Problem Set 10

Due on December 9 (Sat), 2023, noon

To successfully complete this problem set, please follow these steps:

1. Download the contents of the assignment Dropbox link in a folder (directory) on your computer designated for this problem set only. Follow the directory structure, e.g., putting the data folder containing the datasets as a subdirectory of the project directory.
2. Insert your answers in the yellow boxes using Microsoft Word, and prepare a single .R script for what you produce. Save the word document as a.PDF.
3. Please submit the PDF to the designated PS-XX: pdf link and your R Script to the PS-XX: R link.
4. Your name:

|  |
| --- |
|  |

1. Group members, if any:

|  |
| --- |
|  |

1. Compliance with the Academic Code on problem set[[1]](#footnote-1) (sign with an X below)

|  |
| --- |
|  |

Gradescope will have the following packages installed:

library(tidyverse)  
library(fixest)  
library(haven)  
library(glue)  
library(gt)  
library(scales)  
library(modelsummary)

The problem set is worth 8 points, given that PS-09 accounted for 12 points in total. Part of the 8 points was already accounted for in the 11/29 class assignment. The assignment for 11/29 asked you to the read and comment on each of the two writing samples for that class.

# Problem 1 (the only coding question)

In class on Nov. 28, I briefly mentioned the clustering of standard errors. Although I could not get into the internal logic of this procedure, this gives us all the tools we need to replicate one of the regressions we saw in the beginning of the semester: The Graduation program by Banerjee, Duflo, Karlan.   
  
Here, we will revisit this paper one last time. They explained their estimating regression as follows:

|  |
| --- |
|  |

and in the caption of Table 3, they report that “standard errors are clustered at the unit of randomization.”

Using the household level data "data/graduation\_pooled\_hh.dta" whose variables are described below, replicate their main regression examining the ITT effects of the Graduation program on the index of assets at endline 2 (asset\_index\_fup). **Present a readable regression table with two regressions**:

1. The simplest bivariate regression regressing the outcome on a treatment assignment (assignment).
2. A regression that includes the control variables included in the excerpt, including the baseline outcome *Z*. This regression should ***cluster standard errors*** by the unit at which randomization occurred. Its point estimate and standard error should be identical to the one presented in the article (Table 3, Endline 2 Asset Index ITT). The estimates of the control variables should not be shown in your table.

| **Variable name** | **Description** |
| --- | --- |
| rand\_unit | An ID for the unit of randomization assignment. In countries where treatment was assigned at the village level, all households from the same village are assigned a single treatment. In countries where treatment was assigned at the individual level, each household unit gets a single randomization unit. |
|  |  |
| geo\_cluster | An ID for the geographic units that were used as strata in the random assignment. Stands in for and in equation 1 of the paper. These can be used as fixed effects (see appendix). |
| assignment | Assignment to enter into the Graduation program |
|  |  |
| Variables that start with the character control\_ | Control variables measured before treatment assignment |
|  |  |
| m\_asset\_hh\_index\_bsl | Whether or not the asset index value at baseline was missing: 1 if missing, 0 if not missing. Asset index values that are initially missing are imputed with its country mean (a mean of 0). |

## 

|  |
| --- |
|  |

See the formatting reminder and guidelines for tables at the appendix.

## 1.2

Substantively interpret the two model’s coefficient point estimates and standard errors.

|  |
| --- |
|  |

# Problem 2

The following questions ask you to reflect on the course to synthesize your understanding of research design and the papers. Your answers will also help me adjust the course later. Please write something substantive in the text for credit.

## 2.1

In terms of *writing*, what is (if any) the most memorable line, sentence, paragraph or title, from any of the class papers you read? Please quote.

|  |
| --- |
|  |

## 2.2

What are the new *statistical* concept or idea that you might *remember the most* from what you learned through this course? Please list one or several. If there was not much new or insightful, feel free to write that too.

|  |
| --- |
|  |

# Problem 3

I am continually thinking of ways to improve the effectiveness of the course. At the end of this month, you will evaluate the course through Yale’s official course evaluation. The following survey is different: it asks you about specific aspects of the course that might not be reflected in the official course evals. [**https://yalesurvey.ca1.qualtrics.com/jfe/form/SV\_1T6DJsi0X1Kl2XI**](https://yalesurvey.ca1.qualtrics.com/jfe/form/SV_1T6DJsi0X1Kl2XI)

(please take enter this survey after the last day of class on Dec 6).

My hope is that a few minutes of your time now will substantially benefit future students of this course and will help me become a better teacher. I would greatly appreciate it if you could be candid and thoughtful in your answers. Please sign / mark with an X here once you have completed the survey. – SK

|  |  |  |
| --- | --- | --- |
| I have completed the anonymous survey: |  |  |

# Reminders for Github and Final Project

* Final Project Guidelines: <https://yale.instructure.com/courses/89156/assignments/383593>
* Your final project should consist of a PDF document submitted to Canvas, with the first page including a Github link for your repository including code and minimal datasets. This allows us to understand how you created your figures and tables. It should be possible for us to navigate the repository to see each script.
* If you want to keep your Github repository private, please give the teaching team access: @kuriwaki, @seannossek, @changwookju, @robinwyj
* Guidelines on tech resources:
  + Using Github via RStudio: Happy Git with R (<https://happygitwithr.com/>).
  + High-level guidance to Project organization, used in week 1 of our class: R4DS chapter on Workflow (<https://r4ds.hadley.nz/workflow-scripts#projects>).
  + Specific organization for quantitative social science: Gentzkow and Shapiro (2014) <https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf>

# Appendix

## Problem 1

Formatting Guidelines: The following are probably familiar to you by now. Please:

* Omit the estimates of control variables that are not themselves quantities of interest. You can drop them entirely from the regression table, and instead create a row in the table that explains the specification
* Omit statistics like AIC/BIC/Log Likelihood that we have not covered (and which are rarely analyzed).
* If there are multiple specifications, include minimal information about them so that the reader can quickly tell the difference between specifications.

Constructing the table and regressions:

* Recall the use of functions in PS-06 and PS-08, such as glue(), str\_subset(), and str\_c(), to construct long formula specifications
* As mentioned briefly in class 11.1, we specify the level of clustering by the cluster argument in feols.
* To understand how the structure of the stratification variables and the randomization units differ, it is helpful to look up some examples. How many households are in each stratification/randomization unit? What is the fraction of treated vs. control units in each grouping?
* Be careful not to use the IDs like geo\_cluster as numerical variables. In order to take dummy variables for each of the levels, they must be converted to a categorical variable so that, e.g., a value of 101 and 102 are treated as separate levels void of their numerical magnitude. Coercing the variable into a factor is one way to do this. Another approach (and probably the one that you will use here) is to use this as a fixed effect in the fixest notation.

1. You may use the same code from classmates, Ed Discussion Board, instructors, and generative AI. However, you must hand in your own unique written work and code in all cases. Any wholesale copy/paste of another’s work is plagiarism. In other words, you can work with your classmate(s), sitting side-by-side and going through the problem set question-by-question, or use generative AI to provide potential code for you, but you must each type your own answers and your own code. [↑](#footnote-ref-1)