Homework 2b

STAT6306; Due: 10/03/2017 before class

Introduction

A major issue with antiretroviral drugs is the mutation of the virus' genes. Because of its high rate of replication (10^9 to 10^{10} virus per person per day) and error-prone polymerase[1], HIV can easily develop mutations that alter susceptibility to antiretroviral drugs. The emergence of resistance to one or more antiretroviral drugs is one of the more common reasons for therapeutic failure in the treatment of HIV.

In the paper 'Genotypic predictors of human immunodeficiency virus type 1 drug resistance'[2], a sample of in vitro[3] HIV viruses were grown and exposed to a particular antiretroviral therapy. The susceptibility of the virus to treatment and the number of genetic mutations of each virus were recorded.

```
load("hiv.rda")

X = hiv.train$x
Y = hiv.train$y

n = nrow(X)
p = ncol(X)

geneLabels = colnames(X)
```

Let's revisit this problem with the omitted parts. Use/update the forward selection or ridge code you submitted previously and add the requested lasso results.

Question 1

We may have (at least) two goals with a data set such as this:

- inference: can we find some genes whose mutation seems to be most related to viral susceptibility?
- prediction: can we make a model that would predict whether this therapy would be efficacious, given a virus with a set of genetic mutations

(a) Inference

(i)

Find the selected model for:

- forward selection using BIC as the criterion
- lasso
- refitted/relaxed lasso

```
#Forward selection
#SOLUTION
if(!require(leaps)){install.packages('leaps',repos='http://cran.us.r-project.org');require(leaps)}
```

[1] An enzyme that 'stitches' back together DNA or RNA after replication
[2] The entire paper is on the website. Try to see what you can get out of it if you have the time.
[3] Latin for 'in glass', sometimes known colloquially as a test tube

```
## Loading required package: leaps
```

```
outForward      = regsubsets(x=X,y=Y,nvmax=p,method='forward')
```

```
## Warning in leaps.setup(x, y, wt = weights, nbest = nbest, nvmax = nvmax, :
## 12 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
## Warning in rval$lopt[] <- rval$vorder[rval$lopt]: number of items to
## replace is not a multiple of replacement length
```

```
#  note this warning is that the feature matrix
#  isn't full rank.  This is, there are redundant
#  columns in it:
cat('The rank is: ',qr(X)$rank,' while the # of features is: ',p,'\n')
```

```
## The rank is:  196  while the # of features is:  208
```

```
sumForward      = summary(outForward)
model.forward   = sumForward$which[which.min(sumForward$bic),]
S.forward       = model.forward[-1]#get rid of the intercept entry
lm.forward      = lm(Y~X[,S.forward])#regsubsets only scores models, not fit them
betaHat.forward = coef(lm.forward)
```

```
#lasso
if(!require(glmnet)){install.packages('glmnet',repos='http://cran.us.r-project.org');require(glmnet)}
```

```
## Loading required package: glmnet
```

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

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

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

```
#SOLUTION
```

```
#refitted/relaxed lasso
#SOLUTION
```

(ii)

Comparing the selected models for each of the above methods

```
#SOLUTION
geneLabels[S.forward]
```

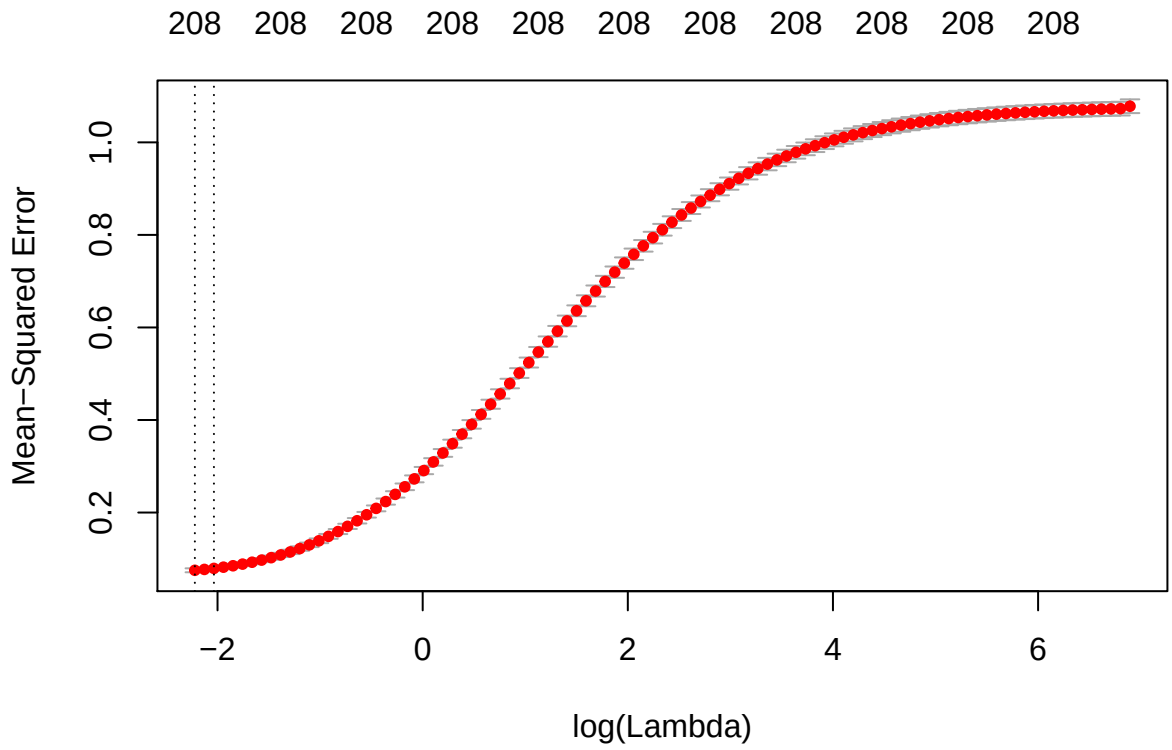
```
##  [1] "p33"  "p54"  "p58"  "p65"  "p67"  "p69"  "p75"  "p90"  "p102" "p115"
## [11] "p117" "p151" "p172" "p184" "p187" "p210" "p215"
```

(b) Prediction

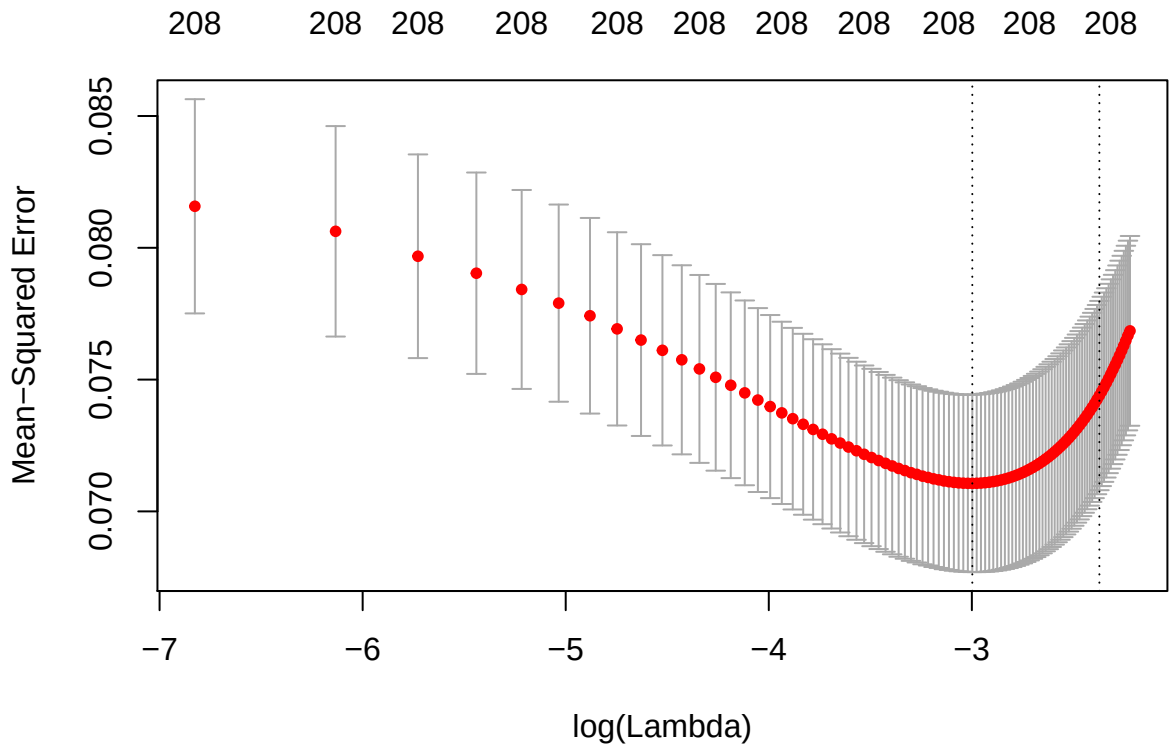
(i) Ridge regression

Now that are looking at prediction, we can use ridge regression (which only addresses prediction):

```
ridge.cv.glmnet = cv.glmnet(X,Y,alpha=0)
plot(ridge.cv.glmnet)
```




```
min.lambda      = min(ridge.cv.glmnet$lambda)
lambda.new      = seq(min.lambda, min.lambda*0.01,length=100)
ridge.cv.glmnet = cv.glmnet(x = X, y = Y, alpha = 0,lambda = lambda.new)
plot(ridge.cv.glmnet) #now it is in middle
```



(ii) Prediction on a test set

Now, let's look at some predictions made by these methods. Use the following for the test set:

```
X_0 = hiv.test$x
Y_0 = hiv.test$y
```

Find an estimate of the risk using the test observations for

- forward selection using BIC as the criterion
- ridge
- lasso
- refitted/relaxed lasso

```
#### Get predictions on test set:
Yhat.test.forward  = X_0[,S.forward] %*% betaHat.forward[-1] + betaHat.forward[1]
Yhat.test.ridge    = predict(ridge.cv.glmnet,X_0,s='lambda.min')
#SOLUTION

# Get estimate of prediction risk via the test set error
Yhat.test.forward   = mean((Yhat.test.forward - Y_0)**2)
pred.error.ridge    = mean((Yhat.test.ridge - Y_0)**2)
#SOLUTION

cat('The prediction error from forward selection + BIC is: \n',
     Yhat.test.forward,'\n')
```

```
## The prediction error from forward selection + BIC is:
##  0.07491952
```

```
cat('The prediction error from ridge is: \n',
     pred.error.ridge,'\n')
```

```
## The prediction error from ridge is:
##  0.09705571
```

SOLUTION: WHICH ONE HAS THE MINIMUM ESTIMATE RISK USING THE TEST ERROR?

Question 2

Using the lasso with CV minimum tuning parameter, which gene mutations are related to susceptibility?

```
#SOLUTION
```

Question 3

At which gene mutation sites are the presence of a mutation associated with a decrease in viral susceptibility to this particular drug? Hint: Consider the signs of the coefficients. What gene site has the largest estimated effect using $\hat{\beta}_{\text{lasso}}(\hat{\lambda})$?

```
#SOLUTION
```

Question 4

Save the HIV feature matrix to your hard drive in the 3-vector format. Read it back into R's memory and verify that you saved/loaded it correctly

```
#SOLUTION
```