Homework 2

STAT6306; Due: 09/19/2017 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.

Question 1

```
hiv = load("hiv.rda")

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

geneLabels = colnames(X)
```

(a)

What are n and p in this problem? What are the features in this problem? What are the observations? What is the supervisor? **Note:** Attempt to answer this question before moving on to the rest of the questions.

```
#SOLUTION
```

SOLUTION

Question 2

Consider the feature matrix $\mathbb{X}$. It is composed of 0's and 1's, with a 1 indicating a mutation in a particular gene. Look at the output for the following chunk of code.

```
table(X)
```

```
## X
##      0      1
## 135589  10843
```

What results do you see? What does this indicate?

[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.

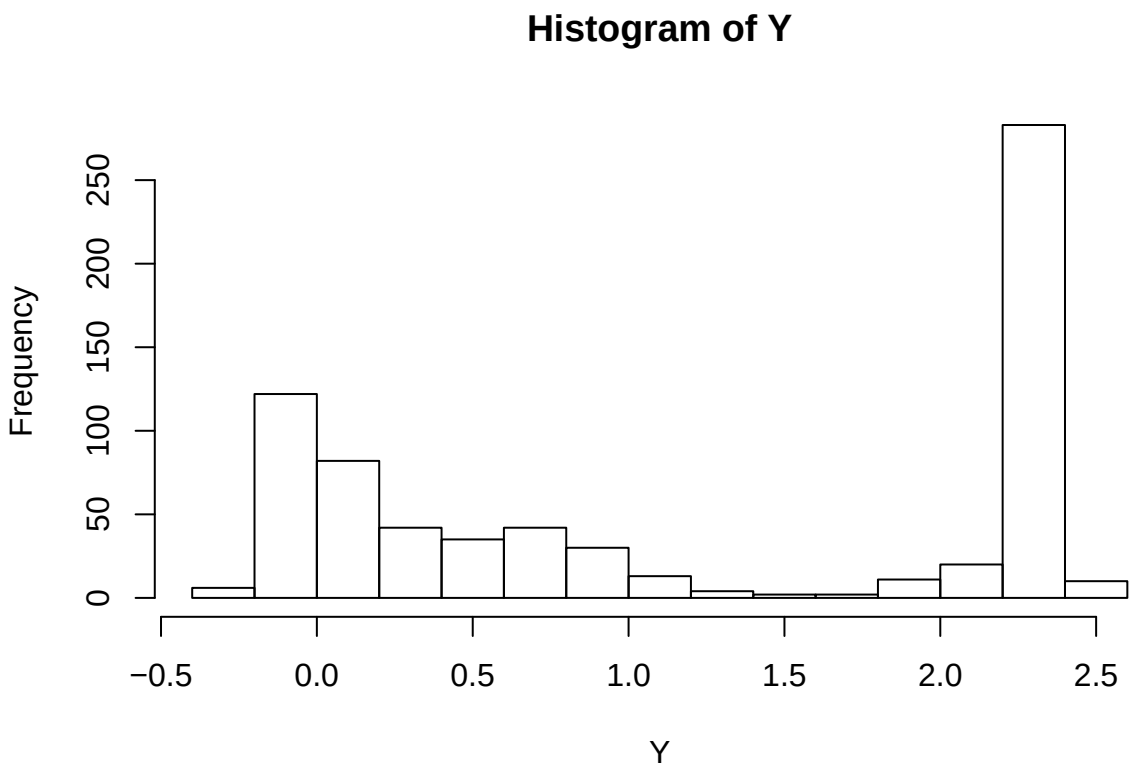
[3] Latin for 'in glass', sometimes known colloquially as a test tube

SOLUTION

Question 3

The supervisor is the log transformed susceptibility of a virus to the considered treatment, with large values indicating the virus is relatively more resistant (that is, not susceptible). Run

```
hist(Y)
```



What plot did you just create? What does this indicate?

SOLUTION

Question 4

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)

Try to find the best subset solution for this problem. Discuss any problems or findings you discover. In particular, how many possible models are there?

SOLUTION

There are 2^p possible different solutions, which, given the size of p (208), gives us 4.1137614e+62 possible solutions, which is way too large to compute all subsets.

(b) Inference

(i)

Find the selected model for:

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

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

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

```
## Warning: package 'leaps' was built under R version 3.3.2
```

```
#SOLUTION
```

```
#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 2.0-5
```

```
#SOLUTION
```

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

(ii)

Comparing the selected models for each of the above methods

```
#SOLUTION
```

(c) Prediction

(i) Ridge regression

Now that are looking at prediction, we can use ridge regression (which only addresses prediction). Using the package glmnet, plot the CV curve over the grid of λ values and indicate the minimum, and finally report the CV estimate of the prediction risk for $\hat{\beta}_{\text{ridge},\hat{\lambda}}$

Note: There is no need to report the p coefficient estimates from the ridge solution. Also, glmnet has a grid problem. Make two plots, one that shows the problem and one that shows it being corrected.

```
#SOLUTION
```

(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

```
#SOLUTION
```

(d)

Challenge Suppose we didn't have access to any test data. How could you provide an estimate of the risk? What are the pros and cons of your proposal?

```
#SOLUTION
```

Question 6

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

```
#SOLUTION
```

Question 7

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
```

Additional challenge problems:

I don't want to overwhelm you with homework problems. However, there are additional topics that are relevant for an interested student. You don't need to do these/turn them in.

Question 8

Derive, implement, and run both "batch" and "stochastic" gradient descent for this HIV data.

Question 9

The LARS algorithm is quite similar to forward selection. Run LARS using the option forward.stagewise and compare it to forward selection using Mallow's Cp.

Question 10

Try and use a GIC-based method instead of K-fold CV for finding $\hat{\lambda}$ using the HIV data.