PDFZilla – Unregistered

PDFZilla - Unregistered

PDFZilla - Unregistered

Module-4
Part - I

Curve Fitting

It is the method of finding equation of a curve that approximates a given set of n data points and this equation is called best fitting equation.

Curve Fitting By The Method of Least squares

1. Fitting of a straight line
2. Fitting of a Second degree curve (Parabola)
3. Fitting of Exponential curve

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

The normal equations of fitting $y = ax + b$ are

$$\Sigma y = a\Sigma x + nb \quad \text{and}$$

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

The normal equations for fitting a straight line $y = a + bx$ are

$$\Sigma y = na + b\Sigma x \quad \text{and}$$

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

Working procedure to fit a straight line $y = ax + b$ or $y = a + bx$

- **Write the normal equations of the given curve.**
- **Prepare the relevant table and find the value of summation present in the normal equation and substitute.**
- **We find the parameters *'a' and 'b' by solving and then substitute in the* given equation.**

WORKED EXAMPLES

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

x	5	10	15	20	25
y	16	19	23	26	30

The normal equations of fitting $y = ax + b$ are

$$\Sigma y = a\Sigma x + nb \quad \text{and}$$

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

We prepare a relevant table as follows

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

Here, $n = 5$, $\Sigma x = 75$, $\Sigma y = 114$, $\Sigma xy = 1885$ and $\Sigma x^2 = 1375$

Substituting these values in the above normal equations, we get

$$114 = 75a + 5b$$

$$98 = 1375a + 75b$$

$$114 = 75a + 5b$$
$$98 = 1375a + 75b$$

Solving these equations, we get

$$a = 0.7, \quad b = 12.3$$

$\therefore$ The equation of the best fitting straight line is $y = 0.7x + 12.3$

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

x	0	1	2	3	4	5	6
y	–4	–2	0	2	4	6	8

The normal equations of fitting $y = ax + b$ are

$$\Sigma y = a\Sigma x + nb \quad \text{and}$$

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

We prepare a relevant table as follows

	x	y	xy	x^2
	0	– 4	0	0
	1	–2	–2	1
	2	0	0	4
	3	2	6	9
	4	4	16	16
	5	6	30	25
	6	8	48	36
Σ	21	14	98	91

Here, $n = 7$, $\Sigma x = 21$, $\Sigma y = 14$, $\Sigma xy = 98$ and $\Sigma x^2 = 91$

Substituting these values in the above normal equations, we get

$$14 = 21a + 7b$$
$$98 = 91a + 21b$$

Solving these equations, we get

$$a = 2, \quad b = -4$$

$\therefore$ The equation of the best fitting straight line is $y = 2x - 4$

Fit a straight line $y = a + bx$ for the data

x	0	1	3	6	8
y	1	3	2	5	4

The normal equations for fitting a straight line $y = a + bx$ are

$$\Sigma y = na + b\Sigma x \quad \text{and}$$

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

We prepare a relevant table as follows

	x	y	xy	x^2
	0	1	0	0
	1	3	3	1
	3	2	6	9
	6	5	30	36
	8	4	32	64
Σ	18	15	71	110

Here, $n = 5$, $\Sigma x = 18$, $\Sigma y = 15$, $\Sigma xy = 71$ and $\Sigma x^2 = 110$

Substituting these values in the above normal equations, we get

$$15 = 5a + 18b$$

$$71 = 18a + 110b$$

$$15 = 5a + 18b$$
$$71 = 18a + 110b$$

Solving these equations, we get

$$a = 1.6, \quad b = 0.38$$

$\therefore$ The equation of the best fitting straight line is $y = 1.6 + 0.38x$

If P is the pull required to lift a load W by means of a pulley block, find a linear law of the form P = mW + c connecting P and W, using the following data:

P	12	15	21	25
W	50	70	100	120

Also find P when W = 150 kgs

The normal equations of fitting $P = mW + c$ are

$$\Sigma P = m\Sigma W + nc \quad \text{and}$$

$$\Sigma WP = m\Sigma W^2 + c\Sigma x$$

We prepare a relevant table as follows

	W	P	WP	W^2
	50	12	600	2500
	70	15	1050	4900
	100	21	2100	10000
	120	25	3000	14400
Σ	340	73	6750	31800

Here, $n = 4$, $\Sigma W = 340$, $\Sigma P = 73$, $\Sigma WP = 6750$ and $\Sigma W^2 = 31800$

Substituting these values in the above normal equations, we get

$$73 = 340m + 4c$$
$$6750 = 31800m + 340c$$

Solving these equations, we get

$$m = 0.1879, \quad c = 2.2759$$

$\therefore$ The best fit of a line is $P = 0.1879W + 2.2759$

When W = 150 kg

$\Rightarrow P = 0.1879(150) + 2.2759 = 30.4635$ kg

EXERCISE

1. Fit a straight line to the following data

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	5

2. Fit a straight line to the following data

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

(VTU 2011)

3. Fit a straight line to the following data

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

and find the expected production in 2006.

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. The data are given below:

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 $Y = a + bP$

5. The results of measurement of electric resistance R of a copper bar at various temperatures $t\,^0C$ are listed below:

t	19	25	30	36	40	45	50
R	76	77	79	80	82	83	85

6. Fit a straight line to the following data

x	0	0.2	0.5	0.7	0.9	1.1	1.3
y	1.5	1.08	0.45	0.03	–0.39	–0.81	–1.23

Module-4
Part - 2

Fitting of a Second degree curve (Parabola) $y = ax^2 + bx + c$

The normal equations of fitting $y = ax^2 + bx + c$ are

$$\Sigma y = a\Sigma x^2 + b\Sigma x + nc$$

$$\Sigma xy = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x \quad \text{and}$$

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

The normal equations for fitting a straight line $y = a + bx + cx^2$ are

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

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \text{and}$$

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

Working procedure to fit a second degree curve $y = ax^2 + bx + c$ or $y = a + bx + cx^2$

- **Write the normal equations of the given curve.**
- **Prepare the relevant table and find the value of summation present in the normal equation and substitute.**
- **We find the parameters *'a', 'b' and 'c' by solving and then substitute in the* given equation.**

WORKED EXAMPLES

Fit a second degree curve $y = ax^2 + bx + c$ for the data

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	1.1	1.3	1.6	2.0	2.7	3.4	4.1

The normal equations of fitting $y = ax^2 + bx + c$ are

$$\Sigma y = a\Sigma x^2 + b\Sigma x + nc$$

$$\Sigma xy = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x \quad \text{and}$$

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

We prepare a relevant table as follows

	x	y	xy	x^2	x^3	x^4	x^2y
	1.0	1.1	1.1	1	1	1	1.1
	1.5	1.3	1.95	2.25	3.375	5.0625	2.925
	2.0	1.6	3.2	4.0	8.0	16.0	6.4
	2.5	2.0	5.0	6.25	15.625	39.0625	12.5
	3.0	2.7	8.1	9.0	27.0	81.0	24.3
	3.5	3.4	11.9	12.25	42.875	150.0625	41.65
	4.0	4.1	16.4	16	64	256	65.6
Σ	17.5	16.2	47.65	50.75	161.875	548.1875	154.475

Here, $n = 7$, $\Sigma x = 17.5$, $\Sigma y = 16.2$, $\Sigma xy = 47.65$, $\Sigma x^2 = 50.75$, $\Sigma x^3 = 161.875$, $\Sigma x^4 = 548.1875$ and $\Sigma x^2 y = 154.475$

Substituting these values in the above normal equations, we get

$$16.2 = 50.75a + 17.5b + 7c$$

$$47.65 = 161.875a + 50.75b + 17.5c$$

$$154.475 = 548.1875a + 161.875b + 50.75c$$

$$16.2 = 50.75a + 17.5b + 7c$$
$$47.65 = 161.875a + 50.75b + 17.5c$$
$$154.475 = 548.1875a + 161.875b + 50.75c$$

Solving these equations, we get

$a = 0.243$, $b = -0.193$ and $c = 1.04$

$\therefore$ The equation of the best fitting curve is $y = 0.243x^2 - 0.193x + 1.04$

Fit a second degree curve $y = a + bx + cx^2$ for the data

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

The normal equations for fitting a straight line $y = a + bx + cx^2$ are

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

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \text{and}$$

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

We prepare a relevant table as follows

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

Here, $n = 5$, $\Sigma x = 10$, $\Sigma y = 12.9$, $\Sigma xy = 37.1$, $\Sigma x^2 = 30$, $\Sigma x^3 = 100$, $\Sigma x^4 = 354$ and $\Sigma x^2 y = 130.3$

Substituting these values in the above normal equations, we get

$$12.9 = 4a + 10b + 30c$$
$$37.1 = 10a + 30b + 100c$$
$$130.3 = 30a + 100b + 354c$$

Solving these equations, we get

$a = 1.42$, $b = -1.07$ and $c = 0.55$

$\therefore$ The equation of the best fitting curve is $y = 1.42 - 1.07x + 0.55x^2$

Fit a second degree curve $y = a + bx + cx^2$ for the data

x	–3	–2	–1	0	1	2	3
y	4.63	2.11	0.67	0.09	0.63	2.15	4.58

The normal equations for fitting a straight line $y = a + bx + cx^2$ are

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

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \text{and}$$

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

We prepare a relevant table as follows

	x	y	xy	x^2	x^3	x^4	x^2y
	–3	4.63	–13.89	9	–27	81	41.67
	–2	2.11	– 4.22	4	–8	16	8.44
	–1	0.67	– 0.67	1	–1	1	0.67
	0	0.09	0	0	0	0	0
	1	0.63	0.63	1	1	1	0.63
	2	2.15	4.3	4	8	16	8.6
	3	4.58	13.74	9	27	81	41.22
Σ	0	14.86	– 0.11	28	0	196	101.23

Here, $n = 7$, $\Sigma x = 0$, $\Sigma y = 14.86$, $\Sigma xy = -0.11$, $\Sigma x^2 = 28$, $\Sigma x^3 = 0$, $\Sigma x^4 = 196$ and $\Sigma x^2 y = 101.23$

Substituting these values in the above normal equations, we get

$$14.86 = 7a + 0b + 28c$$
$$-0.11 = 0a + 28b + 0c$$
$$101.23 = 28a + 0b + 196c$$

$$14.86 = 7a + 0b + 28c$$
$$-0.11 = 0a + 28b + 0c$$
$$101.23 = 28a + 0b + 196c$$

Solving these equations, we get

$a = 0.1329$, $b = -0.00393$ and $c = 0.4975$

$\therefore$ The equation of the best fitting curve is

$$y = 0.1329 - 0.00393x + 0.4975x^2$$

Example **Find the best values of *a*, *b* and *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

The normal equations for fitting a straight line $y = a + bx + cx^2$ are

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

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \text{and}$$

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

We prepare a relevant table as follows

	x	y	xy	x^2	x^3	x^4	x^2y
	–2	-3.150	6.3	4	– 8	16	–12.6
	–1	–1.390	1.39	1	– 1	1	–1.39
	0	0.620	0	0	0	0	0
	1	2.880	2.880	1	1	1	2.880
	2	5.378	10.756	4	8	16	21.512
Σ	0	4.328	21.326	10	0	34	10.402

Here, $n = 5$, $\Sigma x = 0$, $\Sigma y = 4.328$, $\Sigma xy = 21.326$, $\Sigma x^2 = 10$, $\Sigma x^3 = 0$, $\Sigma x^4 = 34$ and $\Sigma x^2 y = 10.402$

Substituting these values in the above normal equations, we get

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

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

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

$$4.338 = 5a + 0b + 10c$$
$$21.326 = 0a + 10b + 0c$$
$$10.402 = 10a + 0b + 34c$$

Solving these equations, we get

$$a = 0.621, \quad b = 2.1326 \text{ and } c = 0.1233$$

$\therefore$ The best values of a, b and c for fitting of $y = a + bx + cx^2$ are

$$a = 0.621, \quad b = 2.1326 \text{ and } c = 0.1233$$

Fit a second degree curve $y = ax^2 + bx + c$ for the data

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

The normal equations for fitting a straight line $y = a + bx + cx^2$ are

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

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \text{and}$$

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

We prepare a relevant table as follows

	x	y	xy	x^2	x^3	x^4	x^2y
	1	10	10	1	1	1	10
	2	12	24	4	8	16	48
	3	13	39	9	27	81	117
	4	16	64	16	64	256	256
	5	19	95	25	125	625	475
Σ	15	61	232	55	225	979	906

Here, $n = 5$, $\Sigma x = 15$, $\Sigma y = 61$, $\Sigma xy = 232$, $\Sigma x^2 = 55$, $\Sigma x^3 = 225$, $\Sigma x^4 = 979$ and

$\Sigma x^2 y = 906$

Substituting these values in the above normal equations, we get

$$61 = 55a + 15b + 5c$$
$$232 = 225a + 55b + 15c$$
$$906 = 979a + 225b + 55c$$

$$61 = 55a + 15b + 5c$$
$$232 = 225a + 55b + 15c$$
$$906 = 979a + 225b + 55c$$

Solving these equations, we get

$a = 0.29$, $b = 0.46$ and $c = 9.43$

$\therefore$ The equation of the best fitting curve is $y = 0.29x^2 + 0.46x + 9.43$

Fit a second degree 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

The normal equations of fitting $y = ax^2 + bx + c$ are

$$\Sigma y = a\Sigma x^2 + b\Sigma x + nc$$

$$\Sigma xy = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x \quad \text{and}$$

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

We prepare a relevant table as follows

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

Here, $n = 9$, $\Sigma x = 45$, $\Sigma y = 74$, $\Sigma xy = 421$, $\Sigma x^2 = 285$, $\Sigma x^3 = 2025$, $\Sigma x^4 = 15333$ and $\Sigma x^2 y = 2771$

Substituting these values in the above normal equations, we get

$$74 = 285a + 45b + 9c$$

$$421 = 2025a + 285b + 45c$$

$$2771 = 15333a + 2025b + 285c$$

Solving these equations, we get

$a = -0.2673$, $b = 3.523$ and $c = -0.9282$

$\therefore$ The equation of the best fitting curve is

$$y = -0.2673x^2 + 3.523x - 0.9282$$

EXERCISE

1. Fit a second degree Parabola to the following data:

x	1	2	3	4	5	6	7	8	9	10
y	124	129	140	159	228	289	315	302	263	210

2. Fit a second degree Parabola to the following data:

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

(VTU 2011)

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

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

(VTU 2013)

4. The velocity V of a liquid is known to vary with temperature according to a quadratic law $V = a + bT + cT^2$. Find the best values of a, b and c for the data:

T	1	2	3	4	5	6	7
V	2.31	2.01	3.80	1.66	1.55	1.47	1.41

5. The following table gives the results of the measurements of train resistances, V is the velocity in miles per hour, R is the resistance in pounds per ton:

V	20	40	60	80	100	120
R	5.5	9.1	14.9	22.8	33.3	46.0

If R is related to V by the relation $R = a + bV + cV^2$, find a, b and c.

6. Fit a second degree Parabola $y = ax^2 + bx + c$ to the following data:

x	10	20	30	40	50	60
y	157	179	210	252	302	361

Module-4
Part - 3

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

$$y = ax^b \qquad \text{----- (1)}$$

Taking *log* on both sides

$$\log y = \log\left(ax^b\right)$$

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

$$Y = A + BX \qquad \text{----- (2)}$$

where $Y = \log y,\ A = \log a,\ B = b$ and $X = \log x$

The normal equations of (2) are

$$\Sigma Y = nA + B\Sigma X \quad \text{and}$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2$$

Solving these equations we obtain A and B from which $a = e^A$ and $b = B$ can be found.

Substituting these values of a and b in (1), we obtain the equation of the best fitting curve.

WORKED EXAMPLES

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

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

$$y = ax^b \qquad \text{----- (1)}$$

Taking *log* on both sides

$$\log y = \log\left(ax^b\right)$$

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

$$Y = A + BX \qquad \text{----- (2)}$$

where $Y = \log y$, $A = \log a$, $B = b$ and $X = \log x$

The normal equations of (2) are

$$\Sigma Y = nA + B\Sigma X \quad \text{and}$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2$$

where $Y = \log y$, $A = \log a$, $B = b$ and $X = \log x$

We prepare a relevant table as follows

	x	y	$X = \log x$	$Y = \log y$	XY	X^2
	1	2.98	0	1.0919	0	0
	2	4.26	0.6931	1.4492	1.0044	0.4804
	3	5.21	1.0986	1.6506	1.8133	1.2069
	4	6.1	1.3863	1.8083	2.5068	1.9218
	5	6.8	1.6094	1.9169	3.0851	2.5909
	6	7.5	1.7918	2.0149	3.6103	3.2105
Σ	21	32.85	6.5792	9.9318	12.0199	9.4098

Here, $n = 6$, $\Sigma X = 6.5792$, $\Sigma Y = 9.9318$, $\Sigma XY = 12.0199$ and $\Sigma X^2 = 9.4098$

Substituting these values in the above normal equations, we get

$$9.9318 = 6A + 6.5792B$$
$$12.0199 = 6.5792A + 9.4098B$$

Solving these equations, we get

$$A = 1.0912, \quad B = 0.5144$$

$$\Rightarrow \quad a = e^A = e^{1.0912} = 2.9778 \quad \text{and } b = B = 0.5144$$

$\therefore$ The equation of the best fitting curve is $y = (2.9778)x^{0.5144}$

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

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

$$y = ax^b \qquad \text{----- (1)}$$

Taking *log* on both sides

$$\log y = \log\left(ax^b\right)$$

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

$$Y = A + BX \qquad \text{----- (2)}$$

where $Y = \log y,\ A = \log a,\ B = b$ and $X = \log x$

The normal equations of (2) are

$$\Sigma Y = nA + B\Sigma X \quad \text{and}$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2$$

where $Y = \log y$, $A = \log a$, $B = b$ and $X = \log x$

We prepare a relevant table as follows

	x	y	$X = \log x$	$Y = \log y$	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	2.5903
Σ	15	27.5	4.7874	6.1092	9.0804	6.1993

Here, $n = 5$, $\Sigma X = 4.7874$, $\Sigma Y = 6.1092$, $\Sigma XY = 9.0804$ and $\Sigma X^2 = 6.1993$

Substituting these values in the above normal equations, we get

$$6.1092 = 5A + 4.7874B$$
$$9.0804 = 4.7874A + 6.1993B$$

Solving these equations, we get

$$A = -0.6929, \quad B = 1.9998$$

$$\Rightarrow \quad a = e^{A} = e^{-0.6929} = 0.5 \qquad \text{and} \quad b = B = 1.9998$$

$\therefore$ The equation of the best fitting curve is $y = (0.5)x^{1.9998}$

An experiment gave the following values:

v (ft/min)	350	400	500	600
t (min)	61	26	7	2.6

It is known that v and t are connected by the relation $v = at^b$. Find the best possible values of a and b.

$$v = at^b \qquad \text{----- (1)}$$

Taking *log* on both sides

$$\log v = \log\left(a t^b\right)$$

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

$$Y = A + BX \qquad \text{----- (2)}$$

where $Y = \log v$, $A = \log a$, $B = b$ and $X = \log t$

The normal equations of (2) are

$$\Sigma Y = nA + B\Sigma X \quad \text{and}$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2$$

where $Y = \log v$, $A = \log a$, $B = b$ and $X = \log t$

We prepare a relevant table as follows

	v	t	$X = \log t$	$Y = \log v$	XY	X^2
	350	61	4.1109	5.8579	24.0812	16.8995
	400	26	3.2581	5.9915	19.5209	10.6152
	500	7	1.9459	6.2146	12.0930	3.7865
	600	2.6	0.9555	6.3969	6.1122	0.9130
Σ	1850	96.6	10.2704	24.4609	61.8073	29.2142

Here, $n = 4$, $\Sigma X = 10.2704$, $\Sigma Y = 24.4609$, $\Sigma XY = 61.8073$ and $\Sigma X^2 = 29.2142$

Substituting these values in the above normal equations, we get

$$24.4609 = 4A + 10.2704B$$
$$61.8073 = 10.2704A + 29.2142B$$

Solving these equations, we get

$$A = 7.0167, \quad B = -0.3511$$

$$\Rightarrow \quad a = e^A = e^{7.0167} = 1115.10 \quad \text{and} \quad b = B = -0.3511$$

EXERCISE

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

x	20	16	10	11	14
y	22	41	120	89	56

(VTU 2015)

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

x	1	2	4	6
y	6	4	2	2

3. Predict y at $x = 3.75$, by fitting a curve $y = ax^b$ for the data

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.10	6.80	7.50

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

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

Module-4
Part – 4
STATISTICAL METHODS

Mean (Arithmetic mean): If $x_1, x_2, x_3, \ldots, x_n$ are set of n values of a variate, then the mean is given by

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\Sigma x_i}{n}.$$

In a frequency distribution, if $x_1, x_2, x_3, \ldots, x_n$ be the mid-values of the class-intervals having frequencies $f_1, f_2, f_3, \ldots, f_n$ respectively, we have

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots + x_n f_n}{f_1 + f_2 + f_3 + \ldots + f_n} = \frac{\Sigma x_i f_i}{\Sigma f_i}$$

Variance: If a variate x take the values $x_1, x_2, x_3, \ldots, x_n$ then the variance V is defined as follows:

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

Also for a grouped data,

$$V = \frac{\Sigma f \left(x - \bar{x}\right)^2}{\Sigma f}$$

Standard deviation (S.D):

$$\sigma = \sqrt{V} \text{ or } \sigma^2 = V$$

$$\Rightarrow \quad \sigma = \sqrt{\frac{\Sigma\left(x-\bar{x}\right)^2}{n}} = \sqrt{\frac{\Sigma X^2}{n}} \text{ where } X = x - \bar{x}$$

$$\text{or} \quad \sigma^2 = \frac{\Sigma\left(x-\bar{x}\right)^2}{n} = \frac{\Sigma X^2}{n}$$

Alternative formula for σ^2

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

Correlation

- **The two variables *x and y are related in such a way that an increase in* one is accompanied by an increase or decrease in the other is called co-variation.**
- **Co-variation of two independent magnitudes is known as correlation.**

Correlation

- **Correlation is Positive when the values increase together and Correlation is Negative when one value decreases as the other increases.**

Coefficient of correlation

The numerical measure of correlation between two variables *x and y is* known as **Karl Pearson's coefficient of correlation** or **simply coefficient of correlation** and is denoted by ***r*** and is defined by

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

$$r = \frac{\Sigma XY}{n\sigma_x \sigma_y}$$

Where X = deviation from the mean = $x - \bar{x}$, Y = deviation from the mean = $y - \bar{y}$, n = number of values of the two variables, σ_x = S.D. of $x-series$, σ_y = S.D. of $y-series$.

Substituting the value of σ_x and σ_y in the above formula, we get

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$$

Another form of the above formula is

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{\left\{n\Sigma x^2 - (\Sigma x)^2\right\} \times \left\{n\Sigma y^2 - (\Sigma y)^2\right\}}}$$

Property of coefficient of correlation

The coefficient of correlation numerically does not exceed '1',

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

Formula for correlation coefficient

The formula to compute coefficient of correlation is

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

Here,

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

$$\bar{x} = \frac{\Sigma x}{n}$$

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

$$\bar{y} = \frac{\Sigma y}{n}$$

$$\sigma_{x-y}^2 = \frac{\Sigma (x-y)^2}{n} - \left(\overline{x-y}\right)^2$$

$$\overline{(x-y)} = \frac{\Sigma (x-y)}{n}$$

Regression

It is an estimation of one independent variable in terms of the other.

Example: The best fitting straight line of the form *y = a + bx* is called the regression of *y* on *x* and *x = a + by* is called the regression of *x* on *y*.

Equation of the Regression lines

The regression line of *y on x* is

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

The regression line of *x on y* is

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

Here $r\frac{\sigma_y}{\sigma_x}$ and $r\frac{\sigma_x}{\sigma_y}$ are the regression coefficients. Their product is r^2

Here $r\frac{\sigma_y}{\sigma_x}$ and $r\frac{\sigma_x}{\sigma_y}$ are the regression coefficients. Their product is r^2

Thus 'r' is the geometric mean of regression coefficients,

$$\text{i.e., } r = \pm\sqrt{(\text{coefficient of } x)\times(\text{coefficient of } y)}$$

Note: If both the coefficients are positive then we have to consider *r value as* positive and if both coefficients are negative then we have to consider *r value as* negative.

Angle between the lines of regression

The angle between the lines of regression is

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

Proof:

If m_1 and m_2 are the slopes of two lines then the angle between the lines is given by

$$\tan\theta = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right| \qquad \text{----- (1)}$$

The lines of regression are

$$y-\bar{y}=r\frac{\sigma_y}{\sigma_x}\left(x-\bar{x}\right) \text{ and } x-\bar{x}=r\frac{\sigma_x}{\sigma_y}\left(y-\bar{y}\right)$$

$$\Rightarrow \quad y-\bar{y}=r\frac{\sigma_y}{\sigma_x}\left(x-\bar{x}\right) \text{ and } y-\bar{y}=\frac{\sigma_y}{r\sigma_x}\left(x-\bar{x}\right)$$

The slopes of regression lines are $m_1=\frac{r\sigma_y}{\sigma_x}$ and $m_2=\frac{\sigma_y}{r\sigma_x}$

Using these in (1), we get

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

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

$\therefore$ The angle between the lines of regression is

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

Its significance when $r = 0$ and $r = \pm 1$

Case – 1: If $r = 0$ then $\tan\theta = \infty$

$\Rightarrow$ $\theta = \frac{\pi}{2}$

$\therefore$ The lines are perpendicular

Case – 2: If $r = \pm 1$ then $\tan\theta = 0$

$\Rightarrow$ $\theta = 0$

$\therefore$ The lines are parallel

Example 1: Find the coefficient of correlation and the lines of regression for the data

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

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

We prepare a relevant table as follows

	x	y	$x - y$	x^2	y^2	$(x-y)^2$
	1	2	−1	1	4	1
	2	5	−3	4	25	9
	3	3	0	9	9	0
	4	8	−4	16	64	16
	5	7	−2	25	49	4
Σ	15	25	−10	55	151	30

Here, $n = 5$, $\Sigma x = 15$, $\Sigma y = 25$, $\Sigma(x-y) = -10$, $\Sigma x^2 = 55$, $\Sigma y^2 = 151$ and $\Sigma(x-y)^2 = 30$.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{15}{5} = 3 \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{25}{5} = 5$$

$$\overline{(x-y)} = \frac{\Sigma(x-y)}{n} = -\frac{10}{5} = -2$$

$$\sigma_x^2 = \frac{\Sigma x^2}{n} - \left(\bar{x}\right)^2 = \frac{55}{5} - (3)^2 = 2$$

$$\sigma_y^2 = \frac{\Sigma y^2}{n} - \left(\bar{y}\right)^2 = \frac{151}{5} - (5)^2 = 5.2$$

$$\sigma_{x-y}^2 = \frac{\Sigma(x-y)^2}{n} - \left(\overline{x-y}\right)^2 = \frac{30}{5} - (-2)^2 = 2$$

Substituting these values in r, we get

$$r = \frac{2 + 5.2 - 2}{2\left(\sqrt{2}\right)\left(\sqrt{5.2}\right)} = 0.81$$

Also, we have the equations of the regression lines are

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}\left(x - \bar{x}\right) \text{ and } x - \bar{x} = r\frac{\sigma_x}{\sigma_y}\left(y - \bar{y}\right)$$

$$y - 5 = (0.81)\frac{\sqrt{5.2}}{\sqrt{2}}(x - 3) \text{ and } x - 3 = (0.81)\frac{\sqrt{2}}{\sqrt{5.2}}(y - 5)$$

$$y - 5 = 1.306(x - 3) \text{ and } x - 3 = 0.502(y - 5)$$

$$y = 1.306x + 1.082 \text{ and } x = 0.502y + 0.49$$

These are the lines of regression.

Example 2: Find the coefficient of correlation and the lines of regression for the data

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

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

We prepare a relevant table as follows

	x	y	$x-y$	x^2	y^2	$(x-y)^2$
	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
Σ	28	77	–49	140	875	347

Here, $n = 7$, $\Sigma x = 28$, $\Sigma y = 77$, $\Sigma(x-y) = -49$, $\Sigma x^2 = 140$, $\Sigma y^2 = 875$ and $\Sigma(x-y)^2 = 347$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{28}{7} = 4 \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{77}{7} = 11$$

$$\overline{(x-y)} = \frac{\Sigma(x-y)}{n} = -\frac{49}{7} = -7$$

$$\sigma_x^2 = \frac{\Sigma x^2}{n} - \left(\bar{x}\right)^2 = \frac{140}{7} - (4)^2 = 4$$

$$\sigma_y^2 = \frac{\Sigma y^2}{n} - \left(\bar{y}\right)^2 = \frac{875}{7} - (11)^2 = 4$$

$$\sigma_{x-y}^2 = \frac{\Sigma(x-y)^2}{n} - \left(\overline{x-y}\right)^2 = \frac{347}{7} - (-7)^2 = 0.57$$

Substituting these values in r, we get

$$r = \frac{4+4-0.57}{2(2)(2)} = 0.93$$

Also, we have the equations of the regression lines are

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

$$y - 11 = (0.93)\frac{2}{2}(x - 4) \text{ and } x - 4 = (0.93)\frac{2}{2}(y - 11)$$

$$y - 11 = 0.93(x - 4) \text{ and } x - 4 = 0.93(y - 11)$$

$$y = 0.93x + 7.28 \text{ and } x = 0.93y - 6.23$$

These are the lines of regression.

Example 3: Find the coefficient of correlation and the lines of regression for the data

x	1	2	3	4	5	6	7	8	9	10
y	10	12	16	28	25	36	41	49	40	50

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

We prepare a relevant table as follows

	x	y	$x-y$	x^2	y^2	$(x-y)^2$
	1	9	–9	1	100	81
	2	12	–10	4	144	100
	3	16	–13	9	256	169
	4	28	–24	16	784	576
	5	25	–20	25	625	400
	6	36	–30	36	1296	900
	7	41	–34	49	1681	1156
	8	49	–41	64	2401	1681
	9	40	–31	81	1600	961
	10	50	–40	100	2500	1600
Σ	55	307	–252	385	11387	7624

Here, $n = 10$, $\Sigma x = 55$, $\Sigma y = 307$, $\Sigma(x-y) = -252$, $\Sigma x^2 = 385$, $\Sigma y^2 = 11387$ and $\Sigma(x-y)^2 = 7624$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{55}{10} = 5.5 \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{307}{10} = 30.7$$

$$\overline{(x-y)} = \frac{\Sigma(x-y)}{n} = -\frac{252}{10} = -25.2$$

$$\sigma_x^2 = \frac{\Sigma x^2}{n} - \left(\bar{x}\right)^2 = \frac{385}{10} - (5.5)^2 = 8.25$$

$$\sigma_y^2 = \frac{\Sigma y^2}{n} - \left(\bar{y}\right)^2 = \frac{11387}{10} - (30.7)^2 = 196.21$$

$$\sigma_{x-y}^2 = \frac{\Sigma(x-y)^2}{n} - \left(\overline{x-y}\right)^2 = \frac{7624}{10} - (-25.2)^2 = 127.36$$

Substituting these values in r, we get

$$r = \frac{8.25 + 196.21 - 127.36}{2\left(\sqrt{8.25}\right)\left(\sqrt{196.21}\right)} = 0.96$$

Also, we have the equations of the regression lines are

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}\left(x - \bar{x}\right) \text{ and } x - \bar{x} = r\frac{\sigma_x}{\sigma_y}\left(y - \bar{y}\right)$$

$$y - 30.7 = (0.96)\frac{\sqrt{196.21}}{\sqrt{8.25}}(x - 5.5) \text{ and } x - 5.5 = (0.96)\frac{\sqrt{8.25}}{\sqrt{196.21}}(y - 30.7)$$

$$y = 4.686x + 4.927 \text{ and } x = 0.197y - 0.548$$

These are the lines of regression.

Example **Find the lines of regression for the following data**

Ages of cars (in years)	2	4	6	7	8	10	12
Annual Maintenance cost (in hundreds)	16	15	18	19	17	21	20

Hence estimate the maintenance cost if the age of a car is 9 years and the age of a car if the maintenance cost is Rs.1550/-

We prepare a relevant table as follows

	x	y	$x-y$	x^2	y^2	$(x-y)^2$
	2	16	−14	4	256	196
	4	15	−11	16	225	121
	6	18	−12	36	324	144
	7	19	−12	49	361	144
	8	17	−9	64	289	81
	10	21	−11	100	441	121
	12	20	−8	144	400	64
Σ	49	126	−77	413	2296	871

Here, $n = 7$, $\Sigma x = 49$, $\Sigma y = 126$, $\Sigma(x - y) = -77$, $\Sigma x^2 = 413$, $\Sigma y^2 = 2296$ and $\Sigma(x - y)^2 = 871$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{49}{7} = 7$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{126}{7} = 18$$

$$\overline{(x-y)} = \frac{\Sigma(x-y)}{n} = -\frac{77}{7} = -11$$

$$\sigma_x^2 = \frac{\Sigma x^2}{n} - \left(\bar{x}\right)^2 = \frac{413}{7} - (7)^2 = 10$$

$$\sigma_y^2 = \frac{\Sigma y^2}{n} - \left(\bar{y}\right)^2 = \frac{2296}{7} - (18)^2 = 4$$

$$\sigma_{x-y}^2 = \frac{\Sigma(x-y)^2}{n} - \left(\overline{x-y}\right)^2 = \frac{871}{7} - (-11)^2 = 3.429$$

Substituting these values in r, we get

$$r = \frac{10 + 4 - 3.429}{2\left(\sqrt{10}\right)\left(\sqrt{4}\right)} = 0.8357$$

Also, we have the equations of the regression lines are

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}\left(x - \bar{x}\right) \text{ and } x - \bar{x} = r\frac{\sigma_x}{\sigma_y}\left(y - \bar{y}\right)$$

$$y - 18 = (0.8357)\frac{2}{\sqrt{10}}(x - 7) \text{ and } x - 7 = (0.8357)\frac{\sqrt{10}}{2}(y - 18)$$

$$y = 0.52857x + 14.3 \text{ and } x = 1.321429y - 16.786$$

These are the lines of regression.

Also,

(i) If the age of the car $x = 9$ then the maintenance cost y is given by the line of regression of y on x is

$$y = 0.52857\ (9) + 14.3 = 19.06 \text{ hundreds.}$$

$\therefore$ The maintenance cost = Rs.1906/-

(ii) If the maintenance cost y is Rs.1550/- i.e., 15.5 hundreds then the age of the car x is given by the line of regression of x on y is

$$x = 1.321429\ (15.5) - 16.786 = 3.7 \text{ years}$$

$\therefore$ Age of car = 3.7 years.

Example : In the following table are recorded data showing the test scores made by salesmen on an intelligence test and their weekly sales:

Salesmen	1	2	3	4	5	6	7	8	9	10
Test scores	40	70	50	60	80	50	90	40	60	60
Sales	2.5	6.0	4.5	5.0	4.5	2.0	5.5	3.0	4.5	3.0

Calculate the regression line of sales on test scores and estimate the most probable weekly sales volume if a salesman makes a score of 70.

Example **In a partially destroyed laboratory record of correlation data, the following results only are available:**
Variance of x is 9 and the regression equations are $4x - 5y + 33 = 0$ and $20x - 9y = 107$. Calculate (i) the mean values of x and y, (ii) standard deviation of y and (iii) the coefficient of correlation between x and y.

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

$$\therefore \quad 4\bar{x} - 5\bar{y} = -33$$

$$20\bar{x} - 9\bar{y} = 107$$

Solving these equations, we get

$$\bar{x} = 13 \quad \text{and} \quad \bar{y} = 17$$

We rewrite the given regression equations as

$$y = \frac{4}{5}x + \frac{33}{5} \quad \text{and} \quad x = \frac{9}{20}y + \frac{107}{20} \qquad \text{----- (1)}$$

We have, the coefficient of correlation between x and y is

$$r = \pm\sqrt{(\text{coefficient of } x)(\text{coefficient of } y)}$$

$$r = \pm\sqrt{\left(\frac{4}{5}\right)\left(\frac{9}{20}\right)} = \pm\frac{3}{5}$$

$$r = \frac{3}{5} \quad (\because \text{ both the coefficients are positive})$$

Given that, $\sigma_x^2 = 9 \Rightarrow \sigma_x = 3$

We have, $$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}\left(x - \bar{x}\right) \qquad \text{----- (2)}$$

Comparing (2) with (1),

we have $$r\frac{\sigma_y}{\sigma_x} = \frac{4}{5}$$

we have $r\frac{\sigma_y}{\sigma_x}=\frac{4}{5}$

$$\sigma_y=\frac{4}{5}\times\frac{\sigma_x}{r}$$

$$=\frac{4}{5}\times\frac{3}{(3/5)}=4$$

Example **Compute $\bar{x}, \bar{y}$ and r from the equations of the regression lines $2x+3y+1=0$ and $x+6y=4$.**

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

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

Solving these equations, we get

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

We rewrite the given regression equations as

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

We have, the coefficient of correlation between x and y is

$$r=\pm\sqrt{(\text{coefficient of } x)(\text{coefficient of } y)}$$

$$r = \pm\sqrt{\left(-\frac{2}{3}\right)(-6)} \quad = \pm\sqrt{4} = \pm 2$$

which is not possible because $-1 \le r \le 1$.

Hence, we rewrite the given regression equations in other form as

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

We have, the coefficient of correlation between x and y is

$$r = \pm\sqrt{(\text{coefficient of } x)(\text{coefficient of } y)}$$

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

$$r = -\frac{1}{2} \quad (\because \text{both the coefficients are negative})$$

Example If $2x - 3y = 0$ and $3x - 2y = 5$ are the lines of regression of the variables x and y. Find the following:

(i) Mean of x and y

(ii) Coefficient of correlation between x and y.

(iii) Angle between the lines of regression.

(iv) Standard deviation of x when variance of y is 2.

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

$$\therefore \quad 2\bar{x} - 3\bar{y} = 0$$

$$3\bar{x} - 2\bar{y} = 5$$

Solving these equations, we get

$$\bar{x} = 3 \quad \text{and} \quad \bar{y} = 2$$

We rewrite the given regression equations as

$$y = \frac{2}{3}x \quad \text{and} \quad x = \frac{2}{3}y + 5 \qquad \text{----- (1)}$$

We have, the coefficient of correlation between x and y is

$$r = \pm\sqrt{(\text{coefficient of } x)(\text{coefficient of } y)}$$

$$r = \pm\sqrt{\left(\frac{2}{3}\right)\left(\frac{2}{3}\right)} = \pm\frac{2}{3}$$

$$r = \frac{2}{3} \quad (\because \text{both the coefficients are positive})$$

The angle between the lines of regression is $\tan\theta = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right|$

Now, $m_1 = -\frac{\text{coefficient of } x}{\text{coefficient of } y} = -\frac{2}{-3} = \frac{2}{3}$ and

$$m_2 = -\frac{\text{coefficient of } x}{\text{coefficient of } y} = -\frac{3}{-2} = \frac{3}{2}$$

$$\therefore \quad \tan\theta = \left|\frac{(2/3)-(3/2)}{1+(2/3)(3/2)}\right| = \left|\frac{-(5/6)}{2}\right| = \frac{5}{12}$$

$$\Rightarrow \qquad \theta = \tan^{-1}\left(\frac{5}{12}\right)$$

Given that, $\sigma_y^2 = 2 \quad \Rightarrow \sigma_y = \sqrt{2}$

We have, $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}\left(x - \bar{x}\right)$ ---- (2)

Comparing (2) with (1),

we have $r\frac{\sigma_y}{\sigma_x} = \frac{2}{3}$

we have $r\frac{\sigma_y}{\sigma_x} = \frac{2}{3}$

$$\left(\frac{2}{3}\right)\frac{\sqrt{2}}{\sigma_x} = \frac{2}{3}$$

$$\sigma_x = \sqrt{2}$$

Example While calculating correlation coefficient between two variables x and y from 25 pairs of observations, the following results were obtained: $n = 25, \Sigma x = 125, \Sigma y = 100, \Sigma x^2 = 650, \Sigma y^2 = 460, \Sigma xy = 508$. Later it was discovered at the time of checking that the pairs of values

x	y
8	12
6	8

were copied down as

x	y
6	14
8	6

Obtain the correct value of correlation coefficient.

To get the correct results, we subtract the incorrect values and add the corresponding correct values.

$\therefore$ The correct results are

$$n = 25,$$
$$\Sigma x = 125 - 6 - 8 + 8 + 6 = 125,$$
$$\Sigma y = 100 - 14 - 6 + 12 + 8 = 100,$$
$$\Sigma x^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650,$$
$$\Sigma y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436,$$
$$\Sigma xy = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520$$

We have, $$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{\left\{n\Sigma x^2 - (\Sigma x)^2\right\} \times \left\{n\Sigma y^2 - (\Sigma y)^2\right\}}}$$

$\Rightarrow$ $$r = \frac{25 \times 520 - 125 \times 100}{\sqrt{\left\{25 \times 650 - (125)^2\right\} \times \left\{25 \times 436 - (100)^2\right\}}} = \frac{20}{\sqrt{25 \times 36}} = \frac{2}{3}$$

EXERCISE

1. Find two lines of regression and coefficient of correlation for the data
   $n=18,\ \Sigma x=12,\ \Sigma y=18,\ \Sigma x^2=6,\ \Sigma y^2=96,\ \Sigma xy=48.$
2. Find the coefficient of correlation and the lines of regression for the data

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

3. Find the coefficient of correlation from the following data

x	78	89	97	69	59	79	68	57
y	125	137	156	112	107	138	123	108

4. Find the coefficient of correlation for the data

x	21	23	30	54	57	58	72	78	87	90
y	60	71	72	83	110	84	100	92	113	135

5. If the coefficient of correlation between two variables x and y is 0.5 and the acute angle between their lines of regression is $\tan^{-1}\left(\frac{3}{8}\right)$, show that $\sigma_x=\frac{1}{2}\sigma_y$.

(VTU 2004)

6. Two random variables have the regression lines with equations $3x+2y=26$ and $6x+y=31$. Find the mean values and the correlation coefficient between x and y.
7. The regression equations of two variables x and y are $x=0.7y+5.2$ and $y=0.3x+2.8$. Find the means of the variables and the coefficient of correlation between them.
8. In a partially destroyed laboratory data, only the equations giving the two lines of regression of y on x and x on y are available and are respectively, $7x-16y+9=0$ and $5y-4x-3=0$. Calculate the coefficient of correlation and mean values of x and y.
9. The two regression equations of the variables x and y are $x=19.13-0.87y$ and $y=11.64-0.50x$. Find the mean values of x and y and the correlation coefficient between x and y. **(VTU 2004)**
10. Psychological tests of intelligence and of engineering ability were applied to 10 students. Here is a record of ungrouped data showing intelligence ratio (I.R) and engineering ratio (E.R). Calculate the coefficient of correlation.

Student	A	B	C	D	E	F	G	H	I	J
I.R	105	104	102	101	100	99	98	96	93	92
E.R	101	103	100	98	95	96	104	92	97	94

Module-4
Part – 5
Rank Correlation

Rank Correlation

- **A group of individuals may be arranged in order to merit with respect to some characteristic. The same group would give different orders for different characteristics.**
- **Considering the orders corresponding to two characteristics A and B, the correlation between these n pairs of ranks is called the rank correlation in the characteristics A and B for that group of individuals.**

Let x_i and y_i be the ranks of the i^{th} individuals in A and B respectively. Assuming that no two individuals are bracketed equal in either case, each of the variables taking the values 1, 2, ... , n,

we have, $$\bar{x} = \bar{y} = \frac{1+2+3+\ldots+n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

If X and Y be the deviations of x and y from their means, then

$$\Sigma X_i^2 = \Sigma\left(x_i - \bar{x}\right)^2 = \Sigma x_i^2 + n\left(\bar{x}\right)^2 - 2\bar{x}\Sigma x_i$$

$$= \Sigma n^2 + \frac{n(n+1)^2}{4} - 2\left(\frac{n+1}{2}\right)\Sigma n$$

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

$$= \frac{1}{12}\left(n^3 - n\right)$$

Similarly, $\Sigma Y_i^2 = \frac{1}{12}\left(n^3 - n\right)$

Now, let $d_i = x_i - y_i$ so that $d_i = \left(x_i - \bar{x}\right) - \left(y_i - \bar{y}\right) = X_i - Y_i$

$$\therefore \qquad \Sigma d_i^2 = \Sigma X_i^2 + \Sigma Y_i^2 - 2\Sigma X_i Y_i$$

$$\Rightarrow \qquad \Sigma X_i Y_i = \frac{1}{2}\left[\Sigma X_i^2 + \Sigma Y_i^2 - \Sigma d_i^2\right]$$

$$\Rightarrow \qquad \Sigma X_i Y_i = \frac{1}{12}\left(n^3 - n\right) - \frac{1}{2}\Sigma d_i^2$$

Hence the correlation coefficient between these variables is

$$r = \frac{\Sigma X_i Y_i}{\sqrt{\Sigma X_i^2 \Sigma Y_i^2}} = \frac{\frac{1}{12}\left(n^3 - n\right) - \frac{1}{2}\Sigma d_i^2}{\frac{1}{12}\left(n^3 - n\right)}$$

$$\Rightarrow \quad r = 1 - \frac{6\Sigma d_i^2}{\left(n^3 - n\right)}$$

This formula is called the rank correlation coefficient or Spearman's Rank Correlation Coefficient and is denoted by ρ.

$$\text{i.e.,} \quad \rho = 1 - \frac{6\Sigma d_i^2}{\left(n^3 - n\right)}$$

- Spearman's rank is probably one of the most useful statistical tests that we can do in Geography to prove a relationship between two different sets of data.
- The Spearman rank correlation coefficient, *is the non-parametric* version of the Pearson correlation coefficient. Data must be ordinal, interval or ratio. Spearman's returns a value from –1 to 1, where:

+1 = a perfect positive correlation between ranks

–1 = a perfect negative correlation between ranks

0 = no correlation between ranks.

Note: To assign the rank for the given set of values, order the scores from greatest to smallest; assign the rank 1 to the highest score, 2 to the next highest and so on. If ranks are tied, i.e., Tied ranks are where two items in a column have the same rank then each tied data point assigned a mean rank.

Example 4.4.1: Calculate the rank correlation coefficient and comment on the following data on sunflower:

Height of a sunflower (in cm)	183	134	234	256	190	89	112
Width of the stem (in mm)	21	14	24	32	29	18	20

We have, the rank correlation coefficient is

$$\rho = 1 - \frac{6\Sigma d_i^2}{\left(n^3 - n\right)}$$

We prepare a relevant table as follows

Height of a sunflower (in cm)	Rank x_i	Width of the stem (in mm)	Rank y_i	$d_i = x_i - y_i$	d_i^2
183	4	21	4	0	0
134	5	14	7	-2	4
234	2	24	3	-1	1
256	1	32	1	0	0
190	3	29	2	1	1
89	7	18	6	1	1
112	6	20	5	1	1
				Σ	8

Here, $n = 7$, $\Sigma d_i^2 = 8$

$$\rho = 1 - \frac{6\Sigma d_i^2}{(n^3 - n)} = 1 - \frac{6(8)}{(7^3 - 7)} = 1 - 0.143 = 0.857$$

As ρ is close to one, we can conclude that the wider the stem to higher the sunflower grows.

Example 4.4.2: The scores for 9 students in Physics and Maths are as follows:

Physics:	35	23	47	17	10	43	9	6	28
Mathematics:	30	33	45	23	8	49	12	4	31

Compute the ranks of students in the two subjects and compute the Spearman's rank correlation.

We have, the rank correlation coefficient is

$$\rho = 1 - \frac{6\Sigma d_i^2}{\left(n^3 - n\right)}$$

We prepare a relevant table as follows

Physics	Rank x_i	Mathematics	Rank y_i	$d_i = x_i - y_i$	d_i^2
35	3	30	5	-2	4
23	5	33	3	2	4
47	1	45	2	-1	1
17	6	23	6	0	0
10	7	8	8	-1	1
43	2	49	1	1	1
9	8	12	7	1	1
6	9	4	9	0	0
28	4	31	4	0	0
				Σ	12

Here, $n = 9$, $\Sigma d_i^2 = 12$

$$\rho = 1 - \frac{6\Sigma d_i^2}{\left(n^3 - n\right)} = 1 - \frac{6(12)}{\left(9^3 - 9\right)} = 1 - 0.1 = 0.9$$

Example 4.4.3: Ten participants in a contest are ranked by two judges as follows:

X	1	6	5	10	3	2	4	9	7	8
Y	6	4	9	8	1	2	3	10	5	7

Calculate the rank correlation coefficient.

We have, the rank correlation coefficient is

$$\rho = 1 - \frac{6\Sigma d_i^2}{\left(n^3 - n\right)}$$

We prepare a relevant table as follows

Rank x_i	Rank y_i	$d_i = x_i - y_i$	d_i^2
1	6	-5	25
6	4	2	4
5	9	-4	16
10	8	2	4
3	1	2	4
2	2	0	0
4	3	1	1
9	10	-1	1
7	5	2	4
8	7	1	1
		Σ	60

Here, $n = 10$, $\Sigma d_i^2 = 60$

$$\rho = 1 - \frac{6\Sigma d_i^2}{\left(n^3 - n\right)} = 1 - \frac{6(60)}{\left(10^3 - 10\right)} = 1 - 0.36 = 0.64$$

Example 4.4.4: Three judges A, B and C give the following ranks. Find which pair of judges has common approach

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

We have, the rank correlation coefficient is

$$\rho = 1 - \frac{6\Sigma d_i^2}{\left(n^3 - n\right)}$$

We prepare a relevant table as follows

Ranks by A x_i	Ranks by B y_i	Ranks by C z_i	d_1 $x_i - y_i$	d_2 $y_i - z_i$	d_3 $z_i - x_i$	d_1^2	d_2^2	d_3^2
1	3	6	-2	-3	5	4	9	25
6	5	4	1	1	-2	1	1	4
5	8	9	-3	-1	4	9	1	16
10	4	8	6	-4	-2	36	16	4
3	7	1	-4	6	-2	16	36	4
2	10	2	-8	8	0	64	64	0
4	2	3	2	-1	-1	4	1	1
9	1	10	8	-9	1	64	81	1
7	6	5	1	1	-2	1	1	4
8	9	7	-1	2	1	1	4	1
					Σ	200	214	60

Here, $n = 10$

$$\rho(x,y) = 1 - \frac{6\Sigma d_1^2}{(n^3 - n)} = 1 - \frac{6(200)}{(10^3 - 10)} = -0.2$$

$$\rho(y,z) = 1 - \frac{6\Sigma d_2^{\,2}}{(n^3 - n)} = 1 - \frac{6(214)}{(10^3 - 10)} = -0.3$$

$$\rho(z,x) = 1 - \frac{6\Sigma d_3^{\,2}}{(n^3 - n)} = 1 - \frac{6(60)}{(10^3 - 10)} = 0.6$$

Since $\rho(z,x)$ is maximum, the pair of judges A and C have the nearest common approach.

EXERCISE

1. Find the rank correlation for the following data:

x	56	42	72	36	63	47	55	49	38	42	68	60
y	147	125	160	118	149	128	150	145	115	140	152	155

2. Calculate the rank correlation coefficient from the following data showing ranks of 10 students in two subjects:

Maths:	3	8	9	2	7	10	4	6	1	5
Physics:	5	9	10	1	8	7	3	4	2	6

3. Find 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

(VTU 2018)

Mat 41

Module - 4

Statistical Methods

Mean (Arithmetic mean) :-

If $x_1, x_2, \ldots, x_n$ be a set of n values of a variate x, the mean denoted by $\bar{x}$ is defined as follows.

$$\bar{x} = \frac{\sum x}{n} \quad \text{(or)} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

for a grouped data in the form of a frequency distribution,

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} \quad \text{(or)} \quad \bar{x} = \frac{\sum fx}{\sum f}$$

where f_i's are the frequency of the classes having corresponding midpoint x_i.

Variance (V) and Standard deviation (SD) :-

If a variate x take values $x_1, x_2, \ldots, x_n$ the variance (V) is defined as follows.

$$V = \frac{\sum (x-\bar{x})^2}{n} \quad \text{(or)} \quad V = \frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n}$$

Also for a grouped data

$$V = \frac{\sum f(x-\bar{x})^2}{\sum f} \quad \text{(or)} \quad V = \frac{\sum_{i=1}^{n} f_i (x_i-\bar{x})^2}{\sum_{i=1}^{n} f_i}$$

Standard deviation (SD), $\sigma = \sqrt{V}$ (or) $\sigma^2 = V$

Alternative expression for σ^2 :-

consider, $\sigma^2 = \frac{1}{n}\sum (x-\bar{x})^2$

$$\sigma^2 = \frac{1}{n}\sum [x^2 - 2x\bar{x} + (\bar{x})^2]$$

$$= \frac{\sum x^2}{n} - 2(\bar{x})\bar{x} + \frac{n(\bar{x})^2}{n}$$

Scanned with CamScanner

Here, $\frac{\sum x}{n} = \bar{x}$ & $(\bar{x})^2$ being a constant added 'n' times gives $n(\bar{x})^2$.

i.e $\sigma^2 = \frac{\sum x^2}{n} - 2(\bar{x})^2 + (\bar{x})^2$

$$\sigma^2 = \frac{\sum x^2}{n} - (\bar{x})^2$$

For a grouped data the expression will be of the form

$$\sigma^2 = \frac{\sum f x^2}{\sum f} - (\bar{x})^2$$

Example :-

① we shall find the mean and standard deviation of a set of observations. 6, 8, 7, 5, 4, 9, 3

Soln :- $\bar{x} = \frac{\sum x}{n}$,

$$\bar{x} = \frac{6+8+7+5+4+9+3}{7} = \frac{42}{7} = 6$$

Thus mean $(\bar{x}) = 6$

$$\text{Var} \text{ (or) } \sigma^2 = \frac{\sum (x-\bar{x})^2}{n} = \frac{1}{7}\{(6-6)^2+(8-6)^2+(7-6)^2 + (5-6)^2+(4-6)^2+(9-6)^2 + (3-6)^2\}$$

$$\text{Var} = \frac{28}{7} = 4$$

$$\therefore \sigma = \sqrt{4} = 2$$

Alternate : $\sigma^2 = \frac{\sum x^2}{n} - (\bar{x})^2$

$$= \frac{6^2+8^2+7^2+5^2+4^2+9^2+3^2}{7} - (6)^2$$

$$= \frac{280}{7} - 36$$

$$= 4$$

Thus $SD = \sigma = 2$.

② Let us find the mean & SD for the following grouped data

class	1-10	11-20	21-30	31-40	41-50	51-60
Frequency	3	16	26	31	16	8

Scanned with CamScanner

class	f	x	fx	$(x-\bar{x})$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
1-10	3	5.5	16.5		702.25	2106.75
11-20	16	15.5	248.0		272.25	4356.00
21-30	26	25.5	663.0		42.25	1098.50
31-40	31	35.5	1100.5		12.25	379.75
41-50	16	45.5	728.0		182.25	2916.00
51-60	8	55.5	444.0		552.25	4418.00
Totals	100		3200			15275

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{3200}{100} = 32$$

$$\sigma^2 = \frac{\sum f(x-\bar{x})^2}{\sum f} = \frac{15275}{100} = 152.75$$

$$\therefore \sigma = \sqrt{152.75} = 12.36$$

Curve fitting :-

* **Fitting of a straight line** : $y = ax+b$

Consider a set of 'n' given values (x, y) for fitting the straight line $y = ax+b$ where 'a' & 'b' are parameters to be determined. The residual $R = y - (ax+b)$ is the difference between the observed and estimated values of 'y'. By the method of least squares we find parameters 'a' & 'b' such that the sum of squares of the residuals is minimum.

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

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	

Scanned with CamScanner

The normal eqns for fitting the straight line $y = ax + b$ are

$$\Sigma y = a\Sigma x + nb \quad (n = 8)$$
$$\Sigma xy = a\Sigma x^2 + b\Sigma x$$

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
$\Sigma x = 56$	$\Sigma y = 40$	$\Sigma xy = 364$	$\Sigma x^2 = 524$

The normal eqns become

$$56a + 8b = 40.$$
$$524a + 56b = 364$$

$\therefore a = 0.63 \approx 0.64, b = 0.54 \approx 0.55$

$\therefore$ Thus by substituting these values in $y = ax + b$ we obtain the eqn.

$$y = 0.64x + 0.55$$

(2) Find the eqn of the best fitting straight line for the following data & hence estimate the value of the dependent variable corresponding to the value 30 of the independent variable.

x	5	10	15	20	25
y	16	19	23	26	30

$\therefore$ $y = ax + b$

$$\Sigma y = a\Sigma x + nb$$
$$\Sigma xy = a\Sigma x^2 + b\Sigma x \quad (n = 5)$$

Scanned with CamScanner

x	y	xy	x^2
5	16	80	25
10	19	190	100
15	23	345	225
20	26	520	400
25	30	750	625

$\Sigma x = 75$ $\Sigma y = 114$ $\Sigma xy = 1885$ $\Sigma x^2 = 1375$

$\therefore 75a + 5b = 114$

$1375a + 75b = 1885$

$\therefore a = 0.7$, $b = 12.3$.

$y = ax + b$

$y = 0.7x + 12.3$

when $x = 30$, we obtain $y = 0.7(30) + 12.3 = 33.3$

(3) 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 & 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 $y = a + bp$ & hence estimate 'y' when 'p' is 150.

Soln:- The normal equations associated with $y = a + bp$ are as follows.

$$\Sigma y = na + b\Sigma p$$

$$\Sigma py = a\Sigma p + b\Sigma p^2 \qquad (n=6)$$

Scanned with CamScanner

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

$\Sigma p = 900 \quad \Sigma y = 3.95 \quad \Sigma py = 621 \quad \Sigma p^2 = 142000$

$\therefore$ Eqns $\Rightarrow$

$$6a + 900b = 3.95$$
$$900a + 142000b = 621$$

$\therefore a = 0.0476 \quad , \quad b = 0.0041$

Thus the required Law is $y = 0.0476 + 0.0041P$

Also when $P = 150$, $y = 0.6626 \approx 0.66$

④ 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.

Soln :- let $x = x - 1981$ & the line of fit with be $y = a + bx$

$$\Sigma y = na + b\Sigma x \qquad (n = 5)$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Scanned with CamScanner

x	y	xy	x^2
-20	8	-160	400
-10	10	-100	100
0	12	0	0
10	10	100	100
20	16	320	400

$\Sigma x = 0 \quad \Sigma y = 56 \quad \Sigma xy = 160 \quad \Sigma x^2 = 1000$

The normal equations become,

$5a = 56$ & $1000b = 160$

$a = 11.2$ & $b = 0.16$

Hence $y = a + bx$, with $x = x - 1981$ becomes

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

Thus $y = -305.76 + 0.16x$ is the required line of fit.

Also when $x = 2006$, $y = -305.76 + 0.16\,(2006)$

$y = 15.2$

Expected production in the year 2006 is 15.2 ton.

(5) Find the eqn of the best filling straight line for the following data.

i)

x	1	2	3	4	5
y	14	13	9	5	2

ii)

x	0	1	2	3	4	5
y	9	8	24	28	26	20

iii)

x	62	64	65	69	70	71	72
y	65.7	66.8	67.2	69.3	69.8	70.5	70.9

iv)

x	1	2	3	4	5	6	7
y	80	90	92	83	94	99	92

Scanned with CamScanner

⑥

Year (x)	1911	1921	1931	1941	1951
production(y) (in thousand tons)	8	10	12	10	6

Soln:- let $X = x - 1931$ & the line of fit will be $y = a + bX$

The normal eqns associated with $y = a + bx$ are as follows.

$$\Sigma y = na + b\Sigma x \qquad (n = 5)$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

X	y	Xy	X^2
-20	8	-160	400
-10	10	-100	100
0	12	0	0
10	10	100	100
20	06	120	400

$\Sigma X = 0 \quad \Sigma y = 46 \quad \Sigma xy = -40 \quad \Sigma x^2 = 1000$

The normal eqns become,

$$46 = 5a \rightarrow ①$$
$$-40 = 1000b \rightarrow ②$$

Eqn ① $a = 9.2$,

Eqn ② $b = -0.04$

Hence $y = a + bX$, with $X = x - 1931$

$$y = 9.2 + (-0.04)(x - 1931)$$
$$= 9.2 - 0.04x + 77.24$$

Thus, $y = 86.44 - 0.04x$ is the required line of fit.

Scanned with CamScanner

* fitting of a second degree parabola $y = ax^2+bx+c$:-

consider a set of 'n' given values (x,y) for fitting the curve $y = ax^2+bx+c$. The residual $R = y-(ax^2+bx+c)$ is the difference b/w the observed & estimated values of y. we have to find parameters a, b, c such that the sum of the squares of the residuals is the least.

$$\Sigma y = a\Sigma x^2 + b\Sigma x + nc$$
$$\Sigma xy = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x$$
$$\Sigma x^2y = a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2$$

*

① Fit a best fitting parabola $y = ax^2+bx+c$ for the following data:

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

and hence estimate at $x=6$.

Soln:- The normal eqns associated with $y = ax^2+bx+c \rightarrow (*)$ are as follows.

$$\Sigma y = a\Sigma x^2 + b\Sigma x + nc \longrightarrow ①$$
$$\Sigma xy = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x \longrightarrow ②$$
$$\Sigma x^2y = a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2 \longrightarrow ③$$

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

$\Sigma x=15$ $\Sigma y=70$ $\Sigma xy=232$ $\Sigma x^2=55$ $\Sigma x^2y=906$ $\Sigma x^3=225$ $\Sigma x^4=979$

eqn ① $\Rightarrow$ $70 = 55a+15b+5c$

eqn ② $\Rightarrow$ $232 = 225a+55b+15c$

eqn ③ $\Rightarrow$ $906 = 979a+225b+55c$

Scanned with CamScanner

$\therefore a = 0.2857 \approx 0.29$, $b = 0.4857 \approx 0.49$, $c = 9.4$.

Thus the required second degree parabola is

$y = 0.29x^2 + 0.49x + 9.4$ Also at $x = 6$

$y = 22.78$

★ ② Fit a parabola $y = a + bx + cx^2$ for the data.

x	0	1	2	3	4
y	1	1.8	1.3	2.5	2.3

Soln:- The normal eqn associated $y = a + bx + cx^2$ → ⊛ are

$\Sigma y = an + b\Sigma x + c\Sigma x^2 \rightarrow ①$

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

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

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	2.3	9.2	16	36.8	64	256
$\Sigma x = 10$	$\Sigma y = 8.9$	$\Sigma xy = 21.1$	$\Sigma x^2 = 30$	$\Sigma x^2 y = 66.3$	$\Sigma x^3 = 100$	$\Sigma x^4 = 354$

Eqn ① ⇒ $8.9 = 5a + 10b + 30c$

Eqn ② ⇒ $21.1 = 10a + 30b + 100c$

eqn ③ ⇒ $66.3 = 30a + 100b + 354c$

$a = 1.0771$, $b = 0.4157$, $c = -0.0214$

Thus the parabola of fit is

$y = 1.0771 + 0.4157x - 0.0214x^2$

★ ③ Fit a second degree parabola to the following data,

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	1.1	1.3	1.6	2.0	2.7	3.4	4.1

★ ④ Fit a second degree parabola to the following data:

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

Scanned with CamScanner

* Fitting of a curve of the form $y = ae^{bx}$:-

Consider, $y = ae^{bx}$. Taking logarithms (to the base e) on both sides we get.

$$\log_e y = \log_e (ae^{bx})$$
$$= \log_e a + \log_e e^{bx}$$
$$= \log_e a + bx \log_e e$$
$$\log_e y = \log_e a + bx \quad (\text{or})$$
$$Y = A + BX \longrightarrow (1)$$

$\log mn = \log m + \log n$
$\log m^n = n \log m$
$\log_e e = 1$

where $Y = \log_e y$, $A = \log_e a$, $B = b$, $X = x$

It is evident that eqn (1) is the eqn of a straight line & the associated normal eqns are as follows.

$$\Sigma Y = nA + B\Sigma X \longrightarrow (2)$$
$$\Sigma XY = A\Sigma X + B\Sigma X^2 \longrightarrow (3)$$

Solving eqn (2) & (3) we obtain 'A' & 'B'. But $\log_e a = A \Rightarrow a = e^A$. Also $b = B$.

Substituting these values in $y = ae^{bx}$ we get the curve of best fit, in the required form.

* (1) Fit a curve of the form $y = ae^{bx}$ to the following data.

x	77	100	185	239	285
y	2.4	3.4	7.0	11.1	19.6

Soln :- consider, $y = ae^{bx} \longrightarrow (*)$

$$y = A + bx$$

The normal eqns are as follows.

$$\Sigma y = nA + b\Sigma x \longrightarrow (1)$$
$$\Sigma xy = A\Sigma x + b\Sigma x^2 \longrightarrow (2)$$

where $y = \log_e y$, $A = \log_e a$

Scanned with CamScanner

x	y	$Y=\log_e y$	xY	x^2
77	2.4	0.8754	67.4058	5929
100	3.4	1.2237	122.37	10000
185	7.0	1.9459	359.9915	34225
239	11.1	2.4069	575.2491	57121
285	19.6	2.9755	848.0175	81225
$\Sigma x=886$	$\Sigma y=43.5$	$\Sigma Y=9.4274$	$\Sigma xy=1973.0339$	$\Sigma x^2=188500$

Eqn ① $\Rightarrow$ $9.4274 = 5A + 886b$

Eqn ② $\Rightarrow$ $1973.0339 = 886A + 188500b$

$A = 0.1838$ $\quad b = 9.6028\times10^{-3}$

$A = \log_e a = 0.1838$

$a = e^A$

$= e^{0.1838}$

$a = 1.2017$, $\quad b = 9.6028\times10^{-3}$

The curve of fit is $y = a e^{bx}$

Thus $y = (1.2017)e^{9.6028\times10^{-3}x}$ is the curve of fit.

(2)★ fit an exponential curve of the form $y = ae^{bx}$ by the method of least squares for the following data.

No of petals	5	6	7	8	9	10
No of flowers	133	55	23	7	2	2

Scanned with CamScanner

x	y	$y=\log_e y$	xy	x^2
5	133	4.8903	24.4515	25
6	55	4.0073	24.0438	36
7	23	3.1355	21.9485	49
8	7	1.9459	15.5672	64
9	2	0.6931	6.2379	81
10	2	0.6931	6.9310	100

$\Sigma x=45$ $\Sigma y=15.3652$ $\Sigma xy=99.1799$ $\Sigma x^2=355$

The normal eqn becomes

$$6A+45b=15.3652$$

$$45A+355b=99.1799$$

$A=9.4433$ & $b=-0.9177$

Thus the required curve of fit is

$$y=(12623.3)\,e^{-0.9177x}.$$

(3) Fit a curve of the form $y=ax^b$ for the data

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

(4) Find the eqn of the best fitting curve in the form $y=ae^{bx}$ for the data

x	0	2	4
y	5.02	10	31.62

Scanned with CamScanner

Correlation and correlation co-efficient :-

co-variation of two independent magnitudes is known as correlation. If two variables 'x' & 'y' are related in such a way that ↑ (or) ↓ in one of them corresponds to ↑ (or) ↓ in the other, we say that the variables are +vely correlated. Also if increase (or) decrease in one of them corresponds to decrease or ↑ in the other, the variables are said to be -vely correlated.

The numerical measure of correlation b/w two variables 'x' & 'y' is known as Pearson's coefficient of correlation usually denoted by 'r' & is defined as follows.

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

This can be put in an alternative form as follows. If $X = x - \bar{x}$, $Y = y - \bar{y}$ we can write.

$$\sigma_x^2 = \frac{\sum (x-\bar{x})^2}{n} = \frac{\sum X^2}{n},$$

$$\sigma_y^2 = \frac{\sum (y-\bar{y})^2}{n} = \frac{\sum Y^2}{n}$$

$$\therefore \sigma_x \sigma_y = \frac{\sqrt{\sum X^2}\sqrt{\sum Y^2}}{\sqrt{n^2}} = \frac{\sqrt{\sum X^2}\sqrt{\sum Y^2}}{n}$$

$$\therefore n \sigma_x \cdot \sigma_y = \sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}$$

Thus eqn (1) becomes

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

Note :- * The coefficient of correlation numerically does not exceed unity. i.e $-1 \le r \le +1$

Scanned with CamScanner

* If $r = \pm 1$ we say that 'x' & 'y' are perfectly correlated & if $r = 0$ we say that 'x' & 'y' are non correlated.

Alternative formula for the correlation co-efficient r :-

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

Proof:- let $z = x - y$

$$\therefore \frac{\Sigma z}{n} = \frac{\Sigma x}{n} - \frac{\Sigma y}{n} \quad \text{(or)} \quad \bar{z} = \bar{x} - \bar{y}$$

Hence, $(z - \bar{z}) = (x - \bar{x}) - (y - \bar{y})$

i,e $(z - \bar{z}) = (x - \bar{x}) - (y - \bar{y})$

Squaring both sides, taking summation & dividing by 'n' we have,

$$\frac{\Sigma (z - \bar{z})^2}{n} = \frac{\Sigma [(x - \bar{x}) - (y - \bar{y})]^2}{n}$$

$$= \frac{\Sigma (x - \bar{x})^2}{n} + \frac{\Sigma (y - \bar{y})^2}{n} - 2\frac{\Sigma (x - \bar{x})(y - \bar{y})}{n}$$

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

i,e $\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$

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

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

Note :- In general if $z = ax + by$ we can obtain as before

$$\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2rab\sigma_x\sigma_y$$

i,e $\sigma_{ax+by}^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2rab\sigma_x\sigma_y$

Scanned with CamScanner

Regression :-

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

The best fitting straight line of the form $y = ax+b$ ('x' being the independent variable) is called the regression line of 'y' on 'x' & $x = ay+b$ ('y' being the independent variable) is called the regression line of 'x' on 'y'. $y-\bar{y} = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$, $x-\bar{x} = r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$

Note :- 1) The lines of regression are

$$y = \frac{\sum xy}{\sum x^2}(x) \quad \text{and} \quad x = \frac{\sum xy}{\sum y^2}(y)$$

where $x = x-\bar{x}$ & $y = y-\bar{y}$.

This form will be useful to find out the coefficient of correlation by first obtaining the lines of regression as we have deduced that

$$\boxed{r = \pm\sqrt{(\text{coeff. of } x)(\text{coeff. of } y)}}$$

2) To compute the coefficient of correlation we prefer to use the formula.

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

where SDs can be found by applying the formula.

$$\sigma^2 = \frac{\sum x^2}{n} - (\bar{x})^2$$

If $\bar{x}$ & $\bar{y}$ are integers computation of 'r' by the formula.

$$r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$$ is convenient where $x = x-\bar{x}$, $y = y-\bar{y}$.

Scanned with CamScanner

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

Soln:- we have $r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y} \rightarrow ①$

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
$\Sigma x = 28$	$\Sigma y = 77$	$\Sigma z = -49$	$\Sigma x^2 = 140$	$\Sigma y^2 = 875$	$\Sigma z^2 = 347$

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

$\bar{x} = \frac{28}{7} = 4$, $\bar{y} = \frac{77}{7} = 11$, $\bar{z} = \frac{-49}{7} = -7$

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

$= \frac{140}{7} - (4)^2 = 4$ $= \frac{875}{7} - (11)^2 = 4$ $= \frac{347}{7} - (-7)^2 = 0.5$

we have, $\sigma_x^2 = 4$, $\sigma_y^2 = 4$, $\sigma_z^2 = \sigma_{x-y}^2 = 0.57$

Eqn ① $\Rightarrow$ $r = \frac{4+4-0.57}{2\sqrt{4}\sqrt{4}} = 0.92875 \approx 0.93$.

Thus $r = 0.93$

The lines of regression are given by

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

$y - 11 = \frac{(0.93)(2)}{2}(x-4)$, $x - 4 = \frac{(0.93)2}{2}(y-11)$

$y - 11 = 0.93(x-4)$, $x - 4 = 0.93(y-11)$

Thus $y = 0.93x + 7.28$ & $x = 0.93y - 6.23$ are the lines of regression.

Scanned with CamScanner

② obtain the lines of regression & hence find the coefficient of correlation for the data.

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

Soln :- Here $\bar{x} = 4$, $\bar{y} = 11$

$\therefore X = x - \bar{x}$, $Y = y - \bar{y}$

$= x - 4$, $= y - 11$

x	y	X	Y	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
				$\Sigma XY = 26$	$\Sigma X^2 = 28$	$\Sigma Y^2 = 28$

we shall consider regression lines in the form.

$$Y = \frac{\Sigma XY}{\Sigma X^2} \cdot X \quad \text{and} \quad X = \frac{\Sigma XY}{\Sigma Y^2} \cdot Y$$

i,e $y - 11 = \frac{26}{28}(x-4)$, $x - 4 = \frac{26}{28}(y-11)$

$y - 11 = 0.93(x-4)$, $x - 4 = 0.93(y-11)$

$y = 0.93x + 7.28$, $x = 0.93y - 6.23$

These are the regression lines and we compute r as the geometric mean of the regression coefficients.

i,e $r = \sqrt{(\text{coeff of } x)(\text{coeff of } y)} = \sqrt{(0.93)(0.93)}$

$r = 0.93$

The sign of r must be +ve since both the regression coefficients are +ve.

Thus $r = 0.93$.

Scanned with CamScanner

③ Find the correlation coefficient & the eqn of the lines of regression for the following values of x & y.

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

Soln:- $n = 5$

x	y	$z = x - y$	x^2	y^2	z^2
1	2	-1	1	4	1
2	5	-3	4	25	9
3	3	0	9	9	0
4	8	-4	16	64	16
5	7	-2	25	49	4
$\Sigma x = 15$	$\Sigma y = 25$	$\Sigma z = -10$	$\Sigma x^2 = 55$	$\Sigma y^2 = 151$	$\Sigma z^2 = 30$

$$\bar{x} = \frac{\Sigma x}{n}, \quad \bar{y} = \frac{\Sigma y}{n}, \quad \bar{z} = \frac{\Sigma z}{n}$$

$$\bar{x} = \frac{15}{5} = 3, \quad \bar{y} = \frac{25}{5} = 5, \quad \bar{z} = \frac{-10}{5} = -2$$

$$\sigma_x^2 = \frac{\Sigma x^2}{n} - (\bar{x})^2 = \frac{55}{5} - (3)^2 = 2$$

$$\sigma_y^2 = \frac{\Sigma y^2}{n} - (\bar{y})^2 = \frac{151}{5} - (5)^2 = 5.2$$

$$\sigma_z^2 = \frac{\Sigma z^2}{n} - (\bar{z})^2 = \frac{30}{5} - (-2)^2 = 2$$

we have, $\sigma_x^2 = 2$, $\sigma_y^2 = 5.2$, $\sigma_z^2 = 2$

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

$$r = \frac{2 + 5.2 - 2}{2\sqrt{2}\sqrt{5.2}} = 0.8062 \approx 0.81$$

Thus $\boxed{r = 0.81}$

The eqn of the regression lines are as follows.

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

$$y - 5 = (0.81)\cdot\frac{\sqrt{5.2}}{\sqrt{2}}(x - 3), \qquad x - 3 = 0.81\cdot\frac{\sqrt{2}}{\sqrt{5.2}}(y - 5)$$

Scanned with CamScanner

$y - 5 = 1.306(x-3)$, $\quad x - 3 = 0.502(y-5)$

Thus $y = 1.306x + 1.082$ & $x = 0.502y + 0.49$

These are the lines of regression.

④ Find the correlation coefficient b/w 'x' & 'y' for the following data. Also obtain the regression lines.

x	1	2	3	4	5	6	7	8	9	10
y	10	12	16	28	25	36	41	49	40	50

Soln:- Here $n = 10$

x	y	$z = x - y$	x^2	y^2	z^2
1	10	-9	1	100	81
2	12	-10	4	144	100
3	16	-13	9	256	169
4	28	-24	16	784	576
5	25	-20	25	625	400
6	36	-30	36	1296	900
7	41	-34	49	1681	1156
8	49	-41	64	2401	1681
9	40	-31	81	1600	961
10	50	-40	100	2500	1600
$\Sigma x = 55$	$\Sigma y = 307$	$\Sigma z = -252$	$\Sigma x^2 = 385$	$\Sigma y^2 = 11387$	$\Sigma z^2 = 7624$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{55}{10} = 5.5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{307}{10} = 30.7$$

$$\bar{z} = \frac{\Sigma z}{n} = \frac{-252}{10} = -25.2$$

Scanned with CamScanner

$\sigma_x^2 = \frac{\Sigma x^2}{n} - (\bar{x})^2 = \frac{385}{10} - (5.5)^2 = 8.25$, $\sigma_x = 2.87$

$\sigma_y^2 = \frac{\Sigma y^2}{n} - (\bar{y})^2 = 1138.7 - (30.7)^2 = 196.21$, $\sigma_y = 14.01$

$\sigma_z^2 = \sigma_{x-y}^2 = \frac{\Sigma z^2}{n} - (\bar{z})^2 = 762.4 - (-25.2)^2 = 127.36$

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

$= \frac{8.25 + 196.21 - 127.36}{2 \times 2.87 \times 14.01} = 0.96$

Thus $\boxed{r = 0.96}$

Equation of the lines of regression are

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

on substituting & simplifying we get,

$y - 30.7 = 0.96 \times \frac{14.01}{2.87}(x - 5.5)$

$\boxed{y = 4.686x + 4.927}$

$x - 5.5 = 0.96 \times \frac{2.87}{14.01}(y - 30.7)$

$\boxed{x = 0.197y - 0.548}$

There are the lines of regression.

⑤ Find the coefficient of correlation for the following data.

x	10	14	18	22	26	30
y	18	12	24	6	30	36

Soln:- we have $\bar{x} = \frac{\Sigma x}{n} = \frac{120}{6} = 20$

$\bar{y} = \frac{\Sigma y}{n} = \frac{126}{6} = 21$

Let $x = x - \bar{x}$, & $y = y - \bar{y}$

$x = x - 20$, $y = y - 21$ &

Scanned with CamScanner

x	y	X	Y	x^2	y^2	XY
10	18	-10	-3	100	9	30
14	12	-6	-9	36	81	54
18	24	-2	3	4	9	-6
22	6	2	-15	4	225	-30
26	30	6	9	36	81	54
30	36	10	15	100	225	150
				$\Sigma x^2 = 280$	$\Sigma y^2 = 630$	$\Sigma XY = 252$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = \frac{252}{\sqrt{280} \times \sqrt{630}} = 0.6$$

★

⑥ Find the correlation coefficient and the eqⁿ of the line of regression for the following.

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

Solⁿ:- Let $n = 5$

x	y	$z = x - y$	x^2	y^2	z^2
1	2	-1	1	4	1
2	5	-3	4	25	9
3	3	0	9	9	0
4	8	-4	16	64	16
5	7	-2	25	49	4
$\Sigma x = 15$	$\Sigma y = 25$	$\Sigma z = -10$	$\Sigma x^2 = 55$	$\Sigma y^2 = 151$	$\Sigma z^2 = 30$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{15}{5} = 3, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{25}{5} = 5, \quad \bar{z} = \frac{\Sigma z}{n} = \frac{-10}{5} = -2$$

$$\sigma_x^2 = \frac{\Sigma x^2}{n} - (\bar{x})^2 = \frac{55}{5} - (3)^2 = 2$$

$$\sigma_y^2 = \frac{\Sigma y^2}{n} - (\bar{y})^2 = \frac{151}{5} - (5)^2 = 5.2$$

$$\sigma_z^2 = \sigma_{x-y}^2 = \frac{\Sigma z^2}{n} - (\bar{z})^2 = \frac{30}{5} - (-2)^2 = 2$$

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

Scanned with CamScanner

$$r = \frac{2+5.2-2}{2\sqrt{2}\sqrt{5.2}} = 0.8062 \approx 0.81$$

Thus $r = 0.81$

The eqns of the regression lines are as follows.

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

$$y-5 = (0.81)\frac{\sqrt{5.2}}{\sqrt{2}}(x-3) \quad , \quad x-3 = (0.81)\frac{\sqrt{2}}{\sqrt{5.2}}(y-5)$$

$$y-5 = 1.306(x-3) \quad , \quad x-3 = 0.502(y-5)$$

Thus $y = 1.306x + 1.082$ & $x = 0.502y + 0.49$

There are the lines of regression.

★ (7) Find the correlation coefficient b/w x & y for the following data. Also obtain the regression lines.

x	1	2	3	4	5	6	7	8	9	10
y	10	12	16	28	25	36	41	49	40	50

Soln:- Here $n = 10$

x	y	$z = x-y$	x^2	y^2	z^2
1	10	-9	1	100	81
2	12	-10	4	144	100
3	16	-13	9	256	169
4	28	-24	16	784	576
5	25	-20	25	625	400
6	36	-30	36	1296	900
7	41	-34	49	1681	1156
8	49	-41	64	2401	1681
9	40	-31	81	1600	961
10	50	-40	100	2500	1600
$\Sigma x = 55$	$\Sigma y = 307$	$\Sigma z = -252$	$\Sigma x^2 = 385$	$\Sigma y^2 = 11387$	$\Sigma z^2 = 7624$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{55}{10} = 5.5, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{307}{10} = 30.7, \quad \bar{z} = \frac{\Sigma z}{n} = \frac{-252}{10} = -25.2$$

Scanned with CamScanner

$\sigma_x^2 = \frac{\sum x^2}{n} - (\bar{x})^2 = \frac{385}{10} - (5.5)^2 = 8.25$, $\sigma_x = 2.87$

$\sigma_y^2 = \frac{\sum y^2}{n} - (\bar{y})^2 = \frac{11387}{10} - (30.7)^2 = 196.21$, $\sigma_y = 14.01$

$\sigma_z^2 = \sigma_{x-y}^2 = \frac{\sum z^2}{n} - (\bar{z})^2 = \frac{7624}{10} - (-25.2)^2 = 127.36$

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

$$r = \frac{8.25 + 196.21 - 127.36}{2 \times 2.87 \times 14.01} = 0.96$$

Thus $r = 0.96$

Eqn of the line of regression are

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

$y - 30.7 = 0.96 \times \frac{14.01}{2.87}(x - 5.5)$, $(x - 5.5) = 0.96 \times \frac{2.87}{14.01}(y - 30.7)$

$y = 4.686x + 4.927$ and $x = 0.197y - 0.548$

There are lines of regression.

★ (8) Find the regression line 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

Soln :- $\bar{x} = \frac{\sum x}{n} = \frac{56}{8} = 7$, $\bar{y} = \frac{\sum y}{8} = \frac{40}{8} = 5$

we denote $X = x - \bar{x}$ and $Y = y - \bar{y}$

$X = x - 7$, $Y = y - 5$

we have lines of regression in the form

$$Y = \frac{\sum XY}{\sum X^2} X \qquad X = \frac{\sum XY}{\sum Y^2} \cdot Y.$$

Scanned with CamScanner

x	y	$x = x - \bar{x}$	$y = y - \bar{y}$	x^2	y^2	xy
1	1	-6	-4	36	16	24
3	2	-4	-3	16	9	12
4	4	-2	-1	4	1	2
6	4	-1	-1	1	1	1
8	5	1	0	1	0	0
9	7	2	2	4	4	4
11	8	4	3	16	9	12
14	9	7	4	49	16	28

$\Sigma x = 56$ $\Sigma y = 40$ $\Sigma x^2 = 132$ $\Sigma y^2 = 56$ $\Sigma xy = 84$

i.e $y - \bar{y} = \frac{\Sigma xy}{\Sigma x^2}(x - \bar{x})$, $x - \bar{x} = \frac{\Sigma xy}{\Sigma y^2}(y - \bar{y})$

$y - 5 = \frac{84}{132}(x - 7)$ $x - 7 = \frac{84}{56}(y - 5)$

$y - 5 = 0.63(x - 7)$ $x - 7 = 1.5(y - 5)$

$y - 5 = 0.63x - 4.41$ $x - 7 = 1.5y - 7.5$

$y = 0.63x - 4.41 + 5$ $x = 1.5y - 7.5 + 7$

$y = 0.66x + 0.59$ $x = 1.5y - 0.5$

These are the lines of regression.

(9) Calculate the karl pearson's coefficient of correlation for 10 students who have obtained the following % of marks in mathematics & Electronics.

Roll No	1	2	3	4	5	6	7	8	9	10
Marks in mathematics	78	36	98	25	75	82	90	62	65	39
Marks in Electronics	84	51	91	60	68	62	86	58	53	47

Soln:- we have $\bar{x} = \frac{\Sigma x}{n} = \frac{650}{10} = 65$

$\bar{y} = \frac{\Sigma y}{n} = \frac{660}{10} = 66$

Let $x = x - \bar{x}$ & $y = y - \bar{y}$

$x = x - 65$ $y = y - 66$

Scanned with CamScanner

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	x^2	y^2	xy
78	84	13	18	169	324	234
36	51	-29	-15	841	225	435
98	91	33	25	1089	625	825
25	60	-40	-6	1600	36	240
75	68	10	2	100	4	20
82	62	17	-4	289	16	68
90	86	25	20	625	400	500
62	58	-3	-8	9	64	24
65	53	0	-13	0	169	0
39	47	-26	-19	676	361	494

$\Sigma x^2 = 5398 \quad \Sigma y^2 = 2224 \quad \Sigma xy = 2840$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = \frac{2840}{\sqrt{5398}\sqrt{2224}} = \frac{2840}{73.47 \times 47.15} = \frac{2840}{3464.1105}$$

$$r = 0.819 \approx 0.8$$

★ (10) In a partially destroyed lab record, only the lines of regression of y on x & x on y are available as $4x - 5y + 33 = 0$ & $20x - 9y = 107$ respectively. calculate $\bar{x}$, $\bar{y}$ and coefficient of correlation b/w x & y.

Soln: we know that regression lines passes through $\bar{x}$ & $\bar{y}$.

$$4\bar{x} - 5\bar{y} = -33$$
$$20\bar{x} - 9\bar{y} = 107$$
$$\bar{x} = 13, \quad \bar{y} = 17$$

we shall now rewrite the eqn of the regression lines to find the regression coefficients.

$5y = 4x + 33$ (or) $y = 0.8x + 6.6 \rightarrow (1)$

$20x = 9y + 107$ (or) $x = 0.45y + 5.35 \rightarrow (2)$

Scanned with CamScanner

From ① & ②

$$r \cdot \frac{\sigma_y}{\sigma_x} = 0.8 \quad , \quad r \frac{\sigma_x}{\sigma_y} = 0.45$$

correlation coefficient $r = \sqrt{0.8 \times 0.45} = \pm 0.6$

Thus $r = 0.6$

★ Show that if θ is the angle b/w the lines of regression, then

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

Soln:- W.K.T if θ is a acute, the angle b/w the lines $y = m_1 x + c_1$ and $y = m_2 x + c_2$ is given by.

$$\tan\theta = \frac{m_2 - m_1}{1 + m_1 m_2}$$

we have the lines of regression,

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \text{ ① \& } x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}).$$

we write the second of the eqn as

$$y - \bar{y} = \frac{\sigma_y}{r\sigma_x}(x - \bar{x}) \longrightarrow ②$$

Slopes of ① & ② are respectively given by

$$m_1 = r\frac{\sigma_y}{\sigma_x} \text{ \& } m_2 = \frac{\sigma_y}{r\sigma_x}$$

Substituting these in the formula for $\tan\theta$ we have,

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

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

Scanned with CamScanner

(2) Find the rank correlation for the data

x	8	5	11	13	10	5	18	15	2	8
y	56	44	79	72	70	54	94	85	33	65

Soln:-

x	y	Rank (x)	Rank (y)	Adjust rank R_x	Adjust rank R_y	$d = R_x - R_y$
8	56	6	7	$\frac{6+7}{2} = 6.5$	7	-0.5
5	44	8	9	$\frac{8+9}{2} = 8.5$	9	-0.5
11	79	4	3	4	3	1
13	72	3	4	3	4	-1
10	70	5	5	5	5	0
5	54	8	8	8.5	8	0.5
18	94	1	1	1	1	0
15	85	2	2	2	2	0
2	33	10	10	10	10	0
8	65	6	6	6.5	6	0.5

d^2	0.25	0.25	1	1	0	0.25	0	0	0	0.25

$\Sigma d^2 = 3$

Now $\rho = 1 - \frac{6[\Sigma d^2 + C.F]}{n(n^2-1)}$

$n = 10$

$i = 2$

$C.F = \frac{1}{12}\sum_{i=1}^{K} m_i(m_i^2-1) = \frac{1}{12}\sum_{i=1}^{2} m_i(m_i^2-1)$

$m_1 = 2$

$m_2 = 2$

$= \frac{1}{12}\left[m_1(m_1^2-1) + m_2(m_2^2-1)\right]$

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

$$\therefore \rho = 1 - \frac{6[3+1]}{10(10^2-1)} = 1 - \frac{24}{990} = 0.976$$

(3) Find the rank correlation coefficient for the given data

x	50	33	40	10	15	15	65	24	15	57
y	12	12	24	6	15	4	20	9	6	18

Sol:

x	y	Rank x	Rank y	Adjust Rank Rx	Adjust Rank Ry	$d = Rx - Ry$	d^2
50	12	3	5	3	$\frac{5+6}{2} = 5.5$	-2.5	6.25
33	12	5	5	5	5.5	-0.5	0.25
40	24	4	1	4	1	3	9
10	6	10	8	10	$\frac{8+9}{2} = 8.5$	1.5	2.25
15	15	7	4	$\frac{7+8+9}{3} = 8$	4	4	16
15	4	8	10	8	10	-2	4
65	20	1	2	1	2	-1	1
24	9	6	7	6	7	-1	1
15	6	7	8	8	8.5	-0.5	0.25
57	18	2	3	2	3	-1	1

$i = 3, \quad m_1 = 3, \quad m_2 = 2, \quad m_3 = 2 \qquad \Sigma d^2 = 41$

$n = 10$

$$\rho = 1 - \frac{6\left(\Sigma d^2 + C.F\right)}{n(n^2-1)}$$

$$C.F = \frac{1}{12}\sum_{i=1}^{3} m_i(m_i^2-1) = \frac{1}{12}\left[m_1(m_1^2-1) + m_2(m_2^2-1) + m_3(m_3^2-1)\right]$$

$$C.F = \frac{1}{12}\left[3(3^2-1) + 2(2^2-1) + 2^2(2^2-1)\right]$$

$$= \frac{1}{12}\left[24 + 6 + 6\right] = \frac{36}{12} = 3$$

$$\therefore \rho = 1 - \frac{6\left[41+3\right]}{10(10^2-1)} = 1 - \frac{264}{990}$$

$$\boxed{\rho = 0.7333}$$

(3) Find the rank correlation for the data

x	68	64	75	50	64	80	75	40	55	64
y	74	58	68	45	61	60	68	48	50	74