SHRINKAGE REGRESSION: RIDGE & LASSO

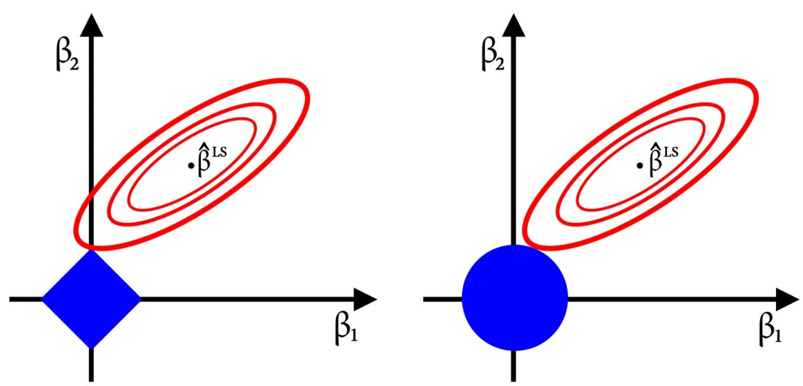
Anaranya Basu

April 2024

1 Shrinkage Regression : An Overview

One of the most important problems in regression analysis is the selection of predictors in the model. The reason behind this is we are often not satisfied with the least squares estimates. However, all of the methods were containing a subset of the predictors. As an alternative to this, it is possible to fit a model containing all p predictors by constraining or regularizing the coefficient estimates or shrinking the coefficient estimates towards zero.

There are two best-known technique's for shrinking the regression coefficients towards zero are ridge and lasso regression



The Mathematical expression for **RIDGE** regression is

$$\min_\beta \left(\sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right)$$

Where:

- λ is the regularization parameter (also known as the ridge penalty parameter).
- The first term $\sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2$ represents the residual sum of squares (RSS), which measures the discrepancy between the observed response and the predicted response.

- The second term $\lambda \sum_{j=1}^{p} \beta_j^2$ is the penalty term, also known as the ridge penalty, which penalizes large values of coefficients β to prevent overfitting.
- The parameter λ controls the trade-off between fitting the data well and keeping the coefficients small.
- The goal of ridge regression is to find the values of β that minimize this objective function.

2 Data Description

Data has been collected from

```
Data: http://www.stat.uchicago.edu/~yibi/s224/data/P236.txt
```

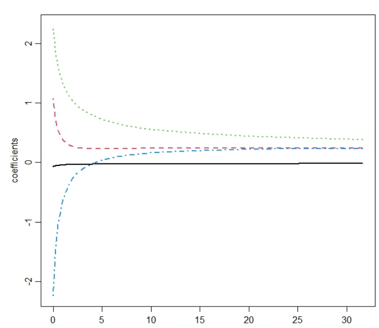
- ACHV: Student achievement index (higher values are better)
- FAM: Faculty credentials index
- PEER: the influence of their peer group in the school
- SCHOOL: School facility/resource index

3 Analysis

The lambda (λ) value(s) must be specified. The following gives the Ridge estimates for the intercept β_0 and the coefficients β_j for FAM, PEER, and SCHOOL for $\lambda = 1$, 5, and 10 respectively.

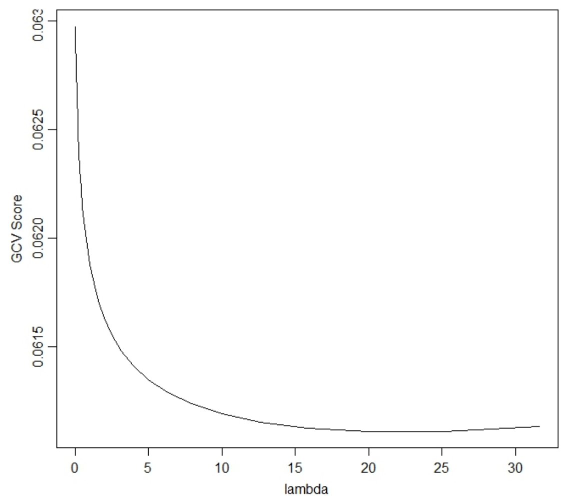
```
 > lm.ridge(ACHV~FAM + PEER + SCHOOL,data=EEO,lambda = c(1,5,10))
                          FAM      PEER     SCHOOL
 1 -0.04055397 0.3768768 1.3205433 -0.6276745
 5 -0.02708020 0.2318348 0.7229545  0.0419616
10 -0.02354968 0.2383780 0.5567732  0.1624044
```

We can try more values of lambda and plot how the coefficients shrink as lambda grows larger:



3.1 Selecting the λ using CV

For each λ, the lm.ridge() function computes the generalized cross-validation (GCV), similar to cross-validation using RMSE based on training data and test data.



The best lambda (among those lambda's specified in EEO.rg) can be selected automatically to be 19.95.

```
 > select(EEO.rg)
modified HKB estimator is 0.3785843
modified L-W estimator is 4.081519
smallest value of GCV  at 19.95262
```

4 Final Prediction

Setting lambda at the optimal value 19.95 that minimize the GCV, the Ridge estimates for coefficients of the EEO data can be obtained as follows:

```
    > lm.ridge(ACHV ~ FAM + PEER + SCHOOL, data=EEO, lambda=19.95)
                    FAM        PEER      SCHOOL
-0.02033817  0.24403336  0.44263995  0.21866867
```

The Ridge estimates of the 3 coefficients are all positive, which makes more sense than the OLS estimates below that asserts better SCHOOL facility has a negative impact on students' performance.

```
    > lm.ridge(ACHV~FAM + PEER + SCHOOL,data=EEO)$coef
      FAM      PEER    SCHOOL
 1.184281  2.132042 -2.317950
```

The 3 Ridge estimates all have smaller magnitudes than corresponding OLS estimates.

5 LASSO REGRESSION

Ridge regression, unlike feature selection methods that explained in this, select models that that involve just a subset of the variables, will include all predictors in the final model. With lambda parameter, it is possible to shrink all of the coefficients towards zero, but it will not set any of them exactly to zero. This is actually not a problem for prediction accuracy, however, it can create a challenge in model interpretation in settings in which number of predictor is quite large.

Lasso Regression is a relatively recent alternative to ridge regression that overcomes this disadvantage. Lasso estimates betas using the values that minimize the following formula:

$$\min_\beta \left(\frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

Given a dataset with n observations and p predictors, let $\mathbf{X}$ be the $n \times p$ matrix of predictor variables, $\mathbf{y}$ be the $n \times 1$ vector of response variables, and $\boldsymbol{\beta}$ be the $p \times 1$ vector of regression coefficients. In the case of lasso, L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when lambda is sufficiently large. In other words, just like feature selection methods, lasso select features. Thus, the model generated by lasso is much easier to interpret. Again, to determine lambda, we can use cross validation.

5.1 Data Description

215 samples of finely chopped pure meat sample has been taken.

A Tecator near-infrared spectrometer was used to measure the spectrum of light transmitted through each sample of meat. The spectrum gives the absorbance at 100 wavelengths in the range 850-1050 nm. Since determining the fat content via analytical chemistry is time consuming, we would like to build a model to predict the fat content of new samples using the 100 absorbance which can be measured more easily. The first 100 variables are the 100 absorbances of different wave lengths. The 101th variable fat is the fat content determined via analytical chemistry. We first split the meatspec data into training data and test data.

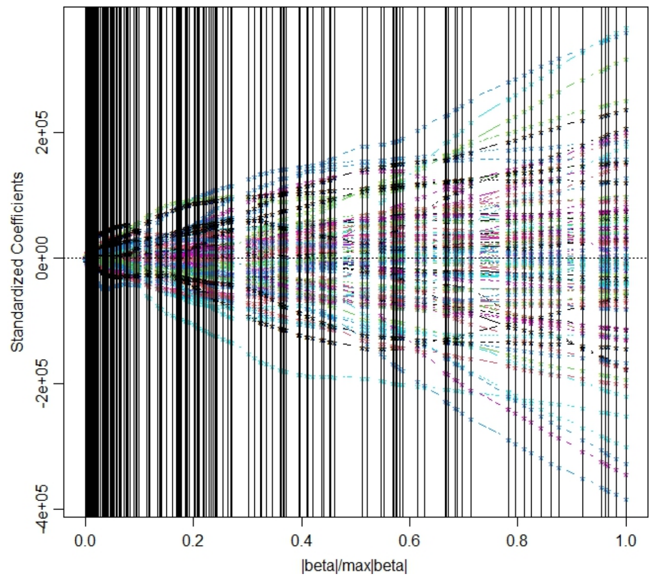
```
train=meatspec[1:172,]
test=meatspec[173:215,]
```

5.2 Analysis

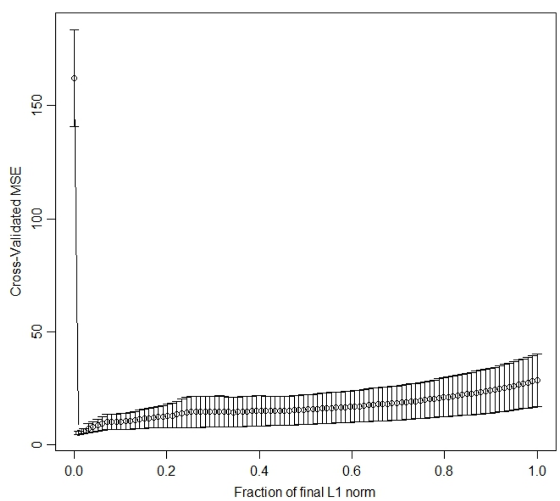
We compute the Lasso fit for the training data :

```
train=meatspec[1:172,]
test=meatspec[173:215,]
trainy = train$fat
trainx = as.matrix(train[,-101])
library(lars)
lassomod = lars(trainx,trainy)
```

Below is the plot of the estimated coefficients as a function of t



Now using the Cross Validation plot we will try to understand how the coefficients shrinks with the optimal lamda value,



Now we need to knife out the optimal value of lambda,

```
> cvout$index[which.min(cvout$cv)]
[1] 0.01010101
```

The best t selected by cross-validation is t = 0.0101

5.3 Final Prediction

Setting t at the optimal value 0.0101 determined by cross-validation, the Lasso estimates for coefficients of the meat data can be obtained as follows:

```
pred$coefficients
        V1          V2          V3          V4          V5          V6
   0.00000  -137.11529     0.00000     0.00000     0.00000     0.00000
        V7          V8          V9         V10         V11         V12
   0.00000     0.00000     0.00000     0.00000     0.00000   249.46803
       V13         V14         V15         V16         V17         V18
   0.00000     0.00000     0.00000     0.00000     0.00000     0.00000
       V19         V20         V21         V22         V23         V24
   0.00000     0.00000     0.00000     0.00000     0.00000  -266.13292
       V25         V26         V27         V28         V29         V30
   0.00000     0.00000     0.00000     0.00000     0.00000  1827.77437
       V31         V32         V33         V34         V35         V36
   0.00000     0.00000     0.00000 -4255.98243     0.00000     0.00000
       V37         V38         V39         V40         V41         V42
1931.22334  1384.07825     0.00000     0.00000     0.00000 -1202.83609
       V43         V44         V45         V46         V47         V48
   0.00000     0.00000   868.02355   323.68648   132.35110     0.00000
       V49         V50         V51         V52         V53         V54
-1102.97154   -15.53693     0.00000     0.00000     0.00000   189.47433
       V55         V56         V57         V58         V59         V60
   0.00000     0.00000     0.00000     0.00000     0.00000     0.00000
       V61         V62         V63         V64         V65         V66
 205.21985     0.00000     0.00000     0.00000     0.00000     0.00000
       V67         V68         V69         V70         V71         V72
   0.00000     0.00000     0.00000     0.00000  -223.73059     0.00000
       V73         V74         V75         V76         V77         V78
   0.00000     0.00000     0.00000     0.00000     0.00000     0.00000
       V79         V80         V81         V82         V83         V84
  80.82735     0.00000     0.00000     0.00000     0.00000     0.00000
       V85         V86         V87         V88         V89         V90
   0.00000     0.00000     0.00000     0.00000    27.26149     0.00000
       V91         V92         V93         V94         V95         V96
   0.00000     0.00000     0.00000     0.00000     0.00000   -96.96579
       V97         V98         V99        V100
   0.00000     0.00000     0.00000    81.72160
```

we can see that only 20 coefficients have non zero LASSO estimates.

```
sum(pred$coefficients!=0)
[1] 20
```

Here are the 20 variables with non zero estimates :

```
> pred$coefficients[pred$coefficients != 0]
         V2         V12         V24         V30         V34         V37 
 -137.11529   249.46803  -266.13292  1827.77437 -4255.98243  1931.22334 
        V38         V42         V45         V46         V47         V49 
 1384.07825 -1202.83609   868.02355   323.68648   132.35110 -1102.97154 
        V50         V54         V61         V71         V79         V89 
  -15.53693   189.47433   205.21985  -223.73059    80.82735    27.26149 
        V96        V100 
  -96.96579    81.72160 
```