ALL BRANCHES

ME,CE,EC,EE,CS

Probability and Statistics

Lecture No- 06

By- Vinay sir

Topics to be Covered

01 Exponential Distribution

02 Linear Regression

03 Correlation Coefficient

04 Joint Probability distributions

Revision

(Successes are Time Independent)

Discrete → Binomial distribution → Bernoulli Trials

Discrete → Poisson distribution.

(Success are Time dependent)

When 'n' is large, Binomial distribution approaches to Poisson distribution.

$$P(X=r) = {}^{n}C_{r} \cdot p^{r} \cdot q^{n-r};$$

Binomial

$$P(x=r) = \frac{e^{-\lambda} \cdot \lambda^{r}}{r!}$$

Poisson

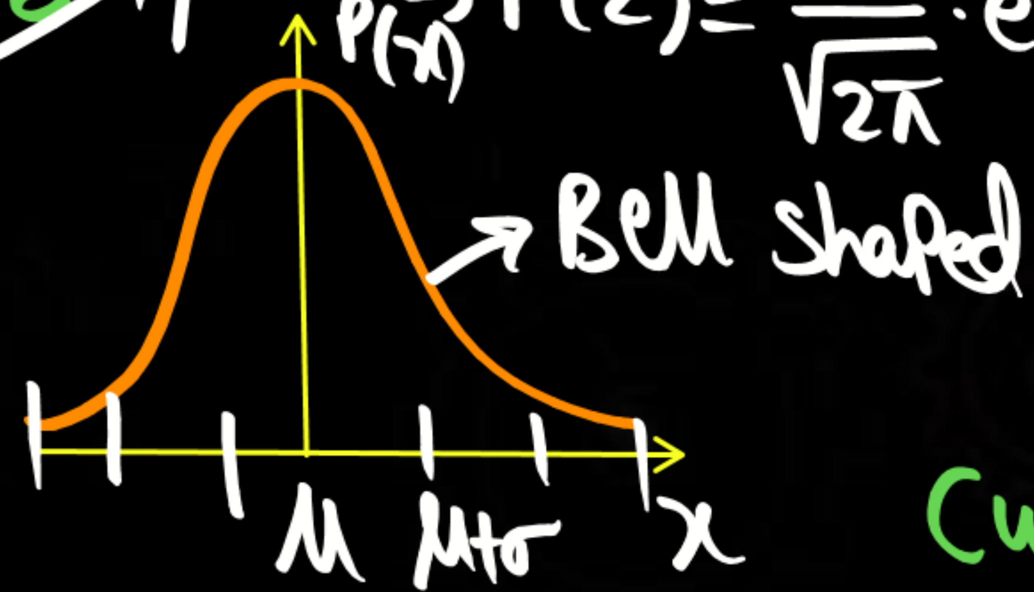
Binomial: Mean $= np$; Variance $= npq$; $\sigma = \sqrt{npq}$.

Poisson: Mean $=$ Variance $= \lambda$; $\sigma = \sqrt{\lambda}$.

Normal/Gaussian distribution: $P(x) = \dfrac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

$$z = \frac{x - \cancel{\mu}^{\,0}}{\cancel{\sigma}^{\,1}}$$

$P(x) \Rightarrow P(z) = \dfrac{1}{\sqrt{2\pi}} \cdot e^{-z^2/2}$



$P(-1 \le z \le 1) = 0.68$

$P(-2 \le z \le 2) = 0.95$

$P(-3 \le z \le 3) = 0.997$

Cumulative function is 'S' shaped.

Exponential Distribution

The Probability density function of a Continuous Random Variable that is exponentially distributed is given as

$$P(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & ; x \geq 0 \\ 0 & ; x < 0 \end{cases}$$

Where Mean of the Variable $= \frac{1}{\lambda}$.

$$\text{Mean} = \int_{-\infty}^{\infty} x \cdot P(x)\, dx$$

$$= \int_{-\infty}^{0} x \cdot (0)\, dx + \int_{0}^{\infty} x \cdot (\lambda \cdot e^{-\lambda x})\, dx$$

$$= \int_{0}^{\infty} (\lambda x) \cdot (e^{-\lambda x})\, dx$$

let $\lambda x = t$

$\Rightarrow \lambda \cdot dx = dt \Rightarrow dx = \frac{dt}{\lambda}$

<u>L.L:</u> $\lambda x = t \Rightarrow \lambda(0) = t \Rightarrow t = 0$

<u>U.L:</u> $\lambda x = t \Rightarrow \lambda(\infty) = t \Rightarrow t \to \infty$

$$\Rightarrow \text{Mean} = \int_{0}^{\infty} t \cdot e^{-t} \cdot \frac{dt}{\lambda}$$

$$= \frac{1}{\lambda} \cdot \int_{0}^{\infty} e^{-t} \cdot t^{2-1} \cdot dt$$

$$= \frac{1}{\lambda} \cdot \Gamma 2 = \frac{1}{\lambda} \cdot 1! = \frac{1}{\lambda}$$

$$\boxed{\therefore \text{Mean} = \frac{1}{\lambda}}$$

→ Variance: σ^2.

$$\sigma^2 = E(x^2) - (E(x))^2.$$

$$E(x^2) = \int_{-\infty}^{\infty} x^2 \cdot P(x)\,dx$$

$$= \int_0^{\infty} x^2 \cdot \lambda \cdot e^{-\lambda x}\,dx$$

let $\lambda x = t \Rightarrow x = \frac{t}{\lambda} \Rightarrow dx = \frac{dt}{\lambda}$.

L.L: $\lambda x = t \Rightarrow \lambda(0) = t \Rightarrow t = 0$

U.L: $\lambda x = t \Rightarrow \lambda(\infty) = t \Rightarrow t \to \infty$

$$\therefore E(x^2) = \int_0^{\infty} \frac{t^2}{\lambda^2} \cdot \cancel{\lambda} \cdot e^{-t} \cdot \frac{dt}{\cancel{\lambda}}$$

$$= \frac{1}{\lambda^2} \int_0^{\infty} e^{-t} \cdot t^2\,dt$$

$$= \frac{1}{\lambda^2} \int_0^{\infty} e^{-\lambda} \cdot t^{3-1}\,dt = \frac{1}{\lambda^2} \cdot \Gamma 3$$

$$= \frac{2!}{\lambda^2} = \frac{2}{\lambda^2}.$$

$$\therefore E(x^2) = \frac{2}{\lambda^2}.$$

$$\therefore \sigma^2 = E(x^2) - (E(x))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2}$$

$$\boxed{\Rightarrow \sigma^2 = \frac{1}{\lambda^2}}$$

$\therefore$ For a exponentially distributed Variable;

$$\text{Mean} = \text{Standard deviation} = \frac{1}{\lambda}.$$

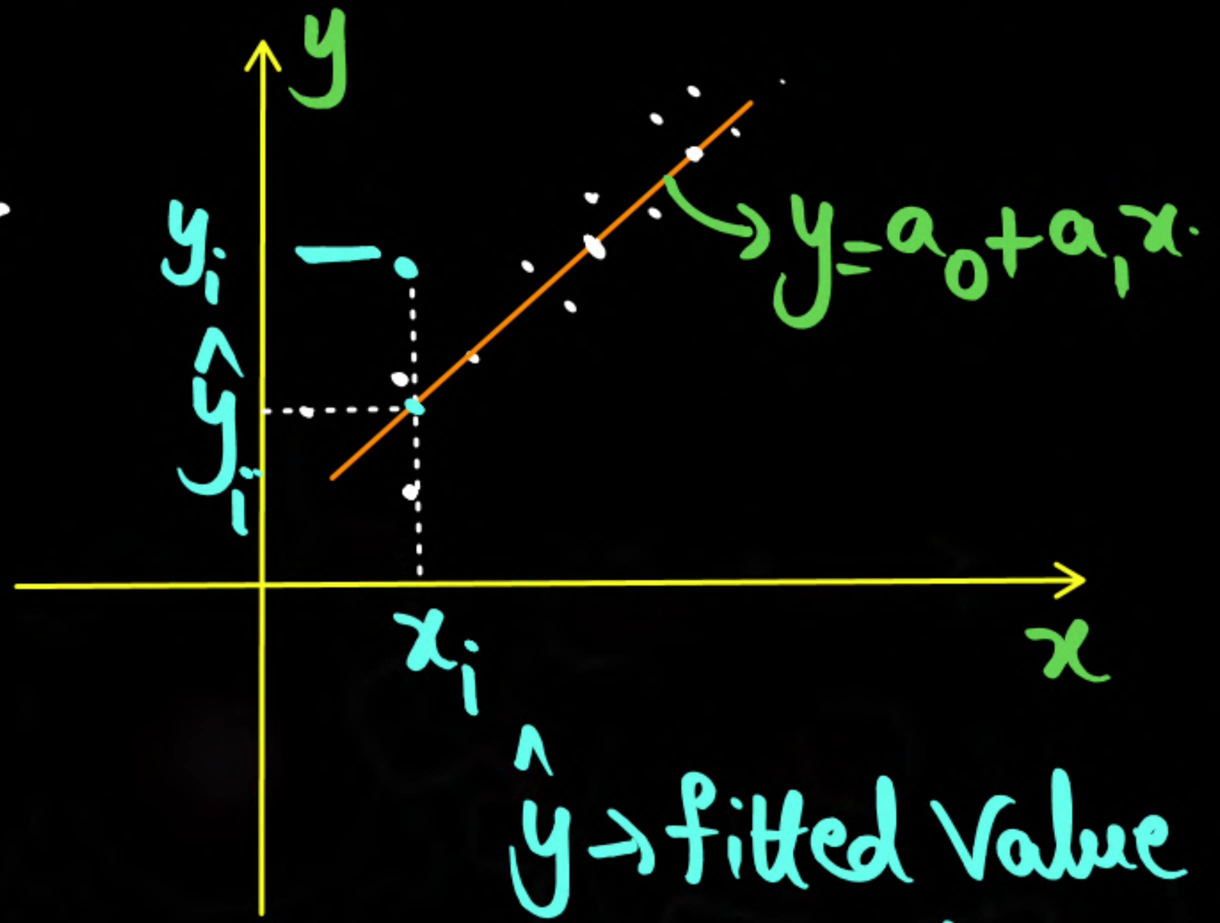
Statistical Attributes: The Parameters that governs the distributio

Distribution	Statistical Attributes	
1. Binomial	n, p	$\longrightarrow 2$
2. Poisson	λ	$\longrightarrow 1$
3. Gaussian/Normal	μ, σ	$\longrightarrow 2$
4. Exponential	λ	$\longrightarrow 1$

Linear Regression

Regression → fitting of a curve b/w a set of Variables.

Let the Data Points be $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots (x_N, y_N)$.



$y_i \to$ Actual Data Point Value.

$\hat{y}_i \to$ fitted Value.

$\hat{y} \to$ fitted Value at x_i

$\Rightarrow \hat{y}_i = a_0 + a_1 x_i$

Error $= e_i = y_i - \hat{y}_i = y_i - (a_0 + a_1 x_i)$

Sum of squares of errors: $\sum_{i=1}^{N} e_i^2$.

$\therefore$ For the fit be a good fit, Sum of squares of errors should be least possible.

$\Rightarrow \sum_{i=1}^{N} e_i^2$ should be least. (Least square Regression) (LSR).

$\Rightarrow \sum_{i=1}^{N} (y_i - a_0 - a_1 x_i)^2 \longrightarrow$ Should be least

$$\Rightarrow f(a_0, a_1) = \sum_{i=1}^{N} (y_i - a_0 - a_1 x_i)^2$$

$$f(a_0, a_1) = \sum_{i=1}^{N} \left(y_i^2 + a_0^2 + a_1^2 . x_i^2 - 2a_0 y_i + 2a_0 a_1 x_i - 2y_i a_1 x_i \right)$$

For $f(a_0, a_1)$ to be minimum,

$$\frac{\partial f}{\partial a_0} = 0 \quad \text{and} \quad \frac{\partial f}{\partial a_1} = 0$$

$$\frac{\partial f}{\partial a_0} = 0 \Rightarrow \sum_{i=1}^{N} \left(\cancel{2}a_0 - \cancel{2}y_i + \cancel{2}a_1 x_i \right) = 0 \Rightarrow \sum_{i=1}^{N} a_0 - \sum_{i=1}^{N} y_i + a_1 . \sum_{i=1}^{N} x_i = 0$$

$$\frac{\partial f}{\partial a_1} = 0 \Rightarrow \sum_{i=1}^{N} \left(\cancel{2}a_1 x_i^2 + \cancel{2}a_0 x_i - \cancel{2}x_i y_i \right) = 0 \Rightarrow \sum_{i=1}^{N} a_1 x_i^2 + \sum_{i=1}^{N} a_0 . x_i - \sum_{i=1}^{N} x_i y_i = 0$$

$$\Rightarrow a_0 \cdot N + a_1 \cdot \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \longrightarrow (1)$$

$$\sum_{i=1}^{N} 1 = 1+1+ \cdots +1 = \underline{N}$$

$$a_0 \cdot \sum_{i=1}^{N} x_i + a_1 \cdot \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i \longrightarrow (2)$$

Solving above two Equations.

$$(1) \times \sum_{i=1}^{N} x_i - (2) \times N$$

$$\Rightarrow a_1 \left\{ \left(\sum_{i=1}^{N} x_i \right)^2 - N \cdot \sum_{i=1}^{N} x_i^2 \right\} = \left(\sum_{i=1}^{N} x_i \right) \left(\sum_{i=1}^{N} y_i \right) - N \cdot \sum_{i=1}^{N} x_i y_i$$

$$\Rightarrow a_1 = \frac{N \cdot \left(\sum_{i=1}^{N} x_i y_i\right) - \left(\sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i\right)}{N \cdot \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}$$

Since $y = a_0 + a_1 x \rightarrow a_1$ is slope of the line.

$$\boxed{\therefore a_1 = \frac{N \cdot \sum xy - (\sum x)(\sum y)}{N \cdot \sum x^2 - (\sum x)^2}}$$

Substituting a_1 in ①

$$\Rightarrow a_0 \cdot N + \left\{ \frac{N \cdot \sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{N \cdot \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} \right\} \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i$$

$$\Rightarrow a_0 \cdot N + \frac{N \cdot \sum_{i=1}^{N} x_i y_i \cdot \sum_{i=1}^{N} x_i - \left(\sum_{i=1}^{N} x_i\right)^2 \left(\sum_{i=1}^{N} y_i\right)}{N \cdot \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \sum_{i=1}^{N} y_i$$

$$\Rightarrow N^{\cancel{2}} \cdot a_0 \cdot \sum_{i=1}^{N} x_i^2 - \cancel{N} \cdot a_0 \left(\sum_{i=1}^{N} x_i\right)^2 + \cancel{N} \cdot \sum_{i=1}^{N} x_i y_i \cdot \sum_{i=1}^{N} x_i - \cancel{\sum_{i=1}^{N} y_i \left(\sum_{i=1}^{N} x_i\right)^2}$$

$$= \cancel{N} \cdot \sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i - \cancel{\left(\sum_{i=1}^{N} x_i\right)^2 \left(\sum_{i=1}^{N} y_i\right)}$$

$$\Rightarrow a_0\left(N\cdot\sum_{i=1}^{N} x_i^2-\left(\sum_{i=1}^{N} x_i\right)^2\right)=\sum_{i=1}^{N} x_i^2\cdot\sum_{i=1}^{N} y_i-\sum_{i=1}^{N}(x_i y_i)\cdot\sum_{i=1}^{N} x_i$$

$$\boxed{\Rightarrow a_0=\frac{\sum_{i=1}^{N} x_i^2\cdot\sum_{i=1}^{N} y_i-\sum_{i=1}^{N}(x_i y_i)\cdot\sum_{i=1}^{N} x_i}{N\cdot\sum_{i=1}^{N} x_i^2-\left(\sum_{i=1}^{N} x_i\right)^2}}$$

slope of the line

$\therefore$ For a linear fit, $y=a_0+a_1x$

$$a_0=\frac{(\Sigma x^2)(\Sigma y)-(\Sigma xy)\cdot(\Sigma x)}{n(\Sigma x^2)-(\Sigma x)^2}\ ;\quad a_1=\frac{n(\Sigma xy)-(\Sigma x)(\Sigma y)}{n(\Sigma x^2)-(\Sigma x)^2}$$

Goodness of the fit: (R^2) R^2 lies in 0 to 1.

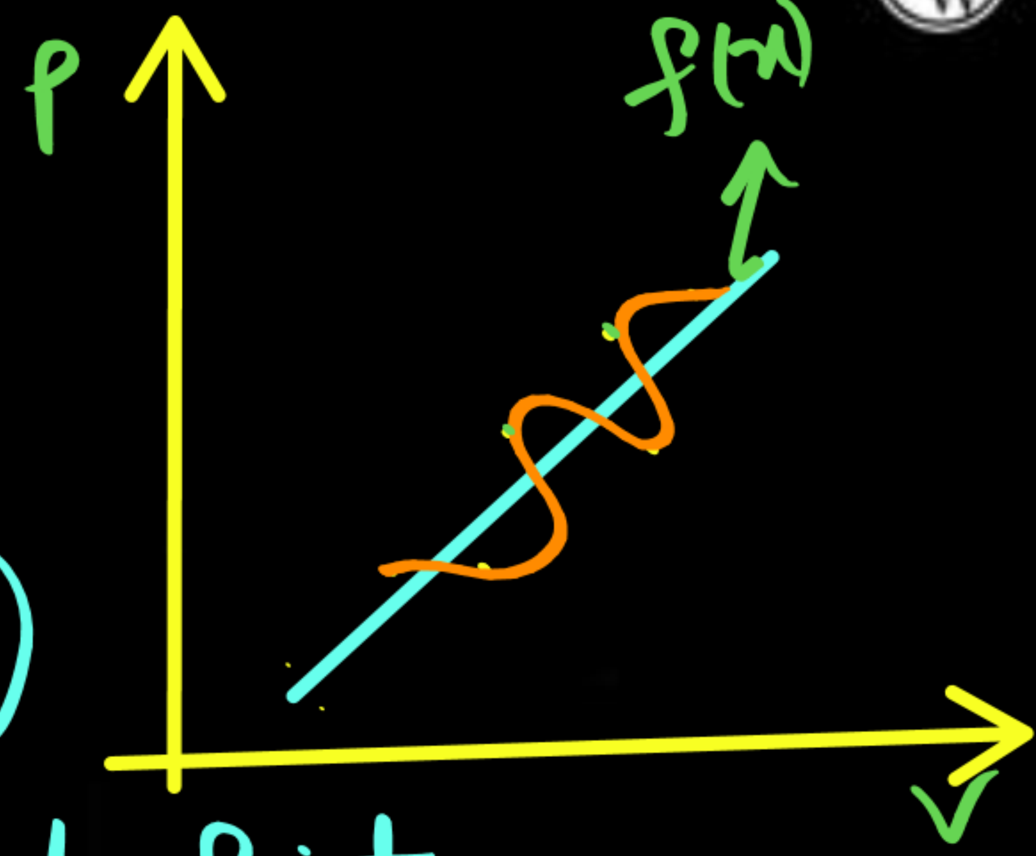
$R^2 \rightarrow$ Coefficient of determination.

$$R^2 = 1 - \frac{SS|_{Res}}{SS|_{Tot}}$$

$$\Rightarrow R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Residue

error $= (y_i - f_i)$



$y_i \rightarrow$ Actual Data Point

$f_i \rightarrow$ fitted Value.

$\bar{y} = \sum_{i=1}^{n} y_i$

$(y_i - \bar{y}) \rightarrow$ Deviation

Correlation Coefficient

$$\text{Correlation coefficient} = r = \frac{\text{cov}(x,y)}{n \cdot \sigma_x \cdot \sigma_y}$$

$$\text{cov}(x,y) = E\left((x-\bar{x})(y-\bar{y})\right)$$

$$= \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i-\bar{x})(y_i-\bar{y})$$

$$\sigma_x = \sqrt{\frac{1}{N} \cdot \sum (x_i-\bar{x})^2} \; ; \; \sigma_y = \sqrt{\frac{1}{N} \cdot \sum (y_i-\bar{y})^2} .$$

$y = a + bx \rightarrow$ Linear fit

$\Sigma y = a\Sigma 1 + b\cdot\Sigma x$

$\Rightarrow \frac{1}{N}\cdot\Sigma y = \frac{a}{\cancel{N}}\cdot\cancel{\Sigma 1} + \frac{b}{N}\cdot\Sigma x$

$\boxed{\Rightarrow \bar{y} = a + b\cdot\bar{x}}$

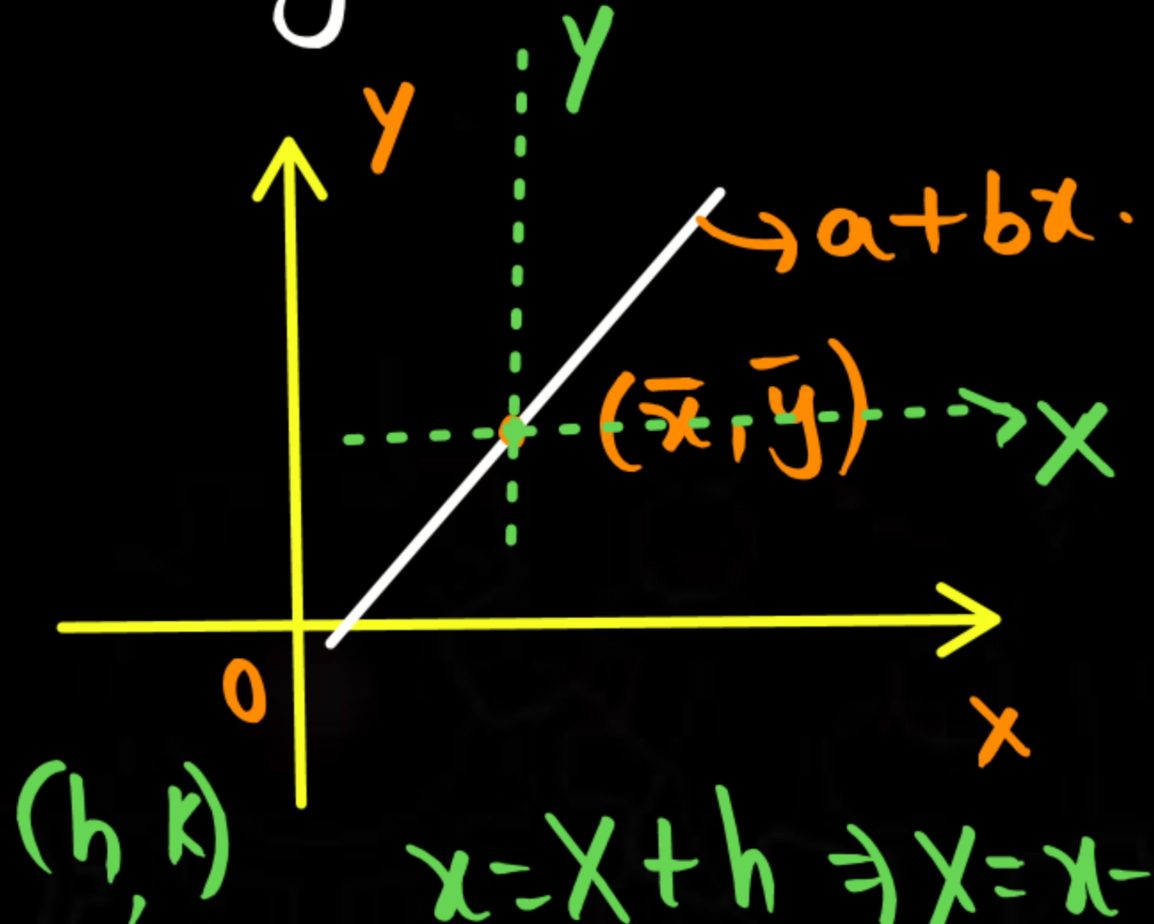
For 'N' data Points

$(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$,

the Mean $(\bar{x}, \bar{y})$ lie on the fit.

since $\Sigma y = a\cdot + b\cdot\Sigma x$

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



$x = X + h \Rightarrow X = x - \bar{x}$

$y = Y + k \Rightarrow Y = y - \bar{y}$

$$\Rightarrow \sum (x-\bar{x})(y-\bar{y}) = a\cdot\cancel{\sum (x-\bar{x})}^{0} + b\cdot\sum (x-\bar{x})^2.$$

(sum of deviations is 0).

$$\Rightarrow \quad b = \text{Slope} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$\sigma_x^2 = \frac{1}{n}\cdot\sum (x-\bar{x})^2$$

$$\Rightarrow b = \text{Slope} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n\cdot\sigma_x^2}$$

$$\Rightarrow b = \frac{r\cdot\cancel{n}\cdot\cancel{\sigma_x}\cdot\sigma_y}{\cancel{n}\cdot\sigma_x^{\cancel{2}}}$$

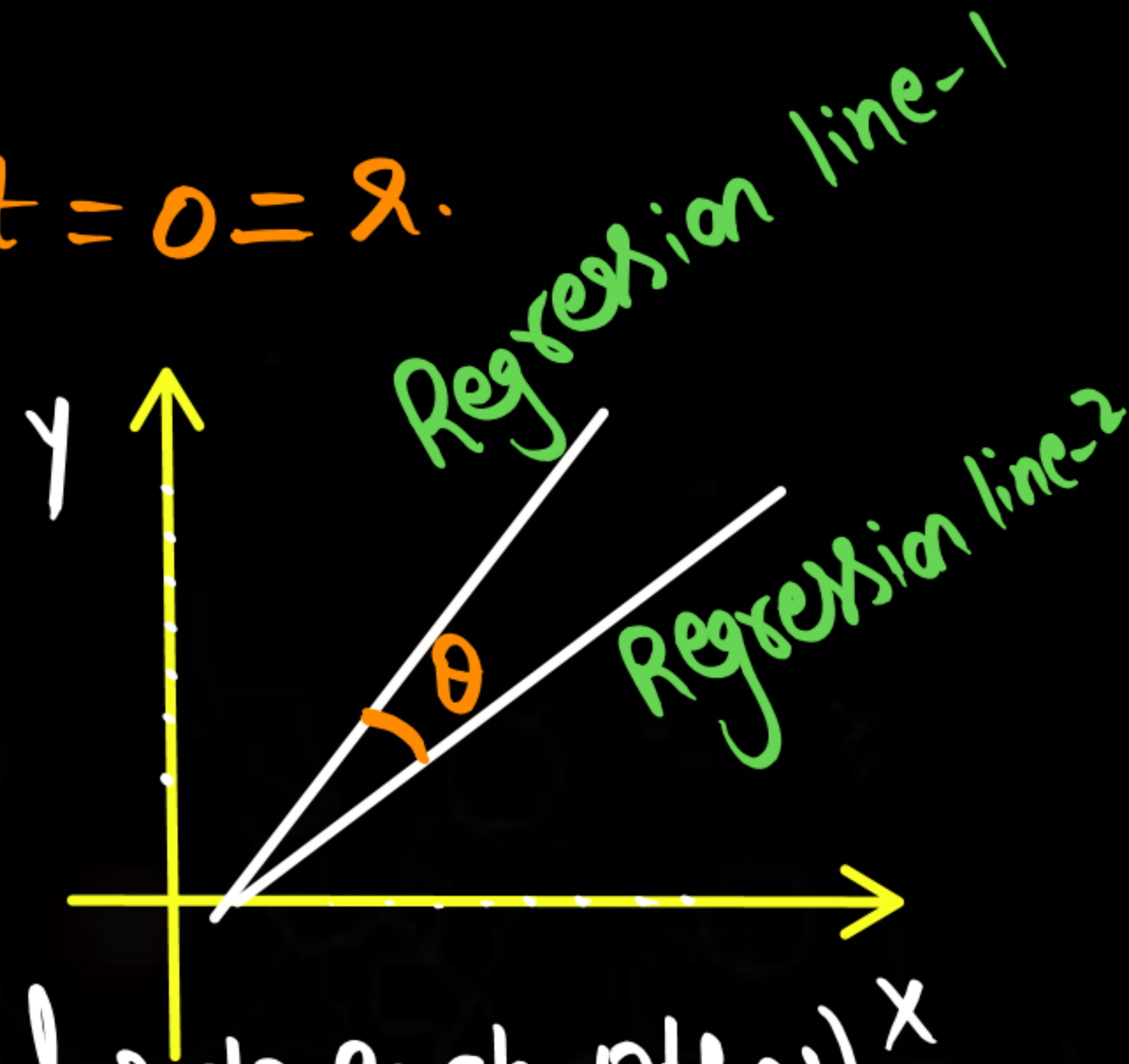
$$\boxed{\Rightarrow \text{Slope} = r\cdot\frac{\sigma_y}{\sigma_x}}$$

For two independent variables,

$$cov(X,Y) = 0$$

Angle b/w the correlation lines $\Rightarrow$ correlation coefficient $= 0 = r$.

$$\boxed{\tan\theta = \frac{1-r^2}{r} \cdot \frac{\sigma_x \cdot \sigma_y}{\sigma_x^2 + \sigma_y^2}}$$



If $r = 0 \Rightarrow \theta = \frac{\pi}{2}$ (lines are Perpendicular to each other)

If $r = 1 \Rightarrow \theta = 0$, (lines are Parallel or coincident)

Thank You!

GW Soldiers