$X = \begin{bmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_n^T \end{bmatrix}$ size $(n, k+1)$ all "i"s for β_0

intercept

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

x_i : sample i $\begin{bmatrix} x_{i0} \\ \vdots \\ x_{ik} \end{bmatrix}$, x_i^T transpose to row vec

$\hat{\beta}$ size $(k+1, 1)$: k features plus 1 intercept

$(n, k+1)(k+1) \Rightarrow (n, 1)$ pred.

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

y : labels in dataset also $(n, 1)$

$\Rightarrow e = y - \hat{y}$

$e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$ → residual for sample $1 = y_i - \hat{y}_i$

$RSS = e^T e$

since $(1, k+1) \times (k+1, 1) = (1, 1)$ residual

$\min RSS = \min e^T e$

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

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

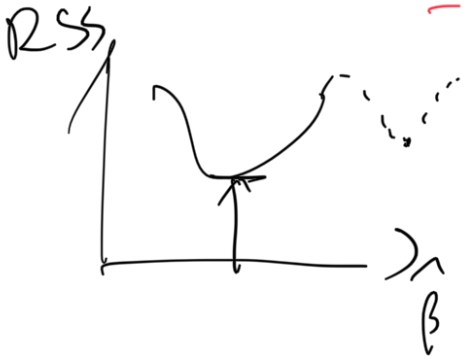
$$(y - X\beta)^T (y - X\beta)$$

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

$$\underbrace{y^T y}_{①} - \underbrace{y^T X \beta}_{②} - \underbrace{\hat{\beta}^T X^T y}_{③} + \underbrace{\hat{\beta}^T X^T X \hat{\beta}}_{④}$$

(1.1) (1.1)

(①n) (n, k+1) (k+1, ①)



$$\frac{\partial e^T e}{\partial \hat{\beta}} = 0$$

① $\frac{\partial y^T y}{\partial \hat{\beta}} = 0$

②+③: $\frac{\partial (② + ③)}{\partial \hat{\beta}} = \frac{\partial (-2 y^T X \hat{\beta})}{\partial \hat{\beta}} = -2 X^T y$

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

② · ③ are 1×1 and thus symmetric

symmetric is equal to its transpose

$\frac{\partial v^T a}{\partial a} = \frac{\partial a^T v}{\partial a}$

Rule 1: $\frac{\partial v}{\partial v} = 1$

$$= -2 \frac{\partial\, y^T X \hat{\beta}}{\partial \hat{\beta}} \qquad y^T X = a^T$$

$$a = X^T y$$

(4) $$\frac{\partial\, \hat{\beta}^T X^T X \hat{\beta}}{\partial \hat{\beta}} = 2 X^T X \hat{\beta}$$

$X^T X = A$

Rule 2: $\frac{\partial\, v^T A v}{\partial v} = 2Av$

$(X^T X)^{-1}$

{ Rank $= k$: independent features

$k \leq n$: # features $\leq$ # samples