Chicago Booth, LSE and UBC

Pre-Doctoral Research Assistant Application

March 2023

Please complete the following data analysis and computation tasks. It should be possible to complete them in 2-3 hours. Please submit a short PDF document (preferably written in LaTeX) describing the outcome of the tasks, including an explanation of what you have done and a summary of the findings. There is no page limit, but please be as concise as possible. You should provide enough information for a trained economist to be able to replicate your findings. Please also submit a copy of your code. It should be readable and reproducible. For all two tasks, feel free to use the statistical software and/or programming language of your choice but make sure the output (graphs, etc) is easy to digest. If any of the instructions below are ambiguous or unclear, please make what you consider to be a reasonable assumption and note the assumption in your document. Similarly, if there are any mistakes in formulas or notation, don't get flustered; just assume what you need to provide an answer. As long as you have a firm grasp of data analysis and scientific computing you should be able to complete this. Good luck!

Task 1: Data Analysis—State Level Inflation

Inflation behaves differently at the regional versus the aggregate level. The goal of this task is to introduce you to state level inflation data in the United States.

1. Read in the .csv file *statecpi_beta.csv* accompanying this task. This file contains quarterly state level data on inflation, starting in 1978. This dataset has been hand coded from archival sources, and may contain mistakes or missing observations. Please briefly summarize any outliers in the data, patterns of missing data, and how you plan to deal with the data issues that arise. Please **keep only** states that have information on inflation in both 1987 and 2017.

2. Inflation has a great deal of dispersion across states. In **one single graph** plot the median, 25th and 75th percentiles of state level inflation for each quarter. Has inflation dispersion been rising over time?

3. In 2009Q4, what share of states had inflation more than 100 basis points away from median state level inflation?

4. What percent of the total variation in state level inflation is due to differences across individual states? What percent of the variation is due to common changes in inflation across all states?

5. How would you go about testing the following hypothesis? "The persistence of inflation in the United States has been falling over time. However, declining persistence is due national and not to state level factors." Write the estimating equation and/or write code that tests the hypothesis.

Task 2: Text to Data

The file "PatentsRawData.csv" was downloaded from Google Patents and contains patents of aircraft producers between the years 1920 and 1960. The objective of this task is to assign the patents to the specific aircraft plant (factory) where the invention was made.

The patent data specifies the name of the *company* that has the property right to the patent (the "assignee"), but companies often have several plants and the patent doesn't indicate in which of the company's plants the invention occurred. The file "PlantLocations.csv" lists all the companies and the location of each of their plants. The patent data specifies the name of the inventor ('inventor/author"), and the place and state of residence of the inventor ("inventorlocation").

You are asked to find a systematic way to assign each of the patents in "PatentsRawData.csv" to one of the specific aircraft plants in "PlantLocations.csv". Note that there is no known "correct" answer to this question. Your job is to find a way to maximize the likelihood that patents will be assigned to the correct plant. This is an advanced computational task. You are NOT asked to provide a good final solution and definitely should not try to do this task manually. Instead, you are asked to write code that makes a first attempt at solving this problem. You may also provide an algorithm that explains how you would go by achieving this task. You will be judged by the clarity of your thinking, not the result.

1. The first practical problem that we face is that the "inventorlocation" variable is just a raw chunk of text which contains the residence of the inventor, but also some other text that we do not need. Also, because the text is a result of automatic text recognition from scanned images, there are many typos. The objective is to create two new variables "inventorcity" and "inventorstate" giving the city and state where the inventor is located. Note that the inventor may live in locations that don't appear in "PlantLocations.csv". Provide code that makes progress towards this goal or an algorithm that would achieve this aim.

2. Now that you have the inventor's location, how would you go by assigning the patent to a specific plant (probably the plant where the inventor works)?