GX5004: HW 2

To the extent that it helps you to learn, you may work with fellow students on this assignment. R and Python have extensive libraries online that can guide you on this assignment. [25 points – each question for 5 points]

1. In words, the "standard normal distribution" is a normal distribution with mean zero and variance one, denoted here as N(0,1). For a random variable X that is distributed as a standard normal, mathematically we write $X \sim N(0,1)$. [5 points]
    a. Using R or Python, write code to draw at random 10 observations from a N(0,1) random variable. Instruct the machine to calculate the mean, variance and standard deviation of your draws.
    b. Repeat this exercise using 10,000 draws from a N(0,1), instructing again the machine to calculate the mean, variance and standard deviation of your draws.
    c. Repeat this exercise with 1,000,000 draws from a N(0,1), instructing again the machine to calculate the mean, variance and standard deviation of your draws.
    d. What conclusions, if any, do you draw from increasing the sample size?
    e. Submit your code and results.
2. We discussed at some length the bivariate linear regression model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. [5 points]
    a. Go to http://www.random.org/integers/ and generate two series of 1,000 random integers with values between 0 and 9. Call one series y and the other x.
    b. Using Python or R, fit the bivariate linear regression model.
    c. Examine your t-statistic to evaluate whether it is greater than two in absolute value. Would you reject or fail to reject that there is *any relationship* between these two series?
    d. Submit your series, your code, and your results.
3. Go to Yahoo!Finance (or the source of your choice, such as Bloomberg) and download a daily price series for a particular publicly-traded stock of your choice for a ten-year time period (don't use Apple), as well as the daily price series on the exchange on which it trades. [5 points]
    a. Using R or Python, calculate the log returns of each series as the natural log of the ratio of (price today/price yesterday). Use the reported closing price for this exercise.
    b. Using R or Python, generate a histogram of log returns of the stock of your choice.
    c. Using R or Python, generate a scatterplot that relates the log returns of your stock of choice to the log returns of the exchange on which it is traded.
    d. Finally, using R or Python, fit a linear model to obtain estimates of what finance folks call the "alpha" and the "beta". Is "alpha" significantly different than zero at a 95% level? Does a 95% confidence level for "beta" include one?
    e. Submit all code and results.
4. Download the file train.dta from the course website. These data are formatted as a Stata dataset. [5 points]

a. Read this dataset into R or Python. (For R, you may find the "foreign" library of use. For Python, check out Pandas. The goal here is to get you familiar with reading datasets with alternative formats.)
b. Generate summary statistics for the following variables in the data:
   - d, which is an indicator for whether a particular email is spam
   - x1, which is an attribute of the email
c. Using least squares, regress d on x1. (For R, check out lm. For Python, check out StatsModels.) Congratulations, you have created a support vector machine that you will use to forecast whether an incoming email with a different attribute is spam.
d. Suppose you set the threshold that an email is spam if the predicted value exceeds 1.[1] I give you a new email with an attribute value 0.65. Would you classify it as spam or not spam?
e. I give you another new email with an attribute value of 0.99. Would you classify it as spam or not spam?
f. Submit code and results.

5. This is a very challenging question, but it addresses several key topics in data analytics. You should work collectively on a solution with the recognition that you may not complete it. The phrase "data generating process" (or DGP) is often used to describe the hypothetical process by which observations arise in the real world. We discussed at some length the bivariate linear regression model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. In this problem, we will work with a specific DGP and evaluate features of $\widehat{\beta_1}$, the least squares estimate of β_1. [5 points]
   a. Suppose your DGP is $y_i = 1 + 2x_i + \epsilon_i$, where $x \sim N(0,1)$ and $\epsilon \sim N(0,1)$.
   b. Using R or Python, write code to generate 1,000 draws for x and ϵ. Use these draws to generate y in accordance with the DGP in a.
   c. Using R or Python, write code to estimate the bivariate model, $y_i = \beta_0 + \beta_1 x_i$ and summarize the findings.
   d. Repeat b. and c. above five times for a new set of random draws for each replication. (This effort is called Monte Carlo simulation.)
   e. Given what you've done in d., Suppose you wrote code to repeat b. and c. above 1,000 times, each time recording the estimated value of β_1. What do you think a histogram of these 1,000 replications of the estimate value of β_1 would show?
   f. Suppose that you were not interested in the estimate of β_1 per se, but instead in some functional transformation, such as the estimate of $\exp(\beta_1)$. What might you do with your 1,000 replications from e. above to inform you about the distribution of the estimate of $\exp(\beta_1)$?
   g. Submit code and results.

[1] Remember that the predicted value is simply the estimated coefficient from your regression times the value of the attribute.