X = Input variable / predictors / independent variables

Y = Output variable / response / dependent variables

N = # of observations

p = # of features

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

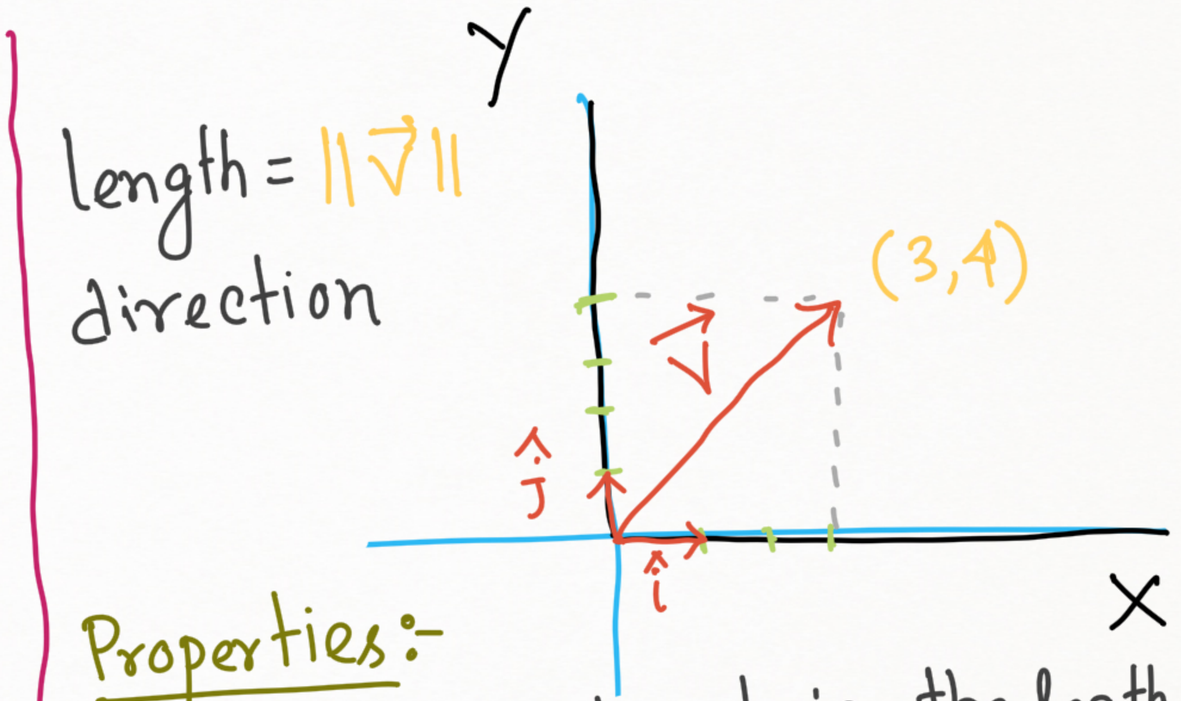
x_{12} = TOEFL Score of the first observation

x_{21} = GRE Score of the second observation

So a single observation can be represented as below,

$$row1 = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \end{pmatrix}$$

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

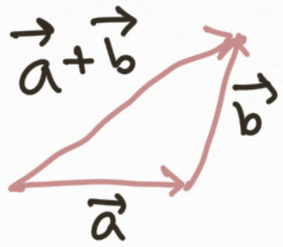


length = $\|\vec{v}\|$

direction

Properties :-

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



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

$\Downarrow$

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

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

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

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

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \qquad x_i^T = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}$$

(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)

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

Vector of length N

Representing each and every feature Column as vector z_j

$$X = \begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix}$$

(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 $\longrightarrow$

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

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

So, Basically we can view the whole dataset as,

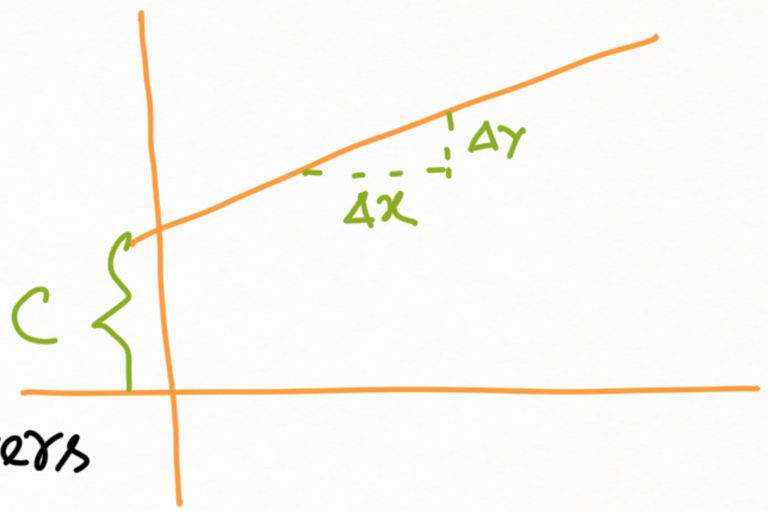
(x_1, y_1), (x_2, y_2), ... (x_N, y_N)

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) $y = mx + C$

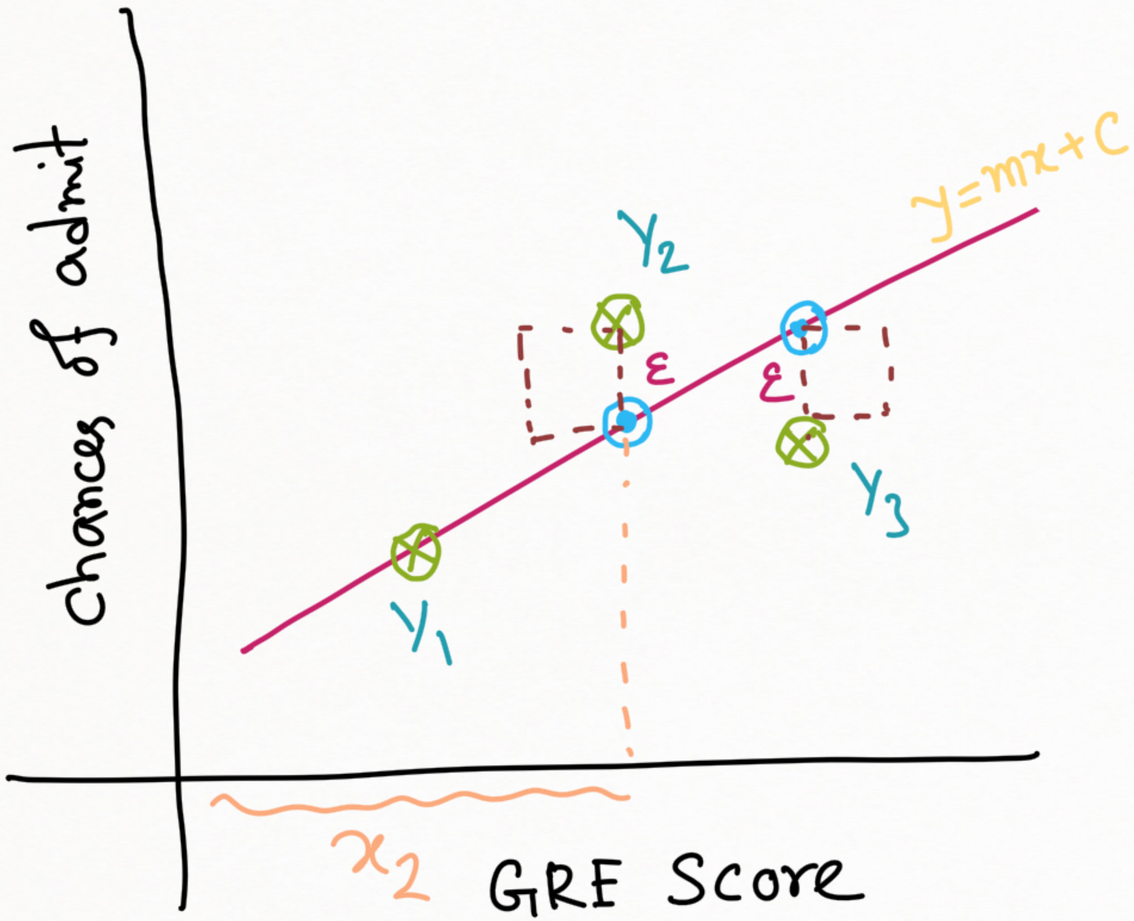


$m = \frac{\Delta y}{\Delta x}$ = Slope

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

$$\varepsilon = (y - y_i)^2$$

This is error for a single point.

⋮

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

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

$\Rightarrow E = \sum (mx_i + c - y_i)^2$ [as the predicted value $y = mx_i + c$]

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

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

$\Rightarrow$ $E = \sum m^2 x_i^2 + \sum c^2 + \sum y_i^2 + \sum 2mx_ic - \sum 2cy_i - \sum 2mx_iy_i$

Let us further simplify it,

Lets say,

$\alpha = \sum y_i^2$

$\beta = \sum x_i^2$

$\gamma = \sum x_i y_i$

$\mu = \sum y_i$

$\theta = \sum x_i$

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

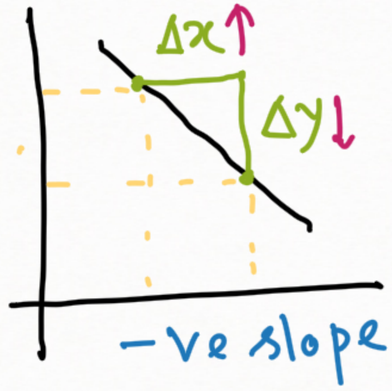
$$E = m^2 \beta + \sum_{i=1}^{N} c^2 + \alpha + 2mc\theta - 2c\mu - 2m\gamma$$

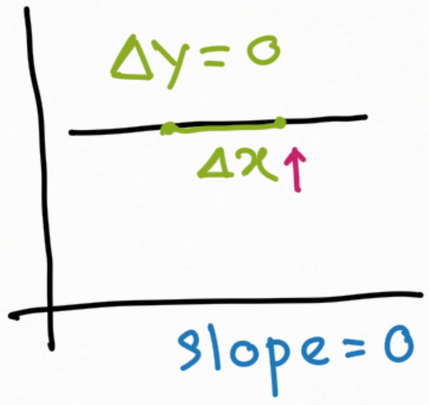
$$E = m^2 \beta + c^2 N + \alpha + 2mc\theta - 2c\mu - 2m\gamma$$

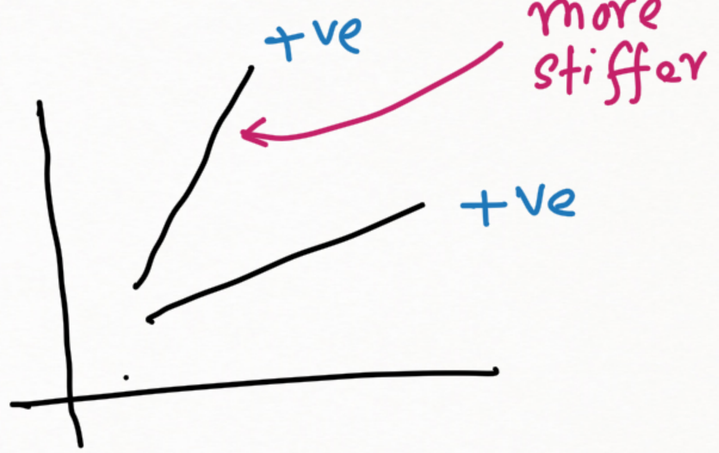
** Now our Job is to find m & c values for which E is min

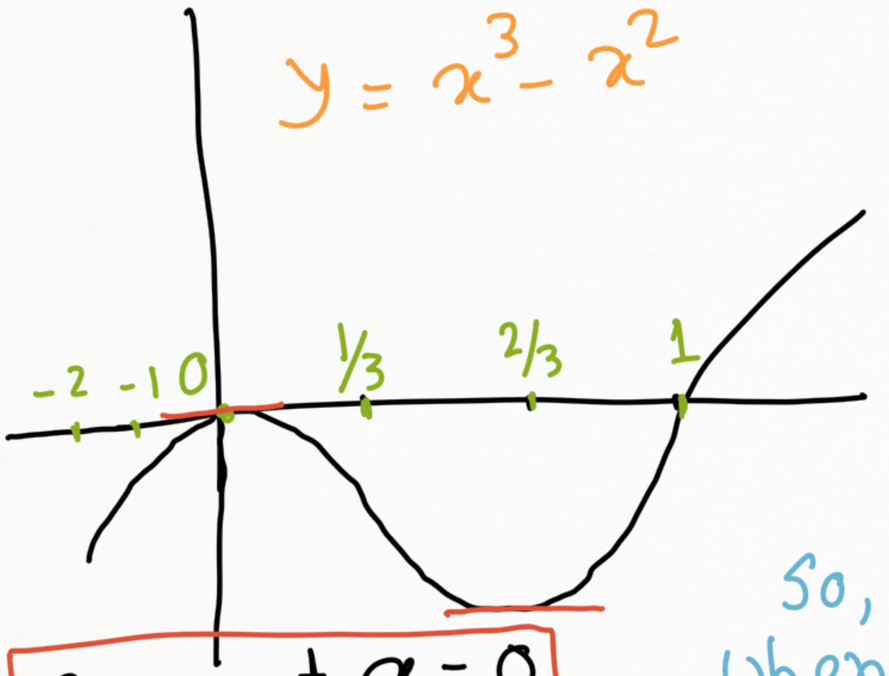
Derivative = **Slope** of tangent line

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









Slope at $x = 0$

$y' = 0$

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

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

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

$y' = 0$

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

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

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

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

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

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

Similarly,

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

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

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

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

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

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

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

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

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

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

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

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

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,

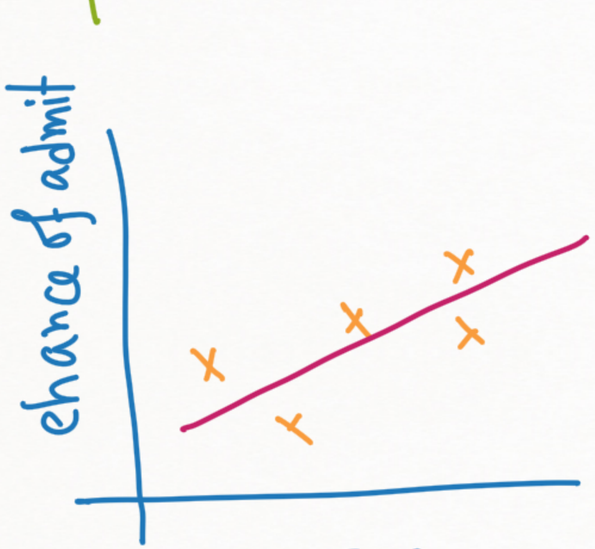
$$C = \frac{\Sigma y_i \Sigma x_i^2 N - \Sigma x_i y_i N \Sigma x_i}{N\left(\Sigma x_i^2 N - (\Sigma x_i)^2\right)}$$

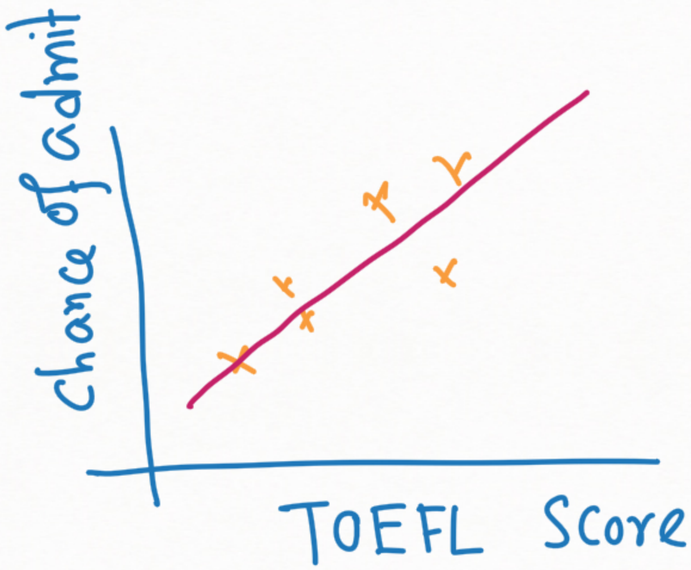
So, Finally we found the equation of our line

$$y = mx + C$$

Multiple Linear Regression :-

Approach 1 Fit seperate Simple linear regression model for each predictor.





Cons:-

1. It is unclear how to make a single prediction of Chances of admit given GRE Score & TOEFL Score
2. Each of the regression equation will ignore the effect of other predictors while forming the regression Co-efficient, which leads to wrong estimate.

Approach 2 Extend the simple linear regression model so that it can directly accomodate multiple predictors. We can do that by giving each predictor a seperate slope coefficient in a single model.

p = # of features / predictor

$$\hat{y} = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p$$

where,

$$z_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Ni} \end{bmatrix}$$

($N \times 1$ Vector)

$\hat{y}$ = Predicted output

Lets define,

$$z_0 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$ ($N \times 1$ vector)

$$\hat{y} = \beta_0 z_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p$$

$$= \sum_{i=0}^{p} \beta_i z_i$$

Lets define,

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

(p×1 vector)

We already established,

$$X = \begin{bmatrix} z_0 & z_1 & z_2 & \cdots & z_p \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{NP} \end{bmatrix}$$

(N×p matrix)

$$X\beta = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

$(N \times p \text{ matrix}) \quad \times \quad (p \times 1 \text{ vector}) = (N \times 1 \text{ vetor})$

$$= \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{N1} + \beta_2 x_{N2} + \cdots + \beta_p x_{Np} \end{bmatrix}$$

$(N \times 1 \text{ vector})$

Lets define,

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}$$

(N x 1 Vector)

So, we can write,

$$X\beta = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}$$

$$X\beta = \hat{Y}$$

$\hat{y}_i$ = predicted output for each observation

We already established,

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

where, Y = Actual output vector

y_i = Actual output for each observation

In Simple LR we were calculating the error for each observation point of training set and then summing it up their square.
Here also we will do the same.

Lets say,
error vector,

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_N - \hat{y}_N \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} = Y - \hat{Y}$$

We will be working on vectorised solution for multiple LR because taking square of a expression of p+1 terms then take derivative p times will be exausting task.

Total Error Vector $= \begin{bmatrix} e_1^2 \\ e_2^2 \\ \vdots \\ e_N^2 \end{bmatrix}$

So, $E = e_1^2 + e_2^2 + \cdots + e_N^2$

$= \sum_{i=1}^{N} e_i^2$

$\Rightarrow \boxed{E = e^T e}$

$\Rightarrow E = (y - \hat{y})^T (y - \hat{y})$

$\Rightarrow \boxed{E = (y - X\beta)^T (y - X\beta)}$

Recap

$\begin{bmatrix} e_1 & e_2 & e_3 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$

$(e^T) \quad (e)$

$= e_1^2 + e_2^2 + e_3^2$

$\boxed{= \sum_{i=1}^{3} e_i^2 = e^T e = e\,e^T}$

Recap

Let, $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ $B = \begin{bmatrix} e & f \\ g & h \end{bmatrix}$

$A^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$ $B^T = \begin{bmatrix} e & g \\ f & h \end{bmatrix}$

$A - B = \begin{bmatrix} a-e & b-f \\ c-g & d-h \end{bmatrix}$, $(A-B)^T = \begin{bmatrix} a-e & c-g \\ b-f & d-h \end{bmatrix}$

$A^T - B^T = \begin{bmatrix} a-e & c-g \\ b-f & d-h \end{bmatrix} = (A-B)^T$

$$E = \left(Y^T - (X\beta)^T\right)\left(Y - X\beta\right)$$

Recap

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad B = \begin{bmatrix} e & f \\ g & h \end{bmatrix}$$

$$A^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \qquad B^T = \begin{bmatrix} e & g \\ f & h \end{bmatrix}$$

$$AB = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}$$

$$B^T A^T = \begin{bmatrix} e & g \\ f & h \end{bmatrix}\begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

$$(AB)^T = \begin{bmatrix} ae+bg & ce+dg \\ af+bh & cf+dh \end{bmatrix} = \begin{bmatrix} ea+bg & ec+gd \\ fa+bh & fc+hd \end{bmatrix}$$

So, $(AB)^T = B^T A^T$ ..... (a)

In general, $\boxed{(ABC)^T = C^T B^T A^T}$

$$E = (y^T - \beta^T x^T)(y - x\beta)$$

$$E = y^T y - y^T x\beta - \beta^T x^T y + \beta^T x^T x\beta$$

So, we got our equation for E. Now we need to minimize E. For simple linear Regression we took derivative w.r.t m & c. So, here we will take derivative w.r.t β

So, we are talking about differentiation of a matrix with respect to a vector.

$$\frac{\partial E}{\partial \beta} = 0$$

$$\Rightarrow \frac{\partial (y^T y - y^T x\beta - \beta^T x^T y + \beta^T x^T x\beta)}{\partial \beta} = 0$$

$$\Rightarrow \frac{\partial (y^T y)}{\partial \beta} - \frac{\partial (y^T x\beta)}{\partial \beta} - \frac{\partial (\beta^T x^T y)}{\partial \beta} + \frac{\partial (\beta^T x^T x\beta)}{\partial \beta} = 0$$

- - - (1)

Matrix Derivative

Recap:-

There 6 common types of matrix derivatives.

Type	Scalar	Vector	Matrix
Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial y}{\partial x}$	$\frac{\partial Y}{\partial x}$
Vetor	$\frac{\partial y}{\partial x}$	$\frac{\partial y}{\partial x}$	
Matrix	$\frac{\partial y}{\partial x}$		

Case 1:- When y & x both are scalar, like $y = x^3 + x^2$

$$y' = \frac{\partial y}{\partial x}$$

Case 2:- When y is vector and x is scalar.

Like, $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_N}{\partial x} \end{bmatrix}$$

Case 3:- When Y is a matrix and x is scalar.

like $$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1P} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NP} \end{bmatrix}$$

$$\frac{\partial Y}{\partial x} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x} & \frac{\partial y_{12}}{\partial x} & \cdots & \frac{\partial y_{1P}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_{N1}}{\partial x} & \frac{\partial y_{N2}}{\partial x} & \cdots & \frac{\partial y_{NP}}{\partial x} \end{bmatrix}$$

..

Case 4 :- When y is scalar and x is a vector.

Like, $x = [x_1\ x_2 \dots x_N] \Rightarrow \dfrac{\partial y}{\partial x} = \left[\dfrac{\partial y}{\partial x_1}\ \dfrac{\partial y}{\partial x_2} \cdots \dfrac{\partial y}{\partial x_N}\right]$

Case 5 :- When y is vector and x is vector.

Like, $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$ $x = [x_1\ x_2 \cdots x_n]$ (Numerator layout notation)

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

Case 6 :- when y is a scalar and X is a Matrix.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NP} \end{bmatrix}$$

$$\frac{\partial y}{\partial X} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{N1}} \\ \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{N2}} \\ \vdots & \cdots & \vdots \\ \frac{\partial y}{\partial x_{1P}} & \cdots & \frac{\partial y}{\partial x_{NP}} \end{bmatrix}$$

Some derivation

Derivation 1 :-

Lets say a is a vector x is another vector.

then $$\boxed{\frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a^T}$$

We already saw $a^T x = x^T a = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$

$$\frac{\partial a^T x}{\partial x} = \frac{\partial (a_1 x_1 + a_2 x_2 + \cdots + a_n x_n)}{\partial x}$$

(This is case 4)

$$= \left[\frac{\partial (a_1 x_1 + a_2 x_2 + \cdots + a_n x_n)}{\partial x_1} \quad \cdots \quad \frac{\partial (a_1 x_1 + a_2 x_2 + \cdots + a_n x_n)}{\partial x_n} \right]$$

$$= \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} = a^T$$

Derivation 2 :- if A is a Matrix and x is a vector the

$$\boxed{\frac{\partial Ax}{\partial x} = A}$$

Lets,

$$Ax = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n \end{bmatrix}$$

$$\frac{\partial Ax}{\partial x} = \begin{bmatrix} \frac{\partial (a_{11}x_1 + \cdots + a_{1n}x_n)}{x_1} & \cdots & \frac{\partial (a_{11}x_1 + \cdots + a_{1n}x_n)}{x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial (a_{n1}x_1 + \cdots + a_{nn}x_n)}{x_1} & \cdots & \frac{\partial (a_{n1}x_1 + \cdots + a_{nn}x_n)}{x_n} \end{bmatrix} \text{(case 5)} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$$= A$$

Derivation 3:- Lets say A and x are matrix and a vector consicutively, then

$$\frac{\partial x^T A x}{\partial x} = x^T (A + A^T)$$

We need to memorise one product rule here,

$$\frac{\partial u^T v}{\partial x} = u^T \frac{\partial v}{\partial x} + v^T \frac{\partial u}{\partial x}$$

Applying the above rule, $u = x$, $v = Ax$

$$\frac{\partial x^T Ax}{\partial x} = x^T \frac{\partial Ax}{\partial x} + (Ax)^T \frac{\partial x}{\partial x}$$

$$= x^T A + (Ax)^T \quad \text{[from derivation 2]}$$

$$\frac{\partial (x^T A x)}{\partial x} = x^T A + x^T A^T$$

$$= x^T (A + A^T) \quad \left[\text{distributive property}\right]$$

now, lets apply whatever we learn so far to compute the derivative of E. From equation (1) we get,

$$\frac{\partial E}{\partial \beta} = \frac{\partial (y^T y)}{\partial \beta} - \frac{\partial (y^T x \beta)}{\partial \beta} - \frac{\partial (\beta^T x^T y)}{\partial \beta} + \frac{\partial (\beta^T x^T x \beta)}{\partial \beta}$$

now, $\boxed{\frac{\partial (y^T y)}{\partial \beta} = 0}$ (There is no β term)

$$\frac{\partial (Y^T X \beta)}{\partial \beta} = \frac{\partial A\beta}{\partial \beta} \qquad \left[A = Y^T X = \text{a matrix} \right]$$

$$= A \qquad [\text{Derivation 2}]$$

$$\Rightarrow \boxed{\frac{\partial (Y^T X \beta)}{\partial \beta} = Y^T X}$$

$$\frac{\partial (\beta^T X^T Y)}{\partial \beta} = \frac{\partial (\beta^T a)}{\partial \beta} \qquad \left[a = X^T Y \right]$$

$$= a^T \qquad [\text{Derivation 1}]$$

$$\Rightarrow \boxed{\frac{\partial (\beta^T X^T Y)}{\partial \beta} = (X^T Y)^T}$$

$$\frac{\partial(\beta^T x^T x \beta)}{\partial \beta} = \frac{\partial(\beta^T A \beta)}{\partial \beta} \quad [A = X^T X]$$

$$= \beta^T (A + A^T) \quad [\text{Derivation 3}]$$

$$= \beta^T (x^T x + (x^T x)^T)$$

$$= \beta^T (x^T x + x^T x)$$

$$\Rightarrow \boxed{\frac{\partial(\beta^T x^T x \beta)}{\partial \beta} = 2\beta^T x^T x}$$

So, $\frac{\partial E}{\partial \beta} = 0 - y^T x - (x^T y)^T + 2\beta^T x^T x$

$= -y^T x - y^T x + 2\beta^T x^T x$

$= 2\beta^T x^T x - 2y^T x$

Setting $\frac{\partial E}{\partial \beta} = 0$

$\Rightarrow 2\beta^T x^T x - 2y^T x = 0$

$\Rightarrow \beta^T x^T x = y^T x$

$\Rightarrow x^T x \beta = x^T y$ [taking transpose both side]

$\Rightarrow \boxed{\beta = (x^T x)^{-1} x^T y}$

Our equation to find co-efficients.