Lecture 13: Regularization
STAT 452, Spring 2021

Introduction

- Last time we discussed **subset selection** methods such as backbwards elimination and stepwise selection. For these types of methods, we identify a subset of the p predictors that we believe are most useful for predicting the response, and then fit a model using least squares on the reduced set of variables.

- Today we discuss an alternative, more modern approach, called **regularization**. For this approach, we fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates.

Review: Least Squares Estimation

The multiple linear regression model is given by:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e \\
&= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e
\end{aligned}
$$

Recall that the regression parameters $\beta_0, \beta_1, \cdots, \beta_p$ can be estimated by minimizing the residuals sum of squares:

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

This process is called *least squares estimation*.

Ridge Regression

In contrast, for ridge regression, the regression parameters $\beta_0, \beta_1, \cdots, \beta_p$ are estimated by minimizing

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

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

- As with least squares estimation, 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 the *shrinkage penalty*, is small when $\beta_1, \cdots, \beta_p$ are close to zero, and so it has the effect of *shrinking* the estimates of β_j towards zero.
- The tuning parameter λ controls the relative impact of these two terms on the regression coefficient estimates.

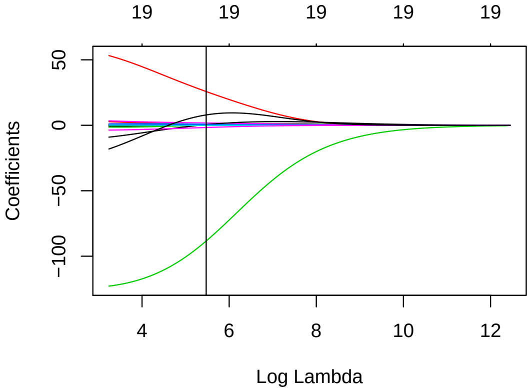
Ridge Regression

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

Properties of ridge regression estimates:

- When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates.
- As $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero.
- Unlike least squares, which generates only one set of coefficient estimates, ridge regression will produce a different set of coefficient estiamtes for each λ.
- λ can be selected (estimated) using software (`glmnet` package in R), which implements some form of cross-validation.

The ridge regression coefficient estimates, for the `Hitters` data set, as a function of the tuning parameter λ. The coefficient estimates corresponding to the value of the λ selected by the software (using cross-validation) is denoted by the black vertical line. Recall for `Hitters` data set, the model is fit with `Salary`, the baseball player's salary (in \$1,000's), as the response, and 19 predictor variables related to the player's performance.



- One disadvantage with ridge regression is that it includes all p predictors in the final model.
- That is, ridge regression will shrink the coefficients towards zero, but it will not set any of them exactly equal to zero.
- This may not be a problem for prediction accuracy, but it can create a challenge in model interpretation when p is large.

The Lasso

The lasso estimates the regression parameters $\beta_0, \beta_1, \cdots, \beta_p$ by minimizing

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

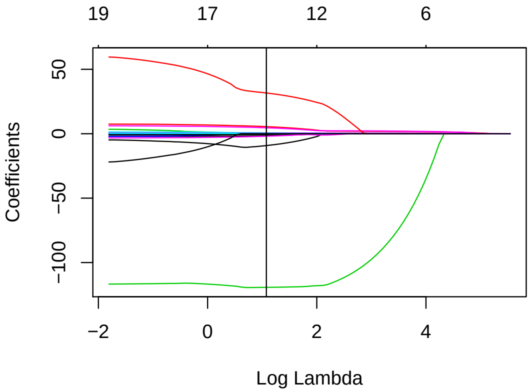
- As with ridge regression, the lasso shrinks the coefficient estimates towards zero.
- However, in the case of the lasso, the penalty term, $\lambda\sum_{j=1}^{p}|\beta_j|$, has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- Hence, the lasso also performs *variable selection*.

The LASSO

Least Absolute Shrinkage and Selection Operator

Alamy/Lisa Dearing

The lasso coefficient estimates, for the `Hitters` data set, as a function of the tuning parameter λ. The coefficient estimates corresponding to the value of λ selected by the software (using cross-validation) is denoted by the black vertical line.



Another Formulation

For the lasso and ridge regression the parameters $\beta_0, \beta_1, \cdots, \beta_p$ are found by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq s$$

and

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}\beta_j^2 \leq s$$

respectively.

Summary

- Both ridge regression and the lasso shrink the regression coefficient estimates towards zero, relative to the least squares estimates.
- The lasso forces some coefficients to zero, and so it performs variables selection as well.
- As a results, models estimated using the lasso tend to be much easier to interpret than ridge regression, especially when p is large.
- Both the lasso and ridge regression can potentially perform better than ordinary least squares on withheld test data. Thus, regularization can improve predictive performance.