Ch 6.2: Shrinkage - Ridge regression

Lecture 18 - CMSE 381

Prof. Elizabeth Munch

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

Mon, Oct 16, 2023

Announcements

Last time:

- Subset selection

This time:

- Ridge regression

Announcements:

- HW #5 due Wednesday
- Be sure to make note of people you worked with and resources you used.

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 $k = 1, 2, \dots p$:
   (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 $\mathcal{M}_0, \dots, \mathcal{M}_p$ 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 $k = 0, \dots, p-1$:
   (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 $\mathcal{M}_0, \dots, \mathcal{M}_p$ 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 $k = p, p-1, \dots, 1$:
   (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 $\mathcal{M}_0, \dots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Fixing the code from the notebook

```python
Ms = []

# For each of sizes k
for k in range(1,5):
    myvars = []
    myscores = []

    # Take all the size k subsets of input variables
    for Xs in combinations(inputvars,k):

        # use the training score as the quality measure
        myvars.append(Xs)
        myscores.append(myscore_train(auto,Xs))

    myResults = pd.DataFrame({'Vars':myvars, 'TrainScore':myscores})
    print('\n k:', k)
    print(myResults)


    indexmin = myResults.idxmin(numeric_only = True)
    Ms.append(myResults.Vars[indexmin].iloc[0])


print('\n---\n')
for k in range(1,5):
    print('M_'+str(k), Ms[k-1])
```

```
 k: 1
                 Vars  TrainScore
0        (cylinders,)   24.020180
1       (horsepower,)   23.943663
2           (weight,)   18.676617
3     (acceleration,)   49.873627

 k: 2
                          Vars  TrainScore
0      (cylinders, horsepower)   20.848190
1          (cylinders, weight)   18.382946
2    (cylinders, acceleration)   23.942447
3         (horsepower, weight)   17.841442
4   (horsepower, acceleration)   22.461644
5       (weight, acceleration)   18.247176

 k: 3
                                      Vars  TrainScore
0          (cylinders, horsepower, weight)   17.763871
1    (cylinders, horsepower, acceleration)   20.055715
2        (cylinders, weight, acceleration)   18.126486
3       (horsepower, weight, acceleration)   17.841430

 k: 4
                                              Vars  TrainScore
0  (cylinders, horsepower, weight, acceleration)     17.7614

---

M_1 ('weight',)
M_2 ('horsepower', 'weight')
M_3 ('cylinders', 'horsepower', 'weight')
M_4 ('cylinders', 'horsepower', 'weight', 'acceleration')
```

More fixing

In [29]:
```
##ANSWER##

testscores = []

# Use kfold cv to get the test score to make the final judgement
for X in Ms:
    testscores.append(myscore_cv(auto,X))

myResultsM = pd.DataFrame({'Vars':Ms, 'TestScore':testscores})

myResultsM
```

Out[29]:

	Vars	TestScore
0	(weight,)	18.844627
1	(horsepower, weight)	18.108011
2	(cylinders, horsepower, weight)	18.200516
3	(cylinders, horsepower, weight, acceleration)	18.316760

In [30]:
```
##ANSWER##
indexmin = myResultsM.idxmin(numeric_only = True)
print('Best Model:', myResultsM.Vars[indexmin])
```

```
Best Model: 1    (horsepower, weight)
Name: Vars, dtype: object
```

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:

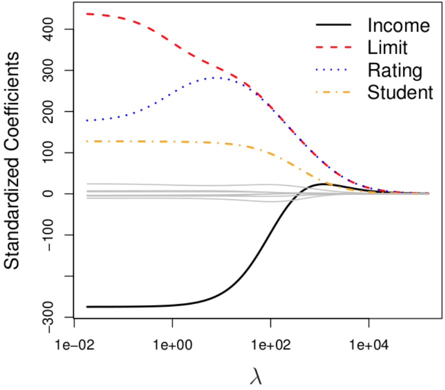
$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

After:

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

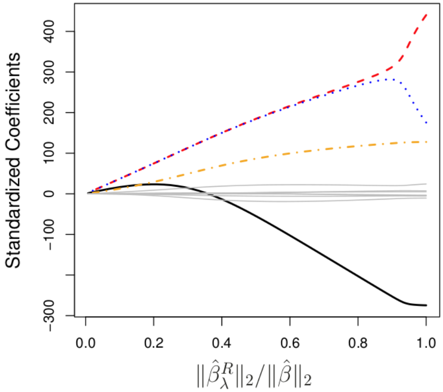
Example from the Credit data

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



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 $1/c$ ($1/c\beta_i$)

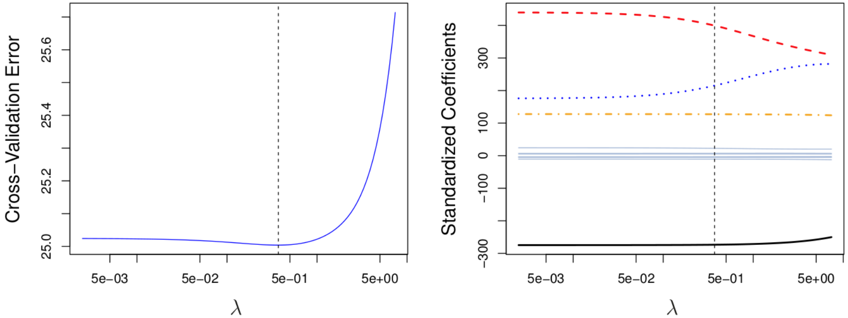
Solution: Standardize predictors

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

Using Cross-Validation to find λ

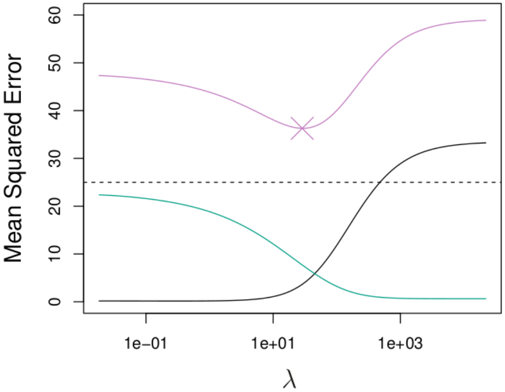
- Choose a grid of λ values
- Compute the (k-fold) cross-validation error for each value of λ
- Select the tuning parameter value λ for which the CV error is smallest.
- The model is re-fit using all of the available observations and the selected value of the tuning parameter.

LOOCV choice of λ for ridge regression and `Credit` data



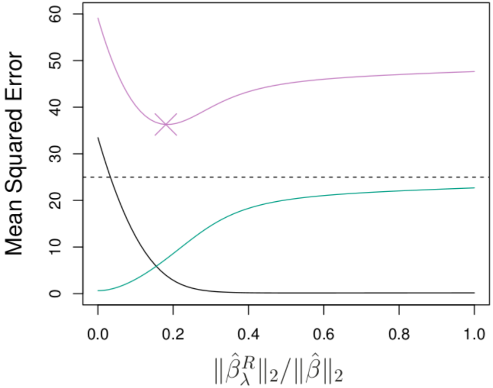
Coding

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

12	Mon	Oct 2	Leave one out CV	5.1.1, 5.1.2	
13	Wed	Oct 4	k-fold CV	5.1.3	
14	Fri	Oct 6	More k-fold CV,	5.1.4-5	
15	Mon	Oct 9	k-fold CV for classification	5.1.5	HW #4 Due
16	Wed	Oct 11	Resampling methods: Bootstrap	5.2	
17	Fri	Oct 13	Subset selection	6.1	
18	Mon	Oct 16	Shrinkage: Ridge	6.2.1	
19	Wed	Oct 18	Shrinkage: Lasso	6.2.2	
	Fri	Oct 20	**Review**		
	Mon	Oct 23	No class - Fall break		
	Wed	Oct 25	**Midterm #2**		
20	Fri	Oct 27	Dimension Reduction	6.3	