Unit-III

Correlation and Regression

Types of Distribution:

There are two types of distributions.

1, Univariate distribution:

A distribution in which where is only one variable, such as heights of students of a class.

2, Bivariate Distribution:

The distribution involving two variables such as heights and weights of students of a class

Correlation:

Whenever two variables x and y are so related that an increase in the one is accompanied by an increase or decrease in the other, then the variables are said to be correlated.

For example, the yield of crop varies with the amount of rain fall.

Types of correlation:

1, Positive correlation:

If an increase in the value of one variable X results in a corresponding increase in value of other variable Y on an average.

OR

If a decrease in the value of one variable X results in a corresponding decrease in value of other variable Y on an average.

The correlation is said to be positive.

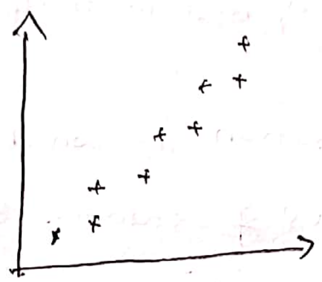
2, Negative correlation:

If the increase (decrease) in the values of one variable X results in a corresponding decrease (increase) in the values of other variable Y.

The correlation between X and Y is said to be negative.

3. Linear correlation:

when all the ploted points lie approximately on a straight line, then the correlation is said to be linear correlation.



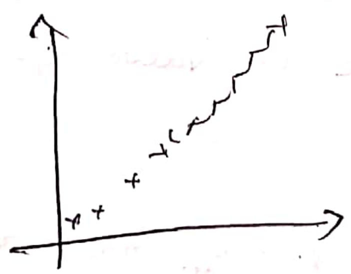
4. Perfect correlation:

If the deviation of one variable X is proportional to the deviation other variable Y, then the correlation is said to be perfect.

In this case the plotted points on a graph lie exactly on a straight line.

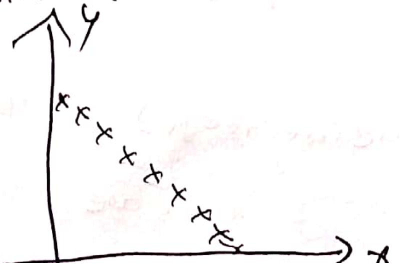
5. positive perfect correlation:

If increase in one variable X is proportional to the increase in the other variable Y. The graph will be exactly straight line.



6. Negative perfect correlation:

If increase in one variable is proportional to the decrease in the other variable. The graph will be exactly a straight line.

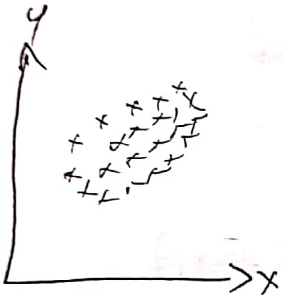


Scatter or DOT-diagram:

When we plot the corresponding values of two variables, taking one on x-axis and the other along y-axis, it shows a collection of dots.

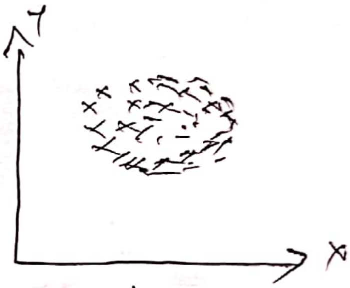
This collection of dots is called a dot diagram or a scatter diagram.

Remarks: Following are the figures of the standard ... for $r>0$, <0, $=0$ and ± 1

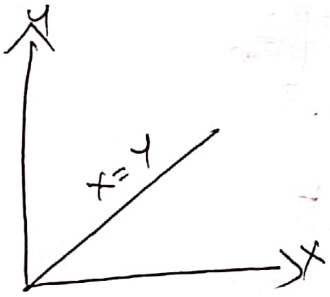


$r>0$

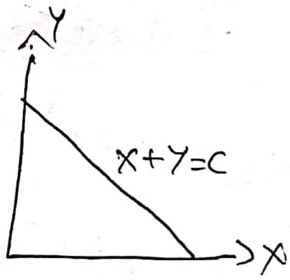
$r<0$



$r=0$



$r=+1$



$r=-1$

Karl pearson's coefficient of correlation:

Correlation coefficient between two random variables X and Y, usually denoted by $r(X,Y)$ or simply r_{XY}, is a numerical measure of linear relationship between them and is defined as:

$$r(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

If (x_i, y_i), $i=1,2 \ldots n$ in the bivariate distribution, then

$$cov(X,Y) = \frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y})$$

$$\sigma_X^2 = \frac{1}{n}\Sigma(x_i-\bar{x})^2$$

$$\sigma_Y^2 = \frac{1}{n}\Sigma(y_i-\bar{y})^2$$

Another convenient form of the above formula is

$$r(X,Y) = \frac{\frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\Sigma x_i^2 - \bar{x}^2\right)\left(\frac{1}{n}\Sigma y_i^2 - \bar{y}^2\right)}}$$

→ calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y):

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

Sol: calculation of correlation coefficient:

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
$\Sigma X = 544$	$\Sigma Y = 552$	$\Sigma X^2 = 37028$	$\Sigma Y^2 = 38132$	$\Sigma XY = 37560$

$\bar{X} = \frac{\Sigma X}{n} = \frac{544}{8} = 68$ $\quad \bar{y} = \frac{\Sigma Y}{n} = \frac{552}{8} = 69$

$$r(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma XY - \bar{X}\bar{Y}}{\sqrt{\left(\frac{1}{n}\Sigma X^2 - \bar{X}^2\right)\left(\frac{1}{n}\Sigma Y^2 - \bar{Y}^2\right)}}$$

$$= \frac{\frac{1}{8}(37560) - (68)(69)}{\sqrt{\left[\frac{1}{8}(37028) - (68)^2\right]\left[\frac{1}{8}(38132) - (69)^2\right]}}$$

Scanned with CamScanner

$$= \frac{4695 - 4692}{\sqrt{(4628.5 - 4624)(4766.5 - 4761)}} = \frac{3}{\sqrt{4.5 \times 5.5}} \approx \underline{\underline{0.603}}$$

→ Ten students got the following percentage of marks in Economics and statistics

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

calculate the coefficient of correlation

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
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	~~400~~ 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
$\Sigma x_i = 650$	$\Sigma y_i = 660$	0	0	5398	2224	2704

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{650}{10} = 65$$

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

$$r_{(x,y)} = \frac{\mathrm{cov}(x,y)}{\sigma_x \, \sigma_y}$$

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

$$= \frac{\frac{1}{10} \cdot 2704}{\sqrt{\frac{5398}{10} \cdot \frac{2224}{10}}} = \frac{2704}{\sqrt{5398 \times 2224}} = \frac{2704}{73.4 \times 47.1}$$

$$= \frac{2704}{3457} = \underline{\underline{0.78}}$$

Scanned with CamScanner

⇒ A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results:

$n=25$, $\Sigma X=125$, $\Sigma X^2=650$, $\Sigma Y=100$, $\Sigma Y^2=460$, $\Sigma XY=508$.

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

X	Y
6	14
8	6

while correct values are

X	Y
8	12
6	8

. obtain the correct value of correlation coefficient.

Solution:

$\Sigma X = 125-6-8+8+6=125$

~~$\Sigma Y = 100-12-8+14+6=$~~

$\Sigma Y = 100-14-6+12+8=100$

$\Sigma X^2 = 650-6^2-8^2+8^2+6^2=650$

$\Sigma Y^2 = 460-14^2-6^2+12^2+8^2=436$

$\Sigma XY = 508-6\times14-8\times6+8\times12+6\times8=520$

$\bar{X}=\frac{\Sigma X}{n}=\frac{125}{25}=5 \quad \bar{Y}=\frac{\Sigma Y}{n}=\frac{100}{25}=4$

$$r(X,Y)=\frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

$$=\frac{\frac{1}{n}\Sigma XY-\bar{X}\bar{Y}}{\sqrt{\left(\frac{1}{n}\Sigma X^2-\bar{X}^2\right)\left(\frac{1}{n}\Sigma Y^2-\bar{Y}^2\right)}}$$

$$=\frac{\frac{1}{25}(520)-(5)(4)}{\sqrt{\left(\frac{1}{25}(650)-25\right)\left(\frac{1}{25}(436)-4^2\right)}}$$

$$=\frac{\frac{520-500}{25}}{\sqrt{\left(\frac{650-625}{25}\right)\left(\frac{436-400}{25}\right)}}=\frac{4/5}{\sqrt{(1)\left(\frac{36}{25}\right)}}=\frac{4/5}{6/5}=\frac{4}{6}$$

$=2/3=0.67$

$\frac{520}{25}-20$

$\frac{520-500}{25}$

$\frac{16}{25}$

$\frac{80}{32}$

$\frac{1}{400}$

note: $\gamma(X,Y) = \dfrac{cov(X,Y)}{\sigma_X \sigma_Y}$

where $cov(X,Y) = E\{[X-E(X)]\{Y-E(Y)\}\} = \frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y})$

$$\sigma_X^2 = E(X-E(X))^2 = \frac{1}{n}\Sigma(x_i-\bar{x})^2$$

$$\sigma_Y^2 = E(Y-E(Y))^2 = \frac{1}{n}\Sigma(y_i-\bar{y})^2$$

→ If $Z = aX+bY$ and r is the correlation coefficient between X and Y then S.T $\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,r\,\sigma_X\sigma_Y$

S.T the correlation coefficient r between two random variables X and Y is given by $r = \dfrac{(\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2)}{2\sigma_X\sigma_Y}$ where σ_X, σ_Y and σ_{X-Y} are the standard deviations of X, Y and X-Y respectively.

Sol: Given $Z = aX+bY$ — ①

Taking expectation on both sides

$E(Z) = aE(X)+bE(Y)$ — ②

① − ② ⇒ $Z - E(Z) = a[X-E(X)] + b[Y-E(Y)]$

taking expection on both sides

$E(Z-E(Z)) = E[a[X-E(X)] + b[Y-E(Y)]]$

Squaring on both sides

$$E(Z-E(Z))^2 = E\left[a(X-E(X)) + b(Y-E(Y))\right]^2$$

$$\sigma_Z^2 = E\left[a^2(X-E(X))^2 + b^2(Y-E(Y))^2 + 2ab(X-E(X))(Y-E(Y))\right]$$

$$\sigma_Z^2 = a^2 E(X-E(X))^2 + b^2 E(Y-E(Y))^2 + 2ab\, E\left[(X-E(X))(Y-E(Y))\right]$$

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,\mathrm{cov}(X,Y)$$

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\,r\,\sigma_X\sigma_Y \quad \left(\because r = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}\right)$$

(2) put $a=1,\ b=-1$

then $Z = X-Y$

Then from above,

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 + (2)(1)(-1)\,r\,\sigma_X\sigma_Y$$

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X\sigma_Y$$

$$r = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2\sigma_X\sigma_Y}$$

⇛ The random variables X and Y are jointly normally distributed and U and V are defined by $U = X\cos\alpha + Y\sin\alpha$, $V = Y\cos\alpha - X\sin\alpha$. show that U and V will be uncorrelated if

$$\tan 2\alpha = \frac{2r\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2}.$$

Sol: Given $U = X\cos\alpha + Y\sin\alpha$ $\quad V = Y\cos\alpha - X\sin\alpha$

$E(U) = E(X)\cos\alpha + E(Y)\sin\alpha$ $\quad E(V) = E(Y)\cos\alpha - E(X)\sin\alpha$

$$[U-E(U)] = (X-E(X))\cos\alpha + (Y-E(Y))\sin\alpha$$

$$[V-E(V)] = (Y-E(Y))\cos\alpha - (X-E(X))\sin\alpha$$

$$\mathrm{cov}(U,V) = E((U-E(U))(V-E(V)))$$

$$= E\left(\left[(X-E(X))\cos\alpha + (Y-E(Y))\sin\alpha\right]\times\left(Y-E(Y))\cos\alpha - (X-E(X))\sin\alpha\right)\right)$$

Scanned with CamScanner

$$= E\Big[(X-E(X))(Y-E(Y))\cos^2\alpha - (X-E(X))^2\sin\alpha\cos\alpha + (Y-E(Y))^2\sin\alpha\cos\alpha - (Y-E(Y))(X-E(X))\sin^2\alpha\Big]$$

$$= E[(X-E(X))(Y-E(Y))]\cos^2\alpha - E(X-E(X))^2\sin\alpha\cos\alpha + E(Y-E(Y))^2\sin\alpha\cos\alpha - E(Y-E(Y))(X-E(X))\sin^2\alpha$$

$$= cov(X,Y)\cos^2\alpha - \sigma_x^2\sin\alpha\cos\alpha + \sigma_y^2\sin\alpha\cos\alpha - cov(X,Y)\sin^2\alpha$$

$$= cov(X,Y)(\cos^2\alpha - \sin^2\alpha) - \sin\alpha\cos\alpha(\sigma_x^2 - \sigma_y^2)$$

Since u and v are uncorrelated then $r(u,v)=0$

$$\Rightarrow cov(X,Y)(\cos^2\alpha - \sin^2\alpha) - \sin\alpha\cos\alpha(\sigma_x^2 - \sigma_y^2) = 0$$

$$\Rightarrow \cos 2\alpha \, cov(X,Y) - \frac{\sin 2\alpha(\sigma_x^2 - \sigma_y^2)}{2} = 0$$

$$\cos 2\alpha \, cov(X,Y) = \frac{\sin 2\alpha(\sigma_x^2 - \sigma_y^2)}{2}$$

$$\frac{2\,cov(X,Y)}{\sigma_x^2 - \sigma_y^2} = \frac{\sin 2\alpha}{\cos 2\alpha}$$

$$\Rightarrow \tan 2\alpha = \frac{2\,cov(X,Y)}{\sigma_x^2 - \sigma_y^2}$$

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

→ The variables X and Y are connected by the equation $aX+bY+c=0$ s.t the correlation between them is -1 if the signs of a and b are alike and $+1$ if they are different.

Sol: $aX+bY+c=0 \Rightarrow aE(X)+bE(Y)+c=0$

$$\Rightarrow a[X-E(X)] + b[Y-E(Y)] = 0 \Rightarrow X-E(X) = -\frac{b}{a}(Y-E(Y))$$

$$cov(X,Y) = E[(X-E(X))(Y-E(Y))]$$

$$= E\left[-\frac{b}{a}(Y-E(Y))(Y-E(Y))\right]$$

$$= E\left[-\frac{b}{a}(Y-E(Y))^2\right] = -b/a\, E(Y-E(Y))^2 = -b/a\, \sigma_y^2$$

and $\sigma_x^2 = E(X-E(X))^2 = \frac{b^2}{a^2}E(Y-E(Y))^2 = \frac{b^2}{a^2}\sigma_y^2$

$$\therefore r = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \cancel{\frac{-b/a\,\sigma_y^2}{\sqrt{\frac{b}{a}\sigma}\;\frac{b}{a}\sigma_y\,\sigma_y}} =$$

$$= \frac{-b/a\,\sigma_y^2}{\sqrt{\frac{b^2}{a^2}\sigma_y^2}\sqrt{\sigma_y^2}} = \frac{-b/a\,\sigma_y^2}{\left|\frac{b}{a}\right|\sigma_y^2}.$$

$$= \begin{cases} +1, & \text{if } b \text{ and } a \text{ are of opposite sign} \\ -1, & \text{if } b \text{ and } a \text{ are of same sign.} \end{cases}$$

→ X and Y are two random variables with variances σ_x^2 and σ_y^2 respectively and r is the coefficient of correlation between them. If $U = X+kY$ and $V = X+\left(\frac{\sigma_x}{\sigma_y}\right)Y$, find the value of k so that U and V are uncorrelated.

Sol: Given $U = X+kY$

$E(U) = E(X)+kE(Y)$

$U - E(U) = (X-E(X)) + k(Y-E(Y))$

$V = X + \frac{\sigma_x}{\sigma_y}Y$

$E(V) = E(X) + \frac{\sigma_x}{\sigma_y}E(Y)$

$V - E(V) = (X-E(X)) + \frac{\sigma_x}{\sigma_y}(Y-E(Y))$

$cov(U,V) = E\left((U-E(U))(V-E(V))\right)$

$= E\left[(X-E(X)) + k(Y-E(Y))\right) \times \left(X-E(X)\right) + \frac{\sigma_x}{\sigma_y}(Y-E(Y))$

$= E\left[(X-E(X))^2 + \frac{\sigma_x}{\sigma_y}(X-E_X)(Y-E_Y) + k\frac{\sigma_x}{\sigma_y}(Y-E(Y))^2 + \frac{\sigma_x}{\sigma_y}k(X-E(X))(Y-E(Y))\right]$

$= \cancel{E\left[\sigma_x^2 E(X-E(X))^2 + \right.}$

$E(X-E(X))^2 + \frac{\sigma_x}{\sigma_y}E((X-E_X)(Y-E_Y)) + k\frac{\sigma_x}{\sigma_y}E(Y-E(Y))^2 + \frac{\sigma_x}{\sigma_y}kE(X-E(X))(Y-E(Y))$

$= \sigma_x^2 + \frac{\sigma_x}{\sigma_y}cov(X,Y) + k\frac{\sigma_x}{\cancel{\sigma_y}}\sigma_y^{\cancel{2}} + k\frac{\sigma_x}{\sigma_y}cov(X,Y)$

$= \cancel{(\sigma_x^2 + k\sigma_x\sigma_y) + \left(\frac{\sigma_x}{\sigma_y} + k\frac{\sigma_x}{\sigma_y}\right)} cov(XY$

Scanned with CamScanner

$$= \sigma_x^2 + \frac{\sigma_x}{\sigma_y} r\sigma_x\sigma_y + k\sigma_x\sigma_y + k\frac{\sigma_x}{\cancel{\sigma_y}} r.\sigma_x\cancel{\sigma_y}$$

$$= \sigma_x^2 + r\sigma_x^2 + k\sigma_x\sigma_y + rk\sigma_x^2$$

$$= \cancel{\sigma_x(1+r}$$

U and V are uncorrelated then $r(U,V)=0$

$$\Rightarrow cov(U,V)=0$$

$$\Rightarrow \sigma_x^2 + r\sigma_x^2 + k\sigma_x\sigma_y + rk\sigma_x^2 = 0$$

$$\Rightarrow \sigma_x^2 + k\sigma_x\sigma_y = 0$$

$$\Rightarrow k = -\frac{\sigma_x^2}{\sigma_x\sigma_y} = -\frac{\sigma_x}{\sigma_y}$$

Rank correlation coefficient:- (Spearman's Formula).

Suppose $(x_1, x_2, \ldots x_n)$ & $(y_1, y_2, \ldots y_n)$ are the ranks of two variables x & y. in the order of merit, with respect to some property. The correlation between these 'n' pairs of ranks is called the rank correlation.

$$\boxed{\rho = 1 - \frac{6 \Sigma D^2}{N(N^2-1)}}$$

ρ = Rank coefficient of correlation

D^2 = sum of the squares of the differences of ~~two~~ ~~factos~~ ranks

N = No of paired observations.

This method is based on rank & is useful in dealing with qualitative characteristics such as morality, character, intelligence & beauty. It cannot be measured quantitatively as in the case of Pearson's coefficient of correlation. It is based on the ranks given to the observations. Rank correlation is application only to the individual observations.

Properties of Rank correlation coefficient.

1) The value of ρ lies between -1 & 1

2) If $\rho = 1$, there is complete agreement in the order if the ranks & the directions of the rank is same.

3) If $\rho = -1$ then there is complete disagreement in the order of the ranks & they are in opposite directions.

Procedure to solve problems :-

1) when the ranks are given.

i) Compute the difference of two ranks & denote it by D.

ii) square D & get ΣD^2

iii) obtain ρ by substituting the figures in the formula

2) when the ranks are not given, but actual data are given, then we must give ranks. we can give ranks by taking the highest as 1 or the lowest value as 1. next to the highest (lowest) as 2 & follow the same procedure for both the variables.

1) The ranks of same 16 students in Mathematics & Physics are as follows. Two numbers within brackets denote the ranks of the students in mathematics & Physics:
(1, 1), (2, 10), (3, 3), (4, 4), (5, 5), (6, 7) (7, 2) (8, 6) (9, 8) (10, 11) (11, 15) (12, 9) (13, 14) (14, 12) (15, 16) (16, 13)

Calculate the rank correlation coefficient for proficiencies of this group in Mathematics & Physics.

Rank correlation coefficient is given by:

$$\boxed{\rho = 1 - \frac{6 \Sigma D^2}{N(N^2-1)}}.$$

Ranks in Maths (x)	Ranks in Physics (Y)	$D = X - Y$	D^2
1	1	0	0
2	10	-8	64
3	3	0	0
4	4	0	0
5	5	0	0
6	7	-1	1
7	2	5	25
8	6	2	4
9	8	1	1
10	11	-1	1
11	15	-4	16
12	9	3	9
13	14	-1	1
14	12	2	4
15	16	-1	1
16	13	3	9
		0	136

$$\boxed{\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)}} = 1 - \frac{6\times136}{16(256-1)} = 1-\frac{1}{5} = \frac{4}{5} = 0.8$$

2) Ten competitions in a musical test were ranked by the 3 judges A, B & C in the following order:

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

using rank correlation method, discuss which pair of judges has the nearest approach to common tings in music.

Ranks by A (x)	Ranks by B (y)	Ranks by C (z)	$D_1 = x-y$	$D_2 = x-z$	$D_3 = y-z$	D_1^2	D_2^2	D_3^2
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	-3	-4	-1	9	16	1
10	4	8	6	2	-4	36	4	16
3	7	1	-4	2	6	16	4	36
2	10	2	-8	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	-9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	-1	1	2	1	1	4
			0	0	0	200	60	214

$$\rho(x,y) = 1 - \frac{6\Sigma D_1^2}{N(N^2-1)} = 1 - \frac{6 \times 200}{10(100-1)} = 1 - \frac{1200}{990} = -\frac{7}{33} = -0.2$$

$$\rho(x,z) = 1 - \frac{6\Sigma D_2^2}{N(N^2-1)} = 1 - \frac{6 \times 60}{10(100-1)} = 1 - \frac{360}{990} = \frac{7}{11}$$

$$\rho(y,z) = 1 - \frac{6\Sigma D_3^2}{N(N^2-1)} = 1 - \frac{6 \times 214}{10(100-1)} = -\frac{49}{165}$$

$\rho(x,z)$ is maximum, we conclude that the pair of judges A & C has the nearest approach to common likings in music.

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 ρ.

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)}.$$

x	Y	D = x - Y	D^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
			60

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 60}{10(99)} = \underline{0.6}$$

4) A Random sample of 5 college students is selected & their grades in Mathematics & statistics are found to be

Mathematics	85	60	73	40	90
statistics	93	75	65	50	80

Calculate Pearman's rank correlation coefficient.

Marks in Mathematics (X)	Ranks (X_1)	Marks in Statistics (Y)	Ranks (Y_1)	Rank difference $D = X_1 - Y_1$	D^2
85	2	93	1	1	1
60	4	75	3	1	1
73	3	65	4	-1	1
40	5	50	5	0	0
90	1	80	2	-1	1
					4

N = 5.

Rank correlation $\rho = 1 - \frac{6 \Sigma D^2}{N(N^2-1)}$

$= 1 - \frac{6 \times 4}{5(25-1)}$.

$= 1 - \frac{24}{120} = 1 - \frac{1}{5} = 1 - 0.2$

$= 0.8$.

Equal or Repeated Ranks

If any two or more persons are bracketed equal in any classification or if there is more than one item with the same value in the series then the spearman's Formula for calculating the rank correlation coefficient breaks down. In this case common ranks are given to repeated items. The common rank is the average of the ranks which these items would have assumed, if they were different from each other & the next item will get the rank next to ranks all ready assumed.

Eg:- If two individuals are placed in the 7^{th} & 8^{th} place, each of them are given the rank 7.5 & the next rank will be 9.

i.e $\frac{7+8}{2} = \frac{15}{2} = 7.5$.

Rank correlation $\boxed{R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}m(m^2-1) + \frac{1}{12}m(m^2-1) + \dots\right]}{N(N^2-1)}}$.

Note:- In the correlation formula, we add the factor $\frac{m(m^2-1)}{12}$ to ΣD^2, where m is the number of items whose ranks are common. This correlation factor is to be added for each repeated value in both the x-series & y-series.

5) From the following data calculate the rank correlation coefficient after making adjustment for tied ranks:

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

∵ it is repeated Ranks ∴

Rank correlation $R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}m(m^2-1) + \frac{1}{12}m(m^2-1) + \dots\right]}{N(N^2-1)}$.

First we have to assign ranks to the variables.

X	Rank of X (X_1)	Y	Rank of Y (Y_1)	$D = X_1 - Y_1$	D^2
48	3	13	5.5	-2.5	6.25
33	5	13	5.5	0.5	0.25
40	4	24	1	3	9
9	10	6	8.5	1.5	2.25
16	8	15	4	4	16
16	8	4	10	-2	4
65	1	20	2	-1	1
24	6	9	7	-1	1
16	8	6	8.5	0.5	0.25
57	2	19	3	-1	1
					41

In X items, 16 is repeated 3 times

$\therefore \frac{1}{12} m(m^2-1) = \frac{1}{12} \times 3(9-1) = \frac{1}{12} \times 3 \times 8 = 2$.

In Y items, 13 & 6 is repeated 2 times

$\therefore \frac{1}{12} m(m^2-1) = \frac{1}{12} \times 2(4-1) = \frac{1}{12} \times 2 \times 3 = \frac{1}{2}$.

$\frac{1}{12} m(m^2-1) = \frac{1}{12} \times 2(4-1) = \frac{1}{2}$.

$\therefore$ Rank correlation $\rho = 1 - \frac{6[\Sigma D^2 + \frac{1}{12} m(m^2-1) + \frac{1}{12} m(m^2-1) + \dots]}{N(N^2-1)}$

$$= 1 - \frac{6[41 + 2 + \frac{1}{2} + \frac{1}{2}]}{10(100-1)}$$

$$= 1 - \frac{6(44)}{990}$$

$$= 0.733$$

In Y series 68 is repeated twice which are in the positions 3rd, 4th ranks.

$\therefore$ Common rank $= \frac{3+4}{2} = \frac{7}{2} = 3.5$

$\therefore$ Common rank 3.5 is to be given for each 68.

In X series 75 repeated twice $(m=2)$ & 64 repeated thrice $(m=3)$

$\therefore$ The total correction for the X-series is.

$\frac{1}{12}m(m^2-1) = \frac{1}{12} \times 2(4-1) + \frac{1}{12} \times 3(9-1) = \frac{1}{12} \times 2 \times 3 + \frac{1}{12} \times 3 \times 8$

$= \frac{6}{12} + \frac{24}{12} = \frac{30}{12} = \frac{5}{2}$.

In Y series 68 repeated twice $(m=2)$.

$\therefore$ The total correction for the Y-series is

$= \frac{1}{12} \times 2 \times (4-1) = \frac{6}{12} = \frac{1}{2}$.

$\therefore$ Rank correlation $\rho = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}m(m^2-1) + \cdots\right]}{N(N^2-1)}$.

$= 1 - \frac{6\left[72 + \frac{5}{2} + \frac{1}{2}\right]}{10 \times 99}$.

$= 1 - \frac{6(72+3)}{10 \times 99}$

$= 1 - \frac{450}{990}$.

$\boxed{\rho = 0.5454}$

Scanned with CamScanner

7) Following are the rank obtained by 10 students in two subjects, Statistics & Mathematics. To what extent the knowledge of the students in two subjects is related?

Statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	2	4	1	5	3	9	7	10	6	8

Ranks in statistics (x)	Ranks in Mathematics (y)	$D = x - y$	D^2
1	2	-1	1
2	4	-2	4
3	1	2	4
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
9	6	3	9
10	8	2	4
			40

$$\rho = 1 - \frac{6 \Sigma D^2}{N(N^2-1)} = 1 - \frac{6 \times 40}{10(99)}$$

$$= 1 - \frac{240}{990}$$

$$= 1 - 0.24$$

$$= 0.76$$

Scanned with CamScanner

8) The following are the marks obtained by 12students in Economics & Statistics.

Economics (x)	78	56	36	66	25	75	82	62
Statistics ~~Economics~~ (y)	84	44	57	58	60	68	62	58

Compute the spearman rank correlation coefficient between x & y.

Economics (x)	Rank of $x = x_1$	Statistics (y)	Rank of $y = y_1$	$D = x_1 - y_1$	D^2
78	2	84	1	1	1
56	6	44	8	-2	4
36	7	57	7	0	0
66	4	58	5.5	-1.5	2.25
25	8	60	4	4	16
75	3	68	2	1	1
82	1	62	3	-2	4
62	5	58	5.5	-0.5	0.25
					28.5

In Y-series 58 is repeated twice.

$\therefore$ correlation factor $= \frac{m(m^2-1)}{12} = \frac{1}{12} \times 2(4-1) = \frac{1}{2} = 0.5$.

$$\text{Rank correlation} = \rho = 1 - \frac{6 \Sigma D^2}{N(N^2-1)}$$

$$= 1 - \frac{6 \times 28.5}{8(64-1)}$$

$$= 1 - \frac{6 \times 28.5}{8 \times 63}$$

$$= 1 - 0.345 = 0.655$$

→ If X and Y are two random variables with variances σ_X^2 and σ_Y^2 respectively and r is the coefficient of correlation between them. If $U = X + kY$ and $V = X + \left(\frac{\sigma_X}{\sigma_Y}\right)Y$. Find the value of k so that U and V are uncorrelated.

Sol: Given $U = X + kY$ $\quad V = X + \frac{\sigma_X}{\sigma_Y} Y$

$$E(U) = E(X) + kE(Y) \quad E(V) = E(X) + \frac{\sigma_X}{\sigma_Y} E(Y)$$

$$U - E(U) = (X - E(X)) + k(Y - E(Y)) \quad V - E(V) = [X - E(X)] + \frac{\sigma_X}{\sigma_Y}[Y - E(Y)]$$

$$cov(U, V) = E[(U - E(U))(V - E(V))]$$

$$= E[(X - E(X)) + k(Y - E(Y))][(X - E(X)] + \frac{\sigma_X}{\sigma_Y}[Y - E(Y)]]$$

$$= E[\{X - E(X)\}^2 + \frac{\sigma_X}{\sigma_Y}(X - E_X)(Y - E(Y)) + k(Y - E(Y))(X - E(X)) + k\frac{\sigma_X}{\sigma_Y}(Y - E(Y))^2]$$

$$= E(X - E(X))^2 + \frac{\sigma_X}{\sigma_Y} E[(X - E_X)(Y - E_Y)] + k E(Y - E(Y))(X - E(X)) + k\frac{\sigma_X}{\sigma_Y} E(Y - E(Y))^2$$

$$= \sigma_X^2 + \frac{\sigma_X}{\sigma_Y} cov(X, Y) + k\, cov(X, Y) + k\frac{\sigma_X}{\sigma_Y}\sigma_Y^2$$

$$= \sigma_X^2 + \frac{\sigma_X}{\cancel{\sigma_Y}} r\sigma_X \cancel{\sigma_Y} + k(r\sigma_X\sigma_Y) + k\sigma_X\sigma_Y$$

$$= \sigma_X^2 + r\sigma_X^2 + kr\sigma_X\sigma_Y + k\sigma_X\sigma_Y$$

$$= \sigma_X^2(1 + r) + k\sigma_X\sigma_Y(r + 1)$$

$$= (1 + r)(\sigma_X^2 + k\sigma_X\sigma_Y)$$

Since U and V are uncorrelated, $cov(U, V) = 0$

$$\Rightarrow (1 + r)(\sigma_X^2 + k\sigma_X\sigma_Y) = 0$$

$$\Rightarrow \sigma_X^2 + k\sigma_X\sigma_Y = 0$$

$$\Rightarrow k = \frac{-\sigma_X^2}{\sigma_X\sigma_Y} = -\frac{\sigma_X}{\sigma_Y}$$

coefficient of correlation for grouped data:

when x and y series are both given as frequency distributions, these can be represented by a two way table known as the correlation table. It is double entry table with one series along the horizontal and the other along the vertical. The coefficient of correlation for such a bivariate frequency distribution is calculated by the formula.

$$\gamma = \frac{n(\Sigma f d_x d_y) - (\Sigma f d_x)(\Sigma f d_y)}{\sqrt{[n \Sigma f d_x^2 - (\Sigma f d_x)^2][n \Sigma f d_y^2 - (\Sigma f d_y)^2]}}$$

where d_x = deviation of the central values from the assumed mean of x-series

d_y = deviation of the central values from the assumed mean of y-series.

f = frequency corresponding to the pair (x, y)

$n = (\Sigma f)$ the total number of frequencies

→ The following table gives, according to age, the frequency of marks obtained by 100 students in an intelligence test:

Age in years → / Marks ↓	18	19	20	21	Total
10-20	4	2	2	-	8
20-30	5	4	6	4	19
30-40	6	8	10	11	35
40-50	4	4	6	8	22
50-60	-	2	4	4	10
60-70	-	2	3	1	6
Total	19	22	31	28	100

calculate the correlation coefficient

$d_x = x_i - A,\quad d_y = \frac{y_i - A}{h}$

$= x_i - 19 \qquad = \frac{y_i - 35}{10}$

Age in years		x	18	19	20	21				
		d_x	-1	0	1	2				
Males y	Mid pt y	d_y					f	fd_y	fd_y^2	fd_xd_y
10-20	15	-2	4 (8)	2 (0)	2 (-4)		8	-16	32	4
20-30	25	-1	5 (5)	4 (0)	6 (-6)	4 (-8)	19	-19	19	-9
30-40	35	0	6 (0)	8 (0)	10 (0)	11 (0)	35	0	0	0
40-50	45	1	4 (-4)	4 (0)	6 (6)	8 (16)	22	22	22	18
50-60	55	2		2 (0)	4 (8)	4 (16)	10	20	40	24
60-70	65	3		2 (0)	3 (9)	1 (6)	6	18	54	15
Total f			19	22	31	28	n=100	25	167	52
fd_x			-19	0	31	56	$\Sigma fd_x = 68$			
fd_x^2			19	0	31	112	$\Sigma fd_x^2 = 162$			
fd_xd_y			9	0	13	30	52			

$$r = \frac{n(\Sigma f d_x d_y) - (\Sigma f d_x)(\Sigma f d_y)}{\sqrt{\left[n\Sigma f d_x^2 - \{\Sigma (f d_x)\}^2\right]\left[n\Sigma f d_y^2 - \{\Sigma (f d_y)\}^2\right]}}$$

$$= \frac{100(52) - (68)(25)}{\sqrt{\left[100(162) - (68)^2\right]\left[100(167) - (25)^2\right]}}$$

$$= \underline{\underline{0.25}}$$

→ calculate the coefficient of correlation for the following table:

x age → / wgts y	0-4	4-8	8-12	12-16	Total
0-5	7	–	–	–	7
5-10	6	8	–	–	14
10-15	–	5	3	–	8
15-20	–	7	2	–	9
20-25	–	–	–	9	9
Total	13	20	5	9	47

Sol: $d_x = \frac{x_i - A}{h} = \frac{x_i - 10}{4}$, $d_y = \frac{y_i - A}{h} = \frac{y_i - 12.5}{5}$

		x age	0-4	4-8	8-12	12-16				
		midpt x	2	6	10	14				
weights y	midpt y	d_x / d_y	-2	-1	0	1	f	fd_y	fd_y^2	fd_xd_y
0-5	2.5	-2	7 (28)				7	-14	28	28
5-10	7.5	-1	6 (12)	8 (8)			14	-14	14	20
10-15	12.5	0		5 (0)	3 (0)		8	0	0	0
15-20	17.5	1		7 (-7)	2 (0)		9	9	9	-7
20-25	22.5	2				9 (18)	9	18	36	18
f			13	20	5	9	n = 47	-1	87	59
fd_x			-26	-20	0	9	-37			
fd_x^2			52	20	0	9	81			
fd_xd_y			40	1		18	59			

$$r = \frac{n(\Sigma f\, d_x d_y) - (\Sigma f\, d_x)(\Sigma f\, d_y)}{\sqrt{\left[n\Sigma f d_x^2 - (\Sigma f d_x)^2\right]\left[n\Sigma f d_y^2 - (\Sigma f d_y)^2\right]}}$$

$$= \frac{47(59) - (-37)(-1)}{\sqrt{\left[47(81) - (-37)^2\right]\left(47(87) - (-1)^2\right)}}$$

$$= 0.87$$

→ Find the correlation coefficient between the age and sum assured from the following table

Sum assured → / Age-group ↓	10,000	20,000	30,000	40,000	50,000	No. of persons
20-30	4	6	3	7	1	21
30-40	2	8	15	7	1	33
40-50	3	9	12	6	2	32
50-60	8	4	2	–	–	14
	17	27	32	20	4	100

Age-group ↓ / Sum-assured →	Midpt y	d_y \ d_x	10,000 ($d_x=-2$)	20,000 ($d_x=-1$)	30,000 ($d_x=0$)	40,000 ($d_x=1$)	50,000 ($d_x=2$)	f	fd_y	fd_y^2	fd_xd_y
20-30	25	-2	(16) 4	(12) 6	(0) 3	(-14) 7	(-4) 1	21	-42	84	10
30-40	35	-1	(4) 2	(8) 8	(0) 15	(-7) 7	(-2) 1	33	-33	33	3
40-50	45	0	(0) 3	(0) 9	(0) 12	(0) 6	(0) 2	32	0	0	0
50-60	55	1	(-16) 8	(-4) 4	(0) 2	-	-	14	14	14	-20
f			17	27	32	20	4	100	-61	131	-7
fd_x			-34	-27	0	20	8	-33			
fd_x^2			68	27	0	20	16	131			
fd_xd_y			4	16	0	-21	-6	-7			

$d_x = \frac{x_i - A}{h} = \frac{x_i - 30{,}000}{h}$ $\qquad d_y = \frac{y_i - A}{h} = \frac{y_i - 45}{10}$

$$r = \frac{n(\Sigma f d_x d_y) - (\Sigma f d_x)(\Sigma f d_y)}{\sqrt{\left[n\Sigma f d_x^2 - \left[\Sigma(f d_x)\right]^2\right]\left[n\Sigma f d_y^2 - \left[\Sigma(f d_y)\right]^2\right]}}$$

$$= \frac{100(-7) - (-33)(-61)}{\sqrt{\left[100(131) - (-33)^2\right]\left[100(131) - (-61)^2\right]}}$$

$$= -0.2556$$

Regression

Introduction:

The study of correlation measures the direction and strength of the relationship between two variables. In correlation we can estimate the value of one variable, when the value of the other variable is given. Since price and supply are correlated, we can find out the expected amount of supply for a given price & we can find out the required price to get a given amount of supply.

In regression, we can estimate the value of one variable with the value of the other variable which is known. The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called regression. The line described in the average relationship between two variables is known as line of regression.

Uses:

1, It is used to estimate the relation between two economic variables like income and expentiture.

2, It is highly valuable tool in Economics & Business

3, widely used for prediction purpose.

4, we can calulate the coefficient of correlation and coefficient of determination with the help of the regression coefficient.

5, It is useful in statistical estimation of demand curves, supply curves, production function, cost function and consumption function etc.

Comparison between correlation and Regression:

The correlation coefficient is a measure of degree of covariability between two variables, while the regression establishes a functional relation between dependent and independent variables. In correlation, both the variables X and Y are random variables where as in Regression, X is a random variable and Y is fixed variable. The coefficient of correlation is a relative measure where as Regression is an absolute figure.

If there is a function relationship between the two variables X and Y the points in the scatter diagram cluster around some curve is called the curve of regression.

If this curve is a straight line, then there is a relationship between the two variables and this straight line is called the line of regression.

The line of regression of Y on X:

If we minimize the sum of squares of the residuals of the ordinates between the points of equal abscissa on the line to be fitted, the line fitted is called the line of regression of Y on X.

The line of regression of X on Y:

If we minimize the sum of the squares of the residuals of abscissa between the points of equal ordinates on the line to be fitted, the fitted line is called the line of regression of X on Y.

Regression coefficient of X on Y:

The slope of the regression line on X of Y is called the regression coefficient of X on Y and is denoted by b_{xy}.

$$\boxed{b_{xy} = r\frac{\sigma_x}{\sigma_y}}$$

Regression coefficient of Y on X:

The slope of the regression line of Y on X is called the regression coefficient of Y on X and is denoted by b_{yx}.

$$\boxed{b_{yx} = r\frac{\sigma_x}{\sigma_y}}$$

The lines of regression can be found in two ways:

1) By finding $\bar{x}, \bar{y}, \sigma_x, \sigma_y$ and coefficient of correlation r and substitute in the equations.

line of regression of X on Y, $x-\bar{x} = r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$

line of regression of Y on X, $y-\bar{y} = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$

2) By least square method.

→ Obtain the equations of two lines of regression for the following data. Also obtain the estimate of X for Y=70.

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

Sol :

x	$X = x - \bar{x}$	X^2	y	$Y = y - \bar{y}$	Y^2	XY
65	-3	9	67	-2	4	6
66	-2	4	68	-1	1	2
67	-1	1	65	-4	16	4
67	-1	1	68	-1	1	1
68	0	0	72	3	9	0
69	1	1	72	3	9	3
70	2	4	69	0	0	0
72	4	16	71	2	4	8
544		36	552		44	24

$$\bar{x} = \frac{\Sigma x}{n} = \frac{544}{8} = 68$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{552}{8} = 69$$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} = \frac{24}{\sqrt{(36)(44)}} = \frac{24}{6 \cdot 2\sqrt{11}} = \frac{2}{\sqrt{11}} = 0.6030$$

The line of regression of Y on X

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

$$\sigma_x^2 = \frac{\Sigma X^2}{n} = \frac{36}{8} = 4.5 \qquad \sigma_x = \sqrt{4.5} = 2.12$$

$$\sigma_y^2 = \frac{\Sigma Y^2}{n} = \frac{44}{8} = 5.5 \qquad \sigma_y = \sqrt{5.5} = 2.35$$

$$(Y-69) = (0.6030)\,\frac{2.35}{2.12}(X-68)$$

$$Y-69 = 0.6637(X-68)$$

$$Y-69 = 0.664X - \cancel{Y} 45.134$$

$$Y = 69 + 0.664X - 45.134$$

$$Y = 0.664X + 23.866$$

The line of regression of X on Y

$$X-\bar{X} = r\frac{\sigma_X}{\sigma_Y}(Y-\bar{Y})$$

$$r\frac{\sigma_X}{\sigma_Y} = (0.603)\times\frac{2.12}{2.35} = \frac{1.272}{2.35} = 0.5412$$

$$X-68 = 0.5412(Y-64)$$

$$X = 68 + 0.54Y - 37.343$$

$$X = 0.54Y + 30.66$$

~~If~~ To estimate of X for $Y=70$, we use the line of regression of X on Y

If $Y=70$, estimate value of X is given by

$$X = 0.54Y + 30.66$$

$$= 0.54(70) + 30.66 = 68.46$$

→ Find the mean values of the variables X and Y and correlation coefficient from the following regression equations

$2Y-X-50=0$ and $3Y-2X-10=0$.

Sol: Given $2Y-X=50$ — ①

$3Y-2X=10$ — ②

By solving above equations, we get $X=130$, $Y=90$

Let $2y-x=50$ and $3y-2x=10$ be the lines of regression of y on x and x on y. These equations can be put in the form

$$2y-x=50 \qquad 3y-2x=10$$

$$2y=x+50 \qquad 2x=3y-10$$

$$y=\frac{1}{2}x+25 \qquad x=\frac{3}{2}y-5$$

$$b_{yx}=r\frac{\sigma_y}{\sigma_x}=\frac{1}{2}<1 \qquad b_{xy}=r\frac{\sigma_x}{\sigma_y}=3/2>1$$

= regression coeff of y on x = Regression coeff. of x on y.

(property of regression coefficients. i.e "if one of the regression coefficients is greater than unity, the other must be less than unity")

$$r^2=b_{yx}\,b_{xy}$$

$$=\frac{1}{2}\cdot\frac{3}{2}=3/4$$

$$r=\sqrt{\frac{3}{4}}=\frac{\sqrt{3}}{2}=0.866$$

→ In a partially destroyed laboratory, records of an analysis of correlation data, the following results only are legible: variance of $x=9$, Regression equations: $8x-10y+66=0$, $40x-18y=214$. what are: i, the mean of x and y ii, the correlation coefficient between x and y iii, the S.D of y.

Sol

i, Given $8x-10y+66=0$ — ①

$40x-18y-214=0$ — ②

By solving ① & ② we get $x=13$, $y=17$

ii, Let $8X-10Y+66=0$, $40X-18Y-214=0$ be the lines of regression of Y on X and X on Y. The eqs can be put in the form $Y=\frac{8}{10}X+\frac{66}{10}$ and $X=\frac{18}{40}Y+\frac{214}{40}$

$\therefore b_{YX}$ = Regression coeff. of Y on $X=\frac{8}{10}=4/5$

b_{XY} = Regression coeff of X on $Y=\frac{18}{40}=9/20$

$\therefore b_{YX} \cdot b_{XY}=\left(r\frac{\sigma_Y}{\sigma_X}\right)\left(r\frac{\sigma_X}{\sigma_Y}\right)=r^2$

$$r^2=(4/5)(9/20)=9/25$$

$$r=\pm 3/5=\pm 0.6$$

Since both the regression coefficients are +ve, we take

$r=0.6$

iii, Standard deviation of Y:

$$b_{YX}=r\frac{\sigma_Y}{\sigma_X}$$

$$4/5=0.6\cdot\frac{\sigma_Y}{3} \quad (\because \text{Variance } \sigma_X^2=9)$$

$$\Rightarrow \sigma_Y=4/5\times\frac{3}{0.6}\Rightarrow \sigma_Y=4$$

Note: If we assumed that $8X-10Y+66=0$, is the eqn of line of regression of X on Y and $40X-18Y=214$ is the eqn of line of regression of Y on X then

$8X-10Y+66=0$

$40X-18Y=214$

$X=\frac{10}{8}Y+\frac{66}{8}$

$Y=\frac{40}{18}X-\frac{214}{18}$

$b_{XY}=10/8$

$b_{YX}=40/18$

$r^2=b_{XY}\cdot b_{YX}=\frac{10}{8}\times\frac{40}{18}=2.78 \Rightarrow r=\sqrt{2.78}=1.66$

Since r^2 always lie between 0 & 1, our assumption is wrong.

→ Find the most likely price in Mumbai corresponding to the price of Rs.70 at kolkata from the following.

	kolkata	Mumbai
Average price	65	67
standard deviation	2.5	3.5

correlation coefficient between the prices of commodities in the two cities is 0.8.

Sol: Let the prices (in Rupees) in kolkata & Mumbai are denoted by X and Y respectively, then

$$\bar{X}=65,\ \cancel{\bar{y}=70}\ \bar{Y}=67,\ \sigma_X=2.5,\ \sigma_y=3.5,\ r=r(X,Y)=0.8$$

Line of regression of Y on X is

$$Y-\bar{Y}=r\frac{\sigma_Y}{\sigma_X}(X-\bar{X})$$

$$Y-67=0.8\times\frac{3.5}{2.5}(X-65)$$

when $X=70$

$$Y=67+1.12(70-65)$$

$$=67+1.12(5)=72.6$$

Hence the most likely price in Mumbai corresponding to the price of 70 Rs at ~~kolkata~~ is Rs 72.60

→ from the following data calculate

i) correlation coefficient ii) standard deviation of $Y(\sigma_y)$

Given $b_{xy}=0.85$ $b_{yx}=0.89$, $\sigma_X=3$

Sol: i) $b_{xy}=0.85$ $b_{yx}=0.89$, $\sigma_X=3$

$$r^2=b_{xy}\cdot b_{yx}=(0.85)(0.89)$$

$$r=\sqrt{(0.85)(0.89)}=0.87$$

ii, Standard deviation of y:

we can use b_{xy} & b_{yx} formula

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$0.89 = 0.87 \times \frac{\sigma_y}{3}$$

$$\Rightarrow \sigma_y = \frac{0.89 \times 3}{0.87} = 3 \times 1.022 = 3.068 \simeq 3.07$$

→ Given the following data, calculate the expected value of y when $x = 12$.

	x	y
Average	7.6	14.8
Standard deviation	3.6	2.5

$r = 0.99$

Sol: Given $\bar{x} = 7.6$, $\bar{y} = 14.8$, $\sigma_x = 3.6$, $\sigma_y = 2.5$, $r = 0.99$

consider the line of regression of y on x is

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

$$(y - 14.8) = 0.99 \cdot \frac{2.5}{3.6}(x - 7.6)$$

$$y = 14.8 + 0.99 \times \frac{2.5}{3.6}(x - 7.6)$$

$$y = 14.8 + 0.99 \times 0.694(x - 7.6)$$

$$= 14.8 + 0.6875(x - 7.6)$$

$$= 14.8 - 5.225 + 0.6875x$$

$$y = 9.575 + 0.688x$$

when $x = 12$, $y = 9.575 + 0.688(12)$

$$= 17.831$$

∴ Expected value of y is 17.831

→ From the data given below find

a, The two regression coefficients

b, The two regression equations

c, The coefficient of correlation between the marks in economics and statistics

d, the most likely marks in statistics, when marks in economics are 30.

Marks in Economics	25	28	35	32	31	36	29	38	34	32
Marks in Statistics	43	46	49	41	36	32	31	30	33	39

sol: Let us denote the marks in economics by the variable x and the marks in statistics by the variable y.

x	y	$X = x - \bar{x}$	$Y = y - \bar{y}$	X^2	Y^2	XY
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	+3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
28	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
320	380	0	0	140	398	-93

$\bar{X} = \frac{\Sigma x}{n} = \frac{320}{10} = 32$

$\bar{y} = \frac{\Sigma y}{n} = \frac{380}{10} = 38$

a, Regression coefficients:

coefficient of regression of y on x = $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

$$b_{yx} = \frac{cov(X,Y)}{\sigma_x \cancel{\sigma_y}} \frac{\cancel{\sigma_y}}{\sigma_x}$$

$$= \frac{cov(X,Y)}{\sigma_x^2}$$

$$= \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2}$$

$$= \frac{\Sigma xy}{\Sigma x^2} = \frac{-93}{140} = -0.6643$$

coefficient of regression of x on Y = $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$$= \frac{cov(X,Y)}{\cancel{\sigma_x}\sigma_y} \cdot \frac{\cancel{\sigma_x}}{\sigma_y} = \frac{cov(X,Y)}{\sigma_y^2} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(y_i-\bar{y})^2}$$

$$= \frac{\Sigma xy}{\Sigma y^2} = \frac{-93}{398} = -0.2337$$

b, Regression equations:

Equation of the line of regression of y on x is

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

$$y-38 = -0.6643(x-32)$$

$$y = 38 - 0.6643x + (0.6643)32$$

$$y = -0.6643x + 59.2576$$

Equation of the line of regression of x on y :

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

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

$$x-32 = -0.2337(y-38)$$

$$x-32 = -0.2337y + 8.8806$$

$$x = -0.2337y + 8.8806 + 32$$

$$x = -0.2337y + 40.8806$$

c) coefficient of correlation:

$$r^2 = b_{yx} \cdot b_{xy} = -0.6643 \times -0.2337 = 0.1552$$

$$r = \pm\sqrt{0.1552} = \pm 0.394$$

both the regression coefficients are negative, r must be negative, Hence $\underline{r = -0.394}$

d) the most likely marks in statistics, when marks in economics are 30.

we use line of regression of y on x

$$y = 0.6643x + 54.2576$$

$$y = 0.6643(30) + 54.2576$$

$$= \underline{39.3236 \simeq 39}$$

→ can $Y = 5 + 2.8X$ and $X = 3 - 0.5Y$ be the estimated regression equations of y on x and x and y respectively? Explain your answer with suitable theoretical arguments.

sol:

Sol: Line of regression of y on X is $Y = 5 + 2.8X \Rightarrow b_{yx} = 2.8$

Line of regression of X on Y is $X = 3 - 0.5Y \Rightarrow b_{xy} = -0.5$

This is not possible, since each of the regression coefficients b_{yx} and b_{xy} must have the same sign, which is same as that of $cov(X,Y)$. If $cov(X,Y)$ is positive, then both the regression coefficients are positive and if $cov(X,Y)$ is negative, then both the regression coefficients are negative.

Hence these eqns cannot be estimated as regression eqns of Y on X and X on Y respectively.

Curve fitting

curve fitting:

the exact mathematical relationship between two variables is given by a simple algebraic expression is called curve fitting.

Method of least squares:

Let $y_1, y_2 \ldots y_m$ be the values corresponding to $x_1, x_2 \ldots x_m$ which satisfy the given curve $y=f(x)$.

Let y_i be the experimental value at $x=x_i$ and the corresponding value on the fitting curve is $f(x_i)$.

Then the residual error e_i at $x=x_i$ is given by

$e_i = y_i - f(x_i)$

$S = \Sigma e_i^2 = (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \ldots + [y_m - f(x_m)]^2$

$= e_1^2 + e_2^2 + \ldots + e_m^2 = \Sigma e_i^2$

then the method of least squares aims to minimize the value of S.

Fitting of a straight line:

Let $y = a + bx$ be a straight line for point (x_i, y_i), $i = 1, 2 \ldots n$.

Let $y_1, y_2 \ldots y_n$ be the experimental values at $x=x_i$ and corresponding value on the fitting curve is $f(x_i)$

Now we have to minimize $S = \Sigma e_i^2 = \Sigma (y_i - f(x_i))^2$

$= \Sigma [y_i - (a + bx_i)]^2$

$\frac{dS}{da} = 0 \Rightarrow 2\Sigma [y_i - (a+bx_i)](-1) = 0$

$-\Sigma [y_i - (a+bx_i)] = 0$

$\Sigma y_i - \Sigma (a + bx_i) = 0$

$\sum y_i = \sum (a + b x_i)$

$\sum y_i = \sum a + b \sum x_i$

$\sum y_i = a \sum 1 + b \sum x_i$

$\sum y_i = an + b \sum x_i$ — (1)

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

$\sum_{i=1}^{n} (-y_i x_i + (a + b x_i) x_i) = 0$

$\sum_{i=1}^{n} (-y_i x_i + a x_i + b x_i^2) = 0$

$-\sum_{i=1}^{n} y_i x_i + a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 = 0$

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

$\Rightarrow \sum y_i x_i = a \sum x_i + b \sum x^2$ — (2)

Here the eqns (1) & (2) are known as normal equations.

By solving (1) & (2), we get the values of a and b, which gives the best fit of a straight line.

Fitting of a parabola:

Let $y = a + bx + cx^2$ be the second degree equation or parabola for (x_i, y_i), $i = 1, 2 .. n$

Let $y_1, y_2 -- y_n$ be the set of experimental values and $f(x_1), f(x_2) .. f(x_n)$ are theoretical values.

Now we have to minimize $S = \sum e_i^2 = \sum (y_i - f(x_i))^2$

$= \sum [y_i - (a + b x_i + c x_i^2)]^2$

consider $\frac{\partial S}{\partial a} = 0 \Rightarrow \sum 2\left[y_i - (a+bx_i+cx_i^2)\right](-1) = 0$

$$\sum (y_i - (a+bx_i+cx_i^2))(-1) = 0$$

$$= \sum(-y_i) + \sum(a+bx_i+cx_i^2) = 0$$

$$\Rightarrow \sum y_i = \sum a + b\sum x_i + c\sum x_i^2$$

$$\sum y_i = a\sum 1 + b\sum x_i + c\sum x_i^2$$

$$\sum y_i = an + b\sum x_i + c\sum x_i^2 \quad \text{— (1)}$$

$$\frac{\partial S}{\partial b} = 0 \Rightarrow \sum 2\left[y_i - (a+bx_i+cx_i^2)\right](-x_i) = 0$$

$$\sum\left[-y_i x_i + ax_i + bx_i^2 + cx_i^3\right] = 0$$

$$-\sum y_i x_i + a\sum x_i + b\sum x_i^2 + c\sum x_i^3 = 0$$

$$\Rightarrow \sum y_i x_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3 \quad \text{— (2)}$$

$$\frac{\partial S}{\partial c} = 0 \Rightarrow \sum 2\left[y_i - (a+bx_i+cx_i^2)\right](x_i^2) = 0$$

$$\sum \left[y_i x_i^2 - ax_i^2 - bx_i^3 - cx_i^4\right] = 0$$

$$\sum y_i x_i^2 - a\sum x_i^2 - b\sum x_i^3 - c\sum x_i^4 = 0$$

$$\sum y_i x_i^2 = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4 \quad \text{— (3)}$$

Here (1), (2) & (3) are normal equations of parabola and by solve these, we get the best values of a, b and c to fit a parabola.

Note:

1. Normal equation of $y = a + bx$ are

$$\cancel{\Sigma y_i = a\Sigma} \quad \Sigma y_i = na + b\Sigma x_i$$

$$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2$$

2. Normal equation of $y = a + bx + cx^2$ are

$$\Sigma y_i = an + b\Sigma x_i + c\Sigma x_i^2$$

$$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 + c\Sigma x_i^3$$

$$\Sigma y_i x_i^2 = a\Sigma x_i^2 + b\Sigma x_i^3 + c\Sigma x_i^4$$

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

x	-2	-1	0	1	2
y	1	2	3	3	4

Sol: Let the straight line be $y = a + bx$

normal equations are $\Sigma y_i = an + b\Sigma x_i$

$$\Sigma y_i x_i = a\Sigma x_i + b\Sigma x_i^2$$

x_i	y_i	x_i^2	~~y_i^2~~	$x_i y_i$
-2	1	4	~~4~~	-2
-1	2	1	~~9~~	-2
0	3	0	~~9~~	0
1	3	1	~~16~~	3
2	4	4	~~39~~	8
0	13	10		7

$\Rightarrow \quad 13 = 5a + 0.6 \Rightarrow a = 13/5 = 2.6$

$7 = 0.a + 10b \Rightarrow b = 7/10 = 0.7$

$\boxed{\therefore a = 2.6}$ $\boxed{b = 0.7}$

$\therefore$ the equation of straight line is $y = a + bx$

$\boxed{y = 2.6 + 0.7x}$

→ ~~Fit the~~

→ Find the values of a and b so that $y = a + bx$ fits the data given in the table.

x	0	1	2	3	4
y	1	2.9	4.8	6.7	8.6

Sol:

Given $y = a + bx$

normal eqns are $\sum y_i = na + b\sum x_i$

$\sum y_i x_i = a\sum x_i + b\sum x_i^2$

x_i	y_i	x_i^2	$x_i y_i$
0	1	0	0
1	2.9	1	2.9
2	4.8	4	9.6
3	6.7	9	20.1
4	8.6	16	34.4
$\sum x_i = 10$	$\sum y_i = 24$	$\sum x_i^2 = 30$	$\sum x_i y_i = 67$

$\therefore 24 = 5a + 10b$

$67 = 10a + 30b$

By solving we get

$a = 1 \quad b = 1.9$

$\therefore$ straight line is

$\boxed{y = 1 + 1.9x}$

→ The table below gives the temperature T in centigrades and lengths l in mm of a heated rod. If $l = a_0 + a_1 T$. Find a_0 and a_1 using least square method.

T	40	50	60	70	80
l	600.5	600.6	600.8	600.9	601

Sol: To fit the line $l = a_0 + a_1 T$

normal equations are $\Sigma l_i = n a_0 + a_1 \Sigma T_i$

$$\Sigma l_i T_i = a_0 \Sigma T_i + a_1 \Sigma T_i^2$$

T_i	l_i	T_i^2	$l_i T_i$
40	600.5	1600	24,020
50	600.6	2500	10030
60	600.8	3600	36048
70	600.9	4900	42063
80	601	6400	48080
300	3003.8	~~180241~~ 19000	~~19000~~ 180241

$$3003.8 = 5a_0 + 300 a_1$$

$$180241 = 300 a_0 + 19000 a_1$$

By solving these eqns we get

$a_0 = 600 \quad a_1 = 0.013$

$\therefore$ eqn is $l = 600 + 0.013 T$

→ If P is the pull required to lift the weight by means of a pulley block, find a linear law of the form $P = a + bw$ connecting P and w, using the following data

w(lb) :	50	70	100	120
P(lb) :	12	15	21	25

Compute P when w = 150 lb.

Sol: Now we have to find $P = a + bw$

normal equations are $\sum P_i = an + b\sum w_i$

$\sum P_i w_i = a\sum w_i + b\sum w_i^2$

w_i	P_i	w_i^2	$P_i w_i$
50	12	2500	600
70	15	4900	1050
100	21	10000	2100
120	25	14400	3000
340	73	31800	6750

$\therefore$ $73 = 4a + 340b$

$6750 = 340a + 31800b$

by solving $a = 2.2785$

$b = 0.1879$

$\therefore$ ~~P = a + bw~~ $P = a + bw$ is

$$\boxed{P = 2.2785 + 0.1879w}$$

If $w = 150$, $P = 2.2785 + 0.1879(150)$

$$\boxed{P = 30.4635 \text{ kg}}$$

→ Find the least squares fit of the form $y = a_0 + a_1 x^2$ to the following data.

x	-1	0	1	2
y	2	5	3	0

Sol: put $x^2 = X$ we have $y = a_0 + a_1 X$

Normal eqns are $\sum y_i = n a_0 + a_1 \sum X_i$

$\sum y_i X_i = a_0 \sum X_i + a_1 \sum X_i^2$

x_i	y_i	$X_i = x_i^2$	X_i^2	$X_i Y_i$
-1	2	1	1	2
0	5	0	0	0
1	3	1	1	3
2	0	4	16	0
	$\sum y_i = 10$	$\sum X_i = 6$	$\sum X_i^2 = 18$	$\sum Y_i X_i = 5$

$10 = 4a_0 + 6a_1$

$5 = 6a_0 + 18a_1$

$30 = 12a_0 + 18a_1$

$5 = 6a_0 + 18a_1$

$(-) \quad (-) \quad (-)$

$25 = 6a_0 \Rightarrow a_0 = 25/6$

⇒ by solving

$a_0 = 4.167, \quad a_1 = -1.111$

∴ ~~Eq of the~~ curve of best fit is

$y = 4.167 + (-1.111) X$

i.e $y = 4.167 - 1.111 x^2$

→ Fit an ~~exponen~~ curve of the form $y = ab^x$ to the following data

x	1	2	3	4	5	6	7	8
y	1.0	1.2	1.8	2.5	3.6	4.7	6.6	9.1

S.No	x_i	y_i	$Y_i = \log y_i$	x_i^2	$Y_i x_i$
1	1	1.0	0.00	1	0
2	2	1.2	0.0792	4	0.1594
3	3	1.8	0.2553	9	0.7659
4	4	2.5	0.3979	16	1.5916
5	5	3.6	0.5563	25	2.7815
6	6	4.7	0.6721	36	4.0326
7	7	6.6	0.8195	49	5.7365
8	8	9.1	0.9590	64	7.6720
	36	30.5	3.7393	204	22.7385

$Y = ab^x \Rightarrow \log y = \log(ab^x) \Rightarrow \log y = \log a + \log b^x$

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

$\Rightarrow Y = A + Bx \Rightarrow$ where $Y = \log y$

$A = \log a$

$B = \log b$

Normal eqns are $\Sigma Y_i = nA + B\Sigma x_i$

$\Sigma Y_i x_i = A\Sigma x_i + B\Sigma x_i^2$

$\Rightarrow 3.7393 = 8A + 36B$

$22.7385 = 36A + 204B$

Solving we get $A = -0.1662$, $B = 0.1408$

$A = \log a \Rightarrow a = 10^A = 10^{(-0.1662)} = 0.6821$,

$B = \log b \Rightarrow b = 10^{B} = 10^{0.1408} = 1.383$

$\therefore$ curve of best fit is

$y = (0.68821)(1.383)^{x}$

$\rightarrow$ An experiment gave the following values

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

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

Sol:

$v = at^{b}$

$\Rightarrow \log v = \log at^{b}$

$\Rightarrow \log v = \log a + \log t^{b}$

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

$Y = A + BX \Rightarrow$ where $Y = \log v$

$A = \log a$

$B = b$

$X = \log t$

v_i (ft/min)	t_i (min)	$X_i = \log t_i$	$Y_i = \log v_i$	X_iY_i	X_i^2
350	61	1.7853 ~~2.5441~~	2.5441	4.542	3.187
400	26	1.4150	2.6021	3.682	2.002
500	7	0.8451	2.6990	2.281	0.714
600	26	0.4150	2.7782	1.153	0.172
		4.4604	10.6234	11.658	6.075

$\Rightarrow$ Normal eqns —
$\Sigma Y_i = nA + B\Sigma X_i$

$\Sigma Y_i X_i = A\Sigma X_i + B\Sigma X_i^2$

$10.623 = 4A + 4.46B$

$11.658 = 4.46A + 6.075B$

By solving we get

$A = 2.845 \quad B = -0.1697$

Since $A = \log_{10} a$ $\qquad B = b$

$\Rightarrow \log_{10} a = 2.845$ $\qquad \Rightarrow b = -0.1697$

$a = 10^{2.845}$

$a = 699.8$

$\therefore$ curve of best fit is $y = (699.8)\,t^{-0.1697}$

$\Rightarrow$ Predict the mean radiation dose at an altitude of 3000 feet by fitting an exponential curve to the given data

Altitude (x):	50	450	780	1200	4400	4800	5300
Dose of radiation (y):	28	30	32	36	51	58	69

Sol: Let $y = ab^x$ be the exponential curve

$\log_{10} y = \log_{10}(ab^x)$

$\log_{10} y = \log_{10} a + \log_{10} b^x$

$\log_{10} y = \log_{10} a + x\log_{10} b$

$Y = A + XB \Rightarrow Y = \log_{10} y, \; A = \log_{10} a, \; X = x, \; B = \log_{10} b$

normal eqs are $\Sigma Y_i = nA + B\Sigma X_i$

$\Sigma Y_i X_i = A\Sigma X_i + B\Sigma X_i^2$

x_i	y_i	$X_i = x_i$	$Y_i = \log_{10} y_i$	$X_i Y_i$	X_i^2
50	28	50	1.447158	72.3579	2500
450	30	450	1.477121	664.7044	202500
780	32	780	1.505150	1174.0170	608400
1200	36	1200	1.556303	1867.5636	1440000
4400	51	4400	1.707570	7513.3080	19360000
4800	58	4800	1.7634 28	8464.4544	23040000
5300	69	5300	1.838849	9745.8997	28090000
16,980		16,980	11.295579	28502.305	72743400

Then $11.295579 = 7A + 16980B$

$28502.305 = 16980A + 72743400B$

By solving $A = 1.4521015$ $B = 0.0000666289$

$A = \log_{10} a$ $B = \log_{10} b$

$\Rightarrow a = 10^A$ $\Rightarrow b = 10^B$

$= 10^{1.4521015}$ $= 10^{(0.0000666289)}$

$=$ $=$

$\therefore$ curve of best fit : $y = (\quad)(\quad)^x$

when $x = 3000$ ft

$y = (\quad)(\quad)^{3000}$

$=$

→ 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 ~~2.98~~	4.26	5.21	6.10	6.80	7.50

$y = 2.978\, x^{0.5143}$, 5.8769

→ Fit a polynomial of second degree to the data points given in the following table

x	0	1	2
y	1	6	17

sol: We have to find eqn of the form $y = a + bx + cx^2$

Normal eqns are $\Sigma y_i = na + b\Sigma x_i + c\Sigma x_i^2$

$\Sigma y_i x_i = a\Sigma x_i + b\Sigma x_i^2 + c\Sigma x_i^3$

$\Sigma y_i x_i^2 = a\Sigma x_i^2 + b\Sigma x_i^3 + c\Sigma x_i^4$

x_i	y_i	x_i^2	x_i^3	x_i^4	$y_i x_i$	$y_i x_i^2$
0	1	0	0	0	0	0
1	6	1	1	1	6	6
2	17	4	8	16	34	68
3	24	5	9	17	40	74

eqns are $24 = 3a + 3b + 5c$

$40 = 3a + 5b + 9c$

$74 = 5a + 9b + 17c$

∴ By solving $a = 1, b = 2, c = 3$

∴ curve is $y = 1 + 2x + 3x^2$

→ Fit a second degree polynomial or parabola of the following

x	0	1	2	3	4
y	1	1.8	1.3	2.5	6.5

Sol: we have to find $y = a + bx + cx^2$

Normal eqns are $\sum y_i = na + b\sum x_i + c\sum x_i^2$

$\sum y_i x_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3$

$\sum y_i x_i^2 = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4$

x_i	y_i	$x_i y_i$	x_i^2	x_i^3	x_i^4	$x_i^2 y$
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	6.5	26	16	64	256	104
10	13.1	37.9	30	100	354	133.5

$\therefore$ $13.1 = 5a + 10b + 30c$

$37.9 = 10a + 30b + 100c$

$133.5 = 30a + 100b + 354c$

$a = 1.43$, $b = -1.14$, $c = 0.58$

$\therefore$ curve is $y = 1.44 - 1.14x + 0.58x^2$

or

$y = 1.44 - 1.14x + 0.58x^2$

→ For 10 randomly selected observations, the following data recorded.

observation no	1	2	3	4	5	6	7	8	9	10
over time (hrs) (x)	1	1	2	2	3	3	4	5	6	7
Additional units (y)	2	7	7	10	8	12	10	14	11	14

Determine the coefficients of regression and regression eqn using the non linear form $y = a + b_1 x + b_2 x^2$

S.NO	x_i	y_i	x_i^2	x_i^3	x_i^4	$y_i x_i$	$y_i x_i^2$
1	1	2	1	1	1	2	2
2	1	7	1	1	1	7	7
3	2	7	4	8	16	14	28
4	2	10	4	8	16	20	40
5	3	8	9	27	81	24	72
6	3	12	9	27	81	36	108
7	4	10	16	64	256	40	160
8	5	14	25	125	625	70	350
9	6	11	36	216	1296	66	396
10	7	14	49	343	2401	98	686
	34	95	154	820	4774	377	1849

$\therefore$ $10a + 34b_1 + 154b_2 = 95$

$34a + 154b_1 + 820b_2 = 377$

$154a + 820b_1 + 4774b_2 = 1849$

By solving $a = 1.80$, $b_1 = 3.48$ $b_2 = -0.27$

$\therefore$ $y = 1.80 + 3.48x - 0.27x^2$

→ Fit a second degree parabola to the following data

x	1989	1990	1991	1992	1993	1994	1995	1996	1997
y	352	356	357	358	360	361	361	360	359

Taking $u = x - 1993$ $\quad v = y - 357$

the eqn $y = a + bx + cx^2$ becomes

$\cancel{v = a + bu}$ $v = A + Bu + Cu^2$

x	$u = x - 1993$	y	$v = y - 357$	uv	u_i^2	u_i^3	u_i^4	$u_i^2 v$
1989	-4	352	-5	20	16	-64	256	-80
1990	-3	360	-1	3	9	-27	81	-9
1991	-2	357	0	0	4	-8	16	0
1992	-1	358	1	-1	1	-1	1	1
1993	0	360	3	0	0	0	0	0
1994	1	361	4	4	1	1	1	4
1995	2	361	4	8	4	8	16	16
1996	3	360	3	9	9	27	81	27
1997	4	359	2	8	16	64	256	32
	$\Sigma u = 0$		$\Sigma v = 11$	51	60	0	708	-9

$\therefore$ normal eqns are

$$\Sigma v_i = 9A + B\Sigma u_i + C\Sigma u_i^2$$

$$\Sigma v_i u_i = A\Sigma u_i + B\Sigma u_i^2 + C\Sigma u_i^3$$

$$\Sigma v_i u_i^2 = A\Sigma u_i^2 + B\Sigma u_i^3 + C\Sigma u_i^4$$

$\therefore 11 = 9A + 60C, \quad 51 = 60B, \quad -9 = 60A + 708C$

$\Rightarrow B = \frac{51}{60} = \frac{17}{20}$

by solving we get $A = \frac{694}{231}$, $B = \frac{17}{20}$; $C = -\frac{247}{924}$

$\therefore \quad v = \frac{694}{231} + \frac{17}{20}u - \frac{247}{924}u^2$

$\therefore \quad y - 357 = \frac{694}{231} + \frac{17}{20}(x - 1993) - \frac{247}{924}(x-1993)^2$

$y = \frac{694}{231} - \frac{33861}{20} - \frac{247}{924}(1993)^2 + \frac{17}{20}x + \frac{247 \times 3866}{924}x$

$- \frac{247}{924}x^2 + 357$

$y = 3 - 1693.05 - 998823.36 + 357 + 0.85x + 1033.44x - 0.267x^2$

$\therefore y = -1000106.41 + 1034.29x - 0.267x^2$

→ Fit a second degree polynomial to the following data

1)

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

2)

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

→ Determine the constants a and b by the method of least squares such that $y = ae^{bx}$ fits the following data

x	2	4	6	8	10
y	4.077	11.084	30.128	81.897	222.62

Sol: we have to find

$\cancel{a = e^{bx}}$

$\cancel{\log a = \log e^{bx}}$

$\cancel{\log a = bx \log e}$

$y = ae^{bx}$

$\log_{10} y = \log_{10} ae^{bx}$

$\log_{10} y = \log_{10} a + \log_{10} e^{bx}$

$\log_{10} y = \log_{10} a + bx \log_{10} e$

$Y = A + BX$ where $Y = \log_{10} y$

$A = \log_{10} a$

$B = b \log_{10} e$

$X = x$

x_i	y_i	$X_i = x_i$	$Y_i = \log_{10} y_i$	X_i^2	$Y_i X_i$
2	4.077	2	0.6103	4	1.2206
4	11.084	4	1.0447	16	4.1788
6	30.128	6	1.479	36	~~6.874~~ 8.874
8	81.897	8	1.9133	64	15.3064
10	222.62	10	2.3476	100	23.476
		30	7.3949	220	~~51.0758~~ 53.0558

Normal eqns ~ $\Sigma Y_i = nA + B\Sigma X_i \Rightarrow 7.3949 = 5A + 30B$

$\Sigma Y_i X_i = A\Sigma X_i + B\Sigma X_i^2 \Rightarrow 53.0558 = 30A + 220B$

~~$A = 0.47302$, $B = 0.16766$~~

$A = 0.17602$, $B = 0.21716$

$A = \log_{10} a \Rightarrow a = 10^{A}$

$\cancel{= 10^{0.47702}}$

$= 10^{0.17602}$

$a = 1.49975$

$B = b \log_{10} e$

$\Rightarrow b = \frac{B}{\log_{10} e}$

$= \frac{0.21716}{0.4343} = 0.5$

$\therefore$ Required curve of best fit is

$y = ae^{bx}$

$= (1.49975) e^{(0.5)x}$

→ Determine the constants a and b by the least square method such that $y = ae^{bx}$ fits the following data

x	1	1.2	1.4	1.6
y	40.17	73.196	133.372	243.02

Sol:

Let $y = ae^{bx}$

$\log_{10} y = \log_{10} (ae^{bx})$

$\log_{10} y = \log_{10} a + \log_{10} e^{bx}$

$\log_{10} y = \log_{10} a + bx \log_{10} e$

$Y = A + BX$ where $Y = \log_{10} y$

$A = \log_{10} a$

$B = b \log_{10} e$

$X = x$

x_i	y_i	$X_i = x_i$	$Y_i = \log_{10} y_i$	X_i^2	$Y_i X_i$
1	40.17	1	1.6039	1	1.6039
1.2	73.196	1.2	1.8645	1.44	2.2374
1.4	133.372	1.4	2.1251	1.96	2.97514
1.6	243.02	1.6	2.3856	2.56	3.81696
		5.2	7.9791	6.96	10.6334

Normal eqns are

$\sum Y_i = nA + B\sum X_i$

$\sum Y_i X_i = A\sum X_i + B\sum X_i^2$

$\Rightarrow 7.9791 = 4A + 5.2B$

$10.6334 = 5.2A + 6.96B$

$A = 0.30107$

$B = 1.30285$

$A = \log_{10} a \Rightarrow a = 10^A = 10^{0.30107}$

$a = 1.9985$

~~A = log~~ $B = b\log_{10} e \Rightarrow b = \frac{B}{\log_{10} e} = \frac{1.30285}{0.4343}$

$= 2.9999$

$\therefore$ curve of best fit is

$y = (1.9985)e^{(2.9999)x}$