① { multiple linear regression }

MLR is used when there are multiple input features and a single output feature. (more than 1 predictors)

Suppose we have n input features. Let them be $X_1, \dots, X_n$ and let the output feature be y. Then MLR is :

$$\boxed{\hat{y} = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}$$

The objective is to find $(\beta_0, \dots, \beta_n)$ $\ni$ we have minimum loss

② { Mathematical formulation }

Let us suppose we have m input features $(X_1, \dots, X_m)$ and n such data points. Let y be the output feature. So our dataset looks like this.

m predictors

X_1	X_2		X_m	y
X_{11}	X_{12}	....	X_{1m}	y_1
X_{21}	X_{22}	....	X_{2m}	y_2
$\vdots$				$\vdots$
X_{n1}			X_{nm}	y_n

n data points

output feature

Then
$$\hat{y}_1 = \beta_0 + \beta_1 X_{11} + \dots + \beta_m X_{1m} \quad - ①$$
$$\vdots \qquad\qquad \vdots$$

$$\hat{y}_n = \beta_0 + \beta_1 x_{n1} + \dots + \beta_m x_{nm} \quad \text{— } (n)$$

These n equations can then be written as :

$$\underbrace{\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}}_{(n \times 1) \text{ shape}} = \underbrace{\begin{bmatrix} \beta_0 + \beta_1 x_{11} + \dots + \beta_m x_{1m} \\ \\ \beta_0 + \beta_1 x_{n1} + \dots + \beta_m x_{nm} \end{bmatrix}}_{(n \times 1) \text{ shape}}$$

$$= \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ & & \vdots & \\ 1 & x_{n1} & \dots & x_{nm} \end{bmatrix}_{n \times (m+1)} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_m \end{bmatrix}_{(m+1) \times 1}$$

Thus, $\boxed{\hat{y} = X\beta}$ $\Longrightarrow$ super important equation.

③ { Mathematical formulation of error function }

Let $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ and $\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}$

We define
$$E = d_1^2 + d_2^2 + \dots + d_n^2$$
$$= (y_1 - \hat{y}_1)^2 + \dots + (y_n - \hat{y}_n)^2$$

$$= \begin{bmatrix} y_1-\hat{y}_1 & \cdots & y_n-\hat{y}_n \end{bmatrix}_{1\times n} \begin{bmatrix} y_1-\hat{y}_1 \\ \vdots \\ y_n-\hat{y}_n \end{bmatrix}_{n\times 1}$$

$$= e^T e \quad , \quad e = \begin{bmatrix} y_1-\hat{y}_1 \\ \\ y_n-\hat{y}_n \end{bmatrix}_{n\times 1}$$

$$= \|e\|^2 \quad \text{(norm of } e\text{)}$$

Thus, error function E can also be represented as a product of two matrices, namely e^T and e.

We may also write e as $y-\hat{y}$

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

$$= \left(y^T - (\hat{y})^T\right)(y-\hat{y})$$

$$= y^T y - \underbrace{y^T\hat{y} - (\hat{y})^T y}_{\text{These two are equal.}} + (\hat{y})^T\hat{y}$$

Claim :- $y^T\hat{y} = (\hat{y})^T y$ i,e $y^T\hat{y}$ is a symmetric matrix

Proof :- y^T has shape $1\times n$ and $\hat{y}$ has shape $n\times 1$.

Thus, $y^T y$ has shape 1×1. Hence, we have shown that it is a scalar matrix.

We know that every scalar matrix is symmetric. Thus

$$y^T \hat{y} = (y^T \hat{y})^T = (\hat{y})^T (y^T)^T$$

$$= (\hat{y})^T y$$

$\boxed{\text{QED}}$

Thus, $E = y^T y - 2 y^T \hat{y} + (\hat{y})^T \hat{y}$

(4) {final derivation}

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

putting this in the equation of E, we get

$$E = y^T y - 2 y^T X \beta + (X\beta)^T (X\beta)$$

$$\boxed{E = y^T y - 2 y^T X \beta + \beta^T X^T X \beta}$$

main equation

Thus, E is a function of β as X and y are already fixed.

$$E(\beta) = y^T y - 2y^T X \beta + \beta^T X^T X \beta$$

We need to find $\beta \ni E(\beta)$ has minimum value.

Here
$$y = (y_1 \dots y_n)^T$$
$$\beta = (\beta_0 \dots \beta_m)^T$$
$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}_{n \times (m+1)}$$

find such value of β matrix for which E is min. for minimum, we should have.

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

Now

$$\frac{\partial E}{\partial \beta} = \frac{\partial}{\partial \beta}\left(y^T y - 2y^T X \beta + \beta^T X^T X \beta\right)$$

$$= 0 - 2y^T X + \underbrace{\frac{\partial}{\partial \beta}\left(\beta^T X^T X \beta\right)}$$

> For this we need to study matrix calculus which is not feasible as of now.
> However, we have a special result in matrix calculus which says that if A is a symme-- matrix and $\alpha = x^T A x$, then
>
> $$\frac{\partial \alpha}{\partial x} = 2x^T A \text{, where } x \text{ is}$$
>
> a $(n \times 1)$ vector.

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

Thus, $$\frac{\partial E}{\partial \beta} = 2\beta^T X^T X - 2y^T X$$

Equating this to 0, we get

$$\beta^T X^T X = y^T X$$

Assuming that $X^T X$ is invertible, we have

$$\beta^T X^T X \left(X^T X\right)^{-1} = y^T X \left(X^T X\right)^{-1}$$

$$\Rightarrow \quad \beta^T = y^T X X^{-1} (X^T)^{-1}$$

$$\Rightarrow \quad (\beta^T)^T = \left\{ y^T X X^{-1} (X^T)^{-1} \right\}^T$$

$$\Rightarrow \quad \beta = \left[(X^T)^{-1} \right]^T (X^{-1})^T X^T (y^T)^T$$

$$\Rightarrow \quad \beta = X^{-1} (X^{-1})^T X^T y$$

$$\Rightarrow \quad \beta = X^{-1} (X^T)^{-1} X^T y$$

$$= \left[X^T X\right]^{-1} X^T y$$

Thus, $\boxed{\beta = (X^T X)^{-1} X^T y}$

$X \rightarrow n \times (m+1)$

$X^T \rightarrow (m+1) \times n$

$[X^T X]^{-1} \rightarrow (m+1) \times (m+1)$

Now $[X^T X]^{-1} X^T$ has shape $(m+1) \times n$. Also, since y has shape $n \times 1$. Thus, $\beta = [X^T X]^{-1} X^T y$ has shape $(m+1) \times 1$. ■

⑤ { Problem with OLS solution }

We have $\beta = (X^T X)^{-1} X^T y$

Matrix multiplication is a computationaly expansive operation. In higher dimension, this is a slow procedure. Thus in higher dimension, we prefer to use non-closed approximation techniques like gradient descent which is computationaly less expansive.