

MODULE-2

STATISTICS

Curve Fitting

This is a method of finding a specific relation connecting the dependent and independent variables for a given data so as to satisfy the data as accurately as possible.

Curve fitting by the method of least squares

The method of least squares provides a relationship $y = f(x)$ such that the sum of the squares of the residuals is least. The difference between observed (y_i) and expected value (Y_i) i.e., $R_i = y_i - Y_i$ are called the residuals.

Fitting of a Straight line $y = ax + b$

Given $y = ax + b$

Normal Equations(N.E's): $a \sum x + nb = \sum y$

$$a \sum x^2 + b \sum x = \sum xy$$

Solving the above equations we get the values of a and b , substituting these values in the given equation leads to the curve of best fit.

Note:

Normal Equations(N.E's) for the st.line $y = a + bx$:

$$\begin{aligned} na + b \sum x &= \sum y \\ a \sum x + b \sum x^2 &= \sum xy \end{aligned}$$

Problems

1. Fit a straight line $y = ax + b$ for the following data

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

Sol: The normal equations for fitting the Straight line $y = ax + b$ --- (*) are

$$a \sum x + nb = \sum y$$

$$a \sum x^2 + b \sum x = \sum xy$$

Here $n=8$

x	y	xy	x^2
-----	-----	------	-------

1	1	1	1
3	2	6	9
4	4	16	16
6	4	24	36
8	5	40	64
9	7	63	81
11	8	88	121
14	9	126	196
$\sum x = 56$	$\sum y = 40$	$\sum xy = 364$	$\sum x^2 = 524$

The N.E's become

$$56a + 8b = 40$$

$$524a + 56b = 364$$

On solving , we get

$$a = 0.6363 \quad \text{and} \quad b = 0.5454$$

Thus, (*) $\Rightarrow y = (0.6363)x + (0.5454)$, is the curve of best fit.

2. Find the equation of the best fitting straight line for the following data and hence estimate the value of dependent variable corresponding to the value 30 of the independent variable.

<i>x</i>	<i>5</i>	<i>10</i>	<i>15</i>	<i>20</i>	<i>25</i>
<i>y</i>	<i>16</i>	<i>19</i>	<i>23</i>	<i>26</i>	<i>30</i>

Sol: Let $y = ax + b$ --- (*) be the equation of the Straight line.

$$\text{N.E's : } a \sum x + nb = \sum y$$

$$a \sum x^2 + b \sum x = \sum xy$$

Here n=5

<i>x</i>	<i>y</i>	<i>xy</i>	<i>x²</i>
-----------------	-----------------	------------------	-----------------------------

5	16	80	25
10	19	190	100
15	23	345	225
20	26	520	400
25	30	750	625
$\sum x = 75$	$\sum y = 114$	$\sum xy = 1885$	$\sum x^2 = 1375$

The N.E's become

$$\begin{aligned} 75a + 5b &= 114 \\ 1375a + 75b &= 1885 \end{aligned}$$

On solving , we get

$$a = 0.7 \quad \text{and} \quad b = 12.3$$

Thus, (*) $\Rightarrow y = (0.7)x + (12.3)$, is the curve of best fit.

When $x = 30 \Rightarrow y = (0.7)(30) + (12.3)$

$$y = 33.3$$

3. Fit a straight line in the least square sense for the following data

x	50	70	100	120
y	12	15	21	25

Sol: Let $y = ax + b$ --- (*) be the equation of the Straight line.

$$\begin{aligned} \text{N.E's : } a \sum x + nb &= \sum y \\ a \sum x^2 + b \sum x &= \sum xy \end{aligned}$$

Here n=4

x	y	xy	x²
----------	----------	-----------	----------------------

50	12	600	2500
70	15	1050	4900
100	21	2100	10000
120	25	3000	14400
$\sum x = 340$	$\sum y = 73$	$\sum xy = 6750$	$\sum x^2 = 31800$

The N.E's become

$$\begin{aligned} 340a + 4b &= 73 \\ 13800a + 340b &= 6750 \end{aligned}$$

On solving , we get

$$a = 0.19 \quad \text{and} \quad b = 2.28$$

Thus, (*) $\Rightarrow y = (0.19)x + (2.28)$, is the curve of best fit.

4. A simply supported beam carries a concentrated load P at its mid point. Corresponding to various values of P the maximum deflection Y is measured and is given in the following table.

P	100	120	140	160	180	200
Y	0.45	0.55	0.60	0.70	0.80	0.85

Find a law of the form $Y = a + bP$ and estimate Y when P is 150.

Sol: Let $Y = a + bP$ --- (*) be the equation of the Straight line.

$$\begin{aligned} \text{N.E's : } na + b \sum P &= \sum Y \\ a \sum P + b \sum P^2 &= \sum PY \end{aligned}$$

Here $n = 6$

P	Y	PY	P^2
-----	-----	------	-------

100	0.45	45	10000
120	0.55	66	14400
140	0.60	84	19600
160	0.70	112	25600
180	0.80	144	32400
200	0.85	170	40000
$\sum P = 900$	$\sum Y = 3.95$	$\sum PY = 621$	$\sum P^2 = 142000$

The N.E's become

$$\begin{aligned} 6a + 900b &= 3.95 \\ 900a + 142000b &= 621 \end{aligned}$$

On solving , we get

$$a = 0.0476 \quad \text{and} \quad b = 0.0041$$

Thus, (*) $\Rightarrow Y = (0.0476) + (0.0041)P$, is the curve of best fit.

$$\text{When } P = 150 \Rightarrow Y = (0.0476) + (0.0041)(150)$$

$$Y = 0.6626$$

5. Fit a straight line to the following data.

Year	1961	1971	1981	1991	2001
Production (in tons)	8	10	12	10	16

Also find the expected production in the year 2006.

Sol: Let year and production respectively be represented by the variables **x** and **y**.

Let the equation of Straight be **y = a + bx**

Let **X = x - 1981** and the line of fit will be **y = a + bX** --- (*)

N.E's are $na + b \sum X = \sum y$
 $a \sum X + b \sum X^2 = \sum Xy$. Here n=5

x	$X = x - 1981$	y	Xy	X^2
1961	-20	8	-160	400
1971	-10	10	-100	100
1981	0	12	0	0
1991	10	10	100	100
2001	20	16	320	400
	$\sum X = 0$	$\sum y = 56$	$\sum Xy = 160$	$\sum X^2 = 1000$

N.E's becomes $5a + 0.b = 56$
 $0.a + 1000b = 160$

On solving , we get

$$a = 11.2 \quad \text{and} \quad b = 0.16$$

Thus, (*) $\Rightarrow y = 11.2 + (0.16)X$

$$y = 11.2 + (0.16)(x - 1981)$$

$y = -305.76 + (0.16)x$, is the curve of best fit.

When $x = 2006$

$$y = -305.76 + (0.16)(2006) = 15.2 \text{ tons}$$

Expected production in the year 2006 is 15.2 tons.

Fitting of a Second degree parabola $y = ax^2 + bx + c$

Given $y = ax^2 + bx + c$

Normal Equations(N.E's): $a \sum x^2 + b \sum x + nc = \sum y$

$$a \sum x^3 + b \sum x^2 + c \sum x = \sum xy$$

$$a \sum x^4 + b \sum x^3 + c \sum x^2 = \sum x^2y$$

Solving the above equations we get the values of a , b and c , substituting these vales in the given equation leads to the curve of best fit.

Note: If $y = a + bx + cx^2$

Normal Equations(N.E's): $na + b \sum x + c \sum x^2 = \sum y$

$$a \sum x + b \sum x^2 + c \sum x^3 = \sum xy$$

$$a \sum x^2 + b \sum x^3 + c \sum x^4 = \sum x^2y$$

Problems

1. Fit a second degree parabola $y = ax^2 + bx + c$ in the least square sense for the following

x	1	2	3	4	5
y	10	12	13	16	19

data and hence estimate y at $x = 6$.

Sol: Given $y = ax^2 + bx + c$

Normal Equations(N.E's): $a \sum x^2 + b \sum x + nc = \sum y$

$$a \sum x^3 + b \sum x^2 + c \sum x = \sum xy$$

$$a \sum x^4 + b \sum x^3 + c \sum x^2 = \sum x^2y$$

Here $n=5$

x	y	xy	x^2	x^2y	x^3	x^4
1	10	10	1	10	1	1
2	12	24	4	48	8	16
3	13	39	9	117	27	81
4	16	64	16	256	64	256
5	19	95	25	475	125	625
$\sum x = 15$	$\sum y = 70$	$\sum xy = 232$	$\sum x^2 = 55$	$\sum x^2y = 906$	$\sum x^3 = 225$	$\sum x^4 = 979$

The N.E's become

$$\begin{aligned} 55a + 15b + 5c &= 70 \\ 225a + 55b + 15c &= 232 \\ 979a + 225b + 55c &= 906 \end{aligned}$$

On Solving, we have

$$a = 0.2857 \quad b = 0.4857 \quad \text{and} \quad c = 9.4$$

Thus, (*) $\Rightarrow y = (0.2857)x^2 + (0.4857)x + (9.4)$, is the curve of best fit .

At $x = 6$,

$$y = (0.2857)(6)^2 + (0.4857)(6) + (9.4)$$

$$y = 22.78$$

2. Fit a second degree parabola $y = a + bx + cx^2$ for the data:

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.3

Sol: Given $y = a + bx + cx^2$

Normal Equations(N.E's): $na + b \sum x + c \sum x^2 = \sum y$

$$a \sum x + b \sum x^2 + c \sum x^3 = \sum xy$$

$$a \sum x^2 + b \sum x^3 + c \sum x^4 = \sum x^2y$$

Here $n=5$

x	y	xy	x^2	x^2y	x^3	x^4
0	1	0	0	0	0	0
1	1.8	1.8	1	1.8	1	1
2	1.3	2.6	4	5.2	8	16
3	2.5	7.5	9	22.5	27	81
4	6.3	25.2	16	100.8	64	256
$\sum x = 10$	$\sum y = 12.9$	$\sum xy = 37.1$	$\sum x^2 = 30$	$\sum x^2y = 130.3$	$\sum x^3 = 100$	$\sum x^4 = 354$

The N.E's become

$$\begin{aligned}5a + 10b + 30c &= 12.9 \\10a + 30b + 100c &= 37.1 \\30a + 100b + 354c &= 130.3\end{aligned}$$

On Solving, we have

$$a = 1.42 \quad b = -1.07 \quad \text{and} \quad c = 0.55$$

Thus, (*) $\Rightarrow y = (1.42) + (-1.07)x + (0.55)x^2$, is the curve of best fit .

3. Fit a parabola $y = ax^2 + bx + c$ in the least square sense for the following data

x	2	4	6	8	10
y	3.07	12.85	31.47	57.38	91.29

Sol: Given $y = ax^2 + bx + c$

Normal Equations(N.E's):

$$\begin{aligned}a \sum x^2 + b \sum x + nc &= \sum y \\a \sum x^3 + b \sum x^2 + c \sum x &= \sum xy \\a \sum x^4 + b \sum x^3 + c \sum x^2 &= \sum x^2y\end{aligned}$$

Here $n=5$

x	y	xy	x^2	x^2y	x^3	x^4
2	3.07	6.14	4	12.28	8	16
4	12.85	51.40	16	205.60	64	256
6	31.47	188.82	36	1132.92	216	1296
8	57.38	459.04	64	3672.32	512	4096
10	91.29	912.90	100	9129.00	1000	10000
$\sum x$ = 30	$\sum y$ = 196.06	$\sum xy$ = 1618.3	$\sum x^2$ = 220	$\sum x^2y$ = 14152.12	$\sum x^3$ = 1800	$\sum x^4$ = 15664

The N.E's become

$$\begin{aligned}220a + 30b + 5c &= 196.06 \\1800a + 220b + 30c &= 1618.3 \\15664a + 1800b + 220c &= 14152.12\end{aligned}$$

On Solving, we have

$$a = 0.992 \quad b = -0.8551 \quad \text{and} \quad c = 0.696$$

Thus, (*) $\Rightarrow y = (0.992)x^2 + (-0.8551)x + (0.696)$, is the curve of best fit.

4. Find the best values of a, b, c if the equation $y = a + bx + cx^2$ is to fit most closely to the following observations

x	-2	-1	0	1	2
y	-3.150	-1.390	0.620	2.880	5.378

Sol: Equation of the parabola $y = a + bx + cx^2$

Normal Equations(N.E's):

$$na + b \sum x + c \sum x^2 = \sum y$$

$$a \sum x + b \sum x^2 + c \sum x^3 = \sum xy$$

$$a \sum x^2 + b \sum x^3 + c \sum x^4 = \sum x^2y$$

Here $n=5$

x	y	xy	x^2	x^2y	x^3	x^4
-2	-3.150	6.30	4	-12.60	-8	16
-1	-1.390	1.39	1	-1.39	-1	1
0	0.620	0	0	0	0	0
1	2.880	2.88	1	2.88	1	1
2	5.378	10.756	4	21.512	8	16
$\sum x = 0$	$\sum y = 4.338$	$\sum xy = 21.326$	$\sum x^2 = 10$	$\sum x^2y = 10.402$	$\sum x^3 = 0$	$\sum x^4 = 34$

The N.E's become

$$5a + 0.b + 10c = 4.338$$

$$0.a + 10b + 0.c = 21.326$$

$$10a + 0.b + 34c = 10.402$$

On Solving, we have

$$a = 0.6210 \quad b = 2.1326 \quad \text{and} \quad c = 0.1233$$

5. Fit a parabola for the following data.

x	1	2	3	4	5	6	7	8	9
y	2	6	7	8	10	11	11	10	9

Sol: Equation of the parabola $y = a + bx + cx^2$

Normal Equations(N.E's):

$$\begin{aligned} na + b \sum x + c \sum x^2 &= \sum y \\ a \sum x + b \sum x^2 + c \sum x^3 &= \sum xy \\ a \sum x^2 + b \sum x^3 + c \sum x^4 &= \sum x^2 y \end{aligned}$$

Here $n=9$

x	y	xy	x^2	x^2y	x^3	x^4
1	2	2	1	2	1	1
2	6	12	4	24	8	16
3	7	21	9	63	27	81
4	8	32	16	128	64	256
5	10	50	25	250	125	625
6	11	66	36	396	216	1296
7	11	77	49	539	343	2401
8	10	80	64	640	512	4096
9	9	81	81	729	729	6561
$\sum x$ = 45	$\sum y$ = 74	$\sum xy$ = 421	$\sum x^2$ = 285	$\sum x^2y = 2771$	$\sum x^3$ = 2025	$\sum x^4$ = 15333

The N.E's becomes

$$\begin{aligned}9a + 45b + 285c &= 74 \\45a + 285b + 2025c &= 421 \\285a + 2025b + 15333c &= 2771\end{aligned}$$

On Solving, we have

$$a = -0.9286 \quad b = 3.5232 \quad \text{and} \quad c = -0.2673$$

Thus, (*) $\Rightarrow y = (-0.9286) + (3.5232)x + (-0.2673)x^2$, is the curve of best fit .

6. Fit a second degree parabola to the following data

x	1	1.5	2	2.5	3	3.5	4
y	1.1	1.3	1.6	2	2.7	3.4	4.1

Sol: Equation of the parabola $y = a + bx + cx^2$

Normal Equations(N.E's):

$$\begin{aligned}na + b \sum x + c \sum x^2 &= \sum y \\a \sum x + b \sum x^2 + c \sum x^3 &= \sum xy \\a \sum x^2 + b \sum x^3 + c \sum x^4 &= \sum x^2y\end{aligned}$$

Here n=7

x	y	xy	x^2	x^2y	x^3	x^4
1	1.1	1.1	1	1.1	1	1
1.5	1.3	1.95	2.25	2.95	3.375	5.0625
2	1.6	3.2	4	6.4	8	16
2.5	2	5	6.25	12.5	15.625	39.0625
3	2.7	8.1	9	24.3	27	81
3.5	3.4	11.9	12.25	41.65	42.875	150.0625
4	4.1	16.4	16	65.6	64	256
$\sum x$ = 17.5	$\sum y$ = 16.2	$\sum xy$ = 47.65	$\sum x^2$ = 50.75	$\sum x^2y$ = 154.475	$\sum x^3$ = 161.875	$\sum x^4$ = 548.1875

The N.E's becomes

$$\begin{aligned}7a + 17.5b + 50.75c &= 16.2 \\17.5a + 50.75b + 161.875c &= 47.65 \\50.75a + 161.875b + 548.1875c &= 154.475\end{aligned}$$

On Solving, we have

$$a = 1.0357 \quad b = -0.1928 \quad \text{and} \quad c = 0.2428$$

Thus, (*) $\Rightarrow y = (1.0357) + (-0.1928)x + (0.2428)x^2$, is the curve of best fit .

Fitting of a curve of the form $y = ax^b$

Consider $y = ax^b$

Taking $\log_e$ on both sides we get,

$$\log_e y = \log_e(ax^b)$$

$$\log_e y = \log_e a + \log_e x^b$$

$$\log_e y = \log_e a + b \log_e x$$

$$Y = A + bX \quad , \text{ where } Y = \log_e y \quad , A = \log_e a \quad X = \log_e x$$

Normal Equations(N.E's)for the st.line $Y = A + bX$:

$$\begin{aligned}nA + b \sum X &= \sum Y \\A \sum X + b \sum X^2 &= \sum XY\end{aligned}$$

Solving above equations we get A, b . But $\log_e a = A \Rightarrow a = e^A$

Problems

1. Fit a least square geometric curve $y = ax^b$ for the following data

x	1	2	3	4	5
y	0.5	2	4.5	8	12.5

Sol: Consider $y = ax^b$ --- (*)

Taking $\log_e$ on both sides we get,

$$\log_e y = \log_e(ax^b)$$

$$\log_e y = \log_e a + \log_e x^b$$

$$\log_e y = \log_e a + b \log_e x$$

$$Y = A + bX \quad , \text{ where } Y = \log_e y \quad , A = \log_e a \quad X = \log_e x$$

Normal Equations(N.E's)for the st.line $Y = A + bX$

$$\begin{aligned} nA + b \sum X &= \sum Y \\ A \sum X + b \sum X^2 &= \sum XY \end{aligned}$$

Here n=5

x	y	$X = \log_e x$	$Y = \log_e y$	XY	X^2
1	0.5	0	-0.6931	0	0
2	2	0.6931	0.6931	0.4804	0.4804
3	4.5	1.0986	1.5041	1.6524	1.2069
4	8	1.3863	2.0794	2.8827	1.9218
5	12.5	1.6094	2.5257	4.0649	2.5902
		$\sum X$ = 4.7874	$\sum Y$ = 6.1092	$\sum XY$ = 9.0804	$\sum X^2$ = 6.1933

N.E's becomes $5A + 4.7874b = 6.1092$
 $4.7874A + 6.1993b = 9.0804$

On Solving, we have

$$A = -0.6932 \quad b = 2$$

But $\log_e a = -0.6932 \Rightarrow a = e^{-0.6932} = 0.5$

Thus, (*) $\Rightarrow y = ax^b$

$y = (0.5)x^2$, is the curve of best fit.

2. An experiment of lifetime t of cutting tool at different cutting speeds v (units) are given below

Speed (v)	350	400	500	600
Life (t)	61	26	7	2.6

Fit a relation of the form $v = at^b$.

Sol: Consider $v = at^b$ --- (*)

Taking $\log_e$ on both sides we get,

$$\log_e v = \log_e(at^b)$$

$$\log_e v = \log_e a + \log_e t^b$$

$$\log_e v = \log_e a + b \log_e t$$

$$V = A + bT \quad , \text{ where } V = \log_e v \quad , A = \log_e a \quad T = \log_e t$$

Normal Equations(N.E's)for the st.line $V = A + bT$

$$\begin{aligned} nA + b \sum T &= \sum V \\ A \sum T + b \sum T^2 &= \sum VT \end{aligned}$$

Here $n=4$

v	t	$V = \log_e v$	$T = \log_e t$	VT	T^2
350	61	5.8579	4.1109	24.0812	16.8995
400	26	5.9915	3.2581	19.5209	10.6152
500	7	6.2146	1.9459	12.093	3.7865
600	2.6	6.3969	0.9555	6.1122	0.913
		$\sum V$ = 24.4609	$\sum T$ = 10.2704	$\sum VT$ = 61.8073	$\sum T^2$ = 32.2142

$$\begin{aligned} \text{N.E's becomes} \quad 4A + 10.2704b &= 24.4609 \\ 10.2704A + 32.2142b &= 61.8073 \end{aligned}$$

Solving above equations we get,

$$A = 6.5539 \quad b = -0.1709$$

$$\text{But } \log_e a = 6.5539 \Rightarrow a = e^{6.5539} = 701.9766$$

$$\text{Thus , (*)} \Rightarrow v = at^b$$

$$v = (701.9766)t^{-0.1709} \quad , \text{ is the curve of best fit.}$$

3. Fit a curve of the form $y = ax^b$ for the following data

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.1	6.8	7.5

Statistical Methods

Statistics is a science of facts and figures which may be readily available or obtained through the process of direct enquiry or enumeration. It deals with the methods of collecting, classifying and analysing the data so as to draw some valid conclusion.

‘Correlation’ can be understood as co-relation. Co-variance of two independent magnitudes is known as correlation. ‘Regression’ is an estimation of one independent variable in terms of the other.

Correlation and Regression

This topic deals with data concerning independent observations.

Examples

1. Marks of individuals in two subjects
2. Height and weight of individuals
3. Amount involved in advertising a product and sale of product etc.,

We discuss the aspect of inter-relation between the independent variables

Correlation and Correlation Coefficient

Co-variation of two independent magnitudes is known as **correlation**. If two variables x and y are related in such a way that increase or decrease in one of them corresponds to increase or decrease in the other, we say that the variables are positively correlated. Also if increase or decrease in one of them corresponds to decrease or increase in the other, the variables are negatively correlated.

The numerical measure of correlation between two variables x and y is known as Karl Pearson’s coefficient of correlation usually by r and is defined as follows

$$r = \frac{\sum XY}{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}} \quad \text{where, } X = x - \bar{x}, Y = y - \bar{y}, \bar{x} = \frac{\sum x}{n}, \bar{y} = \frac{\sum y}{n}$$

Note:

$$1. \sigma_x^2 = \frac{\sum X^2}{n} \quad ; \quad \sigma_y^2 = \frac{\sum Y^2}{n} \quad ; \quad \sigma_x \sigma_y = \frac{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}}{n} \quad \text{or} \quad n \sigma_x \sigma_y = \sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}$$

2. The coefficient of correlation numerically does not exceed unity i.e., $-1 \leq r \leq 1$

3. If $r = \pm 1$, we say that x and y are perfectly correlated and if $r = 0$ we say that x and y are non correlated.

$$4. r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$$

Regression

Regression is an estimation of one independent variable in terms of the other. If x and y are correlated, the best fitting straight line in the least square sense gives reasonably a good relation between x and y .

The best fitting straight line of the form $y = ax + b$ is called the regression line of y on x and $x = ay + b$ is called the regression line of x on y .

The regression line of **y on x** is

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

The regression line of **x on y** is

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

$r \frac{\sigma_x}{\sigma_y}$; $r \frac{\sigma_y}{\sigma_x}$ are known as regression coefficients.

Note:

1. $r = \pm \sqrt{(\text{coeff. of } x) \cdot (\text{coeff. of } y)}$, the sign of r will be positive or negative according as the regression coefficients are positive or negative.
2. The lines of regression are also of the form $Y = \frac{\sum XY}{\sum X^2} (X)$ and $X = \frac{\sum XY}{\sum Y^2} (Y)$

Problems

1. Compute the coefficient of correlation and the equation of the lines of regression for the data

x	1	2	3	4	5	6	7
y	9	8	10	12	11	13	14

Sol: Now $\bar{x} = \frac{\sum x}{n}$, $\bar{y} = \frac{\sum y}{n}$

Here $\sum x = 28$ $\sum y = 77$, $n = 7$

$$\bar{x} = \frac{28}{7}$$

$$\bar{y} = \frac{77}{7}$$

$$\bar{x} = 4$$

$$\bar{y} = 11$$

x	y	X = x - $\bar{x}$ = x - 4	Y = y - $\bar{y}$ = y - 11	XY	X ²	Y ²
1	9	-3	-2	6	9	4
2	8	-2	-3	6	4	9
3	10	-1	-1	1	1	1
4	12	0	1	0	0	1
5	11	1	0	0	1	0
6	13	2	2	4	4	4
7	14	3	3	9	9	9
				$\sum XY$ = 26	$\sum X^2$ = 28	$\sum Y^2$ = 28

Wkt, $r = \frac{\sum XY}{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}}$

$$r = \frac{26}{\sqrt{28} \cdot \sqrt{28}}$$

$$r = \frac{26}{28}$$

$$r = 0.9286$$

The lines regression are given by

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

$$r \frac{\sigma_y}{\sigma_x} = \frac{\sum XY}{\sum X^2} = \frac{26}{28} = 0.9286$$

$$(y - 11) = 0.9286(x - 4)$$

$$y - 11 = 0.9286x - 3.7144$$

$$y = 0.9286x + 7.2856$$

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

$$r \frac{\sigma_x}{\sigma_y} = \frac{\sum XY}{\sum Y^2} = \frac{26}{28} = 0.9286$$

$$(x - 4) = 0.9286(y - 11)$$

$$x - 4 = 0.9286y - 10.2146$$

$$x = 0.9286y - 6.2146$$

2. Obtain the lines of regression and hence find the coefficient of correlation for the following data

x	1	2	3	4	5	6	7
y	9	8	10	12	11	13	14

$$\text{Sol: Now } \bar{x} = \frac{\sum x}{n}, \quad \bar{y} = \frac{\sum y}{n}$$

$$\text{Here } \sum x = 28 \quad \sum y = 77, \quad n = 7$$

$$\bar{x} = \frac{28}{7}, \quad \bar{x} = 4 \quad ; \quad \bar{y} = \frac{77}{7}, \quad \bar{y} = 11$$

x	y	$X = x - \bar{x}$ $= x - 4$	$Y = y - \bar{y}$ $= y - 11$	XY	X^2	Y^2
1	9	-3	-2	6	9	4
2	8	-2	-3	6	4	9
3	10	-1	-1	1	1	1
4	12	0	1	0	0	1
5	11	1	0	0	1	0
6	13	2	2	4	4	4
7	14	3	3	9	9	9
				$\sum XY$ $= 26$	$\sum X^2$ $= 28$	$\sum Y^2$ $= 28$

The lines regression are given by

$$(y - \bar{y}) = \frac{\sum XY}{\sum X^2} (x - \bar{x}) \qquad \frac{\sum XY}{\sum X^2} = \frac{26}{28} = \mathbf{0.9286}$$

$$(y - 11) = 0.9286(x - 4)$$

$$y - 11 = 0.9286x - 3.7144$$

$$\mathbf{y = 0.9286x + 7.2856}$$

$$\text{Also, } (x - \bar{x}) = \frac{\sum XY}{\sum Y^2} (y - \bar{y}) \qquad \frac{\sum XY}{\sum Y^2} = \frac{26}{28} = \mathbf{0.9286}$$

$$(x - 4) = 0.9286(y - 11)$$

$$x - 4 = 0.9286y - 10.2146$$

$$\mathbf{x = 0.9286y - 6.2146}$$

$$\text{Wkt } r = \sqrt{(\text{coeff. of } x) \cdot (\text{coeff. of } y)} \qquad r \text{ must be positive since regression coefficients are +ve}$$

$$r = \sqrt{(0.9286) \cdot (0.9286)}$$

$$\mathbf{r = 0.9286}$$

3. Find the correlation coefficient for the following two groups

A	92	89	87	86	83	77	71	63	53	50
B	86	83	91	77	68	85	52	82	37	57

$$\text{Sol: Now } \bar{x} = \frac{\sum x}{n}, \qquad \bar{y} = \frac{\sum y}{n}$$

$$\text{Here } \sum x = 751 \qquad \sum y = 718 \qquad , \quad n = 10$$

$$\bar{x} = \frac{751}{10}, \quad \bar{x} = \mathbf{75.1} \qquad ; \qquad \bar{y} = \frac{718}{10}, \quad \bar{y} = \mathbf{71.8}$$

x	y	X $= x - \bar{x}$ $= x$ $- 75.1$	Y $= y - \bar{y}$ $= y$ $- 71.8$	XY	X^2	Y^2
92	86	16.9	14.2	239.98	285.61	201.64
89	83	13.9	11.2	155.68	193.21	125.44
87	91	11.9	19.2	228.48	141.61	368.64
86	77	10.9	5.2	56.68	118.81	27.04
83	68	7.9	-3.8	-30.02	62.41	14.44
77	85	1.9	13.2	25.08	3.61	174.24
71	52	-4.1	-19.8	81.18	16.81	392.04
63	82	-12.1	10.2	-123.42	146.41	104.04
53	37	-22.1	-34.8	769.08	488.41	1211.04
50	57	-25.1	-14.8	371.48	630.01	219.04
				$\sum XY$ $= 1774.2$	$\sum X^2$ $= 2086.59$	$\sum Y^2$ $= 2837.6$

Wkt, $r = \frac{\sum XY}{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}}$

$$r = \frac{1774.2}{\sqrt{2086.59} \cdot \sqrt{2837.6}}$$

$$r = \frac{1774.2}{2433.29}$$

$$r = \mathbf{0.7291}$$

4. Find the correlation coefficient and the equation of the lines of regression for the following values of x and y

x	1	2	3	4	5
y	2	5	3	8	7

Sol: Now $\bar{x} = \frac{\sum x}{n}$, $\bar{y} = \frac{\sum y}{n}$

Here $\sum x = 15$ $\sum y = 25$, $n = 5$

$$\bar{x} = \frac{15}{5} , \bar{x} = 3 \quad ; \quad \bar{y} = \frac{25}{5} , \bar{y} = 5$$

x	y	$X = x - \bar{x}$ $= x - 3$	$Y = y - \bar{y}$ $= y - 5$	XY	X^2	Y^2
1	2	-2	-3	6	4	9
2	5	-1	0	0	1	0
3	3	0	-2	0	0	4
4	8	1	3	3	1	9
5	7	2	2	4	4	4
				$\sum XY$ $= 13$	$\sum X^2$ $= 10$	$\sum Y^2$ $= 26$

Wkt, $r = \frac{\sum XY}{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}}$

$$r = \frac{13}{\sqrt{10} \cdot \sqrt{26}}$$

$$r = 0.8062$$

Now, $\sigma_x^2 = \frac{\sum X^2}{n} = \frac{10}{5} = 2$; $\sigma_x^2 = 2$; $\sigma_x = \sqrt{2}$

$$\sigma_y^2 = \frac{\sum Y^2}{n} = \frac{26}{5} = 5.2 \quad ; \quad \sigma_y^2 = 5.2 \quad ; \quad \sigma_y = \sqrt{5.2}$$

The lines regression are given by

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

$$(y - 5) = (0.8062) \frac{\sqrt{2}}{\sqrt{5.2}} (x - 3)$$

$$(y - 5) = (1.3) (x - 3)$$

$$y - 5 = 1.3x - 3.9$$

$$y = 1.3x - 3.9 + 5$$

$$y = 1.3x + 1.1$$

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

$$(x - 3) = (0.8062) \frac{\sqrt{2}}{\sqrt{5.2}} (y - 5)$$

$$x - 3 = 0.5(y - 5)$$

$$x = 0.5y - 2.5 + 3$$

$$x = 0.5y + 0.5$$

5. The following data gives the age of husband (x) and the age of wife (y) in years. Form the two regression lines and calculate the age of husband corresponding to 16 years age of wife

x	36	23	27	28	28	29	30	31	33	35
y	29	18	20	22	27	21	29	27	29	28

Sol: Now $\bar{x} = \frac{\sum x}{n}$, $\bar{y} = \frac{\sum y}{n}$

Here $\sum x = 300$ $\sum y = 250$, $n = 10$

$$\bar{x} = \frac{300}{10} \quad \bar{y} = \frac{250}{10}$$

$$\bar{x} = 30 \quad \bar{y} = 25$$

x	y	X = x - $\bar{x}$ = x - 30	Y = y - $\bar{y}$ = y - 25	XY	X ²	Y ²
36	29	6	4	24	36	16
23	18	-7	-7	49	49	49
27	20	-3	-5	15	9	25
28	22	-2	-3	6	4	9
28	27	-2	2	-4	4	4
29	21	-1	-4	4	1	16
30	29	0	4	0	0	16
31	27	1	2	2	1	4
33	29	3	4	12	9	16
35	28	5	3	15	25	9
				$\sum XY$ = 123	$\sum X^2$ = 138	$\sum Y^2$ = 164

The lines regression are given by

$$(y - \bar{y}) = \frac{\sum XY}{\sum X^2} (x - \bar{x})$$

$$\frac{\sum XY}{\sum X^2} = \frac{123}{138} = 0.8913$$

$$(y - 25) = 0.8913(x - 30)$$

$$y - 25 = 0.8913x - 26.739$$

$$\mathbf{y = 0.8913x - 1.739}$$

$$\text{Also, } (x - \bar{x}) = \frac{\sum XY}{\sum Y^2} (y - \bar{y})$$

$$\frac{\sum XY}{\sum Y^2} = \frac{123}{164} = 0.75$$

$$(x - 30) = 0.75(y - 25)$$

$$x - 30 = 0.75y - 18.75$$

$$x = 0.75y + 11.25$$

$$\text{When } y = 16, x = 0.75(16) + 11.25 \quad x = 23.25 \cong 23$$

Thus husband's age is 23 years corresponding to wife age of 16 years.

6. In a partially destroyed lab record, only the lines of regression of y on x and x on y are available as $4x - 5y + 33 = 0$ and $20x - 9y = 107$ respectively. Calculate the coefficient of correlation between x and y .

Sol: Lines of regression of y on x and x on y are

$$4x - 5y + 33 = 0$$

$$20x - 9y = 107$$

$$5y = 4x + 33$$

$$20x = 9y + 107$$

$$y = \frac{4}{5}x + \frac{33}{5}$$

$$x = \frac{9}{20}y + \frac{107}{20}$$

$$\text{Coefficient of } x = \frac{4}{5}$$

$$\text{Coefficient of } y = \frac{9}{20}$$

The coefficient of correlation is

$$r = \sqrt{(\text{coeff. of } x) \cdot (\text{coeff. of } y)} \quad r \text{ must be positive since regression coefficients are +ve}$$

$$r = \sqrt{\left(\frac{4}{5}\right) \cdot \left(\frac{9}{20}\right)}$$

$$r = \sqrt{\frac{36}{100}}$$

$$r = \frac{6}{10}$$

$$r = 0.6$$

7. Given the equation of the regression lines $x = 19.13 - 0.87y$, $y = 11.64 - 0.5x$.

Compute the mean of x 's, mean of y 's and the coefficient of correlation.

Sol: To find $\bar{x}$, $\bar{y}$ and r .

We know that the regression lines passes through $(\bar{x}, \bar{y})$.

$$\therefore \bar{x} = 19.13 - 0.87\bar{y} \quad \text{and} \quad \bar{y} = 11.64 - 0.5\bar{x}$$

$$\text{That is, } \begin{aligned} \bar{x} + 0.87\bar{y} &= 19.13 \\ 0.5\bar{x} + \bar{y} &= 11.64 \end{aligned}$$

Solving above equations, we get

$$\bar{x} = 15.94 \quad \text{and} \quad \bar{y} = 3.67$$

Rewriting the equations

$$y = -0.5x + 11.64 \quad ; \quad x = -0.87y + 19.13$$

Coefficient of $x = -0.5$

Coefficient of $y = -0.87$

The coefficient of correlation is

$r = -\sqrt{(\text{coeff. of } x) \cdot (\text{coeff. of } y)}$ r must be negative since regression coefficients are -ve

$$r = -\sqrt{(-0.5) \cdot (-0.87)}$$

$$r = -0.6595$$

8. Compute $\bar{x}$, $\bar{y}$ and r from the following equation of the regression lines $2x + 3y + 1 = 0$ and $x + 6y - 4 = 0$

Sol: We know that the regression lines passes through $(\bar{x}, \bar{y})$.

$$\therefore 2\bar{x} + 3\bar{y} = -1 \quad \text{and} \quad \bar{x} + 6\bar{y} = 4$$

Solving above equations , we get

$$\bar{x} = -2 \quad \text{and} \quad \bar{y} = 1$$

Consider the regression lines

$$2x + 3y + 1 = 0 \quad \text{and} \quad x + 6y - 4 = 0$$

$$3y = -2x - 1 \quad \text{and} \quad x = -6y + 4$$

$$y = -\frac{2}{3}x - \frac{1}{3}$$

Coefficient of $x = -\frac{2}{3}$

Coefficient of $y = -6$

The coefficient of correlation is

$r = -\sqrt{(\text{coeff. of } x) \cdot (\text{coeff. of } y)}$ r must be negative since regression coefficients are -ve

$$r = -\sqrt{\left(-\frac{2}{3}\right) \cdot (-6)}$$

$$r = -\sqrt{4}$$

$r = -2$, this is absurd. Since r lies between -1 and +1.

Rewriting the regression lines

$$x + 6y - 4 = 0 \quad \text{and} \quad 2x + 3y + 1 = 0$$

$$6y = -x + 4 \quad \text{and} \quad 2x = -3y - 1$$

$$y = -\frac{1}{6}x + \frac{4}{6} \quad \text{and} \quad x = -\frac{3}{2}y - \frac{1}{2}$$

$$\text{Coefficient of } x = -\frac{1}{6}$$

$$\text{Coefficient of } y = -\frac{3}{2}$$

The coefficient of correlation is

$$r = -\sqrt{(\text{coeff. of } x) \cdot (\text{coeff. of } y)} \quad r \text{ must be negative since regression coefficients are -ve.}$$

$$r = -\sqrt{\left(-\frac{1}{6}\right) \cdot \left(-\frac{3}{2}\right)}$$

$$r = -\sqrt{\frac{1}{4}}$$

$$r = -0.5$$

Thus, $\bar{x} = -2$, $\bar{y} = 1$ and $r = -0.5$

9. Given the following table and $r = 0.8$, write down the equation of the lines of

	x-series	y-series
Mean	18	100
SD	14	20

regression and hence find the most probable value of y when $x = 70$.

Sol: By data $\bar{x} = 18$ and $\bar{y} = 100$; $\sigma_x = 14$ and $\sigma_y = 20$; $r = 0.8$

The equations of regression lines are

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

$$(y - 100) = (0.8) \frac{20}{14} (x - 18) \quad ; \quad (x - 18) = (0.8) \frac{14}{20} (y - 100)$$

$$y - 100 = 1.14(x - 18) \quad ; \quad x - 18 = 0.56(y - 100)$$

$$y = 1.14x + 79.48 \quad ; \quad x = 0.56y - 38$$

$$\text{When } x = 70, \quad y = 1.14(70) + 79.48 \quad ; \quad y = 159.28$$

10. Show that if θ is the angle between the lines of regression, then $\tan \theta = \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \left(\frac{1-r^2}{r} \right)$

Sol:

We know that if θ is acute, the angle between the lines $y = m_1x + c_1$ and $y = m_2x + c_2$ is given by,

$$\tan \theta = \frac{m_2 - m_1}{1 + m_1 m_2} \quad \text{-- (*)}$$

The equations of regression lines are ,

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \quad \text{--- (1)} \quad \text{and} \quad (x - \bar{x}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y}) \quad \text{--- (2)}$$

We rewrite the equation (2) as $(y - \bar{y}) = \frac{\sigma_y}{r \cdot \sigma_x} (x - \bar{x})$

Slopes of (1) and (2) are $m_1 = r \frac{\sigma_y}{\sigma_x}$ and $m_2 = \frac{\sigma_y}{r \cdot \sigma_x}$

Substituting the slopes in (*), we get

$$\tan \theta = \frac{\frac{\sigma_y}{r \cdot \sigma_x} - r \frac{\sigma_y}{\sigma_x}}{1 + r \frac{\sigma_y}{\sigma_x} \cdot \frac{\sigma_y}{r \cdot \sigma_x}}$$

$$\tan \theta = \frac{\frac{\sigma_y}{\sigma_x} \left(\frac{1}{r} - r \right)}{1 + r \cdot \frac{1}{r} \frac{\sigma_y^2}{\sigma_x^2}}$$

$$\tan \theta = \frac{\frac{\sigma_y}{\sigma_x} \left(\frac{1 - r^2}{r} \right)}{1 + \frac{\sigma_y^2}{\sigma_x^2}}$$

$$\tan \theta = \frac{\frac{\sigma_y}{\sigma_x} \left(\frac{1 - r^2}{r} \right)}{\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2}}$$

$$\tan \theta = \frac{\sigma_y}{\sigma_x} \cdot \frac{\sigma_x^2}{(\sigma_x^2 + \sigma_y^2)} \left(\frac{1 - r^2}{r} \right)$$

$$\tan \theta = \frac{\sigma_x \sigma_y}{(\sigma_x^2 + \sigma_y^2)} \left(\frac{1 - r^2}{r} \right)$$

Note: If $r = \pm 1$, $\tan \theta = 0 \Rightarrow \theta = 0$, which implies that the two regression lines coincide and hence the variables are perfectly correlated. Also if $r = 0$, $\tan \theta = \infty \Rightarrow \theta = \frac{\pi}{2}$, which implies that the lines are perpendicular and hence the variables are uncorrelated.

11. In a bivariate distribution $\sigma_x = \sigma_y$ and the angle between the regression lines is $\tan^{-1}(3)$. Find the correlation coefficient.

Sol: Let θ be the angle between the lines of regression, we have

$$\tan \theta = \frac{\sigma_x \sigma_y}{(\sigma_x^2 + \sigma_y^2)} \left(\frac{1 - r^2}{r} \right) \quad \text{---- (1)}$$

By data, $\theta = \tan^{-1}(3)$ or $\tan \theta = 3$ and $\sigma_x = \sigma_y$

$$\text{Hence (1)} \Rightarrow 3 = \frac{\sigma_x \cdot \sigma_x}{(\sigma_x^2 + \sigma_x^2)} \left(\frac{1 - r^2}{r} \right)$$

$$3 = \frac{\sigma_x^2}{2\sigma_x^2} \left(\frac{1-r^2}{r} \right)$$

$$6 = \left(\frac{1-r^2}{r} \right)$$

$$6r = 1 - r^2$$

$$r^2 + 6r - 1 = 0$$

$$r = \frac{-6 \pm \sqrt{6^2 - 4.1.(-1)}}{2.1}$$

$$r = \frac{-6 \pm \sqrt{36 + 4}}{2}$$

$$r = \frac{-6 \pm \sqrt{40}}{2}$$

$$r = \frac{-6 \pm 2\sqrt{10}}{2}$$

$$r = -3 \pm \sqrt{10}$$

$$r = +0.1623 \quad \text{or} \quad r = -6.1623(\text{neglect})$$

$$\text{Since } -1 \leq r \leq 1, r = 0.1623$$

12. If the coefficient of correlation between two variables x and y is 0.5 and the acute angle between their lines of regressions is $\tan^{-1} \left(\frac{3}{5} \right)$, show that $\sigma_x = 2\sigma_y$ or $\sigma_y = 2\sigma_x$.

$$\text{Sol: By data } r = 0.5 \quad \text{and} \quad \theta = \tan^{-1} \left(\frac{3}{5} \right) \quad \text{or} \quad \tan \theta = \frac{3}{5}$$

The angle between the lines of regression is given by

$$\tan \theta = \frac{\sigma_x \sigma_y}{(\sigma_x^2 + \sigma_y^2)} \left(\frac{1-r^2}{r} \right)$$

$$\frac{3}{5} = \frac{\sigma_x \sigma_y}{(\sigma_x^2 + \sigma_y^2)} \left(\frac{1 - \left(\frac{1}{2} \right)^2}{\frac{1}{2}} \right)$$

$$\frac{3}{5} = \frac{\sigma_x \sigma_y}{(\sigma_x^2 + \sigma_y^2)} \left(\frac{1 - \frac{1}{4}}{\frac{1}{2}} \right)$$

$$\frac{3}{5} = \frac{\sigma_x \sigma_y}{(\sigma_x^2 + \sigma_y^2)} \left(\frac{3}{4} \cdot \frac{2}{1} \right)$$

$$\frac{3}{5} = \frac{3\sigma_x \sigma_y}{2(\sigma_x^2 + \sigma_y^2)}$$

$$\frac{1}{5} = \frac{\sigma_x \sigma_y}{2(\sigma_x^2 + \sigma_y^2)}$$

$$2(\sigma_x^2 + \sigma_y^2) = 5\sigma_x \sigma_y$$

$$2\sigma_x^2 + 2\sigma_y^2 = 5\sigma_x \sigma_y$$

$$2\sigma_x^2 - 5\sigma_x \sigma_y + 2\sigma_y^2 = 0$$

$$2\sigma_x^2 - \sigma_x \sigma_y - 4\sigma_x \sigma_y + 2\sigma_y^2 = 0$$

$$\sigma_x(2\sigma_x - \sigma_y) - 2\sigma_y(2\sigma_x - \sigma_y) = 0$$

$$(2\sigma_x - \sigma_y)(\sigma_x - 2\sigma_y) = 0$$

Thus, we get

$$\sigma_y = 2\sigma_x \quad \text{or} \quad \sigma_x = 2\sigma_y$$

13. Compute the coefficient of correlation using the formula $r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x \sigma_y}$ and the lines of regression to the following data

x	1	2	3	4	5	6	7
y	9	8	10	12	11	13	14

Sol: We have $r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x \sigma_y}$

We shall prepare the table showing the columns of x, y, z, x^2, y^2, z^2 where $z = x - y$ so as to compute SD's of x, y, z .

x	y	z = x - y	x ²	y ²	z ²
1	9	-8	1	81	64
2	8	-6	4	64	36
3	10	-7	9	100	49
4	12	-8	16	144	64
5	11	-6	25	121	36

6	13	-7	36	169	49
7	14	-7	49	196	49
$\sum x = 28$	$\sum y = 77$	$\sum z = -49$	$\sum x^2 = 140$	$\sum y^2 = 875$	$\sum z^2 = 347$

Wkt, $\bar{x} = \frac{\sum x}{n}$ $\bar{y} = \frac{\sum y}{n}$, $\bar{z} = \frac{\sum z}{n}$

Here $\sum x = 28$ $\sum y = 77$ $\sum z = -49$, **$n = 7$**

$\bar{x} = \frac{28}{7}$ $\bar{y} = \frac{77}{7}$ $\bar{z} = \frac{-49}{7}$

$\bar{x} = 4$ $\bar{y} = 11$ $\bar{z} = -7$

$\sigma_x^2 = \frac{\sum x^2}{n} - (\bar{x})^2$ $\sigma_y^2 = \frac{\sum y^2}{n} - (\bar{y})^2$ $\sigma_z^2 = \frac{\sum z^2}{n} - (\bar{z})^2$

$\sigma_x^2 = \frac{140}{7} - 16$ $\sigma_y^2 = \frac{875}{7} - 121$ $\sigma_z^2 = \frac{347}{7} - 49$

$\sigma_x^2 = 4$ $\sigma_y^2 = 4$ $\sigma_z^2 = 0.57$

Now, $r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$

$r = \frac{4+4-0.57}{2.2.2}$

$r = .9287$

The lines regression are given by

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

$$(y - 11) = 0.9286(x - 4)$$

$$y - 11 = 0.9286x - 3.7144$$

$$y = 0.9286x + 7.2856$$

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

$$(x - 4) = 0.9286(y - 11)$$

$$x - 4 = 0.9286y - 10.2146$$

$$x = 0.9286y - 6.2146$$

Rank Correlation

The coefficient of correlation in respect of the ranks of some two characteristics of an individual or an observation is called **the rank correlation coefficient** usually denoted by ' ρ '.

$$\rho = 1 - \frac{6 \sum (x-y)^2}{n(n^2-1)} \quad \text{or} \quad \rho = 1 - \frac{6 \sum d^2}{n(n^2-1)}$$

$d = x - y$, being the difference in ranks.

NOTE:

1. If the rankings of x , y are entirely in the same order like for example,

$x: 1, 2, 3, 4, 5$; $y: 1, 2, 3, 4, 5$ then $\sum d^2 = \sum (x - y)^2 = 0$. This will give us $\rho = +1$ and is called **perfect direct correlation**.

If the rankings of x , y are entirely in the opposite order like for example,

$x: 1, 2, 3, 4, 5$; $y: 5, 4, 3, 2, 1$ then $\sum d^2 = \sum (x - y)^2 = (-4)^2 + (-2)^2 + (0)^2 + (2)^2 + (4)^2 = 40$.

This will give us $\rho = 1 - \frac{6(40)}{5(5^2-1)} = -1$ and is called **perfect inverse correlation**.

2. ρ in case of repeated ranks

If two or more magnitudes are repeated, then there will be a 'tie' for a particular rank. In such cases we assign the average rank to all those magnitudes and use the correlation factor $\frac{m(m^2-1)}{12}$ along with $\sum d^2$ in the formula for ρ where m denotes the number of times a magnitude is repeated. The correction factor must be added every time the 'tie' occurs for a particular rank.

$$\rho = 1 - \frac{6 \left[\sum d^2 + \frac{m(m^2-1)}{12} + \dots \right]}{n(n^2-1)}$$

For example, if there is a 'tie' for the second rank in two magnitudes, the rank $\frac{2+3}{2} = 2.5$ is assigned to both the magnitudes thereby weeding both 2nd and 3rd ranks. (Here $m = 2$)

Again, suppose there is a tie for 5th rank among three magnitudes we take the average of the three ranks in a row starting from 5. That is $\frac{5+6+7}{3} = 6$ which will be assigned to all the three magnitudes and the next ranking will only be from 8. (Here $m = 3$)

Problems

1. Ten competitors in a beauty contest are ranked by two judges in the following order. Compute the coefficient of rank correlation.

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

Sol: From the given data, $n = 10$ and we have

$$\sum d^2 = (1-6)^2 + (6-4)^2 + (5-9)^2 + (3-8)^2 + (10-1)^2 + (2-2)^2 + (4-3)^2 + (9-10)^2 + (7-5)^2 + (8-7)^2$$

$$\sum d^2 = 25 + 4 + 16 + 25 + 81 + 0 + 1 + 1 + 4 + 1 = 158$$

$$\text{Wkt, } \rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6(158)}{10(10^2-1)}$$

$$\rho = 0.042$$

2. Ten competitors in music contest are ranked by 3 judges A, B, C in the following order. Use the rank correlation coefficient to decide which pair of judges have the nearest approach to common taste of music.

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

Sol: We shall compute $\rho_{AB}, \rho_{BC}, \rho_{CA}$ with the help of the following table where d is the difference in ranks.

A	B	C	d_{AB}^2	d_{BC}^2	d_{CA}^2
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
			$\sum d_{AB}^2 = 200$	$\sum d_{BC}^2 = 214$	$\sum d_{CA}^2 = 60$

$$\text{Wkt, } \rho = 1 - \frac{6\sum d^2}{n(n^2-1)} \quad \text{and from the given data } n = 10$$

$$\text{Now, } \rho_{AB} = 1 - \frac{6(200)}{10(10^2-1)} = -0.21$$

$$\rho_{BC} = 1 - \frac{6(214)}{10(10^2-1)} = -0.297$$

$$\rho_{CA} = 1 - \frac{6(60)}{10(10^2-1)} = +0.636$$

It may be observed that ρ_{AB} and ρ_{BC} are negative which means that their tastes (A&B ; B&C) are opposite. But ρ_{CA} is positive and is near to 1.(Perfect correlation)

Thus, we conclude that the judges C and A have the nearest approach to common taste of music.

3. Ten students got the following percentage of marks in two subjects x and y . Compute their rank correlation coefficient.

Marks in x	78	36	98	25	75	82	90	62	65	39
Marks in y	84	51	91	60	68	62	86	58	53	47

Sol: We prepare the table consisting of the given data along with the ranks assigned according to their order of the magnitude. In the subject x , 98 will be awarded rank 1, 90 as rank 2 and so on.

Marks in x	Rank (x)	Marks in y	Rank (y)	$d = (x - y)$	$d^2 = (x - y)^2$
78	4	84	3	1	1
36	9	51	9	0	0
98	1	91	1	0	0
25	10	60	6	4	16
75	5	68	4	1	1
82	3	62	5	-2	4
90	2	86	2	0	0
62	7	58	7	0	0
65	6	53	8	-2	4
39	8	47	10	-2	4
					$\sum d^2 = 30$

Wkt, $\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$ and from the given data $n = 10$

$$\rho = 1 - \frac{6(30)}{10(10^2-1)}$$

$$\rho = 0.8182$$

4. Compute the rank correlation coefficient for the following data

x	68	64	75	50	64	80	75	40	55	64
y	62	58	68	45	81	60	68	48	50	70

Sol:

x	Rank (x)	y	Rank (y)	$d = (x - y)$	$d^2 = (x - y)^2$
68	4	62	5	-1	1
64	6	58	7	-1	1
75	2.5	68	3.5	-1	1
50	9	45	10	-1	1
64	6	81	1	5	25
80	1	60	6	-5	25
75	2.5	68	3.5	-1	1
40	10	48	9	1	1
55	8	50	8	0	0
64	6	70	2	4	16
					$\sum d^2 = 72$

Here in x the magnitude 75 is repeated twice for rank 2 and hence the average rank of $2,3 = \frac{2+3}{2} = 2.5$ is assigned to both ($m = 2$) ; Also in x the magnitude 64 is repeated thrice for rank 5 and hence the average rank of $5,6,7 = \frac{5+6+7}{3} = 6$ is assigned to the magnitudes 64 ($m = 3$). Further in y the magnitude 68 is repeated twice for rank 3 and hence the average rank of $3,4 = \frac{3+4}{2} = 3.5$ is assigned to both of them ($m = 2$).

$$\text{Wkt, } \rho = 1 - \frac{6\left[\sum d^2 + \frac{m(m^2-1)}{12} + \dots\right]}{n(n^2-1)} \quad \text{and from the given data } n = 10$$

$$\rho = 1 - \frac{6\left[72 + \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12}\right]}{10(10^2-1)}$$

$$\rho = 1 - \frac{6[72+0.5+2+0.5]}{990}$$

$$\rho = 0.545$$

5. The coefficient of rank correlation obtained by ten students in Statistics and Accountancy was 0.2. It was later discovered that the difference in ranks in the two subjects of one of the students was wrongly taken as 9 instead of 7. Find the correct rank correlation coefficient.

$$\text{Sol: Wkt, } \rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$\text{Here } n = 10 \quad \text{and} \quad \rho = 0.2$$

$$\text{Hence, } 0.2 = 1 - \frac{6\sum d^2}{10(10^2-1)}$$

$$\frac{6\sum d^2}{990} = 1 - 0.2$$

$$\sum d^2 \text{ (incorrect)} = \frac{990 \times 0.8}{6} = 132$$

$$\text{Now, correct } \sum d^2 = 132 - 9^2 + 7^2 = 100$$

$$\therefore \text{ Correct } \rho = 1 - \frac{6(100)}{10(10^2-1)}$$

$$\rho = 0.394$$