Regularization2

Ye

7/10/2017

Genearte the DataSet

```
set.seed(8394756)
Epsilon<-rnorm(500,0,1)
X<-rnorm(500*500,0,2)
dim(X)<-c(500,500)
colnames(X)<-paste0("X",1:500)
slopesSet<-runif(500,-.1,3)
Y<-sapply(2:500,function(z) 1+X[,1:z]%*%slopesSet[1:z]+Epsilon)
```

Separate the Dataset to the training set and test set (1:1) and apply Lasso regression to training set

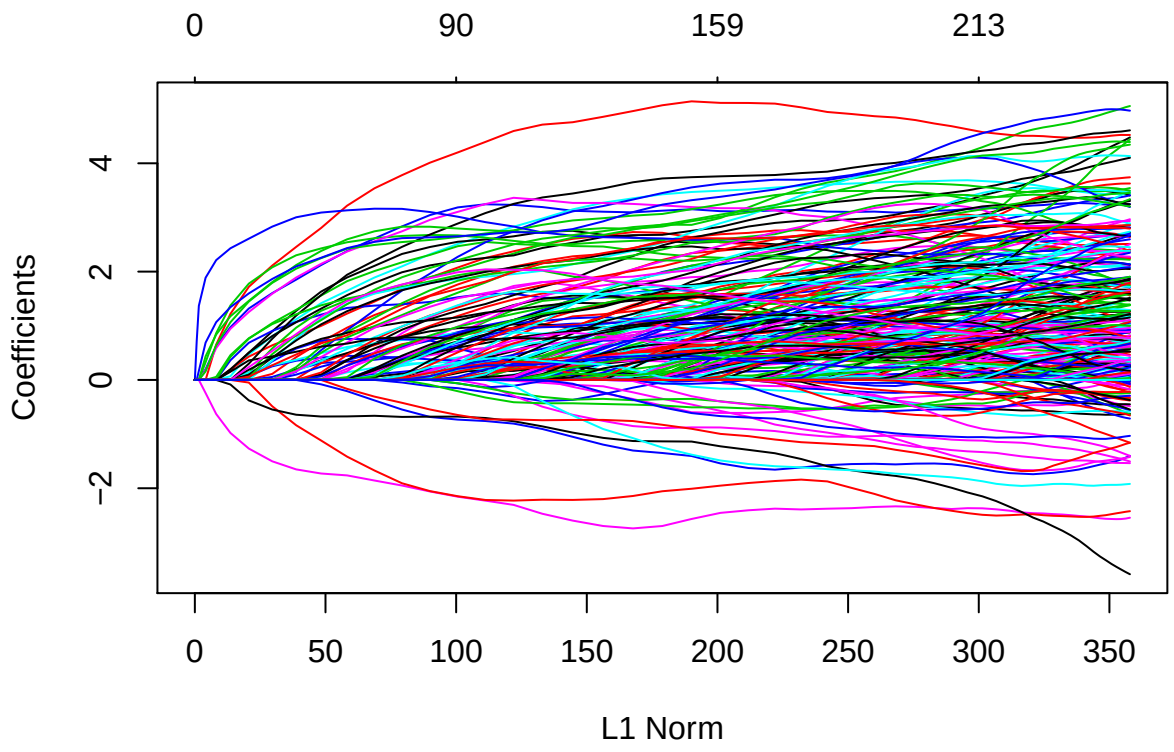
```
set.seed(1)
train = sample(1:nrow(X[,1:491]), nrow(X[,1:491])/2)
test = (-train)
y.test=Y[test,490]
suppressWarnings(library('glmnet'))
```

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

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

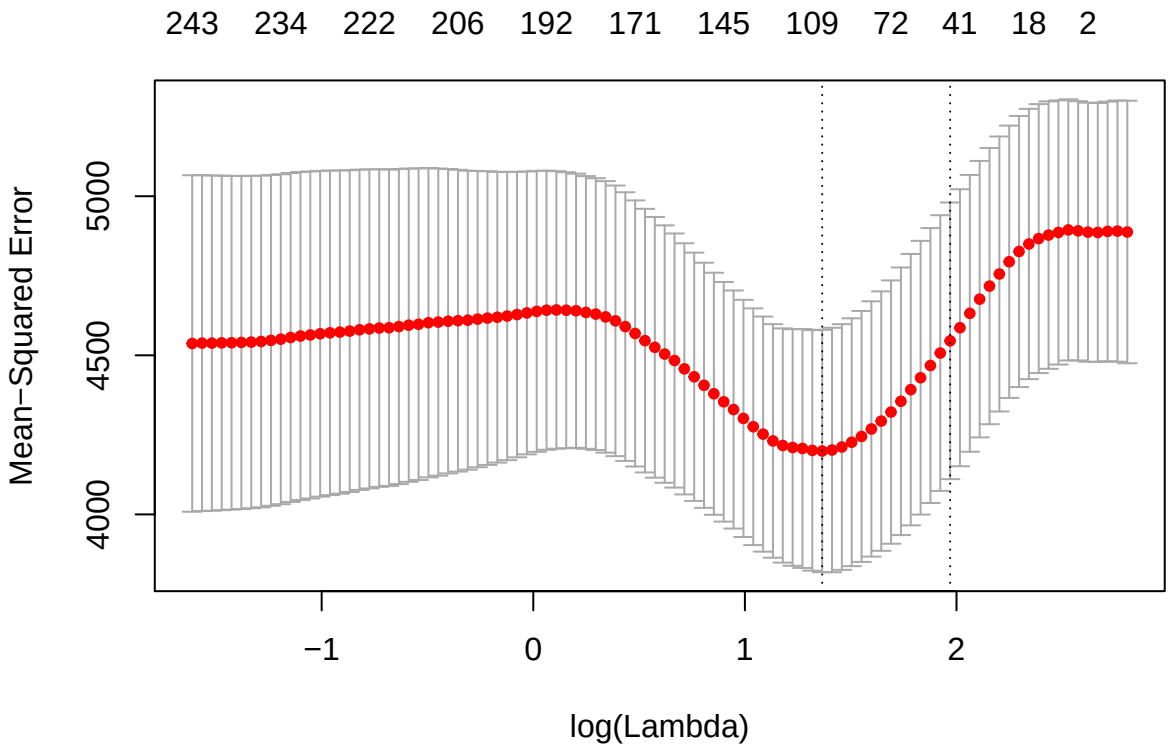
```
## Loaded glmnet 2.0-10
```

```
lasso490=glmnet(x=X[train,1:491],y=Y[train,490],alpha=1,nlambda=100,lambda.min.ratio=.0001)
plot(lasso490)
```



Perform cross-validation on the training set and choose the best lambda

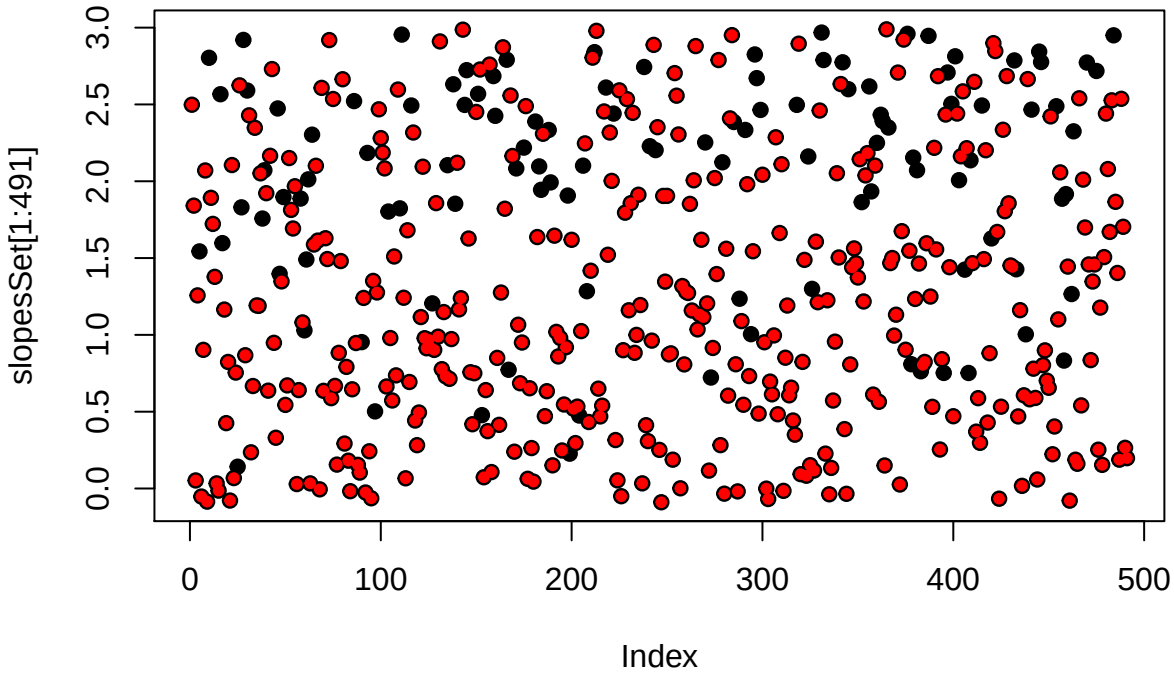
```
cv.out = cv.glmnet(x = X[train, 1:491], Y[train, 490], alpha = 1)
plot(cv.out)
```




```
bestlam = cv.out$lambda.min
print(bestlam)
```

```
## [1] 3.916326
```

```
out=glmnet(x=X[,1:491],y=Y[,490],alpha=1,lambda=bestlam)
lasso.coef=predict(out,type="coefficients",s=bestlam)
removedSlopes<-rep(NA,491)
removedSlopes[lasso.coef[-1]==0]<-slopesSet[1:491][lasso.coef[-1]==0]
plot(slopesSet[1:491],pch=19)
points(removedSlopes,col="red",pch=20)
```

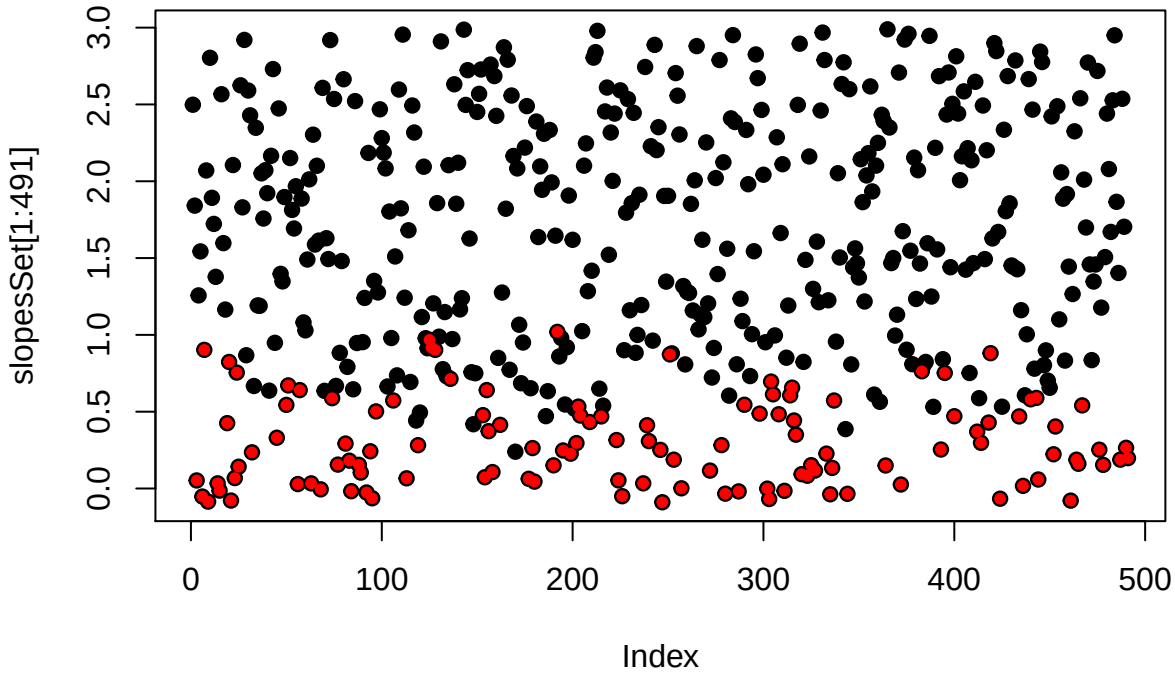


Lasso seems to remove predictors in a random way.

Use linear regreesion and remove predictors with p-values larger than 0.05

```
m490 = lm(Y~., data=data.frame(Y=Y[, 490], X[, 1:491]))
lmRemovedSlopes = rep(NA, 491)
lmRemovedSlopes[coefficients(summary(m490))[-1, 4] > 0.05] =

slopesSet[1:491][coefficients(summary(m490))[-1, 4] > 0.05]
plot(slopesSet[1:491],pch=19)
points(lmRemovedSlopes,col="red",pch=20)
```


slopesSet[1:491]
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
100
200
300
400
500
Index