CHAPTER # 5

SIMPLE REGRESSION AND CORRELATION

5.1 REGRESSION

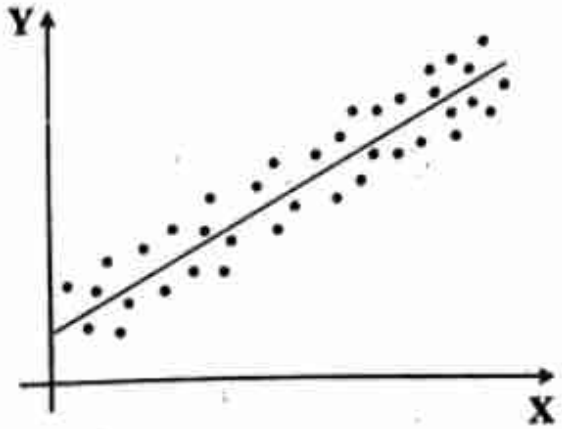
The term regression was introduced by the English biometrician, Sir Francis Galton in 1877. He studied the heights of parents and their children and observed an interesting relationship between them. He found that, though all tall parents have tall children and short parents have short children, the average height of children tends to step back or to regress towards the average height of all men. This tendency towards the average height of all men was called a regression by Galton.

Today, the world regression is used in quite different sense. It investigates the dependence of one variable on other variables, called independent variables and provides an equation to be used for estimating the average value of the dependent variable from the known values of independent variable. For example, if we want to estimate the heights of children on the basis of their ages, the height would be the dependent variable and the ages would be the independent variable.

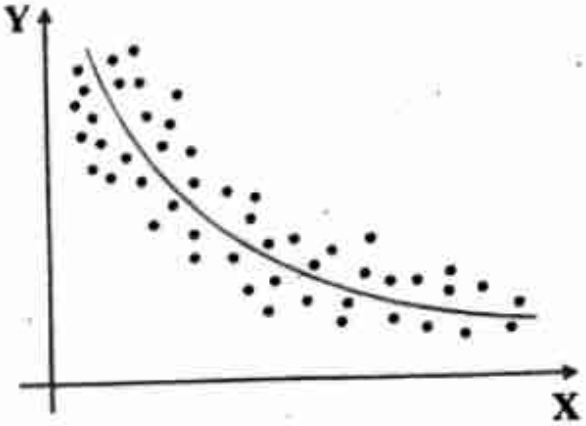
5.2 SCATTER DIAGRAM

A scatter diagram can be used to find whether or not a relationship between two variables exists. The pair of independent - dependent observations as a point on Graph paper using the x-axis for the independent variable and y-axis for the dependent variable. Such a diagram is called a scatter diagram. By looking to the scatter of the various points on the graph we can form an idea as whether variables are related or not. The greater is the scatter of the plotted points, the lesser is the relationship between the variables. If a relationship between the variables exists, then the points in the scatter diagram will show a tendency to cluster around a straight line or some curve. Such a line or curve around which the points cluster, is called the regression line.

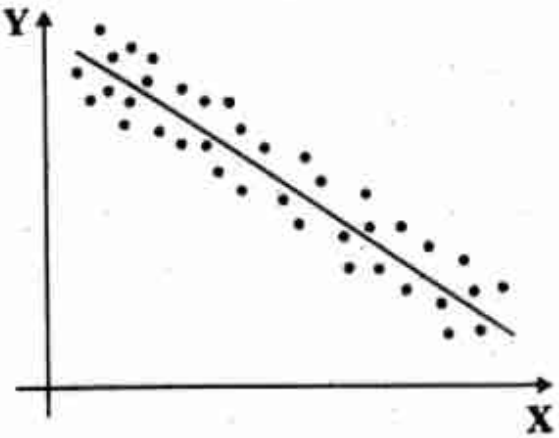
The following scatter diagrams show different types of relation between the variables.



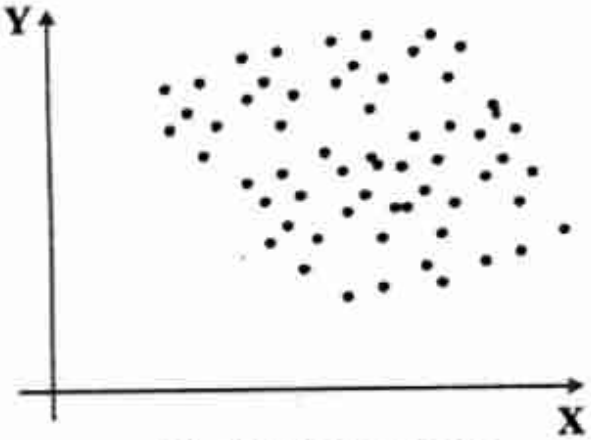
(a) Positive Linear Relationship



(c) Curvilinear Relationship



(b) Negative Linear Relationship



(d) No Relationship

5.3 SIMPLE LINER REGRESSION MODEL

We assume that the linear relationship between the dependent variable Y_i and the value X_i of the regressor X is

$$Y_i = \alpha + \beta X_i + \varepsilon i,$$

where the X_i's are fixed or predetermined values,

the Y_i's are observations randomly drawn from a population,

the ε_i's are error components or random deviations,

α and β are population parameters, α is the intercept and the slope β is called *regression coefficient.*

Furthermore, we assume that

(i) $E(\varepsilon_i) = 0$, *i.e.* the expected value of error term is zero, it implies that the expected value of Y is related to X in the population by a straight line;

(ii) $\text{Var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$ for all i, *i.e.* the variance of error term is constant. It means that the distribution of error has the same variance for all values of X. *(Homoscedasticity assumption);*

(iii) $E(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$, *i.e.* error terms are independent of each other (*assumption of no serial or auto correlation between* ε's);

(iv) $E(X, \varepsilon i) = 0$, *i.e.* X and ε are also independent of each other;

(v) ε_i's are normally distributed with a mean of zero and a constant variance σ^2. This implies that Y values are also normally distributed.

But in practice, we have a sample from a population, therefore we desire to estimate the population regression line from the sample data. The estimated regression line may be written as

$$y = a + bx$$

Where y is the dependent variable and x is the independent variable, while "b" is the regression coefficient and a is an intercept.

Types of Regression Lines

(i) Regression line y on x

$$y = a + b_{yx}$$

$$\text{Where } b_{yx} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

$$\text{Or} \quad b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$\text{and} \quad a = \bar{y} - b\bar{x}$$

Regression equation interms of correlation is

$$(y - \bar{y}) = r\frac{S_y}{S_x}(x - \bar{x})$$

Where r = correlation coefficient

S_y = standard deviation of y

S_x = standard deviation of x

(ii) Regression Equation x on y

$$x = a + b_{xy}y$$

$$\text{Where } b_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

$$\text{Or} \quad b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2}$$

$$\text{and} \quad a = \bar{x} - b\bar{y}$$

Regression equation x on y interms of correlation coefficient is

$$(x - \bar{x}) = r\frac{S_x}{S_y}(y - \bar{y}) \qquad \therefore\ b_{xy} = r\frac{S_x}{S_y}$$

EXAMPLE 5.1

Calculation regression line y on x from the data given below. Also predict y when x = 5.

X	4	3	1	2	6	7	2	3
y	39	38	16	18	41	45	25	38

SOLUTION

x	y	xy	x^2
4	39	156	16
3	38	114	9
1	16	16	1
2	18	36	4
6	41	246	36
7	45	315	49
2	25	50	4
3	38	114	9
28	**260**	**1047**	**128**

We know that regression line y on x is

$$y = a + bx \quad \ldots\ldots (i)$$

$$b = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{8(1047) - 28(260)}{8(128) - (28)^2}$$

$$= \frac{8376 - 7280}{1024 - 784} = \frac{1096}{240} = 4.57$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{260}{8} = 32.5$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{28}{8} = 3.5$$

$$a = \bar{y} - b\bar{x}$$

$$= 32.5 - 4.57(3.5)$$

$$= 32.5 - 15.995 = 16.505$$

Putting the values of a and b in (i), we have

$$\hat{y} = 16.505 + 4.57x$$

Predict y, when x = 5

$$\hat{y} = 16.505 + 4.57\,(5)$$
$$= 16.505 + 22.85$$
$$= 39.36$$

EXAMPLE 5.2

The relationship between money spent on research and development and chemical firms of annual profit are given below.

Year	Million spent on R & D X	Annual profit (millions) Y
1995	5	31
1994	11	40
1993	4	30
1992	5	34
1991	3	25
1990	2	20

Find regression equation y on x

SOLUTION

The estimated regression line y on x is

$$\hat{y} = a + bx$$

where $$b = \frac{n\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2}$$

$$a = \bar{y} - b\bar{x}$$

The necessary calculations are given below.

x	y	xy	x^2
5	31	155	25
11	40	440	121
4	30	120	16
5	34	170	25
3	25	75	9
2	20	40	4
30	**180**	**1000**	**200**

Substituting the values we get

$$b = \frac{6(1000) - 30(180)}{6(200) - (30)^2}$$

$$b = \frac{6000 - 5400}{1200 - 900}$$

$$b = \frac{600}{300} = 2$$

$$\bar{x} = 30/6 = 5, \bar{y} = 180/6 = 30$$

$$a = 30 - 2(5) = 20$$

The estimated regression line is

$$\hat{y} = 20 + 2x$$

5.4 PROPERTIES OF REGRESSION LINE

(i) Least square regression line always passes through mean $(\bar{x}, \bar{y})$ of the data

(ii) The sum of the deviations of the values y from the regression line is always equal to zero.

i.e. $\Sigma(y - \hat{y}) = 0$

(iii) The sum of the square deviations of the observed values from the least square regression line is minimum.

i.e. $\Sigma(y - \hat{y})^2 = \text{minimum}$

(iv) Regression line is the line of best fit, because a and b are unbiased estimates of parameters α and β.

5.5 STANDARD DEVIATION OF REGRESSION OR STANDARD ERROR OF ESTIMATE

The degree of scatter of the observed values about the regression line is called standard error of estimate of y on x. For sample data the standard error of estimate is denoted by S_{yx} and defined as

$$S_{yx} = \sqrt{\frac{\Sigma (y - \hat{y})^2}{n-2}}$$

Or $$= \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}}$$

5.6 COEFFICIENT OF DETERMINATION

The coefficient of determination measures the proportion of variability in the values of dependent variable (y) explained by its linear relationship with the independent variable (x). It is defined as the ratio of explained variation to the total variation. Sample coefficient of determination is denoted by r^2 and defined by

$$r^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2} = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}, \text{ where}$$

$$\Sigma(y - \bar{y})^2 = \Sigma y^2 - a\Sigma y - b\Sigma xy$$

and

$$\Sigma(y - \bar{y})^2 = \Sigma y^2 (\Sigma y)^2/n$$

The alternative formula is

$$r^2 = \frac{a\Sigma y + b\Sigma xy - (\Sigma y)^2/n}{\Sigma y^2 - (\Sigma y)^2/n}$$

EXAMPLE 5.3

Compute the standard error of estimate, coefficient of determination and coefficient of correlation for the data in Example, 5.1

SOLUTION

For the data in example 5.1, we found from the regression calculation, that

$\Sigma y = 260$, $\Sigma y^2 = 9320$ $n = 8$

$a = 16.505$, $b = 4.57$, $\Sigma xy = 1047$

We know that

$$S_{y.x} = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}}$$

$$= \sqrt{\frac{9320 - 16.505(260) - 4.57(1047)}{8 - 2}}$$

$$= \sqrt{\frac{9320 - 4291.3 - 4784.79}{6}}$$

$$= \sqrt{40.65} \qquad \Rightarrow S_{y.x} = 6.376$$

Which is the standard error of estimate. The coefficient of determination is

$$r^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

$$= \frac{a\Sigma y + b\Sigma xy - (\Sigma y)^2/n}{\Sigma y^2 - (\Sigma y)^2/n}$$

$$r^2 = \frac{16.505\,(260) + 4.57\,(1407) - (260)^2/8}{9320 - (260)^2/8}$$

$$= \frac{4291.3 + 4784.79 - 8450}{9320 - 8450} = \frac{626.09}{870} = 0.719$$

The coefficient of correlation is

$$r = \sqrt{0.719} = 0.848$$

5.7 CORRELATION

Correlation measures the strength or closeness of relationship between two variables. The purpose of correlation is to determine whether or not two variables are related. In other words the degree of interdependence between two variables is called correlation. For example, heights and weights of persons ages of husbands and their wives at the time of their marriages, demand and supply of a commodity etc.

5.8 POSITIVE AND NEGATIVE CORRELATION

Positive Correlation

If both the variables tend to increase or decrease together, the correlation is said to be direct or positive e.g. length of an iron bar will increase as the temperature increases.

Negative Correlation

If one variable tends to increase as the other variable decreases the correlation is said to be negative or inverse, e.g. the volume of gas will decrease as pressure increases. Coldness increases as the temperature decreases.

5.9 PERFECT CORRELATION

The correlation coefficient r is a pure number (*i.e.* independent of the units) and it assumes values that can range from +1 for perfect positive linear relationship −1 for perfect negative linear relationship. In case of no correlation, the correlation is said to be zero correlation.

5.10 PEARSON PRODUCT MOMENT CORRELATION CO-EFFICIENT

A numerical measure of strength in the linear relationship between any two variables is known as correlation co-efficient. It is denoted by r is defined as;

$$r = \frac{\Sigma(x - \bar{x})\,(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \,.\, \Sigma\,(y - \bar{y})^2}}$$

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{nS_x S_y}$$

The correlation coefficient r always lies between -1 & $+1$.

If $r = +1$ (Perfect positive correlation)

If $r = -1$ (Perfect negative correlation)

If $r = 0$ (The two variables are independent)

EXAMPLE 5.4

Calculate correlation coefficient between x and y from the following data.

x	85	50	93	35	62	65	87	58	50	45
y	78	38	95	25	75	80	95	63	66	40

SOLUTION

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x-\bar{x})^2$	$(y - \bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
85	78	22	12.5	484	156.25	275.0
50	38	−13	−27.5	169	756.25	357.5
93	95	30	29.5	900	870.25	885.0
35	25	−28	−40.5	784	1640.25	1134.0
62	75	−1	9.5	1	90.25	9.50
65	80	2	14.5	4	210.25	29.0
87	95	24	29.5	576	810.25	708.0
58	63	−5	−2.5	25	6.25	12.5
50	66	−13	0.5	169	0.25	6.5
45	40	−18	−25.5	324	650.25	459.0
630	**655**	**0**	**0**	**34369**	**5250.50**	**3844.0**

Method 1

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 . \Sigma(y - \bar{y})^2}}$$

$$= \frac{3844.00}{\sqrt{(34369)(5250.50)}}$$

$$= 0.91$$

Therefore

$$\bar{x} = \frac{\Sigma x}{n} = \frac{630}{10} = 63.0$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{655}{10} = 65.5$$

Method 2

x	y	xy	x^2	y^2
85	78	6630	7225	6084
50	38	1900	2500	1444
93	95	8835	8649	9025
35	25	875	1225	625
62	75	4650	3844	5625
65	80	5200	4225	6400
87	95	8265	7569	9025
58	63	3654	3364	3969
50	66	3300	2500	4356
45	40	1800	2025	1600
630	655	45109	43126	48153

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][x\Sigma y^2 - (\Sigma y)^2]}}$$

$$= \frac{10(45109) - 630(655)}{\sqrt{[10(43125) - (630)^2][10(48153) - (655)^2]}}$$

$$= \frac{451090 - 412650}{\sqrt{(431260 - 396900)(481530 - 429025)}}$$

$$= \frac{38440}{\sqrt{(34360)(52505)}}$$

$$= \frac{38440}{42474.3664} = 0.91$$

5.11 PROPERTIES OF COEFFICIENT OF CORRELATION

The sample correlation coefficient has the following properties.

(i) The correlation coefficient is symmetrical with respect to variables x and y.

i.e. $r_{xy} = r_{yx}$

(ii) The correlation coefficient lies between −1 and +1.

i.e. $-1 \le r \le 1$

(iii) The correlation coefficient is independent of origin and scale. *i.e.* $r_{xy} = r_{yx}$

(iv) The correlation coefficient is the geometric mean of two regression coefficient.

i.e. $r = \sqrt{(b_{yx})(b_{xy})}$

(v) The correlation coefficient is a pure number. *i.e.* It has no unit.

(vi) For two independent random variables correlation coefficient is zero.

5.12 SHORT CUT METHOD FOR THE CALCULATION OF CORRELATION COEFFICIENT

Let $u = x - a,\ v = y - b$

Then $r_{uv} = \dfrac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{[n\Sigma u^2 - (\Sigma u)^2][n\Sigma v^2 - (\Sigma v)^2]}}$

EXAMPLE 5.5

Calculate the co-efficient of correlation between the variables x and y given below by using short cut method.

x	78	89	97	69	59	79	68	61
y	125	137	156	112	107	136	123	108

SOLUTION

Let $u = x - 69$ and $v = y - 112$

x	y	u=x−69	v = y − 112	u^2	v^2	uv
78	125	9	13	81	169	117
89	137	20	25	400	625	500
97	156	28	44	784	1936	1232
69	112	0	0	0	0	0
59	107	−10	−5	100	25	50
79	136	10	24	100	576	540
68	123	−1	11	1	121	−11
61	108	−8	−4	64	16	32
600	**1004**	**48**	**108**	**1530**	**3468**	**2160**

$$r_{uv} = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{[n\Sigma u^2 - (\Sigma u)^2]\,[n\Sigma v^2 - (\Sigma v)^2]}}$$

$$= \frac{8(2160) - 48(188)}{\sqrt{[8(1530) - (48)^2]\,[8(3468) - (188)^2]}}$$

$$= \frac{17280 - 5184}{\sqrt{(12240 - 2304)\,(27744 - 11664)}}$$

$$= \frac{12096}{\sqrt{(9936)\,(16080)}}$$

$$= \frac{12096}{12640.05} = 0.96$$

5.13 RANK CORRELATION

Sometimes the actual measurements of objects are not available. They are then arranged in order according to some characteristic of interest. Such an ordered arrangements is called a ranking and the order given to an object is called its rank. The correlation between two sets of rankings is called rank correlation.

The spearman rank correlation coefficient is given by

$$r_s = 1 - \frac{6\,\Sigma d^2}{n(n^2 - 1)}, \text{ where } d = x\text{-}y$$

EXAMPLE 5.6

Two Judges in a contest, who were asked to rank 8 candidates A, B, C, D, E, F, G, H is order of their preference. Find the coefficient of rank correlation.

First Judge	A	B	C	D	E	F	G	H
	5	2	8	1	4	5	3	7
Second Judge	4	5	7	3	2	8	1	6

We calculate the coefficient of rank correlation as follows.

x_i	y_i	$d = x_i - y_i$	d^2
5	4	1	1
2	5	–3	9
8	7	1	1
1	3	–2	4
4	2	2	4
6	8	–2	4
3	1	2	4
7	6	1	1
–	–	–	28

Then $$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(28)}{8(64-1)} = 0.67$$

This indicates that Judges agreed well in their choices.

5.14 RANK CORRELATION FOR TIED RANKS

The Spearman's formula applies only when no ties are present. When there are ties in ranks the ranks are adjusted by assigning the means of the ranks which they jointly occupy. For example if fifth, sixth and seventh largest values of variables are the same, we assign each the rank $\frac{5+6+7}{3} = 6$. In tied ranks, one of the following two methods is to be used:

First, for each tie, add a quantity $\frac{1}{12}(t^3 - t)$ to $\sum d^2$ before substituting the values in the Spearman's formula of rank correlation.

Second, use the product moment coefficient of correlation to find the correlation between the two sets of adjusted ranks.

EXAMPLE 5.7

Compute the coefficient of rank correlation for the followings ranks:

X	8	3	6.5	3	6.5	9	3	1	5
Y	8	9	6.5	2.5	4	5	6.5	1	2.5

SOLUTION

We observe that both the sets of ranks have 2,2 ties. In the first set, 3 is repeated 3 times and 6.5 two times. Similarly in the second set 2.5 and 6.5 repeated two times. We add the following $\frac{1}{12}(3^3-3)+\frac{1}{12}(2^3-2)+\frac{1}{12}(2^3-2)+\frac{1}{12}(2^3-2)=2+0.5+0.5+0.5=3.5$ quantities to Σd^2.

x	y	d = x-y	d^2
8	8	0	0
3	9	–6	36
6.5	6.5	0	0
3	2.5	0.5	0.25
6.5	4	2.5	6.25
9	5	4	16
3	6.5	–3.5	12.25
1	1	0	0
5	2.5	2.5	6.25
Total	–	**0**	**77**

hence $r_s = 1 - \frac{6(\Sigma d^2 + 3.5)}{n(n^2-1)}$

$= 1 - \frac{6(77+3.5)}{9(81-1)}$

$= 1 - \frac{483}{720} = 0.3291$

Alternative Method

Let us denote the ranks given by the first member by x and second member by y. The necessary calculations are given below.

x	y	x^2	y^2	xy
8	8	64	64	64
3	9	9	81	27
6.5	6.5	42.25	42.25	42.25
3	2.5	9	6.25	7.5
6.5	4	42.25	16	26
9	5	81	25	45
3	6.5	9	42.25	19.5
1	1	1	1	1
5	2.5	25	6.25	12.5
45	**45**	**282.5**	**284**	**244.75**

Hence the coefficient of rank correlation is

$$r = \frac{n\Sigma xy - \Sigma xy\,\Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\,[n\Sigma y^2 - (\Sigma y)^2]}}$$

$$= \frac{9(244.75) - (45)(45)}{\sqrt{[9(282.5) - (45)^2]\,[9(284) - (45)^2]}}$$

$$r = 0.3391$$

Both the methods gives almost the same answer.

SUMMARY

The formula and methods of computing the correlation and regression are given below:

APPLICATION	FORMULA
Range of Correlation	$-1 \le r \le 1$
CORRELATION	(i) $r = \dfrac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$
	(ii) $r = \dfrac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{[n\Sigma X^2 - (\Sigma X)^2][n\Sigma Y^2 - (\Sigma Y)^2}}$
	(iii) $r = \dfrac{S_{XY}}{S_x S_Y}$
	Where $S_{XY} = \dfrac{\Sigma XY}{n} - (\bar{X}\bar{Y})$
	$S_X = \sqrt{\dfrac{\Sigma X^2}{n} - \left(\dfrac{\Sigma X}{n}\right)^2}$
	$S_Y = \sqrt{\dfrac{\Sigma Y^2}{n} - \left(\dfrac{\Sigma Y}{n}\right)^2}$
Regression line Y on X	$y = a + bx$
	$a = \bar{y} - b\bar{X}$
	Where
	$b_{YX} = \dfrac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2}$
	Or $b_{YX} = \dfrac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$
	Or $b_{YX} = r\dfrac{S_y}{S_x}$

Regression line X on Y	$X = a + by$ Where $b_{XY} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(Y - \bar{Y})^2}$ $b_{XY} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma y^2 - (\Sigma y)^2}$ Or $b_{XY} = r\frac{S_x}{S_y}$ $a = \bar{X} - b\bar{y}$
Relation between regression & correlation	$r = \sqrt{(b_{YX})(b_{XY})}$
Standard error of estimate	$S_{yx} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}}$ Where $\Sigma(y - \hat{y})^2 = \Sigma y^2 - a\Sigma y - b\Sigma xy$
Coefficient of determination	$r^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$ Or $1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - y)^2}$ Or $\frac{a\Sigma y + b\Sigma xy - (\Sigma y)^2/n}{\Sigma y^2 - (\Sigma y)^2/n}$
Rank correlation	$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$

EXERCISES

5.1 **(a)** What do you meant by Regression?

(b) Write a short note on scatter diagram.

5.2 The following sample of 8 grade point averages and marks in matriculation was observed for students from a College.

Score	480	490	510	510	530	550	610	640
G.P.A.	2.7	2.9	3.3	2.9	3.1	3.0	3.2	3.7

Find the least square line. Estimate the mean GPA of students scoring 600 marks.

5.3 The following table shows the marks obtained by students in Math and Physics.

Math (x)	48	40	32	34	30	50	26	50	22	43
Physics (y)	76	56	40	50	34	70	56	68	40	57

(i) Construct scatter diagram for the above data.

(ii) Find the least square regression line of y on x and of x on y.

5.4 Find out coefficient of correlation and regression equations for the following data, x on y and y on x. Calculate the value of y when the value of x is 60.

x	10	15	20	25	30
y	15	17	21	23	26

5.5

x	80	82	86	91	83	85	89	96	93
y	145	140	130	124	133	127	120	110	116

Required

Calculate coefficient of correlation and also the line of regression y on x.

5.6 Fit a regression line of y on x from the following data.

x	5	7	6	12	17	19	20	29
y	22	14	11	9	9	8	6	2

5.7 **(a)** The correlation coefficient between two variables x and y is $r = 0.63$.

If $S_x = 1.50, S_y = 2.00, \bar{x} = 10, \bar{y} = 20$

Find the regression lines y on x and of x on y.

(b) If the equations of the least squares regression lines are:

$y = 20.8 - 0.219\,x$ (y on x)

$x = 16.2 - 0.785\,y$ (x on y)

Find the product moment coefficient of correlation.

5.8 The following data show the son's height and father's height.

Father's height (x)	59	61	63	65	67	69	71	73	75
Son's height (y)	64	66	67	67	68	69	70	72	72

Estimate the regression line $y = a + bx$ using $u = \frac{x - 67}{2}$ and $v = y - 68$.

Predict the mean height of sons whose fathers are 70 inches in height.

5.9 Fit the line of Regression of marks in Maths on marks in Economics from the following data:

Math	5	7	16	12	17	13	10	9
Eco	2	4	17	19	9	8	6	5

5.10 The following table gives the aptitude test score and productivity indices of 10 workers selected at random estimate.

Aptitude score (x)	60	62	65	70	72	48	53	73	65	82
Productivity index (y)	68	60	62	80	85	40	52	62	60	81

Calculate correlation coefficient between aptitude scores and productivity index.

5.11 A financial analyst has gathered the following data bout the relationship between income and investment in securities in respect of 8 randomly selected families.

Income	8	12	9	24	43	37	19	16
Percent Invested	36	25	33	15	28	19	20	20

Find the correlation coefficient and interpret it.

5.12 On the basis of the figured recorded below for supply and price find:

(a) correlation coefficient

(b) Both regression coefficients

Supply(x)	80	82	86	91	83	85	89	96	93
Price (y)	145	140	130	124	133	127	120	110	116

5.13 Find Regression Coefficients of the followings:

$\Sigma x = 17.6$, $\Sigma y = 32.8$, $\Sigma xy = 94.7$

$\Sigma x^2 = 49.64$, $\Sigma y^2 = 182$, $n = 8$

5.14

x	16	72	73	63	83	80	66	66	74	62
y	40	52	43	49	61	58	44	58	50	45

Required: Calculate Coefficient of correlation and comment on the answer.

5.15 Calculate coefficient of correlation between x and y. Find regression equation of y on x. Estimate value of y when x = 75.

x	79	89	97	69	59	79	68	61
y	125	137	156	112	107	136	123	104

5.16 Find out the regression equation for the following data. Calculate the value of y when the value of x is 22.

x	10	15	20	25	30
y	15	17	21	23	26

5.17 Find out:

(i) Regression equation x on y

(ii) Compute trend values and prove that $\Sigma (X - \hat{X}) = 0$

x	5	6	8	10	12	13	15	16	17
y	16	19	23	28	36	40	42	45	50

5.18 (a) Define correlation, write down the properties of correlation.

(b) Calculate coefficient of correlation.

X	1	2	3	4	5	6
Y	2	3	4	3	6	5

(c) The following data show the marks in economics and marks in statistics obtained by ten students.

Student	1	2	3	4	5	6	7	8	9	10
Eco (x)	78	36	96	25	75	82	90	62	65	39
Stat (y)	84	51	91	60	68	62	86	58	53	47

Compute the co-efficient of correlation.

5.19 (a) Calculate coefficient of correlation from the following data.

x	3	5	6	9	10	12	15	20	22	28
y	10	12	15	18	20	22	27	30	32	34

(b) Find the coefficient of correlation between x and y.

x	30	35	40	45	50	60	70	80	90
y	2	4	5	5	8	15	24	30	32

5.20 Given the following data.

x	20	25	30	35	40	45	50	55	60	65
y	90	85	75	65	95	35	20	15	15	5

(i) Calculate the coefficient of correlation between x and y.

(ii) Calculate correlation coefficient between u and v where

$$u = \frac{x - 45}{5} \text{ and } v = \frac{y - 35}{5}$$

5.21 Find the correlation coefficient between x and y, given.

x	5	12	14	16	18	21	22	23	25
y	11	16	15	20	17	19	25	24	21

5.22 (a) For a sample of 20 pairs of observations, we have

$\bar{x} = 2$, $\bar{y} = 8$, $\Sigma x^2 = 180$, $\Sigma y^2 = 3424$, $\Sigma xy = 604$

Calculate the coefficient of correlation.

(b) For a given set of data, we have

$n = 10$, $\Sigma(x - \bar{x})(y - \bar{y}) = 120$, $S_y = 8$

$\Sigma(x - \bar{x})^2 = 90$

Find coefficient of correlation.

5.23 For a set of 50 pairs of observations, the standard deviations of x and y are 4.5 and 3.5 respectively. If the sum of products of deviations of x and y values from their respective means be 420. Find the coefficient of correlation.

5.24 A computer while calculating the coefficient of correlation between two variables x and y from 25 pairs of observations obtained the following sums.

$\Sigma x = 125$, $\Sigma x^2 = 650$, $\Sigma y = 100$, $\Sigma y^2 = 460$

$\Sigma xy = 508$

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

X	Y
6	14
8	6

while the correct values were

x	Y
8	12
6	8

Obtain the correct value of the coefficient of correlation.

5.25 Calculate the coefficient of correlation between supply and demand from the following data.

Supply	400	200	700	100	500	300	600
Demand	60	30	70	10	40	20	50

5.26 The following table gives the aptitude test scores and productivity indices of 10 workers selected at random estimate.

Aptitude. Score (x)	60	62	65	70	72	48	53	73	65	82
Productivity. index (y)	68	60	62	80	85	40	52	62	60	81

Calculate correlation coefficient between aptitude test scores and productivity indices.

5.27 **(a)** What is meant by standard error of estimate?

(b) Explain coefficient of determination.

(c) Compute coefficient of determination and standard error of estimate for the following data.

Income (x) (0000)	10	20	30	40	50	60
Expenditure (y) (0000)	7	21	23	34	36	53

5.28 **(a)** What is rank correlation?

(b) Ten students got the following positions on the basis of their performance in two subjects.

Subject	1	2	3	4	5	6	7	8	9	10
Subject-I	1	5	2	6	4	8	3	7	10	9
Subject-II	3	6	2	7	5	8.	1	4	9	10

Calculate Spearman's rank correlation coefficient

5.29 (a) Find rank correlation coefficient for the following data:

a	4.7	2.9	6.4	2.5	4.9
b	8.6	5.4	6.2	4.8	8.3

(b) The following table shows the marks of six candidates in the subjects.

Candidate	A	B	C	D	E	F
Math (x)	38	62	56	42	59	48
Stat (y)	64	89	84	60	73	69

Calculate the coefficient of rank correlation.

5.30 SELECT THE CORRECT ANSWER:

(i) **The word regression was introduced by:**

(a) Karl Pearson (b) Galton

(c) R.A. Fisher (d) Marshall

(ii) **The diagram of a set of bivariate data is:**

(a) Histogram (b) Historigram

(c) Scatter diagram (d) Bar diagram

(iii) **If the value of regression Coefficient is zero, then two variables are**

(a) Dependent (b) Independent

(c) Both (a) and (b) (d) None of these

(iv) **The signs of regression coefficients (b_{yx} and b_{xy}) are always:**

(a) Different (b) Same

(c) Positive (d) Negative

(v) **The sum of deviations of observed values from regression line is:**

(a) Zero (b) Minimum

(c) Maximum (d) None

(vi) **The sum of square deviations of values from regression line is**

(a) Zero (b) Minimum

(c) Maximum (d) None of these

(vii) **If $y = 10.5 + 6.3\ x$, then the value of regression coefficient is:**

(a) 0 (b) 10.5

(c) 6.3 (d) x

(viii) If $y = 20 + 2.5x$, then the value of intercept is:

(a) 0 (b) 20
(c) 25 (d) None of these

(ix) The correlation coefficient lies between

(a) -1 and 0 (b) 0 and 1
(c) -1 and 1 (d) None of these

(x) The Correlation Coefficient is independent of

(a) Origin (b) Scale
(c) Both (a) and (b) (d) None of these

(xi) If one variable increase and the other variable also increase, then Correlation is

(a) Positive (b) Negative
(c) Perfect (d) None of these

(xii) If $u = x$ and $v = -x$, then correlation γ_{uv} is:

(a) 1 (b) -1
(c) 0 (d) 0.5

(xiii) If both the variables are independent, then correlation will be

(a) No (b) Positive
(c) Negative (d) None of these

(xiv) The Correlation Coefficient is the geometric mean of:

(a) Variables (b) Regression Coefficients
(c) Coefficient of determination
(d) None of these

ANSWERS 5.30

(i)	(b)	(ii)	(c)	(iii)	(b)	(iv)	(b)	(v)	(a)
(vi)	(b)	(vii)	(c)	(viii)	(b)	(ix)	(c)	(x)	(c)
(xi)	(a)	(xii)	(b)	(xiii)	(a)	(xiv)	(b)		

SHORT QUESTION AND ANSWERS

1. Define Regression.

Ans: The dependence of one variable on other is called regression e.g. If we want to estimate the heights of children on the basis of their ages, the height would be the dependent variable and the ages would be the independent variable.

2. What is the standard form of least square regression line?

Ans: $y = a + bx$

where y is dependent variable, x is independent variable, while a and b a constants.

3. Define Scatter diagram.

Ans: The pair of independent-dependent observation as a point on the graph pape using the x-axis for independent variable and the y-axis for the dependen variable. Such a diagram is called scatter diagram.

4. Define Correlation.

Ans: The degree of interdependence between two variables is called correlation e.g ages of husbands and ages of wives at the time of their marriage, demand and supply of a commodity etc.

5. What do you meant by Positive and Negative Correlation?

Ans: **Positive Correlation:**

If both the variables tend to increase or decrease together, the correlation is said to be direct or positive.

Negative Correlation:

If one variable tends to increase as the other variable decreases the correlation is said to be negative or inverse.

6. Write down the properties of correlation.

Ans: (i) Correlation coefficient is symmetrical with respect to be variables x and y.

i.e. $r_{xy} = r_{yx}$

(ii) The correlation coefficient lies between -1 and $+1$ *i.e.*

$-1 \leq r \leq 1$

(iii) The correlation coefficient is independent of origin and scale. *i.e.*

$r_{xy} = r_{uv}$

(iv) The correlation is the geometric mean of two regression coefficients. *i.e.*

$$\sqrt{(b_{yx})\,(b_{xy})}$$

7. What are the properties of Regression?

Ans: (i) Least square regression line always passes through mean $(\overline{x}, \overline{y})$.

(ii) $\Sigma(y - \hat{y}) = 0$

(iii) $\Sigma(y - \hat{y})^2 =$ minimum