Linear Statistical Modeling Methods with SAS

Linear Model Regularization - Part I

Xuemao Zhang
East Stroudsburg University

April 8, 2024

Outline

- Ridge Regression
- Lasso Regression
- An Example of Ridge Regression
- An Example of Lasso Regression

Introduction

- The subset selection methods (Forward, Backward ...) use least squares to fit a linear model that contains a subset of the predictors.
- As an alternative, we can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.
- This is known as regularization or penalization.
- It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance.

Introduction

Crazy Coefficients

- When $p > n$, some of the variables are highly correlated.
- Why does correlation matter?
  - Suppose that X_1 and X_2 are highly correlated with each other... assume $X_1 = X_2$ for the sake of argument.
  - And suppose that the least squares model is

$$\hat{y} = X_1 - 2X_2 + 3X_3$$

  - Then this is also a least squares model (see a simulation study next slide):

$$\hat{y} = 100000001X_1 - 100000002X_2 + 3X_3$$

- Bottom Line: When there are too many variables, the least squares coefficients can get crazy!
- This craziness is directly responsible for poor test error.
- It amounts to too much model complexity.

Introduction

A simulation study of Crazy Coefficients

- Generation of correlated data

```
data SimData;
   call streaminit(123); /* Set random seed for reproducibility */
   do i = 1 to 20;
      x1 = rand('Normal', 0, 15);
      x2 = rand('Normal', x1, 0.001);
      x3 = rand('Uniform');
      y = rand('Normal', -x1 + 3 * x3);
      output;
   end;
run;
```

- Fit a the full model with correlation matrix

```
proc reg data=SimData corr;
   model y = x1 x2 x3 / noint;
run;
```

Ridge Regression

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_p$ using the values that minimize

$$SSE = \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

- In contrast, the ridge regression coefficient estimates $\hat{\beta}^R$ are the values that minimize

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

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

Ridge Regression

- Equivalently, we find $\hat{\beta}^R$ that minimizes

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2$$

subject to the constraint that

$$\sum_{j=1}^{p} \beta_j^2 < s$$

for some s.

Ridge Regression

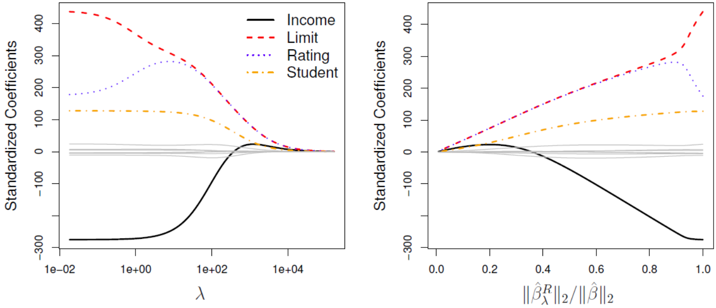
- Ridge regression coefficient estimates minimize

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- When $\lambda = 0$, then ridge regression is just the same as least squares.
- As λ increases, then $\sum_{j=1}^{p} (\hat{\beta}_{\lambda,j}^R)^2$ decreases - i.e. coefficients become shrunken towards zero.
- As $\lambda \to \infty$, $\hat{\boldsymbol{\beta}}^R = 0$.

Ridge Regression

- Ridge Regression As λ Varies: The standardized ridge regression coefficients are displayed for the Credit data set.



Ridge Regression

Ridge regression: scaling of predictors

- The standard least squares coeffcient estimates are scale equivariant: multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$. In other words, regardless of how the jth predictor is scaled, $X_j\hat{\beta}_j$ will remain the same.
- In contrast, the ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the ridge regression objective function.
- Therefore, it is best to apply ridge regression after standardizing the predictors, using the formula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}, i = 1, \ldots, n, j = 1, \ldots, p$$

Ridge Regression In Practice

- Perform ridge regression for a very fine grid of λ values.
- Use cross-validation or the validation set approach to select the optimal value of λ - that is, the best level of model complexity.
- Perform ridge on the full data set, using that value of λ.

Drawbacks of Ridge

- Ridge regression is a simple idea and has a number of attractive properties: for instance, you can continuously control model complexity through the tuning parameter λ.
- But it suffers in terms of model interpretability, since the final model contains all p variables, no matter what.
- We Often want a simpler model involving a subset of the features.
- The lasso involves performing a little tweak to ridge regression so that the resulting model contains mostly zeros.
- In other words, the resulting model is sparse. We say that the lasso performs feature selection.

Lasso Regression

- The lasso involves finding β that minimizes

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

  where $\lambda > 0$ is a tuning parameter.

- Equivalently, we find $\hat{\beta}^L$ that minimizes

$$\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2$$

  subject to the constraint that

$$\sum_{j=1}^{p} |\beta_j| < s$$

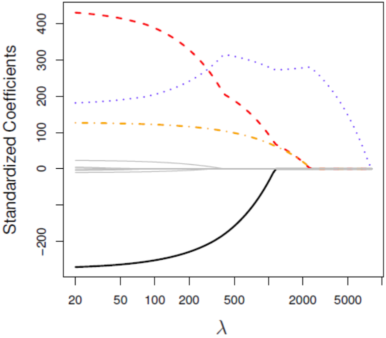
  for some s.

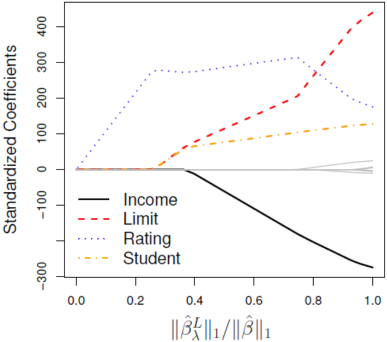
Lasso Regression

- Lasso is a lot like ridge:
  - λ is a nonnegative tuning parameter that controls model complexity.
  - When $\lambda = 0$, we get least squares.
  - When λ is very large, we get $\hat{\beta}^L = 0$.
- But unlike ridge, lasso will give some coefficients exactly equal to zero for intermediate values of λ!
- Hence, much like best subset selection, the lasso performs variable selection.
- We say that the lasso yields sparse models - that is, models that involve only a subset of the variables.

Lasso Regression

- Lasso Regression As λ Varies: The Lasso regression coefficients are displayed for the Credit data set.





Lasso Regression In Practice

- Perform lasso for a very fine grid of λ values.
- Use cross-validation or the validation set approach to select the optimal value of λ - that is, the best level of model complexity.
- Perform the lasso on the full data set, using that value of λ.

Ridge and Lasso: A Geometric Interpretation

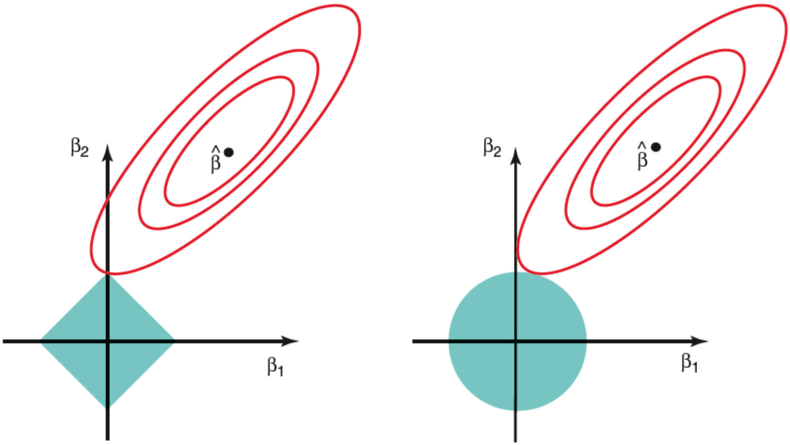


FIGURE *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$*, while the red ellipses are the contours of the RSS.*

Example: Ridge Regression

- We predict `Salary` on the `Hitters` data https://rdrr.io/cran/ISLR/man/Hitters.html.

```
PROC IMPORT
DATAFILE='/home/u5235839/my_shared_file_links/u5235839/Hitters.csv'
    DBMS=CSV
    OUT=Hitters;
    GETNAMES=YES;
RUN;



proc contents data=Hitters;
run;
```

- There are three categorical variables: `Division`, `League` and `NewLeague`

```
proc freq data=Hitters;
tables Division League  NewLeague;
run;
```

Example: Ridge Regression

- Convert the three variables to categorical

```
data Hitters1;
set Hitters;
/* Create binary indicators */
Division_1 = (Division = 'E');
League_A= (League= 'A');
NewLeague_A= (NewLeague='A');
run;
```

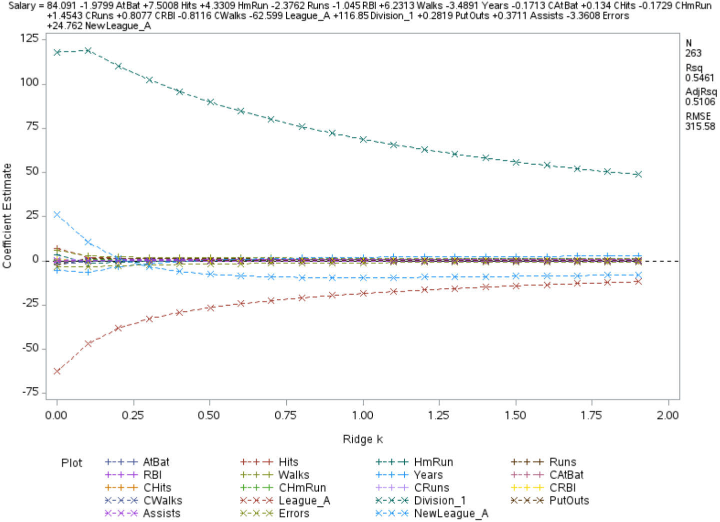
Example: Ridge Regression

- Here, the **outstb** option in the proc statement tells SAS to put the parameter estimates in the output temp.
- The **outvif** option in the proc statement of the regression tells SAS to put the VIF's in the output temp.

```
proc reg data = Hitters1 outest = temp outstb noprint;
model Salary = AtBat    Hits    HmRun   Runs    RBI Walks
Years   CAtBat  CHits   CHmRun  CRuns   CRBI    CWalks  League_A
Division_1  PutOuts Assists Errors  NewLeague_A
 /  ridge = (0.001 to 2 by .1) outvif;
plot / ridgeplot vref=0;
run;
```

- We can see the `ridgeplot` which shows the parameter estimates for different values of λ.

Example: Ridge Regression



Example: Ridge Regression

```
proc print data=temp;
run;
```

- Check parameter estimates

```
proc print data = temp;
where _type_ = 'RIDGESTB';
var _ridge_ AtBat   Hits    HmRun   Runs    RBI Walks
Years   CAtBat  CHits   CHmRun  CRuns   CRBI    CWalks  League_A
Division_1  PutOuts Assists Errors  NewLeague_A;
run;
```

- Check the values of VIF

```
proc print data = temp;
where _type_ = 'RIDGEVIF';
var _ridge_ AtBat   Hits    HmRun   Runs    RBI Walks
Years   CAtBat  CHits   CHmRun  CRuns   CRBI    CWalks  League_A
Division_1  PutOuts Assists Errors  NewLeague_A;
run;
```

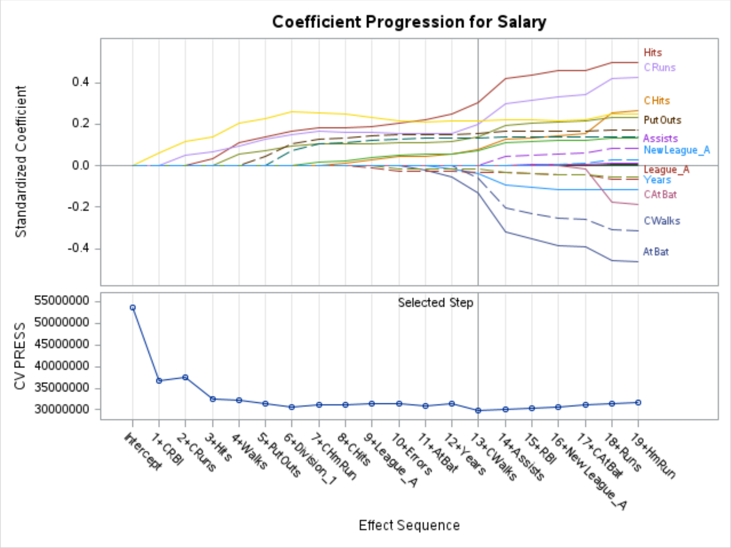
Example: Ridge Regression

- From the ridge plot, we can see that the parameter estiamtes stablize when λ becomes larger
  - Unfortunately, SAS does not provide CV method to choose the best λ for Ridge regression
- **Elastic Net** is a regularization technique that combines both L1 (Lasso) and L2 (Ridge)
  - Check Example 49.6 Elastic Net and External Cross Validation https://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_glmselect_examples06.htm
  - **Note**: Ridge performs **regularization**, but not **variable selection**.
- Fit Ridge (Elastic Net) regression model with model selection

```
proc glmselect data=Hitters1 plots=coefficients;
   model Salary = AtBat Hits    HmRun   Runs    RBI Walks
Years   CAtBat  CHits   CHmRun  CRuns   CRBI    CWalks  League_A
Division_1  PutOuts Assists Errors  NewLeague_A
         /selection=elasticnet(steps=120 choose=CV)
cvmethod=random;
run;
```

Example: Ridge Regression

- Paramter estimates and CV

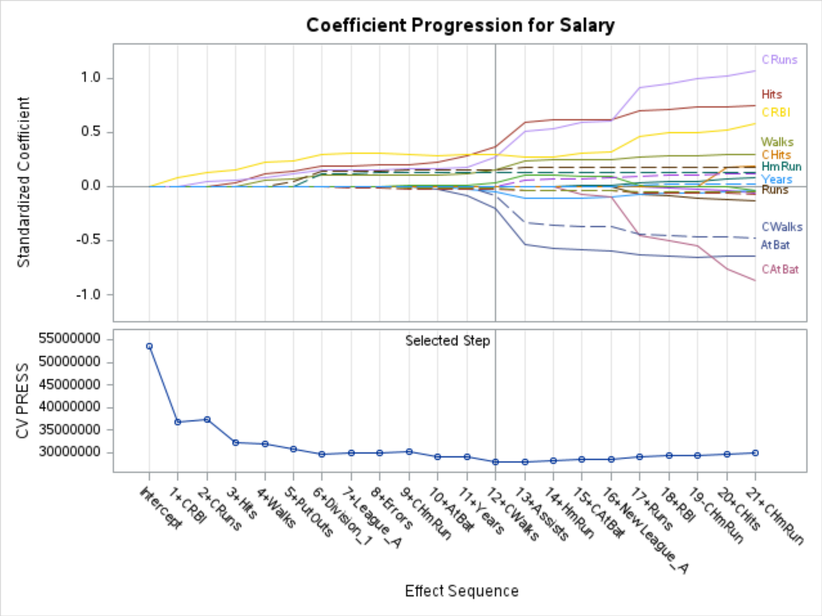


Example: LASSO Regression

- Now we fit a lasso model for the `Hitters` data

```
proc glmselect data=Hitters1 plots=coefficients;
   model Salary = AtBat Hits    HmRun   Runs    RBI Walks
Years   CAtBat  CHits   CHmRun  CRuns   CRBI    CWalks  League_A
Division_1  PutOuts Assists Errors  NewLeague_A
         /selection=LASSO(steps=120  choose=CV) cvmethod=random;
run;
```

Example: LASSO Regression



Example: LASSO Regression

- Lasso regression performs variable selection.
  - It can be seen that some variables are removed from the model.

License