MAT2001
Statistics for Engineers

Module 3
Correlation and Regression

Syllabus

Correlation and Regression

Correlation and Regression – Rank Correlation – Partial and Multiple Correlation – Multiple Regression.

Covariance

$$Var(X) = E\left[(X - E(X)) \cdot (X - E(X))\right]$$

$$Covar(X, Y) = E\left[(X - E(X)) \cdot (Y - E(Y))\right]$$

Correlation

Correlation Co-efficient

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 $p(X, Y)$.

$\dfrac{p(x, y)}{\sigma_x \sigma_y}$ 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
$$r_{XY} = \frac{E\{[X - E(X)][Y - E(Y)]\}}{\sqrt{E\{X - E(X)\}^2 E\{Y - E(Y)\}^2}} \qquad (1)$$

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

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

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

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

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

Properties of Correlation Coefficient

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

   Note: When $0 < r_{XY} \leq 1$, the correlation between X and Y is said to be *positive* or direct.
   When $-1 \leq r_{XY} \leq 0$, the correlation is said to be *negative* or *inverse*.
   When $-1 \leq r_{XY} \leq -0.5$ or $0.5 \leq r_{XY} \leq 1$, 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.

Example:

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

Comment about the nature of correlation.

Solution:

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

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

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

Exercise:

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

Rank Correlation

Rank Correlation Co-efficient

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.

$$D = U - V$$

$$\rho_{XY} = 1 - \frac{6\,\Sigma d^2}{n\left(n^2 - 1\right)}$$

[**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.

Tied or Repeated Ranks

When there is a repetition of ranks, a correction factor $\frac{m(m^2-1)}{12}$ is added to $\sum d^2$ in the Spearman's rank correlation coefficient formula, where m is the number of times a rank is repeated. It is very important to know that this correction factor is added for every repetition of rank in both characters.

Thus, in case of tied or repeated rank Spearman's rank correlation coefficient formula is

$$r_s = 1 - \frac{6\left\{\sum d^2 + \frac{m(m^2-1)}{12} + \ldots\right\}}{n(n^2-1)}$$

Example:

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.

Solution:

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	$: \Sigma d^2 = 26$

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

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

Example:

Suppose we have ranks of 8 students of B.Sc. in Statistics and Mathematics. On the basis of rank we would like to know that to what extent the knowledge of the student in Statistics and Mathematics is related.

Rank in Statistics	1	2	3	4	5	6	7	8
Rank in Mathematics	2	4	1	5	3	8	7	6

Solution: Spearman's rank correlation coefficient formula is

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

Solution:

Let us denote the rank of students in Statistics by R_x and rank in Mathematics by R_y. For the calculation of rank correlation coefficient we have to find $\sum_{i=1}^{n} d_i^2$ which is obtained through the following table:

Rank in Statistics (R_x)	**Rank in Mathematics** (R_y)	**Difference of Ranks** $(d_i = R_x - R_y)$	d_i^2
1	2	−1	1
2	4	−2	4
3	1	2	4
4	5	−1	1
5	3	2	4
6	8	−2	4
7	7	0	0
8	6	2	4
			$\sum d_i^2 = 22$

Here, n = number of paired observations = 8

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 22}{8 \times 63} = 1 - \frac{132}{504} = \frac{372}{504} = 0.74$$

Thus there is a positive association between ranks of Statistics and Mathematics.

Example:

Suppose we have ranks of 5 students in three subjects Computer, Physics and Statistics and we want to test which two subjects have the same trend.

Rank in Computer	2	4	5	1	3
Rank in Physics	5	1	2	3	4
Rank in Statistics	2	3	5	4	1

Solution:

In this problem, we want to see which two subjects have same trend i.e. which two subjects have the positive rank correlation coefficient.

Here we have to calculate three rank correlation coefficients

r_{12s} = Rank correlation coefficient between the ranks of Computer and Physics

r_{23s} = Rank correlation coefficient between the ranks of Physics and Statistics

r_{13s} = Rank correlation coefficient between the ranks of Computer and Statistics

Let R_1, R_2 and R_3 be the ranks of students in Computer, Physics and Statistics respectively.

Rank in Computer (R_1)	**Rank in Physics (R_2)**	**Rank in Statistics (R_3)**	**$d_{12}=R_1-R_2$**	**d_{12}^2**	**$d_{23}=R_2-R_3$**	**d_{23}^2**	**$d_{13}=R_1-R_3$**	**d_{13}^2**
2	5	2	−3	9	3	9	0	0
4	1	3	3	9	−2	4	1	1
5	2	5	3	9	−3	9	0	0
1	3	4	−2	4	−1	1	−3	9
3	4	1	−1	1	−3	9	2	4
Total				32		32		14

Solution (Continued):

Thus,

$$\sum d_{12}^2 = 32,\ \sum d_{23}^2 = 32 \text{ and } \sum d_{13}^2 = 14.$$

Now

$$r_{12s} = 1 - \frac{6\sum d_{12}^2}{n(n^2-1)} = 1 - \frac{6 \times 32}{5 \times 24} = 1 - \frac{8}{5} = -\frac{3}{5} = -0.6$$

$$r_{23s} = 1 - \frac{6\sum d_{23}^2}{n(n^2-1)} = 1 - \frac{6 \times 32}{5 \times 24} = 1 - \frac{8}{5} = -\frac{3}{5} = -0.6$$

$$r_{13s} = 1 - \frac{6\sum d_{13}^2}{n(n^2-1)} = 1 - \frac{6 \times 14}{5 \times 24} = 1 - \frac{7}{10} = \frac{3}{10} = 0.3$$

r_{12s} is negative which indicates that Computer and Physics have opposite trend. Similarly, negative rank correlation coefficient r_{23s} shows the opposite trend in Physics and Statistics. $r_{13s} = 0.3$ indicates that Computer and Statistics have same trend.

Example:

Calculate rank correlation coefficient from the following data:

Expenditure on advertisement	10	15	14	25	14	14	20	22
Profit	6	25	12	18	25	40	10	7

Solution:

Let us denote the expenditure on advertisement by x and profit by y

x	Rank of x (R_x)	y	Rank of y (R_y)	$d = R_x - R_y$	d^2
10	8	6	8	0	0
15	4	25	2.5	1.5	2.25
14	6	12	5	1	1
25	1	18	4	−3	9
14	6	25	2.5	3.5	12.25
14	6	40	1	5	25
20	3	10	6	−3	9
22	2	7	7	−5	25
					$\sum d^2 = 83.50$

$$r_s = 1 - \frac{6\left\{\sum d^2 + \frac{m(m^2-1)}{12} + \ldots\right\}}{n(n^2-1)}$$

Solution (Continued):

Here rank 6 is repeated three times in rank of x and rank 2.5 is repeated twice in rank of y, so the correction factor is

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

Hence rank correlation coefficient is

$$r_s = 1-\frac{6\left\{83.50+\frac{3(3^2-1)}{12}+\frac{2(2^2-1)}{12}\right\}}{8(64-1)}$$

$$r_s = 1-\frac{6\left\{83.50+\frac{3\times 8}{12}+\frac{2\times 3}{12}\right\}}{8\,X\,63}$$

$$r_s = 1-\frac{6(83.50+2.50)}{504}$$

$$r_s = 1-\frac{516}{504}$$

$$r_s = 1-1.024 = -0.024$$

There is a negative association between expenditure on advertisement and profit.

Exercise:

Calculate rank correlation coefficient from the following data:

x	78	89	97	69	59	79	68
y	125	137	156	112	107	136	124

Exercise:

Calculate Spearman's rank correlation coefficient from the following data:

x	20	38	30	40	50	55
y	17	45	30	35	40	25

Exercise:

Calculate rank correlation coefficient from the following data:

x	10	20	30	30	40	45	50
y	15	20	25	30	40	40	40

Exercise:

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.

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:

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

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

the equation of the regression line of Y on X as

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

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

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

Equation of the Regression Line of X on Y:

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

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

$$x - \bar{x} = \frac{p_{xy}}{\sigma_y^2}(y - \bar{y})$$

or

$$x - \bar{x} = \frac{r_{xy}\,\sigma_x}{\sigma_y}(y - \bar{y})$$

X on Y: $(y - \bar{y}) = \dfrac{r_{xy}\sigma_y}{\sigma_x}(x - \bar{x})$

X on Y: $(x - \bar{x}) = \dfrac{r_{xy}\,\sigma_x}{\sigma_y}(y - \bar{y})$

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.

   $\therefore \qquad r_{XY} = \pm \sqrt{b_1\, b_2}$

   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 $\qquad \frac{b_1}{b_2} = \frac{\sigma_Y^2}{\sigma_X^2}$
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.

Example:

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

Example:

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

Find y, when $x = 3.4$

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1				
2				
3				
4				
5				
6				
7				
Total	Σx_i	Σy_i	Σx_i^2	Σy_i^2	$\Sigma x_i y_i$

$\bar{x} = \frac{\Sigma x_i}{n}$ $\quad$ $\bar{y} = \frac{\Sigma y_i}{n}$

$\sigma_x = \sqrt{\frac{1}{n}\Sigma x_i^2 - \left(\frac{1}{n}\Sigma x_i\right)^2}$ $\quad$ $\sigma_y = \sqrt{\frac{1}{n}\Sigma y_i^2 - \left(\frac{1}{n}\Sigma y_i\right)^2}$

$Cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \frac{1}{n}\Sigma x_i \cdot \frac{1}{n}\Sigma y_i$

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

Regression line of Y on X

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

$y = ax + b$ ——— (1)

Put $x = 3.4$

Regression of X on Y

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

$x = ay + b$ ——— (2)

Put $y = 10.4$

Solution:

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

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

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

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

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

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

The regression line of Y on X is

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

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

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

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

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

Example:

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?

Solution:

(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

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

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

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

$\therefore \quad r^2_{XY} = b_{XY} \times b_{YX} = 4$

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

Hence our tentative assumption is wrong.

Solution (Continued):

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

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

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

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

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

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

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

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

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

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

Exercise:

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

Exercise:

Obtain the equations of the regression lines from the following data, using the method of least squares.

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

Exercise:

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?

$(X_1, X_2, X_3) \rightarrow$ Trivariate Distribution

$$r(X_1, X_2) = r_{X_1 X_2} = r_{12} = r_{21}$$

$$r(X_1, X_3) = r_{X_1 X_3} = r_{13} = r_{31}$$

$$r(X_2, X_3) = r_{X_2 X_3} = r_{23} = r_{32}$$

$$r_{xy} = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_x} = r_{yx}$$

Multiple and Partial Correlation

Multiple Correlation

Suppose one variable may be influenced by many other variables. Such a correlation is called multiple correlation.

Multiple Correlation Co-efficient (R)

In a trivariate distribution (X_1, X_2, X_3), the multiple correlation co-efficient of X_1 on X_2 & X_3 is denoted and defined as

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

Similarly $R_{2\cdot 13}$ & $R_{3\cdot 12}$

Note:

* $R^2_{1\cdot 23} \leq 1$ $\qquad |r_{xy}| \leq 1$
* $0 \leq R_{1\cdot 23} \leq 1$

Partial Correlation

The correlation between 2 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, the effect of X_3 on each of X_1 and X_2 were eliminated. Such a correlation is called Partial Correlation.

Partial Correlation Co-efficient:

The partial co-relation co-efficient between X_1 & X_2 after eliminating the linear effect of X_3, is denoted & defined as

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

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

Example:

In a trivariate distribution : $r_{12} = 0{\cdot}77$, $r_{13} = 0{\cdot}72$ and $r_{23} = 0{\cdot}52$

Find the the partial correlation coefficient $r_{12{\cdot}3}$ and multiple correlation coefficient $R_{1{\cdot}23}$.

Solution:

$$r_{12{\cdot}3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r_{13}^2)\,(1 - r_{23}^2)}} = \frac{0{\cdot}77 - 0{\cdot}72 \times 0{\cdot}52}{\sqrt{[1 - (0{\cdot}72)^2][1 - (0{\cdot}52)^2]}} = 0{\cdot}62$$

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

$$= \frac{(0{\cdot}77)^2 + (0{\cdot}72)^2 - 2 \times 0{\cdot}77 \times 0{\cdot}72 \times 0{\cdot}52}{1 - (0{\cdot}52)^2} = 0{\cdot}7334$$

$$\therefore \qquad R_{1{\cdot}23} = + 0{\cdot}8564$$

Exercise:

In a trivariate distribution : $r_{12} = 0{\cdot}7, r_{23} = r_{31} = 0{\cdot}5.$

Find (i) $r_{23.1}$, (ii) $R_{1.23}$,

Multiple Regression

Method of Least Squares

$y, x_1 \propto x_2$

$$\sum_{i=1}^{n} y_i = a \cdot \sum x_{1i} + b \cdot \sum x_{2i} + n \cdot c.$$

$$\sum x_{1i} y_i = a \cdot \sum x_{1i}^2 + b \cdot \sum x_{1i} x_{2i} + c \cdot \sum x_{1i}$$

$$\sum x_{2i} y_i = a \cdot \sum x_{1i} x_{2i} + b \sum x_{2i}^2 + c \cdot \sum x_{2i}$$

Multiple Regression

Normal Estimation Equations for Multiple Linear Regression

$$
\begin{aligned}
nb_0 + b_1 \sum_{i=1}^{n} x_{1i} \quad &+ b_2 \sum_{i=1}^{n} x_{2i} \quad + \cdots + b_k \sum_{i=1}^{n} x_{ki} \quad = \sum_{i=1}^{n} y_i \\
b_0 \sum_{i=1}^{n} x_{1i} + b_1 \sum_{i=1}^{n} x_{1i}^2 \quad &+ b_2 \sum_{i=1}^{n} x_{1i}x_{2i} + \cdots + b_k \sum_{i=1}^{n} x_{1i}x_{ki} = \sum_{i=1}^{n} x_{1i}y_i \\
\vdots \qquad &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
b_0 \sum_{i=1}^{n} x_{ki} + b_1 \sum_{i=1}^{n} x_{ki}x_{1i} &+ b_2 \sum_{i=1}^{n} x_{ki}x_{2i} + \cdots + b_k \sum_{i=1}^{n} x_{ki}^2 \quad = \sum_{i=1}^{n} x_{ki}y_i
\end{aligned}
$$

Example:

The following data represent the chemistry grades for a random sample of 12 freshmen at a certain college along with their scores on an intelligence test administered while they were still seniors in high school. The number of class periods missed is also given.

Student	Chemistry Grade, y	Test Score, x_1	Classes Missed, x_2
1	85	65	1
2	74	50	7
3	76	55	5
4	90	65	2
5	85	55	6
6	87	70	3
7	94	65	2
8	98	70	5
9	81	55	4
10	91	70	3
11	76	50	1
12	74	55	4

(a) Fit a multiple linear regression equation of the form $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$.

(b) Estimate the chemistry grade for a student who has an intelligence test score of 60 and missed 4 classes.

Multiple Regression

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

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

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

$$\omega_{13} = r_{23}\, r_{12} - r_{13}$$

Example:

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

$$\omega_{11} = \begin{vmatrix} 1 & r_{23} \\ r_{32} & 1 \end{vmatrix} = 1 - r_{23}^2 = 1 - (-0.56)^2 = 0.686$$

$$\omega_{12} = -\begin{vmatrix} r_{21} & r_{23} \\ r_{31} & 1 \end{vmatrix} = r_{13}\, r_{23} - r_{21} = -0.576$$

$$\omega_{13} = r_{23}\, r_{12} - r_{13} = (-0.56)\,(0.80) - (-0.40) = -0.048$$

∴ Required equation of plane of regression of X_1 on X_2 and X_3 is given by

$$\frac{0.686}{4.42}(X_1 - 28.02) + \frac{(-0.576)}{1.10}(X_2 - 4.91) + \frac{(-0.048)}{85.00}(X_3 - 594) = 0$$