TP - IMA 205 : Introduction supervised learning - Nanel WAFFA

OLS

We assume: $y = x\beta + \varepsilon$ with $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2 I$

OLS estimator: $\beta^* = (x^T x)^{-1} x^T y = Hy$

$\tilde{\beta} = Cy = (H+D)y$

q1) • $E(\tilde{\beta}) = E[(H+D)y] = (H+D)E(y)$ and $E(y) = X\beta + \underbrace{E(\varepsilon)}_{=0}$

So $E(\tilde{\beta}) = (H+D)X\beta$

Since $\tilde{\beta}$ is unbiased:

$E(\tilde{\beta}) = \beta$ i.e $\underbrace{Hx\beta}_{=I\beta} + Dx\beta = \beta$

i.e $Dx\beta = 0$

i.e $Dx = 0$

We use the assumption that $x \in Ker(D)$

• We have $Var(\beta^*_{OLS}) = \sigma^2 HH^T$

Let's compute $Var(\tilde{\beta})$

$Var(\tilde{\beta}) = Var(Cy)$

$= Var((X\beta + C\varepsilon)$

$= C^T \underbrace{Var(X\beta+\varepsilon)}_{=Var(\varepsilon)=\sigma^2 I} C$

$= \sigma^2 (HH^T + \underbrace{HD^T}_{\to 0} + \underbrace{DH^T}_{\to 0} + DD^T)$

$\to 0$ since $x \in Ker(D)$, $x^T \in Ker(D^T)$

$Var(\tilde{\beta}) = \underbrace{\sigma^2 HH^T}_{=Var(\beta^*)} + \underbrace{\sigma^2 DD^T}_{>0 \text{ since } D \text{ non-zero}} > Var(\beta^*)$

Ridge regression

$$\beta^*_{ridge} = \underset{\beta}{\operatorname{argmin}} \underbrace{(y_c - x_c\beta)^T (y_c - x_c\beta) + \lambda \|\beta\|_2^2}_{f(\beta)}$$

q2) We have the first order optimality condition

$$\frac{\partial f(\beta^*_{ridge})}{\partial \beta} = -2x_c^T(y - x\beta) + 2\lambda\beta = 0$$

i.e $$\beta^*_{ridge} = (x^Tx + \lambda I)^{-1} x^T y$$

Let's compute the bias:

$$E(\beta^*_{ridge}) = E\left[(\lambda I + x^Tx)^{-1} x^Tx\beta\right] + \underbrace{E\left[(\lambda I + x^Tx)^{-1} x^T \varepsilon\right]}_{=0}$$

$$E(\beta^*_{ridge}) = (\lambda I + x^Tx)^{-1} x^Tx\beta$$

$$\neq \beta \quad \text{for } \lambda \neq 0$$

The estimator of ridge regression is biased.

SVD decomposition

q3) We have $x_c = UDV^T$

$$\beta^*_{ridge} = (x^Tx + \lambda I)^{-1} x^T y$$

$$= (VD^TDV^T + \lambda VV^T)^{-1} VD^TU^T y$$

$$= V (D^2 + \lambda)^{-1} \underbrace{V^TV}_{=I} D^TU^T y$$

$$= V (D^2 + \lambda)^{-1} DU^T y$$

since $y = x\beta = UDV^T\beta$:

$$\beta^*_{ridge} = V (D^2 + \lambda)^{-1} D^TDV^T \beta$$

$$\beta^*_{ridge} = V \, \text{diag}\left(\frac{d_i^2}{d_i^2 + \lambda}\right) V^T \beta$$

Variance

q4) We have:

$$\text{Var}(\beta^*_{ridge}) = E\left[(\beta^*_{ridge} - E[\beta^*_{ridge}])(\beta^*_{ridge} - E[\beta^*_{ridge}])^T\right]$$

$$= E\left[\left((x^Tx + \lambda I)^{-1} x^T\varepsilon\right)\left((x^Tx+\lambda I)^{-1}x^T\varepsilon\right)^T\right]$$

$$= (x^Tx+\lambda I)^{-1} x^T \underbrace{E[\varepsilon\varepsilon^T]}_{=\sigma^2 I} x (x^Tx+\lambda I)^{-1}$$

$$\text{Var}(\beta^*_{ridge}) = \sigma^2 V (D^TD+\lambda I)^{-1} D^TD (D^TD+\lambda I)^{-1} V^T$$

$$\text{Var}(\beta^*_{ridge}) = \sigma^2 V \,\text{diag}\left(\frac{d_i^2}{(d_i^2+\lambda)^2}\right) V^T$$

And $\text{Var}(\beta^*_{OLS}) = \sigma^2 (x^Tx)^{-1} = \sigma^2 V \,\text{diag}\left(\frac{1}{d_i^2}\right) V^T$

Thus $\underbrace{\text{Var}(\beta^*_{ridge}) - \text{Var}(\beta^*_{OLS})}_{\leq 0} = \sigma^2 V \,\text{diag}\left(\underbrace{\frac{d_i^2}{(d_i^2+\lambda)^2} - \frac{1}{d_i^2}}\right) V^T$

$$= -\frac{2\lambda d_i^2 + \lambda^2}{d_i^2(d_i^2+\lambda)} \leq 0$$

And so:

$$\text{Var}(\beta^*_{OLS}) \geq \text{Var}(\beta^*_{ridge})$$

q5) Using what we found for the variance and bias for the ridge:

$$E(\beta^*_{ridge}) = V \,\text{diag}\left(\frac{d_i^2}{d_i^2+\lambda}\right) V^T\beta \xrightarrow[\lambda \to 0]{} 0$$

$$\text{Var}(\beta^*_{ridge}) = \sigma^2 V \,\text{diag}\left(\frac{d_i^2}{(d_i^2+\lambda)^2}\right) V^T \xrightarrow[\lambda \to 0]{} 0$$

q6) IF $x_c^T x_c = Id$: $\quad (x^Tx)^{-1} = Id$

$$\beta^*_{ridge} = (\underbrace{x^Tx}_{=I} + \lambda I)^{-1} x^T y = \frac{1}{1+\lambda} \underbrace{I}_{=(x^Tx)^{-1}} x^T y = \frac{1}{1+\lambda}(x^Tx)^{-1}x^Ty$$

$$\beta^*_{ridge} = \frac{\beta^*_{OLS}}{1+\lambda} y$$

'Elastic Net

$$\beta^*_{EENet} = \underset{\beta}{\operatorname{argmin}} \underbrace{(y_c - x_c\beta)^T (y_c - x_c\beta) + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1}_{f(\beta)}$$

97) We have:

$$\frac{\partial f(\beta^*_{EENet})}{\partial \beta} = -2x_c^T(y - x\beta) + 2\lambda\alpha\beta + \lambda(1-\alpha)(\pm 1) = 0$$

$$2\beta(x_c^T x_c + \underbrace{\lambda\alpha}_{\lambda_2}) = 2\underbrace{x^T y}_{(x^Tx)^{-1}x^Ty = \beta^*_{OLS}} \pm \underbrace{\lambda(1-\alpha)}_{\lambda_1}$$

since $(x^Tx)^T = Id$

$$\beta^*_{EENet} = \frac{\beta^*_{OLS} \pm \frac{\lambda_1}{2}}{1 + \lambda_2}$$