Ridge and Lasso(101C)

Seulchan Kim

2019 10 30

Data

```
data <- fivethirtyeight::hate_crimes
data2 <- na.omit(data)
x = model.matrix(avg_hatecrimes_per_100k_fbi~., data = data2)
y = data2$avg_hatecrimes_per_100k_fbi

library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-18
```

Ridge

```
set.seed (1)
grid=10^seq(10,-2,length=100)
ridge.mod=glmnet(x,y,alpha=0,lambda=grid)
dim(coef(ridge.mod))
```

```
## [1] 100 100
```

```
ridge.mod$lambda [50]
```

```
## [1] 11497.57
```

```
# When lamda = 11498
coef(ridge.mod)[,50]
```

```
##              (Intercept)               (Intercept)
##             2.369628e+00              0.000000e+00
##              stateAlaska              stateArizona
##            -1.083680e-04              1.571791e-04
##            stateArkansas           stateCalifornia
##            -2.273161e-04              3.587598e-06
##            stateColorado          stateConnecticut
##             6.509133e-05              2.113427e-04
##            stateDelaware stateDistrict of Columbia
##            -1.366018e-04              1.296322e-03
##             stateFlorida              stateGeorgia
##            -2.532533e-04             -2.964739e-04
##               stateIdaho             stateIllinois
##            -7.284709e-05             -2.009928e-04
##             stateIndiana                 stateIowa
##            -9.310507e-05             -2.738737e-04
##              stateKansas             stateKentucky
##            -3.467657e-05              2.772508e-04
##           stateLouisiana             stateMaryland
##            -1.560126e-04             -1.585933e-04
```

```
##         stateMassachusetts             stateMichigan
##              3.668585e-04              1.249248e-04
##             stateMinnesota             stateMissouri
##              1.871467e-04             -7.022271e-05
##              stateMontana             stateNebraska
##              8.787902e-05              4.727435e-05
##               stateNevada        stateNew Hampshire
##             -3.927237e-05             -4.044599e-05
##           stateNew Jersey           stateNew Mexico
##              3.081650e-04             -7.366936e-05
##             stateNew York       stateNorth Carolina
##              1.100091e-04             -1.679054e-04
##                 stateOhio             stateOklahoma
##              1.310069e-04             -1.952089e-04
##               stateOregon         statePennsylvania
##              1.542440e-04             -2.936169e-04
##         stateRhode Island       stateSouth Carolina
##             -1.649348e-04             -6.596990e-05
##            stateTennessee                stateTexas
##              1.152415e-04             -2.449918e-04
##                 stateUtah              stateVermont
##              1.585954e-06             -7.111904e-05
##             stateVirginia           stateWashington
##             -9.812036e-05              2.181599e-04
##        stateWest Virginia            stateWisconsin
##             -5.081058e-05             -1.891548e-04
##            state_abbrevAL            state_abbrevAR
##             -8.569621e-05             -2.273078e-04
##            state_abbrevAZ            state_abbrevCA
##              1.571732e-04              3.587557e-06
##            state_abbrevCO            state_abbrevCT
##              6.508902e-05              2.113350e-04
##            state_abbrevDC            state_abbrevDE
##              1.296275e-03             -1.365956e-04
##            state_abbrevFL            state_abbrevGA
##             -2.532438e-04             -2.964629e-04
##            state_abbrevIA            state_abbrevID
##             -2.738632e-04             -7.284455e-05
##            state_abbrevIL            state_abbrevIN
##             -2.009855e-04             -9.310178e-05
##            state_abbrevKS            state_abbrevKY
##             -3.467521e-05              2.772406e-04
##            state_abbrevLA            state_abbrevMA
##             -1.560068e-04              3.668452e-04
##            state_abbrevMD            state_abbrevMI
##             -1.585871e-04              1.249203e-04
##            state_abbrevMN            state_abbrevMO
##              1.871399e-04             -7.022004e-05
##            state_abbrevMT            state_abbrevNC
##              8.787588e-05             -1.678994e-04
##            state_abbrevNE            state_abbrevNH
##              4.727256e-05             -4.044452e-05
##            state_abbrevNJ            state_abbrevNM
##              3.081536e-04             -7.366666e-05
```

```
##             state_abbrevNV            state_abbrevNY
##              -3.927081e-05              1.100050e-04
##             state_abbrevOH            state_abbrevOK
##               1.310022e-04             -1.952017e-04
##             state_abbrevOR            state_abbrevPA
##               1.542384e-04             -2.936060e-04
##             state_abbrevRI            state_abbrevSC
##              -1.649287e-04             -6.596738e-05
##             state_abbrevTN            state_abbrevTX
##               1.152374e-04             -2.449827e-04
##             state_abbrevUT            state_abbrevVA
##               1.585984e-06             -9.811660e-05
##             state_abbrevVT            state_abbrevWA
##              -7.111642e-05              2.181519e-04
##             state_abbrevWI            state_abbrevWV
##              -1.891478e-04             -5.080878e-05
##           median_house_inc          share_unemp_seas
##               8.214471e-09              4.330809e-03
##            share_pop_metro              share_pop_hs
##               3.386042e-04              1.092815e-03
##          share_non_citizen       share_white_poverty
##               2.539813e-03             -2.496585e-03
##                 gini_index           share_non_white
##               5.866340e-03              2.294301e-04
##           share_vote_trump hate_crimes_per_100k_splc
##              -1.241740e-03              7.695299e-04
```

```r
sqrt(sum(coef(ridge.mod)[-1,50]^2))
```

```
## [1] 0.008675493
```

```r
ridge.mod$lambda [60]
```

```
## [1] 705.4802
```

```r
# when lambda = 705
coef(ridge.mod)[,60]
```

```
##                (Intercept)               (Intercept)
##               2.306360e+00              0.000000e+00
##                stateAlaska              stateArizona
##              -1.755878e-03              2.546724e-03
##              stateArkansas           stateCalifornia
##              -3.664979e-03              3.673349e-05
##              stateColorado          stateConnecticut
##               1.047622e-03              3.406758e-03
##              stateDelaware stateDistrict of Columbia
##              -2.218951e-03              2.093385e-02
##               stateFlorida              stateGeorgia
##              -4.110238e-03             -4.807917e-03
##                 stateIdaho             stateIllinois
##              -1.159433e-03             -3.269166e-03
##               stateIndiana                 stateIowa
##              -1.491614e-03             -4.433163e-03
##                stateKansas             stateKentucky
##              -5.441306e-04              4.521749e-03
```

```
##            stateLouisiana             stateMaryland
##             -2.515596e-03             -2.594463e-03
##        stateMassachusetts             stateMichigan
##              5.921966e-03              2.026223e-03
##            stateMinnesota             stateMissouri
##              3.023224e-03             -1.129707e-03
##              stateMontana             stateNebraska
##              1.442574e-03              7.854590e-04
##               stateNevada        stateNew Hampshire
##             -6.393535e-04             -6.467394e-04
##           stateNew Jersey           stateNew Mexico
##              4.985764e-03             -1.199918e-03
##             stateNew York       stateNorth Carolina
##              1.763306e-03             -2.718906e-03
##                 stateOhio             stateOklahoma
##              2.138037e-03             -3.144898e-03
##               stateOregon         statePennsylvania
##              2.479890e-03             -4.762134e-03
##         stateRhode Island       stateSouth Carolina
##             -2.682272e-03             -1.060007e-03
##            stateTennessee                stateTexas
##              1.884882e-03             -3.977577e-03
##                 stateUtah              stateVermont
##              3.766151e-05             -1.147070e-03
##             stateVirginia           stateWashington
##             -1.601885e-03              3.520748e-03
##        stateWest Virginia            stateWisconsin
##             -7.973323e-04             -3.057065e-03
##            state_abbrevAL            state_abbrevAR
##             -1.372778e-03             -3.664979e-03
##            state_abbrevAZ            state_abbrevCA
##              2.546724e-03              3.673340e-05
##            state_abbrevCO            state_abbrevCT
##              1.047622e-03              3.406758e-03
##            state_abbrevDC            state_abbrevDE
##              2.093385e-02             -2.218951e-03
##            state_abbrevFL            state_abbrevGA
##             -4.110238e-03             -4.807917e-03
##            state_abbrevIA            state_abbrevID
##             -4.433163e-03             -1.159433e-03
##            state_abbrevIL            state_abbrevIN
##             -3.269166e-03             -1.491614e-03
##            state_abbrevKS            state_abbrevKY
##             -5.441305e-04              4.521749e-03
##            state_abbrevLA            state_abbrevMA
##             -2.515596e-03              5.921966e-03
##            state_abbrevMD            state_abbrevMI
##             -2.594463e-03              2.026223e-03
##            state_abbrevMN            state_abbrevMO
##              3.023224e-03             -1.129707e-03
##            state_abbrevMT            state_abbrevNC
##              1.442574e-03             -2.718906e-03
##            state_abbrevNE            state_abbrevNH
##              7.854591e-04             -6.467393e-04
```

```
##             state_abbrevNJ            state_abbrevNM
##               4.985764e-03             -1.199918e-03
##             state_abbrevNV            state_abbrevNY
##              -6.393536e-04              1.763306e-03
##             state_abbrevOH            state_abbrevOK
##               2.138037e-03             -3.144898e-03
##             state_abbrevOR            state_abbrevPA
##               2.479890e-03             -4.762134e-03
##             state_abbrevRI            state_abbrevSC
##              -2.682272e-03             -1.060007e-03
##             state_abbrevTN            state_abbrevTX
##               1.884882e-03             -3.977577e-03
##             state_abbrevUT            state_abbrevVA
##               3.766152e-05             -1.601885e-03
##             state_abbrevVT            state_abbrevWA
##              -1.147070e-03              3.520748e-03
##             state_abbrevWI            state_abbrevWV
##              -3.057065e-03             -7.973322e-04
##           median_house_inc           share_unemp_seas
##               1.319772e-07              6.970554e-02
##            share_pop_metro               share_pop_hs
##               5.413916e-03              1.766453e-02
##          share_non_citizen       share_white_poverty
##               4.075425e-02             -4.000019e-02
##                 gini_index            share_non_white
##               9.458950e-02              3.647184e-03
##           share_vote_trump hate_crimes_per_100k_splc
##              -1.998286e-02              1.242298e-02
```

```r
sqrt(sum(coef(ridge.mod)[-1,60]^2))
```

```
## [1] 0.1397194
```

```r
predict(ridge.mod,s=50,type="coefficients")[1:20,]
```

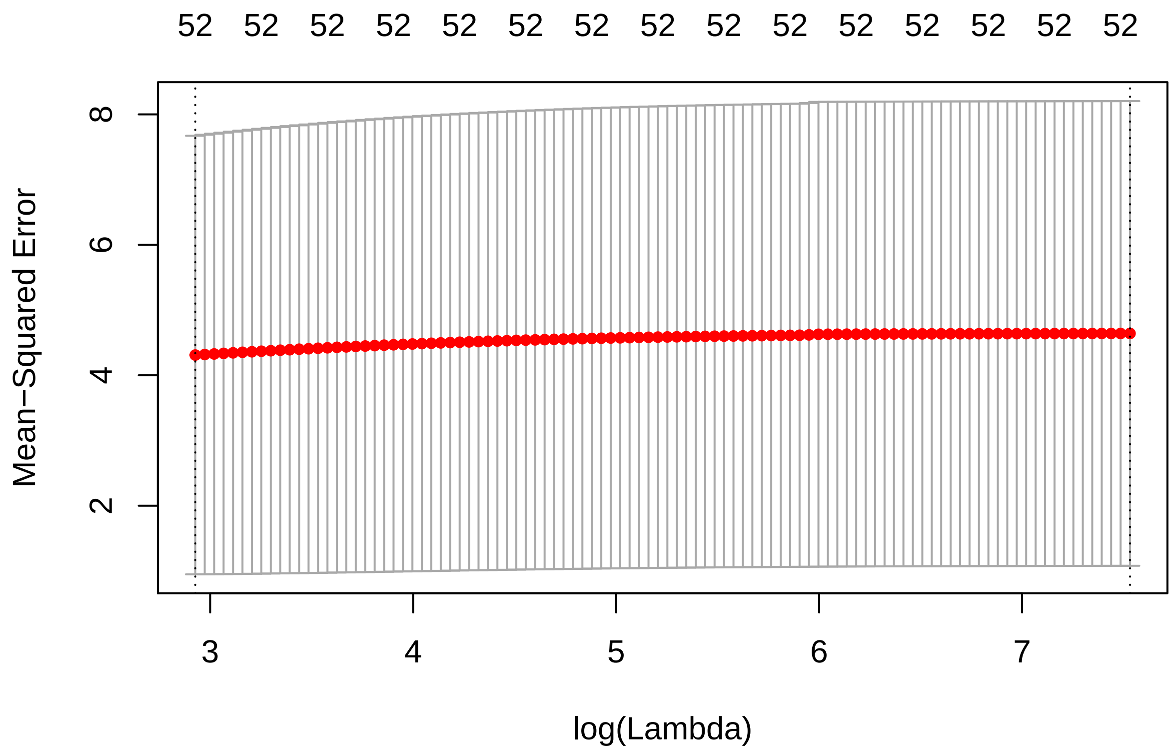
```
##                (Intercept)               (Intercept)
##                1.532861427               0.000000000
##                stateAlaska              stateArizona
##               -0.023333329               0.033963005
##              stateArkansas           stateCalifornia
##               -0.045812501              -0.002783144
##              stateColorado          stateConnecticut
##                0.012796094               0.042488126
##              stateDelaware stateDistrict of Columbia
##               -0.030350812               0.266878224
##               stateFlorida              stateGeorgia
##               -0.055669875              -0.064503440
##                 stateIdaho             stateIllinois
##               -0.012075860              -0.045399714
##               stateIndiana                 stateIowa
##               -0.017148926              -0.058337421
##                stateKansas             stateKentucky
##               -0.004410140               0.064686168
##             stateLouisiana             stateMaryland
##               -0.031467522              -0.038349695
```

```r
# split the samples into a training and a test in order to estime the test error of ridge and lasso
train=sample(1:nrow(x), nrow(x)/2)
test=(-train)
y.test=y[test]

cv.out=cv.glmnet(x[train,],y[train],alpha=0)
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations
## per fold
```

```r
plot(cv.out)
```




```r
bestlam=cv.out$lambda.min
bestlam # lambda that results in the smallest cv error is 1745
```

```
## [1] 18.66568
```

```r
# what is the MSE with this lambda?
ridge.pred=predict(ridge.mod,s=bestlam ,newx=x[test,])
mean((ridge.pred-y.test)^2)
```

```
## [1] 1.029258
```

```r
# refit our ridge regression model on the full data set, using the lambda chosen by cv, and examine the
out=glmnet(x,y,alpha=0)
predict(out,type="coefficients",s=bestlam)[1:20,]
```

```
##         (Intercept)          (Intercept)
##         0.564425852          0.000000000
##         stateAlaska         stateArizona
##        -0.054406549          0.080265456
##       stateArkansas      stateCalifornia
##        -0.099061034         -0.014563616
```

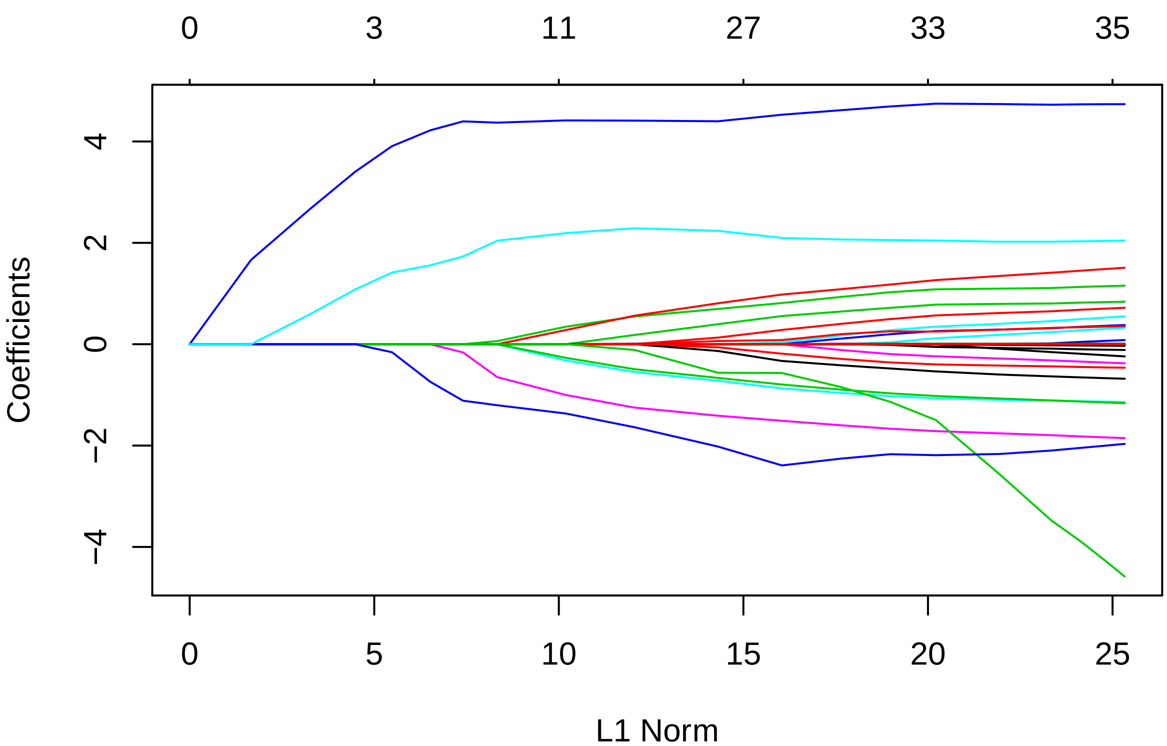
```
##            stateColorado         stateConnecticut
##              0.026782394              0.091660630
##            stateDelaware stateDistrict of Columbia
##             -0.072930474              0.591021612
##             stateFlorida             stateGeorgia
##             -0.132507839             -0.151787213
##               stateIdaho            stateIllinois
##             -0.019105970             -0.110977733
##             stateIndiana                stateIowa
##             -0.032914945             -0.135211179
##              stateKansas            stateKentucky
##             -0.002643702              0.162691724
##           stateLouisiana            stateMaryland
##             -0.068033787             -0.099454259
```

None of the coefficients are zero; ridge regression does not perform variable selection.

Lasso

```r
lasso.mod=glmnet(x[train ,],y[train],alpha=1,lambda=grid)
plot(lasso.mod)
```

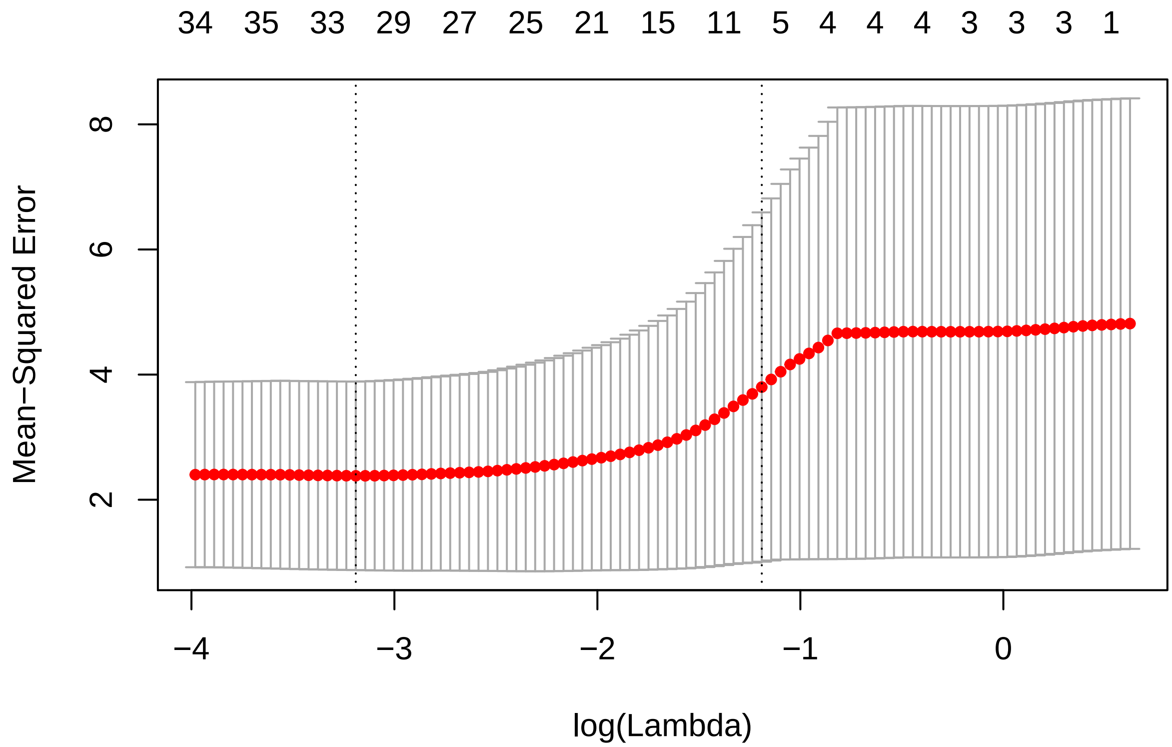
```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to
## unique 'x' values
```




```r
set.seed (1)
cv.out=cv.glmnet(x[train ,],y[train],alpha=1)
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations
## per fold
```

```r
plot(cv.out)
```




```r
bestlam=cv.out$lambda.min
lasso.pred=predict(lasso.mod,s=bestlam ,newx=x[test,])
mean((lasso.pred-y.test)^2)
```

```
## [1] 1.366457
```

```r
out=glmnet(x,y,alpha=1,lambda=grid)
lasso.coef=predict(out,type="coefficients",s=bestlam)[1:20,]
lasso.coef
```

```
##              (Intercept)              (Intercept)
##               0.59259566               0.00000000
##              stateAlaska             stateArizona
##               0.00000000               1.05190096
##            stateArkansas          stateCalifornia
##              -0.32196182               0.00000000
##            stateColorado         stateConnecticut
##               0.00000000               1.06221941
##            stateDelaware stateDistrict of Columbia
##              -0.59401179               4.49283542
##             stateFlorida             stateGeorgia
##              -0.95824359              -1.01887614
##               stateIdaho            stateIllinois
##               0.00000000              -0.64831686
##             stateIndiana                stateIowa
##              -0.03195501              -1.91500154
##              stateKansas            stateKentucky
##               0.20397193               1.67484458
##           stateLouisiana            stateMaryland
##               0.00000000              -0.98113228
```

Lasso Regression has a advantage over Ridge Regression in that the resulting coefficient estimates are sparse.

We can see 4 coefficient estimates are exactly zero. So the lasso model with lambda chosen by cv contains only 4 variables.