Multi - linear regression.

$$RSS = \| \vec{y} - \hat{y} \|_2^2 \qquad \boxed{\hat{y} = X\hat{\beta}}$$

$$= \| \vec{y} - X\hat{\beta} \|_2^2$$

$$\min_{\hat{\beta}} \| \vec{y} - X\hat{\beta} \|_2^2 = \min_{\beta} (\vec{y} - X\hat{\beta})^T (\vec{y} - X\hat{\beta})$$

(inner product of a vector and itself. is the square of L2 norm)

$$\Rightarrow \frac{\partial}{\partial \hat{\beta}} RSS = 0 \quad \text{find optimum.}$$

$$RSS = (\vec{y} - X\hat{\beta})^T (\vec{y} - X\hat{\beta})$$

$$= \vec{y}^T \vec{y} - (X\hat{\beta})^T \vec{y} - \vec{y}^T (X\hat{\beta}) + (X\hat{\beta})^T (X\hat{\beta})$$

note that $\vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$ is a vector.

$X = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_p^{(1)} \\ \vdots & & & \\ 1 & x_1^{(n)} & \cdots & x_p^{(n)} \end{bmatrix}$ is a $(n \times (p+1))$ matrix

$\hat{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}$ is a $(p+1)$ vector.

$\therefore$ $\vec{y}^T \vec{y}$ is a scalar

$X\hat{\beta}$ is a n- vector

$(X\hat{\beta})^T \vec{y}$ is a scalar.

$(\hat{y}^T)(X\hat{\beta})$ is the inner product of $\hat{y}$ and $X\hat{\beta}$, which is a scalar.

$(X\hat{\beta})^T(X\hat{\beta})$ is a scalar.

RSS $\rightarrow$ scalar function of $\hat{\beta}$

$\frac{\partial}{\partial\hat{\beta}}RSS$ in steps:

$$RSS = \underbrace{\vec{y}^T\vec{y}}_{①} - \underbrace{(X\hat{\beta})^T\vec{y}}_{②} - \underbrace{\vec{y}^T(X\hat{\beta})}_{③} + \underbrace{(X\hat{\beta})^T(X\hat{\beta})}_{④}$$

$$\frac{\partial ①}{\partial\hat{\beta}} = \frac{\partial(\vec{y}^T\vec{y})}{\partial\hat{\beta}} = 0$$ (not depend on $\hat{\beta}$)

$$\frac{\partial (X\hat{\beta})^T\vec{y}}{\partial\hat{\beta}} = \frac{\partial \vec{y}^T(X\hat{\beta})}{\partial\hat{\beta}}$$ (because $(X\hat{\beta})^T\vec{y} = \vec{y}^T(X\hat{\beta})^T$) commutative of inner product.

$$= y^TX$$

[☆ matrix differetiation: (proposition 7)

$$\alpha = y^TAx, \quad \frac{\partial\alpha}{\partial x} = y^TA.$$]

Note that ② = ③.

$$\frac{\partial}{\partial\hat\beta}(X\hat\beta)^T(X\hat\beta) = (X\hat\beta)^T\frac{\partial(X\hat\beta)}{\partial\hat\beta} + (X\hat\beta)^T\frac{\partial(X\hat\beta)}{\partial\hat\beta}$$

$$= 2(X\hat\beta)^T X = 2\hat\beta^T(X^TX)$$

Property 10: $\alpha = y^T x$: $\frac{\partial\alpha}{\partial z} = x^T\frac{\partial\alpha}{\partial z} + y^T\frac{\partial\alpha}{\partial z}$.

putting every thing together:

$$\frac{\partial RSS}{\partial\hat\beta} = 0 - 2y^Tx + 2\hat\beta^T(X^TX) = 0$$

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

Transpose $\quad \hat\beta = \left((y^TX)\cdot(X^TX)^{-1}\right)^T$

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

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

Here: X^TX is a symmetric matrix.

$\therefore (X^TX)^{-1}$ is also symmetric

$$\therefore \left((X^TX)^{-1}\right)^T = (X^TX)^{-1}$$

The transpose of a symmetric matrix is itself.