Assignment # 4, Stat 8053, Fall 2013

Problem

These problems are due on Friday, October 28.

1. We will use data on diabetes. Ten baseline variables, age, sex, body mass index, average blood pressure and six blood serum measurements, were obtained for each of $n = 442$ patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline. The statisticians were asked to construct a model that predicted response Y using the ten predictors. There are two goals to analysis: (1) produce accurate predictions of Y from the baseline measurements that can be applied to future patients, and (2) identify the most important predictors of Y. The data are at

   ```
   loc <- "http://www.stat.umn.edu/~sandy/courses/8053/Data/diabetes.txt"
   data <- read.table(loc, header=TRUE)
   ```

   Use `rpart`, `randomForest`, `glmnet` and a reasonable parametric method for these data. Also, do fitting starting with appropriate graphical summaries to guide you through fitting a standard GLM and at least one other method we have discussed. What are the important variables, according to the lasso?

   One published analysis of these data fit the `lasso` using not only the ten predictors but also all two-factor interactions among them for a total of 55 predictors. You can create a matrix as input to `glmnet` with these predictors:

   ```
   big.data <- model.matrix( ~ (. - Y - 1)^2, data)
   ```

   The function `model.matrix` is used to convert a formula and a data frame into a matrix. The specified model uses all columns as predictors except for the one with label Y and the intercept. The `( )^2` generates all main effects and two-factor interactions. If there had been factors, they would have been converted to dummy variables.

2. Repeat question 1, but use the `GlaucomaM` data we have examined previously. This data set will cause problems because the logistic regression model with all 62 predictors gives an exact fit, meaning that there is a hyperplane that separates the cases from the controls. The coordinates of the hyperplane, essentially the regression coefficients, are likely not to be well determines, as there may be many separating hyperplanes. For a comparison glm model, then I suggest starting with

   ```
   > m1 <- glm(class ~ ., familty=binomial, data=GlaucomaM, subset=construction)
   > m2 <- step(m1)
   ```

   which should retain about 40 of the predictors, depending on the `construction` set you use.