542_assign2_2

Tian Ni

9/23/2021

```
library(glmnet)
```

```
## 载入需要的程辑包：Matrix
```

```
## Loaded glmnet 4.1-2
```

```
library(pls)
```

```
##
## 载入程辑包：'pls'
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

```
set.seed(6659)
```

Load the data

```
myData=read.csv("BostonData2.csv")
myData=myData[,-1]
dim(myData)
```

```
## [1] 506  92
```

```
X=data.matrix(myData[,-1])
Y=data.matrix(myData[,1])
```

Then we construct those seven procedure to make it easier to read

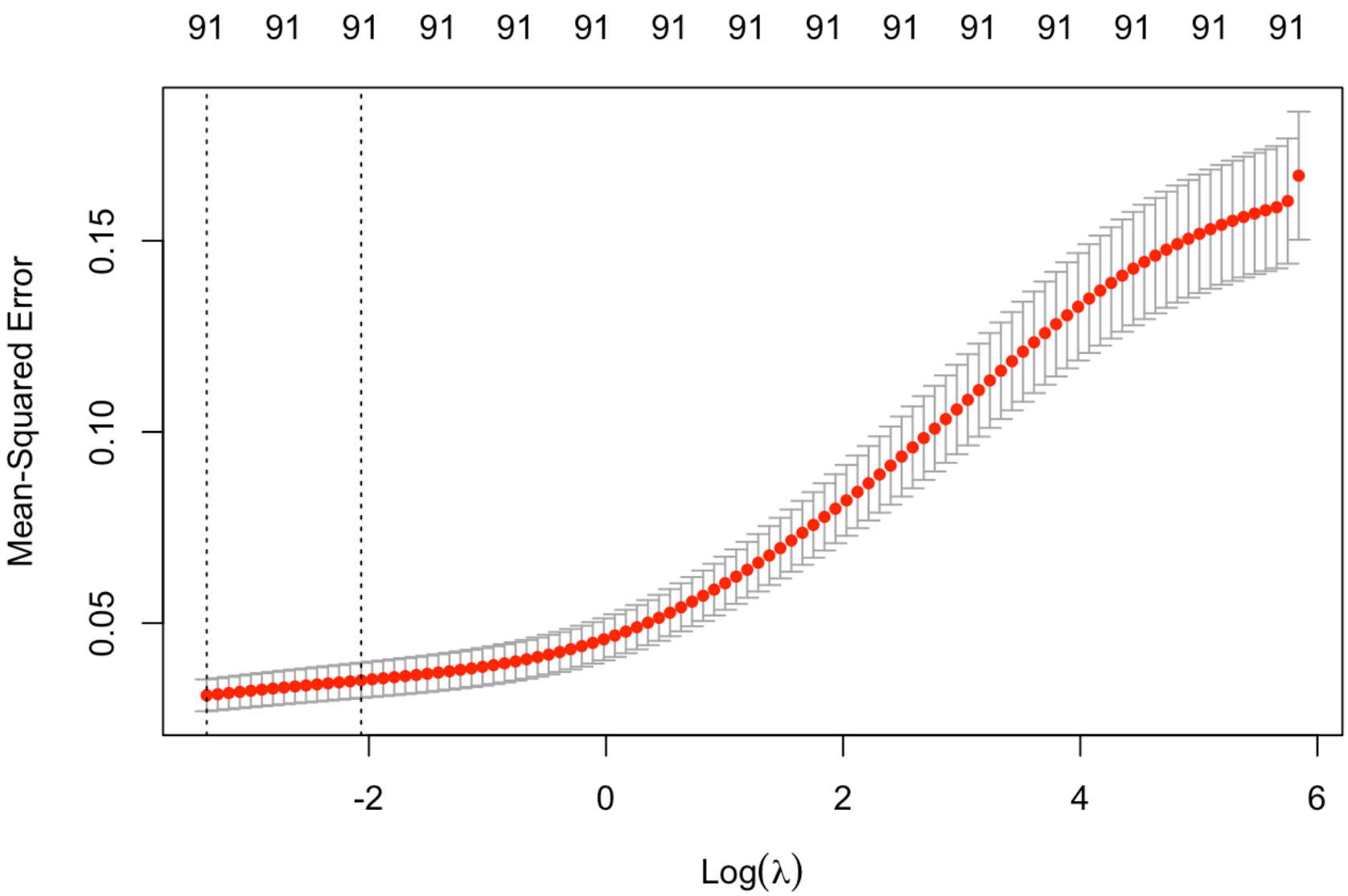
```
T=50
n=length(Y)
ntest=round(n*0.25)
sample=sample(n,ntest)
test=myData[sample,]
train=myData[-sample,]
```

```
full=function(train,test){
  full.model=lm(Y~.,data = train)
  y.pred=predict(full.model,newdata=test)
  MSPE=mean((test$Y-y.pred)^2)
  return(MSPE)
}
```

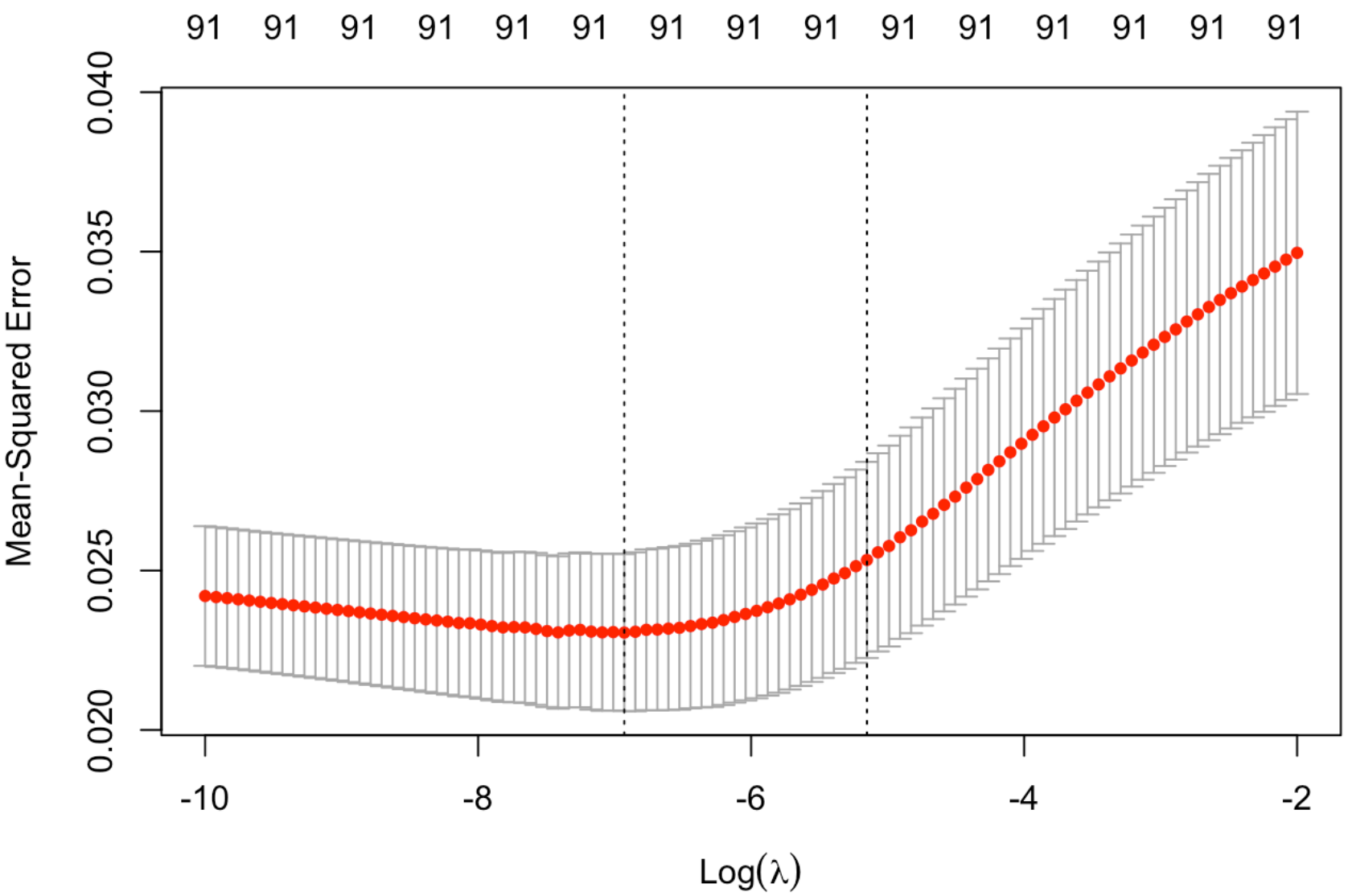
```
## For ridge regression, we need first find the correct range of lambda
cv.out=cv.glmnet(X[-sample,],Y[-sample],alpha=0)
best.lam=cv.out$lambda.min
sum(cv.out$lambda<best.lam)
```

```
## [1] 0
```

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




```
# Try a lower lambda range to get the lowest MSPE
mylasso.lambda.seq=exp(seq(-10,-2,length.out=100))
cv.out=cv.glmnet(X[-sample,],Y[-sample],alpha=0,lambda = mylasso.lambda.seq)
plot(cv.out)
```



Now we found the range of lambda, construct the ridge function

```
ridge=function(xtrain,xtest,ytrain,ytest){
  cv.out=cv.glmnet(xtrain,ytrain,alpha=0,lambda = mylasso.lambda.seq)
  best.lam=cv.out$lambda.min
  Ytest.pred=predict(cv.out,s=best.lam,newx=xtest)
  ridge.min=mean((ytest-Ytest.pred)^2)
  best.lam=cv.out$lambda.1se
  Ytest.pred=predict(cv.out,s=best.lam,newx=xtest)
  ridge.1se=mean((ytest-Ytest.pred)^2)

  return(c(ridge.min,ridge.1se))
}

ridge(X[-sample,],X[sample,],Y[-sample,],Y[sample,])[1]
```

```
## [1] 0.02872024
```

Now we look at the lasso function

```
lasso=function(xtrain,xtest,ytrain,ytest){
  cv.out=cv.glmnet(xtrain,ytrain,alpha=1)
  best.lam=cv.out$lambda.min
  Ytest.pred=predict(cv.out,s=best.lam,newx=xtest)
  lasso.min=mean((ytest-Ytest.pred)^2)

  best.lam=cv.out$lambda.1se
  Ytest.pred=predict(cv.out,s=best.lam,newx=xtest)
  lasso.1se=mean((ytest-Ytest.pred)^2)

  mylasso.coef=predict(cv.out,s=best.lam,type="coefficients")
  var.sel=row.names(mylasso.coef)[which(mylasso.coef != 0)[-1]]
  mylasso.refit=lm(Y~.,myData[-sample,c("Y",var.sel)])
  Ytest.pred=predict(mylasso.refit,newdata=myData[sample,])
  lasso.refit=mean((Ytest.pred-ytest)^2)

  return(c(lasso.min,lasso.1se,lasso.refit))
}

lasso(X[-sample,],X[sample,],Y[-sample,],Y[sample,])
```

```
## [1] 0.02897240 0.03150113 0.03015074
```

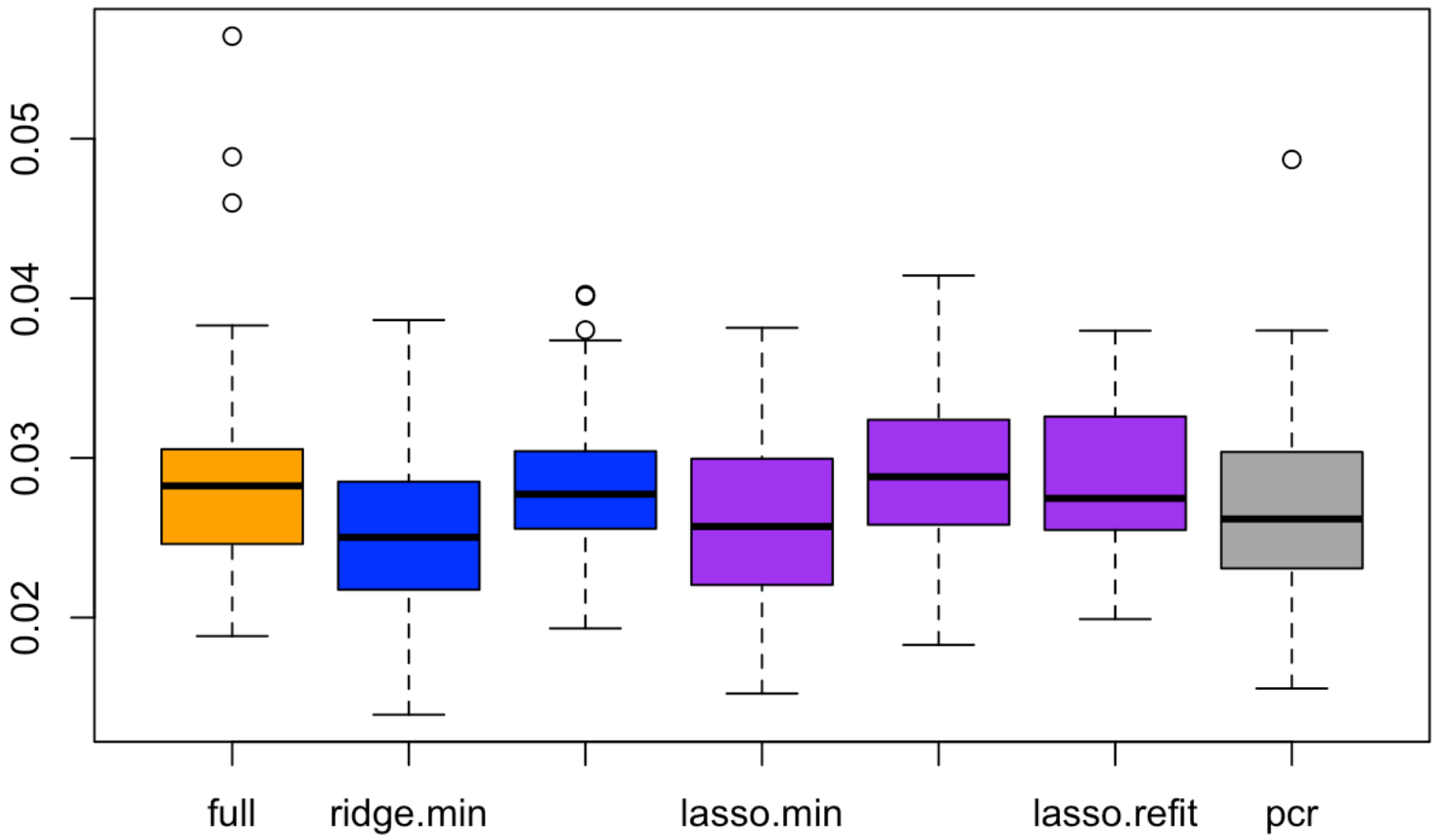
Now we work on the PCR function

```
myPCR=function(train,ytrain,test,ytest){
  mypcr=pcr(Y~., data=train,validation="CV")
  CVerr=RMSEP(mypcr)$val[1, ,]
  adjCVerr=RMSEP(mypcr)$val[2, ,]
  best.ncomp=which.min(CVerr)-1
  if(best.ncomp==0){
    Ytest.pred=mean(myData$Y[-sample])
  }
  else{
    Ytest.pred=predict(mypcr,test,ncomp=best.ncomp)
  }
  pcr=mean((Ytest.pred-ytest)^2)
  return(pcr)
}
myPCR(train,train$Y,test,test$Y)
```

```
## [1] 0.02572769
```

Run the simulation 50 times

```
MSPE=matrix(rep(NA,350),50,7)
for(t in 1:T){
  sample=sample(n,ntest)
  test=myData[sample,]
  train=myData[-sample,]
  MSPE[t,1]=full(train,test)
  MSPE[t,2:3]=ridge(X[-sample,],X[sample,],Y[-sample,],Y[sample,])
  MSPE[t,4:6]=lasso(X[-sample,],X[sample,],Y[-sample,],Y[sample,])
  MSPE[t,7]=myPCR(train,train$Y,test,test$Y)
}
colnames(MSPE)=c("full","ridge.min","ridge.1se","lasso.min","lasso.1se","lasso.ref
it","pcr")
boxplot(MSPE,col=c("orange",rep("blue",2),rep("purple",3),"darkgrey"))
```



Try on another data

Load the newdata

```
myData=read.csv("BostonData3.csv")
myData=myData[,-1]
dim(myData)
```

```
## [1] 506 592
```

```
X=data.matrix(myData[,-1])
Y=data.matrix(myData[,1])
MSPE.new=matrix(rep(NA,300),50,6)
for(t in 1:T){
  sample=sample(n,ntest)
  test=myData[sample,]
  train=myData[-sample,]
  MSPE.new[t,1:2]=ridge(X[-sample,],X[sample,],Y[-sample,],Y[sample,])
  MSPE.new[t,3:5]=lasso(X[-sample,],X[sample,],Y[-sample,],Y[sample,])
  MSPE.new[t,6]=myPCR(train,train$Y,test,test$Y)
}
colnames(MSPE.new)=c("ridge.min","ridge.1se","lasso.min","lasso.1se","lasso.refit"
,"pcr")
boxplot(MSPE.new,col=c(rep("blue",2),rep("purple",3),"darkgrey"))
```