STATISTICAL METHODS IN SOCIOLOGY II
Final project instructions

Thomas Davidson

Spring 2022

REPLICATION PROJECT

The goal of this project is to replicate the statistical analyses in a published journal article. There are several important goals of this analysis. First, it will be an opportunity to get hands-on experience implementing statistical models in sociological research and will give you a greater appreciation of the methodology used in published work than you can get by simply reading a paper. Second, it will allow you to question the original authors' decisions and explore how alternative choices would have impacted the results. Third, depending on your choice of paper, it could be an opportunity to critique or extend existing work. Fourth, and finally, the replication task will highlight the challenges involved in replication and reproducibility.

This document contains information on how to select a viable paper for replication, the tasks required for the project, guidelines for writing up and submitting the analysis, and a timeline for the remainder of the semester.

SELECTING A PAPER TO REPLICATE

I recommend selecting a paper related to your research interests. Ideally, the paper should be published in a sociological journal (or related social science field). The paper must include some form of regression model. I strongly encourage you to find a paper with replication code and data, although the availability will be highly variable. If this is unavailable, you could also email the authors (start with the corresponding author) and politely request the code and/or data. Cristobal Young's 2015 class experiment found that

28% of 53 authors contacted provided students with replication packages.

The Harvard Dataverse and OpenICPSR websites contain replication materials for many recent papers. You can use the search bar on either website to find replication materials for different journals. The replication materials often include some form of README document that explains the structure of the replication data and files. Note that some of the materials posted included data but not replication code. You may want to contact the lead author to request the code in such cases. In addition to these resources, I recommend searching through relevant journals for articles of interest. Some journals include downloadable replication packages on their websites and some authors also host replication code on their personal websites and Github repositories.

Avoid any papers that use restricted data as it is unlikely that you will obtain the necessary data in a timely manner. You should also avoid any papers that require running extensive supplementary code prior to obtaining the replication dataset (e.g. agent-based models) unless you are confident you understand the approach and the relevant code. In some cases, the replication materials might include Stata code (e.g. `do` files) or files written in another programming language. Depending on the complexity of the code, it might be viable to translate this into R. If you are comfortable doing so, you can run code in other languages, but the final analyses (e.g. regression models and any output) should be reported using R in the RMarkdown document (see below for further details).

As a general rule of thumb, you are more likely to find replication data and code for more recently published papers.

TASKS

1. *Replicate a key finding of the paper.* Often multiple analyses are reported in quantitative papers. You do not need to reproduce every result in the paper. At a minimum, you should choose at least one regression model to reproduce.
2. *Estimate a Bayesian version of the original model.* Use `stan_glm` to estimate a Bayesian version of the model. You may use additional code to plot the results and show how Bayesian methods can be used to assess the result in additional ways (e.g. Plot posterior distribution, posterior predictive checks, LOO-CV).
3. *Examine robustness to alternative specifications.* The next step is to assess the robustness of the published result to alternative specifications. Your choices of alternative specifications should be motivated by your domain knowledge and statistical expertise. At a minimum, estimate three additional models, one for each of the following:

- I. A model with alternative variables (e.g. add or remove controls, transformations)
- II. A model with a different subset of the data (e.g. removing outliers)
- III. A model using a different estimator (e.g. Poisson instead of OLS, Probit instead of logit)

4. *Write up the results.* Discuss any challenges related to the replication and whether you were able to reproduce the published results. For parts 2 and 3, discuss the results of the new analyses and if the alternative specifications result in any substantive changes.
5. *(Optional) Extend the analyses of the original paper.* Is there a different kind of model you could estimate using the data that might provide insights into the research topic? Is there something different the author(s) could have done to get at the phenomenon under study? Is there a different question one could ask using the data?

PAPER FORMAT

The replication analyses should be contained in an RMarkdown file. I have provided a template on the course website. This file will contain the code used to replicate the analyses and any associated write up. You will submit the .Rmd file and a rendered PDF. The PDF should include writing, tables, and figures. Raw data and code chunks should *not* be included in the final output unless there is an important reason to show the code. References should be provided in the text using the author-date format (e.g. (McElreath 2020)).

The paper should contain the following sections:

1. *Introduction*: Briefly discuss the paper you have chosen to replicate and the particular results you will be analysing.
2. *Replication*: Present the initial replication and discuss your findings. This section should include a table or figure showing the replicated result (as close to the original paper as possible).
3. *Bayesian replication*: Present the Bayesian replication of the model and discuss your findings. Pay close attention to any discrepancies between the Bayesian and frequentist models. Use tables and/or figures to communicate your results.
4. *Alternative specifications*: Discuss each of the alternative specifications and use tables and/or figures to present your results. Use tables and/or figures to communicate your results.

5. *Discussion*: Discuss your findings and reflect upon the replication exercise.
6. *References*: Provide a section listing works cited.

The final project including the .Rmd file, the PDF, and any associated files and data should be added to a private Github repository. You will submit the project by adding me (`t-davidson`) as a collaborator to the project. I will then read the paper and try to replicate your replication materials by knitting the RMarkdown file.

TIMELINE

- March 25: Select paper to replicate
- May 2: In-class presentation
- May 11: Replication paper due