# Stat 230A Final Project Assignment

## Assignment Description

The goal of this assignment is to apply learned methods from this course to analyze real world datasets and critically appraise claims made in the academic publications. This project is a group assignment with each group consisting of two students. You are welcome to form the group among yourselves and please email me your group formation by **April 1st, 2022**. For those that have not formed groups by the deadline, I will randomly match students to their peers.

Given the goal of this assignment, both the instructor and I highly recommend the group to replicate a published paper, whose original datasets or datasets of similar nature are publicly available, re-analyze results of the paper. We will provide a list of recommended papers towards the end of this assignment document.

If you are concurrently taking a capstone class and writing an empirical report for that class, we also encourage you to combine your capstone project writing effort and this assignment. You will have to work on your own for this final assignment unless the department, in which you are undertaking the capstone class, allows group work.

For other forms of final project assignment, such as a literature review with simulation studies to compare multiple methods, please also write to me or the professor and schedule an office hour appointment with either one of us.

To guide you through this replication assignment, there are two due dates and each counts towards the final grade (40 points) of this assignment.

## Replication Examples

- Original Paper: Alan S. Gerber and Donald P. Green, "The Effects of Canvassing, Telephone Calls, and Direct Mail in Voter Turnout: A Field Experiment", *American Political Science Review*, 2000
- Replication: Kosuke Imai, "Do Get-Out-the-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments", *American Political Science Review*, 2005
- Rebuttal of the Replication: Alan S. Gerber and Donald P. Green, "Correction to Gerber and Green (2000), Replication of Disputed Findings, and Reply to Imai (2005)", *American Political Science Review*, 2005[1]

## Where to Find Academic Papers for Replication

My expertise is in economics, so I would recommend looking for papers published by the American Economic Association, since the Association enacted the policy to require authors to upload their datasets, if the dataset can be shared publicly, and codes onto the publication website in 2005. My recommendation would be looking for empirical papers published in *American Economic Review*, *American Economic Journal: Applied Economics* and *American Economic Journal: Economic Policy*. Three other major outlets for empirical papers in economics that have adopted the data sharing policy within recent five years are *The Quarterly Journal of Economics*, *Journal of*

[1] Here are two more examples of seminal papers with their results being overturned in replication studies. The first one is a re-analysis of the seminar paper "The Colonial Origins of Comparative Development" done by David Albouy. The second one is a re-analysis of "Does Competition among Public Schools Benefit Students and Taxpayers?" done by Jesse Rothstein. Both replication studies were coursework assignments for Econ250A at Berkeley completed by respective authors though the ensuing debates were bitter.

*Political Economy*, and *Journal of Labor Economics*. Most datasets are stored in Stata format in economics. To import such datasets into R, we would recommend using the `foreign()` package in R.

Top journals in other fields that have also adopted and enforced the data sharing policies are *American Journal of Political Science* since 2014 and *Sociological Methods & Research* since 2009.

Other fields are also welcome though I do not have the domain knowledge to provide much help. If you are unsure about the paper or the field, please email me or schedule for an office hour appointment to discuss the paper selection.

### Where to Find Public Datasets

- ICPSR is a user supported data repository at the University of Michigan that contains thousands of datasets. Most major universities subscribe to ICPSR and dataset is downloadable from the website after registration.
- IPUMS is a data repository at the University of Minnesota that contains harmonized census and survey datasets from the around the world.
- Panel Study of Income Dynamics is the longest running longitudinal household survey in the world and has been abundantly used by social scientists.
- National Longitudinal Survey of Youth is a longitudinal data of young adults and again has been abundantly used by social scientists.
- National Health and Nutrition Examination Surveys is a health survey that has been used in epidemiology and public health studies.
- Kaggle a recently emerged repository for the data science and machine learning community. Some recent unpublished work may have used or deposited their datasets onto this website. If you have doubt, consult me.

### Due Dates

For each submission, every individual group member must submit a copy of the assignment from their group in pdf file on bCourse. Unless otherwise stated, all written assignments are due at **1800 Pacific Time** on bCourse on the specified date. No late submissions are allowed. For all written assignments, please attach the code for analysis at the end and the code attachment does not count towards the page limit.

**Format of Submission** The title of the submission should include your group number and your initials, e.g. `Group01_CY.pdf`. Each writeup submission must have your group number and your name in the upper right hand corner, include a brief centered title, use an 11 or 12 pt font, one-half or double spacing, use an 1 inch margin on all sides and use a ragged-right (left-justify) format or justified (for those using LaTeX). For example:

Group 01, Chaoran Yu

Stat 230A Replication Paper Summary

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Donec ac odio tempor orci dapibus. Commodo quis imperdiet massa tincidunt nunc pulvinar. Neque convallis a cras semper auctor neque vitae.

**April 1** **Group Membership**. For those that are not matched by this date, I will randomly match each person to another to form a group. Each group will receive a group number.

**April 11** `[05 pt]` **Paper Summary & Summary Statistics Table**. Maximum two-page writeup. The writeup should summarize the paper's research question and its answer, describe the datasets used towards answering the question, and clean the dataset to replicate and interpret a summary statistics table that presents distributional characteristics (mean, median, IQR, etc.) of key variables and covariates used in the empirical analysis. The summary table need not replicate exactly as the table produced in the paper.[2]

**May 11** `[35 pt]` **Final Project Due**. Maximum 20-page writeup. The following is a rough outline for a replication project.

- **Paper Summary & Summary Statistics Table**: Summarize the paper's research question and its answer, describe the datasets used towards answering the question, and clean the dataset to replicate and interpret a summary statistics table that presents distributional characteristics (mean, median, IQR, etc.) of key variables and covariates used in the empirical analysis;
- `[10 pt]` **Replicate Main Result**:
  1. Describe and state the statistical assumptions for linear models clearly in both English and mathetical symbols;
  2. Replicate their main regression result and interpret it in English;
  3. Critically appraise their statistical assumptions for linear models;
- `[10 pt]` **Replicate Robustness Checks**: In economic and other social sciences, after the main result, there will be a section titled "Robustness Checks" (common in observational studies papers) or "Extensions" (common in experimental papers). The purpose of these sections is to convince readers that the main result holds up against various critiques on the statistical assumptions or to illustrate subtleties in interpreting the main result. Pick at least one of the robustness checks or extensions from the paper, and replicate it. Also include a writeup explaining what this robustness check or extension achieves;
- `[15 pt]` **Re-analyze**: Re-analyze the main result in the paper using methods (such as leverage scores, LOOE, etc.) taught in this class. Justify why these methods can be applied to the setting in the paper. Another strategy would be changing the sample selection used by the authors to produce their results. Compare and contrast with the main result. If results differ significantly, conjecture and/or analyze the source of discrepancies.

## Suggested Papers for Replication

- Maximilian Auffhammer and Ryan Kellogg, "Clearing the Air? The Effects of Gasoline Content Regulation on Air Quality", *American Economic Review*, 2011
- Rudiger Bachmann, Tim O. Berg and Eric R. Sims, "Inflation Expectations and Readiness to Spend: Cross-Sectional Evidence", *American Economic Journal: Economic Policy*, 2015
- Matias D. Cattaneo, Sebastian Galiani, Paul J. Gertler, Sebastian Martinez and Rocio Titiunik, "Housing, Health, and Happiness", *American Economic Journal: Economic Policy*, 2009

[2] It is said that 90% of the effort in an observational empirical paper is cleaning the dataset. It is fine if you are not able to clean the dataset as described in the paper.

- Rafael Di Tella and Ernesto Schargrodsky, "Do Police Reduce Crime? Estimates Using the Allocation of Police Forces After a Terrorist Attack", *American Economic Review*, 2004
- Stefano DellaVigna and Ethan Kaplan, "The Fox News Effect: Media Bias and Voting", *Quarterly Journal of Economics*, 2007
- Stefano DellaVigna and Ulrike Malmendier, "Paying Not to Go to the Gym", *American Economic Review*, 2006
- Joan Esteban, Laura Mayoral and Debraj Ray , "Ethnicity and Conflict: An Empirical Study", *American Economic Review*, 2012
- Melissa S. Kearney and Phillip B. Levine, "Early Childhood Education by Television: Lessons from Sesame Street", *American Economic Journal: Applied Economics*, 2019
- Erzo F. P. Luttmer and Monica Singhal, "Culture, Context, and the Taste for Redistribution", *American Economic Journal: Economic Policy*, 2011
- Stelios Michalopoulos, "The Origins of Ethnolinguistic Diversity", *American Economic Review*, 2012
- Betsey Stevenson and Justin Wolfers, "Bargaining in the Shadow of the Law: Divorce Laws and Family Distress", *Quarterly Journal of Economics*, 2006
- Ebonya L. Washington, "Female Socialization: How Daughters Affect Their Legislator Fathers", *American Economic Review*, 2008