Lecture 8   2/24/21

$$SSE = \overbrace{\vec{y}^\top}^{(1\times n)}\overbrace{\vec{y}}^{(1\times n)} - 2\overbrace{\vec{w}^\top}^{1\times(n+1)}\overbrace{X^\top}^{(p+1)\times n}\underbrace{\vec{y}}_{(n\times 1)} + \underbrace{\vec{w}^\top}_{1\times(p+1)}\overbrace{X^\top}^{(p+1)\times n}\underbrace{X}_{n\times(p+1)}\overbrace{\vec{w}}^{(p+1)}$$

$$\frac{\partial SSE}{\partial \vec{w}} := \begin{bmatrix} \frac{\partial SSE}{\partial w_0} \\ \frac{\partial SSE}{\partial w_1} \\ \vdots \\ \frac{\partial SSE}{\partial w_p} \end{bmatrix} \overset{set}{=} \vec{0}_{p+1} \text{ and solve for } b_0, b_1, \dots, b_p$$

Let $\vec{x} \in \mathbb{R}^n$. Let $a \in \mathbb{R}$ be a constant with respect to $\vec{x}$. $\Rightarrow$

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

Let $\vec{a} \in \mathbb{R}^n$ constant with respect to $\vec{x}$

$$\frac{\partial}{\partial \vec{x}}\Big[\underset{\vec{x}^\top\vec{a}}{\overset{\|}{\vec{a}^\top\vec{x}}}\Big] = \begin{bmatrix} \frac{\partial}{\partial x_1}[a_1x_1 + a_2x_2 + \dots + a_nx_n] \\ \\ \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \vec{a} \neq \vec{a}^\top \quad 1$$

Let $a, b \in \mathbb{R}$ constants with respect to $\vec{x}$.

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

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

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

$$\frac{\partial}{\partial \vec{x}}\left[\vec{x}^\top A\vec{x}\right],\ A\vec{x} = \begin{bmatrix} \leftarrow \vec{a}_1 \rightarrow \\ \leftarrow \vec{a}_2 \rightarrow \\ \vdots \\ \leftarrow \vec{a}_n \rightarrow \end{bmatrix}\begin{bmatrix} \uparrow \\ \vec{x} \\ \downarrow \end{bmatrix} = \begin{bmatrix} \vec{a}_1 \cdot \vec{x} \\ \vec{a}_2 \cdot \vec{x} \\ \vdots \\ \vec{a}_n \cdot \vec{x} \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n \end{bmatrix} \quad 3$$

This scalar expression, $x^\top Ax$ is called a "quadratic form" and it's a common expression and very well-studied.

$$\vec{x}^\top(A\vec{x}) = [x_1 x_2 \dots x_n]\begin{bmatrix}\vec{a}_{1\cdot}\vec{x}\\ \vec{a}_{2\cdot}\vec{x}\\ \vdots \\ \vec{a}_{n\cdot}\vec{x}\end{bmatrix} = x_1\vec{a}_{1\cdot}\vec{x} + x_2\vec{a}_{2\cdot}\vec{x} + \dots + x_n\vec{a}_{n\cdot}\vec{x}$$

$$= x_1(a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n) + x_2(a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n) + \dots + x_n(a_{n1}x_1 + a_{n2}x_2 + \dots a_{nn}x_n)$$

$$\frac{\partial}{\partial x_1}[\;] = (2a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n) + a_{21}x_2 + \dots + a_{n1}x_n = \underbrace{2a_{11}x_1 + 2a_{12}x_2 + \dots + 2a_{1n}x_n}_{= 2\vec{a}_{1\cdot}\vec{x}}$$

a is symmetric thus $a_{12} = a_{21}$, $a_{1n} = a_{n1}$

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

$\vdots$

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

$$\frac{\partial}{\partial\vec{w}}\Big[\overbrace{\vec{y}^\top\vec{y} - 2\vec{w}^\top X^\top\vec{y} + \vec{w}^\top X^\top X\vec{w}}^{SSE}\Big] \overset{\text{rule \#2}}{=} \frac{\partial}{\partial\vec{w}}\big[\cancel{\vec{y}^\top\vec{y}}\big]_{0 \text{ by rule \#0}} - 2\frac{\partial}{\partial\vec{w}}\big[\vec{w}^\top(X^\top\vec{y})\big] + \frac{\partial}{\partial\vec{w}}\big[\vec{w}^\top X^\top X\vec{w}\big] =$$

$$\overset{\text{by rule \#1}}{=} -2X^\top\vec{y} + \frac{\partial}{\partial\vec{w}}\big[\vec{w}^\top(X^\top X)\vec{w}\big] \overset{\text{by rule \#3}}{=} -\cancel{2}X^\top\vec{y} + \cancel{2}\overbrace{X^\top X}^{(X^\top X)^\top = X^\top(X^\top)^\top = X^\top X \Rightarrow \text{symmetric}}\vec{w} \overset{\text{set}}{=} \vec{0}_{p+1}$$ and solve for $\vec{b}$.

$$\Rightarrow \underbrace{\cancel{(X^\top X)^{-1}X^\top X}}_{I_{p+1}}\vec{w} = (X^\top X)^{-1}X^\top\vec{y} \Rightarrow \boxed{\vec{b} = (X^\top X)^{-1}X^\top\vec{y}} \Rightarrow \underbrace{\hat{y}_* = g(\vec{x}_*) = \vec{x}_*\vec{b}}_{\text{predictions}}$$

$A\vec{x} = \vec{b}$

In order to compute the OLS coefficients (vector b), you need $X^\top X$, a $(p+1)\times(p+1)$ square matrix, to be invertible. Equivalently, rank$[X^\top X] = p+1$ i.e. "full rank" i.e. all columns of $X^\top X$ are linearly independent. Since there's a thm: rank$[X^\top X]$ = rank$[X]$, this means rank$[X] = p+1$, i.e. the columns of X are linearly independent.

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

feature measurements of all n subjects.

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

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

$$MSE = \frac{1}{n-(p+1)} SSE, \quad RMSE = \sqrt{MSE},$$

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} = \frac{s_y^2 - s_e^2}{s_y^2} \text{ (same).}$$

you sometimes say the model has $p+1$ ~~"degrees of~~ "degrees of freedom" (i.e. the number of parameters; $w_0, w_1, \ldots, w_p$ is $p+1$) and $p+1 = \dim[\text{colsp}[X]]$.