Jeet Banerjee

Mathematical Formulation (Multiple Linear Regression)

Consider, in a given dataset,
$n \rightarrow$ no. of rows / no. of data points
$m \rightarrow$ no. of i/p or predictor variables
and only has '1' o/p variable

Multiple Linear Regression (MLR) $\downarrow$ more than 1 i/p and single o/p

Hence, the given dataset can be look like:

i/p var-1 (x_1)	i/p var-2 (x_2)	i/p var-3 (x_3)	...	i/p var-m (x_m)	o/p var (y)
x_{11}	x_{12}	x_{13}	...	x_{1m}	y_1
x_{21}	x_{22}	x_{23}	...	x_{2m}	y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{n1}	x_{n2}	x_{n3}	...	x_{nm}	y_n

Hence, the equation of predicted o/p:

$$\boxed{\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_m x_m} \quad -(1)$$

where, β_0 = intercept / offset value
$\beta_1, \beta_2, \beta_3, \ldots, \beta_m$ = coefficients for each of 'm' i/p variables

Hence, to get the prediction hyperplane, our primary motive is to calculate $\beta_0, \beta_1, \beta_2, \beta_3, \ldots, \beta_m$

Let's write down $\hat{y}$ in a matrix multiplication format:

$$\hat{y}_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \cdots + \beta_m x_{1m}$$
$$\hat{y}_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \cdots + \beta_m x_{2m}$$
$$\vdots$$
$$\hat{y}_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \cdots + \beta_m x_{nm}$$

$$\Rightarrow \underbrace{\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}}_{\hat{y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_m \end{bmatrix}}_{\beta} \Rightarrow \boxed{\underbrace{\hat{y}}_{(n\times 1)} = \overbrace{X}^{n\times(m+1)}\underbrace{\beta}_{(m+1)\times 1}} \quad -(2)$$

Jeet Banerjee

Now, loss function is represented as the mean squared error

$$\boxed{L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad -③$$

Consider, the error b/w actual and predicted o/p is represented in matrix format by 'e'; hence,

$$e = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} \Rightarrow e^T e = \begin{bmatrix} y_1 - \hat{y}_1 & y_2 - \hat{y}_2 & \cdots & y_n - \hat{y}_n \end{bmatrix} \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$$

$$\Rightarrow e^T e = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2$$

$$\Rightarrow \boxed{e^T e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad -④$$

From eq. (3) and eq. (4), the loss function can be written as:

$$L = e^T e \text{ ; where } e = y - \hat{y}$$

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

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

$$\Rightarrow L = y^T y \boxed{- y^T \hat{y} - \hat{y}^T y} + \hat{y}^T \hat{y}$$

Let's investigate the terms $y^T \hat{y}$ and $\hat{y}^T y$:

$$\underbrace{\underset{1 \times n}{y^T}\ \underset{n \times 1}{\hat{y}}}_{1 \times 1} \qquad \underbrace{\underset{1 \times n}{\hat{y}^T}\ \underset{n \times 1}{y}}_{1 \times 1}$$

$\Rightarrow$ Hence, both $y^T \hat{y}$ and $\hat{y}^T y$ represent scalar values

- $$y^T \hat{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = y_1 \hat{y}_1 + y_2 \hat{y}_2 + \cdots + y_n \hat{y}_n$$

- $$\hat{y}^T y = \begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \cdots & \hat{y}_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = y_1 \hat{y}_1 + y_2 \hat{y}_2 + \cdots + y_n \hat{y}_n$$

Hence, $\boxed{y^T \hat{y} = \hat{y}^T y}$

Jeet Banerjee

Now, the loss function can be written as:

$$\boxed{L = y^T y - 2y^T \hat{y} + \hat{y}^T \hat{y}} \quad -(5)$$

From eq.(2), substituting $\hat{y} = X\beta$ in above eq.(5):

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

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

So, the loss function is only a function of β (coefficients + intercept) now:

$$\Rightarrow \boxed{L(\beta) = y^T y - 2y^T X\beta + \beta^T X^T X\beta} \quad -(6)$$

Is $X^T X$ a symmetric matrix?

consider, $p = X^T X \Rightarrow p^T = (X^T X)^T = X^T (X^T)^T = X^T X \Rightarrow p^T = p \Rightarrow (X^T X)^T = X^T X$

hence, $X^T X$ is a symmetric matrix

We need to calculate intercept (β_0) and coefficients $(\beta_1, \beta_2, \ldots, \beta_m)$ such that the loss/error would be minimum.

So, $\frac{dL}{d\beta} = 0$

$$\Rightarrow 0 - 2y^T X + 2\beta^T X^T X = 0$$

$$\Rightarrow \boxed{\beta^T X^T X = y^T X} \quad -(7)$$

From eq.(7), we need to find out the value of β. To separate out β, multiplying $(X^T X)^{-1}$ both the sides:

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

considering $p = X^T X$, hence $pp^{-1} = I$

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

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

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

$$\Rightarrow \boxed{\beta = [(X^T X)^{-1}]^T X^T y} \quad -(8)$$

Let's calculate this term

considering $p = X^T X$; hence

$pp^{-1} = I \Rightarrow (pp^{-1})^T = I^T \Rightarrow (p^{-1})^T p^T = I \Rightarrow (p^{-1})^T p = I \Rightarrow (p^{-1})^T pp^{-1} = Ip^{-1}$

as $p = X^T X$ is a symmetric matrix

$$\Rightarrow (p^{-1})^T = p^{-1}$$

$$\Rightarrow \boxed{[(X^T X)^{-1}]^T = (X^T X)^{-1}} \quad -(9)$$

Substituting eq. ⑨ in eq. ⑧, we got:

$$\boxed{\beta = (X^TX)^{-1}X^Ty}$$

$y \to n \times 1$

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

$\Big] \to (m+1) \times 1$

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

$\downarrow$

$(m+1) \times (m+1)$

$(m+1) \times 1$

Hence, the dimension of β is $(m+1) \times 1$, and it justifies as

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}_{(m+1) \times 1}$$

Hence, for MLR, the intercept and coefficients are calculated using the eq.:

$$\boxed{\beta = (X^TX)^{-1}X^Ty} \quad \text{— (10)}$$