Least Absolute Shrinkage and Selection Operator (LASSO)

Bryant Moy

November 28, 2017

Outline

Motivation and Intuition

- Variable Selection

Motivation and Intuition

- Variable Selection
- Too Many Covariates in the Model

Motivation and Intuition

- Variable Selection
- Too Many Covariates in the Model
- Shrinkage Method to Push β's towards zero

Motivation and Intuition

- Variable Selection
- Too Many Covariates in the Model
- Shrinkage Method to Push β's towards zero
- But Why?

LASSO Estimation

- The Lasso coefficients $\hat{\beta}^L_\lambda$ minimizes the following equation:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{i=1}^{P}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{P}|\beta_j|$$

LASSO Estimation

- The Lasso coefficients $\hat{\beta}_\lambda^L$ minimizes the following equation:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{i=1}^{P} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{P} |\beta_j|$$

$$RSS + \lambda \sum_{j=1}^{P} |\beta_j|$$

LASSO Estimation

- The Lasso coefficients $\hat{\beta}_{\lambda}^{L}$ minimizes the following equation:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{i=1}^{P}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{P}|\beta_j|$$

$$RSS + \lambda \sum_{j=1}^{P}|\beta_j|$$

- L 1 Norm Penalty

LASSO Estimation

- The Lasso coefficients $\hat{\beta}_\lambda^L$ minimizes the following equation:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{i=1}^{P}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{P}|\beta_j|$$

$$RSS + \lambda\sum_{j=1}^{P}|\beta_j|$$

- L 1 Norm Penalty
- Ridge Regression is $L2$ Norm $\rightarrow \beta^2$

LASSO vs. Ridge

LASSO

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{i=1}^{P} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{P} |\beta_j|$$

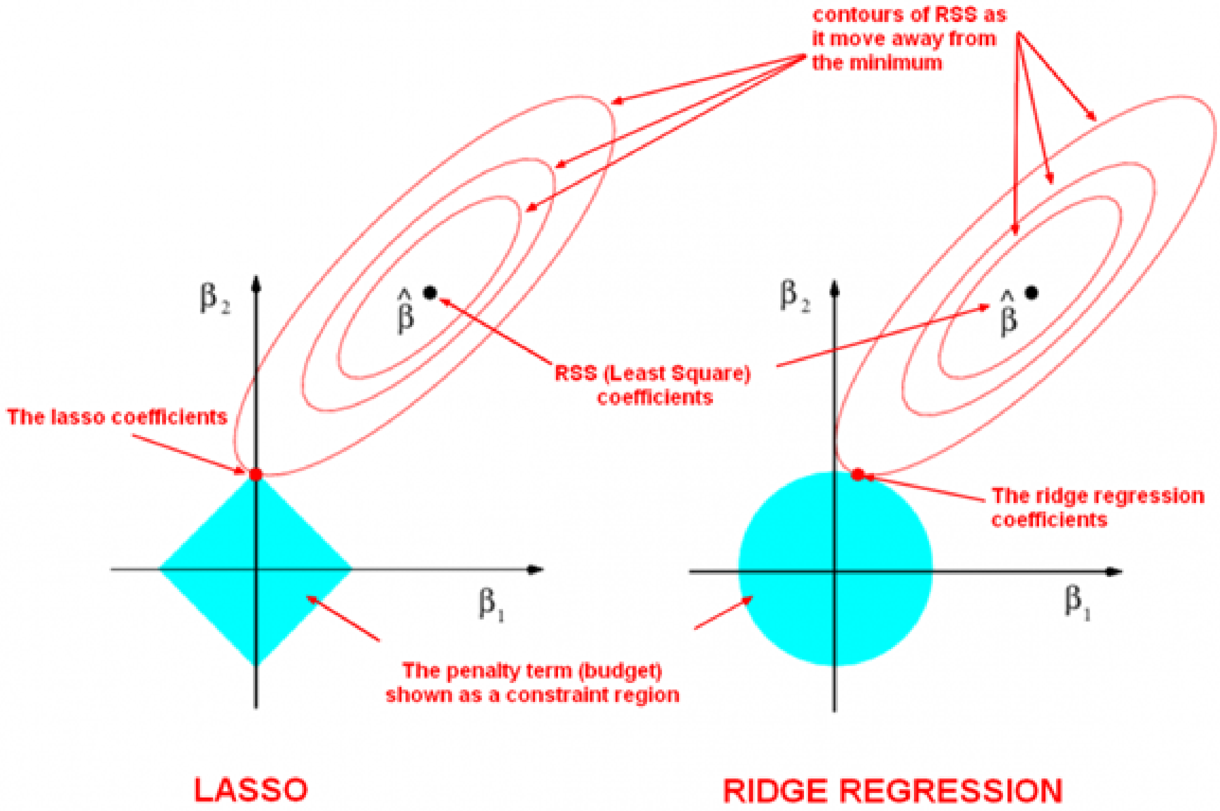
$$RSS + \lambda \sum_{j=1}^{P} |\beta_j|$$

Ridge Regression

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

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

LASSO vs. Ridge

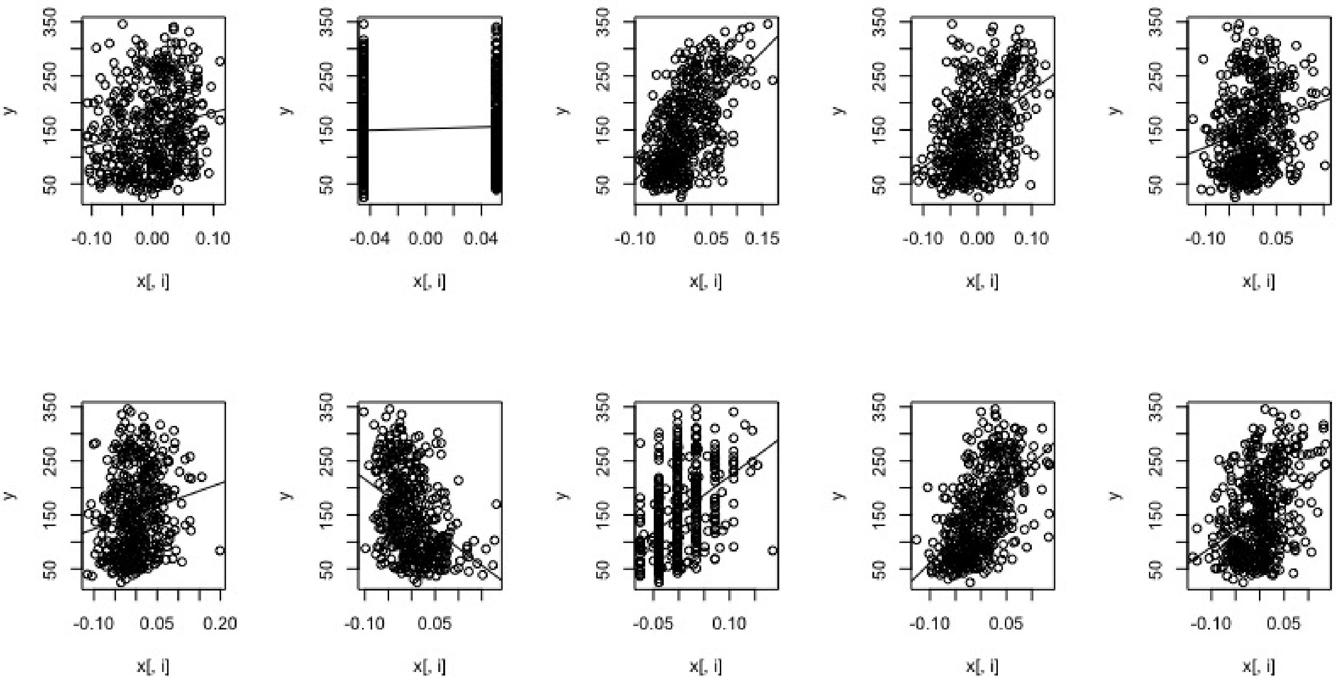


Code and Example

```
library(lars)
library(glmnet)
data(diabetes)
```

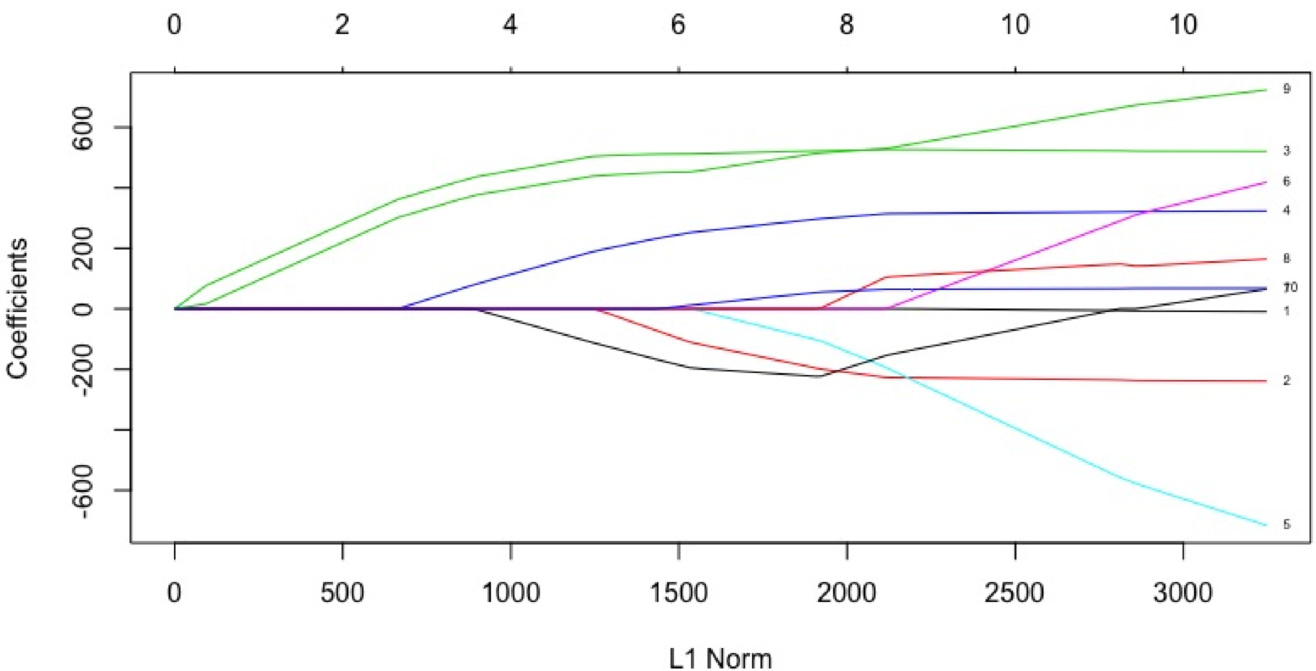
- age
- sex
- bmi
- map
- tc
- ldl
- hdl
- tch
- ltg
- glu

```r
for(i in 1:10){
  plot(x[,i], y)
  abline(lm(y~x[,i]))
  }
```



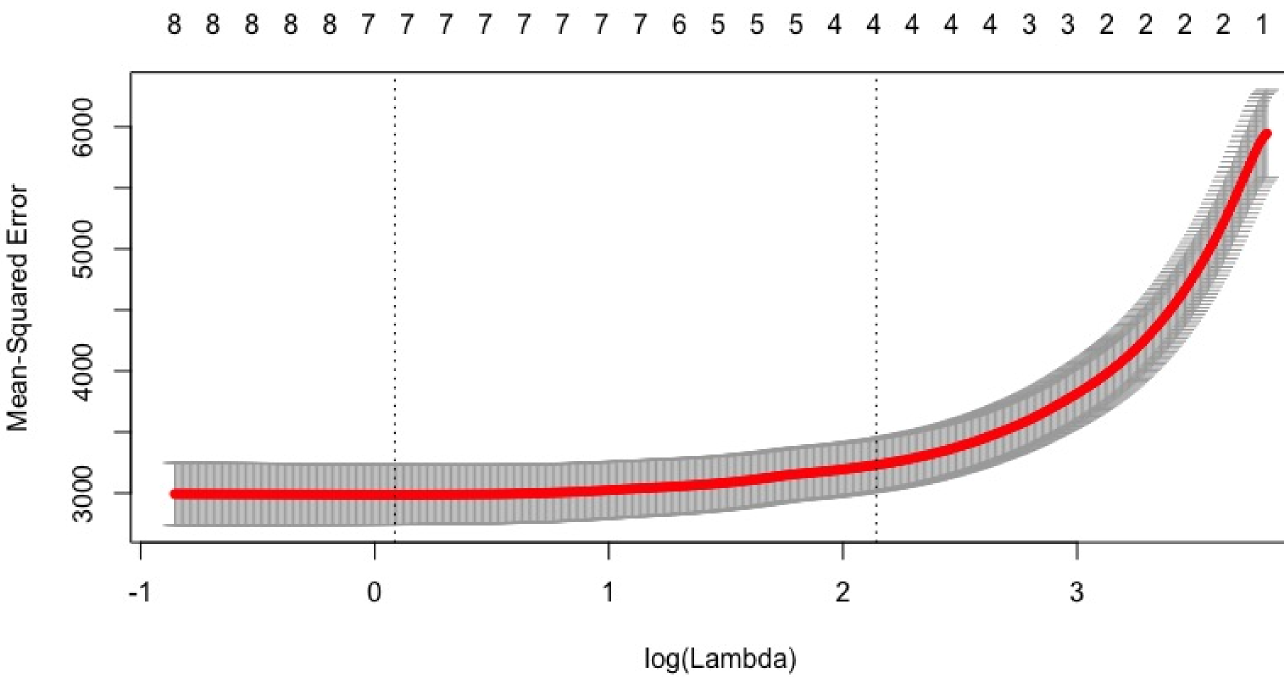
Code and Example

```
model_lasso <- glmnet(x, y)
plot.glmnet(model_lasso, xvar = "norm", label = TRUE)
```



Code and Example

```
cv_fit <- cv.glmnet(x=x, y=y, alpha = 1, nlambda = 1000)
plot.cv.glmnet(cv_fit)
cv_fit$lambda.min # 1.069371
```



Code and Example (OLS)

```
summary(model_ols)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  152.133      2.576  59.061  < 2e-16 ***
xage         -10.012     59.749  -0.168 0.867000
xsex        -239.819     61.222  -3.917 0.000104 ***
xbmi         519.840     66.534   7.813 4.30e-14 ***
xmap         324.390     65.422   4.958 1.02e-06 ***
xtc         -792.184    416.684  -1.901 0.057947 .
xldl         476.746    339.035   1.406 0.160389
xhdl         101.045    212.533   0.475 0.634721
xtch         177.064    161.476   1.097 0.273456
xltg         751.279    171.902   4.370 1.56e-05 ***
xglu          67.625     65.984   1.025 0.305998
---
Residual standard error: 54.15 on 431 degrees of freedom
Multiple R-squared:  0.5177,Adjusted R-squared:  0.5066
F-statistic: 46.27 on 10 and 431 DF,  p-value: < 2.2e-16
```

Code and Example (LASSO)

```
fit <- glmnet(x=x, y=y, alpha = 1, lambda=cv_fit$lambda.min)
fit$beta
10 x 1 sparse Matrix of class "dgCMatrix"
            s0
age          .
sex -193.51972
bmi  521.56400
map  294.92639
tc   -98.53276
ldl          .
hdl -222.65684
tch          .
ltg  511.53755
glu   52.60584
```

Why the LASSO is Uniquely Important: Causal Estimates

Lasso adjustments of treatment effect estimates in randomized experiments

Adam Bloniarz[a,1], Hanzhong Liu[a,1], Cun-Hui Zhang[b], Jasjeet S. Sekhon[a,c], and Bin Yu[a,d,2]

[a]Department of Statistics, University of California, Berkeley, CA 94720; [b]Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ 08854; [c]Department of Political Science, University of California, Berkeley, CA 94720; and [d]Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved December 1, 2015 (received for review June 3, 2015)

$$\hat{\sigma}^2_{e^{(a)}} = \frac{1}{n_A - df^{(a)}} \sum_{i \in A} \left(a_i - \bar{a}_A - (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T \hat{\boldsymbol{\beta}}^{(a)}_{\text{Lasso}} \right)^2,$$

$$\hat{\sigma}^2_{e^{(b)}} = \frac{1}{n_B - df^{(b)}} \sum_{i \in B} \left(b_i - \bar{b}_B - (\mathbf{x}_i - \bar{\mathbf{x}}_B)^T \hat{\boldsymbol{\beta}}^{(b)}_{\text{Lasso}} \right)^2,$$

where $df^{(a)}$ and $df^{(b)}$ are degrees of freedom defined by the following:

$$df^{(a)} = \hat{s}^{(a)} + 1 = \left\| \hat{\boldsymbol{\beta}}^{(a)}_{\text{Lasso}} \right\|_0 + 1; \; df^{(b)} = \hat{s}^{(b)} + 1 = \left\| \hat{\boldsymbol{\beta}}^{(b)}_{\text{Lasso}} \right\|_0 + 1.$$

Define the variance estimate of $\sqrt{n}(\widehat{ATE}_{\text{Lasso}} - ATE)$ as follows:

$$\hat{\sigma}^2_{\text{Lasso}} = \frac{n}{n_A} \hat{\sigma}^2_{e^{(a)}} + \frac{n}{n_B} \hat{\sigma}^2_{e^{(b)}}. \qquad \textbf{[20]}$$

Thanks!