lee #7

⓪ let $\vec{x} \in \mathbb{R}^n$. Let $a \in \mathbb{R}$ be a constant w.r.t. $\vec{x}$

$$\Rightarrow \frac{\partial}{\partial \vec{x}}[a] = \vec{0}_n$$

now let $\vec{a} \in \mathbb{R}^n$ be constant w.r.t. $\vec{x}$

① $$\frac{\partial}{\partial \vec{x}}[\underbrace{\vec{a}^T\vec{x}}_{\vec{x}^T\vec{a}}] = \begin{bmatrix} \partial/\partial x_1 [a_1x_1 + a_2x_2 + \cdots + a_nx_n] \\ \\ \\ \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \vec{a} \neq \vec{a}^T$$

let $a, b \in \mathbb{R}$ be constants w.r.t. $\vec{x}$

② $$\frac{\partial}{\partial \vec{x}}[af(\vec{x}) + bg(\vec{x})] = \begin{bmatrix} \partial/\partial x_1 [af(\vec{x}) + bg(\vec{x})] \\ \vdots \\ \end{bmatrix} = \begin{bmatrix} a\frac{\partial}{\partial x_1}[f(\vec{x})] + b\frac{\partial}{\partial x_1}[g(\vec{x})] \\ \vdots \\ \end{bmatrix}$$

$$= a\frac{\partial}{\partial \vec{x}}[f(\vec{x})] + b\frac{\partial}{\partial \vec{x}}[g(\vec{x})]$$

let $A \in \mathbb{R}^{n \times n}$ be symmetric, constant w.r.t. $\vec{x}$

③ $\frac{\partial}{\partial \vec{x}}[\vec{x}^TA\vec{x}]$ This scalar expression $\vec{x}^TA\vec{x}$ is called a "quadratic form"

and it's a common expression and very-well studied.

$$A\vec{x} = \begin{bmatrix} \leftarrow \vec{a}_1 \rightarrow \\ \leftarrow \vec{a}_2 \rightarrow \\ \vdots \\ \leftarrow \vec{a}_n \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow \\ \vec{x} \\ \downarrow \end{bmatrix} = \begin{bmatrix} \vec{a}_1 \cdot \vec{x} \\ \vec{a}_2 \cdot \vec{x} \\ \vdots \\ \vec{a}_n \cdot \vec{x} \end{bmatrix} = \begin{bmatrix} a_{11}\vec{x}_1 + a_{12}\vec{x}_2 + \cdots a_{1n}\vec{x}_n \\ a_{21}\vec{x}_2 + a_{22}\vec{x}_2 + \cdots a_{2n}\vec{x}_n \\ \vdots \\ a_{n1}\vec{x}_n + a_{n2}\vec{x}_2 + \cdots a_{nn}\vec{x}_n \end{bmatrix}$$

$$\vec{x}^T(A\vec{x}) = [x_1, x_2, \cdots, x_n] \begin{bmatrix} \vec{a}_1 \cdot \vec{x} \\ \vec{a}_2 \cdot \vec{x} \\ \vdots \\ \vec{a}_n \cdot \vec{x} \end{bmatrix} = x_1 a_1 \cdot \vec{x} + x_2 \vec{a}_2 \cdot \vec{x} + \cdots + x_n \vec{a}_n \vec{x}$$

$$= x_1(a_{11}\vec{x}_1 + a_{12}\vec{x}_2 + \cdots + a_{1n}\vec{x}_n) + x_2(a_{21}\vec{x}_1 + a_{22}\vec{x}_2 + \cdots + a_{2n}\vec{x}_n) + \cdots$$
$$+ x_n(a_{n1}\vec{x}_1 + a_{n2}\vec{x}_2 + \cdots + a_{nn}\vec{x}_n)$$

$$\frac{\partial}{\partial x_1}[\;] = 2a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n + a_{21}x_2 + \cdots + a_{n1}x_n = 2a_{11}x_1 + 2a_{12}x_2 + \cdots + 2a_{1n}x_n$$

$$= 2\vec{a}_{1,}\vec{x}$$

$$\frac{\partial}{\partial x_2}[\ ] = a_{12}x_1 + a_{21}x_1 + 2a_{22}x_2 + \cdots + a_{2n}x_n + \cdots + a_{n2}x_n = 2a_{12}x_1 + 2a_{22}x_2 + 2a_{n2}x_n$$
$$= 2\vec{a}_{2\cdot}\cdot\vec{x}$$

$$\vdots$$

$$\frac{\partial}{\partial \vec{x}}[\vec{x}^T A \vec{x}] = \begin{bmatrix} 2\vec{a}_{1\cdot}\cdot\vec{x} \\ \vdots \\ 2\vec{a}_{n\cdot}\cdot\vec{x} \end{bmatrix} = 2A\vec{x}$$

■

$$\frac{\partial}{\partial \vec{w}}[\vec{y}^T\vec{y} - 2\vec{w}^T X^T \vec{y} + \vec{w}^T X^T X \vec{w}] \overset{②}{=} \underbrace{\cancel{\frac{\partial}{\partial \vec{w}}[\vec{y}^T\vec{y}]}}_{\text{by ①}} - 2\frac{\partial}{\partial \vec{w}}[\vec{w}(\vec{x}^T\vec{y})] + \frac{\partial}{\partial \vec{w}}[\vec{w}^T(\vec{x}^T\vec{x})\vec{w}]$$
$$\overset{①}{=} -2X^Ty + \frac{\partial}{\partial \vec{w}}[\vec{w}^T(\vec{x}^T\vec{x})\vec{w}] \overset{③}{=} -2X^Ty + 2X^TX\vec{w}$$

set $\xrightarrow{}$ $\vec{0}_{p+1}$ and solve for $\vec{b}$

$$\Rightarrow -X^Ty + X^TX\vec{w} = 0$$
$$\cancel{(X^TX)^{-1}X^TX}\vec{w} = (X^TX)^{-1}X^Ty$$
$$\boxed{\vec{b} = (X^TX)^{-1}X^Ty}$$

In order to compute the OLS coefficients ($\vec{b}$) you need $\underbrace{X^TX}_{a\,(p+1)\times(p+1)}$ to be invertible,

$$\Rightarrow \underbrace{\vec{\hat{y}}_a = y(\vec{x}_a) = \vec{x}_a\vec{b}}_{\text{predictions}}$$

Equivalently, rank $[X^TX] = p+1$, i.e. "full rank" i.e. all columns are linearly independent. Since there's a thm

rank $[X^TX]$ = rank$[X]$, this means rank $[X] = p+1$

i.e. the columns of X are linearly independent.

$$X = \begin{bmatrix} 1 & \uparrow & \uparrow & & \uparrow \\ 1 & \vec{x}_{\cdot 1} & \vec{x}_{\cdot 2} & \cdots & \vec{x}_{\cdot p} \\ \vdots & & & & \\ 1 & \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

feature measurements on all n subjects.

If X is full rank, that means ... there is no exact data duplication eg. x_1 = height measured in inches, and x_2 = height measured in centimeters. What if you do have a feature that is linearly dependent with other features in X? You just drop it. Then X will be full rank and you're good to estimate the OLS coefficients.

$$\vec{y} = \underset{X\vec{b}}{\underset{\shortparallel}{\hat{\vec{y}}}} + \vec{e} \Rightarrow \vec{e} = \vec{y} - \hat{\vec{y}}, \quad SSE = \sum_{i=1}^{n} e_i^2 = \vec{e}^{\,T}\vec{e}$$

$$MSE = \frac{1}{n-(p+1)} SSE, \quad RMSE = \sqrt{MSE}, \quad R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} = \frac{s_y^2 - s_e^2}{s_y^2}$$

You sometimes say the model has $p+1$ "degrees of freedom"
(i.e. the number of parameters, $w_0, w_1, \cdots, w_p$, is $p+1$)
and $p+1 = \dim[\text{column\_space}[X]]$