**QBS 177: Methods for Statistical Learning for Big Data**

Lab summary for High dimensional regression

The goal of this lab is to perform variable selection and penalized regression on a real dataset of diffuse large B-cell lymphoma (DLBCL) survival times and gene expression data (Rosenwald *et al*., 2002).

1. Use the subset of the data to perform stepwise regression:

data <- read.table("fulldat.txt", header=F)

data[1:3,1:10]

y <- data[,1]

x <- data[,3:50]

workingD <- data.frame(cbind(y,x))

1. Random divide the data into training and testing.

set.seed(1)

train <- sample(1:240,160,replace=FALSE) # randomly divide the data in to training and testing

test <- (1:240)[-train]

1. Perform stepwise regression on training set.
2. Check your training model on testing set.
3. Use the full data to run penalized regression.

y <- data[,1]

x <- data[,3:7401]

workingD <- data.frame(cbind(y,x))

1. Divide the data into training and testing

set.seed(1)

train <- sample(1:240,160,replace=FALSE) # randomly divide the data in to training and testing

test <- (1:240)[-train]

ytrain <- y[train]

ytest <- y[test]

xtrain <- x[train,]

xtest <- x[test,]

1. Perform univariate regression on all genes on training set.
2. Perform penalized regression on training set.
3. Check your training model on testing set