CSC2552 Assignment 1

Instructions

- Your work must be your own.
- This assignment is due at 11:59pm on Wednesday, December 7 (hard deadline).
- Your assignment must be typed. You may use any word-processing software you like (e.g. LaTeX).
- You may use whatever programming language and plotting libraries you like.

Problem 1 [25 pts]

In our second discussion week, we read the widely-circulated paper "Social capital I: measurement and associations with economic mobility" by Chetty et al. Thanks to the era of open and reproducible science we're living in, we can look at the data ourselves. Working through this problem will give you a chance to handle some real-world data, reproduce a state-of-the-art result, and perform a follow-up analysis of it. The authors have put together a very nice resource to walk you through the data and results. Check it out at `https://www.socialcapital.org/` since we'll be using it for this question.

The authors have made all of the data available on their data page: `https://data.humdata.org/dataset/social-capital-atlas`. For this question, we'll only use the data entitled *Social Capital Atlas - US Counties.csv*, which you should download from `https://data.humdata.org/dataset/85ee8e10-0c66-4635-b997-79b6fad44c71/resource/ec896b64-c922-4737-b759-e4bd7f73b8cc/download/social_capital_county.csv`. Unzip the archive on your hard drive. Take a look at how the data is structured and do some quick explorations before you continue.

(a) We will recreate a version of their Extended Data Fig. 4 on page 22. Recall that in this analysis, the authors showed how the county-level clustering coefficient relates to the county-level economic connectedness. This figure only shows the values for certain counties, and emphasizes that there is a lot of heterogeneity—for some counties there is a positive relationship, and for some counties the relationship is negative. You're curious what the effect looks like for all counties. Using the data you downloaded, the goal is to draw a scatterplot where the x-axis is the county-level clustering coefficient, the y-axis is the county-level economic connectedness, and every county is represented by a point in this space. To make this task easier, you can use the replication code the authors have provided (`https://opportunityinsights.org/wp-content/uploads/2022/08/social_capital_replication.zip`), although it is written in Stata. If you aren't familiar with Stata, you can use whatever software you are comfortable with. Don't worry about typesetting or aesthetic considerations.

(b) If the authors hadn't looked at particular counties as in their paper, and only looked at the overall effect as done in part (a), would their interpretations and conclusions have changed? Explain in 2–3 sentences.

Summary. For this question, hand in your plot from part (a), append the code you used to produce your it, and submit a 2–3 sentence answer for part (b).

Problem 2 [20 pts]

In this question, we'll explore ways of approximating experiments with big data through natural experiments.

(a) Explain the difference between a natural experiment and a randomized experiment in 1–2 sentences.

(b) As you've learned, natural experiments are a way to exploit random variation "in the wild" — but in order to do so, you must find a source of random variation. One common such source is the weather (i.e. sometimes it rains and sometimes it doesn't, and this variation is usually external, or "exogenous" as economists say, to the social system being studied). Come up with a research question that you could approach by setting up a natural experiment where the weather is used as the source of as-if random variation, and briefly describe the natural experiment setup you have in mind.

(c) Briefly describe a "counting things" approach to the research question you posed in the previous question.

(d) How do these two approaches compare? Would you prefer the natural experiment or the counting things approach? Explain in 2–3 sentences.

Problem 3 [20 pts]

Enumerate Salganik's ten common characteristics of big data, and for each characteristic mention one of the four papers we read in the Observational Studies section of the course that it applies to. Explain why it was important to the paper you chose in 1 sentence. Refer to Papers 1–4 as "Koenecke et al.", "Hangartner et al.", "Chetty et al.", and "Kleinberg et al.", respectively.

Problem 4 [15 pts]

In class, we focused on crowdsourcing (or "human computation") approaches to mass collaboration. Of the papers we read in class (excluding papers from the Mass Collaboration week and Paper 10, "The Moral Machine experiment"), pick one that you think could have benefitted from including a crowdsourced component. Describe the limitation of the paper that you think crowdsourcing could address, describe your crowdsourcing approach, and describe how your approach addresses the limitation.

Problem 5 [20 pts]

Banksy, known for politically-oriented street graffiti, is one of the most famous contemporary artists in the world. But his precise identity is a mystery. In 2008, the Daily Mail newspaper published an article claiming to identify Banksy's real name. Then, in 2016, Michelle Hauge, Mark Stevenson, D. Kim Rossmo and Steven C. Le Comber (2016) attempted to verify this claim using a Dirichlet process mixture model of geographic profiling. More specifically, they collected the geographic locations of Banksy's public graffiti in Bristol and London. Next, by searching through old newspaper articles and public voting records, they found past addresses of the named individual, his wife, and his football (i.e., soccer) team. The author's summarize the finding of their paper as follows: "With no other serious 'suspects' to investigate, it is difficult to make conclusive statements about Banksy's identity based on the analysis presented here, other than saying the peaks of the geoprofiles in both Bristol and London include addresses known to be associated with [name redacted]."

(a) In 1–2 paragraphs, assess this study using the principles and ethical frameworks in the Ethics chapter of *Bit By Bit*.

(b) The authors included the following ethical note at the end of their paper: "The authors are aware of, and respectful of, the privacy of [name redacted] and his relatives and have thus only used data in the public domain. We have deliberately omitted precise addresses." Does this change your opinion of the paper? If so, how? Do you think the public/private dichotomy makes sense in this case? Discuss in 2–3 sentences.