Lecture September 3

$$\hat{\beta}_{OLS} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \|(y - X\beta)\|_2^2$$

$$\left\{ \begin{array}{l} \|x\|_2^2 = \sum_{i=1}^{n} x_i^2 \\ \|x\|_1 = \sum_{i=1}^{n} |x_i| \end{array} \right.$$

$$\hat{\beta}_{OLS} = \underbrace{(X^T X)^{-1} X^T}\, y$$

$$\beta = \lim_{\lambda \to 0} (X^T X + \lambda \underline{I_p})^{-1} X^T$$

$X \in \mathbb{R}^{n \times p}$

$y \in \mathbb{R}^n \qquad \beta \in \mathbb{R}^p$

$\Rightarrow$ Moore – Penrose pseudo inverse.

<u>SVD</u>

$X = U \Sigma V'$

$U \in \mathbb{R}^{n \times n}$ $\qquad U^T U = U U^T = I_n$

$\Sigma \in \mathbb{R}^{n \times p}$

$V \in \mathbb{R}^{p \times p}$ $\qquad V V^T = V^T V = I_p$

$n \geq p$

example

$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

σ_2

$\sigma_3 = 0$

$\sigma_1 > \sigma_2 > \sigma_3$

$\sigma_0 > \sigma_1 > \dots > \sigma_{p-1} \geq 0$

$\Sigma \Sigma^T \neq \Sigma^T \Sigma$

$$\Sigma^T \Sigma = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$$= \boxed{\begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}}$$

$$\Sigma \Sigma^T = \boxed{\begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}}$$

$$\underline{X^T X} = V \Sigma^T \underbrace{U^T U}_{I_n} \Sigma V^T$$

$[A, D] = 0$

$$V \boxed{\Sigma^T I_n \Sigma} V^T$$

$$= \underline{\Sigma^T \Sigma} = \Sigma^2 \quad (p \times p)$$

$$V V^T = V^T V = I_p$$

$$\tilde{y}_{ocs} = X \hat{\beta}_{ocs}$$

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

$$= U \Sigma \underline{V}^T \left(\frac{1}{\Sigma \cancel{\Sigma}} \right) \underline{V} \Sigma^T U^T y$$

$$= U U^T y = \left(\sum_{i=0}^{n-1} u_i u_i^T \right) y$$

$$U = [u_0\ u_1\ \ldots\ u_{n-1}]$$

$$u_i^T u_j = \delta_{ij}$$

$$X^T X = \Sigma^T \Sigma = \Sigma^2 \in \mathbb{R}^{p \times p}$$

$$= \begin{bmatrix} \sigma_0^2 & & & 0 \\ & \sigma_1^2 & & \\ & & \ddots & \\ 0 & & & \sigma_{p-1}^2 \end{bmatrix}$$

$$\underline{(X^T X) V = \Sigma^2 V}$$

$$(X^T X) v_i = \sigma_i^2 v_i$$

$$V = [v_0, v_1 \ldots v_{p-1}]$$

$$X X^T = U \Sigma V^T V \Sigma^T U^T =$$

$$U (\Sigma \Sigma^T) U^T$$

$$X U$$

$$(XX^T)u = u(\Sigma\Sigma^T)$$

u are the eigenvectors of XX^T

X^TX has σ_i^2 has eigenvalues.

(1) $$C(\beta) = \frac{1}{n} \|(y - X\beta)\|_2^2$$

$$\frac{\partial C}{\partial \beta} = 0 = X^T(y - X\beta)$$

$$\frac{\partial^2 C}{\partial \beta^T \partial \beta} = \boxed{X^T X}$$

eigenvalues = singular values $\sigma_0 > \sigma_1 > \dots \sigma_{p-1}$

> 0

(ii) sample values

$$\mathbb{E}[x] = \frac{1}{n} \sum_{i=0}^{n-1} x_i = \bar{\mu}_x$$

$$\left(= \sum_{i=0}^{n-1} p(x_i) x_i = \mu_x \right)$$

$$var(x) = \frac{1}{n} \sum_{i=0}^{n-1} (x_i - \bar{\mu}_x)^2$$

$$\underline{cov(x,y)} = \underline{\frac{1}{n} \sum_{i=0}^{n-1} (x_i - \bar{\mu}_x)(y_i - \bar{\mu}_y)}$$

$$p(x_1 x_2) = p(x_1)\, p(x_2)$$

$$cov(x,y) = \iint dx\,dy \times (x - \mu_x)(y - \mu_x)\, p(x)\, p(y) = 0$$

$$cov(x,y) \sim \underline{X^T X}$$

$$(cov(x))$$

<u>covariance matrix</u>

$$X = \begin{bmatrix} x_0 & x_1 & \dots & x_{p-1} \end{bmatrix}$$

$$x_i = \begin{bmatrix} x_{0i} \\ x_{1i} \\ \vdots \\ x_{n-1\,i} \end{bmatrix}$$

scaling of matrix

$$\bar{x}_i = \begin{bmatrix} x_{0i} - \bar{\mu}_{x_i} \\ x_{1i} - \bar{\mu}_{x_i} \\ \vdots \\ x_{n-1\,i} - \bar{\mu}_{x_i} \end{bmatrix}$$

$$X^T X \rightarrow \bar{X}^T \bar{X} \quad \left(= X^T X \right)$$

$$\bar{x}_i^T \bar{x}_j$$

$$= \boxed{\sum_{l=0}^{n-1} (x_{li} - \bar{\mu}_{x_i})(x_{lj} - \bar{\mu}_{x_j})}$$

$$= n \cdot \operatorname{cov}(\bar{x}_i, \bar{x}_j)$$

$$X^T X = n \cdot \underline{\mathrm{cov}(X)}$$

$$= \begin{bmatrix} ? & \mathrm{cov}(\bar{x}_0, \bar{x}_1) & \mathrm{cov}(\bar{x}_0, \bar{x}_2) & \cdots \\ \vdots & ? & & \\ & & ? & \\ \mathrm{cov}(\bar{x}_{p-1}, \bar{x}_0) & & & ? \end{bmatrix}$$

is this a symmetric matrix? = Yes

Diagonal = $\mathrm{var}(x_i)$

$$\mathrm{corr}(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}}$$

$$\in [-1, +1]$$

$$\Rightarrow \operatorname{corr}(X) =$$

$$\begin{bmatrix} 1 & \operatorname{corr}(\bar{x}_0, \bar{x}_1) & \cdots & \cdots \\ & 1 & & \\ & & \ddots & \\ \operatorname{corr}(\bar{x}_{p-1}, x_0) & \cdots & \cdots & 1 \end{bmatrix}$$

$$\operatorname{cov}(X) = \frac{1}{n} X^T X$$

Linear regression,
2nd derivative of
MSE = Hessian $= X^T X$

(iii)

$$\operatorname{var}(\beta) \propto (X^T X)^{-1}$$

Ridge Regression

$$\hat{\beta}_{Ridge} = \underset{\beta \in \mathbb{R}^P}{\operatorname{argmin}}$$

$$\frac{1}{\,}\|(y - X\beta)\|_2^2$$

$$\underbrace{\qquad m \qquad}_{MSE}$$

$$+ \underbrace{\lambda \sum_{j=0}^{p-1} \beta_j^2}_{\lambda \|\beta\|_2^2}$$ — Regularization penalty

$$\underbrace{\lambda \|\beta\|_2^2}_{\lambda \beta^T \beta}$$

$$\frac{\partial C}{\partial \beta^T} = -\frac{2}{m} X^T (y - X\beta) + \lambda 2 \beta = 0$$

$$X^T X \beta + \lambda \beta = X^T y \Rightarrow$$

$$\hat{\beta}_{Ridge} = \left(X^T X + \boxed{\lambda I_p}\right)^{-1} X^T y$$

$\lambda > 0$ Hyperparameter

Lasso Regression

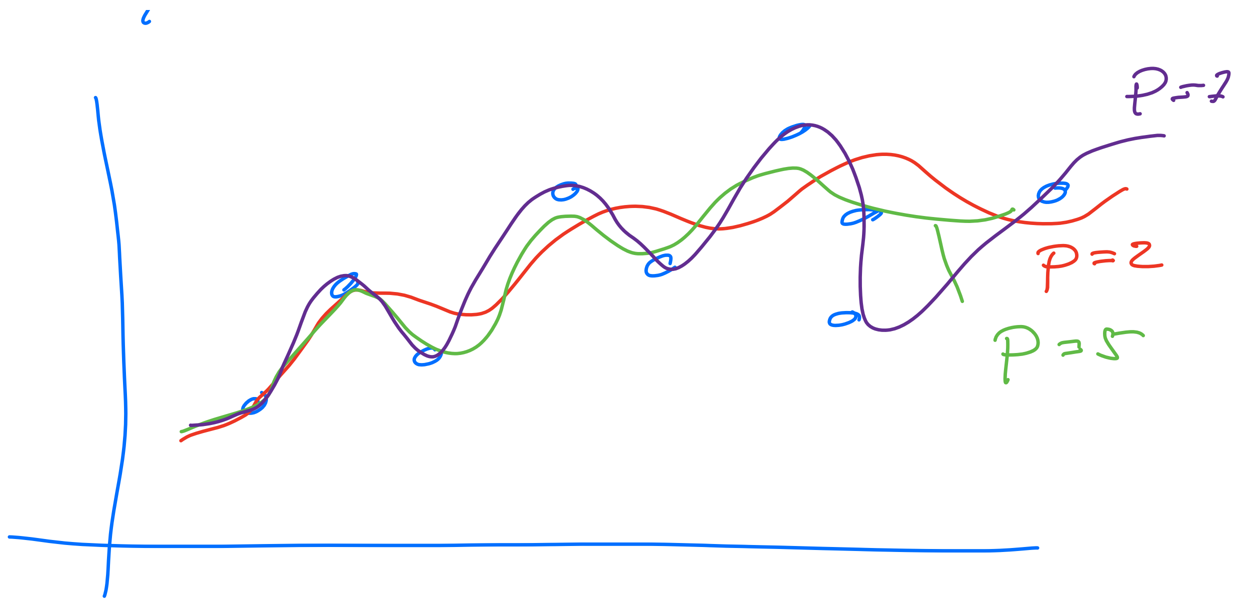
$$\hat{\beta}_{Lasso} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; \|(y - X\beta)\|_2^2 + \lambda \sum_{j=0}^{p-1} |\beta_j|$$

$$\left(\lambda \|\beta\|_1 \right)$$

$$\frac{d|\beta|}{d\beta} := \operatorname{sgn}(\beta) = \begin{cases} 1 & \beta > 0 \\ 0 & \beta = 0 \\ -1 & \beta < 0 \end{cases}$$

$$\frac{\partial C}{\partial \beta} = 0 = -2X^T(y - X\beta) + \lambda \operatorname{sgn}(\beta)$$

why do we introduce λ as an additional parameter?



$p =$	2	5	7
β_1	0.2	~~0.82~~	~~0.87~~
β_2	0.1	~~-0.7~~	~~-0.8~~
β_3		+30.0	+31.0
β_4		-10.0	+20.0
β_5		+2.0	+30.0
β_6			-10.0 ±10
β_7			+5.0

$$f(x) = \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots \beta_7 x^7$$

with a parameter λ,
we can dampen the
fluctuations in β.