Simple Linear Regression (Matrix form)

The Simple Linear Regression (SLR) model in scaler form is represented as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \quad where \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

This can be written for each obeservation in the data

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \epsilon_2 \qquad \forall\ n \in [1, N] \text{ and } p \in [1, p] \\ \vdots \quad & \quad \vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \epsilon_n \end{aligned}$$

The same SLR model can be represented in matrix form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

which can be further broken as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ & & \vdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or simply as

$$\mathbf{y} = X\beta + \epsilon$$

where

- X is called the design matrix.
- β is the vector of coefficients.
- ϵ is the error vector.
- $\mathbf{y}$ is the response or target vector.

Distributional Assumptions in Matrix Form

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

where Σ = covariance matrix

For case of ordinary least square (OLS) where there is a constant variance for all features $\Sigma = \sigma^2 I$, distribution of error can be re-written as

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

and hence distribution of target (y) will be

$$\mathbf{y} \sim \mathcal{N}(X\beta, \sigma^2 I)$$

Therefore,

Covariance of error (ϵ)

$$\sigma_\epsilon^2 = Cov \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

Similarly, Covariance of target ($\mathbf{y}$)

$$\sigma_{\mathbf{y}}^2 = Cov \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sigma^2 I$$

Parameter Estimation

Rearranging the SLR model equation we can get residuals as

$$\epsilon = \mathbf{y} - X\beta$$

.

We want to minimize sum of squared residuals.

$$\text{minimize} \quad \sum \epsilon_i^2 = [\epsilon_1 \ \epsilon_2 \ \cdots \ \epsilon_n] \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \epsilon^T \epsilon$$

or

$$\text{minimize} \quad \epsilon^T \epsilon = (\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta)$$

To find the β which minimize above equation, the differentiation of above equation with respect to β should be equal to zero vector

i.e.

$$\frac{d}{d\beta}(\epsilon^T\epsilon) = \frac{d}{d\beta}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta) = \mathbf{0}$$
$$-2X^T(\mathbf{y} - X\beta) = \mathbf{0}$$
$$X^T\mathbf{y} = X^TX\beta$$

or

$$X^T\mathbf{y} = (X^TX)\beta$$

Left multiplying both side by $(X^TX)^{-1}$ we get

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

therefore,

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

Hat Matrix

$$\hat{\mathbf{y}} = X\beta$$
$$\hat{\mathbf{y}} = X(X^TX)^{-1}X^T\mathbf{y}$$
$$\hat{\mathbf{y}} = H\mathbf{y}$$

where $H = X(X^TX)^{-1}X^T$. We call this the "hat matrix" because it turns $\mathbf{y}$ into $\hat{\mathbf{y}}$.

We can now express residual (ϵ) in terms of hat matrix as

$$\begin{aligned}\epsilon &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - H\mathbf{y} \\ &= (I - H)\mathbf{y}\end{aligned}$$

Notice that the matrices H and $(I - H)$ have two special properties. They are

- Symmetric: $H = H^T$ and $(I - H)^T = (I - H)$.
- Idempotent: $H^2 = H$ and $(I - H)^T(I - H) = (I - H)$

Estimated Covariance Matrix of β

- β is a linear combination of the elements of **y**.
- These estimates are normal if **y** is normal.

Useful theorem

Suppose $U \sim \mathcal{N}(\mu, \Sigma)$, a multivariate normal vector, and $V = c + DU$, a linear transformation of U where c is a vector and D is a matrix. Then $V \sim \mathcal{N}(c + D\mu, D\Sigma D^T)$.

comparing this to SLR, we have

$$\begin{gathered}
U = \mathbf{y} \sim \mathcal{N}(X\beta, \sigma_\epsilon^2 I) \quad and \quad V = \beta = [(X^T X)^{-1} X^T]\mathbf{y} \\
D = (X^T X)^{-1} X^T \\
\mu = X\beta \quad and \quad \Sigma = \sigma_\epsilon^2 I \\
c = \mathbf{0} \\
V = \beta
\end{gathered}$$

Above theorem tells us the vector β is normally distributed with

$$\begin{aligned}
\text{mean} &= (X^T X)^{-1} X^T X\beta \\
&= (X^T X)^{-1} (X^T X)\beta \\
&= \beta \\
\text{Cov} &= ((X^T X)^{-1} X^T)\sigma_\epsilon^2 I((X^T X)^{-1} X^T)^T \\
&= \sigma_\epsilon^2 ((X^T X)^{-1} X^T) I((X^T X)^{-1} X^T)^T \\
&= \sigma_\epsilon^2 (X^T X)^{-1} X^T ((X^T X)^{-1})^T X \\
&= \sigma_\epsilon^2 (X^T X)^{-1} (X^T X)((X^T X)^{-1}) \\
&= \sigma_\epsilon^2 (X^T X)^{-1}
\end{aligned}$$

using the fact that both $X^T X$ and its inverse are symmetric, so $((X^T X)^{-1})^T = (X^T X)^{-1}$

Hence,

$$\beta \sim \mathcal{N}(\beta, \sigma_\epsilon^2 (X^T X)^{-1})$$

Therefore, standard deviation of estimates (β) = $\sqrt{\sigma_\epsilon^2 (X^T X)^{-1}}$