Chapter-17

Correlation & Regression

Nikhil shah

Types of Data

- Univariate Data → one variable at a time
- Bivariate Data → Two variables at same time (pairs)

* **Bivariate Data** - When data are collected on two variables simultaneously, they are known as Biveriate data. and the corresponding frequency distribution derived from it, is known as Biveriate frequency distribution

Ex:- 1) Amount of Investment & profit
2) Advertisement Expenditure & sales
3) Demand & price

Demand (X)	Price (Y)
80	20
90	18
100	16
90	17
90	18
110	15

X is changed due to change in Y.
So, X and Y are Correlated.

* Marginal Distribution

Distribution

It is the frequency of one variable (X or Y) across the other variable's full range of values.

No. of Marginal Distribution in Bivariate Data ⇒ 2

* Conditional Distribution

It is the Distribution of one variable (X or Y) across a particular sub-population of the other variable.

No. of Conditional Distribution ⇒ m+n

m = no. of class interval of X

n = no. of class interval of Y

* Correlation - If change in one variable causes change in another variable either Directly or Inversely, then two variables are known to be Associated or correlated.

Types of correlation

positive correlation

If two variables move in the same direction.

Ex:- price & supply

↑ ↑

(upward sloping curve)

Negative correlation

If two variables move in the opposite direction.

Ex:- price & demand

↑ ↓

(Downward sloping curve)

* Measurement of correlation

What to measure

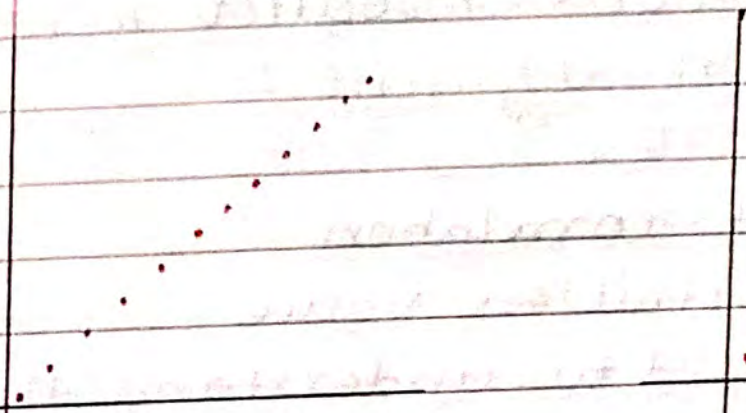
- Nature
- Relative strength (High, low)
- Value

measures of correlation -

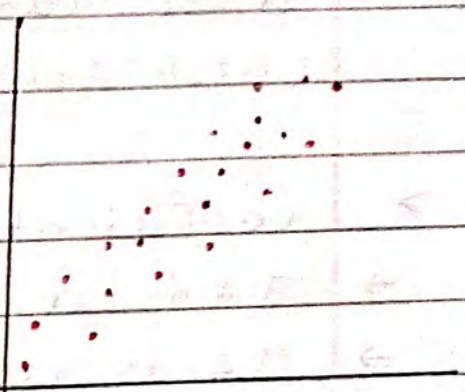
1. → Scatter Diagram
2. → Karl pearson's product moment correlation cofficient
3. → Co-efficient of concurrent deviation
4. → Spareman's Rank correlation Cofficient

1. * Scatter Diagram

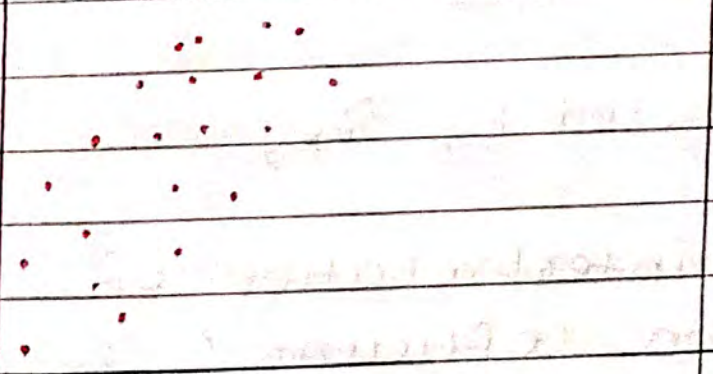
- It helps us to find Nature and Relation Strength of Correlation.
- It is useful for non-linear correlation also
- It cannot be used to determine Value
- Diagrams are time taking



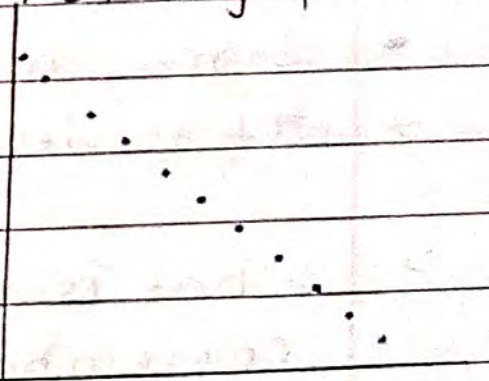
Perfect Positive



Strong positive

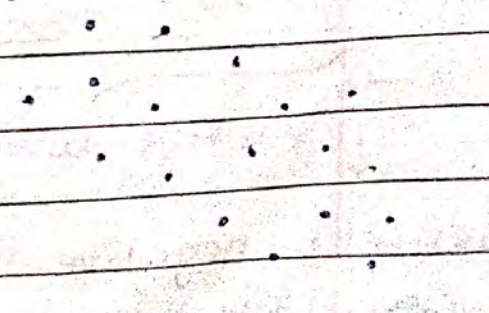


Weak positive



Perfect Negative

Strong Negative



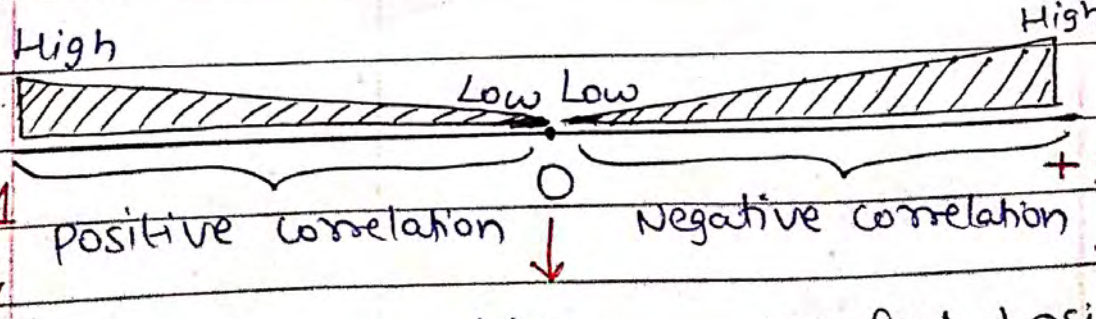
Weak negative

No correlation

2. * Karl Pearson's Product Moment Correlation Cofficient :-

> Cofficient correlation
→ It is a unitless value
→ It is used to understand the nature as well as accurate strength.
→ It is denoted by γ_{xy}

> What is interpretation of correlation cofficient ? γ_{xy}



> How to measure? γ_{xy}

$$\gamma_{xy} = \frac{\text{Covariance of } x \text{ and } y}{\text{SD of } x \times \text{SD of } y}$$

> Covariance coffiicient of x and y (bivariate)

Variance of x $\Rightarrow \frac{\Sigma(x-\bar{x})^2}{n}$ or $\frac{\Sigma x^2}{n} - (\bar{x})^2$

$\Downarrow$ $\Downarrow$

$\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$ or $\frac{\Sigma x.x}{n} - \bar{x}.\bar{x}$

Covariance of x and y $\Rightarrow \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$ or $\frac{\Sigma x.y}{n} - \bar{x}.\bar{y}$

Example - Compute correlation coffiicient b/w x and y from the following data :-

$n = 10$ $\quad \Sigma x^2 = 200$

$\Sigma x = 40$ $\quad \Sigma y^2 = 262$

$\Sigma y = 50$ $\quad \Sigma xy = 220$

Sol^n = $\gamma_{xy} = \frac{Cov(x,y)}{SD_x \cdot SD_y}$

$\bar{x} = \frac{\Sigma x}{n} = \frac{40}{10} = 4$

$\bar{y} = \frac{\Sigma y}{n} = \frac{50}{10} = 5$

$$Cov(x,y) = \frac{220 - 4 \times 5}{10}$$

$$= \frac{200}{10} - 20$$

$$= 2.$$

$$SD \text{ of } x = \sqrt{\frac{200}{10} - \left(\frac{40}{10}\right)^2}$$

$$= \sqrt{20 - 16}$$

$$= \sqrt{4} = \underline{2}$$

$$SD \text{ of } y = \sqrt{\frac{262}{10} - \left(\frac{50}{10}\right)^2}$$

$$= \sqrt{26.2 - 25}$$

$$= \sqrt{1.2} = \underline{1.0954}$$

$$r_{xy} = \frac{Cov(x,y)}{SD \text{ of } x \cdot SD \text{ of } y}$$

$$= \frac{2}{2 \times 1.0954}$$

$$= \underline{0.9128}$$

* Properties of Correlation Coefficient

(1.)

	Change of origin	Change of scale (value)	Change of scale (sign)
Central tendency	✓	✓	✓
Measure of Dispersion	✗	✓	✗
Correlation Coefficient	✗	✗	✓

Original → modified

$x \xrightarrow{\text{origin/scale}} u$

$Y \xrightarrow{\text{origin/scale}} V$

$\pm r_{xy} = r_{uv}$

* If sign of both change of scale are same. $r_{xy} = r_{uv}$

* If sign of both change of scale are different $r_{uv} = -r_{xy}$

Ex:- Given:- $r_{xy} = 0.65$

$u = 5 - 2x$ $\quad$ $v = 10 + 3y$

Change of scale of x = -ive

Change of scale of y = +ive

So,

$r_{uv} = -r_{xy} = -0.65$

(2.) The coefficient of correlation is a unit-free measure.

(3.) Values lies from -1 to $+1$

3. * SPEARMAN'S RANK CORRELATION COFFICIENT

To find Correlation Cofficient in Case of Qualitative data : as Ranks

$$r_R = 1 - \frac{6 \Sigma d^2}{n(n^2-1)}$$

Ex :-

No. of Candidats =	1	2	3	4	5	6	7	8	9	10
Rank by Judge A =	10	5	6	1	2	3	4	7	9	8
Rank by Judge B =	5	6	9	2	8	7	3	4	10	1
Difference =	5	1	3	1	6	4	1	3	1	7
Square of difference $(d)^2$ =	25,	1,	9,	1,	36,	16,	1,	9,	1,	49

$\Sigma d^2 = 148$

$$r_R = 1 - \frac{6 \times 148}{10(10^2-1)} \quad [n = 10]$$

$$r_R = 1 - 0.897$$

$$\boxed{r_R = 0.1030}$$

> In case of Tie in values
- Average Rank to be given in values
- $$r_R = 1 - \frac{6(\Sigma d^2 + \text{adj. value})}{n(n^2-1)}$$

$$\text{adj value} = \sum \frac{t^3 - t}{12} \quad [t = \text{tie length}]$$

Ex:-

Eco marks X	80	56	50	48	50	62	60
Stats marks Y	90	75	75	65	65	50	65
Rank Rx	1	4	5.5	7	5.5	2	3
Ry	1	2.5	2.5	5	5	7	5
d^2	0	2.25	9	4	0.25	25	4

$\Sigma d^2 = 44.5$

Adj value $\Rightarrow \sum \frac{t^3-t}{12} \Rightarrow$

	t	$\frac{t^3-t}{12}$
50-50	2	$\frac{2^3-2}{12}$
75-75	2	$\frac{2^3-2}{12}$
65-65-65	3	$\frac{3^3-3}{12}$

$$\frac{\Sigma t^3 - t}{12} = 3$$

$$r_R = 1 - \frac{6[44.5+3]}{7(48)}$$

$$r_R = 1 - 0.849$$

$$r_R = 0.151$$

* 4. Co-efficient of Concurrent deviations

> Very quick, simple and casual method of finding correlation when we are not serious about the magnitude of two variables.

$$r_c = \pm \sqrt{\pm \frac{2c - m}{m}}$$

m = no. of pairs compared (n-1)
c = no. of concurrent deviations (no. of plus signs)

Ex:-	1990	1991	1992	1993	1994	1995	1996	1997
Price	25	28	30	23	35	38	39	42
Demand	35	34	35	30	29	28	26	43
dx	na	+	+	−	+	+	+	+
dy	na	−	+	−	−	−	−	+
dxdy		−	(+)	(+)	−	−	−	(+)

m = n − 1 = 8 − 1 = 7

c = no. of plus sign in dxdy = 3

$$r_c = \pm \sqrt{\pm \left(\frac{2(3) - 7}{7}\right)}$$

$$r_c = \pm \sqrt{\pm \left(\frac{-1}{7}\right)}$$

$$r_c = -\sqrt{-\left(\frac{-1}{7}\right)}$$

r_c = ~~-0.37~~ -0.3779

* एक जगह $-$ तो दूसरी जगह भी $-$ और अगर एक जगह $+$ तो दूसरी जगह भी $+$

$\frac{2c-m}{m}$ = +ive = + both place

$\frac{2c-m}{m}$ = -ive = - both place

* Regression

Estimation of dependent variable with the given independent variable.

- Estimation of Y when X is given
  Y: Dependent, X: Independent

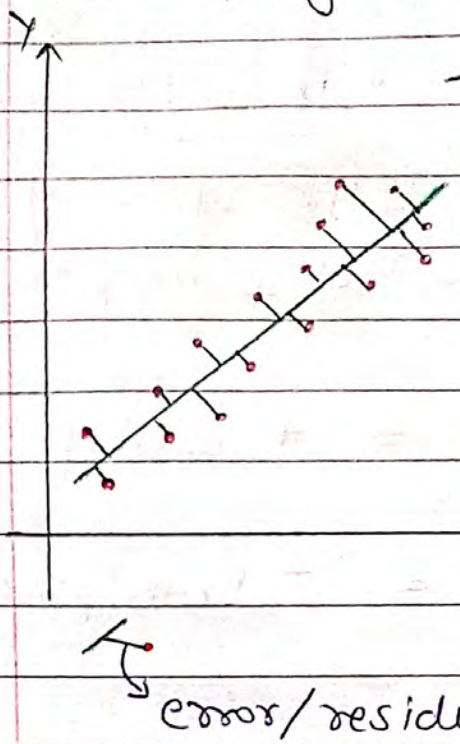
- Estimation of X, when Y is given
  X: Dependent, Y: Independent

1) Perfect Correlation

- When linear eqn exist between two variables, correlation is perfect.

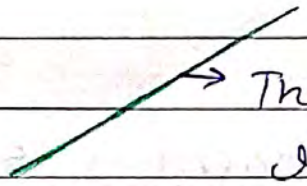
- perfect correlation is represented by linear equation and this eqn can be used for regression purpose directly.

2) Not perfect Correlation

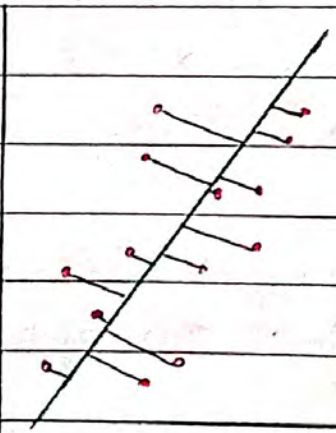


- In case of non perfect correlation point are not in line so we don't have available lineas eqn.

then, in that case we will make a line for purpose of estimation.



These lines are called as regression lines.



Reg. line of X on Y used for estimation of X when Y is given

Reg line of Y on X used for est. of Y when X is given

"Method of Least Square"

Page No.
Date

Objectives

Estimation of Y when X is given

Equation format:

$Y - \bar{Y} = b_{yx}(X - \bar{X})$

$Y = a + bx$

b_{yx} = reg. coff. of Y on X

Estimation of X when Y is given

Equation format:

$X - \bar{X} = b_{xy}(Y - \bar{Y})$

$X = a + by$

b_{xy} = reg. coff. of X on Y

Regression Coefficient = Covariance b/w two variables / Variance of Independent Variable

$$b_{yx} = \frac{Cov(x,y)}{\text{variance of } x}$$

$$b_{xy} = \frac{Cov(x,y)}{\text{variance of } y}$$

Ex^n Find two regression line/eq^n from:

x:	2	4	5	5	8	10
y:	6	7	9	10	12	12

Hence estimate Y when X is 13 and estimate X when Y is 15.

Sol^n → Est. of y on x

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

$y - \frac{56}{6} = b_{yx}\left(x - \frac{34}{6}\right)$

$y - 9.3333 = 0.8145(x - 5.6666)$

$y = 0.8145x - 4.6155 + 9.3333$

$\boxed{y = 0.8145x + 4.7178}$

↳ Regression Eq^n of y on x.

If $x = 13$

$y = 4.7178 + 13 \times (0.8145)$

$\boxed{y = 15.3063}$

$b_{yx} = \frac{\text{Cov}(x,y)}{\text{Var. } x}$

$b_{yx} = \frac{\frac{\Sigma xy}{n} - \bar{x}\cdot\bar{y}}{\frac{\Sigma x^2}{n} - (\bar{x})^2}$

$b_{yx} = \frac{351}{6} - \left(\frac{34}{6} \times \frac{56}{6}\right)$

$\frac{234}{6} - \left(\frac{34}{6}\right)^2$

$b_{yx} = \frac{5.6111}{6.888} = 0.8145$

Estimation of x on y

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

$x - \frac{34}{6} = 1.0745\left(y - \frac{56}{6}\right)$

$x - 5.6666 = 1.0745(y) - (10.086)$

$\boxed{x = 1.0745y - 4.362}$

↳ Reg. eqn of x on y

$b_{xy} = \frac{\text{Cov}(x,y)}{\text{Var. of } y}$

$= \frac{5.6111}{\frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2}$

$= \frac{5.6111}{\left(\frac{554}{6}\right) - \left(\frac{56}{6}\right)^2}$

$= 1.0745$

If ⇒ $y = 15$

$x = 1.0745(15) - 4.362$

$\boxed{x = 11.7555}$

** Properties of Regression lines / Coefficient

> Change of Origin / Scale

Origin : GB (no impact)

Scale : a) value ⟶ Yes
b) sign ⟶ Yes

$x \longrightarrow u$

$y \longrightarrow v$

$$b_{uv} = b_{xy} \times \frac{\text{Change of Scale of } x}{\text{Change of Scale of } y}$$

$$b_{xu} = b_{yx} \times \frac{\text{Change of Scale of } Y}{\text{Change of Scale of } X}$$

Ex~ $u + 3x = 10$

$u = 10 - 3x$

Change of Scale of $x = -3$

$b_{yx} = 0.80$

$b_{uv} = ?$

$2y + 5v = 25$

$5v = 25 - 2y$

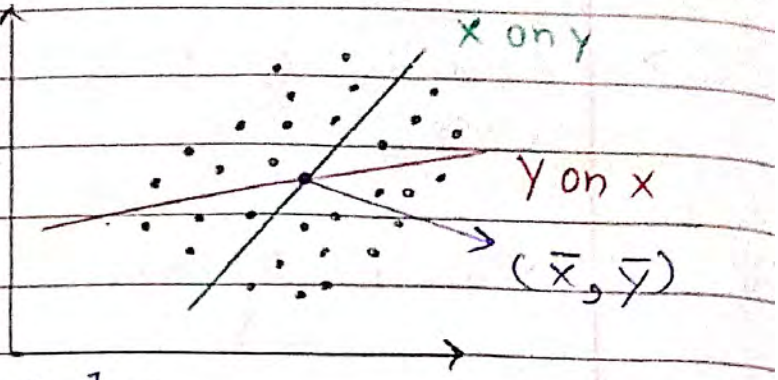
$v = 5 - \frac{2}{5}y$

Change of Scale of $Y = -2/5$

$b_{uv} = 0.80 \times \frac{-2/5}{-3} = \frac{0.80 \times 2}{5 \times 15}$

$b_{uv} = \frac{8}{75} = \underline{\underline{0.1066}}$

> Two lines of regression intersect at $(\bar{x}, \bar{y})$. If lines are not identical.



Ex- $7x - 3y - 18 = 0$]x1

$4x - y - 11 = 0$]x3

$7x - 3y - 18 = 0$	$7x(3) - 3y - 18 = 0$
$12x - 3y - 33 = 0$	$21 - 3y - 18 = 0$
$- \quad + \quad +$	$3 = 3y$
$-5x + 15 = 0$	$\boxed{y = 1}$
$5x = 15$	
$\boxed{x = 3}$	$\bar{x} = 3, \bar{y} = 1$

> Relation b/w Correlation and Regression Coefficient →

$$\boxed{r_{xy} = \pm \sqrt{b_{xy} \times b_{yx}}}$$

Ex- $b_{xy} = -1.06 \quad b_{yx} = -0.73$

$r_{xy} = \sqrt{-1.06 \times -0.73} = 0.879$ ✗

$r_{xy} = \sqrt{-(1.06) \times (-0.73)} = -0.879$

$$\rho_{xy} = \frac{Cov(x,y)}{SD_x\ SD_y} \Rightarrow Cov(x,y) = \rho_{xy} \cdot SD_x \cdot SD_y$$

$$b_{xy} = \frac{Cov(x,y)}{\text{Var. of } y} \qquad b_{yx} = \frac{Cov(x,y)}{\text{Var. of } x}$$

$$b_{xy} = \frac{\rho_{xy} \cdot SD_x \cdot SD_y}{(SD_y)^2} \qquad b_{yx} = \frac{\rho_{xy} \cdot SD_x\ SD_y}{(SD_x)^2}$$

$$b_{xy} = \rho_{xy}\left(\frac{SD_x}{SD_y}\right) \qquad b_{yx} = \rho_{xy}\left(\frac{SD_y}{SD_x}\right)$$

* you are given two regression lines/ eqⁿ and you have to find out which one is [Y on x] and [X on Y.]

Ex- $7x - 3y - 18 = 0$ & $4x - y - 11 = 0$

Coefficient of x / Coefficient of y (Ignore negative Sign)

Coff. of x / Coff. of y (Ignore negative Sign)

$\frac{7}{3} = 2.333$ < $\frac{4}{1} = 4$

So, ↓

Y on X

↓

X on Y

जो बड़ा आएगा वो X on Y and दूसरा वाला Y on X.

Page No.
Date

* **probable error** $= 0.6745 \times \frac{1-r^2}{\sqrt{n}}$

Ex → $r = 0.7$

$n = 64$

probable error $= 0.6745 \times \frac{1-(0.7)^2}{\sqrt{64}}$

$= 0.6745 \times \frac{0.51}{8}$

$= \boxed{0.04299}$

* limits of population correlation coffiricent :-

lower limit	UPPer limit
r + probable error	r − probable error

Ex - $r = 0.7$

$PE = 0.0423$

limit of population correlation coffiricent :-

$\Rightarrow 0.7 + 0.0423$ | $\Rightarrow 0.7 - 0.0423$

$\Rightarrow 0.7423$ | $\Rightarrow 0.6577$

* Coff. of determination / Coff. of non-determination

r^2 / $1 - r^2$

Ex- Corr. Coff. b/w demand (X) and price (Y) is -0.8.

⇒ Demand is change due to price but there are other reasons also.

Reason of Change of demand

Price	Other factors
$(-0.8)^2$	$1 - (-0.8)^2$
$= 0.64$	$= 0.36$
64% Change due to price	36% Change due to other factors
Also known as Explained variance/ Accounting variance	Also known as unexplained variance/ unaccounting variance

* Where no casual (reason) relationship So, it is called "Non-sense correlation" or "Spurious correlation."

* Karl pearson correlation coff. is best.