Multiple Linear Regression

It is a statistical method used to model the relationship between one continuous dependent variable and two or more independent variables, which may be continuous or categorical.

The goal is to predict the dependent variable based on the independent variables by fitting a linear equation to the data.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

- y — Dependent variable
- β_0 — Intercept (Offset)
- β_1, β_2 — Coefficients of independent variable
- x_1, x_2 — Independent variables
- ϵ — error term

Mathematical Foundation

Let's take sample data.

gpa (x_1)	iq (x_2)	salary (y)
8 (x_{11})	80 (x_{12})	75
7 (x_{21})	70 (x_{22})	60
5 (x_{31})	120 (x_{32})	110

$\rightarrow$

$$\hat{y}_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} \dots + \beta_m x_{1m}$$
$$\hat{y}_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_m x_{2m}$$
$$\hat{y}_3 = \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \dots + \beta_m x_{3m}$$
$$\vdots$$
$$\hat{y}_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_m x_{nm}$$

$$\hat{Y} = \underbrace{\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_n \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} \hat{y}_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} \dots + \beta_m x_{1m} \\ \hat{y}_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_m x_{2m} \\ \hat{y}_3 = \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} + \dots + \beta_m x_{3m} \\ \vdots \\ \hat{y}_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_m x_{nm} \end{bmatrix}}_{n \times 1}$$

We can rewrite the matrix in different form by decomposing it.

$$\hat{y} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ \vdots & & & & & \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}}_{\beta}$$

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

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

$$\therefore \boxed{\hat{y} = X\beta} \quad \text{- eq } 1$$

(This equation is true for any number of dimensions)

$$\underbrace{n \times (m+1) \longleftrightarrow (m+1) \times 1}_{n \times 1}$$

$\Downarrow$

$\hat{y}$ (shape of $\hat{y}$)

Error function for MLR

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \qquad \hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

β_0 = offset (reason)
So, if we consider this data

experience β_1	grades β_2	salary

The salary would be

$$\text{Salary} = \underbrace{\text{exp} \times \beta_1}_{} + \underbrace{\text{grades} \times \beta_2}_{} + \beta_0$$

If the value of this are zero, than still there would be some base salary. So, that's what β_0 gives

$$e = y - \hat{y} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix}$$

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

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

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

$\therefore$ $\boxed{E = e^T e}$ (Loss Function) —— eq (2)

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

$$= y^T y - \hat{y}^T y - \hat{y} y^T + \hat{y}^T \hat{y}$$

(first we will prove that this both terms are the same)

↓

(It's a tedious process, but it simply concludes that output is scalar)

$$\boxed{E = y^T y - 2 y^T \hat{y} + \hat{y}^T \hat{y}} \text{ — eq } (3)$$

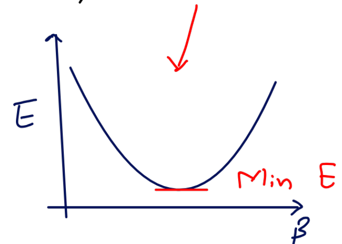
Now, replace value of $\hat{y}$ in eq (3) with eq (1)

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

$$\boxed{E = y^T y - 2 y^T X\beta + \beta^T X^T X \beta} \text{ — eq } (4) \quad \left((AB)^T = B^T A^T \right)$$

This is an important equation because it has β variable.

We have to find value of β Matrix for which E is minimum, So we will now have to differentiate eq (4) w.r.t β



Differentiating the equation

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

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

Because of β^T, the differentiation of $\beta^T X^T X \beta$ is difficult. And it can be only solved by matrix differentiation)

(But the output for this is $\underline{2 \beta^T X^T X}$)

(Note that the answer is only true when the matrix is symmetric, and here $X^T X$ is symmetric).

(Now we compare it with zero)

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

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

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

(Now to find only for β^T, we multiply $(X^T X)$ inverse on both sides to eliminate $X^T X$

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

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

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

Now Transposing both sides

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

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

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

$$\boxed{\beta = (x^T x)^{-1} x^T y} \quad \text{— eq } 5$$

$\therefore$ β values has shape of $(m+1 \times 1)$

This entire method that we used, it's called as OLS (Ordinary Least Square)

$$\left[(x^T x)^{-1}\right]^T = (x^T x)^{-1}$$

(To prove: $(x^T x)^{-1}$ = symmetric)

Assume that $x^T x = A$

$$A A^{-1} = I$$

$$(A A^{-1})^T = I^T$$

$$(A^{-1})^T A^T = I \quad \{ A^T = A$$

$$(A^{-1})^T A = I$$

$$(A^{-1})^T A A^{-1} = I A^{-1}$$

$$(A^{-1})^T I = A^{-1}$$

$$(A^{-1})^T = A^{-1}$$

$\therefore$ Proved that $(A^{-1})^T = A^{-1}$

$$\downarrow$$

$$\left((x^T x)^{-1}\right)^T = (x^T x)^{-1}$$

One of the reasons, Gradient descent is better than OLS is because in OLS, we are inversing a matrix, and in inversing a matrix, it normally takes time complexity of $\boxed{O(n^3)}$ which requires a lot of computation.