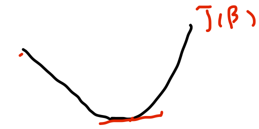
This lecture: estimation of $\underset{(p+1)\times 1}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$

Goal: to find $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}_{(p+1)\times 1}$ such that it minizes the RSS function



$$\underbrace{J(\beta)}_{\text{RSS}} = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 .$$

From calculus, the minimizer of $J(\beta)$ is achieved when

$$\nabla_\beta J(\beta) = 0$$

that is, $$\nabla_\beta J(\beta)\Big|_{\beta=\hat{\beta}} = 0 .$$ $*$ (← $(p+1)$ equations)

<u>Method 1</u>:

$$\underbrace{\nabla_\beta J(\beta)}_{(p+1)\times 1 \text{ dim}} = \nabla_\beta \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 .$$

$x_i : (p+1)\times 1$

$$= \begin{pmatrix} 1 \\ x_{i,1} \\ \vdots \\ x_{i,p} \end{pmatrix}$$

$$= \sum_{i=1}^{n} \nabla_\beta (y_i - x_i^T \beta)^2$$

$$= \sum_{i=1}^{n} 2 \cdot (y_i - x_i^T \beta) \cdot \underbrace{\nabla_\beta (y_i - x_i^T \beta)}_{= -x_i}$$

$$= \underbrace{-2 \sum_{i=1}^{n} (y_i - x_i^T \beta) x_i}_{} . \quad \leftarrow (p+1)\times 1$$

We know from the above disussion that $\hat{\beta}$ satisfies

$$\cancel{-2} \sum_{i=1}^{n} (y_i - x_i^T \hat{\beta}) x_i = 0$$

Use the matrix notation, we have

$$\sum_{i=1}^{n} (y_i - x_i^T \hat{\beta}) x_i = \underbrace{\underset{(p+1)\times n}{[x_1, x_2, \dots\dots, x_n]}}_{} \begin{bmatrix} y_1 - x_1^T \hat{\beta} \\ y_2 - x_2^T \hat{\beta} \\ \vdots \\ y_n - x_n^T \hat{\beta} \end{bmatrix}_{n\times 1}$$

$\uparrow$ $(p+1)\times 1$

$$X^T$$

Recall that
$$\underset{n\times(p+1)}{X} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{np} \end{pmatrix}$$

$$= X^T \cdot \left[\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{Y_{n\times 1}} - \underbrace{\begin{pmatrix} x_1^T\hat\beta \\ \vdots \\ x_n^T\hat\beta \end{pmatrix}}_{=\begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}\hat\beta = \underset{n\times(p+1)}{X}\,\hat\beta_{(p+1)\times 1}} \right]$$

$$= \underbrace{X^T(Y - X\hat\beta) = 0_{(p+1)\times 1}}_{\text{normal equations.}}$$

So $\quad X^TY - X^TX\hat\beta = 0$

$$\Rightarrow \underbrace{(X^TX)}_{(p+1)\times(p+1)}\hat\beta_{(p+1)\times 1} = \underbrace{X^TY}_{(p+1)\times 1}$$

$$\Rightarrow \hat\beta = (X^TX)^{-1}X^TY \quad \text{if } \underline{(X^TX)^{-1} \text{ exists}}$$

OLS estimator

$\text{rank}(X^TX) = (p+1)$.

E.g. • some ~~two~~ columns of $X_{n\times(p+1)}$ are the same $\Rightarrow \text{rank}(X^TX) < (p+1)$

"colinearity" issue. $\longrightarrow$ or a column of X is a linear combination of other columns.

E.g.: $\underline{Y_i = x_i} \rightarrow \beta_0 = 0.\ \beta_1 = 1$

$Y_i = \beta_0 + \beta_1 \underset{1\dim}{\underline{x_i}} + \varepsilon_i \quad i = 1, \cdots, n$

$$\underset{n\times 2}{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

not linearly dependent. if $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \neq \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

high-dimension data • If $p > n$
$\text{rank}(X^TX) \le n < p$

Method 2

We first rewrite $J(\beta)$ using matrix notation.

$$J(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

$$= \underbrace{(y_1 - x_1^T\beta, \ \cdots, \ y_n - x_n^T\beta)_{1\times n}}_{(y - X\beta)^T} \begin{pmatrix} y_1 - x_1^T\beta \\ \vdots \\ y_n - x_n^T\beta \end{pmatrix}_{n\times 1}$$

as shown in Method 1

$$y_{n\times 1} - X_{n\times(p+1)} \beta_{(p+1)\times 1}.$$

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

$Tr(AA^T) = Tr(A^TA)$

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

$$\nabla_\beta J(\beta) = \nabla_\beta \left[Tr\{(y - X\beta)^T (y - X\beta)\}\right]$$

"B" in ②.

$$\beta^T \overbrace{X^TX}^{} \beta$$

$$= \nabla_\beta \left[Tr\left(y^Ty - \underbrace{\overset{y^TX\beta}{\overset{\parallel}{(X\beta)^Ty}}}_{\text{scalar } (1\times 1)} - \underbrace{y^TX}_{\text{"A" in ①}}\beta + \overset{\beta^TX^TX\beta}{\overset{\parallel}{(X\beta)^T(X\beta)}}\right)\right]$$

Recall that from linear alg file

(page 23-24 § 4.3)

① $\nabla_\beta Tr(A\beta) = A^T$ for any A matrix

② $\nabla_\beta Tr\left(\beta^T B \beta\right) = 2B\beta$

for symmetric matrix B.

These results hold generally for

β being vectors or matrix

(HW1. Q5)

$$= 0 - 2\cdot(y^TX)^T$$

$$+ 2(X^TX)\beta.$$

$$= -2X^Ty + 2(X^TX)\beta.$$

Therefore, $\hat{\beta}$ satisfies

$$-\cancel{2}X^Ty + \cancel{2}(X^TX)\hat{\beta} = 0$$

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

iff $(X^TX)^{-1}$ exists.

Hints: • Q5. HW. multivariate response $y_i \in \mathbb{R}^m$. [OLS. $m=1$]

$$\underset{1\times m}{y_i^T} = \underset{1\times d}{x_i^T} \underset{d\times m}{B} + \text{error} \qquad (d = p+1)$$

(a) $$RSS(B) = \sum_{i=1}^{n} (y_i^T - x_i^TB)(y_i^T - x_i^TB)^T$$

Hint: For $A = \begin{bmatrix} -a_1- \\ \vdots \\ -a_n- \end{bmatrix}$ $\quad Tr(AA^T) = \sum_{i=1}^{n} a_i a_i^T$

$RSS(B) = Tr(AA^T)$

with $a_i = y_i^T - x_i^TB$

(b). find $\hat{B}$ s.t $\nabla_B RSS(B) = 0$

$$\nabla_B Tr\left[(Y - XB)(Y-XB)^T\right] = 0$$

$$A = \begin{bmatrix} y_1^T - x_1^T B \\ \vdots \\ y_n^T - x_n^T B \end{bmatrix} = \begin{bmatrix} y_1^T \\ \vdots \\ y_n^T \end{bmatrix} - \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} B = Y_{n\times m} - X_{n\times d} B_{d\times m}$$

Example Simple linear Regression $(p=1)$ note that now $x_1 \cdots x_n \in \mathbb{R}$

(Q4 Hw1)

$$\underset{n\times 1}{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \underset{n\times 2}{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \underset{2\times 1}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

To find OLS $\hat{\beta}$.

Following Method 1, we can have the normal equations.

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\begin{pmatrix} 1 \\ x_i \end{pmatrix} = 0$$

$$\Rightarrow \begin{cases} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad ⓐ \leftarrow \dfrac{\partial J(\beta)}{\partial \beta_0} = 0 \\ \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \quad ⓑ \leftarrow \dfrac{\partial J(\beta)}{\partial \beta_1} = 0 \end{cases}$$

Eq (a) $\overset{\text{(practice)}}{\Longrightarrow}$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

(a) + (b) $\overset{\text{(practice)}}{\Longrightarrow}$ $\hat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

Note that $\hat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})/n-1}{\sum_{i=1}^n (x_i - \bar{x})^2 / n-1}$ ← Sample covariance of x and y ← Sample variance of x

(practice)
$= \text{Sample correlation}(x, y) \cdot \dfrac{\text{Sample Standard deviation }(Y)}{\text{Sample Standard dev }(X)}$.