Lab4

Yilin Li

2019/10/5

Lab 4

```r
library(glmnet)
```

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

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

```
## Loaded glmnet 2.0-16
```

```r
# read in the data set
d <- read.csv("http://andrewpbray.github.io/data/crime-train.csv")

# make sure all data are numerical
d = d[,sapply(d, is.numeric)]

X = model.matrix(ViolentCrimesPerPop~., data = d)[,-1]
Y = d$ViolentCrimesPerPop
lambdas = 10^seq(10,-2, length =100)

# create the ridge model for dataset d and find training MSEs for each lambda
ridge_model = glmnet(x = X, y = Y, alpha = 0, lambda = lambdas, standardize = TRUE)
ridge_MSE = rep(NA,length = length(lambdas))
for (i in 1:length(lambdas)){
  p = predict(ridge_model, s = lambdas[i], newx = X)
  ridge_MSE[i] =  mean((p-Y)^2)
}
lowest_MSE_ridge = min(ridge_MSE)

# create the LASSO model for dataset d and find training MSEs for each lambda
lasso_model = glmnet(x = X, y = Y, alpha = 1, lambda = lambdas)
lasso_MSE = rep(NA,length = length(lambdas))
for (i in 1:length(lambdas)){
  p = predict(lasso_model, s = lambdas[i], newx = X)
  lasso_MSE[i] =  mean((p-Y)^2)
}
lowest_MSE_lasso = min(lasso_MSE)
num_selected_lasso = sum(coef(lasso_model)[,100] > 0) - 1
```

Questions

1. There are 10 variables selected by Lasso Model with the optimal lambda value.
2. The training MSE for ridge is 0.01625487 and the training MSE for lasso is 0.01828888, which are different.

3. With the fact that in the same sequence of lambdas, the optimal lambda value for ridge and lasso regression is the same value, the variance of the ridge regression is lower than the variance of the lasso regression and resulting in a lower training MSE than the lasso regression.

Problem set 4

Exercise 2

a. iii is correct because the added lasso term lowers the variance of the LS model by scaling the coefficients towards zero. Therefore it turns out that it's a less flexible model, which increses the bias and decreases the variance.
b. iii is also correct in the exactly same reason as the lasso regression.

Exercise 3

a. Steadily decrease because as $s = 0$, the $RSS = \sum_{i=1}^{n}(y_i - \beta_0)^2$ which is maxized. As s increses, the RSS will steadily decrease to the RSS of ordinary least square RSS.
b. Decrease initially, and then eventually start increasing in a U shape. Decreasing part has the same reason as part (a), but eventually as s keep increasing, the flexibility of the model will surpass the OLS model and thus the model tends to overfit the training data, which results in increases of the test RSS.
c. Steadily increase. As $s = 0$, the model has lowest variance because it's predicting a constant value. Then increasing s creates more and more possibilities of β_i which steadily increases the complexity of the model.
d. Steadily decrease because the variance of the model is steadily increasing.
e. Constant by the definition of the irreducible error which is indepedent of the structure of the model.

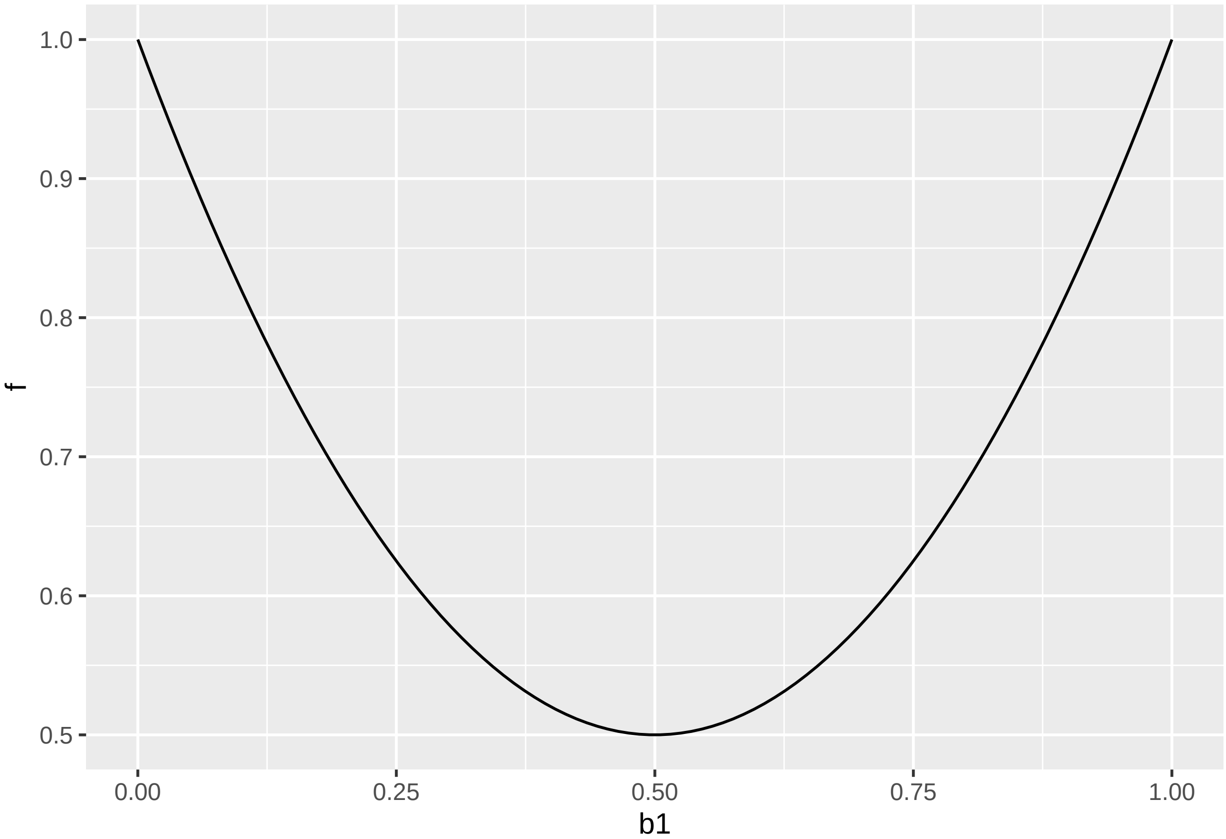
Exercise 4

a. Steadily increase because as $\lambda = 0$, this is the OLS model and has the minimum of the training RSS. As λ increases, β_i tends to be scaled towards zero and thus the training RSS increases steadily.
b. Decrease initially, and then eventually start increasing in a U shape: as $\lambda = 0$, the OLS model tends to fit the training data as good as possible without the regularized term, so as *lambda* increases initially, the test RSS will decrease. But eventually the model will be too simple with such large λ value and thus test RSS increases.
c. Steadily decreases because as λ increases, the fitting coefficients are regularized more and keep decreasing the flexibility of the model. So the variance decreases steadily.
d. Steadily increases as the variance decreases steadily.
e. Constant by the definition of the irreducible error which is indepedent of the structure of the model.

Exercise 6

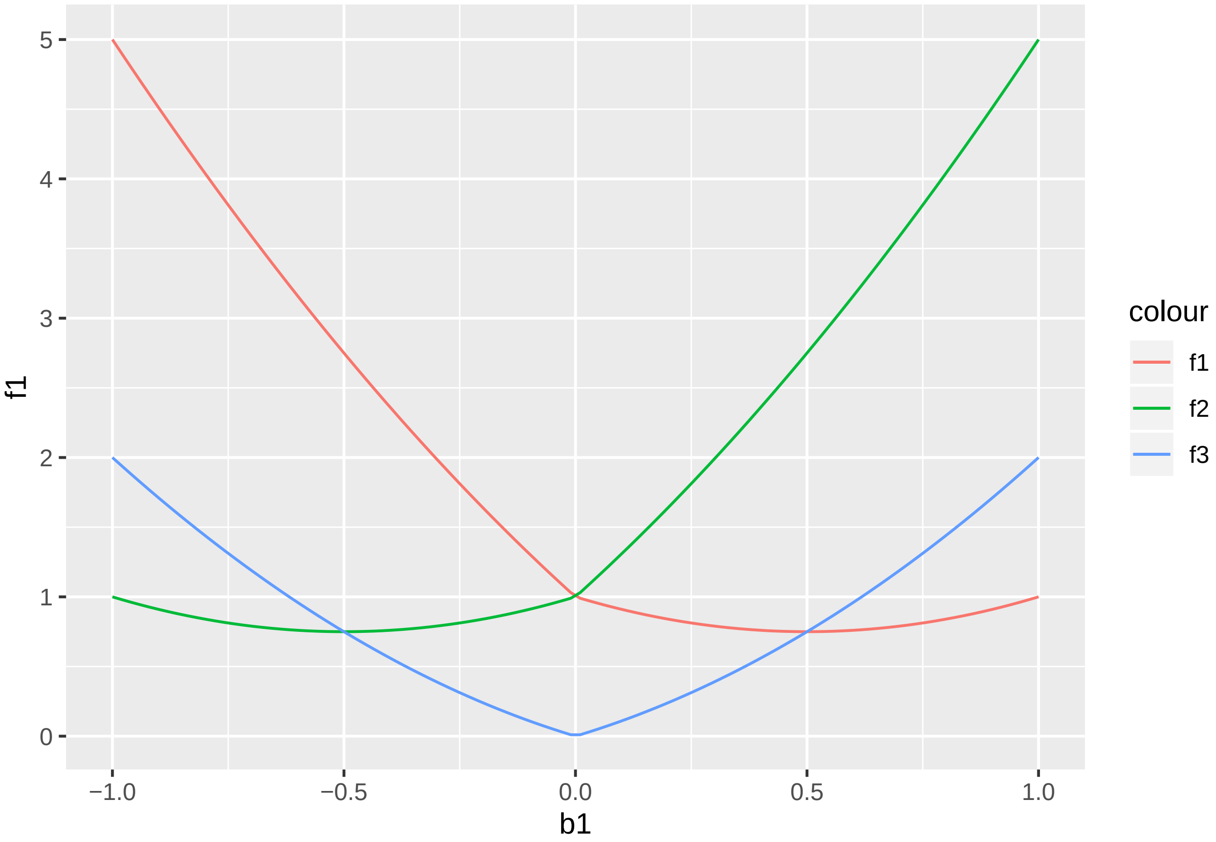
a. Assume $p = 1$, then (6.12) becomes $(y_1 - \beta_1)^2 + \lambda\beta_1^2$. Take $y_1 = 1$ and $\lambda = 1$ and plot (6.12) as a function of β_1 as follows:

```r
library(ggplot2)
b1 = seq(from = 0, to = 1, length = 100)
f = (1-b1)^2 + b1^2
ggplot() + geom_line(aes(x = b1, y = f))
```



It turns out that the optimal $\hat{\beta}_1^R$ is 0.5 which indeed conforms with (6.14) where $\hat{\beta}_1^R = \frac{1}{1+1} = 0.5$ b. Assume $p = 1$ then (6.13) becomes $(y_1 - \beta_1)^2 + \lambda|\beta_1|$. Take $\lambda = 1$ and $y_1 = 1, -1, 0$ seperately and plot (6.13) as a functions of β_1 as follows:

```
b1 = seq(from = -1, to = 1, length = 100)
f1 = c((1-b1[1:50])^2-b1[1:50],(1-b1[51:100])^2+b1[51:100])
f2 = c((-1-b1[1:50])^2-b1[1:50],(-1-b1[51:100])^2+b1[51:100])
f3 = c((-b1[1:50])^2-b1[1:50],(-b1[51:100])^2+b1[51:100])
ggplot() + geom_line(aes(x = b1, y = f1, color = 'f1')) + geom_line(aes(x = b1, y = f2,color = 'f2')) +
```



It turns out that the optimal β_1 is indeed solved by the equation (6.15).

Exercise 5

a. The goal is to minimize:

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{\hat{p}} \hat{\beta_j x_j})^2 + \lambda \sum_{i=1}^{p} \hat{\beta}_i^2$$

$$= (y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(\hat{\beta}_1{}^2 + \hat{\beta}_2{}^2)$$

b. Let $x_{11} = x_{12} = x_1$ and $x_{21} = x_{22} = x_2$. Then, we take the derivatives of RSS with $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively and set them to zerp. We get:

$$\hat{\beta}_1 = \frac{x_1 y_1 + x_2 y_2 - \hat{\beta}_2(x_1^2 + x_2^2)}{x_1^2 + x_2^2 + \lambda}$$

$$\hat{\beta}_2 = \frac{x_1 y_1 + x_2 y_2 - \hat{\beta}_1(x_1^2 + x_2^2)}{x_1^2 + x_2^2 + \lambda}$$

So the equations above show taht $\hat{\beta}_1 = \hat{\beta}_2$

c. The goal is to minimize:

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{\hat{p}} \hat{\beta_j x_j})^2 + \lambda \sum_{i=1}^{p} |\hat{\beta}_i|$$

$$= (y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|)$$

d. As $\hat{\beta_1} \geq 0$ and $\hat{\beta_2} \geq 0$, we have

$$\hat{\beta_1} + \hat{\beta_2} = \frac{2x_1y_1 + 2x_2y_2 - \lambda}{2x_1^2 + 2x_2^2}$$

As $\hat{\beta_1} \leq 0$ and $\hat{\beta_2} \leq 0$, we have

$$\hat{\beta_1} + \hat{\beta_2} = \frac{2x_1y_1 + 2x_2y_2 + \lambda}{2x_1^2 + 2x_2^2}$$

In both cases, for certain λ, $\hat{\beta_1} + \hat{\beta_2}$ is minimized as they move along a 45 degree line. Any combination is possible and thus the solution is not unique.