SIMPLE LINEAR REGRESSION

Two main Objectives,

↳ Establish if there is a RELATIONSHIP b/w two variables

Examples: Income & spending (+ve) [income badhegi to spending]
wage & gender [man earns higher wage than women] -ve
Student height & exam score [no relation]

↳ FORECAST new observations

Can we use what we know about the relationship to forecast unobserved values.

Example: What will our sales be over next quarter?

Dependent Variable: This is the variable whose value we want to explain or forecast

Denoted by : Y

Independent Variable: This is the variable that explains the other one.

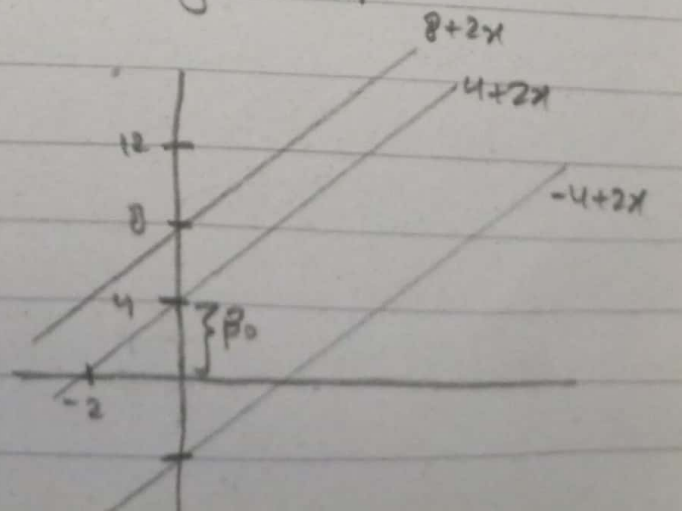
Denoted by X

$$Y = \beta_0 + \beta_1 X$$

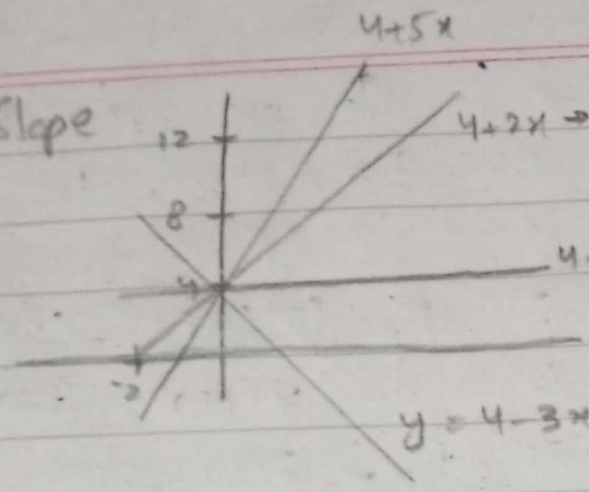
intercept ↳ slope

$$Y = 4 + 2x$$

• Changing intercept



• Changing Slope

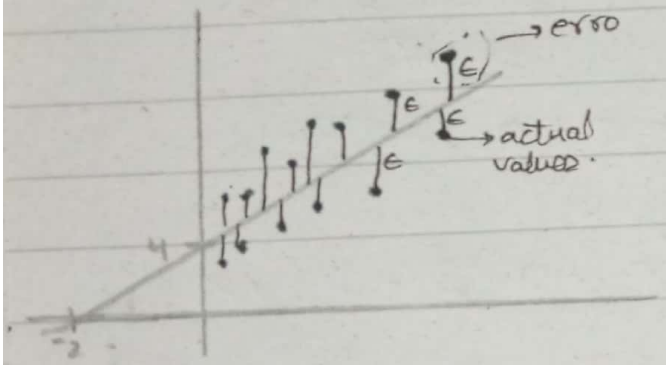


$4+2x$ → means of every one unit of x y changes 2 times

$4+0x$ → no matter what value of x, y remains 4

$y = 4-3x$ → y decreases 3 times for every unit of x

But in real world data is not in a straight lines.



$$Y = \beta_0 + \beta_1 x + \epsilon$$

error term.

Formulas for Finding β_0 & β_1:

$$\beta_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x^2)^2}$$

$$\beta_0 = \frac{\Sigma y}{n} - \frac{\beta_1 (\Sigma x)}{}$$

Q: Write a linear regression equation that "best fits" the data below

X	1	2	3	4	5	6	7
Y	1.5	3.2	6.7	9.0	11.2	13.6	16

x	y	xy	x^2
1	1.5	1.5	1
2	3.8	7.6	4
3	6.7	20.1	9
4	9.0	36	16
5	11.2	56	25
6	13.6	81.6	36
7	16	112	49
Σ 28	61.8	314.8	140

$\Sigma x = 28$, $\Sigma y = 61.8$

$\Sigma xy = 314.8$, $\Sigma x^2 = 140$

$$\beta_1 = \frac{(7)(314.8) - (28)(61.8)}{7(140) - \cancel{(140)}(28)^2}$$

$$\boxed{\beta_1 = 2.4142857}$$

$$\beta_0 = \frac{(61.8) - \beta_1(28)}{7}$$

$$\boxed{\beta_0 = -0.828571}$$

$$Y = -0.83 + 2.414x$$

"CORRELATION"

It is a statistical method used to determine whether a relationship b/w variables exist.

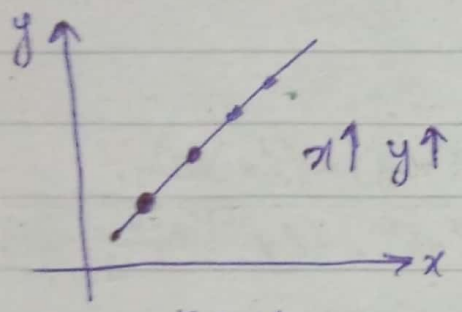
→ The independent & dependant variables can be plotted in a graph called scatter plot.

→ To determine strength of relationship b/w two variables we use coefficent of co-relation. (r)

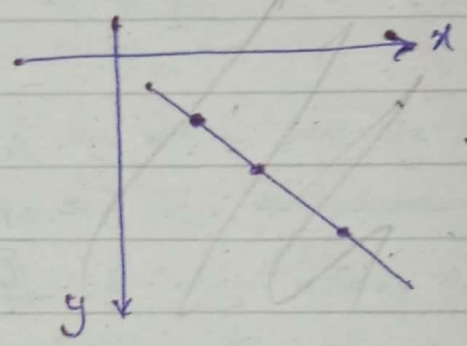
↓ Sample correl

P (population correlation)

+ve Correlation :



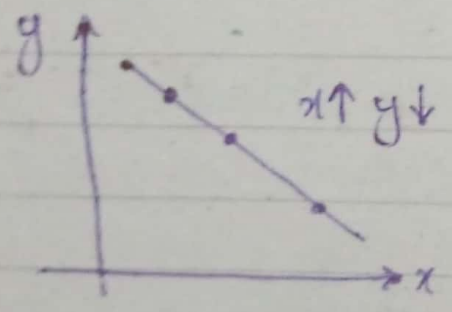
$r = +1$



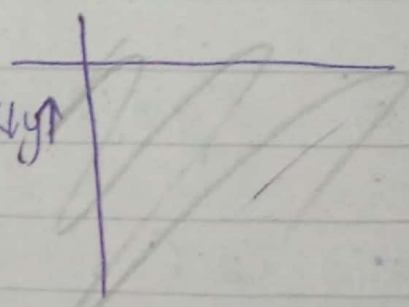
x↓ y↓

Eg. example : height & weight.

-ve Correlation :

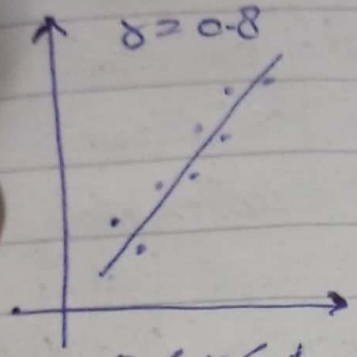


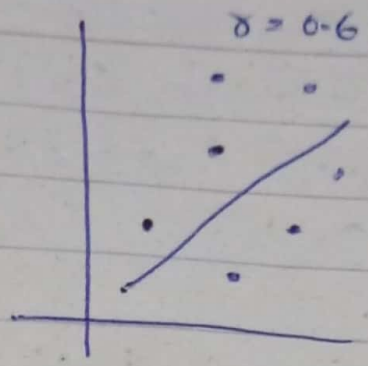
$r = -1$



Example : Price↑ Demand↓

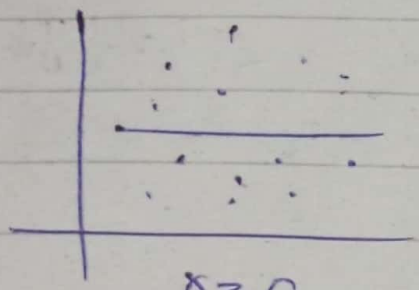
$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}$$





$0 < r \leq 1$

→ jitna line k qareeb honge points utni r ki value 1 k qareeb hogi.



$r = 0$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

Calculate coefficient of correlation:

X	1	2	3	4	5	6
Y	2	4	7	9	12	14

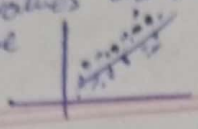
$\Sigma x = 21$, $\Sigma y = 48$ $\Sigma xy = 211$ $\Sigma x^2 = 91$, $\Sigma y^2 = 4$

$$r = \frac{6(211) - (21)(48)}{\sqrt{[(6)(91) - (21)^2][6(490) - (48)^2]}}$$

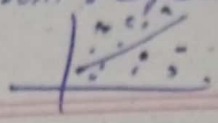
$r = 0.998$ as it is +ve means $x \uparrow y \uparrow$

Coefficent of Determination

if its value is close to 1 means the true values are close to regression line



If close to 0 the true values far from regression line



correlation is symmetric $r_{xy} = r_{yx}$

correlation doesnot depends on unit

It value lies b/w $-1 \leq r \leq +1$

INFERENCE IN CORRELATION

→ Ye check karne k liye ka waqai do variable apas mein correlated hai we perform hypothesis testing on population correlation (ρ).

- $H_0 : \rho = 0$ → This null hypothesis means that there is no correlation b/w x & y variables

$H_1 : \rho \neq 0$ → This alternative hyp means that there is a significant correlation b/w var in popul

$\rho > 0$ → variables are +vely linearly correlated. $x \uparrow y \uparrow$

$\rho < 0$ → " " -vely " " $x \downarrow y \uparrow$ (vice versa)

→ The r is an estimate of ρ.

t - Distribution for Correlation test:

Suppose that variables x & y satisfy the four assumptions for regression inferences & that $\rho = 0$. Then for sample size n, the variable,

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$ has t-dist with $df = n-2$

PROCEDURE:

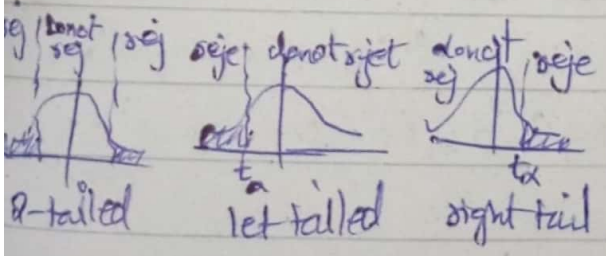
- purpose: To perform a hypothesis test for a population linear correlation coefficient, ρ.
- Assumption: The four assumption for regression inferences
- **Step# 01** The null hypothesis is $H_0: \rho = 0$ & alternative hyp is

  $H_1: \rho \neq 0$ (two-tailed test) or $H_1: \rho < 0$ (left tailed) or $H_1: \rho > 0$ (right tailed)

- **Step# 02** Decide on significance level, α
- **Step# 03** Compute $t_0 = r / \sqrt{(1-r^2)/(n-2)}$

CRITICAL VALUE APP.

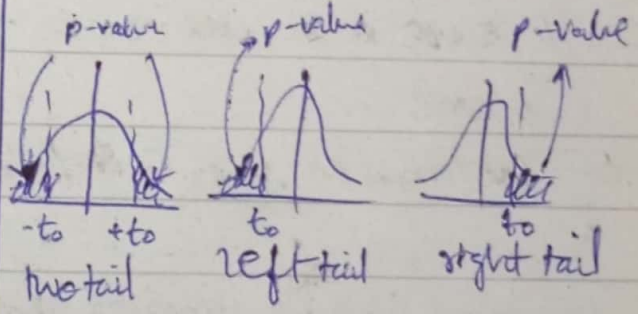
- **Step# 04** The critical values are $\pm t_{\alpha/2}$ (two tailed) or $-t_\alpha$ (left) or t_α (right) with $df = n-2$. Use table IV to find critical values.



- **Step# 05** If t_0 is in reject region, reject H_0.

P-VALUE APP.

- **Step# 04** The t-stat has $df = n-2$. Use table to estimate P-value.



Step# 05 If $p \leq \alpha$ reject H_0.

Ex: 11.1

Q11.1 $\Sigma x_2 = 26{,}591.63$, $\Sigma xy = 65{,}164.04$

$\Sigma x = 778.7$ $\Sigma y = 2050.0$

$n = 25$

$$Y = \beta_0 + \beta_1 x$$

$$\beta_1 = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{25(65{,}164.04) - 778.7 \times 2050}{25(26591.63) - (778.7)^2}$$

$$\boxed{\beta_1 = 0.5609}$$

$$\beta_0 = \frac{\Sigma y - \beta_1 \Sigma x}{n} = \frac{2050 - \beta_1 (778.7)}{25}$$

$$\boxed{\beta_0 = 64.529}$$

$$Y = 64.529 + 0.5609x$$

(b) at $x = 30$

$$Y = 64.529 + 0.5609(30)$$

$$\boxed{Y = 81.35615}$$

Q11.2 $\Sigma x = 707$ $\Sigma y = 658$ $\Sigma xy = 53{,}258$

$\Sigma x^2 = 57{,}557$ $n = 9$

Q11.5: $\Sigma x = 16.5$ $\Sigma y = 100.4$ $\Sigma xy = 152.59$

$\Sigma x^2 = 25.85$

$$\beta_1 = \frac{n\Sigma xy - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{11(152.59) - 16.5 \times 100.4}{11(25.85) - 16.5^2}$$

$\beta_1 = 1.809$

$$\beta_0 = \frac{\Sigma y - \beta_1 \Sigma x}{n} = 6.4137$$

$$\boxed{Y = 6.4137 + 1.809x}$$

at $x = 1.75$

$$\boxed{Y = 9.580}$$

Q4.43 $$r = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$\Sigma x = 464$ $\Sigma y = 476$ $n = 6$

$\Sigma xy = 36926$ $\Sigma x^2 = 36354$ $\Sigma y^2 = 38257$

$r = 0.2397$