Stanford Pre-Doc Application
Common Data Analysis
Fall 2025

1 Overview

Your goal is to write an empirical evaluation of the small business training program described below. You are welcome to use ChatGPT, Copilot, and other AI resources in any way. Please submit your writeup and code in a single zip file to this google form. The zip file name must be YourLastName_YourFirstName.zip.

- Research question: what is the causal effect of the program on business outcomes?
- Writeup: the writeup should read like a short research paper like those published in *Science* or *American Economic Review: Insights*. The section order should probably be (i) introduction, (ii) data, (iii) empirical strategy, (iv) empirical results, and (v) discussion/conclusion. (Don't spend your time on any literature review about business training programs.) The writeup should be in pdf format, ideally generated in LaTeX.
- Code: make sure that the code you submit is clearly commented, and that anyone could execute your code while changing no more than the name of the current directory.

2 The Small Business Training Program

- The program began on January 1st, 2013.
- It consisted of several training sessions for firms' managers.
- The goal was to increase productivity.
- Only firms with 100 or fewer employees were eligible to participate in the program.
- Some eligible firms enrolled in the program, but some others did not.

3 Data

We have sent you three datasets for a broad set of firms:

1. `firm_information.csv` - Basic firm information
2. `aggregate_firm_sales.csv` - Firm sales by month
3. `monthly_data/YYYY-M.csv` - Firm auxiliary data by month

These are administrative datasets and may have reporting errors.

4 Advice

- In the data section, precisely describe any issues with the data and how you addressed them.
- Think hard about what econometric approaches can identify the program's causal effect. For background, you might take a look at Scott Cunningham's free online book, *Causal Inference: The Mixtape*, and think about what chapter(s) could be relevant here.
- Your writeup may have a couple of tables, but we'd like you to focus on creating good figures. You should have figures that
  - describe the data and/or your data cleaning,
  - allow the reader to visually see the program's causal effect, and
  - allow the reader to assess whether your empirical strategy delivers the true causal effect or is instead confounded.
- You may end up with various estimates of the program's causal effect, with different empirical strategies, outcomes, and robustness checks. State which estimate you think of as your "primary" estimate, and explain why you chose that one.
- Policymakers reading your research need to decide whether the program should be offered to more firms in the future. In your conclusion, describe what an economist might recommend, and/or the caveats and additional information one might need.

5 Coding Sample

We would like to see a coding sample that you are proud of. If you have any such sample, please include it in your zip file. Links to GitHub repositories are encouraged if available. The output could be a paper, a figure that you put together, a model that you estimated, or output from a machine learning algorithm, among others. There's no need to do further editing—just send us what you have. If it might be helpful for us to have more context, briefly describe why the problem was challenging and how you dealt with it.