IEOR E4525
Martin Haugh

Due: Friday 26^{th} February 2016

Assignment 3

Answer Questions 1, 2 and 3 and two of the remaining three questions

1. *(Regression Analysis of the CEO Pay Data Set)*

   The data for this exercise is in the csv file `ceo.csv` and a description of the data-set's variables may be found in the text file `ceo.txt`. After reading this text file you should perform the following steps:

   Step 1. Read the data into R as a `data frame`.

   Step 2. Remove the last column in the data frame – it is garbage!

   Step 3. Randomly select 25% of the observations to be in the test set. (Investigate the R function `sample`.)

   Step 4. Use EDA techniques such as scatter plots, box-plots, qq plots etc. to ensure that the split is sufficiently random enough. (You may want to think about what "sufficiently random" means and how you might enforce this to be the case.)

   Step 5. Standardize the independent variables in the training set. (Note that you will want to apply the **"same"** transformation to the test set when you use it to predict the performance of your model. Be sure you understand what the word "same" in this context means.)

   Step 6. Estimate the best *lasso* model that predicts `totcomp` as a linear function of all the other predictors using $k = 10$-fold cross validation. What does the plot for the mean-square error vs the regularization parameter look like?

   Step 7. Estimate the best lasso model that predicts $\log($`totcomp`$)$ as a linear function of all the other predictors using $k = 10$-fold cross validation. What does the plot for the mean-square error look like?

   Step 8. For each of the previous two steps, use the best λ that you found to predict on the test set. Report the mean squared error on the test set in each case.

2. *(Using the Bootstrap to Estimate Standard Errors of Logistic Regression Coefficients)*

   Do Exercise 6 in Chapter 5 of ISLR.

3. *(Some Properties of Ridge Regression)*

   Do Exercise 4 in Chapter 6 of ISLR.

4. *(Leave-One-Out Cross-Validation (LOOCV) with a Simulated Data-Set)*

   Do Exercise 8 in Chapter 5 of ISLR.

5. *(Bootstrapping the Boston Data-Set)*

   Do Exercise 9 in Chapter 5 of ISLR.

6. *(Exercise 3.4 in Bishop: Error in the Predictor $\equiv$ Regularization)*

   Consider a linear model of the form

$$y = w_0 + \sum_{j=1}^{p} w_j x_j$$

together with a sum-of-squares error function of the form

$$\mathcal{E}(\mathbf{w}) = \sum_{i=1}^{N} \left(y_i - w_0 - \sum_{j=1}^{p} w_j x_{ij} \right)^2$$

Now suppose that Gaussian noise η_i with zero mean and variance σ^2 is added independently to each of the *input* variables, x_i. Show that minimizing $\mathbb{E}[\mathcal{E}(\mathbf{w})]$ averaged over the noise distribution is equivalent to a ridge regression where the intercept term w_0 is omitted from the regularizing term.