STAA 577 HW3

Enqun Wang

April 8, 2016

Due: April 11 at noon

A major issue with antiretroviral drugs is the mutation of the virus' genes. Because of its high rate of replication (10^{9} to 10^{1}0 virus per person per day) and error-prone polymerase1, 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 following paper, a sample of in vitro3 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 the data set `hiv.rda` and create

```
load(file = "/Users/ewenwang/Dropbox/CSU-MAS/STAA 577/Homework/hiv.rda")

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

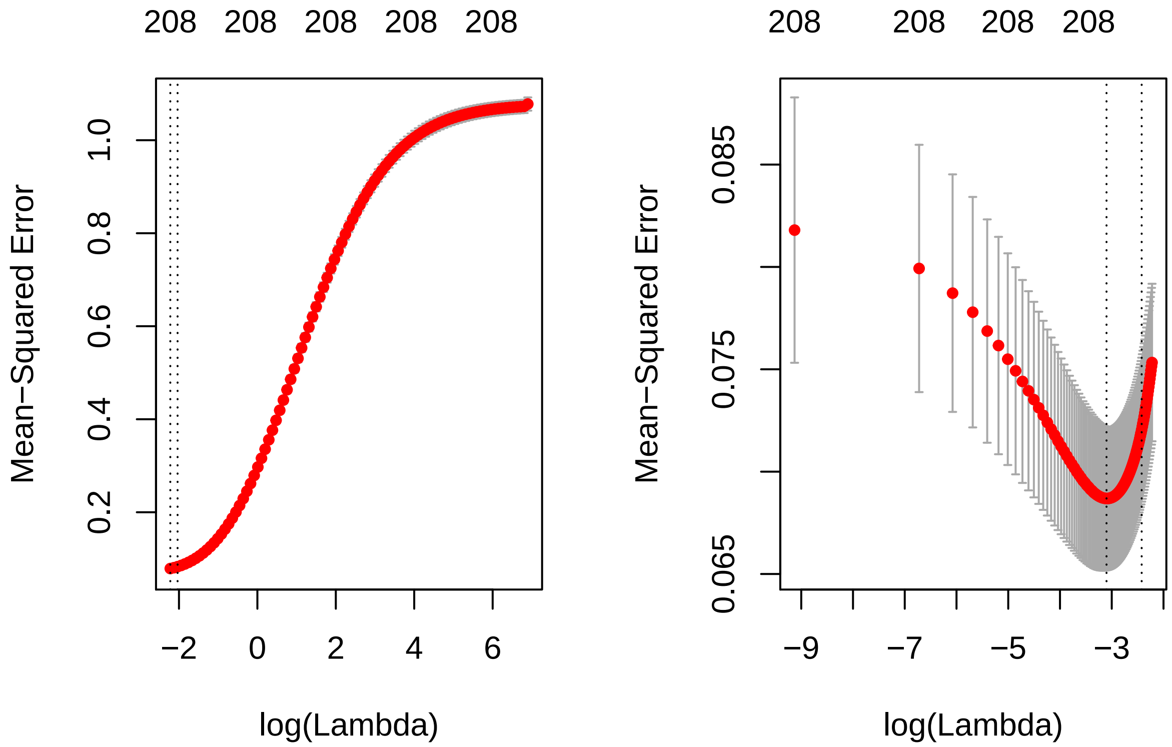
1. Let's return to a question from the previous homework. Let's do ridge regression again. `glmnet` has a grid problem. Make two plots, one that shows the problem and one that shows it being corrected. Include your code for doing the correction.

```
require(glmnet)

par(mfrow = c(1, 2))
ridge.cv = cv.glmnet(x=X,y=Y,alpha=0)
plot(ridge.cv)

ridge.cv = cv.glmnet(x=X,y=Y,alpha=0)
min.lambda = min(ridge.cv$lambda)
lambda.new = seq(min.lambda,min.lambda*.001,length=100)
ridge.cv = cv.glmnet(x=X,y=Y,alpha=0,lambda=lambda.new)
lambda.hat = ridge.cv$lambda[which.min(ridge.cv$cvm)]

plot(ridge.cv)
```



```r
par(mfrow = c(1, 1))
```

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

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

Produce three different predictions on the test set:

- Ridge (using CV minimum λ)
- Lasso (using CV minimum λ)
- Refitted Lasso (using procedure from class)

For instance, for lasso, you would do the following:

```r
lasso.cv.glmnet = cv.glmnet(X,Y,alpha=1)
Yhat.test.lasso = predict(lasso.cv.glmnet,X_0,s='lambda.min')
pred.error.lasso = mean((Yhat.test.lasso - Y_0)**2)
print(pred.error.lasso)
```

Give an estimated risk for the Ridge and Lasso using the training data. You don't have to write the code, but detail how in principle you could find a risk estimate for the refitted lasso. What are the test prediction errors for all three methods? Which method has the lowest risk on the test data?

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

Yhat.train.ridge = predict(ridge.cv.glmnet, X, s='lambda.min')
risk.train.ridge = mean((Yhat.train.ridge - Y)**2)

Yhat.test.ridge = predict(ridge.cv.glmnet, X_0, s='lambda.min')
pred.error.ridge = mean((Yhat.test.ridge - Y_0)**2)

# lasso
lasso.cv.glmnet = cv.glmnet(X,Y,alpha=1)

Yhat.train.lasso = predict(lasso.cv.glmnet, X, s='lambda.min')
risk.train.lasso = mean((Yhat.train.lasso - Y)**2)

Yhat.test.lasso = predict(lasso.cv.glmnet,X_0,s='lambda.min')
pred.error.lasso = mean((Yhat.test.lasso - Y_0)**2)

# refitted lasso
require(scalreg)
lasso.ssr = scalreg(X = X,y = Y,LSE=T)
# the LSE parameter indicates if we want to do refitted lasso

Yhat.train.refitted = X %*% lasso.ssr$lse$coefficients
risk.train.refitted = mean((Yhat.train.refitted - Y)**2)

Yhat.test.refitted = X_0 %*% lasso.ssr$lse$coefficients
pred.error.refitted = mean((Yhat.test.refitted - Y_0)**2)

require(knitr)
training.risk = c(risk.train.ridge, risk.train.lasso, risk.train.refitted)
test.risk = c(pred.error.ridge, pred.error.lasso, pred.error.refitted)
df <- data.frame(training.risk, test.risk)
row.names(df) <- c("Ridge", "LASSO", "Refitted LASSO")
kable(df, align = "l")
```

	training.risk	test.risk
Ridge	0.0446359	0.1021915
LASSO	0.0441975	0.0687893
Refitted LASSO	0.0574791	0.0738412

- According to the summary table above, LASSO has the lowest risk on the test data.

3. Using the lasso, what gene mutations are related to susceptibility?

```r
betahat.lasso = coef(lasso.cv.glmnet,s='lambda.min')
which(abs(betahat.lasso) > 0)
```

```
##  [1]   1  22  34  42  44  54  56  57  61  62  63  65  70  72  74  76  81
## [18]  89  94  97 100 101 102 103 106 110 119 120 124 130 138 147 148 149
## [35] 157 159 160 161 163 164 170 172 179 180 184 188 197
```

- The gene mutations shown above are related to susceptibility.

4. Which gene mutation sites lead to an decrease in viral susceptibility to this particular drug?

Hint: Consider the signs of the coefficients. What gene site has the largest estimated effect using the CV-min lasso?

```r
which(betahat.lasso == max(betahat.lasso))
```

```
## [1] 159
```

- 159 gene mutation sites lead to an decrease in viral susceptibility to this particular drug.