Stat 230, Spring 2016
Homework 12: Ridge Regression, LASSO, Regression Trees
Due Monday 5/2/15 at 11:59pm on bspace.

PART 1 will take you a while to figure out how to do the 3 new methods, but once you write your code for PART 1 you just pretty much just generate new data in PART 2 and run the same code on the new data. Your comments in PART 3 don't have to be deep but I'd like you to think about what you did!

PART 1. Orthonormal variables

Generate 20 independent standard normals to use as predictors. Generate the coefficients in the same way as `makedata()` from PART 1 of HW 11.

1. Repeat (some of) HW#11 part 2 for new methods.
   Use PCR, Ridge Regression, LASSO, and Regression Trees to find the best model. (In HW #11 you did this using PCR, but you'll repeat this to compare to the other methods.) Standardize every variable before you start the analysis.
   Use cross validation (10-fold for all parts of this assignment) to find a suitable λ for Ridge and LASSO. For each method, basically repeat the first three paragraphs of HW #11 Part 2, the first 3 paragraphs, but with slightly different pictures as discussed in b) below.

   Use cross validation to find the best model complexity (number of nodes) for a regression tree. I recommend the `tree` package but `rpart` is also fine. For cross validation, you need to get the fitted values explicitly so that you can compute RSS from each fold as you did in HW#11. Also, you can use the formula $df = n - p$ where n is the number of observations and p is the number of terminal nodes. This is what the tree package does and has the logic that each split is essentially like fitting a dummy variable defined according to the split rule.

2. Make the same picture as on page 62 of HTF for the 4 methods as well as for OLS. Let's do it just like in HTF this time, get "best RSS + 1 SE" based on the CV process outlined on page 18 here:
   `http://www.stat.cmu.edu/ ryantibs/datamining/lectures/18-val1-marked.pdf`. You did the calculation of this in HW #11 although you didn't use it then when you did the picture. For the regression tree picture you can allow the p (number of splits = number of terminal nodes+1) to go up to 20 (or 21) and stop, that will make it easier to compare to the other methods.

3. Which is best?
   Basically this is a contest to see which method can get the smallest cross validated RSS. Replicate the entire analysis 100 times and make adjacent boxplots showing the 100 replications of cross validated RSS from each method. Comment briefly on the model that won (smallest average of the 100 repetitions of cross validated RSS) and whether there is a theoretical and/or intuitive reason why it won.

PART 2. Correlated Explanatory Variables

1. Rewrite the `makedata()` function so that the first 15 vectors of X have the following correlation structure. Group the 15 vectors into 5 groups of 3. Within each group of 3, each vector will have the same correlation with the other 2 vectors. For the 5 groups, those correlations will be .1, .3, .5, .7, and .9. They should still be standard normal. Those correlations will be used for the true model in generating the data, so the values in the data will be approximate. For example, you could generate all the data as independent standard normals and then $X[,2] = 0.1 * X[,1] + \mathrm{sqrt}(1 - .1^2) * X[,2]$ would work for the second vector. Recall that we did this for generating correlated errors in HW 4, you can go back and do it the way we did then as well. Other than this change in X, the rest of the `makedata()` function can be left alone.
2. Repeat PART 1 of this assignment. Again you should end up with 4 side-by-side boxplots of CV RSS for 100 replications as well as new pictures using the correlated data.

PART 3. What happened?

Submit your comments for part 3 in the pdf along with the rest of the lab. A couple of sentences for each question below (answer the question and explain briefly) is sufficient. Think about the results in comparison to the discussion on page 82 of HTF.

1. What methods did better in PARTS 1 and 2?
2. What variables got shrunk the most in PART 2 by ridge and LASSO compared to OLS? Why?
3. What about the effective coefficients in PCR, did these end up getting shrunk similarly to ridge and LASSO?
4. For the regression tree, did the variables used to make splits correspond to the same variables considered to be most important in the other methods?