Lecture 15: Linear Model Selection and Regularization

Nayel Bettache

Department of Statistical Science, Cornell University

The Ridge and Lasso

- The ridge regression estimates $\beta_0, \ldots, \beta_p$ using the values that minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2 = RSS + \lambda \sum_{j=1}^{p}\beta_j^2,$$

- The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| = RSS + \lambda \sum_{j=1}^{p}|\beta_j|,$$

  where $\lambda \geq 0$ is a tuning parameter, to be determined later.

A simple Example (Optional)

- Assume that $n = p$ and X is an identity matrix. We force the intercept term $= 0$.
- The usual least squares problem is to find $\beta_1, \ldots, \beta_p$ that minimize

$$\sum_{j=1}^{p}(y_j - \beta_j)^2.$$

  This gives estimator $\hat{\beta}_j = y_j$.

- The ridge regression is to find $\beta_1, \ldots, \beta_p$ that minimize

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

  This gives estimator $\hat{\beta}_j^R = y_j/(1 + \lambda)$.

A simple Example (Optional)

- The lasso is to find $\beta_1, \ldots, \beta_p$ that minimize

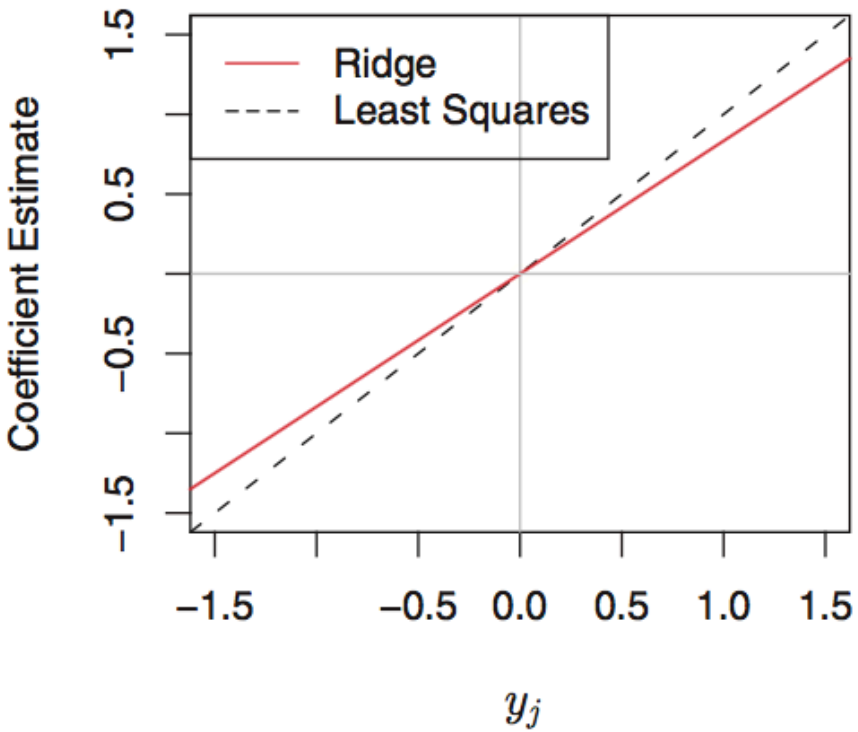
$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

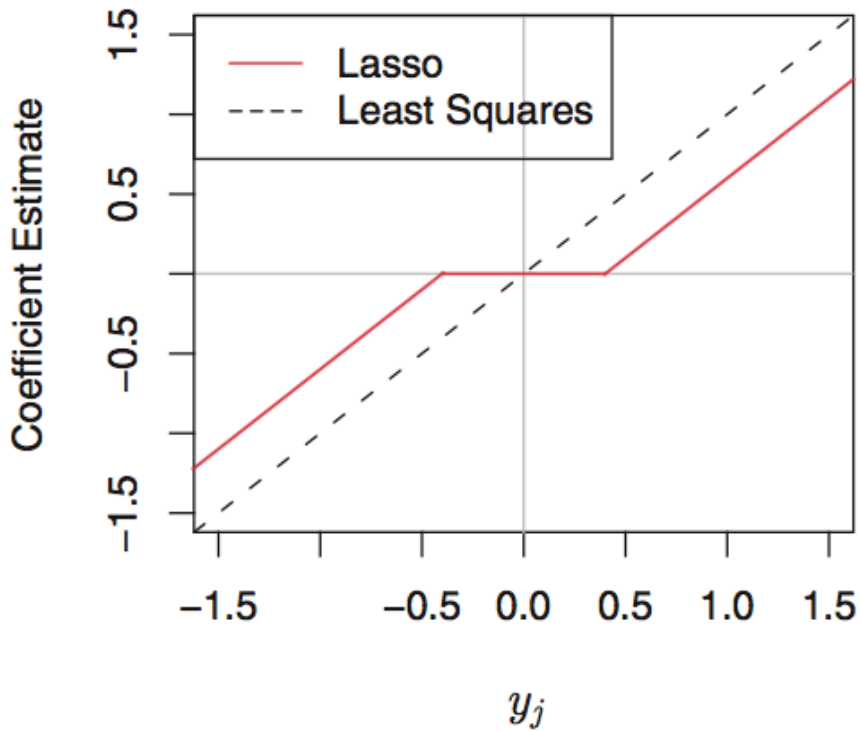
This gives the following estimator

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2; \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2; \\ 0 & \text{if } |y_j| \leq \lambda/2. \end{cases}$$

This is known as **soft-thresholding**.

A simple Example (Optional)

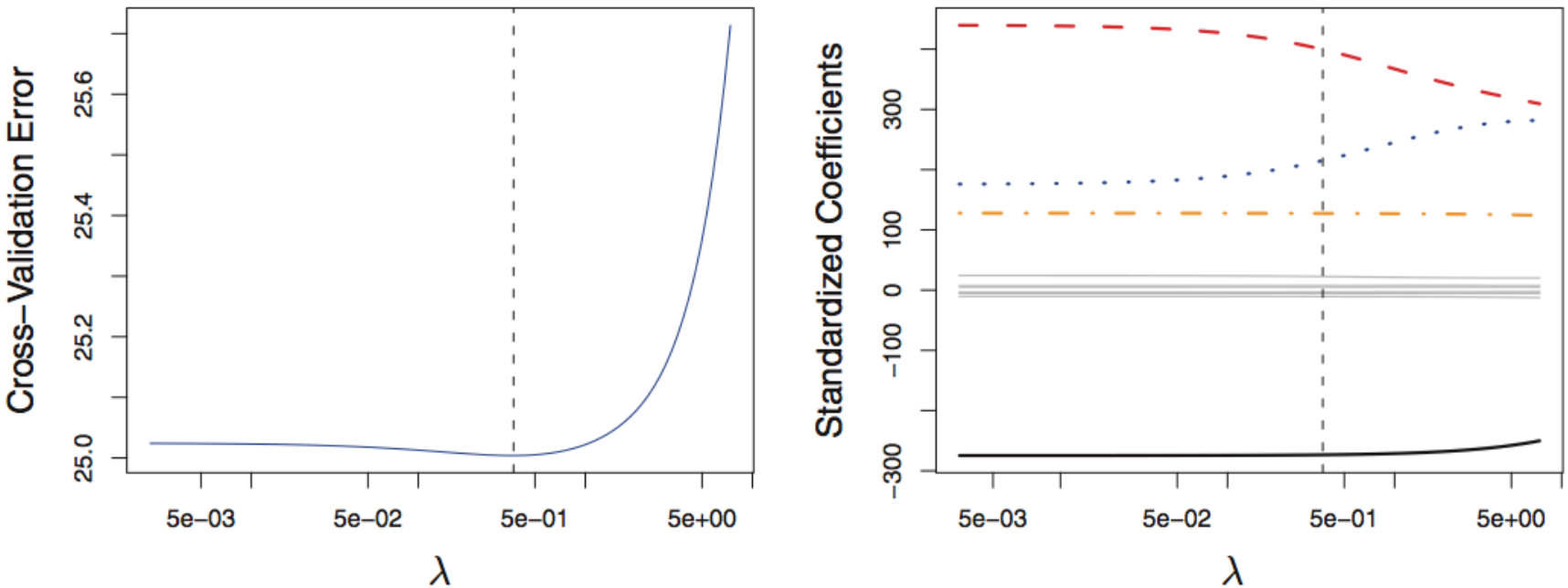




Selecting the Tuning Parameter

- As for subset selection, for ridge regression and lasso we require a method to determine which of the models under consideration is best.
- That is, we require a method selecting a value for the tuning parameter λ or equivalently, the value of the constraint s.
- Cross-validation provides a simple way to tackle this problem. We choose a grid of λ values, and compute the cross-validation error rate for each value of λ.
- We then select the tuning parameter value for which the cross-validation error is smallest.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

Credit Card Data Example



Cross-validation errors that result from applying ridge regression to the Credit data set with various value of λ.

Some Extensions

- The ridge and Lasso regressions can be similarly applied to the logistic regression.
- There are many other different penalty functions. For instance, if we suspect the model is nonlinear, we can add a quadratic terms, say

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \epsilon.$$

We usually use the following **Group Lasso** estimator, which is the minimizer of

$$RSS + \lambda(\sqrt{\beta_1^2 + \beta_2^2} + \sqrt{\beta_3^2 + \beta_4^2}).$$

In this penalty, we view β_1 and β_2 (coefficient of X_1 and X_1^2) as if they belong to the same group. The group Lasso can shrink the parameters in the same group (both β_1 and β_2) exactly to 0.

Regression in High-Dimensional Data

- High dimensional data refers to the data set that the number of features p is much larger than the sample size n.
- Because p is large, we have more extreme collinearity problems.
- Least squares estimates do not exist if $p > n$ or have high variance if $p \approx n$. Don't use least squares for high dimensional data.
- Many of the methods seen in this chapter, such as forward stepwise selection, ridge regression, the lasso, and PCR, may still work in the high-dimensional setting.
- Among these methods, lasso is the most popular and convenient one. Why? Computationally fast; perform feature selection.