Module IV- CURVE FITTING AND STATISTICAL METHODS

CORRELATION

CORRELATION: Covariation of two independent magnitudes is known as correlation

Ex: 1. Field of crop varies with the amount of rainfall.

2. Price of the commodity increases with the reduction of its supply.

If an increase (or decrease) in the values of one variable corresponds to an increase (or decrease) in the other ,the **correlation** is said to be **positive** .

If the increase (or decrease) in one variable corresponds to the decrease (or increase) in the other , the **correlation** is said to be **negative**.

If there is no relationship indicated between the variables, they are said to be **independent or uncorrelated.**

KARL-PEARSON'S COEFFICIENT OF CORRELATION:

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

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

Also r is given by

$$r = \frac{\sum XY}{n\sigma_x\sigma_y}.$$

$$r = \frac{\sigma^2 + \sigma^2 - \sigma^2}{2\sigma_x\sigma_y}.$$

$$r = \frac{n\sum f d_x d_y - (\sum f_x d_x)(\sum f_y d_y)}{\sqrt{[n\sum f_x d_x^2 - (\sum f_x d_x)^2] \times [n\sum f_y d_y^2 - (\sum f_y d_y)^2]}}$$ (step deviation method)

Note:

1. $X = x - \bar{x}$ and $Y = y - \bar{y}$

2. $\bar{x} = \frac{\Sigma S}{n} akd\ \bar{y} = \frac{\Sigma y}{n}$

3. $\sigma_S = \sqrt{\frac{\Sigma(s-\bar{s})^2}{n}} = \frac{\Sigma S^2}{n} - \bar{x}^2$ and $\sigma_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}} = \frac{\Sigma y^2}{n} - \bar{y}^2$

EXAMPLES:

Example 1: Psychological tests of intelligence and of engineering ability were applied to 10 students. Here is a record of ungrouped data showing intelligence ratio (IR) and Engineering ratio (ER). Calculate the coefficient of correlation.

Students	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

Solution:

Student	Intelligence ratio		Engineering ratio		X^2	Y^2	XY
	x	$X = x - \bar{x}$	y	$Y = y - \bar{y}$			
A	105	6	101	3	36	9	18
B	104	5	103	5	25	25	25
C	102	3	100	2	9	4	6
D	101	2	98	0	4	0	0
E	100	1	95	-3	1	9	-3
F	99	0	96	-2	0	4	0
G	98	-1	104	6	1	36	-6
H	96	-3	92	-6	9	36	18
I	93	-6	97	-1	36	1	6
J	92	-7	94	-4	49	16	28
Total	$\sum x = 990$	0	$\sum y = 980$	0	$\sum X^2 = 170$	$\sum Y^2 = 140$	$\sum XY = 92$

$\bar{x} = \frac{\sum S}{n} = \frac{990}{10} = 99$ akd $\bar{y} = \frac{\sum y}{n} = \frac{980}{10} = 98$

Using $r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} = \frac{92}{\sqrt{170 \times 140}} = 0.59.$

Example 2 : While calculating correlation coefficient between two variables x and y from 25 pairs of observations, the following results were obtained: $k = 25; \sum x = 125; \sum x^2 = 650; \sum y = 100; \sum y^2 = 460$; $\sum xy = 508$. 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
8	12
6	8

Obtain the correct value of correlation coefficient.

Solution: $k = 25,$

$\sum x = 125 - 6 - 8 + 8 + 6 = 125$ (6 & 8 were wrongly taken so subtracted; instead 8 & 6 are the correct values so added)

$\sum y = 100 - 14 - 6 + 12 + 8 = 100$ (14 & 6 were wrongly taken so subtracted; instead 12 & 8 are the correct values so added)

Similarly,

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

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

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

$r = \frac{n \sum Sy - (\sum S)(\sum y)}{\sqrt{\{n \sum S^2 - (\sum S)^2\}\{n \sum y^2 - (\sum y)^2\}}} = \frac{2}{3}.$

Example 3: Establish the formula $r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$. Hence calculate r from the following data

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

Solution: Let $z = x - y$ so that $\bar{z} = \bar{x} - \bar{y}$

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

$(z-\bar{z})^2 = (x-\bar{x})^2 + (y-\bar{y})^2 - 2(x-\bar{x})(y-\bar{y})$

Summing up for n terms, and dividing by n

$\frac{\sum(Z-Z)^2}{n} = \frac{\sum(S-S)^2}{n} + \frac{\sum(y-\bar{y})^2}{n} - 2\frac{\sum(S-S)(y-\bar{y})}{n}$

$\sigma_Z^2 = \sigma_S^2 + \sigma_y^2 - 2\sigma_S\sigma_y r$ $\quad \because r == \frac{\sum(S-S)(y-\bar{y})}{n\sigma_x\sigma_y}$

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

x	$X = x - 54$	X^2	y	$Y = y - 100$	Y^2	$z = x - y$	z^2
21	-33	1089	60	-40	1600	-39	1521
23	-31	969	71	-29	841	-48	2304
30	-24	576	72	-28	784	-42	1764
54	0	0	83	-17	289	-29	841
57	3	9	110	10	100	-53	2809
58	4	16	84	-16	256	-26	676
72	18	324	100	0	0	-28	784
78	24	576	92	-8	64	-14	196
87	33	1089	113	13	169	-26	676
90	36	1296	135	35	1225	-45	2025
	$\sum X = 30$	$\sum X^2 = 5936$		$\sum Y = -80$	$\sum X^2 = 5328$	$\sum X = -350$	$\sum z^2 = 13596$

$\sigma_S^2 = \frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2 = 584.6$

$\sigma_y^2 = \frac{\sum Y^2}{n} - \left(\frac{\sum y}{n}\right)^2 = 468.8$

$\sigma_{S-y}^2 = \frac{\sum Z^2}{n} - \left(\frac{\sum Z}{n}\right)^2 = 134.6$

$\therefore r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y} = 0.876$

Example 4: The following table shows the bivariate frequency distribution of marks obtained by 25 students in maths(X) and Computer science(Y). Determine the co-efficient of correlation (r).

Marks in C S (Y)	Marks in Mathematics(X)					
		21-25	26-30	31-35	36-40	41-45
	21-25	1				
	26-30		3	1		
	31-35		2	5	2	
	36-40			1	4	1
	41-45			1	3	
	46-50					1

Solution:

	X (Midpoint)	23	28	33	38	43				
Y (Midpoint)	d_S / d_y	-2	-1	0	1	2	f_y	$f_y d_y$	$f_y {d_y}^2$	Sum of the corner numbers in each row
23	-2	1 [4]					1 (1)	-2	4	4 (4)
28	-1		3 [3]	1 [0]			4 (3+1)	-4	4	3 (3+0)
33	0		2 [0]	5 [0]	2 [0]		9 (2+5+2)	0	0	0 (0+0+0)
38	1			1 [0]	4 [4]	1 [2]	6 (1+4+1)	6	6	6 (0+4+2)
43	2			1 [0]	3 [6]		4 (1+3)	8	16	6 (0+6)
48	3					1 [6]	1 (1)	3	9	6 (6)
f_S		1	5	8	9	2	$\Sigma f_S = \Sigma f_y = 25$	$\Sigma f_y d_y =$ 11	$\Sigma f_y {d_y}^2 =$ 39	$\Sigma f d_S d_y =$ 25
$f_S d_S$		-2	-5	0	9	4	$\Sigma f_S d_S =$ 6			
$f_S {d_S}^2$		4	5	0	9	8	$\Sigma f_S {d_S}^2 =$ 26			

Sum of the corner numbers in each row	4	3	0	10	8	$\sum f d_S d_y =$ 25

Here n = total frequency =25

Substituting the values from the above table in the formula for r i.e

$$r = \frac{n\sum \mathbf{f} d_x d_y - (\sum \mathbf{f}_x d_x)(\sum \mathbf{f}_y d_y)}{\sqrt{[n\sum \mathbf{f}_x {d_x}^2 - (\sum \mathbf{f}_x d_x)^2] \times [n\sum \mathbf{f}_y {d_y}^2 - (\sum \mathbf{f}_y d_y)^2]}} = 0.772$$

Note:

$d_S = \frac{S-a}{h_1}$ & $d_y = \frac{y-b}{h_2}$, a & b are the arbitrary values taken as class marks(assumed values usually the middle or near to middle value in the midpoint), h_1 akd h_2 are the size of class intervals respectively.

Purple coloured are just for detailed information, not to be included in the table.

Try these:

Example 1: Calculate the coefficient of correlation for the following ages of husbands(x) and wives(y)

X	23	27	28	28	29	30	31	33	35	36
Y	18	20	22	27	21	29	27	29	28	29

Answer: 0.82

Example 2: From 10 pairs of observations for x and y the following data is obtained n=10, $\sum x = 66$, $\sum y = 69$, $\sum x^2 = 476$, $\sum y^2 = 521$, $\sum xy = 485$. It was later found that the two pairs of (correct) values

X	Y
4	6
9	8

were copied down as

X	Y
2	3
7	5

. Obtain the correct value of correlation coefficient.

Answer: 0.55

Example 3: The correlation table below shows that the ages of husband and wife of 53 married couples living together om the census night of 1991. Calculate the coefficient of correlation for the following ages of husband and wife.

		Age of wife						Total
		15-25	25-35	35-45	45-55	55-65	65-75	
Age of husband	15-25	1	1					2
	25-35	2	12	1				15
	35-45		4	10	1			15
	45-55			3	6	1		10
	55-65				2	4	2	8
	65-75					1	2	3
Total		3	17	14	9	6	4	53

Answer: 0.91

RANK CORRELATION:

A group of n 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 order corresponding to two characteristics A & B, the correlation between these n pairs of ranks is called the **Rank correlation** in the characteristics A and B for the group of individuals.

Let x_i, y_i be the ranks of the i^{th} individuals in A and B respectively, then the rank correlation coefficient between these variables is denoted by ρ and is given by

$\rho = 1 - \frac{\sum d_i^2}{n^3 - n}$, **where** $d_i = x_i - y_i$

EXAMPLES:

Example 1: 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 ρ.

Solution:

x:	1	6	5	10	3	2	4	9	7	8
y:	6	4	9	8	1	2	3	10	5	7
$d_i = x_i - y_i$	-5	2	-4	2	2	0	1	-1	2	1
d^2	25	4	16	4	4	0	1	1	4	1

$k = 10$

$\sum d_i^2 = 60$

$\therefore \quad \rho = 1 - \frac{6\sum d_i^2}{n^3 - n} = 1 - \frac{6\times 60}{990} = 0.6(kearly).$

Try these:

Example 1:

The marks scored by recruits in the selection test(X) in the proficiency test(Y) are given below

Serial No	1	2	3	4	5	6	7	8	9
X	10	15	12	17	13	16	24	14	22
Y	30	42	45	46	33	34	40	35	39

Answer: 0.4 (hint: take X as x and Y as y)

Example 2:

Three judges A,B,C give the following ranks. Find which pair of judges have 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

Answer: The pair of judges A and C have the nearest common approach($\because \rho_{Z-S}$ $is\ maximum$).

(hint: make table as below

$A = x$	$B = y$	$C = z$	$d_1 = x - y$	$d_2 = y - z$	$d_3 = z - x$	${d_1}^2$	${d_2}^2$	${d_3}^2$

$\rho_{S-y} = -0.2, \quad \rho_{y-Z} = -0.3, \quad \rho_{Z-S} = 0.6$.)

REGRESSION ANALYSIS

Lines of regression

Introduction: 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 between $x\&y$. The best fitting straight line of the form $y = ax + b$ is called the regression line of y ok x and $x = ay + b$ is called the regression line of x ok y.

Equations of the regression lines

The equation (i) $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$ or $Y = \frac{\sum XY}{\sum X^2}(X)$ is called the regression line of y ok x

The equation (ii) $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$ or $X = \frac{\sum XY}{\sum Y^2}(Y)$ is called the regression line of x ok y

where $X = x - \bar{x}$ and $Y = y - \bar{y}$.

The coefficient of x in (i) and the coefficient of y in (ii) respectively given by $r\frac{\sigma_y}{\sigma_x}$ and $r\frac{\sigma_x}{\sigma_y}$ are known as the regression coefficients, their product is equal to r^2 and hence

$r = \sqrt{(coeffof\ x)(coeffof\ y)}$ is the coefficient of correlation

OR $r = \frac{\sum XY}{\sqrt{\sum X^2}\sqrt{\sum Y^2}}$ OR $r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$

where S.Ds can be found by applying the formula $\sigma^2 = \frac{\sum S^2}{n} - (\bar{x})^2$

EXAMPLES:

Example 1: Obtain the lines of regression and hence find the coefficient of correlation for the data

x	1	3	4	2	5	8	9	10	13	15
y	8	6	10	8	12	16	16	10	32	32

Solution: Let $\bar{x} = \frac{\sum S}{n} = \frac{70}{10} = 7$ and $\bar{y} = \frac{\sum y}{n} = \frac{150}{10} = 15$ we note that $X = x - \bar{x}$ and $Y = y - \bar{y}$

The relevant table is as follows

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
1	8	-6	-7	36	49	42
3	6	-4	-9	16	81	36
4	10	-3	-5	9	25	15
2	8	-5	-7	25	49	35
5	12	-2	-3	4	9	6
8	16	1	1	1	1	1

9	16	2	1	4	1	2
10	10	3	-5	9	25	-15
13	32	6	17	36	289	102
15	32	8	17	64	289	136
$\sum = 70$	$\sum y$=150			$\sum X^2$=204	$\sum Y^2$=818	$\sum XY$=360

The equation of regression lines be $Y = \frac{\sum XY}{\sum X^2}(X)$ and $X = \frac{\sum XY}{\sum Y^2}(Y)$ or $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$ and $x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$ that is $y - 15 = \frac{360}{204}(x - 7)$ and $x - 7 = \frac{360}{818}(y - 15)$

$\therefore y = 1.76x + 2.68$ and $x = 0.44y + 0.4$

Coefficient of correlation $r = \pm\sqrt{(coeffof\ x)(coeffof\ y)} = \sqrt{(1.76)(0.44)} = 0.88$

Example 2. Compute the coefficient of correlation and the equation of regression lines for the data

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

Solution:

First we shall prepare the following table of values

x	y	$x - y = z$	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

Now $\bar{x} = \frac{\sum S}{n} = \frac{28}{7} = 4\ akd\ \bar{y} = \frac{\sum y}{n} = \frac{77}{7} = 11\ akd\ \bar{z} = \frac{\sum Z}{n} = \frac{-49}{7} = -7$

$\sigma_S^2 = \frac{\sum S^2}{n} - (\bar{x})^2 = \frac{140}{7} - (4)^2 = 4$, $\sigma_y^2 = \frac{\sum y^2}{n} - (\bar{y})^2 = \frac{875}{7} - (11)^2 = 4$

$\sigma_Z^2 = \frac{\sum Z^2}{n} - (\bar{z})^2 = \frac{347}{7} - (-7)^2 = 0.57$

The coefficient of correlation is given by $r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y} = \frac{4+4-0.57}{2\sqrt{4}\sqrt{4}} = 0.93$

The regression lines are given by

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

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

ie $y - 11 = 0.93(x - 4)$; $x - 4 = 0.93(y - 11)$

Thus $y = 0.93x + 7.28$; $x = 0.93y - 6.23$ are the regression lines

Example 3. if $\theta$is the angle between the two regression lines, show that $\tan\theta = \frac{1-r^2}{r} \times \frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$

Explain the significance when $r = 0\ \ akd\ \ r = \pm 1$

Solution: The equations to the line of regression of y ok x akd x ok y are

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

$\therefore$ Their slopes are $m_1 = r\frac{\sigma_y}{\sigma_x}\ akd\ m_2 = \frac{\sigma_y}{r\sigma_x}$

Thus $\tan\theta = \frac{m_2 - m_1}{1 + m_1 m_2} = \frac{\sigma_y/r\sigma_x - r\sigma_y/\sigma_x}{1 + \sigma_y^2/\sigma_x^2} = \frac{1-r^2}{r} \times \frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$

When $r = 0, \tan\theta \to \infty$ or $\theta = \pi/2$ i. e. when the variables are independent, the two lines of regression are perpendicular to each other.

When $r = \pm 1$, $\tan\theta = 0$ i. e., $\theta = 0$ or π. Thus the lines of regression coincide i.e., there is perfect correlation between the two variables.

Example 3. In a partially destroyed laboratory record, only the lines of regression of y ok x akd x ok y are available as $4x - 5y + 33 = 0\ akd\ 20x - 9y = 107$ respectively. Calculate $\bar{x}$, $\bar{y}$ and the coefficient of correlation between x akd y.

Solution: since the regression lines pass through $(\bar{x}, \bar{y})$, therefore,

$4\bar{x} - 5\bar{y} + 33 = 0,\ 20\bar{x} - 9\bar{y} = 107.$

Solving these equations, we get $\bar{x} = 13$, $\bar{y} = 17$.

Rewriting the line of regression of y ok x as $y = \frac{4}{5}x + \frac{33}{5}$, we get

$b_{yS} = r\frac{\sigma_y}{\sigma_x} = \frac{4}{5}$ ---------- (i)

Rewriting the line of regression of x ok y as $x = \frac{9}{20}y + \frac{107}{9}$ we get

$$b_{Sy} = r\frac{\sigma_x}{\sigma_y} = \frac{9}{20} \text{----------} \text{(ii)}$$

Multiplying (i) and (ii) , we get $r^2 = \frac{4}{5} \times \frac{9}{20} = 0.36 \quad \therefore \quad r = 0.6$

Hence $r = 0.6$, the positive sign being taken as b_{yS} and b_{Sy} both are positive.

Example 4. $8x - 10y + 66 = 0$ akd $40x - 18y = 214$ are the two regression lines. Find the mean of x's , y's and the coefficient of correlation. Find σ_y if $\sigma_S = 3$

Solution: Given that the regression lines: $8x - 10y + 66 = 0 \Rightarrow 8x - 10y = -66$ and $40x - 18y = 214$ we know that the regression lines passes through $\bar{x}$ akd $\bar{y}$

Therefore $8\bar{x} - 10\bar{y} = -66$ ----- (i)

$40\bar{x} - 18\bar{y} = 214$ ---- (ii)

By solving we get $\bar{x} = 13$ akd $\bar{y} = 17$ are means of x's akd y's.

We shall now rewrite the equations of the regression lines to find the correlation coefficient

$10y = 8x + 66$ or $y = 0.8x + 6.6$------(iii) , $40x = 18y + 214$ or $x = 0.45y + 5.35$ --- (iv)

Correlation coefficient $r = \sqrt{(coeffof\ x)(coeffof\ y)} = \sqrt{(0.8)(0.45)} = \pm 0.6$

Also $\sigma_S = 3$ by the data and we have $r\frac{\sigma_y}{\sigma_x} = 0.8$ i. e. , $(0.6)\sigma_y = 2.4$ $thus$ $\sigma_y = 4$

Example 5. Given

	x series	y series
Mean	18	100
S.D	14	20

and $r = 0.8$, write down the equation of the regression lines and hence find the most probable value of y when $x = 70$.

Solution: By data $\bar{x} = 18$ akd $\bar{y} = 100$, $\sigma_S = 14$ akd $\sigma_y = 20$

We have the equations of the regression lines

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

i. e., $y - 100 = 0.8 \times \frac{20}{14}(x - 18)$ akd $x - 18 = 0.8 \times \frac{14}{20}(y - 100)$

i. e. , $y - 100 = 1.14(x - 18)$; $x - 18 = 0.56(y - 100)$

Thus $y = 1.14x + 79.48$; $x = 0.56y - 38$ these are the regression lines .

When $x = 70$, we obtain from the first equation $y = 1.14(70) + 79.48 = 159.28$

Try these

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

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

Example 2. Compute $\bar{x}, \bar{y}$ and r from the following equation of the regression lines:

$$2x + 3y + 1 = 0 \ ; \ x + 6y - 4 = 0$$

Example3. The two regression equations of variables x & y are $x = 19.13 - 0.87y$ akd $y = 11.64 - 0.50x$. Find a) mean of $x's$ b) mean of $y's$ c) the correlation coefficient $betweek\ x\ \&\ y$.

Curve Fitting

The process of determining a curve of best fit is called curve fitting. This method is called as the method of least square sense. It is generally used for curve fitting.

Consider $y = f(x)$ is an approximate relation that fits into a given data (x_i, y_i) i $= 1, 2, 3, 4, 5, ----, n$. Then y_i 's are called the observed values and $Y_i = f(x_i)$ are called the expected values, their difference $R_i = y_i - Y_i$ are called the residuals or estimate errors.

I) <u>Fitting of a straight line $y = a + bx$:</u>

Consider a set of $n -$ given values (x, y) for fitting the straight line $y = a + bx$, where a and b are parameters to be determined. They can be determined by using the following normal equations for fitting the straight line $y = a + bx$ in the least squares sense.

$$n\,a + (\sum x)b = \sum y \quad (1)$$

$$(\sum x)\,a + (\sum x^2)\,b = \sum x\,y \quad (2)$$

Examples:

1) Fit a straight line $y = a + bx$ in the least square sense for the data

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

Answer: Here $n = 8$ values of (x, y) are given.

We know the normal equations for straight line $y = a + bx$.

$$n\,a + (\sum x)b = \sum y \quad (1)$$

$$(\sum x)\,a + (\sum x^2)b = \sum xy \quad (2)$$

We prepare a table with respect to equations (1) and (2)

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

Put these table values in equations (1) and (2), we have

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

Solving these equations simultaneously or by calculator, we get

$$a = 0.5454, \quad b = 0.6363$$

Put these values in $y = a + bx$, we get

$$y = 0.5454 + 0.6363\,x$$

2. Find the equation of the best fitting straight line $y = ax + b$ for the following data

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

Answer: Here $n = 5$ values of (x, y) are given.

We know the normal equations for straight line $y = ax + b$

Before to write normal equations, we write straight line as $y = b + a\,x$

Now normal equations are

$$n\,b + (\sum x)\,a = \sum y \qquad (1)$$
$$\sum x\;b + (\sum x^2)\,a = \sum x\,y \qquad (2)$$

We prepare a table with respect to equations (1) and (2)

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

Put these table values in equations (1) and (2), we have

$$5\,b + 75\,a = 114,$$
$$75\,b + 1375\,a = 1885.$$

Solving these equations simultaneously or by calculator, we get

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

Put these values in $y = ax + b$, we get $y = 0.7\,x + 12.3$.

3. If P is the pull required to lift a load W by means of a pulley block, find a linear law of the form $P = m\mathrm{W} + c$ connecting P and W, using the following data

P	**12**	**15**	**21**	**25**
W	**50**	**70**	**100**	**120**

Where P and W are taken in *kg-wt.* Compute P when $\mathrm{W} = 150kg.wt.$

Answer: Here $n = 4$ values of (W, P) are given.

We know the normal equations for straight line $P = m\mathrm{W} + c$ or $P = c + m\mathrm{W}$.

$$n\,c\ + (\sum \mathrm{W})\ m = \sum P \qquad (1)$$

$$(\sum \mathrm{W})\ c\ + (\sum \mathrm{W}^2)\ m = \sum \mathrm{W}P \qquad (2)$$

We prepare the table with respect to equations (1) and (2), we have

W	P	W^2	WP
50	12	2500	600
70	15	4900	1050
100	21	10000	2100
120	25	14400	3000
$\sum \mathrm{W} = 340$	$\sum P = 73$	$\sum \mathrm{W}^2 = 31800$	$\sum \mathrm{W}P = 6750$

Put these values in equations (1) and (2), we have

$$4\,c\ + 340\,m = 73$$

$$340\,c + 3180\,m = 6750$$

After solving these equations, we get $c = 2.2785$ and $m = 0.1879$

Put c, m in equation $P = m\mathrm{W} + c$, we get

$$P = 0.1879\mathrm{W} + 2.2785. \qquad (3)$$

Now, we find P by putting $\mathrm{W} = 150\ kg$ in equation (3), we get

$$P = 0.1879\ (150) + 2.2785 \qquad \therefore P = 30.4635$$

II <u>Fitting of a second degree parabola</u> $y = a + b\mathrm{x} + c\mathrm{x}^2$**:**

Consider a set of n – given values (x, y) for fitting a second degree parabola $y = a + bx + c\,x^2$, where a, b and c are parameters to be determined. They can be determined by using the following normal equations for fitting the parabola $y = a + bx + c\,x^2$ in the least squares sense.

$$n\,a\ +\ (\sum x)\ b + (\sum x^2)\ \ c = \sum y \qquad (1)$$

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

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

Solving these equations by simultaneously or by calculator, we get $a, b. c.$

Put these in above equations, we get required best fitting parabola

$y = a + bx + c\,x^2.$

Examples:

1. **Fit a parabola of 2nd degree parabola $y = a + bx + c\,x^2$ for the data**

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

Answer: Here $n = 5$ values of (x, y) are given.

We know the normal equations for parabola $y = a + bx + cx^2$

$$n\,a + (\sum x)\,b + (\sum x^2)\,c = \sum y \quad (1)$$

$$(\sum x)\,a + (\sum x^2)b + (\sum x^3)c = \sum xy \quad (2)$$

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

We prepare the table with respect to equations (1), (2), and (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
$\sum x = 10$	$\sum y = 8.9$	$\sum xy = 21.1$	$\sum x^2 = 30$	$\sum x^2y = 66.3$	$\sum x^3 = 100$	$\sum x^4 = 354$

Put these in equations (1), (2), and (3) we have

$$5\,a + 10\,b + 30\,c = 8.9$$

$$10\,a + 30\,b + 100\,c = 21.1$$

$$30\,a + 100\,b + 354\,c = 66.3$$

After solving these equations, we get, $a = 1.078,\ b = 0.414,\ c = -0.021.$

Now, put $a,\ b,\ c,$ in $y = a + bx + cx^2$

We get, $y = 1.078 + 0.414\,x - 0.021\,x^2.$

2. **Fit a parabola of second degree parabola** $y = a + bx + c\,x^2$ **for the data**

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

Estimate y when $x = 4.5$

Answer: Here $n = 9$ values of (x, y) are given.

We know the normal equations for parabola $y = a + bx + cx^2$

$$n\,a + (\sum x)\,b + (\sum x^2)\,c = \sum y \quad (1)$$

$$(\sum x)\,a + (\sum x^2)b + (\sum x^3)c = \sum xy \quad (2)$$

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

We prepare the table with respect to equations (1), (2), and (3)

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
$\sum x = 45$	$\sum y = 74$	$\sum xy = 421$	$\sum x^2 = 285$	$\sum x^2y = 2771$	$\sum x^3 = 2025$	$\sum x^4 = 15333$

Put these in equations (1), (2), and (3) we have

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

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

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

After solving these equations, we get,

$$a = -0.9286,\ b = 3.5232,\ c = -0.2673.$$

Now, put $a,\ b,\ c,$ in $y = a + bx + cx^2$, we get

$$y = -0.9286 + 3.5232\,x - 0.2673\,x^2. \qquad (3)$$

We find y by substituting $x = 4.5$ in equation (3), we have

$$y = -0.9286 + 3.5232\,(4.5) - 0.2673\,(4.5)^2$$

$$\therefore y = 9.5125$$

III Fitting of an exponential curve $y = ax^b$:

Consider $y = ax^b,$ (1)

Taking log on both sides, we have

$$logy = loga + b\,logx,$$

Take $logy = Y,\ loga = A,\ b = B,\ logx = X$, we get

$$Y = A + BX \qquad (2)$$

This is the equation of straight line, the normal equation for straight are

$$n\,A + (\sum X)\,B = \sum Y \qquad (3)$$

$$(\sum X)\,A + (\sum X^2)\,B = \sum XY \qquad (4)$$

Solving these equations (3) and (4), we will get A and B,

But, we have to calculate a and b.

Here, we have taken $loga = A$ $\therefore$ $a = e^A$ and $b = B$

Finally, substitute a and b **in** $y = ax^b$**.** This gives the curve of best fit.

Examples:

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

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

Answer: Here $n = 5$, values of (x, y) are given

Consider $y = ax^b$

Taking log on both sides, we have

$$logy = loga + b\ logx,$$

Take $logy = Y,\ loga = A,\ b = B,\ logx = \mathrm{X}$, we get

$$Y = A + B\mathrm{X} \quad (2)$$

This is the equation of straight line, the normal equation for straight are

$$n\ A\ + (\sum \mathrm{X})\ B = \sum Y \quad (3)$$

$$(\sum \mathrm{X})\ A\ + (\sum \mathrm{X}^2)\ B = \sum \mathrm{X}Y \quad (4)$$

We prepare the table with respective equations (3) and (4)

x	$\mathrm{X} = logx$	y	$Y = logy$	X^2	$\mathrm{X}Y$
1	0	0.5	-0.6931	0	0
2	0.6931	2	0.6931	0.4804	0.4804
3	1.0986	4.5	1.5041	1.2069	1.6524
4	1.3863	8	2.0794	1.9218	2.8827
5	1.6094	12.5	2.5257	2.5902	4.0649
	$\sum X = 4.7874$		$\sum Y = 6.1092$	$\sum X^2 = 6.1993$	$\sum XY = 9.0804$

Put these values in equations (3) and (4), we have

$$5\ A\ + 4.7874\ B = 6.1092$$

$$4.7874\ A\ + 6.1993\ B = 9.0804$$

After solving these equations, we get $A = -0.6929$ and $B = 1.9998$.

Here, we have taken, $loga = A$ $\therefore a = e^A$

$\therefore a = e^{-0.9629}$ $\therefore a = 0.5001 \approx 0.5$

and $b = B$ $\therefore b = 1.998 \approx 2$**,**

Put a and b in $y = ax^b$, **we get**

$$y = 0.5\mathrm{x}^2$$

This is the curve of best fit.

2. An experiment gave the following values:

$v\,(ft/min)$	**350**	**400**	**500**	**600**
$t\,(\quad)min$	**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.

Answer: Here $n = 4$, values of (t, v) are given

Consider $v = at^b$

Taking log on both sides, we have

$$logv = loga + b\,logt,$$

Take $logv = Y,\ loga = A,\ b = B,\ logt = \mathrm{X}$, we get

$$Y = A + B\mathrm{X} \qquad (2)$$

This is the equation of straight line, the normal equation for straight are

$$n\,A\ + (\sum \mathrm{X})\,B = \sum Y \qquad (3)$$

$$(\sum \mathrm{X})\,A\ + (\sum \mathrm{X}^2)\,B = \sum \mathrm{X}Y \qquad (4)$$

We prepare the table with respective equations (3) and (4)

t	$\mathrm{X} = logt$	v	$Y = logv$	X^2	$\mathrm{X}Y$
61	1.7853	350	2.5441	3.187	4.542
26	1.4150	400	2.6021	2.002	3.682
7	0.8451	500	2.6990	0.714	2.281
2.6	0.4150	600	2.7782	0.172	1.153
	$\sum X = 4.4604$		$\sum Y = 10.6234$	$\sum X^2 = 6.075$	$\sum XY = 11.658$

Put these data in equations (3) and (4), we have

$$4\,A + 4.4604\,B = 10.6234$$

$$4.4604\,A\ + 6.075\,B = 11.658$$

After solving these equations, we get $A = 2.845$ and $B = -0.1697$.

Here, we have taken, $loga = A \quad \therefore a = e^A$

$\therefore a = e^{2.845} \qquad \therefore a = 699.8$

and $b = B \qquad \therefore b = -0.1697,$

Practice Problems:

1. Predict y at $x = 3.75$, by fitting a power curve $y = ax^b$ to the given data:

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

Answer: $y = 2.978x^{0.5143}$ $and\ y = 5.8769\ at\ x = 3.75.$

2. Fit a parabola $y = a + bx + cx^2$ to the following data:

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

Answer: $y = 0.34 - 0.78x + 0.99x^2$

3. Fit a second degree parabola

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

Answer: $y = 1.42 - 1.07x + 0.55x^2$

4. Fit a second degree parabola

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

Answer: $y = 1.04 - 0.198x + 0.244x^2$

5. The results of measurement of electric resistance R of a copper bar at various temperature $t^{\circ}C$ are listed below:

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

Find a relation $R = a + bt$ where a and b are constants to be determined.
Answer: $R = 70.052 + 0.292t.$

6. Find the best possible curve of the form $y = a + bx$, using the method of least squares for the data:

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

Answer: $y = 0.545 + 0.636x$