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. How was the data collected?

   (b) Impute missing data-points with their mean. What is n and p?

   (c) Standardize the numerical predictors using equation (6.6) in the ISLR book.

   - 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_{learn} to fit lasso, elastic-net $\alpha = 0.5$, ridge, and random forrest.

   (c) Tune the λs using 10-fold CV.

   (d) For each estimated model calculate

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

4. Create a presentation with less than 6 slides. Your objective is to be clear and concise. Hence I recommend the following:

   (a) a brief description of the nature of the data, shape, etc as discussed above. (1 slide)

   (b) Show the side-by-side boxplots of R^2_{test}, R^2_{train}. (1 slide)

   (c) For one on the 100 samples, create 10-fold CV curves for lasso, elastic-net $\alpha = 0.5$, ridge. (1 slide).

   (d) For one on the 100 samples, show the side-by-side boxplots of train and test residuals (1 slide). Comment on the distribution and size of the residuals.

   (e) Present bar-plots (with bootstrapped error bars) of the estimated coefficients, and the importance of the parameters. If you have something interesting to say about coefficients that are (or are not important) say it. (1 slide)

   (f) Summary slide: summarize the performance and the time need to train each model in a table and comment on it. (1 slide)

5. Submission guidelines:

- Create an account on https://github.com/ and upload your code, data and the pdf of your presentation. The code and data should be such that if I download it in one folder, I should be able to run it, and create the figures you present in your pdf. Reproducibility is an important step in any data analysis project.
- Upload your 5 minute video on vimeo (with or without a password, depending on what you are more comfortable). One way to create the video is to full-screen your pdf presentation, and talk over it while Quicktime is recording your screen and sound.
- Submit a SINGLE PDF with hyperlinks that take me to your github page, and the vimeo page where your presentation is uploaded. See the template.