

Lecture 23: Mean squared error of the ridge estimator; the Hoerl–Kennard theorem

Econometrics ■ Unit II: Multicollinearity and heteroskedasticity

Kapil Kumar

Department of Statistics
Chaudhary Charan Singh University, Meerut

B.Sc. (Hons) Statistics, Semester V
Academic Session 2026–2027

Previous lecture

Lecture 22: Ridge regression: the estimator, constrained least squares, bias and covariance.

Today

- Ridge regression
 - Mean squared error of the ridge estimator
 - Existence of a better ridge estimator

Use the eigen decomposition $X'X = V\Lambda V'$ and put $\alpha = V'\beta$.

Step 1. Since $VV' = I$,

$$\begin{aligned}X'X + \delta I &= V\Lambda V' + \delta VV' = V(\Lambda + \delta I)V', \\(X'X + \delta I)^{-1} &= V(\Lambda + \delta I)^{-1}V'.\end{aligned}$$

Step 2 (variance part).

$$\begin{aligned}\text{tr Var}(\hat{\beta}_{\text{ridge}}) &= \sigma^2 \text{tr}\{V(\Lambda + \delta I)^{-1}V' V \Lambda V' V(\Lambda + \delta I)^{-1}V'\} \\ &= \sigma^2 \text{tr}\{V(\Lambda + \delta I)^{-1}\Lambda(\Lambda + \delta I)^{-1}V'\} \\ &= \sigma^2 \text{tr}\{(\Lambda + \delta I)^{-1}\Lambda(\Lambda + \delta I)^{-1}V'V\} \quad (\text{Rule P5}) \\ &= \sigma^2 \sum_{j=1}^k \frac{\lambda_j}{(\lambda_j + \delta)^2}.\end{aligned}$$

Step 3 (bias part).

$$\begin{aligned}\text{Bias}' \text{Bias} &= \delta^2 \beta' (X'X + \delta I)^{-1} (X'X + \delta I)^{-1} \beta \\ &= \delta^2 \beta' V (\Lambda + \delta I)^{-2} V' \beta = \delta^2 \alpha' (\Lambda + \delta I)^{-2} \alpha \\ &= \delta^2 \sum_{j=1}^k \frac{\alpha_j^2}{(\lambda_j + \delta)^2}.\end{aligned}$$

Step 4 (total).

$$\text{MSE}(\hat{\beta}_{\text{ridge}}) = \sigma^2 \sum_{j=1}^k \frac{\lambda_j}{(\lambda_j + \delta)^2} + \delta^2 \sum_{j=1}^k \frac{\alpha_j^2}{(\lambda_j + \delta)^2}. \quad (3.1)$$

In matrix form (3.1) is

$$\text{MSE}(\hat{\beta}_{\text{ridge}}) = \sigma^2 \text{tr}\{(X'X + \delta I)^{-1}X'X(X'X + \delta I)^{-1}\} \\ + \delta^2 \beta'(X'X + \delta I)^{-2}\beta.$$

- At $\delta = 0$ the first term is $\sigma^2 \sum_j 1/\lambda_j = \text{MSE}(b)$ and the second is zero.
- As δ increases, the variance term decreases and the squared-bias term increases.
- The choice of δ balances the two.

Theorem 3.6 (Hoerl and Kennard (1970))

There always exists $\delta > 0$ such that $\text{MSE}(\hat{\beta}_{\text{ridge}}) < \text{MSE}(b)$. In particular, this holds for every δ with

$$0 < \delta < \frac{\sigma^2}{\max_j \alpha_j^2},$$

and hence for every $0 < \delta < \sigma^2/(\beta'\beta)$.

Proof.

Let $m(\delta)$ denote the right side of (3.1). Differentiate each term with respect to δ .

Proof (cont.)

Step 1 (variance term).

$$\frac{d}{d\delta} \left\{ \frac{\sigma^2 \lambda_j}{(\lambda_j + \delta)^2} \right\} = \sigma^2 \lambda_j \cdot (-2)(\lambda_j + \delta)^{-3} = -\frac{2\sigma^2 \lambda_j}{(\lambda_j + \delta)^3}.$$

Step 2 (bias term). By the quotient rule,

$$\begin{aligned} \frac{d}{d\delta} \left\{ \frac{\delta^2}{(\lambda_j + \delta)^2} \right\} &= \frac{2\delta(\lambda_j + \delta)^2 - \delta^2 \cdot 2(\lambda_j + \delta)}{(\lambda_j + \delta)^4} \\ &= \frac{2\delta(\lambda_j + \delta) - 2\delta^2}{(\lambda_j + \delta)^3} = \frac{2\delta\lambda_j}{(\lambda_j + \delta)^3}. \end{aligned}$$

Proof (cont.)

Step 3 (add).

$$m'(\delta) = \sum_{j=1}^k \frac{-2\sigma^2\lambda_j + 2\delta\lambda_j\alpha_j^2}{(\lambda_j + \delta)^3} = 2 \sum_{j=1}^k \frac{\lambda_j(\delta\alpha_j^2 - \sigma^2)}{(\lambda_j + \delta)^3}.$$

Step 4 (sign). If $0 < \delta < \sigma^2 / \max_j \alpha_j^2$, then $\delta\alpha_j^2 < \sigma^2$ for every j , so every term is negative and $m'(\delta) < 0$. Hence m is strictly decreasing on this interval, and $m(\delta) < m(0) = \text{MSE}(b)$ for every such δ .

Step 5. Since V is orthogonal, $\sum_j \alpha_j^2 = \boldsymbol{\alpha}'\boldsymbol{\alpha} = \boldsymbol{\beta}'VV'\boldsymbol{\beta} = \boldsymbol{\beta}'\boldsymbol{\beta}$, so $\max_j \alpha_j^2 \leq \boldsymbol{\beta}'\boldsymbol{\beta}$ and $\sigma^2 / \boldsymbol{\beta}'\boldsymbol{\beta} \leq \sigma^2 / \max_j \alpha_j^2$. □

- Note also that

$$m'(0) = -2\sigma^2 \sum_j 1/\lambda_j^2 < 0 :$$

even a very small δ reduces the MSE.

- The theorem proves that a good δ exists, but the bound involves the unknown σ^2 and β .
- In practice δ must be chosen from the data.

Practice

Re-work the derivations and examples of this lecture without looking at the slides.

Next lecture

Lecture 24: Why ridge regression works; choice of the biasing parameter; ridge trace and HKB.