NBA 4920/6921 Lecture 13

Shrinkage Methods: Ridge Regression

Murat Unal

Johnson Graduate School of Management

10/14/2021

Agenda

Ridge regression

Application in R

Shrinkage methods

Recall with **subset-selection** methods we

1. Algorithmically search for the best subset of p predictors
2. Use least squares to fit the selected model

Shrinkage methods

In what follows we consider alternatives to least squares because doing so can improve:

1. **Prediction accuracy**, especially for $p > n$
2. **Model interpretability**, by assigning 0 to coefficient estimates of irrelevant features

Shrinkage methods

- Fit a model that contain all p predictors
- At the same time constrain or **regularize** the coefficient estimates
- Regularization **shrinks** the coefficients towards zero

Regularization

- Regularization plays an important role in ML
- ML algorithms typically have a regularizer associated with them
- It allows to measure the complexity of a function/learner
- By choosing the level of regularization appropriately, we can have some benefits of flexible functional forms without having those benefits be overtaken by overfit
- As we regularize less, we do a better job at approximating the in-sample variation, but for the same reason, the wedge between in-sample and out-of-sample fit will typically increase

Shrinkage methods

1. Ridge regression
2. Lasso
3. Elastic net

Ridge regression

Recall that we estimate coefficients $\beta_0, \beta_1, \cdots, \beta_p$ by minimizing RSS

$$\min_{\hat{\beta}} \mathsf{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

Ridge regression makes a small change by adding a shrinkage penalty, the sum of squared coefficients ($\lambda \sum_j \beta_j^2$)

$$\min_{\hat{\beta}^R} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_j \beta_j^2 = \min_{\hat{\beta}} \mathsf{RSS} + \lambda \sum_j \beta_j^2$$

Ridge regression

$$\min_{\hat{\beta^R}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_j \beta_j^2$$

$\lambda >= 0$ is a tuning parameter that determines the magnitude of the penalty

$\lambda = 0 \rightsquigarrow$ no penalty $\rightsquigarrow$ back to least squares

Ridge regression

- Similar to least squares, Ridge regression seeks coefficient estimates that fit the data well, by making the RSS small
- But the shrinkage penalty is small when the coefficients are close to zero, thus it has the effect of shrinking the estimates towards zero
- Each value of λ results in different coefficient estimates, thus selecting a good value is critical
- We typically use cross-validation to choose the optimal λ

Ridge regression

- How does shrinking coefficients towards zero help?
- It's all about the **bias-variance trade-off**.
- Shrinking coefficients reduces the model's variance.

Ridge regression

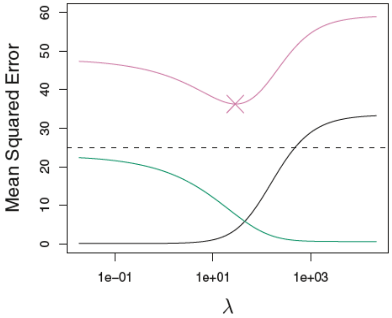
- How does shrinking coefficients towards zero help?
- It's all about the **bias-variance trade-off**.
- Shrinking coefficients reduces the model's variance.
- Think about the extreme case. What happens if all coefficients are zero?

Ridge regression

- How does shrinking coefficients towards zero help?
- It's all about the **bias-variance trade-off**.
- Shrinking coefficients reduces the model's variance.
- Think about the extreme case. What happens if all coefficients are zero?
- We would use the mean outcome to make new predictions. This has zero variance, but large bias.

Ridge regression

- The optimal penalty balances reduced variance with increased bias. $p = 45, n = 50$
- OLS, $\lambda = 0$, will have low bias but high variance
- Ridge regression works best in situations where the least squares estimates have high variance



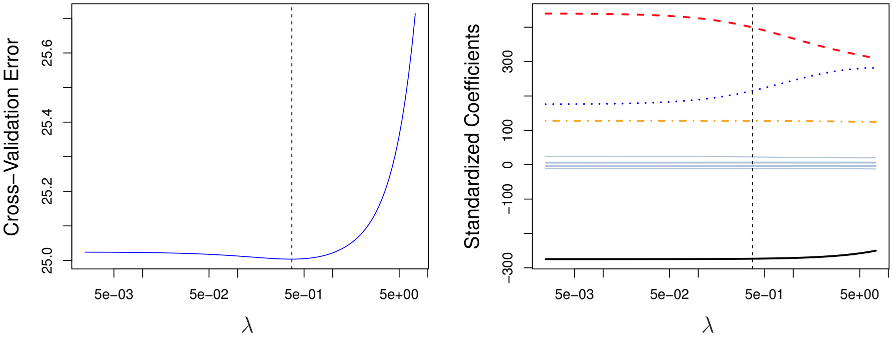
Source: ISL

Selecting the tuning parameter λ

- We perform cross-validation to find the optimum λ
- Start by defining a grid of λ values, and compute the cross-validation error rate for each value of λ
- Select the tuning parameter value for which the cross-validation error is smallest
- Finally, the model is fit again using all of the available observations and the selected value of the tuning parameter.

Selecting the tuning parameter

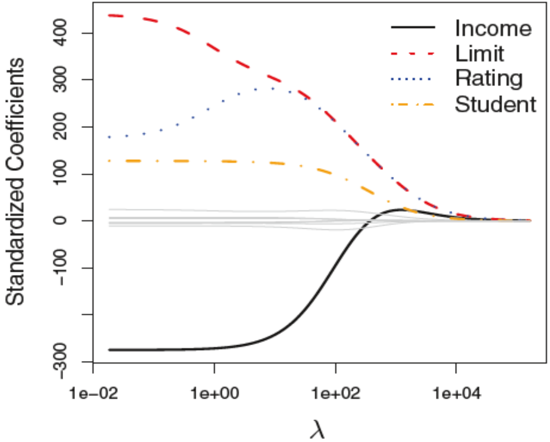
λ Cross-validation errors that result from applying ridge regression to the Credit data set with various value of λ



Source: ISL

Ridge regression

Ridge regression coefficients for the Credit data set, as a function of λ



Source: ISL

Ridge regression

- Least squares estimates are **scale invariant:** multiplying X_j by a constant c leads to a scaling of the coefficient estimates by a factor of $1/c$.
- Regardless how the jth predictor is scaled, $X_j \hat{\beta}_j$ will remain the same.
- If X_j is 1,000 grams and $\beta_j = 5$ then when X_j is 1 kg, $\beta_j = 5000$

Ridge regression

- The same does not apply to Ridge regression.
- Predictors' units can substantially affect ridge regression results.
- Ridge regression pays larger penalty for $\beta_j = 5000$ then $\beta_j = 5$
- Solution: standardize all variables so they are all on the same scale and have standard deviation of 1!

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2017)
An Introduction to Statistical Learning
Springer.
https://www.statlearning.com/

Ed Rubin (2020)
Economics 524 (424): Prediction and Machine-Learning in Econometrics
Univ, of Oregon.