X = Input variable / predictors / independent variables

Y = Output variable / response / dependent variables

N = # of observations

P = # of features

$x_{\text{observation\# feature\#}}$ ⟹ x_{11} = GRE Score of the first observation

x_{12} = TOEFL Score of the first observation

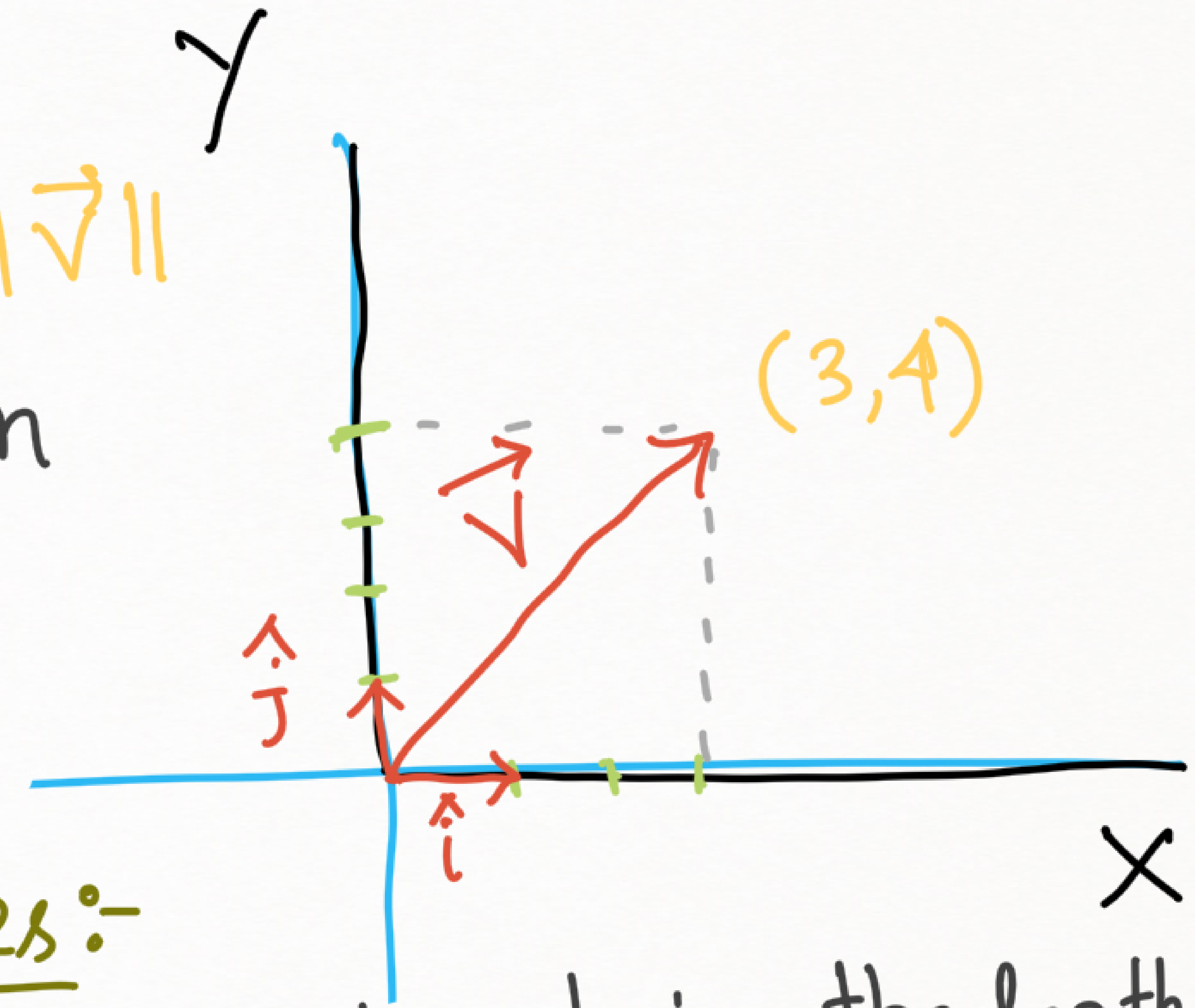
x_{21} = GRE Score of the second observation

So a single observation can be represented as below,

$$\text{row1} = \left(x_{11} \quad x_{12} \quad x_{13} \quad \cdots \quad x_{1P} \right)$$

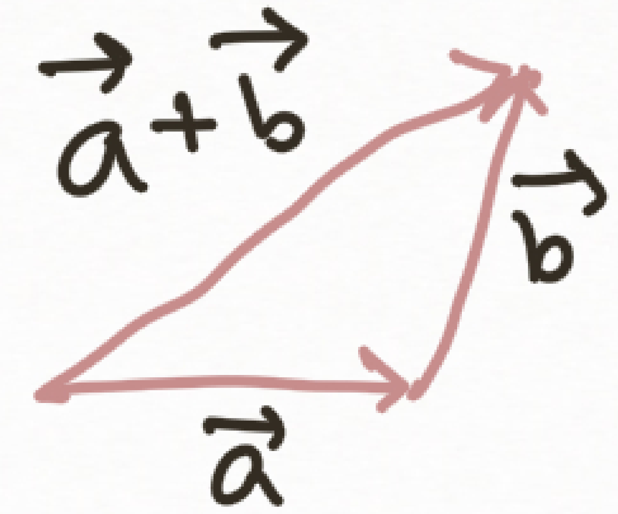
$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}$$

length $= \|\vec{v}\|$
direction



Properties:-

$2\vec{v} =$ A vector twice the legth of $\vec{v}$ but in same direction



$\Rightarrow \quad 3\hat{i} + 4\hat{j} = \vec{v}$

$\Downarrow$

(column vector) $\begin{bmatrix} 3 \\ 4 \end{bmatrix}$

Transpose:- (row becomes cols & cols becomes rows)

$\vec{v} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$

$\vec{v}^{\,T} = \begin{bmatrix} 3 & 4 \end{bmatrix}$

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$$

$$x_i^T = [\, x_{i1} \quad x_{i2} \quad \cdots \quad x_{ip} \,]$$

(Vector of length p)

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}$$

(Row view of the training inputs)

Representing each and every feature column as vector Z_j

$$z_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{Nj} \end{bmatrix}$$

Vector of length N

$$X = [\, z_1 \quad z_2 \quad \cdots \quad z_p \,]$$

(Column view of inputs)

Similarly, the output Y can be represented like below,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Based on # of features LR can be sub-divided in two categories,

LR

Simple/Univariet LR

Multivariable LR

For Simple LR $\longrightarrow$

$$X = \begin{bmatrix} z_1 \end{bmatrix} = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{N1} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

So, Basically we can view the whole dataset as,

(x_1, y_1), (x_2, y_2), ... (x_N, y_N)

Questions we should ask:-

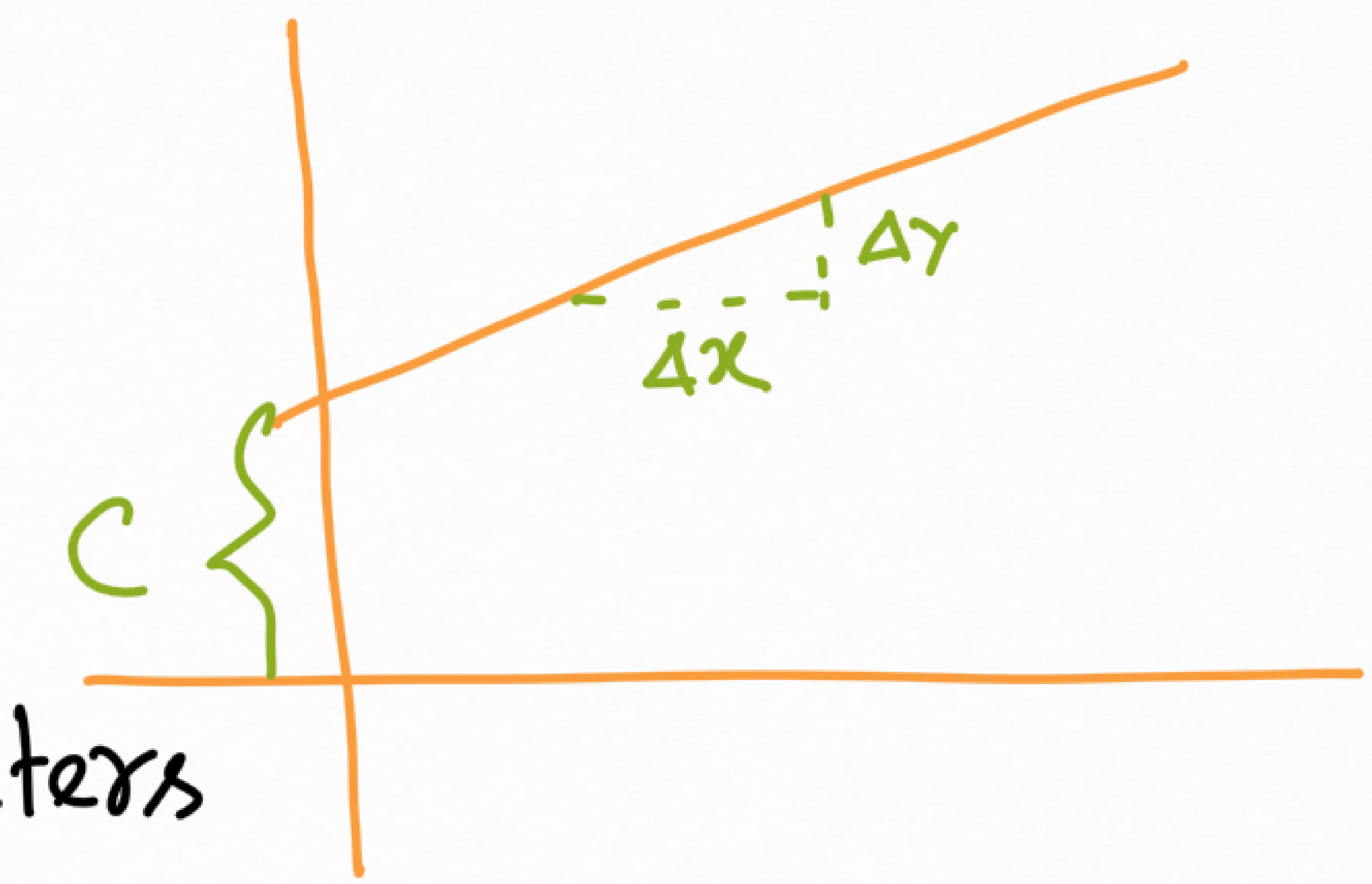
1) Is there a relationship between GRE Score & chance of admit?

2) How strong is the relationship? Given a GRE Score can we predict chance of admit?

3) Is the relationship linear? ***

Lets do the math ☺

(model) $\boxed{y = mx + C}$



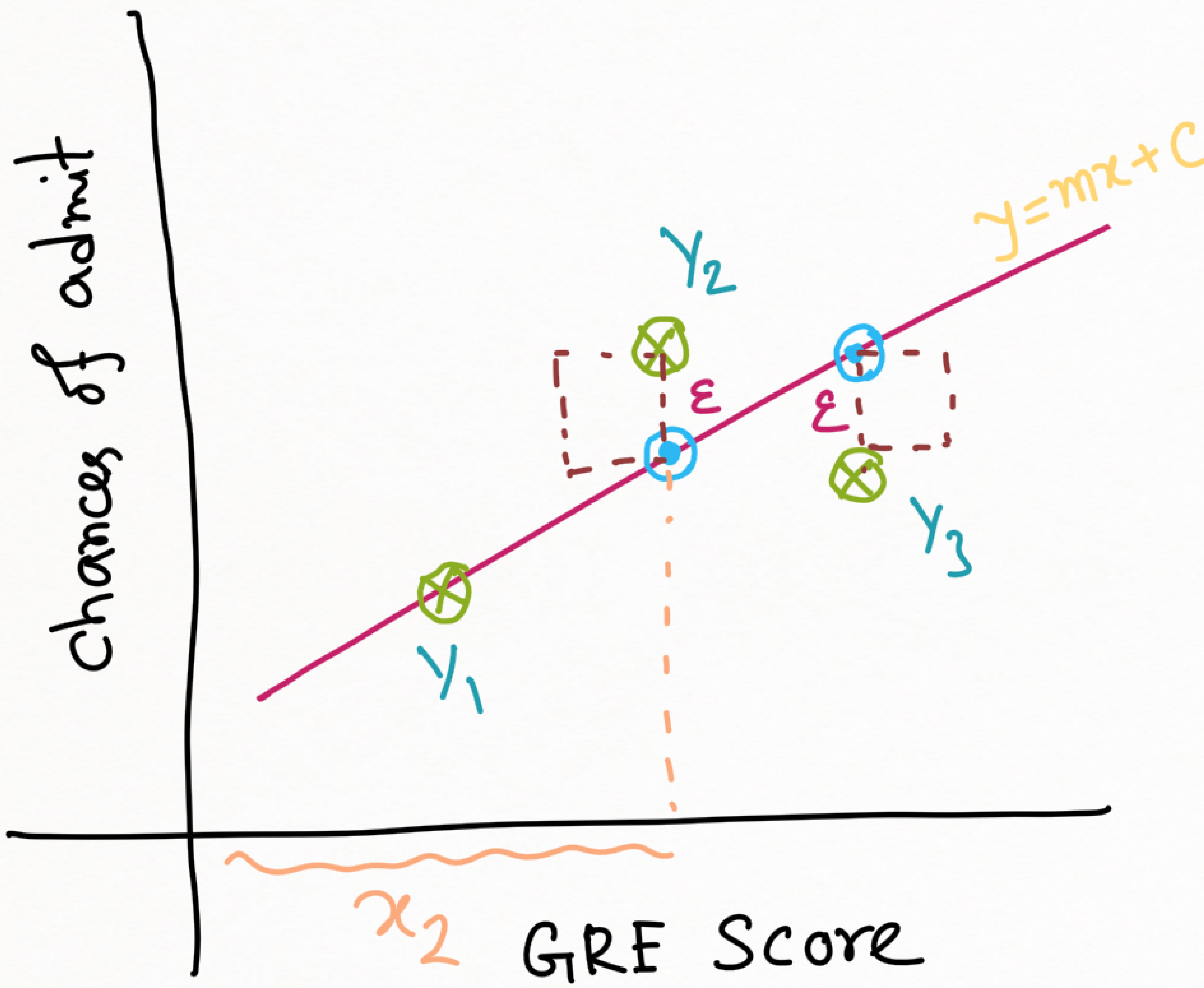
$m = \frac{\Delta y}{\Delta x}$

= Slope

C = Y-axis Intercept

In ML, m & C are called parameters or co-efficients

$$(\text{chances of Admit}) = m\,(\text{GRE Score}) + c$$



First think we need to measure how goodness the fit is.

We will define ε as a error measuring function for single data point with below rule,

1. If the line pass through that point error (ε) is zero.
2. ε can't be negative

$$\boxed{\varepsilon = (y - y_i)^2}$$

This is error for a single point.

⋮

$$\text{Total Error } (E) = \sum_{i=1}^{N} \varepsilon \quad \text{(Sum of all errors)}$$

$$E = \sum (y - y_i)^2$$

[as $\varepsilon = (y - y_i)^2$ and we will assume $\sum = \sum_{i=1}^{N}$]

$$\Rightarrow E = \sum (mx_i + c - y_i)^2$$

[as the predicted value $y = mx_i + c$]

$$\Rightarrow E = \sum (m^2 x_i^2 + c^2 + y_i^2 + 2mx_i c - 2cy_i - 2mx_i y_i)$$

[as, $(a+b+c)^2 = a^2 + b^2 + c^2 + 2ab + 2bc + 2ca$]

$$\Rightarrow \boxed{E = \sum m^2 x_i^2 + \sum c^2 + \sum y_i^2 + \sum 2mx_i c - \sum 2cy_i - \sum 2mx_i y_i}$$

Let us further simplify it,

Lets say,

$\alpha = \sum y_i^2$

$\beta = \sum x_i^2$

$\gamma = \sum x_i y_i$

$\mu = \sum y_i$

$\theta = \sum x_i$

$$E = m^2 \sum x_i^2 + \sum c^2 + \sum y_i^2 + 2mc \sum x_i - 2c \sum y_i - 2m \sum x_i y_i$$

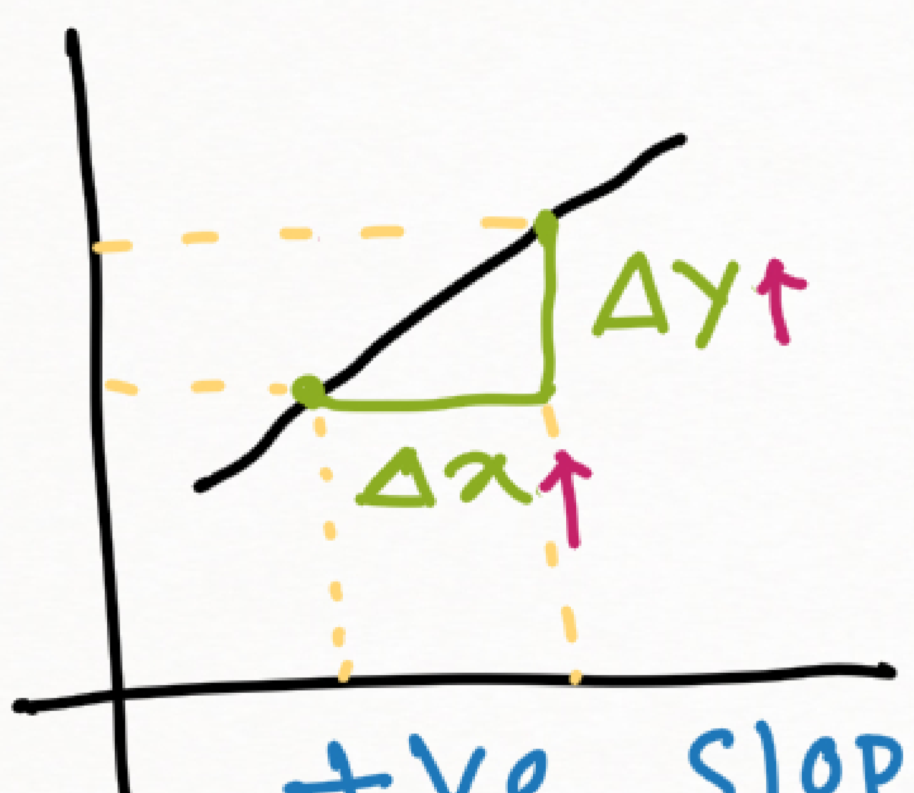
$$E = m^2 \beta + \sum_{i=1}^{N} c^2 + \alpha + 2mc\theta - 2c\mu - 2m\gamma$$

$$\boxed{E = m^2 \beta + c^2 N + \alpha + 2mc\theta - 2c\mu - 2m\gamma}$$

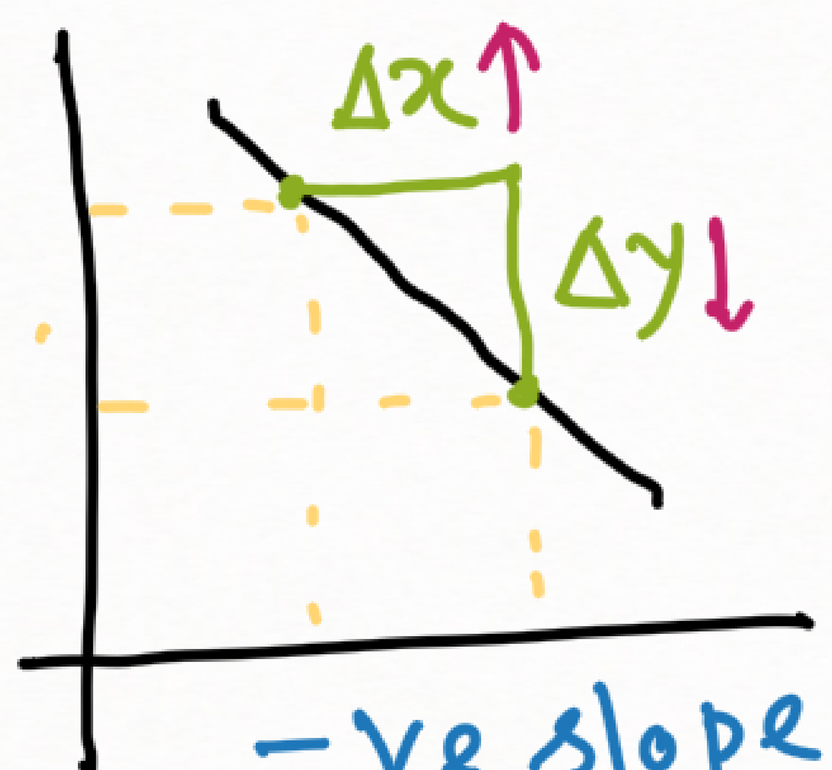
** Now our Job is to find m & c values for which E is min

$\boxed{\text{Derivative}}$ = $\boxed{\text{Slope}}$ of tangent line

$$\text{Slope} = \frac{\Delta y}{\Delta x}$$ (when we go from one point to another point)

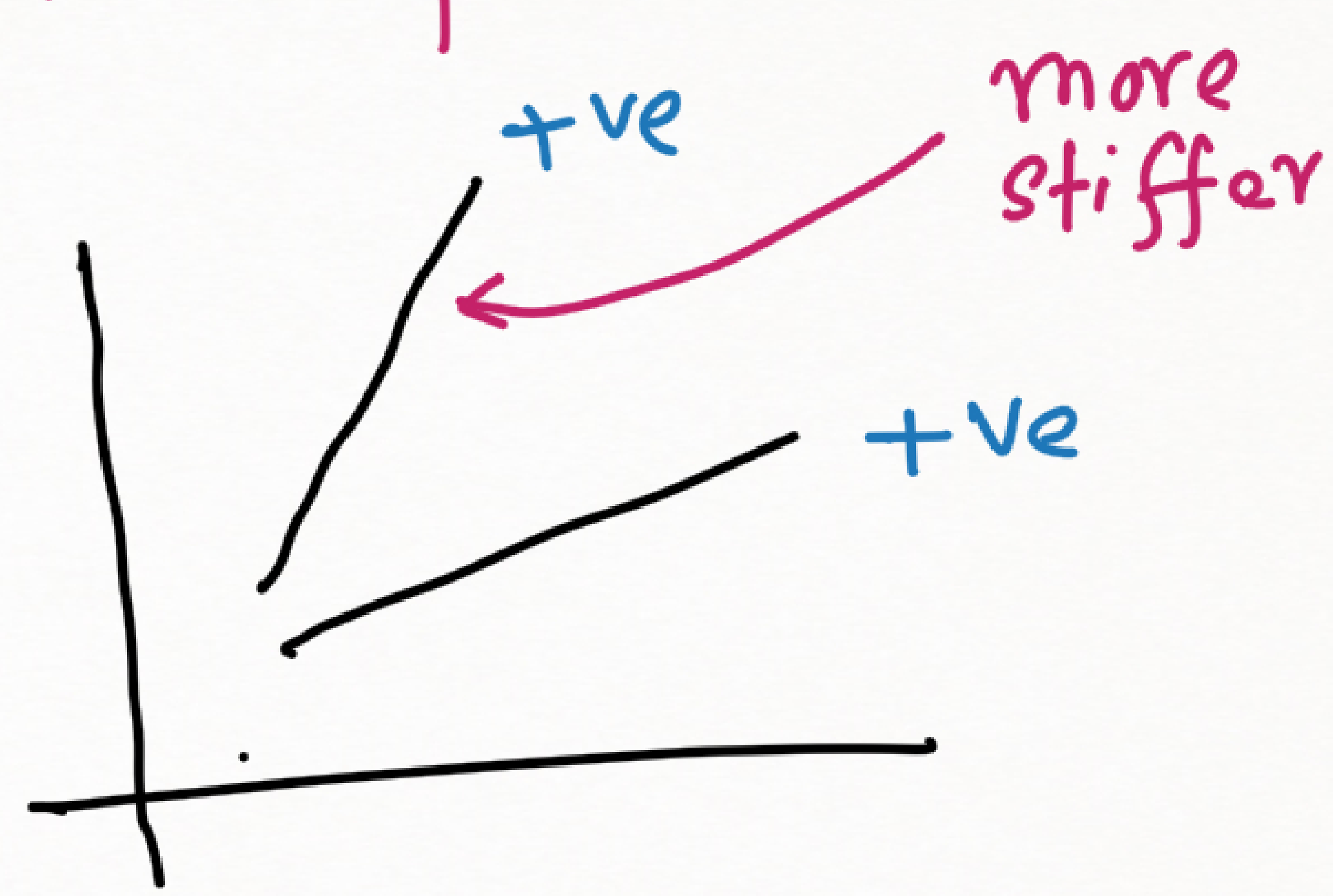


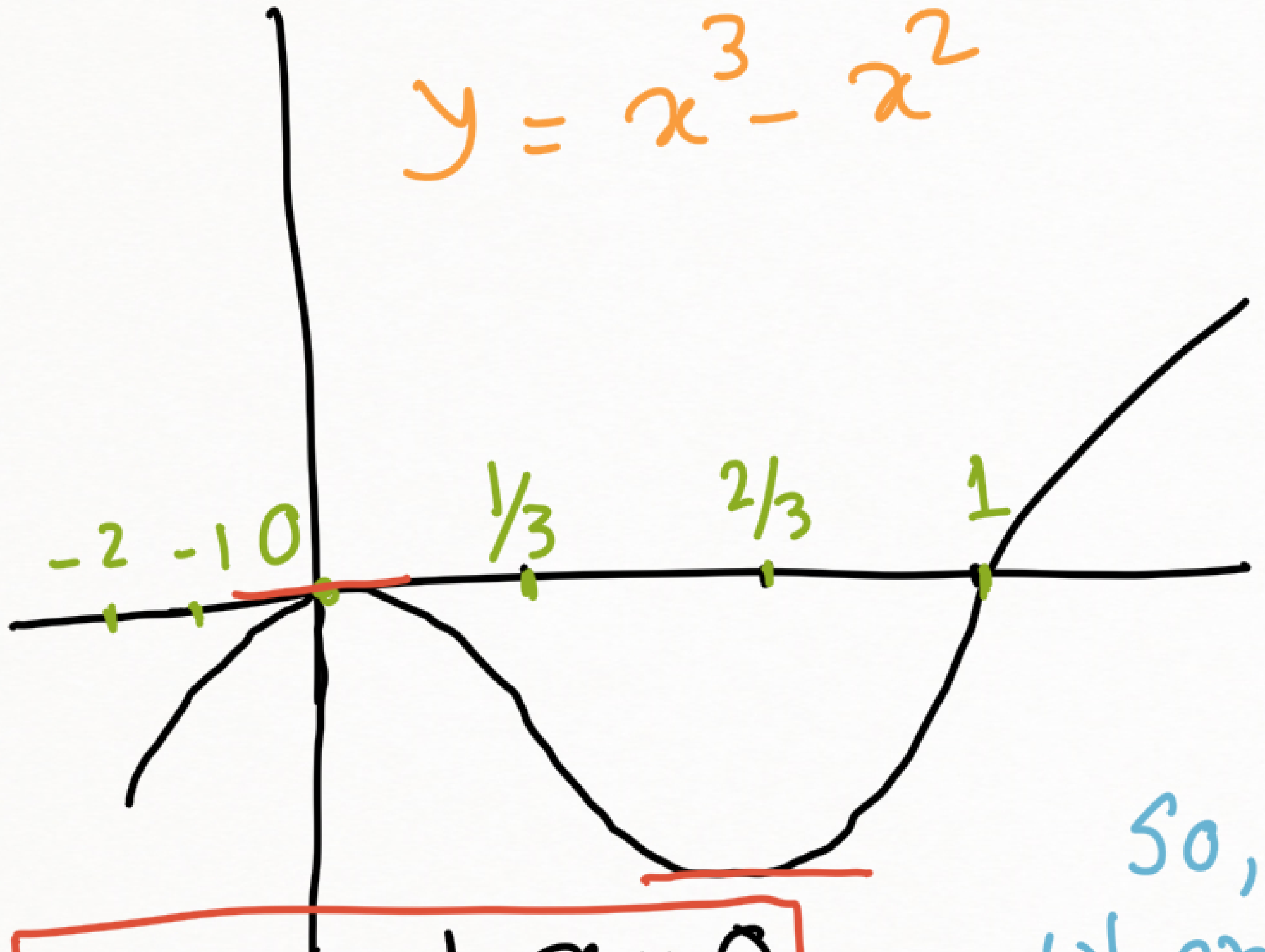
+ve slope



−ve slope

slope=0





$\boxed{\text{Slope at } x = 0}$

$y' = 0$

Slope at $x = \frac{1}{3}$

$y' = -\frac{1}{3}$

$\boxed{\text{Slope at } x = \frac{2}{3}}$

$y' = 0$

$\frac{dy}{dx} = y' = 3x^2 - 2x$

$\left[\frac{d}{dx} x^n = n . x^{n-1} \right]$

Slope at $x = -2$

$y' = 3(-2)^2 - 2(-2) = 16$

Slope at $x = -1$

$y' = 3(-1)^2 - 2(-1) = 5$

So, we are dealing with maxima or minima when derivative $= 0$. To Find whether its maxima or minima we need to take second derivative.

$\frac{d^2y}{dx^2} = y'' = 6x - 2$

at $x = 0$

$y'' = -2 < 0$

(maxima)

at $x = \frac{2}{3}$

$y'' = 2 > 0$

(minima)

Now we reach a point where we need to take derivative of E

As E is a function of m & c So we need to take partial derivative

$$E = m^2\beta + c^2 N + \alpha + 2mc\theta - 2c\mu - 2m\gamma$$

$$\frac{\partial E}{\partial m} = 2m\beta + 0 + 0 + 2c\theta - 0 - 2\gamma = 2m\beta + 2c\theta - 2\gamma$$

To get min first we will set this to zero

$$\boxed{\frac{\partial^2 E}{\partial m^2} = 2\beta > 0}$$

$$\frac{\partial E}{\partial m} = 0 \Rightarrow 2m\beta + 2c\theta - 2\gamma = 0$$

$$\Rightarrow m = \frac{\gamma - c\theta}{\beta} \quad \text{- - - - (i)}$$

Similarly,

$$\boxed{\frac{\partial^2 E}{\partial c^2} = 2N > 0}$$

$$\frac{\partial E}{\partial c} = 0 + 2cN + 0 + 2m\theta - 2\mu = 0$$

$$\Rightarrow c = \frac{\mu - m\theta}{N} \quad \text{- - - - (ii)}$$

$$m = \frac{\gamma - c\theta}{\beta}$$

$$\Rightarrow m\beta = \gamma - \left(\frac{\mu - m\theta}{N}\right)\theta$$

$$\Rightarrow m\beta = \gamma - \frac{\mu\theta - m\theta^2}{N}$$

$$\Rightarrow m\beta - \gamma + \frac{\mu\theta - m\theta^2}{N} = 0$$

$$\Rightarrow \frac{m\beta N - \gamma N + \mu\theta - m\theta^2}{N} = 0$$

$$\Rightarrow m(\beta N - \theta^2) = \gamma N - \mu\theta$$

$$\Rightarrow m = \frac{\gamma N - \mu\theta}{\beta N - \theta^2}$$

Putting back $\gamma, \mu, \theta, \beta$ we get,

$$\boxed{m = \frac{N\sum x_i y_i - \sum y_i \sum x_i}{N\sum x_i^2 - (\sum x_i)^2}}$$

Similarly,

$$C = \frac{\mu - m\theta}{N}$$

$$\Rightarrow CN = \mu - \left(\frac{\gamma N - \mu\theta}{\beta N - \theta^2}\right)\theta$$

$$\Rightarrow CN = \mu - \frac{\gamma N\theta - \mu\theta^2}{\beta N - \theta^2}$$

$$\Rightarrow CN = \frac{\mu\beta N - \cancel{\mu\theta^2} - \gamma N\theta + \cancel{\mu\theta^2}}{\beta N - \theta^2}$$

$$\Rightarrow C = \frac{\mu\beta N - \gamma N\theta}{N(\beta N - \theta^2)}$$

Putting back again,

$$\boxed{C = \frac{\Sigma y_i \Sigma x_i^2 N - \Sigma x_i y_i N \Sigma x_i}{N(\Sigma x_i^2 N - (\Sigma x_i)^2)}}$$

So, Finally we found the equation of our line

$$\boxed{y = mx + C}$$