Ridge regression setup

Setup: $Y, X_1, \ldots, X_p$ with $p \gg 0$ or even $p > n$

Model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ as "usual"

$\hat{\beta}_0 = \bar{Y}$, $\underset{p \times 1}{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$

$Y_i \to \underset{\text{new } Y}{Y_i - \bar{Y}} \quad \left[= Y_i - \hat{\beta}_0 \right]$ ($\to$ old Y)

$x_i \to \dfrac{x_i - \bar{x}}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} (x_j - \bar{x})^2}}$ (x_i old x) $\Rightarrow$ new xs

$\Rightarrow$ $\bar{x} = 0$ and $\sum_{i=1}^{n} x_i^2 = n$

$\text{new } x_i$

X = design matrix of new features without 1st column of 1s, so is $n \times p$.

Ridge regression

" " estimate of β minimizes

$$\sum_{i=1}^{n} \left(y_i - (\beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad , \quad \lambda \geq 0$$

$$= (y - X\beta)^T (y - X\beta) + \lambda \|\beta\|_2^2 \quad \rightarrow \text{same}$$

$$= \cdots\cdots + \lambda\, \beta^T \beta$$

Ex $p=1$, $y = \beta x_1 + \varepsilon$. OLS minimizes

$$\sum (y_i - \beta x_i)^2 \Rightarrow \hat{\beta}_{ols} = \frac{\sum_i x_i y_i}{\sum_{i,j} x_j^2} = \frac{\sum_{i=1}^{n} x_i y_i}{n}$$

$$\hat{\beta}_{ols}$$

ridge minimizes

$$\sum (y_i - \beta x_i)^2 + \lambda \beta^2$$

$$\frac{\partial}{\partial \beta} \Big| \Rightarrow -2\sum_{i=1}^{n} x_i(y_i - \beta x_i) + 2\lambda\beta$$

$$= -2\sum x_i y_i + 2\beta \underbrace{\sum x_i^2}_{n} + 2\lambda\beta$$

$$= -2\sum x_i y_i + 2\beta(n+\lambda) \overset{\text{set}}{=} 0$$

$$\Rightarrow \boxed{\hat{\beta}_{Ridge} = \frac{\sum x_i y_i}{n+\lambda}}$$

$\Rightarrow \hat{\beta}_{Ridge}$ is shrinking $\hat{\beta}_{OLS}$ toward 0

Note Ridge regression is a shrinkage estimator : in that the estimated βs tend to be shrunk toward 0.

- $\lambda = 0 \Rightarrow \hat{\beta}_{Ridge} = \hat{\beta}_{OLS}$
- $\lambda \to \infty \Rightarrow \hat{\beta}_{Ridge} = 0 \Rightarrow$ model reduces to $Y = \beta_0 + \varepsilon$

Solution

$\hat{\beta}_{ridge}$ minimizes $(y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta$

$$\frac{\partial}{\partial \beta} \Big| \Rightarrow -2X^T(y - X\beta) + 2\lambda\beta$$

$$= -2X^T y + 2X^T X\beta + 2\lambda\beta$$

$$= -2X^T y + 2(X^T X + \lambda I)\beta \overset{\text{set}}{=} 0$$

$$\Rightarrow (X^T X + \lambda I)\beta = X^T y$$

$$\Rightarrow \boxed{\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T y}$$

Aside $(X^T X)$ is a nonnegative definite matrix (semi positive definite).

Choose any $\underline{a}$

show $0 \leq \underline{a}^T (X^T X) \underline{a} = (\underline{a}^T X^T)(X \underline{a}) = (X\underline{a})^T (X\underline{a})$

$[X\underline{a} = \text{some vector} = \underline{b}]$

$= \underline{b}^T \underline{b} = \sum_{i=1}^{p} b_i^2 \geq 0.$ ✓

That implies that $(X^T X)$ has nonnegative eigenvalues.

But if $(X^T X)$ is not full rank, it has some eigenvalues that are exactly 0. $\Rightarrow$ $(X^T X)$ is not always invertible. $\hat{\beta}_{ols} = (X^T X)^{-1} X^T y$

The effect of ridge's $+\lambda I$? Bumps up all eigenvalues of $X^T X$ by λ.

Side: If $\underline{v}$ is eigenvector of (X^TX) with eigenvalue α. Then

$$(X^TX + \lambda I)\underline{v} = (X^TX)\underline{v} + \lambda \underline{v} = \alpha \underline{v} + \lambda \underline{v}$$

$$= (\alpha + \lambda)\underline{v}$$

$\Rightarrow$ $\underline{v}$ is still an eigenvec of $(X^TX + \lambda I)$ but it has eigenval $\alpha + \lambda$

$\Rightarrow$ All zero eigenvalues of (X^TX) are now $\lambda > 0$

$\Rightarrow$ $(X^TX + \lambda I)$ is invertible

$\Rightarrow$ unique solution & allows us to fit models with more parameters than data.

Properties

- $E\,\hat{\beta}_{ridge} = E\left((X^TX + \lambda I)^{-1} X^T y\right)$

$= (X^TX + \lambda I)^{-1} X^T E\,y$

$= (X^TX + \lambda I)^{-1} X^T X \beta \neq \beta$

$\Rightarrow$ $\hat{\beta}_{ridge}$ is biased

- $\mathrm{Var}\,\hat{\beta}_{ridge} = \sigma^2 (X^TX + \lambda I)^{-1} (X^TX) (X^TX + \lambda I)^{-1}$

$\mathrm{Var}\,\hat{\beta}_{OLS} = \sigma^2 (X^TX)^{-1}$

nano

$\mathrm{Var}\,\hat{\beta}_{ridge}$

$\mathrm{Var}\,\hat{\beta}_{OLS}$

$\frac{1}{a}$

$(a+\lambda)^{-1}\, a\, (a+\lambda)^{-1} = \frac{a}{(a+\lambda)^2}$

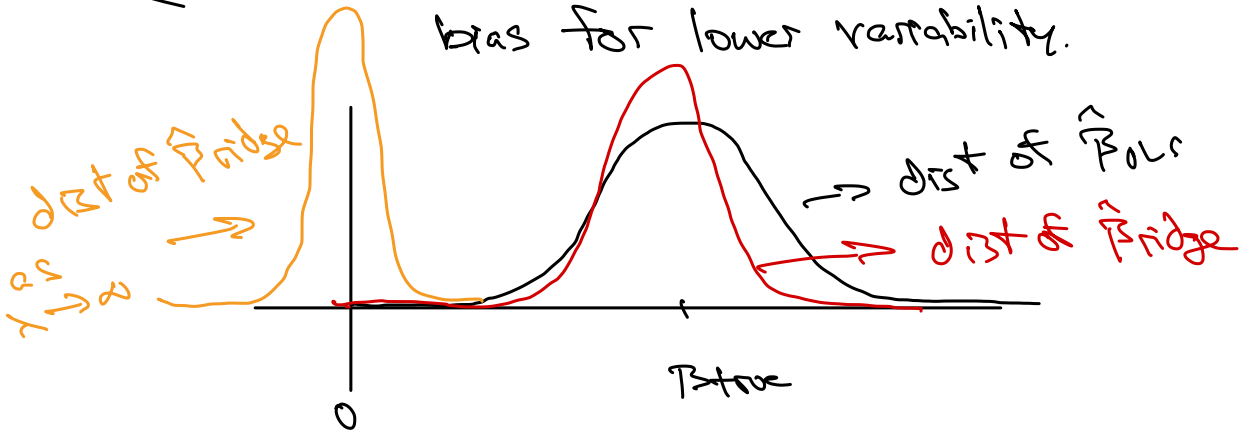
Is $\text{Var}_{\text{Ridge}} < \text{Var}_{\text{OLS}}$?

$$\frac{a}{(a+\lambda)^2} \overset{?}{<} \frac{1}{a}$$

$$\frac{a^2}{(a+\lambda)^2} \overset{?}{<} 1$$

Take home point

Ridge regression sacrifices some bias for lower variability.



Recall The mean squared error (MSE) is

$$E\left[\sum_{j=1}^{p}(\beta_j - \hat{\beta}_j)^2\right] = E\left[(\beta - \hat{\beta})^T(\beta - \hat{\beta})\right]$$

For ridge:

$$MSE(\lambda) = \beta^T(I - M(\lambda)X)^T(I - M(\lambda)X)\beta + \sigma^2 \text{tr}[M(\lambda)^T M(\lambda)]$$

trace

where $M(\lambda) = (X^TX + \lambda I)^{-1}X^T$

For certain choices of λ, $MSE_{ridge} < MSE_{OLS}$

$MSE = (\text{Squared bias}) + (\text{Var})$