

Prediction

Ridge regression uses the minimizer of a penalized squared error loss function to estimate the regression coefficients:

$$\hat{\beta} \equiv \operatorname{argmin}_{\beta} \|Y - X\beta\|^2 + \lambda\beta'D\beta.$$

Typically D is a diagonal matrix with 0 in the 1,1 position and ones on the rest of the diagonal. In this case,

$$\beta'D\beta = \sum_{j \geq 1} \beta_j^2.$$

This makes most sense when the covariates have been standardized so it is reasonable to penalize the β_j equally.

Ridge regression is a compromise between fitting the data as well as possible (by making $\|Y - X\beta\|^2$ small), while not allowing any one fitted coefficient to get very large.

For example, if $\beta = (3, 3)$ and $\beta = (4, 2)$ fit the data approximately equally well, then $(3, 3)$ is preferred since it has a smaller penalty.

Ridge regression and colinearity

Suppose X_1 and X_2 are highly positively associated, and the population slopes are β_1 and β_2 .

We have seen that fits of the form

$$(\beta_1 + \gamma)X_1 + (\beta_2 - \gamma)X_2 = EY + \gamma(X_1 - X_2)$$

are similar in MSE, since $X_1 - X_2$ is small in this case.

OLS can't easily distinguish among these fits, but ridge regression strongly prefers the fits with smaller values of γ . The penalty

$$(\beta_1 + \gamma)^2 + (\beta_2 - \gamma)^2$$

is minimized at $\gamma = (\beta_2 - \beta_1)/2$

For a given value $\lambda > 0$, ridge regression is no more difficult computationally than ordinary least squares, since

$$\frac{\partial}{\partial \beta} \|Y - X\beta\|^2 + \lambda \beta' D \beta = -2X'Y + 2X'X\beta + 2\lambda D\beta,$$

so the ridge estimate $\hat{\beta}$ solves

$$(X'X + \lambda D)\beta = X'Y.$$

This equation can have a unique solution even when $X'X$ is singular. Thus one application of ridging is to produce regression estimates for colinear design matrices.

Ridge regression bias and variance

Ridge regression estimates are biased, but may be less variable than OLS estimates. If $X'X$ is non-singular, the ridge estimator can be written

$$\begin{aligned}\hat{\beta}_\lambda &= (X'X + \lambda D)^{-1} X'Y \\ &= (I + \lambda(X'X)^{-1}D)^{-1}(X'X)^{-1}X'Y \\ &= (I + \lambda(X'X)^{-1}D)^{-1}(X'X)^{-1}X'(X\beta + \epsilon) \\ &= (I + \lambda(X'X)^{-1}D)^{-1}\beta + (I + \lambda(X'X)^{-1}D)^{-1}(X'X)^{-1}X'\epsilon.\end{aligned}$$

Thus the bias is

$$E\hat{\beta}_\lambda - \beta = ((I + \lambda(X'X)^{-1}D)^{-1} - I)\beta$$

and the variance is

$$\text{var}\hat{\beta}_\lambda = \sigma^2(I + \lambda(X'X)^{-1}D)^{-1}(X'X)^{-1}(I + \lambda(X'X)^{-1}D)^{-T}.$$

Ridge regression bias and variance (continued)

Next we will show that $\text{var}\hat{\beta} \geq \text{var}\hat{\beta}_\lambda$, in the sense that $\text{var}\hat{\beta} - \text{var}\hat{\beta}_\lambda$ is non-negative definite.

First let $M = \lambda(X'X)^{-1}D$, and note that

$$\begin{aligned}v'(\text{var}\hat{\beta} - \text{var}\hat{\beta}_\lambda)v &\propto v' \left((X'X)^{-1} - (I + M)^{-1}(X'X)^{-1}(I + M)^{-T} \right) v \\ &= u' \left((I + M)(X'X)^{-1}(I + M)' - (X'X)^{-1} \right) u \\ &= u' \left(M(X'X)^{-1} + (X'X)^{-1}M' + M(X'X)^{-1}M' \right) u\end{aligned}$$

where $u = (I + M)^{-T}v$. By direct calculation, all three matrices in the last line are non-negative definite.

We can conclude that for any fixed vector θ ,

$$\text{var}\theta'\hat{\beta}_\lambda \leq \text{var}\theta'\hat{\beta}.$$

Ridge regression effective degrees of freedom

Like OLS, the fitted values under ridge regression are a linear function of the observed values

$$\hat{Y}_\lambda = X(X'X + \lambda D)^{-1}X'Y$$

In OLS regression, the degrees of freedom is the number of free parameters in the model, which is equal to the trace of the projection matrix P that satisfies $\hat{Y} = PY$.

Fitted values in ridge regression are not a projection of Y , but the matrix

$$X(X'X + \lambda D)^{-1}X'$$

plays an analogous role.

Ridge regression effective degrees of freedom (continued)

The “effective degrees of freedom” for ridge regression is defined as

$$\text{EDF}_\lambda = \text{tr} [X(X'X + \lambda D)^{-1}X'] .$$

The trace can be easily computed using the identity

$$\text{trace} (X(X'X + \lambda D)^{-1}X') = \text{trace} ((X'X + \lambda D)^{-1}X'X) .$$

Ridge regression effective degrees of freedom (continued)

EDF_λ is monotonically decreasing in λ . To see this we will use the following fact about matrix derivatives

$$\partial \text{tr}(A^{-1}B) / \partial A = -A^{-T} B' A^{-T}.$$

By the chain rule, letting $A = X'X + \lambda D$, we have

$$\begin{aligned} \partial \text{tr}(A^{-1}X'X) / \partial \lambda &= \sum_{ij} \frac{\partial \text{tr}(A^{-1}X'X)}{\partial A_{ij}} \cdot \frac{\partial A_{ij}}{\partial \lambda} \\ &= - \sum_{ij} [A^{-T}(X'X)A^{-T}]_{ij} \cdot D_{ij} \\ &= - \sum_i [A^{-T}(X'X)A^{-T}]_{ii} \cdot D_{ii} \\ &\leq 0. \end{aligned}$$

Ridge regression effective degrees of freedom (continued)

EDF_λ equals $\text{rank}(X)$ when $\lambda = 0$. To see what happens as $\lambda \rightarrow \infty$, we can apply the Sherman-Morrison-Woodbury identity

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

Let $G = X'X$, and write $D = FF'$, where F has independent columns (usually F will be $p + 1 \times p$ as we do not penalize the intercept).

Ridge regression effective degrees of freedom (continued)

Applying the SMW identity and letting $\lambda \rightarrow \infty$ we get

$$\begin{aligned}\text{tr} [(G + \lambda D)^{-1} G] &= \text{tr} [(G^{-1} - G^{-1} F (I/\lambda + F' G^{-1} F)^{-1} F' G^{-1}) G] \\ &= \text{tr} [I_{p+1} - G^{-1} F (I/\lambda + F' G^{-1} F)^{-1} F'] \\ &\rightarrow \text{tr} I_{p+1} - \text{tr} [G^{-1} F (F' G^{-1} F)^{-1} F'] \\ &\rightarrow \text{tr} I_{p+1} - \text{tr} [(F' G^{-1} F)^{-1} F' G^{-1} F] \\ &= p + 1 - \text{rank}(F).\end{aligned}$$

Therefore in the usual case where F has rank p , EDF_λ converges to 1 as λ grows large, reflecting the fact that all coefficients other than the intercept are forced to zero.

Ridge regression tuning parameter

There are various ways to set the ridge parameter λ .

Generalized cross validation, which minimizes the following over λ , is by far the most common approach.

$$\text{GCV}(\lambda) = \frac{\|Y - \hat{Y}_\lambda\|^2}{(n - \text{EDF}_\lambda)^2}$$