Ch 6.2: Shrinkage - Ridge regression

Lecture 17 - CMSE 381

Prof. Elizabeth Munch

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Fri, Oct 11, 2024

Announcements

Last time:

- Subset selection

This time:

- Ridge regression

Announcements:

- HW #5 due Friday 10/18

Lec #	Date			Reading	HW
12	Mon	9/30	Leave one out CV	5.1.1, 5.1.2	
13	Wed	10/2	k-fold CV	5.1.3	
14	Fri	10/4	More k-fold CV,	5.1.4-5	
15	Mon	10/7	k-fold CV for classification	5.1.5	
16	Wed	10/9	Subset selection	6.1	HW #4 Due Weds 10/9
17	Fri	10/11	Shrinkage: Ridge	6.2.1	
18	Mon	10/14	Shrinkage: Lasso	6.2.2	
19	Wed	10/16	Dimension Reduction	6.3	
20	Fri	10/18	Overflow, Possibly more dimension reduction?		HW #5 Due Fri 10/18
	Mon	10/21	No class - Fall break		
	Wed	10/23	**Review**		
	Fri	10/25	**Midterm #2**		

Section 1

Last time

Subset selection

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k - 1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Section 2

Ridge Regression

Goal

- Fit model using all p predictors
- Aim to constrain (regularize) coefficient estimates
- Shrink the coefficient estimates towards 0

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Ridge regression

Before:

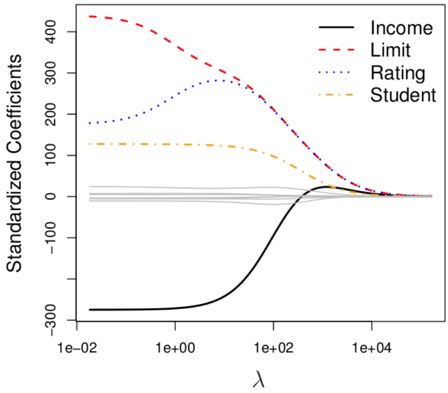
$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

After:

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

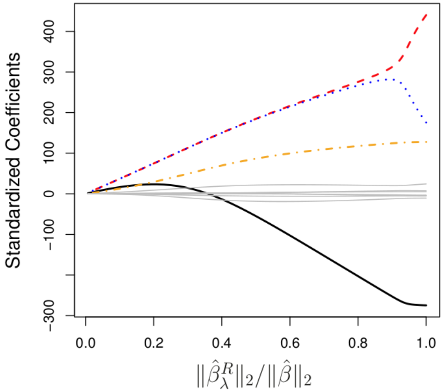
Example from the `Credit` data

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$



Same Setting, Different Plot

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad \|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$$



Scale equivavariance (or lack thereof)

Scale equivariant: Multiplying a variable by c (cX_i) just returns a coefficient multiplied by $1/c$ ($1/c\beta_i$)

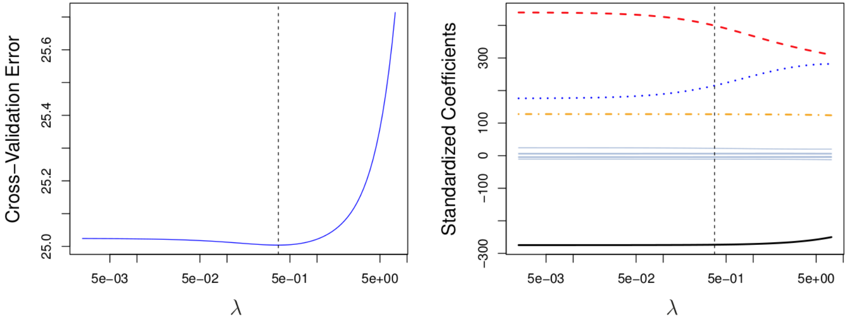
Solution: Standardize predictors

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

Using Cross-Validation to find λ

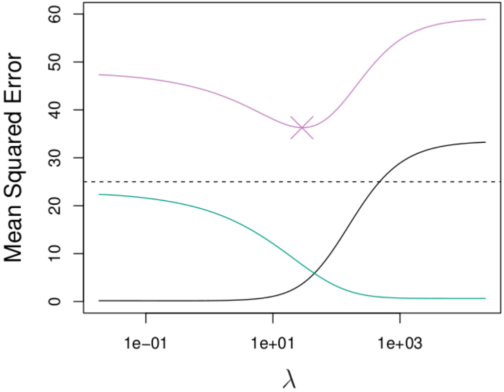
- Choose a grid of λ values
- Compute the (k-fold) cross-validation error for each value of λ
- Select the tuning parameter value λ for which the CV error is smallest.
- The model is re-fit using all of the available observations and the selected value of the tuning parameter.

LOOCV choice of λ for ridge regression and Credit data



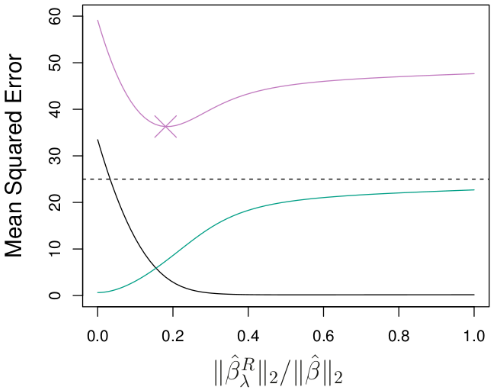
Coding

Bias-Variance tradeoff



Squared bias (black), variance (green), and test mean squared error (purple) for simulated data.

More Bias-Variance Tradeoff



Squared bias (black), variance (green), and test mean squared error (purple) for simulated data.

Advantages of Ridge

Ridge vs. Least Squares:

Ridge vs. Subset Selection:

Next time

Lec #	Date			Reading	HW
12	Mon	9/30	Leave one out CV	5.1.1, 5.1.2	
13	Wed	10/2	k-fold CV	5.1.3	
14	Fri	10/4	More k-fold CV,	5.1.4-5	
15	Mon	10/7	k-fold CV for classification	5.1.5	
16	Wed	10/9	Subset selection	6.1	HW #4 Due Weds 10/9
17	Fri	10/11	Shrinkage: Ridge	6.2.1	
18	Mon	10/14	Shrinkage: Lasso	6.2.2	
19	Wed	10/16	Dimension Reduction	6.3	
20	Fri	10/18	Overflow, Possibly more dimension reduction?		HW #5 Due Fri 10/18
	Mon	10/21	No class - Fall break		
	Wed	10/23	**Review**		
	Fri	10/25	**Midterm #2**		