Shrinkage Methods

Rosalba Radice

Analytics Methods for Business

Ridge Regression and Lasso

- The subset selection methods use least squares to fit a linear model that contains a subset of the predictors.
- As an alternative, we can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.
- We will be looking at two shrinking methods: ridge regression and lasso.

Ridge Regression

- Recall that the least squares fitting procedure estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ using the values that minimize

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

- In contrast, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

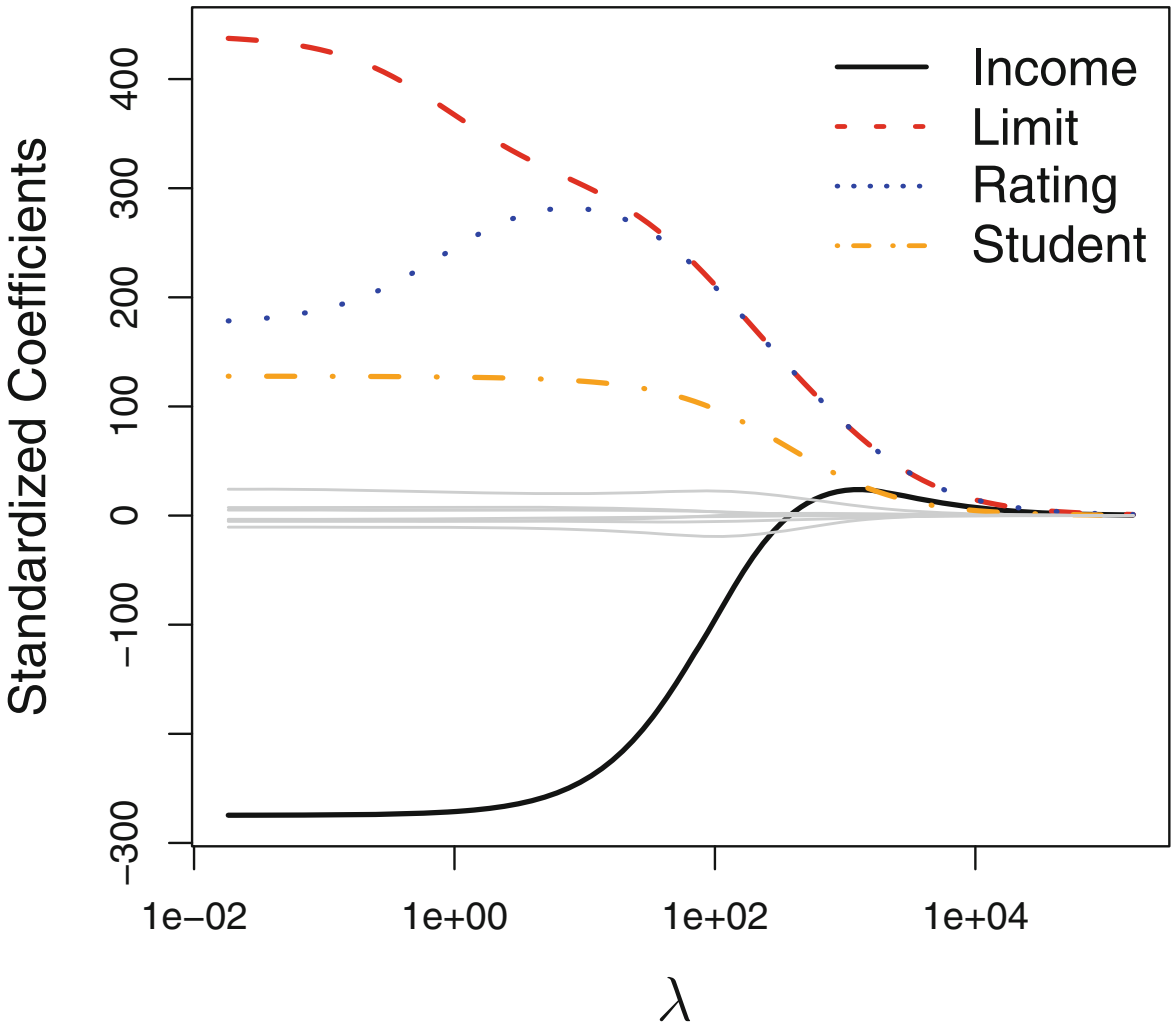
$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

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

Ridge Regression - continued

- As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small.
- However, the second term, $\lambda \sum_{j=1}^{p} \beta_j^2$, called a shrinkage penalty, is small when $\beta_1, \ldots, \beta_p$ are close to zero, and so it has the effect of shrinking the estimates of β_j towards zero.
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.
- Selecting a good value for λ is critical; cross-validation is used for this.

Credit Data Example



The standardized ridge regression coefficients are displayed for the `Credit` data set, as a function of λ.

Ridge Regression: Scaling of Predictors

- The standard least squares coefficient estimates are scale equivariant: multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$. In other words, regardless of how the jth predictor is scaled, $X_j\hat{\beta}_j$ will remain the same.
- In contrast, the ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.
- Therefore, it is best to apply ridge regression after standardizing the predictors, using the formula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}.$$

The Lasso

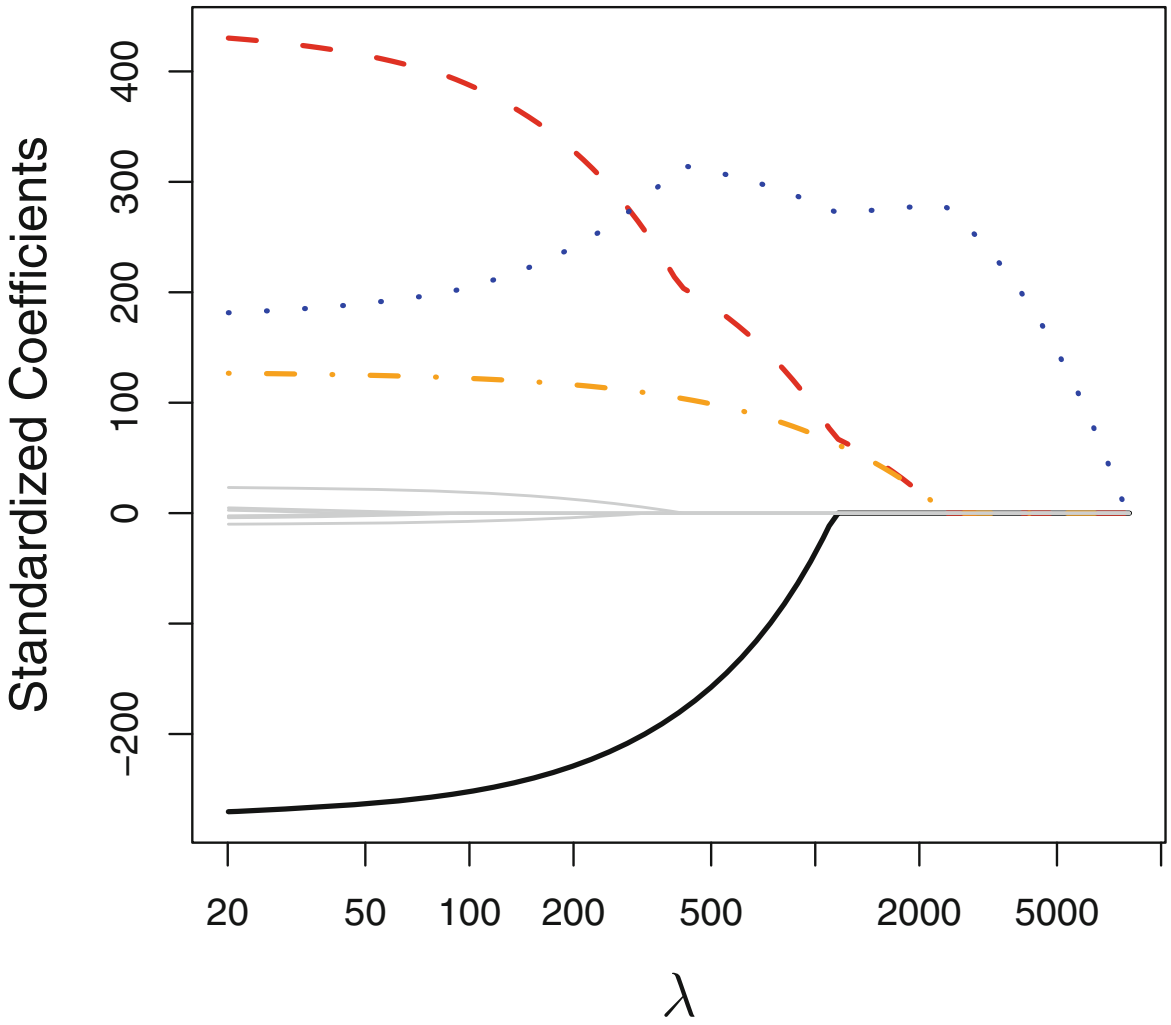
- Ridge regression does have one disadvantage: unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all p predictors in the final model
- The lasso is an alternative to ridge regression that overcomes this disadvantage. The estimated lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity

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

The Lasso - continued

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the lasso, the penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- Hence, much like best subset selection, the lasso performs variable selection.
- We say that the lasso yields sparse models - that is, models that involve only a subset of the variables.
- As in ridge regression, selecting a good value of λ for the lasso is critical; cross-validation is again the method of choice.

Example: Credit Dataset



The standardized lasso coefficients on the `Credit` data set are shown as a function of λ.

Lasso or Ridge Regression?

- In general, one might expect the lasso to perform better when the response is a function of only a relatively small number of predictors.
- However, the number of predictors that is related to the response is never known a priori for real data sets.
- A technique such as cross-validation can be used in order to determine which approach is better on a particular data set.

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.
- 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.

Cross-Validation

- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- Idea is to randomly divide the data into k equal-sized parts. We leave out a part, fit the model to the other $k-1$ parts (combined), and then obtain predictions for the left-out part.
- This is done in turn for each part $j = 1, 2, \ldots, k$, and then the results are combined.

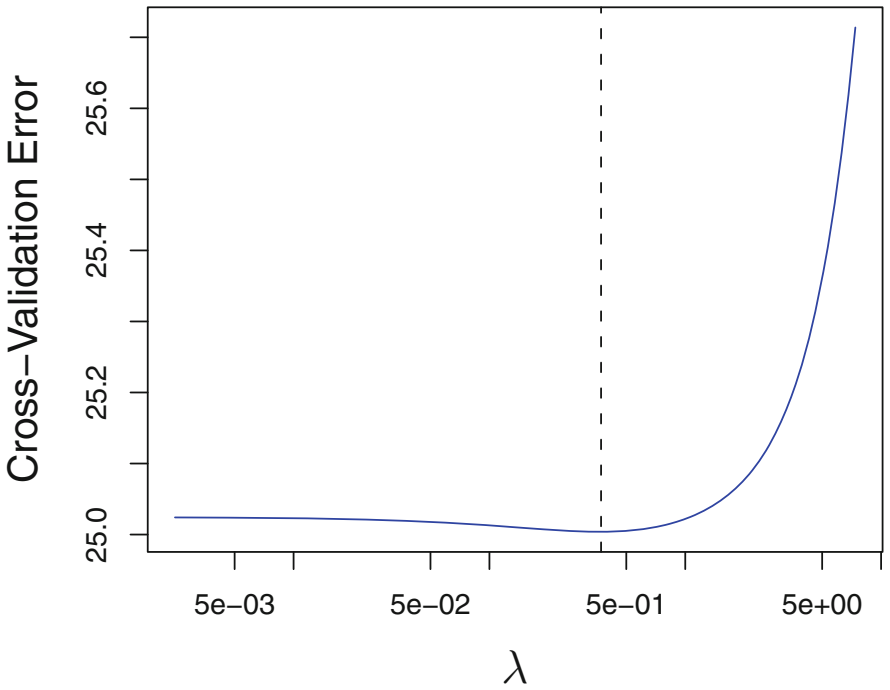
The Details

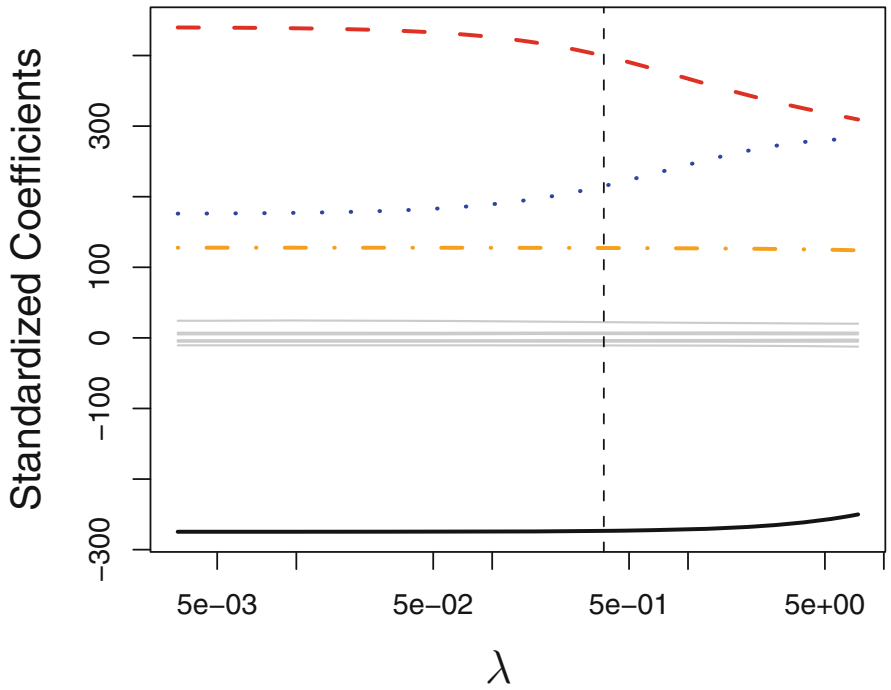
- This approach involves randomly dividing the set of observations into k groups of approximately equal size. The first group is treated as a validation set, and the method is fitted on the remaining $(k-1)$ groups.
- The mean squared error, MSE, is then computed on the observations in the held-out group.
- This procedure is repeated k times; each time, a different group of observations is treated as a validation set.
- This process results in k estimates of $\text{MSE1}_1, \text{MSE1}_2, \ldots, \text{MSE1}_k$. The cross-validation (CV) estimate is computed by averaging these values, that is

$$\text{CV}_k = \frac{1}{k} \sum_{j=1}^{k} \text{MSE}_j.$$

- This procedure is called k-fold cross-validation.
- Setting $k = n$ yields n-fold or leave-one out cross-validation.

Credit Data Example





Left: Cross-validation errors that result from applying ridge regression to the `Credit` data set with various value of λ. Right: The coefficient estimates as a function of λ. The vertical dashed lines indicate the value of λ selected by cross-validation.