DATA 2050 Proposal

Student: Zhirui Li

Research Professor: Alice Paul

Instructor: Andras Zsom

Brown University DSI

June 7th, 2022

Access to research data can help with various important scientific tasks, such as verification, discovery, and evidence synthesis. Data availability is becoming more recognized as vital to an efficient and progressive scientific ecosystem that produces trustworthy findings. A minimum level of credibility we would expect of all published findings is that the described analyses and results could be reproduced using publicly available datasets, a concept known as computational or analytic reproducibility. As a result, researchers should practice replicating the results of a research paper or a journal using publicly available data because replicability keeps researchers honest and can give readers confidence in research.

This project tries to replicate and extend a paper that involves public health using publicly available data. The first step is to identify a topic and several research papers related to that topic with a good amount of depth so that replication is manageable. The second step is reproducing the model using publicly available datasets and comparing my results with the original paper. In the end, I will go beyond the scope of the original paper, conducting more visualizations and analyses.

This project's main challenge is settling down on a particular topic and searching for appropriate research papers. Currently, I am searching for proper journals on websites such as Plos ONE, JAMA Network, Journal of Applied Statistics, and American Journal of Epi, among topics such as COVID-19, diabetes, abortion, gun legislation, nursing homes, foodborne diseases, vaping, and opioids. For each topic, I want to find several papers where the statistical analysis method is mainly regression-based or simulation-based, and the data used in the research is not publicly available. After I figure out a particular research paper I want to reproduce, I will collect publicly available datasets related to the topic on websites such as Kaggle, Google Dataset Search, and UCI Machine Learning Repository. All the statistical analyses will be performed using RStudio.

The first milestone for this project is to successfully replicate a paper from the area listed above and compare my result with the paper's result. The second milestone for this project is to do original research by extending the replication analysis, possibly by adding more complementary data, shifting the question, or performing a simulation study.

The structure for this project is similar to the PHP 2550 final project, where the authors try to replicate the work done by Brown B.P., evaluating the association between abortion rates and restrictive legislation. The publicly available datasets they used for the replication included a dataset containing county demographic information, a

dataset containing county election results, a dataset containing county-level abortion rates from 1988 to 2019, and a dataset containing states' abortion policies.

In their project, the authors first estimated the propensity score by constructing a linear regression between the binary variable indicating if there was highly restrictive legislation towards abortion and other demographic covariates. Then, they fitted a linear regression model where the response is the abortion rate, and the predictor is the binary variable indicating whether legislation towards abortion is restrictive or not, weighted by the propensity score. Finally, they built the same model, again adjusting for distance, and tested whether this predictor serves as a mediator in the framework by creating an additional linear regression model between space and the binary predictor of whether the county has highly restrictive legislation toward abortion.

The result for their replication is different compared to the original paper despite similar demographic characteristics data. In particular, the coefficients for the binary variable indicating legislative restrictiveness are insignificant in both models (one estimated change in abortion rates, and the other one estimated the same thing but adjusted for distance), whereas, in the original paper, both the coefficients are significant at the 5% significance level.

The authors conducted a simulation study in the extension part to estimate the association between highly restrictive legislation and abortion rates under four different distance definitions. The first scenario is that women will always go to the nearest facility in their state; the second scenario is that women will always go to the nearest facility or may travel to other states if the nearest one is not available; the third scenario is that women will randomly go to any facilities within 50 miles radius from their location; the last method is that women will go to the facility with the optimized combination between the distance and the level of legislative restrictiveness. The first step in this simulation study was to randomly generate the location for each woman and calculate the minimum Euclidean distance between each woman to all facilities subject to the four scenarios. Then, fit a propensity score-weighted regression model using each of the four distance measures to assess the association between distance to facilities and the level of legislative restrictiveness.

The result of the extension part proposed a new interpretation for the original paper's conclusion. In the original article, Brown B.P. concluded that a highly restrictive legislative climate would significantly lower abortion rates. Still, in this simulation study, the authors proved that the association found in the original paper is likely to be a result-driven by an indirect effect, where the highly restrictive legislative climate eventually

lowers the abortion rates by significantly increasing the distance that the women must travel to obtain abortion care.

For my project, since I will conduct extensive research to settle down on one research paper to replicate and compare my results with the paper's results, it fulfills the requirement to propose a question, perform analysis, and draw an appropriate conclusion. Moreover, since many machine learning techniques, such as regression and simulation, will be used in the replication process, the ability to apply statistics and machine learning knowledge in a real-world setting will also be satisfied. Since I don't have access to the original dataset, I can analyze what kind of bias would occur if using publicly available data. Lastly, after reproducing the original paper, we can discuss the societal impacts my findings can lead to.