IMA205 - Introduction Supervised Learning

Júlia Togashi de Miranda

March 2021

1 OLS

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

Calculating expected value and variance of $\tilde{\beta}$:

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

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

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

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

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

  A matrix DD^T is always symmetric and semi positive, that means $DD^T \geq 0$, if it's equal to 0, is equivalent to the OLS, to conclude:

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

  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 $Var(\beta^* < Var(\tilde{\beta})$ holds. That is assuming that x is deterministic and $E[\epsilon] = 0$ (normality assumption holds, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

2 Ridge regression

- The explicit Ridge solution can be written as follows: $\beta^*_{ridge} = (x_c^T x_c + \lambda I)^{-1} x_c^T y_c$.
  So the expected value is given by: $E[\beta^*_{ridge}] = E[(x_c^T x_c + \lambda I)^{-1} x_c^T y_c] = [(x_c^T x_c + \lambda I)^{-1} x_c^T] E[y_c]$

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

  Which is different from β unless $lambda = 0$ (the OLS case), meaning is a biased estimator.

- The SVD decomposition for the Ridge estimator can be written as:

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

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

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

  The manipulations to get the result above use that U and V are orthogonal matrix (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 $\frac{eigenvalue}{(eigenvalue^2 + \lambda)}$.

- The variance of the Ridge estimator can be calculated as: $Var(\beta^*_{ridge}) = Var((x_c^T x_c + \lambda I)^{-1} x_c^T y_c)$

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

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

  For a positive λ, $(x_c^T x_c + \lambda I)$ will always be greater than $x_c^T x_c$, as consequence $(x_c^T x_c + \lambda I)^{-1} x_c^T x_c (x_c^T x_c + \lambda I)^{-1}$ will always be smaller than $(x_c^T x_c)^{-1}$, meaning that $Var(\beta^*_{OLS}) \geq Var(\beta^*_{Ridge})$.

- 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 0 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 variance.

- As: $\beta^*_{ridge} = (x_c^T x_c + \lambda I_d)^{-1} x_c^T y_c$.If $x_c^T x_c = I_d$,
  Therefore $\beta^*_{ridge} = (I_d + \lambda I_d)^{-1} x_c^T y_c = ((1 + \lambda) I_d)^{-1} x_c^T y_c$.

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

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

3 Elastic Net

Rewriting equation 2 from the exercise list:

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

As the function is strictly convex, the minimum can be obtained equaling the subgradient to zero ($\lambda_1||\beta||_1$ is not differentiable in 0.)

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

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

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

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

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

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

Giving the expected value proved by this demonstration.