ML4Bio – 2016 - Assignment #4 - Regression

Due Date: April 1st, 2016. Email your completed assignments to ml4bio@gmail.com

Please include any R code you used for answering the questions, as well as a short description of what you did to answer those questions and any figures that you generated. We can read PS, PDF or Word documents.

Question #1: Regressions of random data

Statistical theory predicts that if there is no relationship between X and Y (and X and Y are normally distributed) then the estimate of b1 in a simple linear regression (where you try to predict Y based on X) has a normal distribution where the mean is 0, and the variance is proportional to 1/n, where n is the number of observations in X and Y.

(a) Use R to generate random observations and confirm that this is true (hint: linear regression in R can be done using the `lm` function, and `qqnorm` can be used to test whether some numbers follow a normal distribution. You'll need to do a bunch of linear regressions to see the distribution, and you'll need to try a bunch of n values).

(b) How important is the assumption that X and Y are normally distributed? (hint: repeat part (a) using data that is not normally distributed)

(c) Statistical theory also predicts a relationship between the estimates of b0 and b1. Can you see any relationship in your random regressions? What is it? Does it depend on the assumption that X and Y are normally distributed?

(d) Say you did a multiple regression on 500 dimensional random data. Based on what you found in (c), would you predict the variance of the b parameters to be more or less than the b1 estimates you get from doing 500 univariate regressions? Explain briefly. Test your prediction using R.

Question #2: Regularized regression as a method to find eQTLs

Download and install the `glmnet` package for R and download the eQTL data from Brem et al. on the course website.

(a) Use L1 regularized regression to find the markers that are most strongly associated with expression of AMN1. (hints: use `plot()` to see the coefficients in your `glmnet` model, and replace missing data for both genotypes and expression levels. Missing data is indicated as a '2' for the genotypes)

(b) Make a multiple regression model using the markers you found in part (a). How much variance does each marker explain? Make a plot showing the predicted and observed gene expression values based on this model.

(c) Try using standard multiple regression to find the markers associated with AMN1 expression. Explain what you find.