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

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

This lecture

Previous lecture

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

Today

- Ridge regression
  - Why ridge regression works: the idea in eigenvalue form
  - Choice of the biasing parameter

Why ridge regression works

Multicollinearity arises because some eigenvalues of $X'X$ are close to zero.

In the canonical form (Section 3.5.5 of the notes), the OLS estimator of $\boldsymbol{\alpha} = V'\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\alpha}} = \Lambda^{-1}V'X'y$, so $\widehat{\alpha}_j = (V'X'y)_j/\lambda_j$. For ridge,

$$V'\widehat{\boldsymbol{\beta}}_{\text{ridge}} = V'V(\Lambda + \delta I)^{-1}V'X'y = (\Lambda + \delta I)^{-1}V'X'y,$$

so the jth canonical ridge coefficient is

$$\frac{(V'X'y)_j}{\lambda_j + \delta} = \frac{\lambda_j}{\lambda_j + \delta}\,\widehat{\alpha}_j.$$

Why ridge regression works (cont.)

- OLS divides by λ_j; ridge divides by $\lambda_j + \delta$.
- A small δ is added to each λ_j, so that even if $\lambda_j = 0$, the quantity $1/(\lambda_j + \delta)$ remains meaningful.
- The shrinkage factor $\lambda_j/(\lambda_j + \delta)$ lies between 0 and 1 and is smallest for the smallest λ_j: the directions with the least information are shrunk the most.

Choice of the biasing parameter

The value of δ can be chosen using criteria such as stability of the estimates, reasonable signs and size of the coefficients, and the residual sum of squares. Two common methods are as follows.

(a) Ridge trace. The ridge trace is a graph of the ridge estimates of all coefficients against δ.

- Under severe multicollinearity the estimates change rapidly near $\delta = 0$ and then settle down.
- We choose the smallest δ at which all the curves have become stable (nearly flat).
- Moving to a larger δ after that point adds bias without much further reduction in variance.

Choice of the biasing parameter (cont.)

(b) Hoerl–Kennard–Baldwin choice (Hoerl et al., 1975). Use

$$\widehat{\delta} = \frac{k\widehat{\sigma}^2}{b'b},$$

where b and $\widehat{\sigma}^2$ are the OLS estimates from the standardised model.

This copies the bound in the theorem above, with the unknown quantities replaced by their estimates.

Example 3.5 (Ridge trace and the HKB choice)

Use the standardised data of Example 3.3:

$$X'X = \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}, \qquad X'y = \begin{pmatrix} 0.8 \\ 0.7 \end{pmatrix}, \qquad n = 20.$$

OLS ($\delta = 0$). $|X'X| = 1 - 0.9025 = 0.0975$, so

$$b = \frac{1}{0.0975}\begin{pmatrix} 1 & -0.95 \\ -0.95 & 1 \end{pmatrix}\begin{pmatrix} 0.8 \\ 0.7 \end{pmatrix}$$
$$= \frac{1}{0.0975}\begin{pmatrix} 0.8 - 0.665 \\ 0.7 - 0.76 \end{pmatrix} = \begin{pmatrix} 1.3846 \\ -0.6154 \end{pmatrix}.$$

The second coefficient has the wrong (negative) sign although x_2 is positively correlated with y.

Example 3.5 (cont.)

Ridge for a general δ. $X'X + \delta I = \begin{pmatrix} 1+\delta & 0.95 \\ 0.95 & 1+\delta \end{pmatrix}$ with determinant $(1+\delta)^2 - 0.9025$. Hence

$$\widehat{\beta}_1(\delta) = \frac{0.8(1+\delta) - 0.665}{(1+\delta)^2 - 0.9025}, \qquad \widehat{\beta}_2(\delta) = \frac{0.7(1+\delta) - 0.76}{(1+\delta)^2 - 0.9025}.$$

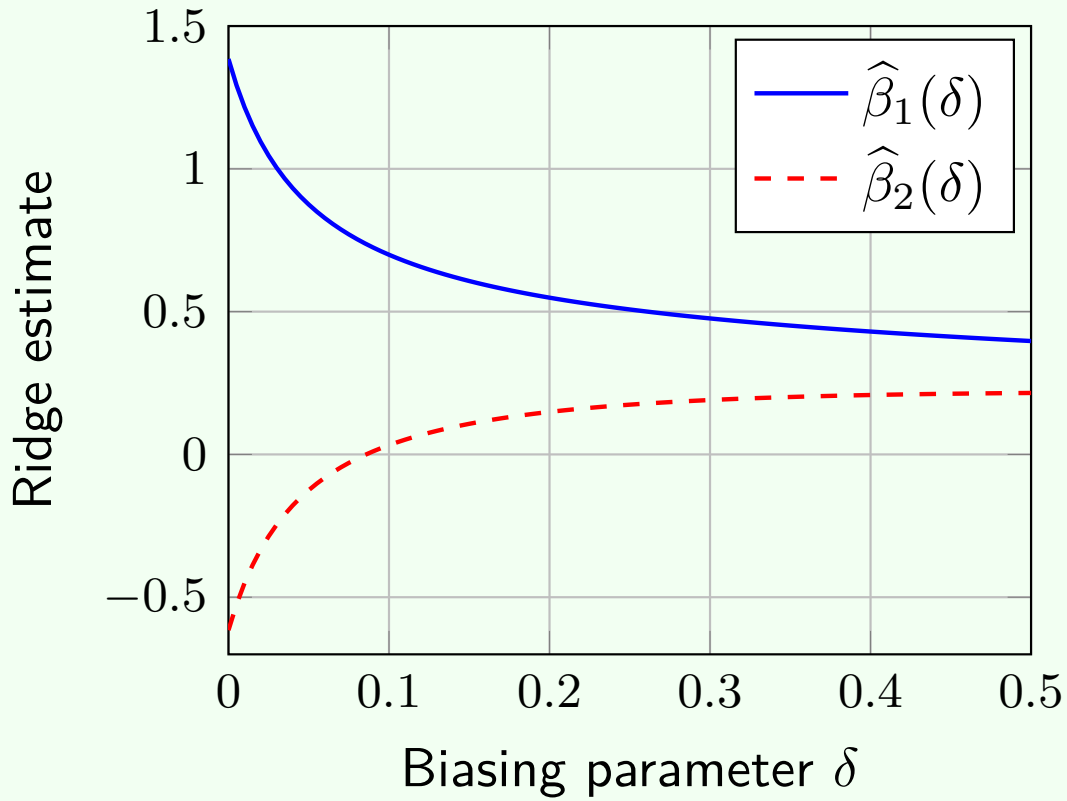
For $\delta = 0.10$: determinant $1.21 - 0.9025 = 0.3075$,

$$\widehat{\beta}_1 = \frac{0.88 - 0.665}{0.3075} = 0.6992, \qquad \widehat{\beta}_2 = \frac{0.77 - 0.76}{0.3075} = 0.0325.$$

Ridge trace and the HKB choice (cont.)

Example 3.5 (cont.): the ridge trace

δ	$\widehat{\beta}_1(\delta)$	$\widehat{\beta}_2(\delta)$
0	1.3846	−0.6154
0.01	1.2160	−0.4507
0.05	0.8750	−0.1250
0.10	0.6992	0.0325
0.20	0.5488	0.1488
0.50	0.3970	0.2152



Ridge trace and the HKB choice (cont.)

Example 3.5 (cont.)

The estimates change rapidly for δ below about 0.1 and are nearly stable after about 0.2. The ridge trace suggests $\delta \approx 0.2$, giving $(0.5488,\ 0.1488)'$: both coefficients now have sensible positive signs.

HKB choice. In the standardised model the total sum of squares is 1 and the explained sum of squares is

$$b' X' y = 1.3846(0.8) - 0.6154(0.7) = 0.6769.$$

So the residual sum of squares is $1 - 0.6769 = 0.3231$, with $n - k - 1 = 20 - 2 - 1 = 17$ degrees of freedom (one is lost for the mean of y).

Ridge trace and the HKB choice (cont.)

Example 3.5 (cont.)

Hence $\widehat{\sigma}^2 = 0.3231/17 = 0.0190$. Also $b'b = 1.3846^2 + 0.6154^2 = 2.2959$. Therefore

$$\widehat{\delta} = \frac{2(0.0190)}{2.2959} = 0.0166,$$

and the ridge estimates are $(1.1326,\ -0.3699)'$. This much smaller δ shrinks less than the ridge trace choice, an example of limitation 3 below.

Exam tip

"Is the ridge estimator unbiased?" No. Write $\mathrm{Bias}(\widehat{\boldsymbol{\beta}}_{\mathrm{ridge}}) = -\delta(X'X + \delta I)^{-1}\boldsymbol{\beta}$, and add that its MSE can still be smaller than that of OLS (Hoerl–Kennard theorem).

After this lecture

Practice

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

Next lecture

Lecture 25: Limitations of ridge regression; summary of key formulas.