\}}} \tag{6}$$

or
$$r_{XY} = \frac{n\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{\left\{n\Sigma x^2 - (\Sigma x)^2\right\}\left\{n\Sigma y^2 - (\Sigma y)^2\right\}}} \tag{7}$$

Properties of Correlation Coefficient

1. $-1 \le r_{XY} \le 1$ or $|\text{Cov}(X, Y)| \le \sigma_X \cdot \sigma_Y$.

Let us consider

$$E[a\{X - E(X)\} + \{Y - E(Y)\}]^2 = a^2 \sigma_x^2 + 2a\, C_{XY} + \sigma_Y^2 \tag{1}$$

The R.H.S. expression is a quadratic expression in a, that is a real quantity. It is positive, as it is the expected value of a perfect square.

Hence, by the property of quadratic expressions, the discriminant of the R.H.S. ≤ 0

i.e., $$4\, C_{XY}^2 - 4\, \sigma_X^2 \sigma_Y^2 \le 0$$

i.e., $$C_{XY}^2 \le \sigma_X^2 \cdot \sigma_Y^2 \tag{2}$$

i.e., $$\frac{C_{XY}^2}{\sigma_X^2 \cdot \sigma_Y^2} \le 1$$

i.e., $$r_{XY}^2 \le 1$$

$\therefore$ $$|r_{XY}| \le 1 \text{ or } -1 \le r_{XY} \le 1$$

From step (2), it is clear that $|C_{XY}| \le \sigma_X \cdot \sigma_Y$

Note: When $0 < r_{XY} \le 1$, the correlation between X and Y is said to be *positive* or direct.

When $-1 \le r_{XY} \le 0$, the correlation is said to be *negative* or *inverse*.

When $-1 \le r_{XY} \le -0.5$ or $0.5 \le r_{XY} \le 1$, the correlation is assumed to be high, otherwise the correlation is assumed to be poor.

2. Correlation coefficient is independent of change of origin and scale.

i.e., If $U = \frac{X-a}{h}$ and $V = \frac{Y-b}{k}$, where $h, k > 0$, then $r_{XY} = r_{UV}$.

By the transformations, $X = a + hU$ and $Y = b + kV$

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

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

Then $C_{XY} = E[h\{U - E(U)\} \cdot k\{V - E(V)\} = hk\, C_{UV}$

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

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

$$\therefore \quad r_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\sigma_X^2 \cdot \sigma_Y^2}} = \frac{hk\, C_{UV}}{\sqrt{h^2 \cdot \sigma_U^2 \cdot k^2\ \sigma_V^2}}$$

$$= \frac{C_{UV}}{\sigma_U \cdot \sigma_V} = r_{UV}$$

Note *[If X and Y take considerably large values, computation of r_{XY} will become difficult. In such problems, we may introduce change of origin and scale and compute r using the above property.]*

3. Two independent R.V.'s X and Y are uncorrelated, but two uncorrelated R.V.'s need not be independent.
When X and Y are independent, $E(XY) = E(X) \cdot E(Y)$.
$\therefore C_{XY} = 0$ and hence $r_{XY} = 0$
viz., X and Y are uncorrelated.
The converse is not true, since $E(XY) = E(X) \cdot E(Y)$, when $r_{XY} = 0$.
This does not imply that X and Y are independent, as X and Y are independent only when $f(x, y) = f_X(x) \cdot f_Y(y)$.

Note *Note When $E(xy) = 0$, X and Y are said to be orthogonal R.V.'s.*

4. $$r_{XY} = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{(X-Y)}^2}{2\sigma_X\ \sigma_Y}$$

Let $Z = X - Y$. Then $E(Z) = E(X) - E(Y)$

$\therefore Z - E(Z) = [X - E(X)] - [Y - E(Y)]$

$$\therefore \quad \sigma_Z^2 = E[Z - E(Z)]^2 = E[\{X - E(X)\} - \{Y - E(Y)\}]^2$$

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

$$= \sigma_X^2 + \sigma_Y^2 - 2C_{XY}$$

$$\therefore \quad C_{XY} = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_Z^2}{2}$$

$$\therefore \quad r_{XY} = \frac{C_{XY}}{\sigma_X\ \sigma_Y} = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{(X-Y)}^2}{2\sigma_X\ \sigma_Y}$$

Similarly we can prove that

$$\sigma_{(X+Y)}^2 = \sigma^2{}_X + \sigma^2{}_Y + 2C_{XY}$$

and hence $r_{XY} = \dfrac{\sigma_{(X+Y)}^2 - \sigma_X^2 - \sigma_Y^2}{2\sigma_X\ \sigma_Y}$

Rank Correlation Coefficient

Sometimes the actual numerical values of X and Y may not be available, but the positions of the actual values arranged in order of merit (ranks) only may be available. The ranks of X and Y will in general, be different and hence may be considered as random variables. Let them be denoted by U and V. The correlation coefficient between U and V is called *the rank correlation coefficient* between (the ranks of) X, Y and denoted by ρ_{XY}.

Let us now derive a formula for ρ_{XY} or r_{UV}. Since U represents ranks of n values of X, U takes the values 1, 2, 3, $\cdots$, n.

Similarly V takes the same values 1, 2, 3, $\cdots$, n in a different order.

$$E(U) = E(V) = \frac{1}{n}(1 + 2 + \cdots + n) = \frac{n+1}{2}$$

$$E(U^2) = E(V^2) = \frac{1}{n}\ (1^2 + 2^2 + \cdots + n^2) = \frac{(n+1)\,(2n+1)}{6}$$

$$\therefore \quad \sigma_U^2 = \sigma_V{}^2 = E(U^2) - E^2(U)$$

$$= \frac{(n+1)\,(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$= \frac{(n+1)}{12}\ \{2(2n+1) - 3(n+1)\}$$

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

Let $D = U - V \quad \therefore E(D) = 0$

and $\sigma_D^2 = E(D^2)$

By property (4) given above,

$$\rho_{XY} = r_{UV} = \frac{\sigma_U^2 + \sigma_V^2 - \sigma_D^2}{2\sigma_U\,\sigma_V}, \text{ where } D = U - V$$

$$= \frac{\left(\frac{n^2-1}{6}\right) - \sigma_D^2}{2\left(\frac{n^2-1}{12}\right)}$$

$$= 1 - \frac{6}{n^2-1}\sigma_D^2 \text{ or } 1 - \frac{6E(D^2)}{n^2-1}$$

$$= 1 - \frac{6\Sigma d^2}{n(n^2-1)} \quad \left[\because E(D^2) = \frac{1}{n}\Sigma d^2\right]$$

[**Note:** The formula for the rank correlation coefficient is known as *spearman's formula*. The values of r_{XY} and ρ_{XY} (or r_{UV}) will be, in general, different.

Worked Example 4(B)

Example 1

Compute the coefficient of correlation between X and Y, using the following data:

X:	1	3	5	7	8	10
Y:	8	12	15	17	18	20

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	8	1	64	8
3	12	9	144	36
5	15	25	225	75
7	17	49	289	119
8	18	64	324	144
10	20	100	400	200
34	90	248	1446	582

Thus $n = 6$

$\Sigma x_i = 34,\ \Sigma y_i = 90$

$\Sigma x_i^2 = 248,\ \Sigma y_i^2 = 1446$

$\Sigma x_i y_i = 582$

$$r_{XY} = \frac{n\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{\left\{n\Sigma x^2 - (\Sigma x)^2\right\}\left\{n\Sigma y^2 - (\Sigma y)^2\right\}}}$$

$$= \frac{6 \times 582 - 34 \times 90}{\sqrt{\left\{6 \times 248 - (34)^2\right\}\left\{6 \times 1446 - (90)^2\right\}}}$$

$$= \frac{432}{\sqrt{332 \times 576}} = 0.9879$$

Example 2

Compute the coefficients of correlation between X and Y using the following data:

X:	65	67	66	71	67	70	68	69
Y:	67	68	68	70	64	67	72	70

We effect change of origin in respect of both X and Y. The new origins are chosen at or near the average of extreme values. Thus we take $\frac{65+71}{2} = 68$ as the new origin for X and $\frac{64+72}{2} = 68$ as the new origin for Y. viz., we put $u_i = (x_i - 68)$ and $v_i = y_i - 68$ and find r_{UV}.

$X = x_i$	$Y = y_i$	$u_i = x_i - 68$	$v_i = y_i - 68$	u_i^2	v_i^2	$u_i v_i$
65	67	−3	−1	9	1	3
67	68	−1	0	1	0	0
66	68	−2	0	4	0	0
71	70	3	2	9	4	6
67	64	−1	−4	1	16	4
70	67	2	−1	4	1	−2
68	72	0	4	0	16	0
69	70	1	2	1	1	2
	Total	−1	2	29	39	13

$$r_{XY} = r_{UV} = \frac{n\Sigma uv - \Sigma u \cdot \Sigma v}{\sqrt{\left\{n\Sigma u^2 - (\Sigma u)^2\right\}\left\{n\Sigma v^2 - (\Sigma v)^2\right\}}}$$

$$= \frac{8 \times 13 - (-1) \times 2}{\sqrt{(8 \times 29 - 1)(8 \times 39 - 4)}} = \frac{106}{\sqrt{231 \times 308}} = 0.3974$$

Example 3

Find the coefficient of correlation between X and Y using the following data:

X:	5	10	15	20	25
Y:	16	19	23	26	30

As the values of X are in arithmetic progression, we make the change of origin and scale, by choosing the middle most value 15 as the new origin and the common difference 5 as the new scale.

i.e., we put $U = \dfrac{X-15}{5}$

As the values of Y are not in A.P., we are content with effecting a change of origin only i.e., we put $V = Y - \left(\dfrac{30+16}{2}\right) = Y - 23$.

x	y	$u = \frac{x-15}{5}$	$v = y-23$	u^2	v^2	uv
5	16	−2	−7	4	49	14
10	19	−1	−4	1	16	4
15	23	0	0	0	0	0
20	26	1	3	1	9	3
25	30	2	7	4	49	14
	Total	0	−1	10	123	35

$$r_{XY} = r_{UV} = \frac{n\Sigma uv - \Sigma u \cdot \Sigma v}{\sqrt{\left\{n\Sigma u^2 - (\Sigma u)^2\right\}\left\{n\Sigma v^2 - (\Sigma v)^2\right\}}}$$

$$= \frac{5 \times 35 - 0 \times (-1)}{\sqrt{(5 \times 10 - 0)(5 \times 125 - 1)}}$$

$$= \frac{175}{\sqrt{50 \times 624}} = 0.9907$$

Example 4

The following table gives the bivariate frequency distribution of marks in an intelligence test obtained by 100 students according to their age:

Age (x) in yrs / *Marks (y)*	*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

Calculate the coefficient of correlation between age and intelligence.

Since the frequencies of various values of x and y are not equal to 1 each the formula for the computation of r_{XY} is taken with a slight modification as given below:

$$r_{XY} = r_{UV} = \frac{N\Sigma f_{XY}\, uv - \Sigma f_X u \cdot \Sigma f_Y v}{\sqrt{\left\{N\Sigma f_X u^2 - (\Sigma f_X u)^2\right\}\left\{N\Sigma f_Y v^2 - (\Sigma f_Y v)^2\right\}}} \quad (1)$$

where $u = x - 20$, $v = \dfrac{y-35}{10}$, f_X represents frequencies of X-distribution, f_Y represents frequencies of Y-distribution and f_{XY} are the cell frequencies.

Mid y/mid x	*18*	*19*	*20*	*21*	f_Y	v	$f_Y v$	$f_Y v^2$	$\Sigma f_{XY}\, uv$
15	4	2	2	–	8	−2	−16	32	20
25	5	4	6	4	19	−1	−19	19	10
35	6	8	10	11	35	0	0	0	0
45	4	4	6	8	22	1	22	22	−4
55	–	2	4	4	10	2	20	40	4
65	–	2	3	1	6	3	18	54	−3
f_X	19	22	31	28	100	Total	5	167	27
u	−2	−1	0	1	Total				
$f_X u$	−38	−22	0	28	−32				
$f_X u^2$	76	22	0	28	126				
$\Sigma f_{XY}\, uv$	18	−0	0	15	27				

Note *Σf_{XY} uv for the first row of the table is computed as follows.*

$$S f_{XY}\, uv = f_{11}\, u_1\, v_1 + f_{12}\, u_2\, v_1 + f_{13}\, u_3\, v_1 + f_{14}\, u_4\, v_1$$
$$= 4(-2)(-2) + 2(-1)(-2) + 2(0)(-2) + 0\,(1)\,(-2)$$
$$= 20$$

Similarly other $\Sigma f_{XY} uv$ values are computed. Value of $(\Sigma\Sigma f_{XY} uv)$ obtained as the total of the entries of the last column and as that of the last row must tally.

Using the relevant values obtained in the table in (1), we have

$$r_{XY} = \frac{100 \times 27 - (-32) \times 5}{\sqrt{\left\{100 \times 126 - (-32)^2\right\}\left\{100 \times 167 - 5^2\right\}}}$$

$$= \frac{2860}{\sqrt{13624 \times 16675}} = 0.1897$$

Example 5

Calculate the correlation coefficient for the following ages of husbands (X) and wives (Y), using only standard deviations of X and Y:

X:	23	27	28	28	29	30	31	33	35	36
Y:	18	20	22	27	21	29	27	29	28	29

x	y	$u = x - 30$	$v = y - 24$	u^2	v^2	$d = x - y$	d^2
23	18	– 7	– 6	49	36	5	25
27	20	– 3	– 4	9	16	7	49
28	22	– 2	– 2	4	4	6	36
28	27	– 2	3	4	9	1	1
29	21	– 1	– 3	1	9	8	64
30	29	0	5	0	25	1	1
31	27	1	3	1	9	4	16
33	29	3	5	9	25	4	16
35	28	5	4	25	16	7	49
36	29	6	5	36	25	7	49
	Total	0	10	138	174	50	306

$$\sigma_X^2 = \frac{1}{n}\Sigma u^2 - \left(\frac{1}{n}\Sigma u\right)^2 = \frac{1}{10} \times 138 = 13.8$$

$$\sigma_Y^2 = \frac{1}{n}\Sigma v^2 - \left(\frac{1}{n}\Sigma v\right)^2 = \frac{1}{10} \times 174 - \left(\frac{10}{10}\right)^2 = 16.4$$

$$\sigma_{(X-Y)}^2 = \sigma_D^2 = \frac{1}{n}\Sigma d^2 - \left(\frac{1}{n}\Sigma d\right)^2 = \frac{1}{10} \times 306 - \left(\frac{50}{10}\right)^2 = 5.6$$

$$r_{XY} = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_D^2}{2\sigma_X\,\sigma_Y} = \frac{13.8 + 16.4 - 5.6}{2 \times \sqrt{13.8} \times \sqrt{16.4}}$$

$$= \frac{24.6}{30.0879} = 0.8176$$

Example 6

If the independent random variables X and Y have the variances 36 and 16 respectively, find the correlation coefficient between $(X + Y)$ and $(X - Y)$.

Let $U = X + Y$ and $V = X - Y$

$$E(U) = E(X) + E(Y);\ E(V) = E(X) - E(Y)$$

$$E(UV) = E(X^2 - Y^2) = E(X^2) - E(Y^2)$$

$$E(U^2) = E\{(X + Y)^2\} = E(X^2) + E(Y^2) + 2E(XY)$$

$$E(V^2) = E(X^2) + E(Y^2) - 2E(XY)$$

$$C_{UV} = E(UV) - E(U) \cdot E(V)$$

$$= E(X^2) - E(Y^2) - \{E^2(X) - E^2(Y)\}$$

$$= [E(X^2) - E^2(X)] - [E(Y^2) - E^2(Y)]$$

$$= \sigma_X^2 - \sigma_Y^2 = 36 - 16 = 20$$

$$\sigma_U^2 = E(U^2) - E^2(U)$$

$$= \{E(X^2) + E(Y^2) + 2E(XY)\} - \{E^2(X) + E^2(Y) + 2E(X) \cdot E(Y)\}$$

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

$$= 36 + 16 + 2 \times 0$$

[Θ X and Y are independent and hence uncorrelated]

$$= 52$$

Similarly, $\sigma_V^2 = 52$

Now $$r_{UV} = \frac{C_{UV}}{\sigma_U \cdot \sigma_V} = \frac{20}{52} = \frac{5}{13}$$

Example 7

If X, Y and Z are uncorrelated R.V.'s with zero means and standard deviations 5, 12 and 9 respectively and if $U = X + Y$ and $V = Y + Z$, find the correlation coefficient between U and V.

$$E(X) = E(Y) = E(Z) = 0$$

$$\text{Var}(X) = E(X^2) - E^2(X) = 25 \quad \therefore E(X^2) = 25$$

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

X and Y are uncorrelated

$\therefore$ $r_{XY} = 0$. i.e., $E(XY) - E(X) \cdot E(Y) = 0$

$\therefore$ $E(XY) = 0$. Similarly $E(YZ) = 0$; $E(ZX) = 0$

Now $E(U) = E(X + Y) = 0$ and $E(V) = 0$

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

$$= 25 + 144 + 2 \times 0 = 169$$

$$E(V^2) = E(Y^2 + Z^2 + 2YZ)$$

$$= 144 + 81 + 2 \times 0 = 225$$

$\therefore$ $\sigma_U^2 = E(U^2) - E^2(U) = 169$

and $\sigma_V^2 = E(V^2) - E^2(V) = 225$

$$E(UV) = E\{(X + Y)(Y + Z)\}$$

$$= E(XY) + E(XZ) + E(YZ) + E(Y^2)$$

$$= 0 + 0 + 0 + 144 = 144$$

$$r_{UV} = \frac{E(UV) - E(U)\cdot E(V)}{\sigma_U \cdot \sigma_V} = \frac{144}{13 \times 15} = \frac{48}{65}$$

Example 8

If X and Y are two R.V.'s with variances σ_X^2 and σ_Y^2 respectively, find the value of k, if $U = X + kY$ and $V = X + \frac{\sigma_X}{\sigma_Y} \cdot Y$ are uncorrelated.

U and V are uncorrelated.

$\therefore$ $\quad \text{Cov}(U, V) = 0$

i.e., $\quad E(UV) - E(U) \cdot E(V) = 0$

i.e., $\quad E\left\{(X + kY)\left(X + \frac{\sigma_X}{\sigma_Y} \cdot Y\right)\right\} - E(X + kY) \cdot E\left(X + \frac{\sigma_X}{\sigma_Y} Y\right) = 0$

i.e. $\quad E\left\{X^2 + k\frac{\sigma_X}{\sigma_Y} \cdot Y^2 + \left(k + \frac{\sigma_X}{\sigma_Y}\right) XY\right\}$

$$- \left[\{E(X) + kE(Y)\}\left\{E(X) + \frac{\sigma_X}{\sigma_Y} E(Y)\right\}\right] = 0$$

i.e., $\quad E(X^2) + k\frac{\sigma_X}{\sigma_Y} E(Y^2) + \left(k + \frac{\sigma_X}{\sigma_Y}\right) E(XY)$

$$- \left[E^2(X) + k\frac{\sigma_X}{\sigma_Y} E^2(Y) + \left(k + \frac{\sigma_X}{\sigma_Y}\right) E(X) \cdot E(Y)\right] = 0$$

i.e., $\quad \{E(X^2) - E^2(X)\} + \frac{k\sigma_X}{\sigma_Y} \{E(Y^2) - E^2(Y)\}$

$$+ \left(k + \frac{\sigma_X}{\sigma_Y}\right) \{E(XY) - E(X) \cdot E(Y)\} = 0$$

i.e., $\quad \sigma_X^2 + k\frac{\sigma_X}{\sigma_Y} \cdot \sigma_Y^2 + \left(k + \frac{\sigma_X}{\sigma_Y}\right) \text{Cov}(X, Y) = 0$

Dividing throughout by σ_X,

$$(\sigma_X + k\sigma_Y) + (\sigma_X + k\sigma_Y) \cdot \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

i.e., $\quad (\sigma_X + k\sigma_Y)(1 + r_{XY}) = 0$

Assuming that $r_{XY} \neq -1$, we get

$$\sigma_X + k\sigma_Y = 0$$

$\therefore$ $\quad k = -\frac{\sigma_X}{\sigma_Y}$

Example 9

If (X, Y) is a two-dimensional RV uniformly distributed over the triangular region R bounded by $y = 0$, $x = 3$ and $y = 4/3\ x$. Find $f_X(x), f_Y(y)$, $E(X)$, Var(X), $E(Y)$, Var(Y) and ρ_{XY}.

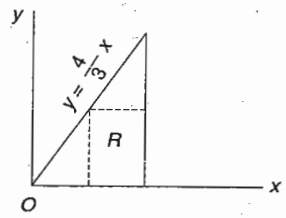

Fig. 4.3

Since (X, Y) is uniformly distributed, $f(x, y)$ = a constant = k

Now $\quad \int_0^4 \int_{3y/4}^3 f(x, y)\, dx\, dy = 1$

i.e., $\quad \int_0^4 \int_{3y/4}^3 k\, dx\, dy = 1$

i.e., $\quad k \int_0^4 \left(3 - \frac{3y}{4}\right) dy = 1$

i.e., $\quad 6k = 1$

$\therefore$ $\quad k = \frac{1}{6}$

$$f_Y(y) = \int_{3y/4}^{3} \frac{1}{6} dx = \frac{1}{8}(4 - y),\ 0 < y < 4$$

$$f_X(x) = \int_{3}^{4x/3} \frac{1}{6} dy = \frac{2}{9} x,\ 0 < x < 3$$

$$E(X) = \int x f_X(x)\, dx = \int_0^3 \frac{2}{9} x^2\, dx = 2$$

$$E(Y) = \int y f_Y(y)\, dy = \int_0^4 \frac{y}{8} \times (4 - y)\, dy = \frac{4}{3}$$

$$E(X^2) = \int_0^3 \frac{2}{9} \times x^3\, dx = \frac{9}{2}$$

$$E(Y^2) = \int_0^4 \frac{y^2}{8} \times (4 - y)\, dy = \frac{8}{3}$$

$$\text{Var}(X) = E(X^2) - \{E(X)\}^2 = \frac{9}{2} - 4 = \frac{1}{2}$$

$$\text{Var}(Y) = E(Y^2) - \{E(Y)\}^2 = \frac{8}{3} - \frac{16}{9} = \frac{8}{9}$$

$$E(XY) = \int_0^4 \int_{3y/4}^3 \frac{1}{6} xy\, dx\, dy$$

$$= \frac{3}{64} \int_0^4 (16 - y^2) y\, dy = 3$$

$$\rho_{XY} = \frac{E(XY) - E(X) \times E(Y)}{\sigma_x \sigma_y} = \frac{3 - 2 \times \frac{4}{3}}{\frac{1}{\sqrt{2}} \times 2 \frac{\sqrt{2}}{3}} = \frac{1}{2}$$

Example 10

Find the correlation co-efficient between X and Y, which are jointly normally distributed with

$$f(x, y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1 - r^2}} \exp\left\{-\frac{1}{2(1 - r^2)}\left(\frac{x^2}{\sigma_x^2} - \frac{2rxy}{\sigma_x \sigma_y} + \frac{y^2}{\sigma_y^2}\right)\right\}$$

$$\frac{x^2}{\sigma_x^2} - \frac{2rxy}{\sigma_x \sigma_y} + \frac{y^2}{\sigma_y^2} = \left(\frac{x}{\sigma_x} - \frac{ry}{\sigma_y}\right)^2 + (1 - r^2) \frac{y^2}{\sigma_y^2}$$

$$= \frac{1}{\sigma_x^2} \left(x - \frac{ry\sigma x}{\sigma_y}\right)^2 + (1 - r^2) \frac{y^2}{\sigma_y^2}$$

$$E(XY) = \frac{1}{2\pi\sigma_x \sigma_y \sqrt{1 - r^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp$$

$$\left\{-\frac{1}{2(1 - r^2)}\left[\frac{1}{\sigma_x^2}\left(x - \frac{ry\sigma_x}{\sigma_y}\right)^2 + \frac{(1 - r)^2 y^2}{\sigma_y^2}\right]\right\} xy\, dx\, dy$$

$$= \frac{1}{\sigma_y \sqrt{2\pi}} \int_{-\infty}^{\infty} y \exp\left(\frac{-y^2}{2\sigma_y^2}\right) \int_{-\infty}^{\infty} \frac{x}{\left(\sigma_x \sqrt{1 - r^2}\right)\sqrt{2\pi}}$$

$$\exp\left\{\frac{-\left(x - \frac{ry\sigma_x}{\sigma_y}\right)^2}{2(1 - r^2)\sigma_x^2}\right\} dx\, dy \qquad (1)$$

The inner integral is the mean of the normal distribution with mean $\frac{ry\sigma_x}{\sigma_y}$ and variance $(1 - r^2)\, \sigma_X^2$.

$\therefore$ the inner integral $= \frac{ry\sigma_x}{\sigma_y}$

Using this value in (1),

$$E(XY) = \left(\frac{r\sigma_x}{\sigma_y}\right) \times \frac{1}{\sigma_y \sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 \exp\left(-\frac{y^2}{2\sigma_y^2}\right) dy$$

$$= \frac{r\sigma_x}{\sigma_y} E(Y^2) \text{ for } N(0, \sigma_Y)$$

$$= \frac{r\sigma_x}{\sigma_y} \times \sigma_Y^2$$

$$= r\sigma_X \sigma_Y$$

$\therefore$ $$\rho_{XY} = \frac{E(XY) - E(X)E(y)}{\sigma_x \sigma_y} = r$$

Example 11

Ten students got the following percentage of marks in Mathematics and Physical sciences:

Students:	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics:	78	36	98	25	75	82	90	62	65	39
Marks in Phy. Sciences:	84	51	91	60	68	62	86	58	63	47

Calculate the rank correlation coefficient.

Denoting the ranks in Mathematics and in Phy. Sciences by U and V, we have the following values of U and V:

U:	4	9	1	10	5	3	2	7	6	8	
V:	3	9	1	7	4	6	2	8	5	10	
D:	1	0	0	3	1	−3	0	−1	1	−2	
D^2:	1	0	0	9	1	9	0	1	1	4	$: \Sigma d^2 = 26$

$$\rho_{XY} = r_{UV} = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 26}{10 \times 99} = 0.8424$$

Example 12

Ten competitors in a beauty contest were ranked by three judges as follows:

	Competitors									
Judges	1	2	3	4	5	6	7	8	9	10
A:	6	5	3	10	2	4	9	7	8	1
B:	5	8	4	7	10	2	1	6	9	3
C:	4	9	8	1	2	3	10	5	7	6

Discuss which pair of judges have the nearest approach to common taste of beauty.

Rank by A (*U*)	*Rank by B* (*V*)	*Rank by C* (*W*)	$d_1 = U - V$	$d_2 = V - W$	$d_3 = U - W$	d_1^2	d_2^2	d_3^2
6	5	4	1	1	2	1	1	4
5	8	9	−3	−1	−4	9	1	16
3	4	8	−1	−4	−5	1	16	25
10	7	1	3	6	9	9	36	81
2	10	2	−8	8	0	64	64	0
4	2	3	2	−1	1	4	1	1
9	1	10	8	−9	−1	64	81	1
7	6	5	1	1	2	1	1	4
8	9	7	−1	2	1	1	4	1
1	3	6	−2	−3	−5	4	9	25
					Total:	157	214	158

$$r_{UV} = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)} = 1 - \frac{6 \times 157}{10 \times 99} = 0.0485$$

$$r_{VW} = 1 - \frac{6\Sigma d_2^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = -0.2970$$

$$r_{UW} = 1 - \frac{6\Sigma d_3^2}{n(n^2-1)} = 1 - \frac{6 \times 158}{10 \times 99} = 0.0424$$

Since r_{UV} is maximum, the judges *A* and *B* may be considered to have common taste of beauty to some extent compared to other pairs of judges.

Exercise 4(B)

Part A

(Short Answer Questions)

1. What is a scatter diagram? What is its role in correlation analysis?
2. What do you mean by correlation between two random variables?
3. What is linear correlation? How will you find that two R.V.'s are linearly correlated?
4. Define covariance of X, Y and coefficient of correlation between X and Y.
5. Why is r_{XY} preferred for measuring the degree of linear correlation to Cov (X, Y)?
6. State the properties of correlation coefficient.
7. State two different formulas used to compute r_{XY}.
8. Define rank correlation coefficient and write down the formula for computing it.
9. Prove that $-1 \le r_{XY} \le 1$.
10. Prove that $\sigma^2_{(X+Y)} - \sigma^2_{(X-Y)} = 4$ Cov (X, Y)
11. If C_{XY} is the covariance of X and Y, prove that $C_{XY} = E(XY) - E(X) \cdot E(Y)$.
12. If X and Y are indejpendent R.V.'s prove that $r_{XY} = 0$. Is the converse true?
13. If X and Y are uncorrelated, prove that $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
14. When are two R.V.s said to be orthogonal?

Part B

15. Ten students got the following marks in Mathematics and Basic Engineering:

Marks in Mathematics	78	36	98	25	75	82	90	62	65	39
Marks in Basic Engg.	84	51	91	60	68	62	86	58	53	47

Calculate the coefficient of correlation.

16. Calculate the correlation coefficient between X and Y from the following data:

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

17. Find the coefficient of correlation between X and Y using the following data:

X:	5.5	3.6	2.6	3.4	3.1	2.7	3.0	3.1	3.2	3.8
Y:	27	36	39	39	32	35	40	36	44	36

18. Compute the coefficient of correlation between X and Y from the following data:

X:	80	45	55	56	58	60	65	68	70	75	85
Y:	82	56	50	48	60	62	64	65	70	74	90

19. Find the coefficient of correlation between X and Y from the following data:

X:	10	14	18	22	26	30
Y:	18	12	24	6	30	36

20. Calculate the coefficient of correlation between X and Y, by finding variances only, from the following data:

X:	21	23	30	54	57	58	72	78	87	90
Y:	60	71	72	83	110	84	100	92	113	135

21. Calculate r_{XY} from the following data, where X represents production (in crore tons) and Y represents exports (in crore tons), using only the variances.

X:	55	56	58	59	60	60	62
Y:	35	38	38	39	44	43	44

22. The following table gives the frequency of scores obtained by 65 students in a general knowledge test according to age groups. Measure the degree of linear relationship between age and general knowledge:

Test scores	*Age in years*			
	19	*20*	*21*	*22*
225	4	4	2	1
275	3	5	4	2
325	2	6	8	5
375	1	4	6	8

23. Compute the value of r_{XY} between X, the ages of husbands and Y the ages of wives from the following data:

X / Y	*15–25*	*25–35*	*35–45*	*45–55*	*55–65*	*65–75*	*Total*
15–25	1	1	–	–	–	–	2
25–35	2	12	1	–	–	–	15
35–45	–	4	10	1	–	–	15
45–55	–	–	3	6	1	–	10
55–65	–	–	–	2	4	2	8
65–75	–	–	–	–	1	2	3
Total	3	17	14	9	6	4	53

24. Find the rank correlation coefficient between the ranks of the variable X and Y:

X:	10	15	12	17	13	16	24	14	22
Y:	30	42	45	46	33	34	40	35	39

25. The competitors in a musical contest were ranked by the three judges A, B, C in the following order:

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

Using rank correlation technique, find which pair of judges have more or less the same taste in music.

26. If X, Y, Z are uncorrelated R.V.'s having the same variance, find the correlation coefficient between $(X + Y)$ and $(Y + Z)$.
27. If X and Y are two uncorrelated R.V.'s with zero means, prove that $U = X \cos \alpha + Y \sin \alpha$ and $V = X \sin \alpha - Y \cos \alpha$ are also uncorrelated.
28. X and Y are independent R.V.'s with means 5 and 10 and variances 4 and 9 respectively. Obtain the correlation coefficient between U and V, where $U = 3X + 4Y$ and $V = 3X - Y$.
29. If X_1, X_2, X_3 are three uncorrelated R.V.'s having variances v_1, v_2, v_3 respectively, obtain the coefficient of correlation between $(X_1 + X_2)$ and $(X_2 + X_3)$.
30. Show that (i) $E\{aX + bY\} = aE(X) + bE(Y)$ and (ii) $\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2ab\,C(X, Y)$, where $C(X, Y)$ is the covariance of (X, Y).
31. It two R.V.'s are uncorrelated, prove that the variance of their sum is equal to the sum of their variances.
32. If the joint density fucntion of (X, Y) is given by $f(x, y) = 2 - x - y$, $0 \le x, y \le 1$, find $E(X)$, $E(Y)$, var (X), var (Y) and r_{XY}.
33. If the two dimensional R.V. (X, Y) is uniformly distributed in $0 \le x < y \le 1$, find $E(X)$, $E(Y)$, var (X), var (Y) and r_{XY}.
34. If the two dimensional R.V. (X, Y) is uniformly distributed over R, where R is defined by $\{(x, y)/x^2 + y^2 \le 1, y \ge 0\}$, find r_{XY}.
35. If the joint pdf of (X, Y) is given by $f(x, y) = x + y$, $0 \le x, y \le 1$, find r_{XY}.
36. Let X be a R.V. with mean value = 3 and variance = 2. Find the second moment of X about the origin. Another R.V. Y is defined by $Y = -6X + 22$. Find the mean value of Y and the correlation of X and Y.

REGRESSION

When the random variables X and Y are linearly correlated, the points plotted on the scatter diagram, corresponding to n pairs of observed values of X and Y, will have a tendency to cluster round a straight line. This straight is called *the regression line*. The regression line can be taken as the best fitting straight line for the observed pairs of values of X and Y in the least square sense, with which the students are familiar.

When two R.V.'s X and Y are linearly correlated, we may not know which variable takes independent values. If we treat X as the independent variable and hence assume that the values of Y depend on those of X, the regression line is called the *regression line of Y on X*. If we assume that the values of X depend on those of the independent variable Y, *the regression line of X on Y* is obtained. Thus in situations where the distinction cannot be made between the R.V.'s X and Y as to which is the independent variable and which is the dependent variable, there will be two regression lines. However, when the value of $Y(X)$ is to be predicted corresponding to a specified value of $X(Y)$, we should make use of the regression line of $Y(X)$ on $X(Y)$.

Equation of the Regression Line of Y on X

The regression line of Y on X is the best-fitting straight line for the observed pairs of values (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n), based on the assumption that x is the independent variable and y is the dependent variable. Hence, let the equation of the regression line of Y on X be assumed as $y = ax + b$. (1)

By the principle of least squares, the normal equations which give the values of a and b.

are $$\Sigma y_i = a \Sigma x_i + nb \quad (2)$$

and $$\Sigma x_i y_i = a \Sigma x_i^2 + b \Sigma x_i \quad (3)$$

Dividing equation (2) by n, we get

$$\bar{y} = a\bar{x} + b \quad (4)$$

where $\bar{x} = E(X)$ and $\bar{y} = E(Y)$. (1)–(4) gives the required equation as

$$y - \bar{y} = a(x - \bar{x}) \quad (5)$$

Eliminating b between equations (2) and (3)

we get $$a = \frac{n\Sigma x_i y_i - \Sigma x_i \cdot \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$= \frac{\frac{1}{n}\Sigma x_i y_i - \left(\frac{1}{n}\Sigma x_i\right)\cdot\left(\frac{1}{n}\Sigma y_i\right)}{\frac{1}{n}\Sigma x_i^2 - \left(\frac{1}{n}\Sigma x_i\right)^2}$$

or $$a = \frac{E(XY) - E(X)\cdot E(Y)}{E(X^2) - E^2(X)} = \frac{p_{XY}}{\sigma_X^2} \quad (6)$$

Using (6) in (5), we get the equation of the regression line of Y on X as

$$y - \bar{y} = \frac{p_{XY}}{\sigma_X^2}(x - \bar{x}) \quad (7)$$

or $$y - \bar{y} = \frac{r_{XY}\,\sigma_Y}{\sigma_X}(x - \bar{x}) \quad (8)$$

$$\left[\because r_{XY} = \frac{p_{XY}}{\sigma_X\,\sigma_Y}\right]$$

In a similar manner, assuming the equation of the regression line of X and Y as $x = ay + b$ and using the equations

$$\Sigma x_i = a\Sigma y_i + nb \text{ and } \Sigma x_i y_i = a\Sigma y_i^2 + b\Sigma y_i,$$

we can get the equation of the regression line of X on Y as

$$x - \bar{x} = \frac{p_{XY}}{\sigma_Y^2}(y - \bar{y}) \quad (9)$$

or $$x - \bar{x} = \frac{r_{XY}\,\sigma_X}{\sigma_Y}(y - \bar{y}) \quad (10)$$

Note:

1. $\frac{p_{XY}}{\sigma_X^2}$ or $\frac{r_{XY}\,\sigma_Y}{\sigma_X}$ is called *the regression coefficient of Y on X* and denoted by b_1 or b_{YX}. $\frac{p_{XY}}{\sigma_Y^2}$ or $\frac{r_{XY}\,\sigma_X}{\sigma_Y}$ is called *the regression coefficient of X on Y* and denoted by b_2 or b_{XY}.
2. Clearly $b_1\,b_2 = r_{XY}^2$, i.e., r_{XY} is the geometric mean of b_1 and b_2.
   $\therefore \quad r_{XY} = \pm\sqrt{b_1\,b_2}$

   The sign of r_{XY} is the same as that of b_1 or b_2, as $b_1 = r_{xy}\frac{\sigma_Y}{\sigma_X}$ and $b_2 = r_{XY}\frac{\sigma_Y}{\sigma_X}$ have the same sign as r_{XY} (Θ σ_X and σ_Y are positive).

   Also $$\frac{b_1}{b_2} = \frac{\sigma_Y^2}{\sigma_X^2}$$
3. When there is perfect linear correlation between X and Y, viz., when $r_{XY} = \pm 1$, the two regression lines coincide.
4. The point of intersection of the two regression lines is clearly the point whose co-ordinates are $(\bar{x}, \bar{y})$.
5. When there is no linear correlation between X and Y, viz., when $r_{XY} = 0$, the equations of the regression lines become $y = \bar{y}$ and $x = \bar{x}$, which are at right angles.

Standard Error of Estimate of Y

Although we use the regression line of Y on X to predict the value of Y corresponding to a specified value of X we may also use it to estimate the value of Y corresponding to an observed value of $X = x_i$, say. The value of Y estimated in this manner need not, in general, be equal to the corresponding observed value of Y, namely, y_i. Hence the difference between Y and Y_E is called *the error of estimate of Y*. This error will vary from one observed value to the other and a random variable. The standard deviation of this R.V. $(Y - Y_E)$ is called *the standard error of estimate of Y* and denoted by S_Y.

Now $E\{Y - Y_E\} = E\left[Y - \left\{\bar{y} + \frac{r_{XY}\,\sigma_Y}{\sigma_X}(X - \bar{x})\right\}\right]$

$= (\bar{y} - \bar{y}) - \frac{r_{XY}\,\sigma_Y}{\sigma_X}(\bar{x} - \bar{x})$

$= 0$

$\sigma^2_{(Y - Y_E)} = E\{(Y - Y_E)^2\} - E^2(Y - Y_E)$

$= E\left[Y - \left\{\bar{y} + \frac{r_{XY}\,\sigma_Y}{\sigma_X}(X - \bar{x})\right\}\right]^2$

$= E\left[(Y - \bar{y})^2 + \frac{r^2_{XY}\,\sigma^2_Y}{\sigma^2_X}(X - \bar{x})^2 - \frac{2r_{XY}\,\sigma_Y}{\sigma_X}(X - \bar{x})(Y - \bar{y})\right]$

(i.e.) $S^2_Y = \sigma^2_Y + \frac{r^2_{XY}\,\sigma^2_Y}{\sigma^2_X}\sigma^2_X - \frac{2r_{XY}\,\sigma_Y}{\sigma_X}\text{Cov}(X, Y)$

$= \sigma^2_Y + r^2_{XY} \cdot \sigma^2_Y - 2r_{XY}\,\sigma^2_Y$

[Θ Cov $(X, Y) = r_{XY}\,\sigma_X\,\sigma_Y$]

$= (1 - r^2_{XY})\,\sigma^2_Y$ or $S_Y = \sqrt{1 - r^2_{XY}}\,\sigma_Y$ (1)

Similarly, the standard error of estimate of X, denoted by S_X is given by

$$S^2_X = (1 - r^2_{XY})\,\sigma^2_X \text{ or } S_X = \sqrt{1 - r^2_{XY}}\,\sigma_X \quad (2)$$

[**Note:** We may use (1) or (2) to prove that $|r_{XY}| \le 1$.

From (1), $S_Y = \sqrt{1 - r^2_{XY}}\,\sigma_Y$

Since S_Y and σ_Y are positive, $1 - r^2_{XY} \ge 0$

$\therefore$ $r^2_{XY} \le 1$

i.e., $|r_{XY}| \le 1$ or $-1 \le r_{XY} \le 1$]

Worked Example 4(C)

Example 1

Obtain the equations of the lines of regression from the following data:

X:	1	2	3	4	5	6	7
Y:	9	8	10	12	11	13	14

X	Y	$U = X - 4$	$V = Y - 11$	U^2	V^2	UV
1	9	−3	−2	9	4	6
2	8	−2	−3	4	9	6
3	10	−1	−1	1	1	1
4	12	0	1	0	1	0
5	11	1	0	1	0	0
6	13	2	2	4	4	4
7	14	3	3	9	9	9
	Total	0	0	28	28	26

$$\bar{x} = E(X) = 4 + \frac{1}{n}\Sigma u = 4$$

$$\bar{y} = E(Y) = 11 + \frac{1}{n}\Sigma v = 11$$

$$\sigma^2_X = \frac{1}{n}\Sigma u^2 - \left(\frac{1}{n}\Sigma u\right)^2 = \frac{1}{7} \times 28 = 4$$

$$\sigma^2_Y = \frac{1}{n}\Sigma v^2 - \left(\frac{1}{n}\Sigma v\right)^2 = \frac{1}{7} \times 28 = 4$$

$$C_{XY} = \frac{1}{n}\Sigma uv - \left(\frac{1}{n}\Sigma u\right)\cdot\left(\frac{1}{n}\Sigma v\right) = \frac{1}{7} \times 26 = 3.7$$

The regression line of Y on X is

$$y - \bar{y} = \frac{p_{XY}}{\sigma^2_X}(x - \bar{x})$$

i.e., $y - 11 = \frac{3.7}{4}(x - 4)$

i.e., $3.7x - 4y + 29.2 = 0$

The regression line of X on Y is

$$x - \bar{x} = \frac{p_{XY}}{\sigma^2_X}(y - \bar{y})$$

i.e., $x - 4 = \frac{3.7}{4}(y - 11)$

i.e., $4x - 3.7y + 24.7 = 0$

Example 2

Obtain the equations of the regression lines from the following data, using the method of least squares. Hence find the coefficient of correlation between X and Y. Also estimate the value of (i) Y, when $X = 38$ and (ii) X, when $Y = 18$.

X:	22	26	29	30	31	31	34	35
Y:	20	20	21	29	27	24	27	31

Put $U = X - 29$ and $V = Y - 27$.

Let the equation of the regression line of Y on X be $y = Ax + B$ or equivalently $v = au + b$ (1)

The normal equations for finding a and b are

$$a\,\Sigma u + nb = \Sigma v \quad (2)$$

and $$a\,\Sigma u^2 + b\Sigma u = \Sigma uv \quad (3)$$

x	y	$u = x - 29$	$v = y - 27$	u^2	v^2	uv
22	20	−7	−7	49	49	49
26	20	−3	−7	9	49	21
29	21	0	−6	0	36	0
30	29	1	2	1	4	2
31	27	2	0	4	0	0
31	24	2	−3	4	9	−6
34	27	5	0	25	0	0
35	31	6	4	36	16	24
	Total	6	−17	128	163	90

Using the relevant values from the table in (2) and (3), we have

$$6a + 8b = -17 \quad (2)'$$

$$128a + 6b = 90 \quad (3)'$$

Solving (2)′ and (3)′, we get

$$a = 0.83;\ b = -2.75$$

Hence the regression line of Y on X is

$$y - 27 = 0.83\,(x - 29) - 2.75$$

i.e., $$y = 0.83x + 0.18 \quad (4)$$

Let the equation of the regression line of X on Y be $x = Cy + D$ or equivalently $u = cv + d$ (5)

The normal equations for finding c and d are

$$c\Sigma v + nd = \Sigma u \quad (6)$$

and $$c\Sigma v^2 + d\Sigma v = \Sigma uv \quad (7)$$

Using the relevant values from the table in (6) and (7), we have

$$-17c + 8d = 6 \quad (6)'$$

$$163c - 17d = 90 \quad (7)'$$

Solving (6)′ and (7)′, we get

$$c = 0.81;\ d = 2.47$$

Hence the regression line of X on Y is

$$x - 29 = 0.81(y - 27) + 2.47$$

i.e., $$x = 0.81y + 9.60 \quad (8)$$

Comparing equation (4) with

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

We get $$r\frac{\sigma_Y}{\sigma_X} = 0.83 \quad (9)$$

Comparing equation (8) with

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

We get $$r\frac{\sigma_X}{\sigma_Y} = 0.81 \quad (10)$$

From (9) and (10), we get $r^2 = 0.83 \times 0.81$

$$\therefore \quad r = 0.82 \left(\because b_1 = \frac{r\,\sigma_Y}{\sigma_X} \text{ and } b_2 = \frac{r\,\sigma_X}{\sigma_Y} \text{ are both positive} \right)$$

We use equation (4) to estimate the value of Y when $X = 38$.

$$\therefore \quad Y = 0.83 \times 38 + 0.18 = 31.72$$

Using equation (8) to estimate the value of X when $Y = 18$, we have

$$X = 0.81 \times 18 + 9.60 = 24.18$$

Example 3

A study of prices of rice at Chennai and Madurai gave the following data:

	Chennai	*Madurai*
Mean	19.5	17.75
S.D.	1.75	2.5

Also the coefficient of correlation between the two is 0.8. Estimate the most likely price of rice (i) at Chennai corresponding to the price of 18 at Madurai and (ii) at Madurai corresponding to the price of 17 at Chennai.

Let the prices of rice at Chennai and Madurai be denoted by X and Y respectively. Then from the data,

$$\bar{x} = 19.5,\ \bar{y} = 17.75,\ \sigma_X = 1.75,\ \sigma_Y = 2.5 \text{ and } r_{XY} = 0.8.$$

Regression line of X on Y is

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

i.e., $$x - 19.5 = \frac{0.8 \times 1.75}{2.5}(y - 17.75)$$

$\therefore$ When $y = 18$,

$$x = 19.5 + \frac{0.8 \times 1.75}{2.5} \times 0.25$$

$$= 19.64$$

Regression line of Y on X is

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

i.e., $$y - 17.75 = \frac{0.8 \times 2.5}{1.75}(x - 19.5)$$

$\therefore$ When $x = 17$,

$$y = 17.75 + \frac{0.8 \times 2.5}{1.75} \times (-2.5)$$

$$= 14.89$$

Example 4

In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible: Variance of $X = 1$. The regression equations are $3x + 2y = 26$ and $6x + y = 31$. What were (i) the mean values of X and Y? (ii) the standard deviation of Y? and (iii) the correlation coefficient between X and Y?

(i) Since the lines of regression intersect at $(\bar{x}, \bar{y})$, we have $3\bar{x} + 2\bar{y} = 26$ and $6\bar{x} + \bar{y} = 31$

Solving these equations, we get $\bar{x} = 4$ and $\bar{y} = 7$.

(ii) Which of the two equations is the regression equation of Y on X and which one is the regression equation of X on Y are not known.

Let us tentatively assume that the first equation is the regression line of X on Y and the second equation is the regression line of Y on X. Based on this assumption, the first equation can be re-written as

$$x = -\frac{2}{3}y + \frac{26}{3} \quad (1)$$

and the other as $$y = -6x + 31 \quad (2)$$

Then $b_{XY} = -\frac{2}{3}$ and $b_{YX} = -6$

$\therefore$ $$r_{XY}^2 = b_{XY} \times b_{YX} = 4$$

$\therefore$ $$r_{XY} = -2,$$ which is absurd.

Hence our tentative assumption is wrong.

$\therefore$ The first equation is the regression line of Y on X and re-written as

$$y = -\frac{3}{2}x + 13 \quad (3)$$

The second equation is the regression line of X on Y and re-written as

$$x = -\frac{1}{6}y + \frac{31}{6} \quad (4)$$

Hence the correct $b_{YX} = -\frac{3}{2}$ and the correct $b_{XY} = -\frac{1}{6}$

$\therefore$ $$r_{XY}^2 = b_{YX} \cdot b_{XY} = \frac{1}{4}$$

$\therefore$ $$r_{XY} = -\frac{1}{2}$$ ($\because$ both b_{YX} and b_{XY} are negative)

(iii) Now $$\frac{\sigma_Y^2}{\sigma_X^2} = \frac{b_{YX}}{b_{XY}} = \frac{-\frac{3}{2}}{-\frac{1}{6}} = 9$$

$\therefore$ $$\sigma_Y^2 = 9 \times \sigma_x^2 = 9$$

$\therefore$ $$\sigma_Y = 3$$

Example 5

Given that $x = 4y + 5$ and $y = kx + 4$ are the regression lines of X on Y and of Y on X respectively, show that $0 \le k \le \frac{1}{4}$. If $k = \frac{1}{16}$, find the means of X and Y and r_{XY}.

From the given equations, we note that

$$b_{YX} = k \text{ and } b_{XY} = 4$$

$$r_{XY}^2 = b_{XY} \cdot b_{YX} = 4k$$

Since $0 \le r_{XY}^2 \le 1$, we get $0 \le 4k \le 1$

$$\therefore \quad 0 \le k \le \frac{1}{4}.$$

When $k = \frac{1}{16}$, $r_{XY}^2 = \frac{1}{4}$

$$\therefore \quad r_{XY} = \pm \frac{1}{2}$$

But both b_{YX} and b_{XY} are positive.

$$\therefore \quad r_{XY} = \frac{1}{2}$$

When $k = \frac{1}{16}$, the regression equations become

$$x = 4y + 5 \tag{1}$$

and

$$y = \frac{1}{16}x + 4 \tag{2}$$

Solving equations (1) and (2), we get

$$x = 28 \text{ and } y = 5.75$$

$$\therefore \quad \bar{x} = 28 \text{ and } \bar{y} = 5.75$$

Example 6

Find the angle between the two lines of regression. Deduce the condition for the two lines to be (i) at right angles and (ii) coincident.

The equations of the regression lines

are

$$y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(x - \bar{x}) \tag{1}$$

and

$$x - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(y - \bar{y}) \tag{2}$$

Slope of line (1) $= r\frac{\sigma_Y}{\sigma_X} = m_1$, say.

Slope of line (2) $= \frac{\sigma_Y}{r\sigma_X}$, m_2, say.

If θ is the acute angle between the two lines, then $\tan\theta = \frac{|m_1 - m_2|}{1 + m_1 m_2}$

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

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

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

The two regression lines are at angles when $\theta = \frac{\pi}{2}$, i.e., $\tan\theta = \infty$

i.e., $r = 0$

$\therefore$ When the linear correlation between X and Y is zero, the two lines of regression will be at right angles.

The two regression lines are coincident, when $\theta = 0$, i.e., when $\tan\theta = 0$

i.e., when $r = \pm 1$.

$\therefore$ When the correlation between X and Y is perfect, the two regression lines will coincide.

Example 7

For two R.V.'s X and Y with the same mean, the two regression equations are $y = ax + b$ and $x = cy + d$. Find the common mean, ratio of the standard deviations and also show that $\frac{b}{d} = \frac{1-a}{1-c}$.

If μ is the common mean, the point (μ, μ) lies on $y = ax + b$ and $x = cy + d$ [Θ They intersect at $(\bar{x}, \bar{y})$]

$$\therefore \quad \mu = a\mu + b \tag{1}$$

$$\mu = c\mu + d \tag{2}$$

From (1),

$$\mu = \frac{b}{1-a}$$

From (2), $\mu = \dfrac{d}{1-c}$

$\therefore \quad \dfrac{b}{1-a} = \dfrac{d}{1-c}$

$\therefore \quad \dfrac{b}{d} = \dfrac{1-a}{1-c}$

Now $\dfrac{\sigma_Y^2}{\sigma_X^2} = \dfrac{b_{YX}}{b_{XY}} = \dfrac{a}{c}$ $\qquad \therefore \dfrac{\sigma_Y}{\sigma_X} = \sqrt{\dfrac{a}{c}}$

Example 8

Find the standard error of estimate of Y on X and of X on Y from the following data:

X:	1	2	3	4	5
Y:	2	5	9	13	14

x	y	x^2	y^2	xy
1	2	1	4	2
2	5	4	25	10
3	9	9	81	27
4	13	16	169	52
5	14	25	196	70
15	43	55	475	161

$$r_{XY} = \frac{n\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{\{n\Sigma x^2 - (\Sigma x)^2\}\{n\Sigma y^2 - (\Sigma y)^2\}}}$$

$$= \frac{5 \times 161 - 15 \times 43}{\sqrt{\{5 \times 55 - (15)^2\}\{5 \times 475 - (43)^2\}}}$$

$$= \frac{160}{\sqrt{50 \times 526}} = 0.9866$$

$$\sigma_X^2 = \frac{1}{n}\Sigma x^2 - \left(\frac{1}{n}\Sigma x\right)^2$$

$$= \frac{1}{5} \times 55 - \left(\frac{1}{5} \times 15\right)^2 = 2$$

$\therefore \quad \sigma_X = 1.4142$

$$\sigma_Y^2 = \frac{1}{n}\Sigma y^2 - \left(\frac{1}{n}\Sigma y\right)^2$$

$$= \frac{1}{5} \times 475 - \left(\frac{1}{5} \times 43\right)^2$$

$$= 21.04$$

$\therefore \quad \sigma_Y = 4.5869$

$$S_Y = \sqrt{1 - r_{XY}^2} \cdot \sigma_Y = \sqrt{1 - (0.9866)^2} \times 4.5869$$

$$= 0.7484$$

$$S_X = \sqrt{1 - r_{XY}^2} \cdot \sigma_X = \sqrt{1 - (0.9866)^2} \times 1.4142$$

$$= 0.2307$$

Exercise 4 (C)

Part A

(Short Answer Question)

1. What do you mean by regression line? What is its use?
2. For a given data of n pairs of values of X and Y, why should there be two regression lines?
3. Write down the analytic equations of the regression lines?
4. When will the two regression lines be (i) at right angles, (ii) coincident?
5. Define regression coefficients.
6. Prove that the correlation coefficient is the geometric mean of the regression coefficients.
7. Find the co-ordinates of the point of intersection of the regression lines.
8. What do you mean by standard error of estimate?
9. Write down the formulas for the standard errors of estimate of Y and X.
10. In the usual notation prove that
    (i) $r_{XY} \cdot S_X S_Y = (1 - r_{XY}^2)\, C_{XY}$ and
    (ii) $b_1 S_X^2 = b_2 S_Y^2$.

Part B

11. Find the equations of the regression lines from the following data. Hence calculate the coefficient of correlation between X and Y.

X:	62	64	65	69	70	71	72	74
Y:	126	125	139	145	165	152	180	208

12. Find the equations of the regression lines from the following data. Also estimate the value of Y when $X = 71$ and the value of X when $Y = 70$.

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

13. Find the equation of the regression line of Y on X using the method of least squares from the following data. Find the value of Y corresponding to $X = 18$.

X:	5	10	15	20	25
Y:	16	19	23	26	30

14. Obtain the line of regression of X on Y using the method of least squares from the following data. Find the value of X when $Y = 45$.

X:	4.7	8.2	12.4	15.8	20.7	24.9	31.9	35.0	39.1	38.8
Y:	4.0	8.0	12.5	16.0	20.0	25.0	31.0	36.0	40.0	40.0

15. Find the most likely price in Mumbai corresponding to the price of Rs. 70 at Chennai and that in Chennai corresponding to the price of Rs. 75 at Mumbai from the following:

	Chennai	Mumbai
Mean	65	67
S.D.	2.5	3.5

Coefficient of correlation between the prices in the two cities is 0.8.

16. In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible.
Variance of $X = 9$. Regression equations are $8x - 10y + 66 = 0$ and $40x - 18y = 214$. What were (i) the mean values of X and Y?
(ii) the correlation coefficient between X and Y and (iii) the standard deviation of Y?
17. The equations of two regression lines got in a correlation analysis are $3x + 12y = 19$ and $3y + 9x = 46$. Obtain (i) the correlation coefficient between X and Y, (ii) the mean values of X and Y and (iii) the ratio of the coefficient of variation of X to that of Y.
18. The equations of lines of regression are given by $x + 2y - 5 = 0$ and $2x + 3y - 8 = 0$ and variance of X is 12. Compute the values of $\bar{x}$, $\bar{y}$, σ_Y^2 and r_{XY}.
19. The regression lines of Y on X and of X on Y are respectively $y = a + bx$ and $x = c + dy$. Find the values of $\bar{x}$, $\bar{y}$ and r_{XY}. Can you find S_X snd S_Y from them ?
20. If the lines of regression of Y on X and X on Y are respectively $a_1x + b_1y + c_1 = 0$ and $a_2x + b_2y + c_2 = 0$, prove that $a_1b_2 \le a_2b_1$. Find also the coefficient of correlation between X and Y and the ratio of the coefficient of variability of Y to that of X.

Characteristic Function

Although higher order moments of a RV X may be obtained directly by using the definition of $E(X^n)$, it will be easier in many problems to compute them through the characteristic function or equivalently through the moment generating function of the RV X. While the characteristic function always exists, the moment generating function need not.

Moment Generating Function (MGF) of a RV X (discrete or continuous) is defind as $E(e^{tX})$, where t is a real variable and denoted as $M(t)$.

If X is discrete, then $M(t) = \sum_r e^{tx_r} p_r$,

where X takes the values $x_1, x_2, x_3, \ldots$, with probabilities $p_1, p_2, p_3, \ldots$

If X is a continuous RV with density function $f(x)$, then

$$M(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

Properties of MGF

(Proofs of the properties are omitted, as the proofs of the corresponding properties of characteristic function will be given later.)

1. $M(t) = \sum_{n=0}^{\infty} t^n E(X^n)/\lfloor n$

   i.e., $E(X^n) = \mu'_n$ is the co-efficient of $\frac{t^n}{\lfloor n}$ in the expansion of $M(t)$ in series of powers of t.

2. $\mu'_n = E(X^n) = \left[\frac{d^n}{dt^n} M(t)\right]_{t=0}$.
3. If the MGF of X is $M_X(t)$ and if $Y = aX + b$, then $M_Y(t) = e^{bt} M_X(at)$.
4. If X and Y are independent RVs and $Z = X + Y$, then $M_Z(t) = M_X(t) M_Y(t)$.

Characteristic function of a RV X (discrete or continuous) is defined as $E(e^{i\omega X})$ and denoted as $\phi(\omega)$.

If X is a discrete RV that can take the values $x_1, x_2, \ldots$, such that $P(X = x_r) = p_r$, then

$$\phi(\omega) = \sum_r e^{i\omega x_r} p_r$$

If X is a continuous RV with density function $f(x)$, then

$$\phi(\omega) = \int_{-\infty}^{\infty} e^{i\omega x} f(x)\, dx$$

Properties of Characteristic Function

1. $\mu'_n = E(X^n)$ = the coefficient of $\frac{i^n \omega^n}{n!}$ in the expansion of $\phi(\omega)$ in series of ascending powers of $i\omega$.

and $$W^2 + Z^2 = \frac{1}{(1-\rho^2)}\left[\frac{X^2}{\sigma_1^2} - \frac{2\rho XY}{\sigma_1\sigma_2} + \frac{Y^2}{\sigma_2^2}\right]$$

Deduce that the joint probability differential of W and Z is

$$dP = \frac{1}{2\pi} \cdot \exp\left[-\frac{1}{2}(w^2 + z^2)\right] dw\, dz$$

and hence that W, Z are independent normal variates with zero means and unit S.D.'s *[Meerut Univ. M.Sc., 1993]*

Hence or otherwise obtain the m.g.f. of the bivariate normal distribution.

22. From a standard bivariate normal population, a random sample of n observations (X_i, Y_i), $(i = 1, 2, \ldots, n)$ is drawn. Show that the distribution of

$$Z_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \text{ and } Z_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

has the moment generating function :

$$\text{Constant}\left[\left(1 - \frac{2t_1}{n}\right)\left(1 - \frac{2t_2}{n}\right) - \frac{4\rho^2 t_1 t_2}{n^2}\right]^{-n/2}$$

Hint. $M_{Z_1, Z_2}(t_1, t_2) = \left[E \exp\left(\frac{t_1 x^2}{n} + \frac{t_2 y^2}{n}\right)\right]^n$

$$= \left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[x^2\left(\frac{t_1}{n} - \frac{1}{2(1-\rho^2)}\right) + \left(\frac{\rho}{1-\rho^2}\right)xy + y^2\left(\frac{t_2}{n} - \frac{1}{2(1-\rho^2)}\right)\right] dx\,dy\right\}^n$$

Now use the result

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp[-(ax^2 + 2hxy + by^2)]\, dx\,dy = \frac{\pi}{\sqrt{ab - h^2}}$$

and simplify.

10·11. Multiple and Partial Correlation. When the values of one variable are associated with or influenced by other variable, *e.g.*, the age of husband and wife, the height of father and son, the supply and demand of a commodity and so on, Karl Pearson's coefficient of correlation can be used as a measure of linear relationship between them. But sometimes there is interrelation between many variables and the value of one variable may be influenced by many others, *e.g.*, the yield of crop per acre say (X_1) depends upon quality of seed (X_2), fertility of soil (X_3), fetilizer used (X_4), irrigation facilities (X_5), weather conditions (X_6) and so on. Whenever we are interested in studying the joint effect of a group of variables upon a variable not included in that group, our study is that of *multiple correlation and multiple regression.*

Suppose in a trivariate or multi-variate distribution we are interested in the relationship between two variables only. The are two alternatives, *viz.*, (*i*) we

consider only those two members of the observed data in which the other members have specified values or (*ii*) we may eliminate mathematically the effect of other variates on two variates. The first method has the disadvantage that it limits the size of the data and also it will be applicable to only the data in which the other variates have assigned values. In the second method it may not be possible to eliminate the entire influence of the variates but the linear effect can be easily eliminated. The correlation and regression between only two variates eliminating the linear effect of other variates in them is called the *partial correlation and partial regression.*

10·11·1. Yule's Notation. Let us consider a distribution involving three random variables X_1, X_2 and X_3. Then the equation of the plane of regression of X_1 on X_2 and X_3 is

$$X_1 = a + b_{12\cdot3}X_2 + b_{13\cdot2}X_3 \qquad \ldots(10\cdot28)$$

Without loss of generality we can assume that the variables X_1, X_2 and X_3 have been measured from their respective means, so that

$$E(X_1) = E(X_2) = E(X_3) = 0$$

Hence on taking expectation of both sides in (10·28), we get $a = 0$.

Thus the plane of regression of X_1 on X_2 and X_3 becomes

$$X_1 = b_{12\cdot3}X_2 + b_{13\cdot2}X_3 \qquad \ldots(10\cdot28a)$$

The coefficients $b_{12\cdot3}$ and $b_{13\cdot2}$ are known as the *partial regression coefficients* of X_1 on X_2 and of X_1 on X_3 respectively.

$$e_{1\cdot23} = b_{12\cdot3}X_2 + b_{13\cdot2}X_3$$

is called the estimate of X_1 as given by the plane of regression (10·28*a*) and the quantity

$$X_{1\cdot23} = X_1 - b_{12\cdot3}X_2 - b_{13\cdot2}X_3,$$

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

In the general case of n variables $X_1, X_2, \ldots, X_n$ the equation of the plane of regression of X_1 on $X_2, X_3, \ldots, X_n$ becomes

$$X_1 = b_{12\cdot34\ldots n}X_2 + b_{13\cdot24\ldots n}X_3 + \ldots + b_{1n\cdot23\cdots(n-1)}X_n$$

The error of estimate or residual is given by

$$X_{1\cdot23\ldots n} = X_1 - (b_{12\cdot34\ldots n}X_2 + b_{13\cdot24\ldots n}X_3 + \ldots + b_{1n\cdot23\cdots(n-1)}X_n)$$

The notations used here are due to Yule. The subscripts before the dot (.) are known as *primary subscripts* and those after the dot are called *secondary subscripts.* The order of a regression coefficient is determined by the number of secondary subscripts, *e.g.*,

$$b_{12\cdot3}, b_{12\cdot34}, \ldots, b_{12\cdot34\ldots n}$$

are the regression coefficients of order 1, 2, ... $(n-2)$ respectively. Thus in general, a regression coefficient with p-secondary subscripts will be called a regression co-efficient of order 'p'. It may be pointed out that the order in which the secondary subscripts are written is immaterial but the order of the primary subscripts is important, *e.g.*, in $b_{12\cdot34\ldots n}$, X_2 is independent while X_1 is dependent variable but in $b_{21\cdot34\ldots n}$, X_1 is independent while X_2 is dependent

variable. Thus of the two primary subscripts, former refers to dependent variable and the latter to independent variable.

The order of a residual is also determined by the number of secondary subscripts in it, *e.g.*, $X_{1\cdot23}, X_{1\cdot234}, \ldots, X_{1\cdot23\ldots n}$ are the residuals of order 2, 3, ..., $(n-1)$ respectively.

Remark. In the following sequences we shall assume that the variables under consideration have been measured from their respective means.

10·12. Plane of Regression. The equation of the plane of regression of X_1 on X_2 and X_3 is

$$X_1 = b_{12\cdot3} X_2 + b_{13\cdot2} X_3 \qquad \ldots(10\cdot29)$$

The constants b's in (10·29) are determined by the principle of least squares, *i.e.*, by minimising the sum of the squares of the residuals, *viz.*,

$$S = \Sigma X_{1\cdot23}^2 = \Sigma(X_1 - b_{12\cdot3} X_2 - b_{13\cdot2} X_3)^2,$$

the summation being extended to the given values (N in number) of the variables.

The normal equations for estimating $b_{12\cdot3}$ and $b_{13\cdot2}$ are

$$\left.\begin{aligned} \frac{\partial S}{\partial b_{12\cdot3}} &= 0 = -2\,\Sigma X_2(X_1 - b_{12\cdot3} X_2 - b_{13\cdot2} X_3) \\ \frac{\partial S}{\partial b_{13\cdot2}} &= 0 = -2\,\Sigma X_3(X_1 - b_{12\cdot3} X_2 - b_{13\cdot2} X_3) \end{aligned}\right\} \qquad \ldots(10\cdot30)$$

i.e.,
$$\Sigma X_2 X_{1\cdot23} = 0 \quad \text{and} \quad \Sigma X_3 X_{1\cdot23} = 0 \qquad \ldots(10\cdot30a)$$

$$\Rightarrow \left.\begin{aligned} \Sigma X_1 X_2 - b_{12\cdot3}\,\Sigma X_2^2 - b_{13\cdot2}\,\Sigma X_2 X_3 &= 0 \\ \Sigma X_1 X_3 - b_{12\cdot3}\,\Sigma X_2 X_3 - b_{13\cdot2}\,\Sigma X_3^2 &= 0 \end{aligned}\right\} \qquad \ldots(10\cdot30b)$$

Since X_i's are measured from their respective means, we have

$$\left.\begin{aligned} \sigma_i^2 &= \frac{1}{N}\Sigma X_i^2,\ \text{Cov}\,(X_i, X_j) = \frac{1}{N}\Sigma X_i X_j \\ \text{and} \qquad r_{ij} &= \frac{\text{Cov}\,(X_i, X_j)}{\sigma_i \sigma_j} = \frac{\Sigma X_i X_j}{N \sigma_i \sigma_j} \end{aligned}\right\} \qquad \ldots(10\cdot30c)$$

Hence from (10·30*b*), we get

$$\left.\begin{aligned} r_{12}\sigma_1\sigma_2 - b_{12\cdot3}\sigma_2^2 - b_{13\cdot2} r_{23}\sigma_2\sigma_3 &= 0 \\ r_{13}\sigma_1\sigma_3 - b_{12\cdot3} r_{23}\sigma_2\sigma_3 - b_{13\cdot2}\sigma_3^2 &= 0 \end{aligned}\right\} \qquad \ldots(10\cdot30d)$$

Solving the equations (10·30*d*) for $b_{12\cdot3}$ and $b_{13\cdot2}$, we get

$$b_{12\cdot3} = \frac{\begin{vmatrix} r_{12}\sigma_1 & r_{23}\sigma_3 \\ r_{13}\sigma_1 & \sigma_3 \end{vmatrix}}{\begin{vmatrix} \sigma_2 & r_{23}\sigma_3 \\ r_{23}\sigma_2 & \sigma_3 \end{vmatrix}} = \frac{\sigma_1}{\sigma_2} \cdot \frac{\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} \qquad \ldots(10\cdot31)$$

Similarly, we will get

$$b_{13\cdot2} = \frac{\sigma_1}{\sigma_3} \cdot \frac{\begin{vmatrix} 1 & r_{12} \\ r_{23} & r_{13} \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} \qquad \ldots(10\cdot31a)$$

If we write

$$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} \qquad (10\cdot32)$$

and ω_{ij} is the cofactor of the element in the ith row and jth column of ω, we have from (10·31) and (10·31a)

$$b_{12\cdot3} = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} \quad \text{and} \quad b_{13\cdot2} = -\frac{\sigma_1}{\sigma_3} \cdot \frac{\omega_{13}}{\omega_{11}} \qquad \ldots(10\cdot33)$$

Substituting these values in (10·29), we get the required equation of the plane of regression of X_1 on X_2 and X_3 as

$$X_1 = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} \cdot X_2 - \frac{\sigma_1}{\sigma_3} \cdot \frac{\omega_{13}}{\omega_{11}} \cdot X_3$$

$$\Rightarrow \qquad \frac{X_1}{\sigma_1} \cdot \omega_{11} + \frac{X_2}{\sigma_2} \cdot \omega_{12} + \frac{X_3}{\sigma_3} \cdot \omega_{13} = 0 \qquad \ldots(10\cdot34)$$

Aliter. Eliminating the coefficient $b_{12\cdot3}$ and $b_{13\cdot2}$ in (10·29) and (10·30d), the required equation of the plane of regression of X_1 on X_2 and X_3 becomes

$$\begin{vmatrix} X_1 & X_2 & X_3 \\ r_{12}\sigma_1\sigma_2 & \sigma_2^2 & r_{23}\sigma_2\sigma_3 \\ r_{13}\sigma_1\sigma_3 & r_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{vmatrix} = 0$$

Dividing C_1, C_2 and C_3 by σ_1, σ_2 and σ_3 respectively and also R_2 and R_3 by σ_2 and σ_3 respectively, we get

$$\begin{vmatrix} \frac{X_1}{\sigma_1} & \frac{X_2}{\sigma_2} & \frac{X_3}{\sigma_3} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix} = 0$$

$$\Rightarrow \qquad \frac{X_1}{\sigma_1}\omega_{11} + \frac{X_2}{\sigma_2}\omega_{12} + \frac{X_3}{\sigma_3}\omega_{13} = 0$$

where ω_{ij} is defined in (10·32).

10·12·1. Generalisation. In general, the equation of the plane of regression of X_1 on $X_2, X_3, \ldots X_n$ is

$$X_1 = b_{12\cdot34\ldots n}X_2 + b_{13\cdot24\ldots n}X_3 + \ldots + b_{1n\cdot23\ldots(n-1)}X_n \qquad \ldots(10\cdot35)$$

The sum of the squares of residuals is given by

$$S = \Sigma X^2_{1\cdot23\ldots n}$$

$$= \Sigma (X_1 - b_{12\cdot34\ldots n} X_2 - b_{13\cdot24\cdots n} X_3 - \ldots - b_{1n\cdot23\ldots(n-1)} X_n)^2$$

Using the principle of least squares, the normal equations for estimating the $(n-1)$, b's are

$$\left.\begin{aligned}
\frac{\partial S}{\partial b_{12\cdot34\ldots n}} &= 0 = -2\Sigma X_2(X_1 - b_{12\cdot34\ldots n} X_2 - b_{13\cdot24\ldots n} X_3 - \ldots - b_{1n\cdot23\ldots(n-1)} X_n) \\
\frac{\partial S}{\partial b_{13\cdot24\ldots n}} &= 0 = -2\Sigma X_3(X_1 - b_{12\cdot34\ldots n} X_2 - b_{13\cdot24\ldots n} X_3 - \ldots - b_{1n\cdot23\ldots(n-1)} X_n) \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
\frac{\partial S}{\partial b_{1n\,23\ldots(n-1)}} &= 0 = -2\Sigma X_n(X_1 - b_{12\cdot34\ldots n} X_2 - b_{13\cdot24\ldots n} X_3 - \ldots - b_{1n\cdot23\ldots(n-1)} X_n)
\end{aligned}\right\} \quad \ldots(10\cdot36)$$

$$i.e., \qquad \Sigma X_i\, X_{1\cdot23\ldots n} = 0, \ (i = 2, 3, \ldots, n) \qquad \ldots(10\cdot36a)$$

which on smplification after using (10·30c), give

$$\left.\begin{aligned}
r_{12}\sigma_1\sigma_2 &= b_{12\cdot34\ldots n}\sigma_2^2 + b_{13\cdot24\ldots n} r_{23}\sigma_2\sigma_3 + \ldots + b_{1n\cdot23\ldots(n-1)} r_{2n}\sigma_2\sigma_n \\
r_{13}\sigma_1\sigma_3 &= b_{12\cdot34\ldots n} r_{23}\sigma_2\sigma_3 + b_{13\cdot24\ldots n}\sigma_3^2 + \ldots + b_{1n\cdot23\ldots(n-1)} r_{3n}\sigma_3\sigma_n \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
r_{1n}\sigma_1\sigma_n &= b_{12\cdot34\ldots n} r_{2n}\sigma_2\sigma_n + b_{13\cdot24\ldots n} r_{3n}\sigma_3\sigma_n + \ldots + b_{1n\cdot23\ldots(n-1)}\sigma_n^2
\end{aligned}\right\} \quad \ldots(10\cdot36b)$$

Hence the eliminant of b's between (10·35) and (10·36b) is

$$\begin{vmatrix}
X_1 & X_2 & X_3 & \ldots & X_n \\
r_{12}\sigma_1\sigma_2 & \sigma^2_2 & r_{23}\sigma_2\sigma_3 & \ldots & r_{2n}\sigma_2\sigma_n \\
r_{13}\sigma_1\sigma_3 & r_{23}\sigma_2\sigma_3 & \sigma_3^2 & \ldots & r_{3n}\sigma_3\sigma_n \\
\vdots & \vdots & \vdots & & \vdots \\
r_{1n}\sigma_1\sigma_n & r_{2n}\sigma_2\sigma_n & r_{3n}\sigma_3\sigma_n & \ldots & \sigma_n^2
\end{vmatrix} = 0 .$$

Dividing $C_1, C_2, \ldots, C_n$ by $\sigma_1, \sigma_2, \ldots \sigma_n$ respectively and $R_2, R_3, \ldots R_n$ by $\sigma_2, \sigma_3, \ldots, \sigma_n$ respectively, we get

$$\begin{vmatrix}
\frac{X_1}{\sigma_1} & \frac{X_2}{\sigma_2} & \frac{X_3}{\sigma_3} & \ldots & \frac{X_n}{\sigma_n} \\
r_{12} & 1 & r_{32} & \ldots & r_{2n} \\
r_{13} & r_{23} & 1 & \ldots & r_{3n} \\
\vdots & \vdots & \vdots & & \vdots \\
r_{1n} & r_{2n} & r_{3n} & & 1
\end{vmatrix} = 0 \qquad \ldots(10\cdot37)$$

If we write

$$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} \\ r_{21} & 1 & r_{23} & \dots & r_{2n} \\ r_{31} & r_{32} & 1 & \dots & r_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \dots & 1 \end{vmatrix} \qquad \dots(10\cdot38)$$

and ω_{ij} is the cofactor of the element in the ith row and jth column of ω, we get from (10·37)

$$\frac{X_1}{\sigma_1} \cdot \omega_{11} + \frac{X_2}{\sigma_2}\omega_{12} + \frac{X_3}{\sigma_3}\omega_{13} + \dots + \frac{X_n}{\sigma_n}\omega_{1n} = 0 \qquad \dots(10\cdot39)$$

as the required equation of the plane of regression of X_1 on $X_2, X_3, \dots, X_n$.

Equation (10·39) can be re-written as

$$X_1 = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} X_2 - \frac{\sigma_1}{\sigma_3} \cdot \frac{\omega_{13}}{\omega_{11}} X_3 - \dots - \frac{\sigma_1}{\sigma_n} \cdot \frac{\omega_{1n}}{\omega_{11}} X_n \qquad \dots(10\cdot39a)$$

Comparing (10·39a) with (10·35), we get

$$\left.\begin{aligned} b_{12\cdot34\dots n} &= -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} \\ b_{13\cdot24\dots n} &= -\frac{\sigma_1}{\sigma_3} \cdot \frac{\omega_{13}}{\omega_{11}} \\ &\vdots \qquad\qquad \vdots \\ b_{1n\cdot23\dots(n-1)} &= -\frac{\sigma_1}{\sigma_n} \cdot \frac{\omega_{1n}}{\omega_{11}} \end{aligned}\right\} \qquad \dots(10\cdot40)$$

Remarks 1. From the symmetry of the result obtained in (10·40), the equation of the plane of regression of X_i, (say), on the remaining variables X_j $(j \neq i = 1, 2, \dots, n)$, is given by

$$\frac{X_1}{\sigma_1}\omega_{i1} + \frac{X_2}{\sigma_2}\omega_{i2} + \dots + \frac{X_i}{\sigma_i}\omega_{ii} + \dots + \frac{X_n}{\sigma_n}\omega_{in} = 0 \; ; \; i = 1, 2, , n \qquad \dots(10\cdot41)$$

2. We have

$$b_{12\cdot34\dots n} = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}}$$

and

$$b_{21\cdot34\dots n} = -\frac{\sigma_2}{\sigma_1} \frac{\omega_{21}}{\omega_{22}}$$

Since each of σ_1, σ_2, ω_{11} and ω_{22} is non-negative and $\omega_{12} = \omega_{21}$, [*c.f.* Remarks 3 and 4 to §10·14, page 10·113], the sign of each regression coefficient $b_{12\cdot34\dots n}$ and $b_{21\cdot34\dots n}$ depends on ω_{12}.

10·13. Properties of residuals

Property 1. *The sum of the product of any residual of order zero with any other residual of higher order is zero, provided the subscript of the former occurs among the secondary subscripts of the latter.*

The normal equations for estimating b's in trivariate and n-variate distributions, as obtained in equations (10·30a) and (10·36a), are

$$\Sigma X_2 X_{1\cdot 23} = 0, \Sigma X_3 X_{1\cdot 23} = 0$$

and $$\Sigma X_i X_{1\cdot 23\ldots n} = 0 \; ; \; i = 2, 3, \ldots, n$$

respectively. Here X_i, $(i = 1, 2, 3, \ldots, n)$ can be regarded as a residual of order zero. Hence the result.

Property 2. *The sum of the product of any two residuals in which all the secondary subscripts of the first occur among the secondary subscripts of the second is unaltered if we omit any or all of the secondary subscripts of the first. Conversely, the product sum of any residual of order 'p' with a residual of order $p + q$, the 'p' subscripts being the same in each case is unaltered by adding to the secondary subscripts of the former any or all the 'q' additional subscripts of the latter.*

Let us consider

$$\Sigma X_{1\cdot 2} X_{1\cdot 23} = \Sigma (X_1 - b_{12} X_2) X_{1\cdot 23} = \Sigma X_1 X_{1\cdot 23} - b_{12} \Sigma X_2 X_{1\cdot 23}$$
$$= \Sigma X_1 X_{1\cdot 23} \qquad (c.f.\ \text{Property 1})$$

Also $$\Sigma X_{1\cdot 23}^2 = \Sigma X_{1\cdot 23} X_{1\cdot 23} = \Sigma (X_1 - b_{12\cdot 3} X_2 - b_{13\cdot 2} X_3) X_{1\cdot 23}$$
$$= \Sigma X_1 X_{1\cdot 23} - b_{12\cdot 3} \Sigma X_2 X_{1\cdot 23} - b_{13\cdot 2} \Sigma X_3 X_{1\cdot 23}$$
$$= \Sigma X_1 X_{1\cdot 23} \qquad (c.f.\ \text{Property 1})$$

$$\therefore \quad \Sigma X_{1\cdot 23}^2 = \Sigma X_{1\cdot 2} X_{1\cdot 23} = \Sigma X_1 X_{1\cdot 23}$$

Again $\Sigma X_{1\cdot 34\ldots n} X_{2\cdot 34\ldots n}$

$$= \Sigma [(X_1 - b_{13\cdot 4\ldots n} X_3 - b_{14\cdot 35\ldots n} X_4 - \ldots - b_{1n\cdot 34\ldots(n-1)} X_n) X_{2\cdot 34\ldots n}]$$
$$= \Sigma X_1 X_{2\cdot 34\ldots n} \qquad (c.f.\ \text{Property 1})$$

Hence the property ?

Property 3. *The sum of the product of two residuals is zero if all the subscripts (primary as well as secondary) of the one occur among the secondary subscripts of the other, e.g.,*

$$\Sigma X_{1\cdot 2} X_{3\cdot 12} = \Sigma (X_1 - b_{12} X_2) X_{3\cdot 12} = \Sigma X_1 X_{3\cdot 12} - b_{12} \Sigma X_2 X_{3\cdot 12} = 0$$
$$(c.f.\ \text{Property 1})$$

$\Sigma X_{2\cdot 34\ldots n} X_{1\cdot 23\ldots n}$

$$= \Sigma [(X_2 - b_{23\cdot 4\ldots n} X_3 - b_{24\cdot 35\ldots n} X_4 - \ldots - b_{2n\cdot 34\ldots(n-1)} X_n) X_{1\cdot 23\ldots n}]$$
$$= \Sigma X_2 X_{1\cdot 23\ldots n} - b_{23\cdot 4\ldots n} \Sigma X_3 X_{1\cdot 23\ldots n} - b_{24\cdot 35\ldots n} \Sigma X_4 X_{1\cdot 23\ldots n}$$
$$\ldots - b_{2n\cdot 34\ldots(n-1)} \Sigma X_n X_{1\cdot 23\ldots n}$$
$$= 0 \qquad (c.f.\ \text{Property 1})$$

Hence the property 3.

10·13·1. Variance of the Residual. Let us consider the plane of regression of X_1 on $X_2, X_3, \ldots, X_n$ *viz.*,

$$X_1 = b_{12\cdot34\ldots n} X_2 + b_{13\cdot24\ldots n} X_3 + \ldots + b_{1n\cdot23\ldots(n-1)} X_n$$

Since all the X_i's are measured from their respective means, we have

$$E(X_i) = 0;\ i = 1, 2, \ldots, n \ \Rightarrow\ E(X_{1\cdot23\ldots n}) = 0$$

Hence the variance of the residual is given by

$$\sigma^2_{1\cdot23\ldots n} = \frac{1}{N}\Sigma[X_{1\cdot23\ldots n} - E(X_{1\cdot23\ldots n})]^2 = \frac{1}{N}\Sigma X^2_{1\cdot23\ldots n}$$

$$= \frac{1}{N}\Sigma X_{1\cdot23\ldots n} X_{1\cdot23\ldots n} = \frac{1}{N}\Sigma X_1 X_{1\cdot23\ldots n},$$

(*c.f.* Property 2 § 10·13)

$$= \frac{1}{N}\Sigma X_1 (X_1 - b_{12\cdot34\ldots n} X_2 - b_{13\cdot24\ldots n} X_3 - \ldots - b_{1n\cdot23\ldots(n-1)} X_n)$$

$$= \sigma_1^2 - b_{12\cdot34\ldots n} r_{12}\sigma_1\sigma_2 - b_{13\cdot24\ldots n} r_{13}\sigma_1\sigma_3 - \ldots - b_{1n\cdot23\ldots(n-1)} r_{1n}\sigma_1\sigma_n$$

$$\Rightarrow \quad \sigma_1^2 - \sigma^2_{1\cdot23\ldots n} = b_{12\cdot34\ldots n} r_{12}\sigma_1\sigma_2 - b_{13\cdot24\ldots n} r_{13}\sigma_1\sigma_3 - \ldots - b_{1n\cdot23\ldots(n-1)} r_{1n}\sigma_1\sigma_n \quad \ldots(10\cdot42)$$

Eliminating the b's in equations (10·42) and (10·36*b*), we get

$$\begin{vmatrix} \sigma_1^2 - \sigma^2_{1\cdot23\ldots n} & r_{12}\sigma_1\sigma_2 & \ldots & r_{1n}\sigma_1\sigma_n \\ r_{12}\sigma_1\sigma_2 & \sigma_2^2 & \ldots & r_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & & \vdots \\ r_{1n}\sigma_1\sigma_n & r_{2n}\sigma_2\sigma_n & \ldots & \sigma_n^2 \end{vmatrix} = 0$$

Dividing $R_1, R_2, \ldots, R_n$, by $\sigma_1, \sigma_2, \ldots, \sigma_n$ respectively and also $C_1, C_2, \ldots, C_n$ by $\sigma_1, \sigma_2, \ldots, \sigma_n$ respectively, we get

$$\begin{vmatrix} 1 - \frac{\sigma^2_{1\cdot23\ldots n}}{\sigma_1^2} & r_{12} & \ldots & r_{1n} \\ r_{12} & 1 & \ldots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{1n} & r_{2n} & \ldots & 1 \end{vmatrix} = 0$$

$$\Rightarrow \quad \begin{vmatrix} 1 - \frac{\sigma^2_{1\cdot23\ldots n}}{\sigma_1^2} & r_{12} & \ldots & r_{1n} \\ r_{12} + 0 & 1 & \ldots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{1n} + 0 & r_{2n} & \ldots & 1 \end{vmatrix} = 0$$

$$\begin{vmatrix} 1 & r_{12} & \dots & r_{1n} \\ r_{12} & 1 & \dots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{1n} & r_{2n} & \dots & 1 \end{vmatrix} - \begin{vmatrix} \frac{\sigma^2_{1\cdot23\dots n}}{\sigma_1^2} & r_{12} & \dots & r_{1n} \\ 0 & 1 & \dots & r_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & r_{2n} & \dots & 1 \end{vmatrix} = 0$$

$$\Rightarrow \quad \omega - \frac{\sigma^2_{1\cdot23\dots n}}{\sigma_1^2}\,\omega_{11} = 0$$

$$\therefore \quad \sigma^2_{1\cdot23\dots n} = \sigma_1^2 \frac{\omega}{\omega_{11}} \qquad \dots(10\cdot43)$$

Remark. In a tri-variate distribution,

$$\sigma_{1\cdot23}^2 = \sigma_1^2 \frac{\omega}{\omega_{11}} \qquad \dots(10\cdot43a)$$

where ω and ω_{11} are defined in (10·32).

10·14. Coefficient of Multiple Correlation. In a tri-variate distribution in which each of the variables X_1, X_2, and X_3 has N observations, the multiple correlation coefficient of X_1 on X_2 and X_3, usually denoted by $R_{1\cdot23}$, is the simple correlation coefficient between X_1 and the joint effect of X_2 and X_3 on X_1. In other words $R_{1\cdot23}$ is the correlation coefficient between X_1 and its estimated value as given by the plane of regression of X_1 on X_2 and X_3 *viz.*,

$$e_{1\cdot23} = b_{12\cdot3}X_2 + b_{13\cdot2}X_3$$

We have $X_{1\cdot23} = X_1 - b_{12\cdot3}X_2 - b_{13\cdot2}X_3 = X_1 - e_{1\cdot23}$

$$\Rightarrow \quad e_{1\cdot23} = X_1 - X_{1\cdot23}$$

Since X_i's are measured from their respective means, we have

$$E(X_{1\cdot23}) = 0 \text{ and } E(e_{1\cdot23}) = 0 \qquad (\because E(X_i) = 0;\ i = 1, 2, 3)$$

By def.,

$$R_{1\cdot23} = \frac{\text{Cov}\,(X_1, e_{1\cdot23})}{\sqrt{V(X_1)\,V(e_{1\cdot23})}}. \qquad \dots(10\cdot44)$$

$$\text{Cov}\,(X_1, e_{1\cdot23}) = E[\{X_1 - E(X_1)\}\{e_{1\cdot23} - E(e_{1\cdot23})\}] = E(X_1\, e_{1\cdot23})$$

$$= \frac{1}{N}\Sigma X_1\, e_{1\cdot23} = \frac{1}{N}\Sigma X_1\,(X_1 - X_{1\cdot23})$$

$$= \frac{1}{N}\Sigma X_1^2 - \frac{1}{N}\Sigma X_1 X_{1\cdot23} = \frac{1}{N}\Sigma X_1^2 - \frac{1}{N}\Sigma X^2_{1\cdot23}$$

$$= \sigma_1^2 - \sigma_{1\cdot23}^2 \qquad (c.f.\ \text{Property 2, § 10·13})$$

Also $V(e_{1\,23}) = E(e_{1\cdot23}^2) = \frac{1}{N}\Sigma e_{1\cdot23}^2 = \frac{1}{N}\Sigma\,(X_1 - X_{1\cdot23})^2$

$$= \frac{1}{N}\Sigma\,(X_1^2 + X_{1\cdot23}^2 - 2\,X_1 X_{1\cdot23})$$

$$= \frac{1}{N}\Sigma X_1^2 + \frac{1}{N}\Sigma X_{1\cdot23}^2 - \frac{2}{N}\Sigma X_1 X_{1\cdot23}$$

$$= \frac{1}{N}\Sigma X_1^2 + \frac{1}{N}\Sigma X_{1\cdot 23}^2 - \frac{2}{N}\Sigma X_{1\cdot 23}^2$$

$$= \sigma_1^2 - \sigma_{1\cdot 23}^2 \qquad (c.f.\ \text{Property 2, § 10·13})$$

$$\therefore \qquad R_{1\cdot 23} = \frac{\sigma_1^2 - \sigma_{1\cdot 23}^2}{\sqrt{\sigma_1^2(\sigma_1^2 - \sigma_{1\cdot 23}^2)}}$$

$$\Rightarrow \qquad R^2_{1\cdot 23} = \frac{\sigma_1^2 - \sigma_{1\cdot 23}^2}{\sigma_1^2} = 1 - \frac{\sigma_{1\cdot 23}^2}{\sigma_1^2}$$

$$\Rightarrow \qquad 1 - R^2_{1\cdot 23} = \frac{\sigma_{1\cdot 23}^2}{\sigma_1^2}$$

Using (10·43*a*), we get

$$1 - R^2_{1\cdot 23} = \frac{\omega}{\omega_{11}} \qquad \ldots(10\cdot 45)$$

where

$$\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23} \text{ (On simplification)}.$$

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

Hence from (10·45), we get

$$R^2_{1\cdot 23} = 1 - \frac{\omega}{\omega_{11}} = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}\,r_{13}\,r_{23}}{1 - r_{23}^2} \qquad \ldots(10\cdot 45a)$$

This formula expresses the multiple correlation coefficient in terms of the total correlation coefficients between the pairs of variables.

Generalisation. In case of n-variate distribution, the multiple correlation coefficient of X_1 on $X_2, X_3, \ldots, X_n$, usually denoted by $R_{1\cdot 23\ldots n}$, is the correlation coefficient between X_1 and

$$e_{1\cdot 23\ldots n} = X_1 - X_{1\cdot 23\ldots n}$$

$$\therefore \qquad R_{1\cdot 23\ldots n} = \frac{\text{Cov}\,(X_1, e_{1\cdot 23\ldots n})}{\sqrt{V(X_1)\,V(e_{1\cdot 23\ldots n})}}$$

$$\text{Cov}\,(X_1, e_{1\cdot 23\ldots n}) = \frac{1}{N}\Sigma X_1\, e_{1\cdot 23\ldots n} = \frac{1}{N}\Sigma X_1(X_1 - X_{1\cdot 23\ldots n})$$

$$= \frac{1}{N}\Sigma X_1^2 - \frac{1}{N}\Sigma X_1 X_{1\cdot 23\ldots n}$$

$$= \frac{1}{N}\Sigma X_1^2 - \frac{1}{N}\Sigma X^2_{1\cdot 23\ldots n} = \sigma_1^2 - \sigma^2_{1\cdot 23\ldots n} \qquad \ldots(*)$$

$$V(e_{1\cdot 23\ldots n}) = \frac{1}{N}\Sigma e^2_{1\cdot 23\ldots n} = \frac{1}{N}\Sigma (X_1 - X_{1\cdot 23\ldots n})^2$$

$$= \frac{1}{N}\Sigma\,(X_1^2 + X^2_{1\cdot 23\ldots n} - 2X_1X_{1\cdot 23\ldots n})$$

$$= \frac{1}{N}\Sigma X_1^2 + \frac{1}{N}\Sigma\, X^2_{1\cdot 23\ldots n} - 2\,\frac{1}{N}\,\Sigma X_1 X_{1\cdot 23\ldots n}$$

$$= \frac{1}{N}\Sigma\, X_1^2 + \frac{1}{N}\Sigma\, X^2_{1\cdot 23\ldots n} - \frac{2}{N}\,\Sigma X^2_{1\cdot 23\ldots n}$$

$$= \sigma_1^2 - \sigma^2_{1\cdot 23\ldots n}$$

$$\therefore \quad R_{1\cdot 23\ldots n} = \frac{\sigma_1^2 - \sigma^2_{1\cdot 23\ldots n}}{\sqrt{\sigma_1^2(\sigma_1^2 - \sigma^2_{1\cdot 23\ldots n})}} = \left(\frac{\sigma_1^2 - \sigma^2_{1\cdot 23\ldots n}}{\sigma_1^2}\right)^{1/2}$$

$$R^2_{1\cdot 23\ldots n} = 1 - \frac{\sigma^2_{1\cdot 23\ldots n}}{\sigma_1^2} = 1 - \frac{\omega}{\omega_{11}} \qquad \ldots(10\text{·}45c)$$

where ω and ω_{11} are defined in (10·38).

Remarks 1. It may be pointed out here that multiple correlation coefficient can never be negative, because from (*) and (**), we get

$$\text{Cov}\,(X_1, e_{1\cdot 23\ldots n}) = \sigma_1^2 - \sigma^2_{1\cdot 23\ldots n} = \text{Var}\,(e_{1\cdot 23\ldots n}) \geq 0$$

Since the sign of $R_{1\cdot 23\ldots n}$ depends upon the covariance term $\text{Cov}\,(X_1, e_{1\cdot 23\ldots n})$, we conclude that $R_{1\cdot 23\ldots n} \geq 0$.

2. Since $R^2_{1\cdot 23\ldots n} \geq 0$, we have :

$$1 - \frac{\omega}{\omega_{11}} \geq 0 \quad \Rightarrow \quad \omega \leq \omega_{11} \qquad \ldots(10\text{·}45d)$$

3. Also, $R^2_{1\cdot 23\ldots n} \leq 1 \quad \Rightarrow \quad 1 - \frac{\omega}{\omega_{11}} \leq 1$

$$\Rightarrow \qquad 0 \leq \frac{\omega}{\omega_{11}} \quad \Rightarrow \quad \frac{\omega}{\omega_{11}} \geq 0 \quad \Rightarrow \quad \omega \geq 0 \qquad \ldots(10\text{·}45e)$$

From the above results, we get

$$\omega_{11} \geq \omega \geq 0 \qquad \ldots(10\text{·}45f)$$

In general, we have

$$\omega_{ii} \geq 0\,;\ i = 1, 2, \ldots, n$$

4. Since ω is symmetric in r_{ij}'s, we have

$$\omega_{ij} = \omega_{ji}\,;\ i \neq j = 1, 2, \ldots, n \qquad \ldots(10\text{·}45g)$$

10·14·1. Properties of Multiple Correlation Coefficient

1. Multiple correlation co-efficient measures the closeness of the association between the observed values and the expected values of a variable obtained from the multiple linear regression of that variable on other variables.

2. Multiple correlation coefficient between observed values and expected values, when the expected values are calculated from a linear relation of the variables determined by the method of least squares, is always greater than that where expected values are calculated from any other linear combination of the variables.

3. Since $R_{1\cdot23}$ is the simple correlation between X_1 and $e_{1\cdot23}$, it must lie between -1 and $+1$. But as seen in Remark 1 above, $R_{1\cdot23}$ is a non-negative quantity and we conclude that $0 \leq R_{1\cdot23} \leq 1$.

4. If $R_{1\cdot23} = 1$, then association is perfect and all the regression residuals are zero, and as such $\sigma^2_{1\cdot23} = 0$. In ths case, since $X_1 = e_{1\cdot23}$, the predicted value of X_1, the multiple linear regression equation of X_1 on X_2 and X_3 may be said to be a perfect prediction formula.

5. If $R_{1\cdot23} = 0$, then all total and partial correlations involving X_1 are zero. [See Example 10·37). So X_1 is completely uncorrelated with all the other variables in this case and the multiple regression equation fails to throw any light on the value of X_1 when X_2 and X_3 are known.

6. $R_{1\cdot23}$ is not less than any total correlation coefficient, *i.e.*,

$$R_{1\cdot23} \geq r_{12}, r_{13}, r_{23}$$

10·15. Coefficient of Partial Correlation. Sometimes the correlation between two variables X_1 and X_2 may be partly due to the correlation of a third variable, X_3 with both X_1 and X_2. In such a situation, one may want to know what the correlation between X_1 and X_2 would be if the effect of X_3 on each of X_1 and X_2 were eliminated. This correlation is called the *partial correlation* and the correlation coefficient between X_1 and X_2 after the linear effect of X_3 on each of them has been eliminated is called the *partial correlation coefficient.*

The residual $X_{1\cdot3} = X_1 - b_{13}X_3$, may be regarded as that part of the variable X_1 which remains after the linear effect of X_3 has been eliminated. Similarly, the residual $X_{2\cdot3}$ may be interpreted as the part of the variable X_2 obtained after eliminating the linear effect of X_3. Thus the partial correlation coefficient between X_1 and X_2, usually denoted by $r_{12\cdot3}$, is given by

$$r_{12\cdot3} = \frac{\text{Cov}\,(X_{1\cdot3}, X_{2\cdot3})}{\sqrt{\text{Var}\,(X_{1\cdot3})\,\text{Var}\,(X_{2\cdot3})}} \qquad \ldots(10\cdot46)$$

We have

$$\text{Cov}(X_{1\cdot3}, X_{2\cdot3}) = \frac{1}{N}\sum X_{1\cdot3}X_{2\cdot3} = \frac{1}{N}\sum X_1 X_{2\cdot3}$$

$$= \frac{1}{N}\sum X_1(X_2 - b_{23}X_3) = \frac{1}{N}\sum X_1X_2 - b_{23}\frac{1}{N}\sum X_1X_3$$

$$= r_{12}\sigma_1\sigma_2 - r_{23}\frac{\sigma_2}{\sigma_3}\cdot(r_{13}\sigma_1\sigma_3)$$

$$= \sigma_1\sigma_2(r_{12} - r_{13}r_{23})$$

$$\text{Also}\quad V(X_{1\cdot3}) = \frac{1}{N}\sum X_{1\cdot3}^2 = \frac{1}{N}\sum X_{1\cdot3}X_{1\,3}$$

$$= \frac{1}{N}\sum X_1X_{1\cdot3} = \frac{1}{N}\sum X_1(X_1 - b_{13}X_3)$$

$$= \frac{1}{N}\Sigma X_1^2 - b_{13} \cdot \frac{1}{N}\Sigma X_1 X_3$$

$$= \sigma_1^2 - r_{13}\frac{\sigma_1}{\sigma_3} r_{13}\sigma_1\sigma_3$$

$$= \sigma_1^2(1 - r_{13}^2)$$

Similarly, we shall get

$$V(X_{2\cdot3}) = \sigma_2^2(1 - r_{23}^2)$$

Hence

$$r_{12\cdot3} = \frac{\sigma_1\sigma_2(r_{12} - r_{13}\, r_{23})}{\sqrt{\sigma_1^2(1 - r_{13}^2)\,\sigma_2^2(1 - r_{23}^2)}} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)\,(1 - r_{23}^2)}} \quad \ldots(10\cdot46a)$$

Aliter. We have

$$0 = \Sigma X_{2\cdot3} X_{1\cdot23}$$

$$= \Sigma X_{2\cdot3}\,(X_1 - b_{12\cdot3} X_2 - b_{13\cdot2} X_3)$$

$$= \Sigma X_1 X_{2\cdot3} - b_{12\cdot3}\Sigma X_{2\cdot3} X_2 - b_{13\cdot2}\,\Sigma X_{2\cdot3} X_3$$

$$= \Sigma X_{1\cdot3} X_{2\cdot3} - b_{12\cdot3}\Sigma X_{2\cdot3} X_{2\cdot3}$$

$$\therefore \qquad b_{12\cdot3} = \frac{\Sigma X_{1\cdot3} X_{2\cdot3}}{\Sigma X_{2\cdot3}^2}$$

From this it follows that $b_{12\cdot3}$ is coefficient of regression of $X_{1\cdot3}$ on $X_{2\cdot3}$.

Similarly, $b_{21\cdot3}$ is the coefficient of regression of $X_{2\cdot3}$ on $X_{1\cdot3}$.

Since correlation coefficient is the geometric mean between regression coefficients, we have

$$r^2_{12\cdot3} = b_{12\cdot3} \times b_{21\cdot3}$$

But by def.,

$$b_{12\cdot3} = -\frac{\sigma_1}{\sigma_2} \cdot \frac{\omega_{12}}{\omega_{11}} \quad \text{and} \quad b_{21\cdot3} = -\frac{\sigma_2}{\sigma_1} \cdot \frac{\omega_{21}}{\omega_{22}}$$

$$\therefore \qquad r^2_{12\cdot3} = \left(-\frac{\sigma_1}{\sigma_2}\,\frac{\omega_{12}}{\omega_{11}}\right)\left(-\frac{\sigma_2}{\sigma_1} \cdot \frac{\omega_{21}}{\omega_{22}}\right) = \frac{\omega_{12}^2}{\omega_{11}\,\omega_{22}}$$

$$(\because \omega_{12} = \omega_{21})$$

$$\Rightarrow \qquad r_{12\cdot3} = -\frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}},$$

the negative sign being taken since the sign of regression coefficients is the same as that of $(-\omega_{12})$.

Substituting the values of ω_{12}, ω_{11} and ω_{22} from (10·32), we get

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

Remarks 1. The expressions for $r_{13\cdot2}$ and $r_{23\cdot1}$ can be similarly obtained, to give

$$r_{13\cdot2} = \frac{r_{13} - r_{12}\,r_{32}}{\sqrt{(1 - r_{12}^2)\,(1 - r_{32}^2)}} \quad \text{and} \quad r_{23\cdot1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}}$$

2. If $r_{12\cdot3} = 0$, we have then $r_{12} = r_{13}\,r_{23}$, it means that r_{12} will not be zero if X_3 is correlated with both X_1 and X_2. Thus, although X_1 and X_2 may be uncorrelated when effect of X_3 is eliminated, yet X_1 and X_2 may appear to be correlated because they carry the effect of X_3 on them.

3. Partial correlation coefficient helps in deciding whether to include or not an additional independent variable in regression analysis.

4. We know that $\sigma_1^2(1 - r_{12}^2)$ and $\sigma_1^2(1 - r_{13}^2)$ are the residual variances if X_1 is estimated from X_2 and X_3 individually, while $\sigma_1^2\,(1 - R_{1\cdot23}^2)$ is the residual variance if X_1 is estimated from X_2 and X_3 taken together. So from the above remark and $R_{1\cdot23}^2 \geq r_{12}^2$ and r_{13}^2, it follows that inclusion of an additional variable can only reduce the residual variance. Now inclusion of X_3 when X_2 has already been taken for predicting X_1, is worthwhile only when the resultant reduction in the residual variance is substantial. This will be the case when $r_{13\cdot2}$ is sufficiently large. Thus in this respect partial correlation coefficient has its significance in regression analysis.

10·15·1. Generalisation. In the case of n variables $X_1, X_2 \ldots, X_n$ the partial correlation coefficient $r_{12\cdot34\ldots n}$ between X_1 and X_2 (after the linear effect of $X_3, X_4, \ldots, X_n$ on them has been eliminated), is given by

$$r^2_{12\cdot34\ldots n} = b_{12\cdot34\ldots n} \times b_{21\cdot34\ldots n}$$

But, we have

and
$$\left.\begin{aligned} b_{12\cdot34\ldots n} &= -\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}} \\ b_{21\cdot34\ldots n} &= -\frac{\sigma_2}{\sigma_1}\cdot\frac{\omega_{21}}{\omega_{22}} \end{aligned}\right\} \text{[}c.f.\text{ Equation (10·40)]}$$

$$\therefore \quad r^2_{12\cdot34\ldots n} = \left(-\frac{\sigma_1}{\sigma_2}\cdot\frac{\omega_{12}}{\omega_{11}}\right)\left(-\frac{\sigma_2}{\sigma_1}\cdot\frac{\omega_{21}}{\omega_{22}}\right) = \frac{\omega_{12}^2}{\omega_{11}\omega_{22}}$$

$$\Rightarrow \quad r_{12\cdot34\ldots n} = -\frac{\omega_{12}}{\sqrt{\omega_{11}\omega_{22}}} \qquad (10\text{·}46b)$$

negative sign being taken since the sign of the regression coefficient is same as that of $(-\omega_{12})$.

10·16. Multiple Correlation in Terms of Total and Partial Correlations.

$$1 - R_{1\cdot23}^2 = (1 - r_{12}^2)(1 - r_{13\cdot2}^2) \qquad \ldots(10\text{·}46c)$$

Proof. We have

$$1 - R_{1\cdot23}^2 = 1 - \frac{r_{12}^2 + r_{13}^2 - 2r_{12}\,r_{13}\,r_{23}}{1 - r_{23}^2}$$

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

Also

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

Hence the result.

Theorem. *Any standard deviation of order 'p' may be expressed in terms of a standard deviation of order (p – 1) and a partial correlation coefficient of order (p – 1).*

Proof. Let us consider the sum :

$$\Sigma X^2_{1\cdot 23\,\ldots\, n} = \Sigma X_{1\cdot 23\ldots n} X_{1\cdot 23\ldots n}$$

$$= \Sigma[X_{1\,23\ \ldots\,(n-1)} X_{1\cdot 23\ldots n}$$

(*c.f.* Property 2, § 10·13)

$$= \Sigma[X_{1\cdot 23\ldots(n-1)} (X_1 - b_{12\cdot 34\ldots n} X_2 - \ldots - b_{1(n-1)\cdot 23\ldots n} X_{n-1} - b_{1n\cdot 23\ldots(n-1)} X_n)]$$

$$= \Sigma X_{1\cdot 23\ldots(n-1)} X_1 - b_{1n\cdot 23\ldots(n-1)} \Sigma X_{1\cdot 23\ldots(n-1)} X_n$$

(*c.f.* Property 2 § 10·13)

$$= \Sigma X^2_{1\cdot 23\ldots(n-1)} - b_{1n\,23\ldots(n-1)} \Sigma X_{1\cdot 23\ldots(n-1)} X_{n\cdot 23\ldots(n-1)}$$

Dividing both sides by N (total number of observations), we get $\sigma^2_{1\cdot 23\ldots n} = \sigma^2_{1\,23\ldots(n-1)} - b_{1n\cdot 23\ \ldots(n-1)}\ \mathrm{Cov}\ (X_{1\,23\ldots(n-1)}, X_{n\cdot 23\ldots(n-1)})$

The regression coefficient of $X_{n\,23\ldots(n-1)}$ on $X_{1\cdot 23\ldots(n-1)}$ is given by

$$b_{n1\cdot 23\ldots(n-1)} = \frac{\mathrm{Cov}\ (X_{1\cdot 23\ldots(n-1)}, X_{n\cdot 23\ldots(n-1)})}{\sigma^2_{1\cdot 23\ldots(n-1)}}$$

$$\therefore \quad \sigma^2_{1\cdot 23\ldots n} = \sigma^2_{1\cdot 23\ldots(n-1)} [1 - b_{1n\cdot 23\ldots(n-1)} b_{n1\cdot 23\ldots(n-1)}]$$

$$= \sigma^2_{1\cdot 23\ldots(n-1)} [1 - r^2_{n1\,23\ldots(n-1)}], \qquad \ldots(10\cdot 47)$$

a formula which expresses the standard deviation of order $(n - 1)$ in terms of standard deviation of order $(n - 2)$ and partial correlation coefficient of order $(n - 2)$. If we take $p = (n - 1)$, the theorem is established.

Cor. 1. From (10·47), we have

$$\sigma^2_{1\,23\ldots(n-1)} = \sigma^2_{1\cdot 23\ldots(n-2)} (1 - r^2_{1(n-)\cdot 23\ldots(n-2)}) \qquad \ldots(10\cdot 47a)$$

and so on. Thus the repeated application of (10·47) gives

$$\sigma^2_{1\cdot 23\ \ldots n} = \sigma_1{}^2(1 - r_{12}{}^2)\,(1 - r_{13\cdot 2}{}^2)\,(1 - r_{14\cdot 32}{}^2)\ldots(1 - r^2_{1n\cdot 23\ldots(n-1)}) \qquad \ldots(10\cdot 47b)$$

Since partial correlation coefficients cannot exceed unity numerically, we get from (10·47), (10·47*a*), and so on,

$$\left.\begin{array}{lcl} \sigma^2_{1\cdot23\ldots n} & \leq & \sigma^2_{1\cdot23\ldots(n-1)} \\ \sigma^2_{1\cdot23\ldots(n-1)} & \leq & \sigma^2_{1\cdot23\ldots(n-2)} \\ \vdots & & \vdots \\ \sigma_{1\cdot23}{}^2 & \leq & \sigma_{1\cdot2}{}^2 \\ \sigma_{1\cdot2}{}^2 & \leq & \sigma_1{}^2 \end{array}\right\}$$

$$\Rightarrow \qquad \sigma_1 \geq \sigma_{1\cdot2} \geq \sigma_{1\cdot23} \geq \ldots \geq \sigma_{1\cdot23\ldots n} \qquad \ldots(10\text{·}47c)$$

Cor. 2. Also, we have

$$\sigma^2_{1\cdot23\ldots n} = \sigma_1{}^2(1 - R^2_{1\cdot23\ldots n})$$

On using (10·47*b*), we get

$$1 - R^2_{1\cdot23\ldots n} = (1 - r_{12}{}^2)(1 - r_{13\cdot2}{}^2)\ldots(1 - r^2_{1n\cdot3\ldots(n-1)}) \qquad \ldots(10\text{·}47d)$$

This is the generalisation of the result obtained in (10·46*c*).

Since $\quad |r_{ij.(s)}| \leq 1;\ s = 0, 1, 2, \ldots, (n-1),$

where $r_{ij.(s)}$ is a partial correlation coefficient of order s. we get from (10·47*d*)

$$1 - R^2_{1\cdot23\ldots n} \leq 1 - r_{12}{}^2$$
$$1 - R^2_{1\cdot23\ldots n} \leq 1 - r^2_{13\cdot2},$$

and so on.

i.e., $$R^2_{1\cdot23\ldots n} \geq r_{12}{}^2, r^2_{13\cdot2}, \ldots, r^2_{1n\cdot23\ldots(n-1)} \qquad \ldots(10\text{·}47e)$$

Since $R_{1\cdot23\ldots n}$ is symmetric in its secondary subscripts, we have

$$\left.\begin{array}{l} R^2_{1\cdot23\ldots n} \geq r_{1i}{}^2,\ (i = 2, 3, \ldots, n) \\ R^2_{1\cdot23\ldots n} \geq r_{1i.j}\ (i \neq j = 2, 3, \ldots, n) \end{array}\right\} \qquad \ldots(10\text{·}47f)$$

and so on

10·17. Expression for Regression Coefficients in Terms of Regression Coefficients of Lower Order. Consider

$$\begin{aligned} \Sigma X_{1\cdot34\ldots n} X_{2\cdot34\ldots n} &= \Sigma X_{1\cdot34\ldots(n-1)} X_{2\cdot34\ldots n} \\ &= \Sigma X_{1\cdot34\ldots(n-1)}(X_2 - b_{23\cdot4\ldots n}X_3 - \ldots - b_{2n\cdot34\ldots(n-1)}X_n) \\ &= \Sigma X_{1\cdot34\ldots(n-1)} X_2 - b_{2n\cdot34\ldots(n-1)} \Sigma X_{1\cdot34\ldots(n-1)} X_n \\ &= \Sigma X_{1\cdot34\ldots(n-1)} X_{2\cdot34\ldots(n-1)} \\ &\qquad - b_{2n\cdot34\ldots(n-1)} \Sigma X_{1\cdot34\ldots(n-1)} X_{n\cdot34\ldots(n-1)} \end{aligned}$$

Dividing both sides by N, the total number of observations, we get

$$\begin{aligned} \text{Cov}\,(X_{1\cdot34\ldots n}, X_{2\cdot34\ldots n}) &= \text{Cov}\,(X_{1\cdot34\ldots(n-1)}, X_{2\cdot34\ldots(n-1)}) \\ &\qquad - b_{2n\cdot34\ldots(n-1)}\,\text{Cov}\,(X_{1\cdot34\ldots(n-1)}, X_{n\cdot34\ldots(n-1)}) \\ b_{12\cdot34\ldots n}\,\sigma^2_{2\cdot34\ldots n} &= b_{12\cdot34\ldots(n-1)}\,\sigma^2_{2\cdot34\ldots(n-1)} \\ &\qquad - b_{2n\cdot34\ldots(n-1)}\, b_{1n\cdot34\ldots(n-1)}\,\sigma^2_{n\cdot34\ldots(n-1)} \end{aligned}$$

On using (10·47), we get

$$b_{12\cdot34\ldots n}\,\sigma^2_{2\cdot34\ldots(n-1)}\,\{1-r^2_{2n\cdot34\ldots(n-1)}\}$$
$$=\sigma^2_{2\cdot34\ldots(n-1)}\left[b_{12\cdot34\ldots(n-1)}-b_{2n\cdot34\ldots(n-1)}\,b_{1n\cdot34\ldots(n-1)}\times\frac{\sigma^2_{n\cdot34\ldots(n-1)}}{\sigma^2_{2\cdot34\ldots(n-1)}}\right]\quad\ldots(*)$$

In the case of two variables, we have

$$b_{ij}\,\sigma_j^2=b_{ji}\,\sigma_i^2=\text{Cov}\,(X_i,X_j)\quad\Rightarrow\quad b_{ij}=\frac{\sigma_i^2}{\sigma_j^2}\,b_{ji}$$

$$\therefore\qquad b_{2n\cdot34\ldots(n-1)}\;\frac{\sigma^2_{n\cdot34\ldots(n-1)}}{\sigma^2_{2\cdot34\ldots(n-1)}}=b_{n2\cdot34\ldots(n-1)}$$

Hence from (*), we get

$$b_{12\cdot34\ldots n}\,\sigma^2_{2\cdot34\ldots(n-1)}\,\{1-r^2_{2n\cdot34\ldots(n-1)}\}$$
$$=\sigma^2_{2\cdot34\ldots(n-1)}\,[b_{12\cdot34\ldots(n-1)}-b_{1n\cdot34\ldots(n-1)}\,b_{n2\cdot34\ldots(n-1)}]$$

$$\therefore\quad b_{12\cdot34\ldots n}=\left[\frac{b_{12\cdot34\ldots(n-1)}-b_{1n\cdot34\ldots(n-1)}\,b_{n2\cdot34\ldots(n-1)}}{1-r^2_{2n\cdot34\ldots(n-1)}}\right]\qquad\ldots(10\text{·}48)$$

$$\Rightarrow\quad b_{12\cdot34\ldots n}=\left[\frac{b_{12\cdot34\ldots(n-1)}-b_{1n\cdot34\ldots(n-1)}\cdot b_{n2\cdot34\ldots(n-1)}}{1-b_{2n\cdot34\ldots(n-1)}\cdot b_{n2\cdot34\ldots(n-1)}}\right]\qquad\ldots(10\text{·}48a)$$

10·18. Expression for Partial Correlation Coefficient in Terms of Correlation Coefficients of Lower Order. By definition, we have

$$b_{ij.lm\ldots t}=r_{ij.lm\ldots t}\times\frac{\sigma_{i.lm\ldots t}}{\sigma_{j.lm\ldots t}}\qquad\ldots(*)$$

$$\therefore\qquad b_{1n\cdot34\ldots(n-1)}\cdot b_{n2\cdot34\ldots(n-1)}$$
$$=r_{1n\cdot34\ldots(n-1)}\frac{\sigma_{1\cdot34\ldots(n-1)}}{\sigma_{n\cdot34\ldots(n-1)}}\times r_{n2\cdot34\ldots(n-1)}\frac{\sigma_{n\cdot34\ldots(n-1)}}{\sigma_{2\cdot34\ldots(n-1)}}$$
$$=r_{1n\cdot34\ldots(n-1)}\cdot r_{n2\cdot34\ldots(n-1)}\cdot\frac{\sigma_{1.34\ldots(n-1)}}{\sigma_{2\cdot34\ldots(n-1)}}\qquad\ldots(**)$$

Hence from (10·48), on using (*) and (**), we get

$$r_{12\cdot34\ldots n}\times\frac{\sigma_{1\cdot34\ldots n}}{\sigma_{2\cdot34\ldots n}}$$
$$=\left[\frac{\{r_{12\cdot34\ldots(n-1)}-r_{1n\cdot34\ldots(n-1)}\,r_{n2\cdot34\ldots(n-1)}\}}{1-r^2_{2n\cdot34\ldots(n-1)}}\;\;\frac{\sigma_{1\cdot34\ldots(n-1)}}{\sigma_{2\cdot34\ldots(n-1)}}\right]\qquad\ldots(***)$$

Also on using (10·47), we get

$$\frac{\sigma_{1\cdot34\ldots n}}{\sigma_{2\cdot34\ldots n}}=\frac{\sigma_{1\cdot34\ldots(n-1)}}{\sigma_{2\cdot34\ldots(n-1)}}\times\left[\frac{1-r^2_{1n\cdot34\ldots(n-1)}}{1-r^2_{2n\cdot34\ldots(n-1)}}\right]^{1/2}$$

Hence from (***), we get

$$r_{12\cdot34\ldots n}\left[\frac{1-r^2_{1n\cdot34\ldots(n-1)}}{1-r^2_{2n\cdot34\ldots(n-1)}}\right]^{\frac{1}{2}}$$

$$=\left[\frac{r_{12\cdot34\ldots(n-1)}-r_{1n\cdot34\ldots(n-1)}\,r_{n2\cdot34\ldots(n-1)}}{1-r^2_{2n\cdot34\ldots(n-1)}}\right]$$

$$\Rightarrow \quad r_{12\cdot34\ldots n}=\frac{r_{12\cdot34\ldots(n-1)}-r_{1n\cdot34\ldots(n-1)}\,r_{n2\cdot34\ldots(n-1)}}{(1-r^2_{1n\cdot34\ldots(n-1)})^{1/2}\,(1-r^2_{n2\cdot34\ldots(n-1)})^{1/2}} \quad \ldots(10\text{-}49)$$

which is an expression for the correlation coefficient of order $p=(n-2)$ in terms of the correlation coefficient of order $(p-1)=(n-3)$.

Example 10·33. *From the data relating to the yield of dry bark (X_1), height (X_2) and girth X_3 for 18 cinchona plants the following correlation coefficients were obtained :*

$$r_{12}=0\cdot77,\ r_{13}=0\cdot72 \text{ and } r_{23}=0\cdot52$$

Find the the partial correlation coefficient $r_{12\cdot3}$ and multiple correlation coefficient $R_{1\cdot23}$.

Solution.

$$r_{12\cdot3}=\frac{r_{12}-r_{13}\,r_{23}}{\sqrt{(1-r_{13}^2)\,(1-r_{23}^2)}}=\frac{0\cdot77-0\cdot72\times0\cdot52}{\sqrt{[1-(0\cdot72)^2][1-(0\cdot52)^2]}}=0\cdot62$$

$$R_{1\cdot23}^2=\frac{r_{12}^2+r_{13}^2-2r_{12}\,r_{13}\,r_{23}}{1-r_{23}^2}$$

$$=\frac{(0\cdot77)^2+(0\cdot72)^2-2\times0\cdot77\times0\cdot72\times0\cdot52}{1-(0\cdot52)^2}=0\cdot7334$$

$$\therefore \quad R_{1\cdot23}=+\,0\cdot8564$$

(since multiple correlation coefficient is non-negative).

Example 10·34. *In a trivariate distribution :*

$$\sigma_1=2,\ \sigma_2=\sigma_3=3,\ r_{12}=0\cdot7,\ r_{23}=r_{31}=0\cdot5.$$

Find (i) $r_{23\cdot1}$, (ii) $R_{1\cdot23}$, (iii) $b_{12\cdot3}$, $b_{13\cdot2}$, and (iv) $\sigma_{1\cdot23}$.

Solution. We have

(i) $$r_{23\cdot1}=\frac{r_{23}-r_{21}r_{31}}{\sqrt{(1-r_{21}^2)\,(1-r_{31}^2)}}=\frac{0\cdot5-(0\cdot7)(0\cdot5)}{\sqrt{(1-0\cdot49)(1-0\cdot25)}}=0\cdot2425$$

(ii) $$R_{1\cdot23}^2=\frac{r_{12}^2+r_{13}^2-2r_{12}\,r_{13}\,r_{23}}{1-r_{23}^2}$$

$$=\frac{0\cdot49+0\cdot25-2(0\cdot7)(0\cdot5)(0\cdot5)}{1-0\cdot25}=0\cdot52$$

$$\therefore \quad R_{1\cdot23}=+\,0\cdot7211$$

(iii) $$b_{12\cdot3}=r_{12\cdot3}\frac{\sigma_{1\cdot3}}{\sigma_{2\cdot3}} \quad \text{and} \quad b_{13\cdot2}=r_{13\cdot2}\frac{\sigma_{1\cdot2}}{\sigma_{3\cdot2}}$$

$$r_{12\cdot3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}=0\cdot6,\ r_{13\cdot2}=\frac{r_{13}-r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}}=0\cdot2425$$

$$\sigma_{1\cdot3} = \sigma_1\sqrt{(1-r_{13}^2)} = 2\sqrt{(1-0{\cdot}25)} = 1{\cdot}7320$$

$$\sigma_{2\cdot3} = \sigma_2\sqrt{(1-r_{23}^2)} = 3\sqrt{(1-0{\cdot}25)} = 2{\cdot}5980$$

$$\sigma_{1\cdot2} = \sigma_1\sqrt{(1-r_{12}^2)} = 2\sqrt{(1-0{\cdot}49)} = 1{\cdot}4282$$

$$\sigma_{3\cdot2} = \sigma_3\sqrt{(1-r_{32}^2)} = 3\sqrt{(1-0{\cdot}25)} = 2{\cdot}5980$$

Hence $b_{12\cdot3} = 0{\cdot}4$ and $b_{13\cdot2} = 0{\cdot}1333$

(*iv*) $\sigma_{1\cdot23} = \sigma_1\sqrt{\dfrac{\omega}{\omega_{11}}}$

where $\omega = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23} = 0{\cdot}36$

and $\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 = 1 - 0{\cdot}25 = 0{\cdot}75$

$\therefore \quad \sigma_{1\cdot23} = 2 \times \sqrt{0{\cdot}48} = 2 \times 0{\cdot}6928 = 1{\cdot}3856$

Example 10·35. *Find the regression equation of X_1 on X_2 and X_3 given the following results :—*

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

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

X_3 = *Accumulated temperature above 42°F.*

Solution. Regression equation of X_1 on X_2 and X_3 is given by

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

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

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

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

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

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

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

Example 10·36. *Five hundred students were examined in three subjects I, II and III, each subject carrying 100 marks. A student getting 120 or more but less than 150 marks was put in pass class. A student getting 150 or more but less than 180 marks was put in second class and a student getting 180 or more marks was put in the first class. The following marks were obtained :*

	I	*II*	*III*
Mean :	*35·8*	*52·4*	*48·8*
S.D. :	*4·2*	*5·3*	*6·1*
Correlation :	$r_{12} = 0·6$,	$r_{13} = 0·7$	$r_{23} = 0·8$

(i) Find the number of students in each of the three classes.

(ii) Find the total number of students with total marks lying between 120 and 190.

(iii) Find the probability that a student gets more that 240 marks.

(iv) What should be the correlation between marks in subjects I and II among students who scored equal marks in subject III ?

(v) If r_{23} was not known, obtain the limits within which it may lie from the values of r_{12} and r_{13} (ignoring sampling errors).

Solution. If Z denotes the total marks of the students in the three subjects and X_1, X_2, X_3 the total, marks of the students in subjects I, II and III respectively, then

$$Z = X_1 + X_2 + X_3$$

$$\therefore \quad E(Z) = E(X_1) + E(X_2) + E(X_3) = 35·8 + 52·4 + 48·8 = 137$$

$$V(Z) = V(X_1) + V(X_2) + V(X_3) + 2[\text{Cov}(X_1, X_2) + \text{Cov}(X_2, X_3) + \text{Cov}(X_3, X_1)]$$

$$= 17·64 + 28·09 + 37·21 + 26·712 + 35·868 + 51·728$$

$$= 197·248 \qquad [\text{Using Cov}(X_i, X_j) = r_{ij}\sigma_i\sigma_j]$$

$$\Rightarrow \quad \sigma_Z^2 = 197·248 \quad \text{or} \quad \sigma_Z = 14·045$$

Now $\xi = \dfrac{Z - E(Z)}{\sigma_Z} \sim N(0, 1)$

Z	$\xi = \dfrac{Z - 137}{14·045}$	$p = \displaystyle\int_{-\infty}^{\xi} p(\xi)d\xi$	*Class*	*Area under the curve in this class (A)*	*Frequency* $500 \times (A)$
120 –	1·21050	0·11314	120 – 150	0·70937	354·685
150	0·92567	0·82251	150 – 180	0·17639	88·195
180	3·06180	0·99890	180 –	0·00102	0·510
190	3·77400	0·99992	120 – 190	0·88678	443·390
240	7·33410	1·00000	240 –	0·00000	0·000

(i) The number of students in first, second and third class respectively are 355, 88 and 0 (approx.)

(ii) Total number of students with total marks between 120 and 190 is 443.

(iii) Probability that a student gets more than 240 marks is zero.

(iv) The correlation coefficient between marks in subjects I and II of the students who secured equal marks in subject III is $r_{12·3}$ and is given by

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\,r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = \frac{0{\cdot}04}{\sqrt{(1 - 0{\cdot}49)(1 - 0{\cdot}64)}} = 0{\cdot}0934$$

(*v*) We have.

$$r_{12\cdot3}^2 = \frac{(r_{12} - r_{13}\,r_{23})^2}{(1 - r_{13}^2)(1 - r_{23}^2)} \le 1$$

$$\therefore \quad \frac{(0{\cdot}6 - 0{\cdot}7a)^2}{(1 - 0{\cdot}49)(1 - a^2)} \le 1, \text{ where } a = r_{23}.$$

$$\Rightarrow \quad 0{\cdot}36 + 0{\cdot}49a^2 - 0{\cdot}84a \le 0{\cdot}51\,(1 - a^2)$$

$$\Rightarrow \quad a^2 - 0{\cdot}84a - 0{\cdot}15 \le 0$$

Thus 'a' lies between the roots of the equation :

$$a^2 - 0{\cdot}84a - 0{\cdot}15 = 0,$$

which are 0·99 and – 0·15.

Hence r_{23} should lie between – 0·15 and 0·99.

Example 10·37. *Show that*

$$1 - R_{1\cdot23}^2 = (1 - r_{12}^2)(1 - r_{13\cdot2}^2)$$

Deduce that

(*i*) $R_{1\cdot23} \ge r_{12}$, (*ii*) $R_{1\cdot23}^2 = r_{12}^2 + r_{13}^2$, *if* $r_{23} = 0$

(*iii*) $1 - R_{1\cdot23}^2 = \dfrac{(1 - \rho)(1 + 2\rho)}{(1 + \rho)}$, *provided all the coefficients of zero order are equal to* ρ.

(*iv*) *If* $R_{1\cdot23} = 0$, X_1 *is uncorrelated with any of other variables, i.e.,* $r_{12} = r_{13} = 0$. [*Delhi Univ. B.Sc. (Stat. Hons.), 1989*]

Solution. (*i*) Since $|\,r_{13\cdot2}\,| \le 1$, we have from (10·46*c*)

$$1 - R_{1\cdot23}^2 \le 1 - r_{12}^2 \quad \Rightarrow \quad R_{1\cdot23} \ge r_{12}$$

(*ii*) We have

$$r_{13\cdot2} = \frac{r_{13} - r_{12}\,r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = \frac{r_{13}}{\sqrt{1 - r_{12}^2}} \qquad (\text{if } r_{23} = 0)$$

$\therefore$ From (10·46*c*), we get

$$1 - R_{1\cdot23}^2 = (1 - r_{12}^2)\left[1 - \frac{r_{13}^2}{1 - r_{12}^2}\right] = 1 - r_{12}^2 - r_{13}^2$$

Hence $R_{1\cdot23}^2 = r_{12}^2 + r_{13}^2$, if $r_{23} = 0$.

(*iii*) Here, we are given that $r_{12} = r_{13} = r_{23} = \rho$

$$\therefore \quad r_{13\cdot2} = \frac{\rho - \rho^2}{\sqrt{(1 - \rho^2)(1 - \rho^2)}} = \frac{\rho(1 - \rho)}{(1 - \rho^2)} = \frac{\rho}{1 + \rho}$$

Hence from (10·46*c*), we have

$$1 - R_{1\cdot23}^2 = (1 - \rho^2)\left[1 - \frac{\rho^2}{(1 + \rho)^2}\right] = \frac{(1 - \rho)(1 + 2\rho)}{(1 + \rho)}$$

(*iv*) If $R_{1\cdot23} = 0$, (10·46*c*) gives

$$1 = (1 - r_{12}^2)(1 - r_{13\cdot2}^2) \qquad \ldots(*)$$

Since $0 \le r_{12}^2 \le 1$ and $0 \le r_{13\cdot2}^2 \le 1$, (*) will hold if and only if

$$r_{12} = 0 \quad \text{and} \quad r_{13\cdot2} = 0$$

Now $r_{13\cdot2} = 0 \Rightarrow \dfrac{r_{13} - r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{32}^2)}} = 0$

$$\Rightarrow \quad \frac{r_{13}}{\sqrt{1-r_{32}^2}} = 0 \qquad (\because r_{12} = 0)$$

$$\Rightarrow \quad r_{13} = 0$$

Thus if $R_{1\cdot23} = 0$, then $r_{13} = r_{12} = 0$, *i.e.*, X_1 is uncorrelated with X_2 and X_3.

Example 10·38. *Show that the correlation coefficient between the residuals $X_{1\cdot23}$ and $X_{2\cdot13}$ is equal and opposite to that between $X_{1\cdot3}$ and $X_{2\cdot3}$.*

[*Poona Univ. B.Sc., 1991*]

Solution. The correlation coefficient between $X_{1\cdot23}$ and $X_{2\cdot13}$ is given by

$$\frac{\text{Cov}\,(X_{1\cdot23}, X_{2\cdot13})}{\sigma_{1\cdot23}\,\sigma_{2\cdot13}} = \frac{\Sigma X_{1\cdot23}\,X_{2\cdot13}}{N\,\sigma_{1\cdot23}\,\sigma_{2\cdot13}} = \frac{\frac{1}{N}\Sigma X_{2\cdot13}(X_1 - b_{12\cdot3}X_2 - b_{13\cdot2}X_3)}{\sigma_{1\cdot23}\,\sigma_{2\cdot13}}$$

$$= -\,b_{12\cdot3}\frac{\Sigma X_{2\cdot13}\,X_2}{N\,\sigma_{1\cdot23}\,\sigma_{2\cdot13}} \qquad \text{(c.f. Property 1, § 10·13)}$$

$$= -\,b_{12\cdot3}\frac{\Sigma X_{2\cdot13}^2}{N\sigma_{1\cdot23}\,\sigma_{2\cdot13}} \qquad \text{(c.f. Property 2, § 10·13)}$$

$$= -\,b_{12\cdot3}\frac{\sigma_{2\cdot13}}{\sigma_{1\cdot23}} = -\,b_{12\cdot3}\frac{(\sigma_2\sqrt{\omega/\omega_{22}})}{(\sigma_1\sqrt{\omega/\omega_{11}})}$$

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

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 \quad \text{and} \quad \omega_{22} = \begin{vmatrix} 1 & r_{13} \\ r_{31} & 1 \end{vmatrix} = 1 - r_{13}^2$$

$$\therefore \; r(X_{1\cdot23}, X_{2\cdot13}) = -\,b_{12\cdot3}\frac{\sigma_2}{\sigma_1}\cdot\sqrt{\frac{1-r_{23}^2}{1-r_{13}^2}} = -\,b_{12\cdot3}\frac{\sigma_{2\cdot3}}{\sigma_{1\cdot3}}$$

[since $\sigma_{2\cdot3}^2 = \sigma_2^2(1-r_{23}^2)$ and $\sigma_{1\cdot3}^2 = \sigma_1^2(1-r_{13}^2)$]

$$\therefore \; r(X_{1\cdot23}, X_{2\cdot13}) = -\frac{\text{Cov}\,(X_{1\cdot3}, X_{2\cdot3})}{\sigma_{2\cdot3}^2}\cdot\frac{\sigma_{2\cdot3}}{\sigma_{1\cdot3}}$$

$$= -\frac{\text{Cov}\,(X_{1\cdot3}, X_{2\cdot3})}{\sigma_{2\cdot3}\,\sigma_{1\cdot3}} = -\,r(X_{1\cdot3}, X_{2\cdot3})$$

Hence the result.

Example 10·39. *Show that if $X_3 = aX_1 + bX_2$, the three partial correlations are numerically equal to unity, $r_{13\cdot2}$ having the sign of a, $r_{23\cdot1}$, the sign of b and $r_{12\cdot3}$, the opposite sign of a/b.* **[*Kanpur Univ. M.Sc., 1992*]**

Solution. Here we may regard X_3 as dependent on X_1 and X_2 which may be taken as independent variables. Since X_1 and X_2 are independent, they are uncorrelated.

Thus $r(X_1, X_2) = 0 \quad \Rightarrow \quad \text{Cov}\,(X_1, X_2) = 0$

$$V(X_3) = V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2) + 2ab\ \text{Cov}\,(X_1, X_2)$$
$$= a^2\sigma_1^2 + b^2\sigma_2^2,$$

where $V(X_1) = \sigma_1^2, V(X_2) = \sigma_2^2$.

Also $X_1X_3 = X_1(aX_1 + bX_2) = a X_1^2 + bX_1X_2$

Assuming that X_i's are measured from their means, on taking expectations of both sides, we get

$$\text{Cov}\,(X_1, X_3) = a\sigma_1^2 + b\ \text{Cov}\,(X_1, X_2) = a\sigma_1^2$$

$$\therefore \quad r_{13} = \frac{\text{Cov}\,(X_1, X_3)}{\sqrt{V(X_1)\,V(X_3)}} = \frac{a\,\sigma_1^2}{\sqrt{\sigma_1^2(a^2\sigma_1^2 + b^2\sigma_2^2)}} = \frac{a\sigma_1}{k},$$

where $k^2 = a^2\sigma_1^2 + b^2\sigma_2^2$.

Similarly, we will get

$$r_{23} = \frac{\text{Cov}\,(X_2, X_3)}{\sqrt{V(X_2)\,V(X_3)}} = \frac{b\sigma_2}{k}$$

Hence

$$r_{13\cdot2} = \frac{r_{13} - r_{12}\,r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} = \frac{a\sigma_1}{k}\,\frac{k}{\sqrt{k^2 - b^2\sigma_2^2}} = \frac{a\sigma_1}{\sqrt{a^2\sigma_1^2}} = \pm\frac{a\sigma_1}{a\sigma_1} = \pm 1$$

according as 'a' is positive or negative. Hence $r_{13\cdot2}$ has the same sign as 'a'.

Again

$$r_{23\cdot1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}} = \frac{b\sigma_2}{k}\,\frac{k}{\sqrt{k^2 - a^2\sigma_1^2}} = \frac{b\sigma_2}{\sqrt{b^2\sigma_2^2}} = \pm 1,$$

according as 'b' is positive or negative. Hence $r_{23\cdot1}$ has the same sign as 'b'.

Now

$$r_{12\cdot3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)\,(1 - r_{23}^2)}} = -\frac{a\sigma_1}{k}\cdot\frac{b\sigma_2}{k}\cdot\frac{k^2}{\sqrt{(k^2 - a^2\sigma_1^2)\,(k^2 - b^2\sigma_2^2)}}$$

$$= -\frac{ab\sigma_1\sigma_2}{\sqrt{b^2\sigma_2^2 \times a^2\sigma_1^2}} = -\frac{ab}{\sqrt{a^2b^2}} = -\frac{a/b}{\sqrt{a^2/b^2}} = \frac{-(a/b)}{\pm(a/b)} = \mp 1,$$

according as (a/b) is positive or negative. Hence $r_{12\cdot3}$ has the sign opposite to that of (a/b).

Example 10·40. *If all the correlation coefficients of zero order in a set of p-variates are equal to* ρ, *show that*

(*i*) *Every partial correlation of s'th order is* $\dfrac{\rho}{1 + s\rho}$ ...(*)

(*ii*) *The coefficient of multiple correlation R of a variate with the other* $(p - 1)$ *variates is given by*

$$1 - R^2 = (1 - \rho)\left[\frac{1 + (p - 1)\,\rho}{1 + (p - 2)\,\rho}\right]$$

[*Delhi Univ. M.Sc. (Maths'), 1990*]

Solution. We are given that

$$r_{mn} = \rho, (m, n = 1, 2, \ldots, p\ ;\ m \neq n)$$

We have

$$r_{ij\cdot k} = \frac{r_{ij} - r_{ik}\, r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}\ ,\ (i, j, k = 1, 2, \ldots, p;\ i \neq j \neq k)$$

$$= \frac{\rho - \rho \cdot \rho}{\sqrt{(1 - \rho^2)(1 - \rho^2)}} = \frac{\rho}{1 + \rho} \qquad \ldots(**)$$

Thus every partial correlation coefficient of first order is $\rho/(1 + \rho)$.

$\Rightarrow$ (*) is true for $s = 1$.

The result will be established by the principle of mathematical induction. Let us suppose that every partial correlation coefficient of order s is given by $\rho/(1 + s\rho)$. Then the partial correlation coefficient of order $(s + 1)$ is given by

$$r_{ij\cdot km\ldots t} = \frac{r_{ij\cdot(s)} - r_{ik\cdot(s)} r_{jk\cdot(s)}}{\sqrt{(1 - r^2_{ik\cdot(s)})\,(1 - r^2_{jk\cdot(s)})}}$$

where $k, m, \ldots, t$ are $(s + 1)$ secondary subscripts and $r_{ij\cdot(s)}$, $r_{ik\cdot(s)}$, $r_{jk\cdot(s)}$, are partial correlation coefficients of order s. Thus

$$r_{ij\cdot km\ldots t} = \frac{\dfrac{\rho}{1 + s\rho} - \left(\dfrac{\rho}{1 + s\rho}\right)^2}{1 - \left(\dfrac{\rho}{1 + s\rho}\right)^2} = \frac{\dfrac{\rho}{1 + s\rho}\left(1 - \dfrac{\rho}{1 + s\rho}\right)}{\left(1 - \dfrac{\rho}{1 + s\rho}\right)\left(1 + \dfrac{\rho}{1 + s\rho}\right)} \frac{\rho}{1 + (s + 1)\rho}$$

Using (**) and (***), the required result follows by induction.

(ii) We have $1 - R^2 = \dfrac{\omega}{\omega_{11}}$

where R is the multiple correlation coefficient of a variable with other $(p - 1)$ variables and

$$\omega = \begin{vmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \rho & \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \vdots & & \vdots \\ \rho & \rho & \rho & \ldots & 1 \end{vmatrix}\ ,\ \text{a determinant of order '}p\text{' and}$$

$$\omega_{11} = \begin{vmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \rho & \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \vdots & & \vdots \\ \rho & \rho & \rho & \ldots & 1 \end{vmatrix}\ \text{a determinant of order } (p - 1).$$

We have

$$\omega = [1 + (p-1)\rho] \begin{vmatrix} 1 & \rho & \rho & \rho \ldots & \rho \\ 1 & 1 & \rho & \rho \ldots & \rho \\ 1 & \rho & 1 & \rho \ldots & \rho \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \rho & \rho & \rho \ldots & 1 \end{vmatrix} \quad \text{(On adding } C_2, C_3 \ldots, C_p \text{ to } C_1\text{).}$$

$$\Rightarrow \quad \omega = [1 + (p-1)\rho] \begin{vmatrix} 1 & \rho & \rho & \rho & \ldots & \rho \\ 0 & (1-\rho) & 0 & 0 & \ldots & 0 \\ 0 & 0 & (1-\rho) & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & \ldots & (1-\rho) \end{vmatrix}$$

[On operating $R_i - R_1$, $(i = 2, 3, \ldots p)$].

$$\therefore \qquad \omega = [1 + (p-1)\rho](1-\rho)^{p-1}$$

Similarly, we will have

$$\omega_{11} = [1 + (p-2)\rho](1-\rho)^{p-2},$$

$$\therefore \qquad 1 - R^2 = \frac{\omega}{\omega_{11}} = (1-\rho)\left[\frac{1 + (p-1)\rho}{1 + (p-2)\rho}\right]$$

Example 10·41. *In a p-variate distribution, all the total (order zero) correlation coefficients are equal to $\rho_0 \neq 0$. Let ρ_k denote the partial correlation coefficient of order k and R_k be the multiple correlation coefficient of one variate on k other variates. Prove that*

(i) $\rho_0 \geq -\frac{1}{(p-1)}$, *(ii)* $\rho_k - \rho_{k-1} = -\rho_k \rho_{k-1}$, *and*

(iii) $R_k^2 = \frac{k\rho_0^2}{1 + (k-1)\rho_0}$. **[*Delhi Univ. M.Sc. (Stat.) 1987*]**

Solution. *(i)* We have proved in Example 10·40, that

$$\rho_k = \frac{\rho_0}{1 + k\rho_0}$$

In the case of *p*-variate distribution, the partial correlation coefficient of the highest order is ρ_{p-2} and is given by

$$\rho_{p-2} = \frac{\rho_0}{1 + (p-2)\rho_0}$$

Since $\quad |\rho_{p-2}| \leq 1 \Rightarrow -1 \leq \rho_{p-2} \leq 1,$

we have (on considering the lower limit)

$$-1 \leq \frac{\rho_0}{1 + (p-2)\rho_0} \quad \text{or} \quad -[1 + (p-2)\rho_0] \leq \rho_0$$

$$\Rightarrow \qquad \rho_0 \geq -\frac{1}{(p-1)}$$

(*ii*) L.H.S. $= \rho_k - \rho_{k-1} = \dfrac{\rho_0}{1 + k\rho_0} - \dfrac{\rho_0}{1 + (k-1)\rho_0}$

$$= \rho_0\left[\frac{-\rho_0}{(1 + k\rho_0)[1 + (k-1)\rho_0]}\right]$$

$$= -\left(\frac{\rho_0}{1 + k\rho_0}\right)\left(\frac{\rho_0}{1 + (k-1)\rho_0}\right) = -\rho_k\rho_{k-1}$$

(*iii*) Taking $\rho = \rho_0$ and $k = p - 1$ in part (*ii*) Example 10·40, we get

$$1 - R_k^2 = (1 - \rho_0)\left[\frac{1 + k\rho_0}{1 + (k-1)\rho_0}\right]$$

$$\therefore \quad R_k^2 = 1 - \frac{(1 - \rho_0)(1 + k\rho_0)}{1 + (k-1)\rho_0} = \frac{k\,\rho_0^2}{1 + (k-1)\rho_0} \text{ (On simplification).}$$

Example 10·42. *If r_{12} and r_{13} are given, show that r_{23} must lie in the range :* $r_{12}\,r_{13} \pm (1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2)^{1/2}$

If $r_{12} = k$ and $r_{13} = -k$, show that r_{23} will lie between -1 and $1 - 2k^2$.

[***Sardar Patel Univ. B.Sc. Oct., 1992; Madras Univ. B.Sc. (Stat. Main) 1991***]

Solution. We have

$$r_{12\cdot 3}^2 = \left[\frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)\,(1 - r_{23}^2)}}\right]^2 \le 1$$

$$\therefore \quad (r_{12} - r_{13}r_{23})^2 \le (1 - r_{13}^2)(1 - r_{23}^2)$$

$$\Rightarrow \quad r_{12}^2 + r_{13}^2 r_{23}^2 - 2r_{12}\,r_{13}\,r_{23} \le 1 - r_{13}^2 - r_{23}^2 + r_{13}^2 r_{23}^2$$

$$\Rightarrow \quad r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}\,r_{13}\,r_{23} \le 1 \qquad \ldots(*)$$

This condition holds for consistent values of r_{12}, r_{13} and r_{23}. (*) may be rewritten as :

$$r_{23}^2 - (2r_{12}r_{13})r_{23} + (r_{12}^2 + r_{13}^2 - 1) \le 0.$$

Hence, for given values of r_{12} and r_{13}, r_{23} must lie between the roots of the quadratic (in r_{23}) equation

$$r_{23}^2 - (2r_{12}\,r_{13})\,r_{23} + (r_{12}^2 + r_{13}^2 - 1) = 0,$$

which are given by :

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

Hence

$$r_{12}\,r_{13} - \sqrt{1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2} \le r_{23} \le r_{1.2}\,r_{13} + \sqrt{(1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2)} \qquad \ldots(**)$$

In other words, r_{23} must lie in the range

$$r_{12}\,r_{13} \pm \sqrt{(1 - r_{12}^2 - r_{13}^2 + r_{12}^2 r_{13}^2)}$$

In particular, if $r_{12} = k$ and $r_{13} = -k$, we get from (**)

$$-k^2 - \sqrt{(1 - k^2 - k^2 + k^4)} \le r_{23} \le -k^2 + \sqrt{(1 - k^2 - k^2 + k^4)}$$

$$\Rightarrow \quad -k^2-(1-k^2) \le r_{23} \le -k^2+(1-k^2)$$

$$\therefore \quad -1 \le r_{23} \le 1-2k^2$$

EXERCISE 10(g)

1. (*a*) Explain partial correlation and multiple correlation.

(*b*) Explain the concepts of multiple and partial correlation coefficients.

Show that the multiple correlation coefficient $R_{1\cdot23}$ is, in the usual notations given by : $R_{1\cdot23}^2 = 1 - \frac{\omega}{\omega_{11}}$

2 (*a*) In the usual notations, prove that

$$R_{1\cdot23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{31}}{1-r_{23}^2} \le r_{12}^2$$

(*b*) If $R_{1\cdot23} = 1$, prove that $r_{2\cdot13}$ is also equal to 1. If $R_{1\cdot23} = 0$, does it necessarily mean that $R_{2\cdot13}$ is also zero ?

3. (*a*) Obtain an expression for the variance of the residual $X_{1\cdot23}$ in terms of the correlations r_{12}, r_{23} and r_{31} and deduce that $R_{1(23)} \ge r_{12}$ and r_{13}.

(*b*) Show that the standard deviation of order p may be expressed in terms of standard deviation of order $(p-1)$ and a correlation coefficient of order $(p-1)$. Hence deduce that :

(*i*) $\sigma_1 \ge \sigma_{1\cdot2} \ge \sigma_{1\cdot23} \ge \ldots \ge \sigma_{1\cdot23\ldots n}$

(*ii*) $1 - R_{1\cdot23\ldots n}^2 = (1-r_{12}^2)(1-r_{13\cdot2}^2)\ldots(1-r_{1n\cdot23\ldots(n-1)}^2)$

[Delhi Univ. M.Sc. (Stat.) 1987]

4. (*a*) In a p-variate distribution all the toal (zero order) correlation coefficients are equal to $\rho_0 \neq 0$. If ρ_k denotes the partial correlation coefficient of order k, find ρ_k. Hence deduce that :

(*i*) $\rho_k - \rho_{k-1} = -\rho_k \rho_{k-1}$

(*ii*) $\rho_0 \ge -1/(p-1)$. *[Delhi Univ. M.Sc. (Stat.), 1989]*

(*b*) Show that the multiple correlation coefficient $R_{1\cdot23\ldots j}$ between X_1 and $(X_2, X_3, \ldots, X_j)$, $j = 2, 3, \ldots, p$ satisfies the inequalities :

$$R_{1\cdot2} \le R_{1\cdot23} \le \ldots \le R_{1\cdot23\ldots p}$$

[Delhi Univ. M.Sc. (Maths.), 1989]

5. (*a*) $X_0, X_1, \ldots, X_n$ are $(n+1)$ variates. Obtain a linear function of $X_1, X_2, \ldots, X_n$ which will have a maximum correlation with X_0. Show that the correlation R of X_0 with the linear function is given by

$$R = \left(1 - \frac{\omega}{\omega_{00}}\right)^{\frac{1}{2}}$$

where $$\omega = \begin{vmatrix} 1 & r_{01} & r_{02}.....r_{0n} \\ r_{10} & 1 & r_{12}.....r_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ r_{n0} & r_{n1} & r_{n2}.....1 \end{vmatrix}$$

and ω_{00} is the determinant obtained by deleting the first row and the first column of ω.

(*b*) With the usual notations, prove that

$$\sigma^2_{1\cdot234\ldots n} = \frac{\omega}{\omega_{11}}\sigma_1^2 = \sigma_1^2(1-r_{12}^2)(1-r_{13\cdot2}^2)\ldots(1-r^2_{1n\cdot23\ldots n-1})$$

(*c*) For a trivariate distribution, prove that

$$r_{12\cdot3} = \frac{r_{12} - r_{13}\,r_{23}}{\sqrt{(1-r_{13}^2)\,(1-r_{23}^2)}}$$

6. (*a*) The simple correlation coefficients between temperature (X_1), corn yield (X_2) and rainfall (X_3) are, $r_{12} = 0{\cdot}59$, $r_{13} = 0{\cdot}46$ and $r_{23} = 0{\cdot}77$.

Calculate the partial correlation coefficients $r_{12\cdot3}$, $r_{23\cdot1}$ and $r_{31\cdot2}$. Also calculate $R_{1\cdot23}$.

(*b*) If $r_{12} = 0{\cdot}80$, $r_{13} = -0{\cdot}40$ and $r_{23} = -0{\cdot}56$, find the values of $r_{12\cdot3}$, $r_{13\cdot2}$ and $r_{23\cdot1}$. Calculate further $R_{1(23)}$, $R_{2(13)}$ and $R_{3(12)}$.

7. (*a*) In certain investigation, the following values were obtained :

$$r_{12} = 0{\cdot}6,\ r_{13} = -0{\cdot}4 \text{ and } r_{23} = 0{\cdot}7$$

Are the values consistent ?

(*b*) Comment on the consistency of

$$r_{12} = \frac{3}{5},\ r_{23} = \frac{4}{5},\ r_{31} = -\frac{1}{2}.$$

(*c*) Suppose a computer has found, for a given set of values of X_1, X_2 and X_3,

$$r_{12} = 0{\cdot}91,\quad r_{13} = 0{\cdot}33 \text{ and } r_{32} = 0{\cdot}81$$

Examine whether the computations may be said to be free from error.

8. (*a*) Show that if $r_{12} = r_{13} = 0$, then $R_{1(23)} = 0$. What is the significance of this result in regard to the multiple regression equation of X_1 on X_2 and X_3 ?

(*b*) For what value of $R_{1\cdot23}$ will X_2 and X_3 be uncorrelated ?

(*c*) Given the data : $r_{12} = 0{\cdot}6$, $r_{13} = 0{\cdot}4$, find the value of r_{23} so that $R_{1\cdot23}$, the multiple correlation coefficient of X_1 on X_2 and X_3 should be unity.

9. From the heights (X_1), weights (X_2) and ages (X_3) of a group of students the following standard deviations and correlation coefficients were obtained : $\sigma_1 = 2{\cdot}8$ inches, $\sigma_2 = 12$ lbs, and $\sigma_3 = 1{\cdot}5$ years, $r_{12} = 0{\cdot}75$, $r_{23} = 0{\cdot}54$, and $r_{31} = 0{\cdot}43$. Calculate (*i*) partial regression coefficients and (*ii*) partial correlation coefficients.

10. For a trivariate distribution :

$\bar{X}_1 = 40$	$\bar{X}_2 = 70$	$\bar{X}_3 = 90$
$\sigma_1 = 3$	$\sigma_2 = 6$	$\sigma_3 = 7$
$r_{12} = 0{\cdot}4$	$r_{23} = 0{\cdot}5$	$r_{13} = 0{\cdot}6$

Find

(*i*) $R_{1\cdot23}$, (*ii*) $r_{23\cdot1}$, (*iii*) the value of X_3 when $X_1 = 30$ and $X_2 = 45$.

11. (*a*) In a study of a random sample of 120 students, the following results are obtained :

$$\bar{X}_1 = 68, \quad \bar{X}_2 = 70, \quad \bar{X}_3 = 74$$
$$S_1^2 = 100, \quad S_2^2 = 25, \quad S_3^2 = 81,$$
$$r_{12} = 0{\cdot}60, \quad r_{13} = 0{\cdot}70, \quad r_{23} = 0{\cdot}65$$

$[S_i^2 = \text{Var}\,(X_i)]$, where X_1, X_2, X_3 denote percentage of marks obtained by a student in I test, II test and the final examination respectively.

(*i*) Obtain the least square regression equation of X_3 on X_1 and X_2.

(*ii*) Compute $r_{12\cdot3}$ and $R_{3\cdot12}$.

(*iii*) Estimate the percentage marks of a student in the final examination if he gets 60% and 67% in I and II tests respectively.

(*b*) X_1 is the consumption of milk per head, X_2 the mean price of milk, and X_3, the per capita income. Time series of the three variables are rendered trend free and the standard deviations and correlation coefficients calculated :

$$s_1 = 7{\cdot}22, \quad s_2 = 5{\cdot}47, \quad s_3 = 6{\cdot}87$$
$$r_{12} = -0{\cdot}83, \quad r_{13} = 0{\cdot}92, \quad r_{23} = -0{\cdot}61$$

Calculate the regression equation of X_1 on X_2 and X_3 and interpret the regression as a demand equation.

12. (*a*) Five thousand candidates were examined in the subjects (*a*), (*b*), (*c*); each of these subjects carrying 100 marks. The following constants relate to these data :

	Subjects		
	(*a*)	(*b*)	(*c*)
Mean	39·46	52·31	45·26
Standard deviation	6·2	9·4	8·7
	$r_{bc} = 0{\cdot}47$	$r_{ca} = 0{\cdot}38$	$r_{ab} = 0{\cdot}29$

Assuming normally correlated population, find the number of candidates who will pass if minimum pass marks are an aggregate of 150 marks for the three subjects together.

(*b*) Establish the equation of plane of regression for variates X_1, X_2, X_3 in the determinant form

$$\begin{vmatrix} X_1/\sigma_1 & X_2/\sigma_2 & X_3/\sigma_3 \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix} = 0$$

[Delhi Univ. B.Sc. (Maths. Hons.), 1986]

13. (*a*) Prove the identity

$$b_{12\cdot3}\, b_{23\cdot1}\, b_{31\cdot2} = r_{12\cdot3}\, r_{23\cdot1}\, r_{31\cdot2}$$ ***[Gujarat Univ. B.Sc., 1992]***

(*b*) Prove that

$$R_{1\cdot 23}^2 = b_{12\cdot 3}\, r_{12}\frac{\sigma_2}{\sigma_1} + b_{13\cdot 2}\, r_{13}\frac{\sigma_3}{\sigma_1}$$

[*Sardar Patel Univ. B.Sc., 1991*]

14. (*a*) If $X_3 = aX_1 + bX_2$ for all sets of values of X_1, X_2, and X_3, find the value of $r_{23\cdot 1}$.

(*b*) If the relation $aX_1 + bX_2 + cX_3 = 0$ holds for all sets of values X_1, X_2 and X_3, what must be the partial correlation coefficients ?

15. (*a*) If $r_{12} = r_{23} = r_{31} = \rho \neq 1$, then

$$r_{12\cdot 3} = r_{23\cdot 1} = r_{31\cdot 2} = \frac{\rho}{1+\rho} \text{ and } R_{1(23)} = R_{2(13)} = R_{3(12)} = \frac{\rho\sqrt{2}}{\sqrt{(1+\rho)}}$$

(*b*) Y_1, Y_2, Y_3 are uncorrelated standard variates. $X_1 = Y_2 + Y_3$, $X_2 = Y_3 + Y_1$, and $X_3 = Y_1 + Y_2$. Find the multiple correlation coefficient between X_3 and (X_1 and X_2).

16. X, Y, Z are independent random variables with the same variance. If

$$X_1 = \frac{1}{\sqrt{2}}(X - Z),\ X_2 = \frac{1}{\sqrt{3}}(X + Y + Z),\ X_3 = \frac{1}{\sqrt{6}}(X + 2Y + Z),$$

show that X_1, X_2, X_3 have equal variances. Calculate $r_{12\cdot 3}$ and $R_{1(23)}$.

17. (*a*) If X_1, X_2 and X_3 are three variables measured from their respective means as origin and if e_1 is the expected value of X_1 for given values of X_2 and X_3 from the linear regression of X_1 on X_2 and X_3, prove that

$$\text{Cov}\,(X_1, e_1) = \text{Var}\,(e_1) = \text{Var}\,(X_1) - \text{Var}\,(X_1 - e_1)$$

(*b*) If $r_{12} = k$ and $r_{23} = -k$, show that r_{13} will lie between -1 and $1 - 2k^2$.

18. (*a*) For three variables X, Y and Z, prove that

$$r_{XY} + r_{YZ} + r_{ZX} \geq -\frac{3}{2} \qquad \ldots(*)$$

Hint. Let us transform X, Y, Z to their standard variables U, V and W, (say), respectively, where

$$U = \frac{X - E(X)}{\sigma_X},\ V = \frac{Y - E(Y)}{\sigma_Y},\ W = \frac{Z - E(Z)}{\sigma_Z}$$

so that

$$\left.\begin{aligned} &E(U) = E(V) = E(W) = 0 \\ \text{and}\quad &\sigma_U^2 = \sigma_V^2 = \sigma_W^2 = 1 \Rightarrow E(U^2) = E(V^2) = E(W^2) = 1 \end{aligned}\right\} \ldots(**)$$

$$\left.\begin{aligned} &r_{UV} = \frac{\text{Cov}\,(U, V)}{\sigma_U \sigma_V} = \frac{E(UV) - E(U)\,E(V)}{\sigma_U\,\sigma_V} = E(UV) \\ \text{and}\quad &r_{UW} = E(UW);\ r_{VW} = E(VW) \end{aligned}\right\} \ldots(***)$$

Since correlation coefficient is independent of change of origin and scale, proving (*) is equivalent to proving

$$r_{UV} + r_{VW} + r_{UW} \geq -3/2 \qquad \ldots(****)$$

To establish (****) let us consider the $E(U + V + W)^2$, which is always non-negative *i.e.*, $E(U + V + W)^2 \geq 0$, and use (**) and (***).

(*b*) X,Y,Z are three reduced (standard) variates and $E(YZ) = E(ZX) = -1/2$, find the limits between which the coefficient of correlation $r(X, Y)$ is necessarily placed.

Hint. Consider $E(X + Y + Z)^2 \geq 0 \Rightarrow r \geq -\frac{1}{2}$.

(*c*) If r_{12}, r_{23} and r_{31} are correlation coefficients of any three random variables X_1, X_2 and X_3 taken in pairs (X_1, X_2), (X_2, X_3) and (X_3, X_1) respectively, show that $1 + 2r_{12}\, r_{23}\, r_{31} \geq r_{12}^2 + r_{13}^2 + r_{23}^2$

19. (*a*) If the relation $aX_1 + bX_2 + cX_3 = 0$, holds for all sets of values of X_1, X_2 and X_3, where X_1, X_2 and X_3 are three standardised variables, find the three total correlation coefficients r_{12}, r_{23} and r_{13} in terms of a, b and c. What are the values of partial correlation coefficients if a, b and c are positive ?

(*b*) Suppose X_1, X_2 and X_3 satisfy the relation $a_1X_1 + a_2X_2 + a_3X_3 = k$.

(*i*) Determine the three total correlation coefficients in terms of standard deviations and the constants a_1, a_2 and a_3.

(*ii*) State what the partial correlation coefficients would be.

20. (*a*) Show that the multiple correlation between Y and $X_1, X_2, \ldots, X_p$ is the maximum correlation between Y and any linear function of $X_1, X_2, \ldots, X_p$.

(*b*) Show that for p variates there are ${}^pC_2$ correlation coefficients of order zero and ${}^{p-2}C_s \cdot {}^pC_2$ of order s. Show further that there are ${}^pC_2 \cdot 2^{p-2}$ correlation coefficients altogether and ${}^pC_2 \cdot 2^{p-1}$ regression coefficients.

ADDITIONAL EXERCISES ON CHAPTER X

1. Find the correlation coefficient between

(*i*) $aX + b$ and Y, (*ii*) $lx + mY$ and $X + Y$, when correlation coefficient between X and Y is ρ.

2. If X_1 and X_2 are independent normal variates and U and V are defined by

$$U = X_1 \cos\alpha + X_2 \sin\alpha, \quad V = X_2 \cos\alpha - X_1 \sin\alpha,$$

show that the correlation coefficient ρ between U and V is given by

$$\rho^2 = 1 - \frac{4\sigma_1^2\sigma_2^2}{4\sigma_1^2\sigma_2^2 + (\sigma_1^2 - \sigma_2^2)\sin^2 2\alpha},$$

where σ_1^2 and σ_2^2 are variances of X_1 and X_2 respectively.

3. The variables X and Y are normally correlated, and ξ, η are defined by

$$\xi = X\cos\theta + Y\sin\theta, \quad \eta = Y\cos\theta - X\sin\theta$$

Obtain θ so that the distributions of ξ and η are independent.

4. A set of n observations of simultaneous values of X and Y are made by an observer and the standard deviations and product moment coefficient about the mean are found to be σ_X, σ_Y and ρ_{XY}. A second observer repeating the same observations made a constant error e in observing each X and a constant error E in observing each Y. The two sets of observations are combined into a single set and coefficient of correlation calculated from it. Show that its value is

$$(\rho_{XY} + \tfrac{1}{4}eE) + \sqrt{(\sigma_X^2 + \tfrac{1}{4}e^2)(\sigma_Y^2 + \tfrac{1}{4}E^2)}$$