Chapter-2
Correlation Analysis

2.1 Simple Correlation Analysis

In a bivariate distribution, if the change in one variable is accompanied by a change in the other variable in such a way that an increase in one variable results in an increase or decrease in the other, then the two variables are said to be correlated. For example, income and expenditure, heights and weights of students in a class, price and demand of certain commodities.

If the increase (or decrease) in one variable results in a corresponding increase (or decrease) in the other, correlation is said to be direct or positive. But if the increase (or decrease) in one variable results in a corresponding decrease (or increase), in the other, correlation is said to be negative. If two variables vary in such a way that their ratio is always constant, then the correlation is said to be perfect.

When we plot the corresponding values of two variables, taking one on X-axis and the other along Y-axis, it shows a collection of dots. This collection of dots is called a dot diagram or a scatter diagram.

If all the plotted points lie in a straight line and show an upward trend, then the correlation is perfect positive. If all the plotted points lie in a straight line and show a downward trend, then the correlation is perfect negative.

If the plotted points are not on a straight line but seem to be scattered around a straight line, the variables are correlated. Closer the scatter of points around a line, higher is the degree of correlation. If the plotted points are not clustered around a straight line but are widely scattered over the diagram, then there is a very low degree of correlation between the variables. If the plotted points show no trend at all, then the variables are independent and are not correlated.

2.2 Uses of Correlation

Correlation studies, if used appropriately, are important to science. In an experiment, it is important to determine a correlation because then a hypothesis can be proven or disproven. Without correlation, a theory is just a theory. There is no real life facts or proof.

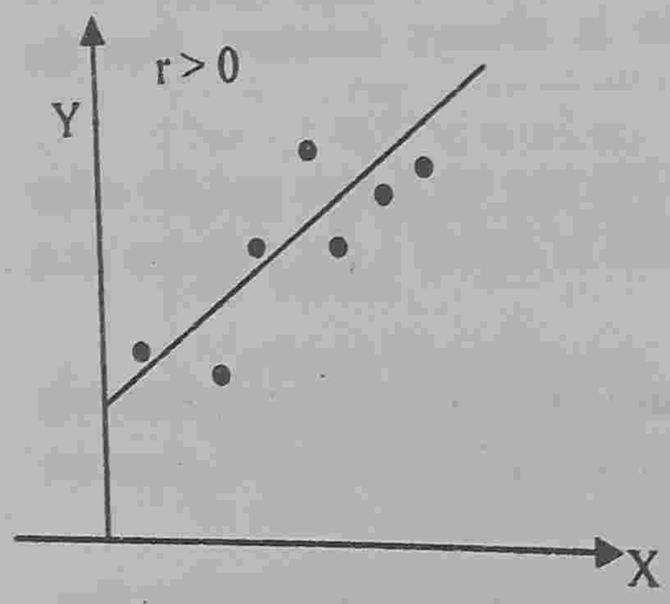
Once correlation is known it can be used to make predictions. When we know a score on one measure we can make a more accurate prediction of another measure that is highly related to it. The stronger the relationship between/among variables the more accurate the prediction is.

Correlation and regression analysis can help business to investigate the determinants of key variables such as their sales.

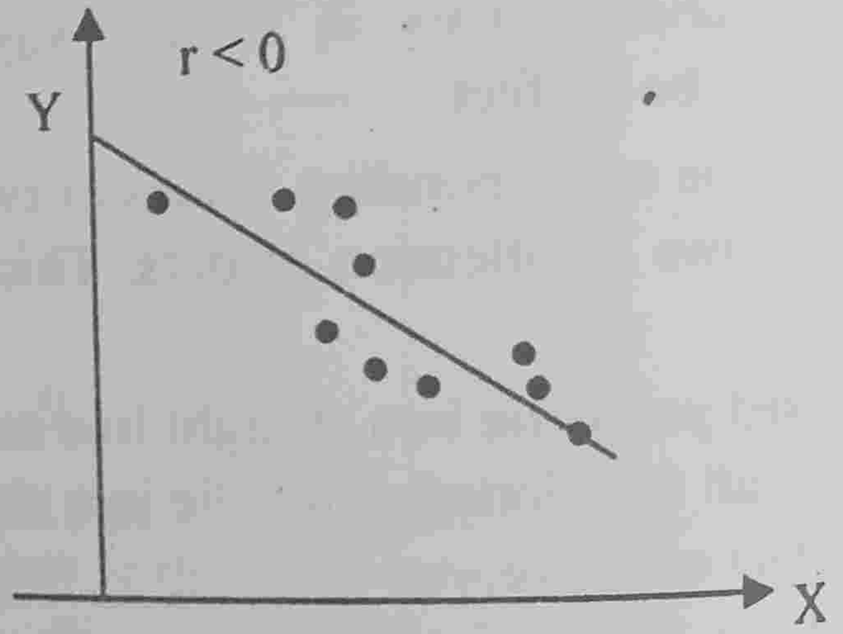
1. Most of the variables show some kind of relationship. For instance, there is relationship between price and supply, income and expenditure etc. With the help of correlation analysis we can measure in one figure the degree of relationship.
2. Once we know that two variables are closely related, we can estimate the value of one variable given the value of another. This is known with the help of regression.
3. Correlation analysis contributes to the understanding of economic behavior, aids in locating the critically important variables on which others depend.
4. Progressive development in the methods of science and philosophy has been characterized by increase in the knowledge of relationship.
5. The effect of correlation is to reduce the range of uncertainty. The prediction based on correlation analysis is likely to be more variable and near to reality.

2.3 Different Types of Association between Variables

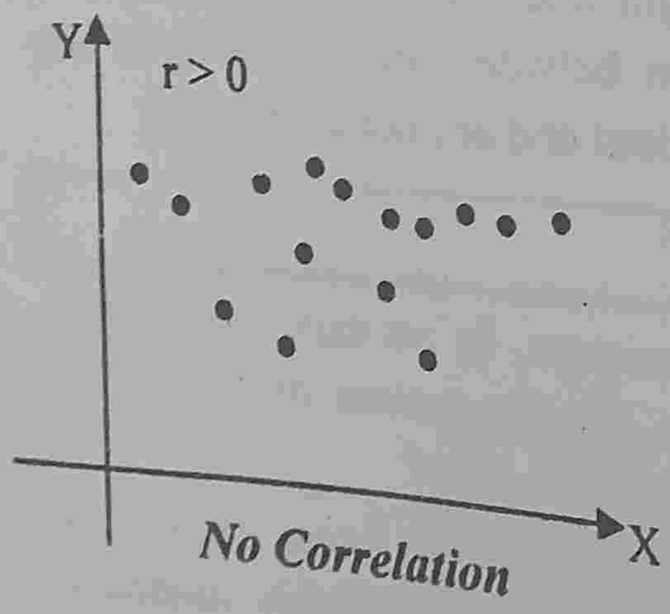
A *scatter diagram* shows the relationship between two variables. If the points lie more or less around a straight line or around a curve, there is *high correlation*. By studying the scatter diagram, we can study whether there is any association between the variables or not and upto what extent, if it does.



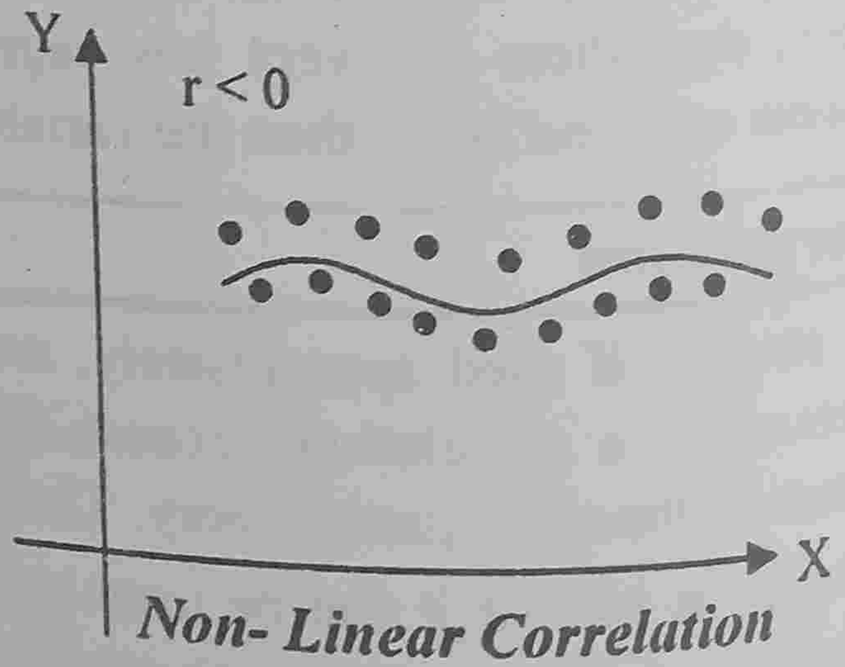
Positive Linear Correlation



Negative Linear Correlation



No Correlation



Non- Linear Correlation

If the correlation involves only two variables, then it is called as ***simple correlation***. If the correlation involves more than two variables, then it is called as ***multiple correlation***.

2.4 Karl Pearson's Coefficient of Correlation

The correlation coefficient $r(x, y)$ between two variables x and y is given by

$$r(x,y)=\frac{\text{cov}(x,y)}{\sqrt{\text{var}iance(x)}\sqrt{\text{var}iance(y)}}=\frac{\text{cov}(x,y)}{\sigma_x\sigma_y}$$

$r(x, y)$ is also denoted by $\rho(x, y)$ or r_{xy} or simply by r.

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

or

$$r(x,y)=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-(\sum_{i=1}^{n}y_i)^2}}$$

$$r(x,y)=\frac{\sum_{i=1}^{n}x_iy_i-\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n}}{\sqrt{\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}\sqrt{\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}}}$$

If the values of x_i and y_i's are large or involve fractions, then define

$$u_i=\frac{x_i-a}{h} \text{ and } v_i=\frac{y_i-b}{k}$$

where a and b arc assumed means of x-series and y-series respectively, h and k are constants. This property is known as change of origin and scale. Correlation coefficient is independent of change of origin and scale. In this $r(x, y)$ is given by the formula:

$$r(x,y)=r(u,v)=\frac{n\sum_{i=1}^{n}u_iv_i-\sum_{i=1}^{n}u_i\sum_{i=1}^{n}v_i}{\sqrt{n\sum_{i=1}^{n}u_i^2-(\sum_{i=1}^{n}u_i)^2}\sqrt{n\sum_{i=1}^{n}v_i^2-(\sum_{i=1}^{n}v_i)^2}}$$

2.5 Properties of Correlation Coefficient

1. If $r=0$, *then there is no correlation between x and y.*
2. If $0<r\leq 1$, *then there is positive correlation between x and y*
3. *The value of r always lies between* -1 *and* $+1$, *i.e.,* $-1\leq r\leq +1$
4. If $-1\leq r<0$, *then there is negative correlation between x and y.*
5. If $r=1$, *then there is perfect positive correlation between x and y.*
6. If $r=-1$, *then there is perfect negative correlation between x and y.*
7. *The correlation is said to be linear if the scatter points near a line of best fit.*
8. *In case of curvilinear relationship, the scatter points do not lie around a straight line but around a curve.*

2.6 Spearman's Rank Correlation

Many a times, in a bivariate distribution the two characteristics of an individual instead expressed in numbers, are expressed in terms of ranks. Sometimes we have to deal with problems in which data cannot be measured quantitatively but qualitative assessment is possible e.g., beauty, honesty, morality etc. In such cases we assign ranks to the individuals possessing these attributes or characteristics. Ranks are assigned starting from lowest to highest value or from highest to lowest value. The coefficient of rank correlation r is given by

$$r(x,y)=1-\frac{6\sum_{i=1}^{n}d_i^{\,2}}{n(n^2-1)}$$

where $r(x,y)$ denotes the coefficient of rank correlation d_i is the difference of corresponding rank and n is the number of pairs of observations.

2.7 Repeated Rank Correlation

In some cases it is sometimes necessary to assign equal ranks to two or more individuals. In such cases, it is customary to assign each individual an average rank. When two or more items have same values, it is difficult to assign different ranks to them. In such a case average ranks (same ranks) are assigned to each of them. In this case the following modified formula is used to find the rank coefficient of correlation:

$$r(x,y)=1-\frac{6\left[\sum_{i=1}^{n}d_i^{\,2}+\frac{m_1(m_1^2-1)}{12}+\frac{m_2(m_2^2-1)}{12}+\ldots\ldots\right]}{n(n^2-1)}$$

where $m_1, m_2, \ldots$ are the number of times the item is repeated in a series.

2.8 Properties of Rank Correlation

1. If $r = 0$, *then there is no correlation between the two characteristics.*
2. If $r = 1$, *then there is perfect positive correlation between the two characteristics.*
3. If $r = -1$, *then there is perfect negative correlation between the two characteristics.*
4. If $r > 0$, *then high rank in one characteristics corresponds to high rank in the other and then low rank in one characteristics corresponds to low rank in the other.*
5. If $r < 0$, *then high rank in one characteristics corresponds to low rank in the other and then low rank in one characteristics corresponds to high rank in the other.*

Illustrative Examples

Example 1. Find the coefficient of correlation for the following data:

$$n = 10,\ \sum x = 50,\ \sum y = -30,\ \sum x^2 = 290,\ \sum y^2 = 300,\ \sum xy = -115$$

Solution

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

$$= \frac{10 \times (-115) - (50) \times (-30)}{\sqrt{10 \times 290 - (50)^2}\sqrt{10 \times 300 - (-30)^2}}$$

$$= \frac{-1150 + 1500}{\sqrt{400}\sqrt{2100}}$$

$$= \frac{350}{200\sqrt{21}} = 0.3819$$

Example 2. Find the coefficient of correlation if $cov(x, y) = 5520$, $V(x) = 9685$, $V(y) = 3420$

Solution

$$covariance(x, y) = 5520$$

$$\sigma_x = \sqrt{9685},\ \sigma_y = 3420$$

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

$$= \frac{5520}{\sqrt{9685} \times \sqrt{3420}}$$

$$= \frac{5520}{98.41 \times 58.48}$$

$$= \frac{5520}{5755.23} = 0.959$$

Example 3. Karl Pearson's coefficient of correlation between two variables X and Y is 0.35. their covariance is + 8. If the variance of x is 9, then find the standard deviation of Y series.

Solution Let the variance of Y be denoted by V(x), and Karl Pearson's coefficient of correlation between two variables X and Y is given by

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

$$0.35=\frac{8}{\sqrt{9}\times\sqrt{\text{var}(Y)}}$$

$$\sqrt{\text{var}(Y)}=\frac{8}{0.35\times 3}=\frac{8}{1.05}$$

$$\sqrt{\text{var}(Y)}=7.62$$

$$\text{var}(Y)=58.06$$

Example 4. A computer while calculating correlation coefficient between two variables x and from 25 pairs of observations obtained the following results:

$n=25, \sum x=125, \sum y=100, \sum x^2=650, \sum y^2=460, \sum xy=508.$

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

x	6	8
y	14	6

while the correct values were

x	8	6
y	12	8

Obtain the correct value of the correlation coefficient.

Solution Corrected $\sum x$ = given $\sum x$ - (sum of incorrect values) + (sum of correct values)

Corrected $\sum x = 125 - (6+8) + (8+6) = 125$

Corrected $\sum y = 100 - (14+6) + (12+8) = 100$

Corrected $\sum x^2 = 650 - (6^2+8^2) + (8^2+6^2) = 650$

Corrected $\sum y^2 = 460 - (14^2+6^2) + (12^2+8^2) = 436$

Corrected $\sum xy = 508 - (6\times 14 + 8\times 6) + (8\times 12 + 6\times 8) = 520$

$$\text{Corrected } r(x,y) = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

$$= \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2}\sqrt{25 \times 436 - (100)^2}}$$

$$= \frac{500}{25 \times 30} = 0.66$$

Example 5. Calculate the Karl Pearson's coefficient of correlation for the following data:

x	2	4	6	8	10
y	20	12	18	10	40

Solution

x	y	x^2	y^2	xy
2	20	4	400	40
4	12	16	144	48
6	18	36	324	108
8	10	64	100	80
10	40	100	1600	400
$\sum x = 30$	$\sum y = 100$	$\sum x^2 = 220$	$\sum y^2 = 2568$	$\sum xy = 676$

Here, $n = 5$

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

$$= \frac{5 \times 676 - 30 \times 100}{\sqrt{5 \times 220 - (30)^2}\sqrt{5 \times 2568 - (100)^2}}$$

$$= \frac{3380 - 3000}{\sqrt{1100 - 900}\sqrt{12840 - 10000}}$$

$$= \frac{380}{\sqrt{200}\sqrt{2840}} = 0.5042$$

Example 6. Calculate the Karl Pearson's coefficient of correlation between x and y for the following data:

x	150	153	154	155	157	160	163	164
y	65	66	67	70	68	53	70	63

Solution

x	y	$u_i = x_i - 155$	$v_i = y_i - 68$	u^2	v^2	uv
150	65	-5	-3	25	9	15
153	66	-2	-2	4	4	4
154	67	-1	-1	1	1	1
155	70	0	2	0	4	0
157	68	2	0	4	0	0
160	53	5	-15	25	225	-75
163	70	8	2	64	4	16
164	63	9	-5	81	25	-45
Total		16	-22	204	272	-84

Here, $n = 8$

$$r(x,y) = r(u,v) = \frac{n\sum uv - \sum u \sum v}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}}$$

$$= \frac{8\times(-84) - 16\times(-22)}{\sqrt{8\times 204 - (16)^2}\sqrt{8\times 272 - (-22)^2}}$$

$$= \frac{-672 + 352}{\sqrt{1376}\sqrt{1692}} = -0.2097$$

Example 7. Calculate the rank correlation coefficient for the following data:

Student	I	II	III	IV	V	VI	VII	VIII	IX	X
Rank in Phys.	9	10	6	5	7	2	4	8	1	3
Rank in Stats.	1	2	3	4	5	6	7	8	9	10

Solution Here the ranks are given and $n = 10$

Student	R_1	R_2	$d = R_1 - R_2$	d^2
I	9	1	8	64
II	10	2	8	64
III	6	3	3	9
IV	5	4	1	1
V	7	5	2	4
VI	2	6	-4	16
VII	4	7	-3	9
VIII	8	8	0	0
IX	1	9	-8	64
X	3	10	-7	49
				$\sum d^2 = 280$

The rank correlation coefficient is given as:

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

$$= 1 - \frac{6 \times 280}{10(10^2-1)}$$

$$= 1 - \frac{1680}{990}$$

$$= 1 - 1.697 = -0.697$$

Example 8. Calculate the rank correlation coefficient from the following data:

X	20	24	16	36	44	48	70
Y	30	20	12	52	34	26	20

Solution Here, the values of x and y are given and n =10. First find the ranks and then differences.

X	Y	$d_x = X - 40$	$d_y = Y - 28$	d_x^2	d_y^2	$d_x d_y$
20	30	-20	2	400	4	-40
24	20	-16	-8	256	64	128
16	12	-24	-16	576	256	384
36	52	-4	24	16	576	96
44	34	4	6	16	36	24
48	26	8	-2	64	4	-16
70	20	30	-8	900	64	-240
258	194	-22	-2	2228	1004	144

In series x, 16 is repeated 3 times, hence $m_1 = 3$. In series y, 13 and 6 repeated 2 times each, hence $m_2 = 2$ and $m_3 = 2$.

Therefore, the correlation coefficient for repeated rank is given by

$$r(x,y) = \frac{n\sum d_x d_y - \left(\sum d_x\right)\left(\sum d_y\right)}{\sqrt{n\sum d_x^2 - \left(\sum d_x\right)^2}\sqrt{n\sum d_y^2 - \left(\sum d_y\right)^2}}$$

$$= \frac{7 \times 144 - (-22) \times (-2)}{\sqrt{7 \times 2228 - (-22)^2}\sqrt{7 \times 1004 - (-2)^2}}$$

$$= \frac{1008 - 44}{\sqrt{15596 - 484}\sqrt{7028 - 4}}$$

$$= \frac{964}{\sqrt{15112}\sqrt{7024}}$$

$$= \frac{964}{122.93 \times 83.81} = 0.0936$$

Example 9. Calculate the coefficient of correlation for the following data:

x	9	12	8	13	7	10	11
y	8	14	6	12	3	9	11

Solution

x	y	x^2	y^2	xy
9	8	81	64	72
12	14	144	196	168
8	6	64	36	48
13	12	169	144	156
7	3	49	9	21
10	9	100	81	90
11	11	121	121	121
$\sum x = 70$	$\sum y = 63$	$\sum x^2 = 728$	$\sum y^2 = 651$	$\sum xy = 676$

Here, $n = 7$

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

$$= \frac{676 - \frac{70 \times 63}{7}}{\sqrt{728 - \frac{70^2}{7}} \sqrt{651 - \frac{63^2}{7}}}$$

$$= \frac{676 - 630}{\sqrt{728 - 700}\sqrt{651 - 567}}$$

$$= \frac{46}{\sqrt{28}\sqrt{84}}$$

$$= \frac{47}{5.29 \times 9.17} = \frac{47}{48.51} = 0.97$$

Example 10. Ten competitors in a beauty contest are ranked by three judges in the following order:

I^st Judge	I	II	III	IV	V	VI	VII	VII	IX	X
II^nd Judge	3	5	8	4	7	10	2	1	6	9
III^d Judge	6	4	9	8	1	2	3	10	5	7

ing the rank correlation method, discuss which pair of judges has the nearest approach to common taste in beauty?

Solution Let R_1, R_2 and R_3 be the ranks given by three judges.

R_1	R_2	R_3	$d_{12} = R_1-R_2$	$d_{13} = R_1-R_3$	$d_{23} = R_2-R_3$	d_{12}^2	d_{13}^2	d_{23}^2
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	-9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	1	2	1	1	4
						200	60	214

Here, $n = 10$

Rank correlation coefficient between first and second judges

$$r_{12} = 1 - \frac{6\sum d_{12}^2}{n(n^2-1)} = 1 - \frac{6\times 200}{10(10^2-1)} = 1 - \frac{40}{33} = -0.212$$

$$= 1 - \frac{6\times 200}{10(10^2-1)}$$

$$r_{12} = 1 - \frac{40}{33} = -0.212$$

Rank correlation coefficient between first and third judges

$$r_{13} = 1 - \frac{6\sum d_{13}^2}{n(n^2-1)}$$

$$= 1 - \frac{6\times 60}{10(10^2-1)}$$

$$r_{13} = 1 - \frac{4}{11} = 0.636$$

Rank correlation coefficient between second and third judges

$$r_{23} = 1 - \frac{6\sum d_{23}^{\ 2}}{n(n^2-1)} = 1 - \frac{6\times 214}{10(10^2-1)} = 1 - \frac{214}{165} = -0.297$$

$$= 1 - \frac{6\times 214}{10(10^2-1)}$$

$$r_{23} = 1 - \frac{214}{165} = -0.297$$

Since r_{13} is maximum, therefore, the pair of judges first and third has the approach to common tastes in beauty.

Example 11. The marks obtained by 9 students in two subjects A and B are given below:

Marks in A	35	23	47	17	10	43	9	6	28
Marks in B	30	33	45	23	8	49	12	4	31

Compute the rank correlation coefficient.

Solution Here the marks are given. First find the ranks and then differences.

Marks in A (X)	*Marks in B (Y)*	*Ranks in X (R_1)*	*Ranks in Y (R_2)*	*$d = R_1 - R_2$*	*d^2*
35	30	3	5	-2	4
23	33	5	3	2	4
47	45	1	2	-1	1
17	23	6	6	0	0
10	8	7	8	-1	1
43	49	2	1	1	1
9	12	8	7	1	1
6	4	9	9	0	0
28	31	4	4	0	0
					$\sum d^2 = 12$

Here, $n = 9$

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

$$=1-\frac{6\times 12}{9(9^2-1)}$$

$$=1-\frac{72}{720}=1-0.1=0.9$$

Example 12. Calculate the rank correlation coefficient from the following data:

X	48	33	40	9	16	16	65	24	16	57
Y	13	13	24	6	15	4	20	9	6	19

Solution Here, the values of x and y are given and n=10. First find the ranks and then differences.

X	Y	*Ranks in* $X(R_1)$	*Ranks in* $Y(R_2)$	$d=R_1-R_2$	d^2
48	13	3	5.5	-2.5	6.25
33	13	5	5.5	-0.5	0.25
40	24	4	1	3	9.00
9	6	10	8.5	1.5	2.25
16	15	8	4	4	16.00
16	4	8	10	-2	4.00
65	20	1	2	-1	1.00
24	9	6	7	-1	1.00
16	6	8	8.5	-0.5	0.25
57	19	2	3	-1	1.00
					$\sum d^2=41$

In series x, 16 is repeated 3 times, hence $m_1=3$. In series y, 13 and 6 are repeated 2 times each, hence $m_2=2$ and $m_3=2$.

Therefore, the correlation coefficient for repeated rank is given by

$$r(x,y)=1-\frac{6\left[\sum_{i=1}^{n}d_i^2+\frac{m_1(m_1^2-1)}{12}+\frac{m_2(m_2^2-1)}{12}+\frac{m_3(m_3^2-1)}{12}\right]}{n(n^2-1)}$$

$$=1-\frac{6\left[41+\frac{3(3^2-1)}{12}+\frac{2(2^2-1)}{12}+\frac{2(2^2-1)}{12}\right]}{10(10^2-1)}$$

$$= 1 - \frac{6[41 + 2 + 0.5 + 0.5]}{990}$$

$$= 1 - \frac{264}{990} = 0.733$$

Example 13. The following table gives the scores obtained by 11 students in Statistics and Mathematics. Calculate the Spearman's rank correlation coefficient.

Statistics Scores	40	46	54	60	70	80	82	85	85	90
English Scores	45	45	50	43	40	75	55	72	65	42

Solution Here, the values of x and y are given and n=10. First find the ranks and then the differences.

Statistics Scores (x)	***English Scores (y)***	***Ranks of x (R_1)***	***Ranks of y (R_2)***	***$d = R_1 - R_2$***
40	45	11	7.5	3.5
46	45	10	7.5	2.5
54	50	9	6	3
60	43	8	9	-1
70	40	7	11	-4
80	75	6	1	5
82	55	5	5	0
85	72	3.5	2	1.5
85	65	3.5	4	-0.5
90	42	2	10	-8
95	70	1	3	-2

In series x, 85 is repeated 2 times, hence $m_1 = 2$. In series y, 45 is repeated 2 times, hence $m_2 = 2$. Therefore, the correlation coefficient for repeated rank is

$$r(x,y) = 1 - \frac{6\left[\sum_{i=1}^{n} d_i^2 + \frac{m_1(m_1^2 - 1)}{12} + \frac{m_2(m_2^2 - 1)}{12}\right]}{n(n^2 - 1)}$$

$$= 1 - \frac{6\left[140 + \frac{2(2^2-1)}{12} + \frac{2(2^2-1)}{12}\right]}{11(11^2-1)}$$

$$= 1 - \frac{6[140 + 0.5 + 0.5]}{1320}$$

$$= 1 - \frac{846}{1320} = 0.359$$

xample 14. Following are the scores of ten students in a class and their IQ:

Students	1	2	3	4	5	6	7	8	9	10
Scores	35	40	25	55	85	90	65	55	45	50
I.Q.	100	100	110	140	150	130	100	120	140	110

Calculate the Spearman's rank correlation coefficient.

olution Here, the values of x and y are given and n=10. First find the ranks and then differences.

Student	***Score***	***I.Q.***	***Ranks for Scores (R_1)***	***Ranks for I.Q. (R_2)***	***$d = R_1 - R_2$***	***d^2***
1	35	100	9	9	0	0.00
2	40	100	8	9	-1	1.00
3	25	110	10	6.5	3.5	12.25
4	55	140	4.5	2.5	2	4.00
5	85	150	2	1	1	1.00
6	90	130	1	4	-3	9.00
7	65	100	3	9	-6	36.00
8	55	120	4.5	5	-0.5	0.25
9	45	140	7	2.5	4.5	20.25
10	50	110	6	6.5	-0.5	0.25
						84

Since in scores, 55 is repeated 2 times, thus $m_1 = 2$.

in I.Q., 140 and 110 repeated 2 times each, thus $m_2 = 2$ and $m_3 = 2$,

and 100 is repeated 3 times, hence $m_4 = 3$.

Therefore, the correlation coefficient for repeated rank is given by

$$r(x,y)=1-\frac{6\left[\sum_{i=1}^{n}d_i^2+\frac{m_1(m_1^2-1)}{12}+\frac{m_2(m_2^2-1)}{12}+\frac{m_3(m_3^2-1)}{12}+\frac{m_4(m_4^2-1)}{12}\right]}{n(n^2-1)}$$

$$=1-\frac{6\left[84+\frac{2(2^2-1)}{12}+\frac{2(2^2-1)}{12}+\frac{2(2^2-1)}{12}+\frac{3(3^2-1)}{12}\right]}{10(10^2-1)}$$

$$=1-\frac{6[84+0.5+0.5+0.5+2.0]}{990}$$

$$=1-\frac{525}{990}=0.47$$

EXERCISE

1. If the number of observation is 10 and correlation coefficient is 0.9, then significance of correlation.

   Ans. Highly sig

2. Karl Pearson's coefficient of correlation between two variables X and Y is 0.6, covariance is + 7. If the variance of x is 9, then find the variance of Y series.

3. If the number of observation is 100 and correlation coefficient is 0.4, then significance of correlation.

   Ans. Sig

4. If the number of observation is 10 and correlation coefficient is 0.4, then significance of correlation.

   Ans. Not sig

5. If the number of observation is 100 and correlation coefficient is 0.9, then significance of correlation.

   Ans. Highly sig

6. Find the coefficient of correlation if $\text{cov}(x,y)=5520,\ V(x)=9685,\ V(y)=3429$

7. A student calculates the value of correlation coefficient as 0.7 when the number of observations is 25. Find the limits within which r lies for another sample from the population.

   Ans. 0.76 to 0.

8. Find the coefficient of correlation if $\sum(y-15)^2=215, \sum(x-10)^2=180, \sum$ $n=10, \sum x=140$ $\sum(x-10)(y-15)=60$

9. Find the coefficient of correlation for the following data:

$$n=13, \sum x=117, \sum y=260, \sum x^2=1313, \sum y^2=6580, \sum xy=2827.$$

10. The coefficient of rank correlation between marks in Statistics and marks in Mathematics obtained by certain group of students is 0.8. If the sum of squares of the differences in marks is 33, find the number of students in the group.

Ans. **10**

11. The correlation coefficient between x and y for 20 items is 0.3. Mean of x is 15 and mean of y is 20 while standard deviations for x and y are 4 and 5 respectively. At the time of calculations one item 27 has wrongly been taken as 17 in case of x series and 35 instead of 30 in case of y series. Find the corrected coefficient of correlation.

Ans. **5. 0.515**

12. The coefficient of rank correlation of the marks obtained by 10 students in Statistics and Mathematics was found to be 0.5. It was then detected that the difference in ranks in the two subjects for one particular student was wrongly taken to be 3 in place of 7. What should be the corrected rank correlation coefficient?

Ans. **0.2576**

13. Calculate the Karl Pearson's correlation coefficient from the following data:

x	80	90	100	110	120	130	140	150	160
y	15	19	16	19	17	18	16	18	15

Ans. **-0.11547**

14. Calculate the Karl Pearson's correlation coefficient between height of father and height of son from the following data:

Height of father (in inches)	64	65	66	67	68	69	70
Height of son (in inches)	66	67	65	68	70	68	72

Ans. **0.81**

15. Calculate the Karl Pearson's correlation coefficient from the following data using 20 as the working mean for price and 70 as the working mean for demand:

Price	14	16	17	18	19	20	21	22	23
Demand	84	78	70	75	66	67	62	58	60

Ans. **-0.9542**

16. Calculate the correlation coefficient from the following results:

$$n=10, \sum x=140, \sum y=150, \sum (x-10)^2=180, \quad \sum (y-15)^2=215, \sum (x-10)(y-15)=60.$$

Ans. **0.9151**

17. Find the Pearson's coefficient of correlation for the following data.

x	40	42	46	48	50
y	10	12	15	23	27

18. Calculate the :Pearson's coefficient of correlation for the following data.

Price (₹)	10	12	15	14
Demand (tonnes)	40	41	48	60

19. Calculate the coefficient of rank correlation from the following data:

x	10	12	8	15	20	25
y	15	10	6	25	16	12

20. Calculate the coefficient of rank correlation from the following data:

x	4	6	8	10	12	14	16
y	10	15	20	25	30	35	40

21. In a beauty contest two judges rank the ten competitors in the following order:

Competitors	A	B	C	D	E	F	G	H	I
Judge I	6	4	3	1	2	7	9	8	10
Judge II	4	1	6	7	5	8	10	9	3

Determine if the two judges have the same taste in beauty?

Ans. **Yes, the two judges have the same taste in**

22. Calculate the coefficient of rank correlation from the following data:

x	68	64	75	50	64	80	75	40	55
y	62	58	68	45	81	60	68	48	50

23. The age of newly married couples are given as follows:

Husbands									
Wives	30	28	27	22	24	28	21	25	22
	25	26	25	21	22	23	20	23	22

24. Calculate the coefficient of rank correlation from the following data:

x	13	11	8	9	12	10	7
y	12	11	6	8	14	9	3

***Ans.* 0.95**

25. Calculate the coefficient of rank correlation from the following data:

x	20	25	35	40	45	30
y	16	10	20	5	10	8

***Ans.* -0.32**

26. Calculate the coefficient of rank correlation from the following data relating to price and supply of a commodity:

Price (₹)	11	12	13	14	15	16	17	18	19	20
Supply (Kg)	16	10	20	5	10	8	16	10	20	5

***Ans.* -0.96**

27. The data on price and quantity purchased relating to an item for 5 months are given as follows:

Price (₹)	48	33	40	9	16	16	65	25	15	57
Quantity (Kg)	13	13	24	6	15	4	20	9	6	19

Calculate the coefficient of rank correlation

***Ans.* 0.754, there is high degree of positive correlation.**

28. Two housewives were asked to express their preferences to different kinds of washing powders. The ranks assigned by them were:

Washing Powder	A	B	C	D	E	F	G	H	I	J
X	1	2	4	3	7	8	6	5	9	10
Y	2	3	5	4	7	6	1	10	8	9

***Ans.* 0.64**

29. Calculate the Pearson's coefficient of correlation for the following data.

Price (₹)	22	24	26	28	30	32	34	36	38	40
Demand (Kg)	60	58	58	50	48	48	42	36	32	9

***Ans.* 0.9674**

30. State whether there is any correlation between demand for rice and its price from the following data:

Demand (tons)	470	510	560	620	600	480	490	520	550
Price (₹)	15	16	17	20	19	19	20	25	27

31. The height of fathers and sons are given below:

Fathers (inches)	73	71	69	68	67	67	66
Sons (inches)	70	69	70	72	68	64	68

Calculate the Karl Pearson's coefficient correlation

32. The following entries were permitted for baby show and the ranks were given by the judges as under:

Entries	A	B	C	D	E	F	G	H	IC	J	K
I^st Judge	8	2	1	9	3	12	11	4	10	6	5
II^nd Judge	4	1	3	11	2	12	10	5	9	7	8

Calculate the rank coefficient correlation

33. The marks of the same 15 students in two different subjects X and Y are given as:
(15, 13), (14, 12), (13, 14), (12, 5), (11, 9), (10, 15), (9, 11),
(8, 1), (7, 3), (6, 8), (5, 4), (4, 6), (3, 2), (2, 7), (1, 10),
Use Spearman's formula to find rank correlation coefficient.

2.9 Regression Analysis

Correlation only indicates the degree and direction of relationship between two variables but does not necessary gives the cause and effect of the relationship. Regression analysis also helps to establish the nature of relationship between the variables so that one can predict or estimate the value of one variable for the given value of the other variable. Regression measures the nature and the degree of correlation. Correlation and regression analysis are constructed under different assumptions and it is not always clear as to which measure should be used. Correlation coefficient is a measure of degree of covariability between x and y, whereas the objective of regression analysis is to describe the nature of relationship between the variables so that we may be able to predict the value of one on the basis of another.

2.10 Linear Regression

If the variable in a bivariate distribution are correlated, then points in scatter diagram will be more or less concentrated round a curve. This curve is called the ***curve of regression***. If the curve is a straight line, it is called a ***line of regression*** and the regression is said to be linear. Since the line of regression gives the best estimate to the value of dependent variable for any given value of the independent variable, therefore, it is called the line of best fit which is obtained by the method of least squares. Since any one of the two variables x and y can be taken as the independent variable and the other as a dependent variable. Therefore, there are two regression lines, one as the line of regression of y on x and the other as the line of regression of x on y. The linear regression does not test whether the data are linear. It assumes that the data are linear, and finds the slope and intercept that make a straight line best fit to our data. The regression line of y on x is drawn in such a way that it minimizes the total of squares of the ***vertical deviations*** and the line of x on y minimizes the total of squares of the ***horizontal deviations***.

2.11 Lines of Regression

Let the equation of line of regression of y on x be

$$y = a + bx \tag{1}$$

then

$$\bar{y} = a + b\bar{x} \tag{2}$$

Now subtracting (2) from (1), we have

$$y - \bar{y} = b(x - \bar{x}) \tag{3}$$

The normal equations for the equation (1) are

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2 \tag{4}$$

Shifting the origin to $(\bar{x}, \bar{y})$, (4) becomes

$$\sum (x - \bar{x})(y - \bar{y}) = a\sum (x - \bar{x}) + b\sum (x - \bar{x})^2 \tag{5}$$

We know that

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

$$= \frac{\frac{1}{n}\sum (x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y}$$

Where $\sum (x - \bar{x}) = 0$ and $\sigma_x^2 = \frac{1}{n}\sum (x - \bar{x})^2$

From (5), we have

$$nr\sigma_x\sigma_y = a.0 + b.n\sigma_x^2 \text{ or } b = r\frac{\sigma_y}{\sigma_x}$$

Hence, from (3), the line of regression of y on x is given by

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

Similarly, the line of regression of x on y is given by

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

$r\frac{\sigma_y}{\sigma_x}$ is called the **regression coefficient** of y on x and is denoted by b_{yx}

$$b_{yx}=r\frac{\sigma_y}{\sigma_x}=\frac{Cov(x,y)}{\sigma_x^2}=\frac{n\sum xy-\sum x\sum y}{n\sum x^2-(\sum x)^2}$$

$r\frac{\sigma_x}{\sigma_y}$ is called the **regression coefficient** of x on y and is denoted by b_{xy}

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

$$b_{xy}=\frac{n\sum xy-\sum x\sum y}{n\sum y^2-(\sum y)^2}$$

Note : The line of regression of y on x is used to estimate the value of y for given valu
The line of regression of x on y is used to estimate the value of x for given value of y.

2.12 Properties of Regression Coefficients

Property 1. The geometric mean of the two regression coefficients is the coeffic correlation i.e.,

$$r=\sqrt{b_{yx}\times b_{xy}}$$

Proof We know that the two regression coefficients are

Now, $b_{yx}=r\frac{\sigma_y}{\sigma_x}$ and $b_{xy}=r\frac{\sigma_x}{\sigma_y}$

$$\text{Geometric mean}=\sqrt{b_{yx}\times b_{xy}}$$

$$=\sqrt{r\frac{\sigma_y}{\sigma_x}\times r\frac{\sigma_x}{\sigma_y}}$$

$$=\sqrt{r^2}=\pm r$$

Note: r will be positive (+ve) if b_{yx} and b_{xy} are positive, r will be negative (-ve) if b_{yx} and b_{xy} are negative.

Property 2. The arithmetic mean of the regression coefficients is greater than or equal to the coefficient of correlation i.e.,

$$\frac{b_{yx}+b_{xy}}{2} \geq r$$

Proof We know that the two regression coefficients are

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \text{ and } b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

Now, $$\text{Arithmetic mean} = \frac{b_{yx}+b_{xy}}{2} \geq r$$

or $$\frac{r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y}}{2} \geq r$$

or $$\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} - 2 \geq 0$$

or $$\frac{1}{\sigma_x\sigma_y}\left[\sigma_x^2 + \sigma_y^2 - 2\sigma_x\sigma_y\right] \geq 0$$

or $$\frac{1}{\sigma_x\sigma_y}\left[\sigma_x - \sigma_y\right]^2 \geq 0 \text{ which is true}$$

$\because \sigma_x, \sigma_y \geq 0$

$$\therefore \frac{b_{yx}+b_{xy}}{2} \geq r$$

Property 3. If one of the regression coefficient is greater than unity, the other must be less than unity i.e , If $b_{yx} > 1$ then $b_{xy} < 1$.

Proof The two regression coefficients are b_{yx} and b_{xy}.

Let $b_{yx} > 1$ then $\frac{1}{b_{yx}} < 1$

We know that $b_{yx} \times b_{xy} = r^2 \leq 1$ or $b_{yx} \times b_{xy} \leq 1$ or $b_{xy} \leq \frac{1}{b_{yx}} < 1$

Similarly, if $b_{xy} > 1$ then $b_{yx} < 1$.

Property 4. Regression coefficients are independent of the origin but not of scale.

Proof Let $u = \frac{x-a}{h}$ and $v = \frac{y-b}{k}$ where a, b, h and k are constants.

Now, $b_{yx} = r\frac{\sigma_y}{\sigma_x} = r\frac{k\sigma_v}{h\sigma_u} = \frac{k}{h}\left(r\frac{\sigma_v}{\sigma_u}\right) = \frac{k}{h}b_{vu}$ $(\because \sigma_x^2 = h^2\sigma_u^2, \sigma_y^2 = k^2$

Similarly, $b_{xy} = \frac{h}{k}b_{uv}$

2.13 Difference between Correlation and Regression

Correlation makes no priori assumption as to whether one variable is dependent on the and is not concerned with the relationship between variables; instead it gives an estimate of degree of association between the variables. In fact, correlation analysis tests for interdepen of the variables.

As regression attempts to describe the dependence of a variable on one or more explan variables; it implicitly assumes that there is a one-way causal effect from the explan variable (s) to the response variable, regardless of whether the path of effect is direct or in There are advanced regression methods that allow a non-dependence based relationship described.

Regression analysis involves identifying the relationship between a dependent variable one or more independent variables. A model of the relationship is hypothesized, and estima the parameter values are used to develop an estimated regression equation. Various tests a employed to determine if the model is satisfactory. If the model is deemed satisfactor estimated regression equation can be used to predict the value of the dependent variable values for the independent variables.

Correlation and regression analysis are related in the sense that both deal with relation among variables. The correlation coefficient is a measure of linear association betwe variables. Neither regression nor correlation analyses can be interpreted as establishing and-effect relationships. They can indicate only how or to what extent variables are ass with each other. The correlation coefficient measures only the degree of linear asso between two variables. Any conclusions about a cause-and-effect relationship must be ba the judgment of the analyst.

2.14 Standard Error of Estimate

By the use of regression equations, the calculations of perfect prediction is not possible, and for this it is better to find the likely error in the estimated values of the y or x we use standard error of estimate. The standard error of estimate measures the dispersi

an average line, called the regression line. The standard error of estimate of y and x is given by the formula:

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

$$\text{or } S_{xy} = \sqrt{\frac{\sum (x - x_e)^2}{n}}$$

where y and x are the original values and y_e is the estimated values and n is number of observations.

$$\text{i.e., } S_{yx} = \sqrt{\frac{\textit{Unexplained Variation in } y}{n}} \text{ or}$$

$$S_{yx} = \sqrt{\frac{\textit{Unexplained Variation in } x}{n}}$$

The simpler way to find the values of S_{yx} and S_{xy} is as follows:

$$S_{yx} = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n}}$$

and

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

The standard error of estimate measures the accuracy of the estimated figures. Thus whenever the value of the standard error of estimate is small, the dots of the scattered diagram will be closer to the regression line. If the value of the standard error is zero, then there is no variation about the line and the correlation will be perfect.

Illustrative Examples

Example 1. The two regression lines of the variables x and y are:

$$x = 19.13 - 0.87y \text{ and } y = 11.64 - 0.50x$$

Find

i) mean of x and mean of y

ii) correlation coefficient between x and y

Solution

(i) Since the mean of x and mean of y lie on the regression lines, we have

$$\bar{x} = 19.13 - 0.87\bar{y} \text{ or } \bar{x} + 0.87\bar{y} = 19.13$$

and $\bar{y} = 11.64 - 0.50\bar{x}$ or $0.50\bar{x} + \bar{y} = 11.64$

Now, on solving the above equations for $\bar{x}$ and $\bar{y}$, we have

$$\bar{x} = 15.935 \text{ and } \bar{y} = 3.67$$

Mean of $x = 15.935$ and mean of $y = 3.67$

(ii) Let the regression line of y on x be

$$y = 11.64 - 0.50x$$

$$b_{yx} = -0.50 \quad \text{(coefficient of } x\text{)}$$

$\therefore$

and the regression line of x on y be

$$x = 19.13 - 0.87y$$

$$b_{xy} = -0.87 \quad \text{(coefficient of y)}$$

$\therefore$

We know that $r = \pm\sqrt{b_{yx} \times b_{xy}}$

$$\therefore \quad r = \pm\sqrt{-0.50 \times -0.87} = -0.66$$

(r is negative, since b_{yx} and b_{xy} are nega

Example 2. From the following information on values of two variables x and y, find the regression lines and the correlation coefficient between x and y.

$n = 10, \sum x = 20, \sum y = 40, \sum x^2 = 240, \sum y^2 = 410, \sum xy = 200.$

Solution We know that

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10 \times 200 - 20 \times 40}{10 \times 240 - (20)^2} = \frac{20 - 8}{24 - 4} = \frac{3}{5}$$

and

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2} = \frac{10 \times 200 - 20 \times 40}{10 \times 410 - (40)^2} = \frac{20 - 8}{41 - 16} = \frac{12}{25}$$

$$\bar{x} = \frac{1}{n}\sum x = \frac{20}{10} = 2, \quad \bar{y} = \frac{1}{n}\sum y = \frac{40}{10} = 4$$

The two regression lines are

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 4 = \frac{3}{5}(x - 2)$$

$$y = 0.6x + 2.8$$

and

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 2 = \frac{12}{25}(y - 4)$$

$$x = 0.48y + 0.08$$

We know that $r = \pm\sqrt{b_{yx} \times b_{xy}}$

$$r = \sqrt{\frac{3}{5} \times \frac{12}{25}} = \sqrt{\frac{36}{125}} = 0.536$$

Example 3. For 100 students of a class, the regression equation of marks in Statistics (x) on the marks in Mathematics (y) is $30y - 50x + 1800 = 0$. The mean marks in Mathematics is 60 and variance of marks in Statistics is $\left(\frac{16}{25}\right)^{th}$ of the variance of marks in Mathematics. Find the mean marks in Statistics and the coefficient of correlation between marks in the two subjects.

Solution Since the given line of regression is x on y so we have

$$30y - 50x + 1800 = 0.$$

$$x = \frac{3}{5}y + \frac{180}{5} = \frac{3}{5}y + 36$$

We have $b_{xy} = \frac{3}{5} = r\frac{\sigma_x}{\sigma_y}$

Given variance of $x = \left(\frac{16}{25}\right)$ variance of y

$$\frac{\text{variance of } x\ (\sigma_x^{\,2})}{\text{variance of } y\ (\sigma_y^{\,2})} = \frac{16}{25} \qquad \Rightarrow \frac{\sigma_x}{\sigma_y} = \frac{4}{5}$$

So, $\frac{3}{5} = r \times \frac{4}{5}$ $\Rightarrow r = \frac{3}{4} = 0.75$

($\because b_{xy}$ is positive, r is positive)

Since the mean of x and mean of y lie on the regression lines, we have

$$\bar{x} = \frac{3}{5}\bar{y} + 36$$

$$\Rightarrow \bar{x} = \frac{3}{5} \times 60 + 36 = 72 \qquad (\because \bar{y} = 60)$$

Example 4. The lines of regression of y on x and x on y are $y = x + 5$ and $16x - 9y = 94$ respectively. Find the variance of x if the variance of y is 16. Also find the covariance of x and y.

Solution Regression equation of y on x is $y = x + 5 \Rightarrow b_{yx} = 1$ (coefficient of x)

Regression equation of x on y is $16x - 9y = 94$ i.e. $x = \frac{9}{16}y + \frac{94}{16}$

$$\Rightarrow b_{xy} = \frac{9}{16} \quad \text{(coefficient of y)}$$

$$r = \pm\sqrt{b_{yx} \times b_{xy}}$$

We know that

$r = \sqrt{1 \times \frac{9}{16}} = \frac{3}{4} = 0.75$ (r is positive, since b_{yx} & b_{xy} are positive

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \Rightarrow \sigma_x = \frac{b_{xy} \times \sigma_y}{r} = \frac{\left(\frac{9}{16}\right) \times 4}{\left(\frac{3}{4}\right)} = 3 \qquad (\because \sigma_y^2 =$$

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

$$\Rightarrow \quad Cov.(x, y) = r \times \sigma_x \times \sigma_y = \frac{3}{4} \times 3 \times 4 = 9 .$$

Example 5. The equations of two lines of regression are $4x + 3y + 7 = 0$ and $3x + 4y + 8 =$

Find

i) the mean values of x and y

ii) the regression coefficients b_{yx} and b_{xy}

iii) the correlation coefficient between x and y

iv) the standard deviation of y, if the variance of x is 16

v) the value of y for $x = 15$.

Solution

(i) Since the mean of x and mean of y lie on the regression lines, we have

$4\bar{x} + 3\bar{y} + 7 = 0$ or $4\bar{x} + 3\bar{y} = -7$

and $3\bar{x} + 4\bar{y} + 8 = 0$ or $3\bar{x} + 4\bar{y} = -8$

Now, on solving the above equations for $\bar{x}$ and $\bar{y}$, we have

$$\bar{x} = -\frac{4}{7} \text{ and } \bar{y} = -\frac{11}{7}$$

Mean of $x = -\frac{4}{7}$ and mean of $y = -\frac{11}{7}$

(ii) Let the regression line of y on x be

$$3x+4y+8=0 \text{ or } y=-\frac{3}{4}x-2 \qquad \text{(coefficient of x)}$$

and the regression line of x on y be

$$4x+3y+7=0 \text{ or } x=-\frac{3}{4}y-\frac{7}{4}$$

$$\therefore \qquad b_{xy}=-\frac{3}{4} \qquad \text{(coefficient of y)}$$

Since $b_{yx}\times b_{xy}=-\frac{3}{4}\times-\frac{3}{4}=\frac{9}{16}<1$

Hence, the choice of regression lines is correct.

So $b_{yx}=-\frac{3}{4}$ and $b_{xy}=-\frac{3}{4}$

(iii) We know that $r=\pm\sqrt{b_{yx}\times b_{xy}}$

$$\therefore \qquad r=\pm\sqrt{\left(\frac{-3}{4}\right)\times\left(\frac{-3}{4}\right)}=\pm\frac{3}{4}=-0.75$$

(r is negative, since b_{yx} and b_{xy} are negative)

(iv) We have $\sigma_x^2=16 \qquad \Rightarrow \sigma_x=4$

Now, $$b_{yx}=-\frac{3}{4} \text{ or } r\frac{\sigma_y}{\sigma_x}=-\frac{3}{4}$$

$$\left(-\frac{3}{4}\right)\times\frac{\sigma_y}{2}=-\frac{3}{4} \qquad \Rightarrow \sigma_y=2$$

(v) Since we have to find y when x is given, we use line of regression of y on x

$$y=-\frac{3}{4}x-2$$

Putting $x=15$, we have

$$y=-\frac{3}{4}\times 15-2=13.25$$

Example 6. In a correlation study the following values are obtained as:

	x	y
Mean	75	77
Standard deviation	2.6	3.6

If the correlation coefficient between x and y is found to be 0.81, then find the two regression equations that are associated with the above values.

Hence we can conclude that the equation $16x - 6y + 60 = 0$, is the equation of y on x

$$y = \frac{16}{6}y + \frac{60}{6} \qquad \Rightarrow y = \frac{8}{3}y + 10$$

thus $$r\frac{\sigma_x}{\sigma_y} = \frac{8}{3}$$

and the equation x on y as $30x - 9y - 150 = 0$

$$x = \frac{9}{30}y + \frac{150}{30} \qquad \Rightarrow x = \frac{3}{10}y + 5$$

thus $$r\frac{\sigma_y}{\sigma_x} = \frac{3}{10}$$

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

$$r^2 = \frac{8}{3} \times \frac{3}{10} = \frac{8}{10} \quad (0.8 < 1),$$

$\Rightarrow$ $$r = 0.894$$

Example 8. Find the means of the variables x and y and the correlation coefficient for the following.

Regression equation of y on x is: $2y = x + 50$ and

Regression equation of x on y is: $3y = 2x + 10$

Solution Means of x and y is obtained by solving the given equations and obtaining the values of x and y.

$$2y = x + 50 \qquad \ldots.(i)$$

$$3y = 2x + 10 \qquad \ldots(ii)$$

Multiplying equation (i) by 2 and subtracting equation (ii) from it, we get

$y - 90 = 0 \qquad \Rightarrow y = 90 \qquad \Rightarrow \bar{y} = 90$

Now substituting the value of $\bar{y} = 90$, in equation (i), we get

$180 - x - 50 = 0 \qquad \Rightarrow -x = -180 + 50 \qquad \Rightarrow \bar{x} = 130$

So the mean values are 130 and 90 respectively for x and y.

Now $2y = x + 50 \qquad \Rightarrow y = \frac{1}{2}x + 25 \qquad \Rightarrow b_{yx} = 0.5$

and $3y = 2x + 10 \qquad \Rightarrow x = \frac{3}{2}y - 5 \qquad \Rightarrow b_{xy} = 1.5$

$\therefore$

$$r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{0.5 \times 1.5} = 0.87$$

Example 9. The following table gives age (x) in years of scooters and its annual main... cost (y) in hundred rupees:

x	1	3	5	7	9
y	15	18	21	23	22

Estimate the maintenance cost for a 4 year old scooter after finding the appro... line of regression.

Solution

x	y	x^2	xy
1	15	1	15
3	18	9	54
5	21	25	105
7	23	49	161
9	22	81	198
$\sum x = 25$	$\sum y = 99$	$\sum x^2 = 165$	$\sum xy = 533$

Here, $n = 5$

$$\bar{x} = \frac{1}{n}\sum x = \frac{25}{5} = 5, \quad \bar{y} = \frac{1}{n}\sum y = \frac{99}{5} = 19.8$$

Now,

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{5 \times 533 - 25 \times 99}{5 \times 165 - (25)^2} = \frac{2665 - 2475}{825 - 625} = \frac{190}{200} = 0.9$$

The line of regression of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 19.8 = 0.95(x - 5)$$

$$y = 0.95x + 15.05$$

When x = 4 years, we have

$$y = 0.95 \times 4 + 15.05 = 18.85 \text{ hundred rupees} = ₹ 188$$

Example 10. Find the equation of two lines of regression for the following data:

x	10	20	30	40	50
y	70	60	50	40	30

and hence find an estimate of y for x = 35 from the appropriate line of regr...

Correlati...

Solution

H...

Example 1...

Solution

x	y	x^2	y^2	xy
10	70	100	4900	700
20	60	400	3600	1200
30	50	900	2500	1500
40	40	1600	1600	1600
50	30	2500	900	1500
$\sum x = 150$	$\sum y = 250$	$\sum x^2 = 5500$	$\sum y^2 = 13500$	$\sum xy = 6500$

Here, $n = 5$

$$\bar{x} = \frac{1}{n}\sum x = \frac{150}{5} = 30, \quad \bar{y} = \frac{1}{n}\sum y = \frac{250}{5} = 50$$

Now, $$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{5 \times 6500 - 150 \times 250}{5 \times 5500 - (150)^2} = \frac{32500 - 37500}{27500 - 22500} = -1$$

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2} = \frac{5 \times 6500 - 150 \times 250}{5 \times 13500 - (250)^2} = \frac{32500 - 37500}{67500 - 62500} = -1$$

So, the line of regression of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 50 = -1(x - 30)$$

$$y = 50 - x + 30$$

$$y = -x + 80$$

and the line of regression of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 30 = -1(y - 50)$$

$$x = -y + 80$$

To estimate the value of y when x is given, we use the line of regression of y on x, i.e.

$$y = -x + 80$$

Now, substitute x = 35, we have

$$y = -35 + 80 = 45$$

Example 11. Obtain a regression plane by using multiple linear regression to fit the data given below:

x_1	1	2	3	4
x_2	0	1	2	3
y	12	18	24	30

Solution Let the equation of regression plane be

$$y = a_0 + a_1 x_1 + a_2 x_2$$

The normal equations for equation (1) are

$$\sum_{i=1}^{n} y_i = na_0 + a_1 \sum_{i=1}^{n} x_{i1} + a_2 \sum_{i=1}^{n} x_{i2}$$

$$\sum_{i=1}^{n} x_{i1} y_i = a_0 \sum_{i=1}^{n} x_{i1} + a_1 \sum_{i=1}^{n} x_{i1}^2 + a_2 \sum_{i=1}^{n} x_{i1} x_{i2}$$

and $$\sum_{i=1}^{n} x_{i2} y_i = a_0 \sum_{i=1}^{n} x_{i2} + a_1 \sum_{i=1}^{n} x_{i1} x_{i2} + a_2 \sum_{i=1}^{n} x_{i2}^2$$

x_1	x_2	y	x_1^2	x_2^2	$x_1 x_2$	$x_1 y$	$x_2 y$
1	0	12	1	0	0	12	0
2	1	18	4	1	2	36	18
3	2	24	9	4	6	72	48
4	3	30	16	9	12	120	90
10	6	84	30	14	20	240	156

Now the equations (2), (3) and (4) becomes

$$84 = 4a_0 + 10a_1 + 6a_2$$

$$240 = 10a_0 + 30a_1 + 20a_2$$

$$156 = 6a_0 + 20a_1 + 14a_2$$

On solving these equations, we have

$$a_0 = 10,\ a_1 = 2 \text{ and } a_2 = 4$$

Putting these values in equation (1) we have

$$y = 10 + 2x_1 + 4x_2$$

Example 12. Obtain a regression equations and calculate the standard error of the estim the following data:

x	6	5	8	6	8	3
y	9	9	7	8	7	5

Solution

Calculations of Regression Equations

x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x}).(y-\bar{y})$
6	9	0	0	0	0	0
5	9	-1	1	0	0	0
8	7	2	4	-2	4	-4
6	8	0	0	-1	1	0
8	7	2	4	-2	4	-4
3	5	-3	9	-4	16	12
36	45	0	18	-9	25	4

Means $\bar{x} = \frac{\sum x}{n} = \frac{36}{6} = 6$

$\bar{y} = \frac{\sum y}{n} = \frac{45}{5} = 9$

Regression equation of y on x is: $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

$r\frac{\sigma_y}{\sigma_x} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{4}{18}$

$y - 9 = \frac{4}{18}(x - 6)$

$y - 9 = 0.22(x - 6)$

Thus Regression equation of y on x is: $y = 0.22x + 10.33$

Regression equation of y on x is: $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$

$r\frac{\sigma_x}{\sigma_y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2} = \frac{4}{25}$

$x - 6 = \frac{4}{25}(y - 9)$

$x - 6 = 0.16(x - 9)$

Thus Regression equation of y on x is: $x = 0.16y + 7.44$

Calculations of Standard Error of Estimate

x	y	y_c	x_c	$(y-y_c)^2$	$(x-x_c)^2$
6	9	11.65	8.88	135.72	78.85
5	9	11.43	8.88	130.64	78.85
8	7	12.09	7.62	146.17	58.06
6	8	11.65	8.72	135.72	76.04
8	7	12.09	7.62	146.17	58.06
3	5	10.93	8.24	119.46	67.89
36	45	69.84	49.96	813.88	417.75

Here $n = 6$

The standard error of estimate of y on x given by the formula:

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

$$S_{yx} = \sqrt{\frac{813.88}{6}} = \sqrt{135.65} = 11.65$$

The standard error of estimate of x and y is given by the formula:

$$S_{xy} = \sqrt{\frac{\sum (x - x_e)^2}{n}}$$

$$S_{xy} = \sqrt{\frac{417.75}{6}} = \sqrt{69.63} = 8.34$$

EXERCISE

1. The two regression lines of the variables x and y are:
   $4x + y - 9 = 0$ and $x + 9y - 11 = 0$
   Find the means and correlation coefficient between x and y

   Ans. 2,1

2. Find the line of regressions for the following data:
   $n = 3, \sum x = 6, \sum y = 15, \sum x^2 = 14, \sum y^2 = 77, \sum xy = 31$

   Ans. $\mathbf{y = 0.5x + 4}$ **and** x=0.5

3. Find the line of regressions for the following data:
   $n = 5, \sum x = 30, \sum y = 40, \sum x^2 = 220, \sum y^2 = 340, \sum xy = 214$

   Ans. $\mathbf{y = 0.537x + 6.38}$ **and** x=4.7

4. The lines of regression of y on x and x on y are $y = x + 15$ and $16x - 9y = 94$ respe
   Find the variance of x if the variance of y is 16. Also find the covariance of x and y.

   Ans. $\mathbf{r = 0.75}$, Cov(

5. The two regression lines of the variables x and y are:
   $8x - 3y = -30$ and $10x - 3y = 50$
   Find the correlation coefficient between x and y

6. The equations of two lines of regression are $8x - 10y + 66 = 0$ and $40x - 18y = 21$
   Find
   i) the mean values of x and y
   ii) the correlation coefficient between x and y
   iii) the standard deviation of y, if the variance of x is 9.

   (ii) 0.6

7. Two random variables have the regression lines $3x + 2y - 26 = 0$ and $6x + y - 31 = 0$. Find the mean values of x and y and the coefficient of correlation. If the variance of x is 25, find standard deviation of y from the data given.

Ans. $\bar{x} = 4, \bar{y} = 7, \sigma_y = 15, r = -0.5$

8. Find the regression line of x on y and estimate the value of x, when y = 5 from the following data:

$n = 25, \sum x = 125, \sum y = 100, \sum x^2 = 1650, \sum y^2 = 1500, \sum xy = 50.$

Ans. $x = -\frac{9}{22}y + \frac{146}{22}; 4.591$

9. Find the regression coefficient b_{xy} for the following data:

$n = 6, \sum x = 30, \sum y = 42, \sum x^2 = 184, \sum y^2 = 318, \sum xy = 199.$

Ans. -0.46

10. Find the regression coefficient b_{yx} and b_{xy} between x and y for the following data:

$n = 7, \sum x = 24, \sum y = 12, \sum x^2 = 374, \sum y^2 = 97, \sum xy = 157.$

Ans. $b_{yx} = 0.397; b_{xy} = 1.516; r = 0.776$

Find the line of regression of y on x for the following data:

x	10	9	8	7	6	4	3
y	8	12	7	10	8	9	6

Ans. $y = \frac{1}{3}x + \frac{133}{21}$

11. The correlation coefficient between the variables x and y is $r = 0.60$, if $\sigma_x = 1.5$, $\sigma_y = 2$, $\bar{x} = 10, \bar{y} = 20$. Then find both the equation of the regression lines.

Ans. $y = 0.8x + 12;\ x = 0.45y + 1$

12. Find the line of regression of y on x for the following data:

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

Estimate the value of y, when x = 10.

Ans. $y = \frac{7}{11}x + \frac{6}{11}; 6.91$

13. Find the regression lines for the following data:

x	6	2	10	4	8
y	9	11	5	8	7

Ans. $y = 11.9 - 0.65x; x = 16.4 - 1.3y$

24. You are given below the following data regarding advertisement and sales as:

	Advertisement (thousand)	Sales (thousand)
Mean	10	90
Standard deviation	3	12

If the correlation coefficient between x and y is found to be 0.80, then estimate the sales when the advertisement expenditure is Rs 15,000 and estimate the value of advertisement expenditure when the sales target is Rs 120,000

Ans. 106,60

2.15 Curve Fitting

Let $(x_i, y_i), i = 1,2,...,n$ be a given set of n pairs of values, where 'x' be an independent variable and y be a dependent variable. These pairs of values of x and y give us 'n' points on a known curve whose equation is

$$y = f(x).$$

Curve fitting means an exact relationship between two variables by algebraic equation. This relationship is the equation of the curve. Therefore, curve fitting means to form an equation of the curve from the given data. Curve fitting has very much importance in theoretical as well as practical statistics.

Theoretically, it is useful in study of correlation and regression and particularly it enables us to represent the relationship between two variables by simple algebraic expressions such as polynomials, exponential or logarithmic functions.

2.16 Method of Least Squares

The method of least squares is the most systematic procedure to fit a unique curve to a given data points and is widely used in practical computations. Let (x_i, y_i), $i = 1,2,...,n$ be a given set of n pairs of values and suppose we want to fit a curve $y = f(x)$ to the given set of values. Let Y_i be the value of y corresponding to the value of x_i of x as determined by $y = f(x)$. The value Y_i is called the estimated value of the given value y_i corresponding to x_i. If e_i is the error of approximation at $x = x_i$, then

$$e_i = y_i - Y_i = y_i - f(x_i)$$

If we minimize the sum of the squares of the errors then it is called least squares method. If S be the sum of squares of errors then,

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2$$

Now, we have to minimize S.

2.17 Fitting of a Straight Line by Method of Least Squares

Let $y = a + bx$ be the straight line to be fitted to the given data (x_i, y_i), $i = 1,2,\ldots,n$. The problem is to determine 'a' and 'b' so that the line is the line of best fit.

Let $P_i(x_i, y_i)$ be any general point in the scatter diagram. Draw P_iR perpendicular to x-axis meeting the line in at Q_i. Abscissa of Q_i is x_i and since Q_i lies on the line, its ordinate is $a + bx_i$. Hence the coordinate of Q_i are $(x_i, a + bx_i)$.

$$P_iQ_i = P_iR - Q_iR$$

$= y_i - (a + bx_i)$ is called the **error of estimate** or the **residual** for y_i. According to the principle of least squares, we have to determine a and b so that

$$S = \sum_{i=1}^{n} (P_iQ_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

is minimum.

From the principle of maxima and minima, the partial derivatives of S, with respect to a and b should vanish separately, i.e.,

$$\frac{\partial S}{\partial a} = 0 \qquad \Rightarrow -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\frac{\partial S}{\partial b} = 0 \qquad \Rightarrow -2\sum_{i=1}^{n} (y_i - a - bx_i)(x_i) = 0$$

We have, $$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

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

These equations are known as the normal equations for estimating a and b.

All the quantities $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i y_i$ can be obtained from the given set of points (x_i, y_i), $i = 1,2,\ldots,n$ and the normal equations can be solved for a and b. With the values of a and b so obtained equation $y = a + bx$ is the line of best fit to the given set of points (x_i, y_i)

2.17.1 Effect of Change of Origin and Scale

Sometimes the magnitude of the variables in the given data is so big that the calculations become very much tedious. The size of the data can be considerably reduced by assuming some convenient origin for x and y series in the given data. The problem is further simplified by taking suitable scale when the values of x are given at equally spaced intervals.

Let 'h' be the width of the interval at which the values of x are given and let the origin and y at the point x_0, y_0 respectively, then putting $u = \frac{x - x_0}{h}$, (h is the width of interval the values of x) and $v = y - y_0$.

In case of change of origin, if n is **odd** then

$$u = \frac{x - (middle\ term)}{h}$$

and if n is **even** then

$$u = \frac{x - (mean\ of\ two\ middle\ terms)}{h/2}$$

Similar transformations can be applied to polynomials of higher degree.

Illustrative Examples

Example 1. Fit a straight line by the method of least squares to the following data:

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

Solution Let the straight line of best fit be

$$y = a + bx \qquad (1)$$

The normal equations are

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

and

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \qquad (3)$$

where $n = 5$ and the normal equations can be solved by using the values of $\sum x, \sum y, \sum x^2$ and $\sum xy$ from the following table:

Calculations for Normal Equations

x	y	x^2	xy
0	1	0	0
1	1.8	1	1.8
2	3.3	4	6.6
3	4.5	9	13.5
4	6.3	16	25.2
$\sum x = 10$	$\sum y = 16.9$	$\sum x^2 = 30$	$\sum xy = 47.1$

Now the equations (2) and (3) becomes

$$16.9 = 5a + 10b$$

$$47.1 = 10a + 30b$$

On solving these equations, we have $a = 0.72$ and $b = 1.33$

Putting these values in (1), we get the straight line as

$$y = 0.72 + 1.33x$$

Example 2. Fit a straight line by the method of least squares to the following data:

x	1	2	3	4	5
y	15	70	140	250	380

Solution Here, $n = 5$ i.e., odd and $h = 1$.

Thus, we take $$u = \frac{x - 3}{1} = x - 3$$

Let the equation of straight line be

$$y = a + bu \qquad (1)$$

The normal equations are

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} u_i \qquad (2)$$

and

$$\sum_{i=1}^{n} u_i y_i = a\sum_{i=1}^{n} u_i + b\sum_{i=1}^{n} u_i^2 \qquad (3)$$

Now, where $n = 5$ and the normal equations can be solved by using the values of $\sum u, \sum y, \sum u^2$ and $\sum uy$ are calculated in the following table:

Calculations for Normal Equations

x	y	u	u^2	uy
1	15	-2	4	-30
2	70	-1	1	-70
3	140	0	0	0
4	250	1	1	250
5	380	2	4	760
$\sum x = 15$	$\sum y = 855$	$\sum u = 0$	$\sum u^2 = 10$	$\sum uy = 910$

Now the equations (2) and (3) becomes

$$855 = 5a + 0 \times b$$

$$910 = 0 \times a + 10b$$

On solving these equations, we have $a = 171$ and $b = 91$