## MATH 456 — Mathematical Modeling

**Assignment** #6: LASSO Regression for Diabetes Data Analysis **Due Date:** March 25, 2022, 11:59 PM

In the attached diabetes dataset, there are 10 baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline. This dataset was originally used in the paper below.

Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani (2004) "Least Angle Regression," Annals of Statistics (with discussion), 407-499.

Note that Each of these 10 feature variables have been mean centered and scaled by the standard deviation times n_samples (i.e. the sum of squares of each column totals 1).

**Goal**: Apply the LASSO regression model to study the effects of baseline variables to the disease progression.

Here are some concrete steps to achieve this goal.

1. Split the dataset randomly into a training set and a test set. For example you may take 300 obs in training and 142 obs in test.

2. Solve the LASSO model with soft thresholding. Note that there are 10 features and one response in this case. To quantify the prediction accuracy, you will need to use the root mean squared error (RMSE) defined by

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{m} |y^i - \hat{y}^i|^2}.$$

   Here $\hat{y}^i$ are the predicted responses and $y^i$ are the true response in the test set. Compute the RMSEs by varying the regularization parameter $\lambda$ in the LASSO model, and plot the values RMSEs versus $\lambda$'s. Discuss your findings.

**Note: The paper above is mentioned just for referencing purpose and does not contain essential stuff relevant to this question. Please implement LASSO regression using either Python or Matlab. Document your results in Word/PDF/Markdown/Jupyter Notebook. You need to upload the document along with the codes.**