

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

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 21: Remedies: more data, dropping variables, prior information, generalized inverse, principal components.

Today

- Remedies for multicollinearity
 - Ridge regression
- Ridge regression
 - Idea: accept a small bias to gain a large reduction in variance
 - The ridge estimator
 - Ridge regression as constrained least squares
 - Bias of the ridge estimator
 - Covariance matrix of the ridge estimator

Ridge regression (Hoerl and Kennard, 1970) is the most widely used remedy.

- It is treated in detail in the next section, following the level of Shalabh (n.d.) and Montgomery et al. (2021).

OLS has the smallest variance among *unbiased* linear estimators. If we allow a small bias, we may find an estimator $\hat{\beta}$ with much smaller variance, and hence smaller mean squared error.

For a vector estimator, the mean squared error is

$$\text{MSE}(\hat{\beta}) = E(\hat{\beta} - \beta)'(\hat{\beta} - \beta).$$

Decomposition. Add and subtract $E(\hat{\beta})$:

$$\hat{\beta} - \beta = \{\hat{\beta} - E(\hat{\beta})\} + \{E(\hat{\beta}) - \beta\}.$$

Squaring the length,

$$(\hat{\beta} - \beta)'(\hat{\beta} - \beta) = \{\hat{\beta} - E\hat{\beta}\}'\{\hat{\beta} - E\hat{\beta}\} + 2\{E\hat{\beta} - \beta\}'\{\hat{\beta} - E\hat{\beta}\} \\ + \{E\hat{\beta} - \beta\}'\{E\hat{\beta} - \beta\}.$$

Take expectations. The middle term has expectation zero because $E\{\hat{\beta} - E\hat{\beta}\} = 0$. The first term has expectation $\sum_j \text{Var}(\hat{\beta}_j) = \text{tr Var}(\hat{\beta})$. Hence

$$\text{MSE}(\hat{\beta}) = \text{tr Var}(\hat{\beta}) + \text{Bias}(\hat{\beta})' \text{Bias}(\hat{\beta}), \quad \text{Bias}(\hat{\beta}) = E(\hat{\beta}) - \beta.$$

So the MSE can be made smaller than the variance of OLS by introducing a small bias. Ridge regression does this.

Definition 3.5

The **ridge regression estimator** is obtained by modifying the normal equations to $(X'X + \delta I)\hat{\beta}_{\text{ridge}} = X'y$:

$$\hat{\beta}_{\text{ridge}} = (X'X + \delta I)^{-1}X'y, \quad \delta \geq 0,$$

where δ is a constant called the **biasing parameter** (or ridge parameter).

- If $\delta = 0$, $\hat{\beta}_{\text{ridge}} = (X'X)^{-1}X'y = b$, the OLS estimator.
- As $\delta \rightarrow \infty$, $\hat{\beta}_{\text{ridge}} = \frac{1}{\delta}(\frac{1}{\delta}X'X + I)^{-1}X'y \rightarrow 0$, because $\frac{1}{\delta}X'X \rightarrow 0$ and $\frac{1}{\delta} \rightarrow 0$.

The ridge estimator (cont.)

- So δ controls the amount of shrinkage towards zero. A larger δ shrinks more.
- Ridge regression is applied to standardised data (as in Section 3.1.1 of the notes), so that a change of measurement units does not change the amount of shrinkage of a coefficient. The final estimates are converted back to the original scale.

Since $X'y = X'Xb$, the ridge estimator can also be written in terms of OLS:

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

Remember

$$\hat{\beta}_{\text{ridge}} = (X'X + \delta I)^{-1} X'y, \quad \text{MSE}(\hat{\beta}) = \text{tr Var}(\hat{\beta}) + \text{Bias}' \text{Bias}.$$

Suppose we want the least squares estimate of β subject to the restriction that its squared length is fixed, $\beta'\beta = c$. Using a Lagrange multiplier δ , minimise

$$S(\beta) = (y - X\beta)'(y - X\beta) + \delta(\beta'\beta - c).$$

Step 1 (expand).

$$S(\beta) = y'y - 2\beta'X'y + \beta'X'X\beta + \delta\beta'\beta - \delta c.$$

Step 2 (differentiate). By Rules P1 and P2 (note $\beta'\beta = \beta'I\beta$),

$$\frac{\partial S}{\partial \beta} = -2X'y + 2X'X\beta + 2\delta\beta.$$

Step 3 (equate to zero). $X'X\beta + \delta\beta = X'y$, that is, $(X'X + \delta I)\beta = X'y$, so

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

Step 4 (minimum). The second-derivative matrix $2(X'X + \delta I)$ is positive definite for $\delta > 0$.

- So ridge regression is least squares with a penalty on large coefficients.
- The penalty prevents the coefficient vector from becoming very long, which is exactly what OLS does under multicollinearity (Section 3.3.4 of the notes).

Since $E(y) = X\beta$,

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

Hence

$$\begin{aligned}\text{Bias}(\hat{\beta}_{\text{ridge}}) &= E(\hat{\beta}_{\text{ridge}}) - \beta \\ &= (X'X + \delta I)^{-1}X'X\beta - (X'X + \delta I)^{-1}(X'X + \delta I)\beta \\ &= (X'X + \delta I)^{-1}\{X'X - X'X - \delta I\}\beta \\ &= -\delta(X'X + \delta I)^{-1}\beta.\end{aligned}$$

So the ridge estimator is biased for every $\delta > 0$ (unless $\beta = 0$).

Step 1.

$$\begin{aligned}\hat{\beta}_{\text{ridge}} - E(\hat{\beta}_{\text{ridge}}) &= (X'X + \delta I)^{-1}X'y - (X'X + \delta I)^{-1}X'X\beta \\ &= (X'X + \delta I)^{-1}X'(y - X\beta) = (X'X + \delta I)^{-1}X'\epsilon.\end{aligned}$$

Step 2. By Rule P4 with $\text{Var}(\epsilon) = \sigma^2 I$,

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

Practice

- Exercises 3.6 and 3.7 (Exercises and Solutions booklet).

Next lecture

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