Ima205 : Answer to theoratical questions

HAITHEM DAGHMOURA

March 2024

OLS

We know that OLS estimator is defined as $\hat{\beta}^* = (x^T x)^{-1} x^T y = Hy$. Let β Another linear unbiased estimator of β defined as $\tilde{\beta} = Cy$, where C is a matrix $d \times n$ and $C = H + D$, D being a non-zero matrix.

First let's calculate the expected value and variance of $\tilde{\beta}$:

- $E[\tilde{\beta}] = E[Cy] = (Id + Dx)\beta$, as $\tilde{\beta}$ is unbiased, it can be concluded that Dx must be equal to zero.
- $\text{Var}(\tilde{\beta}) = \text{Var}(Cy) = C\text{Var}(y)C^T = \sigma^2 CC^T$.

$$\sigma^2 CC^T = \sigma^2((x^T x)^{-1} x^T + D)(x(x^T x)^{-1} + D^T)$$

$$\sigma^2 CC^T = \sigma^2 (x^T x)^{-1} + \sigma^2 (x^T x)^{-1}(Dx)^T + \sigma^2 (Dx)(x^T x)^{-1} + \sigma^2 (DD^T)$$

As proved in the expected value, Dx must be equal to 0 so the estimator is unbiased, this means that:

$$\text{Var}(\tilde{\beta}) = \sigma^2 (x^T x)^{-1} + \sigma^2 (DD^T)$$

A matrix DD^T is always symmetric and semi-positive, meaning $DD^T \geq 0$. If it's equal to 0, it is equivalent to the OLS. To conclude:

$$\text{Var}(\tilde{\beta}) = \text{Var}(\hat{\beta}^*) + \sigma^2 (DD^T)$$

As the second term is greater than 0, the variance of this new estimator is greater than the OLS.

Given the calculations above, the assumption that $\text{Var}(\hat{\beta}^*) < \text{Var}(\tilde{\beta})$ holds. This is assuming that x is deterministic and $E[\epsilon] = 0$ (normality assumption holds, $\epsilon \sim N(0, \sigma^2 I)$.

Ridge Regression

- The Ridge solution is given by: $\hat{\beta}^*_{\text{ridge}} = (x_c^T x_c + \lambda I)^{-1} x_c^T y_c$.
- So the expected value is given by: $E[\hat{\beta}^*_{\text{ridge}}] = E[(x_c^T x_c + \lambda I)^{-1} x_c^T y_c] = [(x_c^T x_c + \lambda I)^{-1} x_c^T] E[y_c]$
- Which is different from β unless $\lambda = 0$ (the OLS case), meaning it is a biased estimator.
- The SVD decomposition for the Ridge estimator can be written as:

  $$\hat{\beta}^*_{\text{ridge}} = (x_c^T x_c + \lambda I)^{-1} x_c^T y_c = ([UDV^T]^T [UDV^T] + \lambda I)^{-1} (UDV^T)^T y_c$$

  $$= (VD^T U^T UDV^T + \lambda I)^{-1} VD^T U^T y_c = (VD^T DV^T + \lambda I)^{-1} VD^T U^T y_c = V(D^T D + \lambda I)^{-1} V^T VD^T U^T y_c$$

  $$\hat{\beta}^*_{\text{ridge}} = V(D^T D + \lambda I)^{-1} D^T U^T y_c$$

  The manipulations to get the result above use that U and V are orthogonal matrices (the inverse is equal to the transpose). Using this transformation might be computationally useful because there is no need to invert a matrix, as $(D^T D + \lambda I)^{-1} D^T$ is equal to a diagonal matrix, where each element is equal to eigenvalue (eigenvalue$^2 + \lambda$).
- The variance of the Ridge estimator can be calculated as:

  $$\text{Var}(\hat{\beta}^*_{\text{ridge}}) = \text{Var}((x_c^T x_c + \lambda I)^{-1} x_c^T y_c)$$

  $$\text{Var}(\hat{\beta}^*_{\text{ridge}}) = ((x_c^T x_c + \lambda I)^{-1} x_c^T y_c) \text{Var}(y_c) ((x_c^T x_c + \lambda I)^{-1} x_c^T)^T$$

  $$\text{Var}(\hat{\beta}^*_{\text{ridge}}) = \sigma^2 (x_c^T x_c + \lambda I)^{-1} x_c^T x_c (x_c^T xc + \lambda I)^{-1}$$

  For a positive λ, $(x_c^T x_c + \lambda I)$ will always be greater than $x_c^T x_c$, as a consequence $(x_c^T x_c + \lambda I)^{-1} x_c^T xc (x_c^T x_c + \lambda I)^{-1}$ will always be smaller than $(x_c^T x_c)^{-1}$, meaning that $\text{Var}(\hat{\beta}^*_{\text{OLS}}) \geq \text{Var}(\hat{\beta}^*_{\text{ridge}})$.
- The Ridge estimator promotes a trade-off between Bias and Variance. As λ increases, the Bias becomes bigger, and the variance becomes smaller. This is logical given that if we take a λ really close to zero, the solution will tend to the OLS solution, with zero bias and high variance, and if λ is close to infinity, the solution will be all parameters equal to zero, meaning zero variance, but high bias.
- As: $\hat{\beta}^*_{\text{ridge}} = (x_c^T x_c + \lambda Id)^{-1} x_c^T y_c$. If $x_c^T x_c = Id$,

  Therefore $\hat{\beta}^*_{\text{ridge}} = (Id + \lambda Id)^{-1} x_c^T y_c = ((1 + \lambda) Id)^{-1} x_c^T y_c$.

  Remembering: $\hat{\beta}^*_{\text{OLS}} = (x_c^T x_c)^{-1} x_c^T yc$, where too $x_c^T x_c = Id$, than $\hat{\beta}^*_{\text{OLS}} = x_c^T y_c$

  Substituting, it's demonstrated that $\hat{\beta}^*_{\text{ridge}} = \frac{\hat{\beta}^*_{\text{OLS}}}{1+\lambda}$

Elastic Net

Rewriting equation 2 from the exercise list:

$$\hat{\beta}^*_{\text{ElNet}} = \text{argmin}_\beta (y_c - x_c\beta)^T (y_c - x_c\beta) + \lambda_2 ||\beta||_2^2 + \lambda_1 ||\beta||_1$$

As the function is strictly convex, the minimum can be obtained by equating the subgradient to zero according to Fermat's law ($\lambda_1 ||\beta||_1$ is not differentiable at 0).

$$\frac{\partial f}{\partial \beta} = 2x_c^T (y_c - x_c\beta) + 2\lambda_2\beta + \lambda_1 \begin{cases} -1 & \text{if } \beta < 0 \\ 1 & \text{if } \beta > 0 \\ [-1, 1] & \text{if } \beta = 0 \end{cases}$$

$$2x_c^T (y_c - x_c\beta) + 2\lambda_2\beta \pm \lambda_1 = 0$$

$$2x_c^T y_c - 2x_c^T x_c\beta + 2\lambda_2\beta \pm \lambda_1 = 0$$

Remembering that $x_c^T x_c = I_d$, so $\hat{\beta}^*_{\text{OLS}} = x_c^T y_c$.

$$2\beta^*_{\text{OLS}} - 2\beta(1 - \lambda_2) \pm \lambda_1 = 0$$

$$\beta = \frac{\beta^*_{\text{OLS}} \pm \frac{\lambda_1}{2}}{(1 - \lambda_2)}$$

This gives the expected value proved by this demonstration.