Project description: Regression

1. Find a new dataset you like to study for regression analysis from here, here, here, or any other source. Something nobody has posted on blackboard.

2. Submit a proposal on the Discussion Board on Blackboard in which you:

   (a) Describe the response variable and the predictors.

   (b) When you remove the missing values, what is n and p?

   (c) How many categorical predictors and how many numerical predictors?

   - The number of features p is at least 40.
   - The sample size n should be at least ten times the number of features p.

3. For each $n_{train} = 0.8n$, repeat the following 100 times, do the following for the different models mentioned below.

   (a) Randomly split the dataset into two mutually exclusive datasets D_{test} and D_{train} with size n_{test} and n_{train} such that $n_{train} + n_{test} = n$.

   (b) Use D_{train} to fit lasso, ridge, and random forrest.

   (c) Tune the λs using 10-fold CV.

   (d) For each estimated model calculate

   $$R^2_{test} = 1 - \frac{\frac{1}{n_{test}} \sum_{i \in D_{test}} (y_i - \hat{y}_i)^2}{\frac{1}{n_{test}} \sum_{i \in D_{test}} (y_i - \bar{y}_{test})^2},$$

   and R^2_{train}.

4. Create a presentation with less than 8 slides. Your objective is to be clear and concise. Hence I recommend the following:

   (a) a brief description of the nature of the data as discussed in part 2 above. (1 slide)

   (b) Show the side-by-side boxplots of R^2_{test}, R^2_{train}. We want to see two panels. One for training, and the other for testing. (1 slide)

   (c) For one on the 100 samples, create 10-fold CV curves for lasso and ridge. Record and present the time it takes to cross-validate ridge/lasso regression. Please do not more two digits to present the time. (1 slide).

5. For all the data do the following:

   - Using 10-fold cross validation, fit ridge and lasso . Also fit random forrest.
   - Also record the time it takes to fit a single ridge/lasso regression (including the time needed to perform cross-validation parameter tuning), and random forrest. Create a table 3 × 2 table, the 3 rows corresponding to the 3 methods, and the two columns for test R^2 and time (using no more than two digits). Specifically,

the first column should show a 90% test R^2 interval based on the 100 samples, and the second column the time it takes to fit the model on all the data (as described in the sentences above). Is there a trade-off between the time it takes to train a model and it's predictive performance? (1 slide).