Lasso Regression

Jeffrey Woo

MSDS, University of Virginia

Drawback of Ridge Regression

- All k predictors are left in the model. Model interpretation can be challenging when k is large.
- The lasso is an alternative to ridge regression that tackles this drawback.

The Lasso

The lasso coefficient estimates, denoted by $\hat{\beta}^L_\lambda$, are found by minimizing

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{k}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{k}|\beta_j| = SSres+\lambda\sum_{j=1}^{k}|\beta_j|, \quad (1)$$

Shrinkage Penalty

- The ℓ_2 penalty for ridge regression is now replaced by the ℓ_1 penalty $\|\beta\|_1 = \sum |\beta_j|$.
- The lasso shrinks the coefficient estimates towards 0.
- Unlike ridge regression, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be 0 when λ is large enough. The lasso performs variable selection.

Ridge vs Lasso

It can be shown that finding the coefficients for lasso and ridge regression is the same as solving

- minimize residual sum of squares subject to $\sum |\beta_j| \leq s$ and
- minimize residual sum of squares subject to $\sum \beta_j^2 \leq s$ respectively.

Consider a case when we have two predictors:

- The lasso estimates have the smallest RSS that lie within a diamond defined by $|\beta_1| + |\beta_2| \leq s$.
- The ridge estimates have the smallest RSS that lie within a circle defined by $\beta_1^2 + \beta_2^2 \leq s$.

Ridge vs Lasso

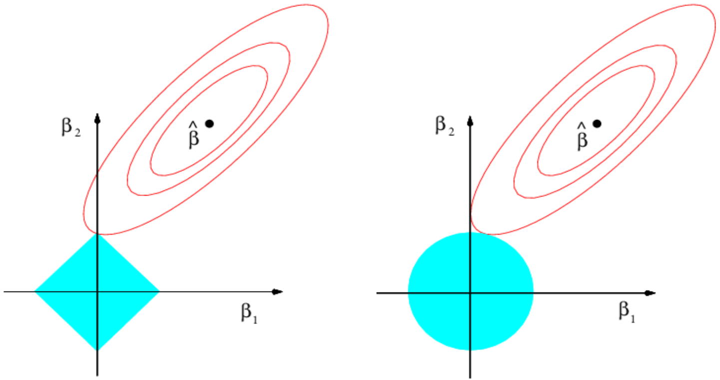

Figure: Comparison of coefficients from lasso vs ridge.

[1]Taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013)

Lasso Vs Ridge Regression

- Both lasso and ridge regression are useful in generating models with smaller variances than least squares by introducing some bias, thus improving model accuracy.
- The lasso produces **simpler and more interpretable** models than ridge regression, which always leaves all predictors in the model.

Lasso Vs Ridge Regression

1. Lasso will generally perform better when a small number of predictors are significant, with the other predictors being very small or equal to 0.
2. Ridge regression will generally perform better when the response variable is a function of many predictors..

The number of predictors that relate to the response variable is rarely known before-hand in real data sets. We could use cross-validation to determine which approach works better in a particular data set.

Worked Example: Gas Mileage

We will use the mtcars data set that is available in R. The data set contains information about fuel consumption and 10 aspects of automobile design and performance for 32 classic vehicles.

Worked Example: Gas Mileage

```
> head(mtcars, 8)
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
```

Worked Example: Gas Mileage

```
mpg:  Miles/(US) gallon
cyl:  Number of cylinders
disp: Displacement (cu.in.)
hp:   Gross horsepower
drat: Rear axle ratio
wt:   Weight (1000 lbs)
qsec: 1/4 mile time
vs:   Engine (0 = V-shaped, 1 = straight)
am:   Transmission (0 = automatic, 1 = manual)
gear: Number of forward gears
carb: Number of carburetors
```

Worked Example: Gas Mileage

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
xcyl        -0.11144    1.04502  -0.107   0.9161
xdisp        0.01334    0.01786   0.747   0.4635
xhp         -0.02148    0.02177  -0.987   0.3350
xdrat        0.78711    1.63537   0.481   0.6353
xwt         -3.71530    1.89441  -1.961   0.0633 .
xqsec        0.82104    0.73084   1.123   0.2739
xvs          0.31776    2.10451   0.151   0.8814
xam          2.52023    2.05665   1.225   0.2340
xgear        0.65541    1.49326   0.439   0.6652
xcarb       -0.19942    0.82875  -0.241   0.8122

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Worked Example: Gas Mileage

Use Lasso on the same splits for training and test data. Use cross-validation to find the λ that is optimal.

```
set.seed(12)
cv.out<-cv.glmnet(x.train,y.train,alpha=1)
plot(cv.out)
bestlam<-cv.out$lambda.min
bestlam
[1] 0.6398081
lasso.pred<-predict(lasso.mod,s=bestlam,newx=x.test)
mean((lasso.pred-y.test)^2)
[1] 7.998606
```

Worked Example: Gas Mileage

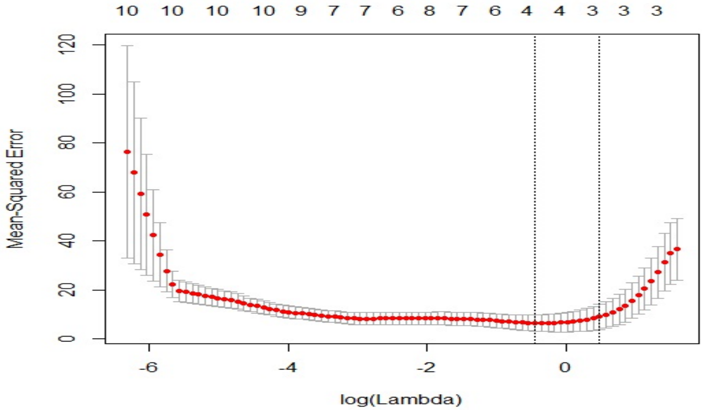

Figure: MSE with lasso against various values of tuning parameter (in logarithm).

Worked Example: Gas Mileage

Fit lasso, ridge regression using all observations using optimal λ, and compare coefficients with least squares.

```
cbind(coefficients(out.lasso), coefficients(out.ridge), coefficients(out.ols))
(Intercept) 36.42838465 21.164674099 12.30334580
cyl         -0.88528211 -0.371614420 -0.11143684
disp          .         -0.005238229  0.01333516
hp          -0.01330367 -0.011645632 -0.02148211
drat          .          1.052609540  0.78711436
wt          -2.77782157 -1.244587522 -3.71529660
qsec          .          0.162770783  0.82103979
vs            .          0.763638514  0.31776368
am           0.06978905  1.635361847  2.52022649
gear          .          0.545524257  0.65541591
carb          .         -0.552818552 -0.19942219
sqrt(sum(coefficients(out.lasso)[-1]^2))
[1] 2.916344
sqrt(sum(coefficients(out.ridge)[-1]^2))
[1] 2.585053
sqrt(sum(coefficients(out.ols)[-1]^2))
[1] 4.693825
```

Worked Example: Gas Mileage

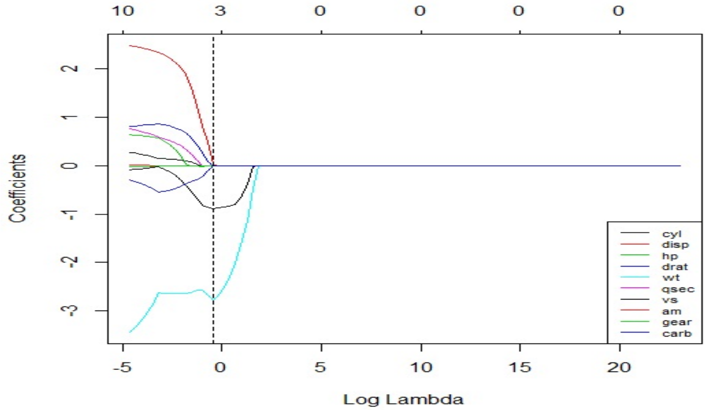

Figure: Lasso coefficients against various values of tuning parameter (in logarithm).

Worked Example: Gas Mileage

- Test MSE for ridge, lasso, least squares: 7.40, 8.00, 12.67.
- For lasso, 6 of the 10 coefficients are 0, leading to a simpler model than ridge which keeps all predictors in the model.