Midterm (Total points = 100)

Econ 294A

April 24, 2023

Important

- **Deadline of submission**: April 30 till midnight
- Use any IDE for the coding. Upload Jupyter code (.ipynb) or Python code file (.py) file on Canvas. Use comment option or markdown inside Python code for the answers.
- **Use only base python, Pandas, Numpy, matplotlib, seaborn, scipy for your answer. Solutions based on any other packages will not be considered as valid**.
- The grading will be done based on coding technique and the efficiency of code.

Question 1: Use "JSTdatasetR3.xlsx" for this question. (*Total marks: 30*)

i Load the data using Pandas and extract the summary statistics of all variables

ii Create a new data set with only a few variables from the full data - country, exports, imports, ca, gdp, xrusd.

iii Create two new variable - (I) ratio of CA to GDP (II) Openness = (exports + Imports)/GDP. Examine the relation between ca to GDP ratio with Openness for USA, Germany, France and Japan using different plots

iv Select a random sample of size 50 with replacement for each country and check the average value of CA/GDP ratio and Openness.

Question 2: Use numpy and pandas for this question (*Total marks: 30*)

i Generate a random sample of 1000 from beta distribution of a = 10, b = 1 (*denote by* x) and generate another random sample of size 1000 from normal distribution with mean 0 and sd 10 (*denote by* ϵ).

ii Create a new numpy array using the following formula

$$y = 0.5 + 0.3 * x + \epsilon \tag{1}$$

Now, create a scatter plot of x and y using seaborn package. Fit a regression line with a confidence band in the plot. Do not use statmodels or scikit-learn.

iii Now, create 5000 samples (with replacement) of size 1000 from x and y. Calculate the median value of x and y. Plot the median values of x & y on a scatter plot (Don't include the regression line in the plot).

iv Now, create 5000 samples (with replacement) of size 1000 from x and y. Create a function to calculate each sample's inter-quartile range (IQR) and standard deviation (SD) of x and y values. Use your function to calculate the IQR and SD for each sample and provide a summary of the IQR & SD values.

Question 3: Generate a random sample of size 10,000 from a normal distribution with mean 0 and SD 1. (*Total marks: 40*)

i Calculate the sample mean and median using numpy

ii Calculate the bootstrap sample mean and median using 50,000 samples (samples are drawn with replacement and sample size is 10,000). Store those bootstrap sample mean and median in a data frame. Also, calculate the bootstrap estimate of mean by taking mean over all bootstrap mean values. Similarly, calculate the bootstrap estimate of median using median of all bootstrapped median values (*Do not use "Bootstrap" function from scipy*)

iii Now, determine the mean and median values from the sample using following logic

- Remove one observation each time and calculate the mean & median values based on remaining 9,999 observations from the original sample

$$\bar{x}_i = \frac{1}{n-1} \sum_{j \neq i, j=1}^{1000} x_j \text{ and } Me(x)_i = Median_{j \neq i} x_j \tag{2}$$

- For instance, remove the first observation and calculate mean and median values from the remaining 9,999 observations. Then remove the second observation and calculate mean & median on the remaining 9,999 observations. Continue this approach till the 10,000th observation.
- After completing the process, you should have 10,000 mean and median values from the sample.

Now, calculate grand mean of those 10,000 mean values and grand median of those 10,000 median values. We call it Jackknife resampling.

iv Lastly, use the "Bootstrap" function from scipy to estimate the sample mean and median.

v Create a table with all of these estimates of mean & median values (using estimates from part i - iv of this question). Each row should have values of mean and median estimates from each method. Label the rows suitably.

End