ISE 529 Engineering Data Analytics Exam 1 Fall 2017

Last Name _____ First Name _____ Masters Program _____

Exam duration 150 minutes. Laptop computer with WiFi, calculator, textbooks, and class notes are allowed. Submit a pdf to `acostame@usc.edu` with the output and code for each question.

1. (40 pts.) Fit a linear regression model for Ford Motor Co. daily returns using Standard and Poor's 500 Index (SPY) returns as the predictor variable. The slope of that regression (called the *beta* of the company) measures how sensitive the stock's return is to changes in the returns of the overall market (measured by SPY). If the slope is greater than 1, we say that the stock is more volatile than the market. For instance, if the slope is 1.5, then a 1% increase in the SPY index, would result in an average increase of 1.5% in the stock's return.

   The R^2 measures the proportion of the total risk that is market-related. For instance, if $R^2 = 0.4$ we would conclude that 40% of the variation in Ford returns is explained by the variation in SPY returns (market-related risk). The remaining 60%, is the proportion of risk that is specific to Ford, and not market-related (firm-specific risk).

   If the market is expected to rise an investor would seek companies with large betas. If market is expected to fall companies with small betas are preferable.

   a) Download daily prices from Ford Motor Co. and Standard and Poor's 500 Index (SPY) from 01-01-2015 to 01-01-2017. Report first 3 and last 3 rows of each.
   b) Find the daily returns (use `Adj.Close` price) for each of them. Use `head()` and `tail()` to report first and last six rows of each.
   c) Fit a linear regression model. What is the *beta* of Ford Motor Co.? Interpret Ford's R^2.
   d) Create a scatterplot of Ford (y-axis) vs. SPY (x-axis) daily returns. Show the fitted equation as a dash line on the scatterplot. Use same boundaries for both axes.
   e) Identify the day of the largest outlier. Label that day on the outlier in the scatterplot.

2. (30 pts.) Life insurance companies are keenly interested in predicting how long their customers will live because their premiums and profitability depend on such numbers. An actuary gathered data from 100 recently deceased male customers. He recorded the age at death, whether he was a smoker (1 for smoker, 0 for non-smoker), plus the ages at death of his mother and father, the mean ages at death of his grandmothers and grandfathers (see file `insurance.csv`).

   a) Fit a regression model `m1` with all predictors. Use `m2=stepAIC(m1)` to simplify the model. For `m2`
      i. Find a 90% CI on the mean longevity of smokers whose mothers lived to 75 years, whose fathers lived to 65 years, whose grandmothers averaged 85 years, and whose grandfathers averaged 75 years.
      ii. Use `set.seed(2)` to divide the data set into a training and a test set (50%). Compare the $\sqrt{MSPE}$ of `m1` and `m2`.

3. (30 pts.) The data set `stockdata` from library `huge` consists of the price and company information of S&P 500 stock shares. The data set consists of two dataframes `names.csv` and `prices.csv`. They are available on Blackboard.

   a) Find the correlation matrix `C` of these prices. Then use `which(C==max(C),arr.ind=T)` to find the largest correlation, and their row and column numbers in `C`. Identify the companies with the largest correlation. Report their full name.
   b) Build a scatterplot of their prices
   c) How many `Health Care` companies are in the full dataset?
   d) Report the name of the two most correlated companies in the Financial Sector.