Recalling regression model is represented mathematically as

$$Y_i = \beta^T X_i + E_i, \quad i = 1, 2, \dots, n,$$

where Y_i, X_i, β_i, and E_i, are our response, predictor, parameter, and error of the i^{th} sample, respectively. In matrix form, the above can be expressed as

$$Y = X\beta + E,$$

where

$$Y = \begin{pmatrix} Y_1^T \\ \vdots \\ Y_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}$$

$$X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} \in \mathbb{R}^{n \times d}$$

$$E = \begin{pmatrix} E_1^T \\ \vdots \\ E_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}$$

and

$$\beta = \begin{pmatrix} \beta_1^T \\ \vdots \\ \beta_n^T \end{pmatrix} \in \mathbb{R}^{d \times p}$$

which is the multivariate case (i.e., β is a $d - vector$ in univariate case).

From this, let's look at the solution for minimizing the above via $L^2 - norm$:

$$\sum_{i=1}^{n} \|Y_i - \beta^T X_i\|^2 = \sum_{i=1}^{n} (Y_i - \beta^T X_i)^T (Y_i - \beta^T X_i) = trace\big((Y_i - \beta^T X_i)^T (Y_i - \beta^T X_i)\big)$$

Let's differentiate wrt β, then set equal to 0, while using the following matrix rules:

$$\frac{\partial trace(A)}{\partial A} = I$$

$$\frac{\partial (A^T Z A)}{\partial A} = 2AZ$$

for any square matrices *A* and compatible matrix *Z*. Thus, we can derive the normal equation:

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