PROBLEM SET 1

**Due on Tuesday, February 14, 2023**

I - INSTRUCTIONS

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

1. Download this Word document file into your computer
2. Insert all your answers into this Word document. Guidance [here](https://www.dropbox.com/s/ox9fhmbpvy2viw5/How%20to%20incorporate%20handwritten%20work%2C%20Stata%20output%2C%20and%20screenshot%20images.pdf?dl=0) on how to insert non-Word objects such as handwritten work or screenshot images in your answers.
3. **Once your document is complete, please save it as a PDF**. This is important to make sure all your work is preserved in the process of submission to Canvas.
4. Please submit an electronic copy of the PDF and your **replicable R script** to the Canvas assignment page.

II - IDENTIFICATION

1. Your information

|  |  |
| --- | --- |
| Your Last Name: |  |
| Your First Name: |  |

**(2**) Group Members (please list below the classmates you worked with on this problem set):

|  |
| --- |
|  |

1. Compliance with Harvard Kennedy School Academic Code[[1]](#footnote-1) (mark with an X below)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | **Yes** | **No** | | I certify that my work in this problem set complies with the Harvard Kennedy School Academic Code |  |  | |

For this problem set, we will be examining the methods used in the following paper:

# Fafchamps, Marcel, David McKenzie, Simon Quinn and Christopher Woodruff. 2014. “Microenterprise growth and the flypaper effect: Evidence from a randomized experiment in Ghana.” Journal of Development Economics, 106: 211-226

# Conceptual Questions (30 points + 8 extra points)

1. Read the paper. Clearly state the primary research question that the authors are trying to answer. What makes this an interesting question? (2 points)

|  |
| --- |
|  |

1. Explain the main finding of the paper (including what the ‘flypaper effect’ is) using non-technical jargon, as if you were writing a brief policy memo. Provide two versions:  
   1. In 3-5 sentences (without copying the article abstract or the text): (2 points)

|  |
| --- |
|  |

* 1. In a tweet, i.e. 280 characters or less: (1 point)

|  |
| --- |
|  |

1. Summarize the specific details of the treatment that participating businesses in this country underwent. (2 points)

|  |
| --- |
|  |

1. The authors note that their experimental design is very similar to an earlier study in Sri Lanka. What reasons do they give to conduct a separate study, rather than expecting the findings of that research to apply in this context? (2 points)

|  |
| --- |
|  |

1. The authors used a randomized control trial because they believed an observational analysis of similar policies would be insufficient. Imagine that another country implemented the investment program without randomizing treatment, and that you were trying to understand the effect of this program on food security in that country. What are two possible confounders (omitted variables) that would bias the results from your observational analysis? Explain the mechanism of the omitted variable and use the omitted variable bias formula to argue whether it would lead to an *understatement* or *overstatement* of the true effect. (3 points)

|  |
| --- |
|  |

1. Let be the business profits in the absence of the grants program and let *Di* = 1 denote participation in the program. If there were no RCT and individuals were allowed to opt-in to the program, critics might point out that: (3 points)

*E(|Di = 1) < E(| Di = 0)*

Explain this equation in words, explain why it is a problem, and give a plausible scenario in which that may be the case.

|  |
| --- |
|  |

1. Defining treatment as being assigned to the grants program, what is the difference between the ITT and TOT in this context? Which do the authors report and why do they make this decision? Write an equation (using potential outcomes notation) that shows what the authors are trying to estimate. (3 points)

|  |
| --- |
|  |

1. At what level do the authors clustered their standard errors of the main results of the paper (if at all)? Briefly note why the authors cluster the standard errors and why this is the appropriate level to cluster at. (2 extra points)

|  |
| --- |
|  |

1. To assess whether treatment was actually randomly assigned, we can examine the results of a balance test, presented in Table 2. Do the results in this table make you more or less confident about the validity of the paper’s results? Interpret one of the *p*-values from column (5). (3 points)

|  |
| --- |
|  |

1. Attrition in experiments like this one is often a concern for internal validity. Does the particular sort of attrition mentioned in this article give you reason to be concerned about the validity in this study? Describe using particular aspects of the experiment or its implementation. (1 point)

|  |
| --- |
|  |

1. What other threats to internal validity may have affected this this experiment? Choose one threat and explain how it might bias the coefficient of interest. (2 points)

|  |
| --- |
|  |

1. Describe four specific problems involved with generalizing the results of this study as a result of using an RCT. Hint: review the Muralidharan and Niehaus (2017) or Banerjee et al. (2017) papers discussed in class. (4 points)

|  |
| --- |
|  |

1. List at least two strategies the authors use to address some of the concerns you described above. (2 points)

|  |
| --- |
|  |

1. Why do the authors include Table 4? (2 extra points)

|  |
| --- |
|  |

1. Do you think the main results would be the same if this experiment were expanded to slightly larger businesses? Describe plausible scenarios in which providing the same transfer to these different businesses could both lead to (i) a larger increase in profits, and (ii) a smaller increase in profits. (2 extra points)

|  |
| --- |
|  |

1. If you were a researcher at the World Bank interested in scaling up one or several of these treatments, what follow-up study would you propose to expand on these findings? Explain in 4-5 sentences as if you were trying to convince a policymaker of the need for additional research. (2 extra points)

|  |
| --- |
|  |

# Data Analysis Questions (22 points + 2 extra points)

In this part, you will replicate the central results from the paper. You will be asked to analyze the data, present results, and submit a replicable R script separately.

In the problem set link, we have provided a lightly cleaned version of their main analysis files: ReplicationDataGhanaJDE.dta. Keep the file in a subfolder called data of your problem set Rstudio Project; this will facilitate our submission verification described at the end of this part. The data we are using comes from the public [study page](https://doi.org/10.48529/0gsj-b583) in the World Bank microdata library.

There are several new programming topics that this part will require:

1. Defining your own function
2. Fixed effects and clustered standard errors
3. String concatenation and extraction

i. To get started on writing your own function, please go through:

* RStudio [Primer for writing functions](https://rstudio.cloud/learn/primers/6.2), and
* An additional [screencast on functions](https://vimeo.com/channels/1534187) tailored to this problem

ii. And to familiarize with fixed effects and clustered standard errors, please make use of our screencasts walking through lfe::felm(), which is the function we recommend in the HKS [cheat-sheet](https://bit.ly/HKS-R).

* [Fixed effects with felm](https://vimeo.com/388825261)
* [Clustered standard errors with felm](https://vimeo.com/388825307)

Please also make use of the Appendix of this problem set for more details about the data.

1. Produce a well-organized descriptive statistics table that includes (i) the number of households, (ii) the number of geographic units, (iii) the number of units of randomization, alongside (iv) the sample mean and standard error of the *income and revenue index* in the control group, and (v) the same sample mean and standard error of the same index but in the treatment group. In other words, the table should have one row (countries), and six columns (including country).
   1. Print your table below. (6 points)

|  |
| --- |
|  |

* 1. Are the differences in baseline profits between the control and treatment groups significant at the 0.05 level? (3 points)

|  |
| --- |
|  |

1. Reproduce the coefficient estimate and standard error estimates in columns (1-2) of **Table 3:** these correspond to the main pooled OLS specifications (Equation 5 in the paper) without and with sample trimming respectively. (3 points)

|  |
| --- |
|  |

1. Represent the values of column (5), rows 3 onward in Table 3 as a well-labeled coefficient plot. This corresponds to the main (untrimmed) OLS analysis finding treatment effects by gender and treatment type (Equation 6 in the paper). The figure must: (6 points)  
   * Print the rounded coefficients and standard errors next to each point.
   * Define and use your own function for at least one part to avoid repetition.
   * Be estimated from the regression specification described in the paper. Therefore, all the numbers should match exactly as reported in Table 3.[[2]](#footnote-2)
   * Be clean and well-labeled (i.e., have clear axis labels, no chartjunk, understandable to a reader who has not read the paper).

|  |
| --- |
|  |

Try implementing the fixed effects specification either for question 18 or 19 above, and report your coefficients as a table or coefficient plot respectively. (2 extra points)

|  |
| --- |
|  |

1. Submit your R script to the Canvas assignment as a separate .R file (or .Rmd file, if you used Rmarkdown). We may pass submissions through a program to check if they run. To pre-test your code, you can verify it runs on our environment. Go to the [math camp](https://rstudio.cloud/spaces/18236/projects) space and copy the project API-210\_PS-01\_eval. The dataset is already uploaded in that project, so if you copy your R code and source/run everything after aligning the file paths, it should produce the correct figures and tables. (4 points)

# RCTs in Your Own Work (8 points)

1. Propose a specific policy question that could best be answered using an RCT. Explain the question in non-technical terms in no more than 3-5 sentences. Write out the empirical specification you would use. (4 points)

|  |
| --- |
|  |

1. Describe your treatment group. Propose a comparison group and explain why you chose that group. (2 points)

|  |
| --- |
|  |

1. Is it possible that the control group could be “contaminated” as a result of interacting with the treatment group? Explain which measures you would take to limit this contamination. (2 points)

|  |
| --- |
|  |

# 

# Appendix for Data Analysis

# Packages to install

Code for the data analysis portion will necessarily vary by student, and there are multiple packages available which could be used to construct solutions. However, the following packages have been tested by the teaching staff and are used in the sample solutions:

# library(haven)

# library(broom)

# library(lfe)

# library(tidyverse)

# library(forcats)

# library(glue)

# library(ggplot2)

# Documentation embedded in Stata dta files

The dataset is a Stata .dta file. Stata .dta files often encode the description of each variable in a separate attribute, which are often very useful to understand what each variable represents. When one reads in a .dta file via haven::read\_dta(), these variable attributes are stored as separate attributes associated with each variable. These can be checked by, for example, applying str() to each vector of a variable, but other packages provide convenience functions to more easily tabulate this data. We recommend the labelled::lookfor() function.

## Variables in Regression

Equation (5) in the paper defines the main regression to be run.

1. We abide by the Harvard Kennedy School Academic [code](https://www.hks.harvard.edu/educational-programs/academic-calendars-policies/student-handbook/general-regulations-and-1) for all aspects of the course. In terms of problem sets, unless explicitly written otherwise, the norms are the following: You are free (and encouraged) to discuss problem sets with your classmates. However, you must hand in your own unique written work and code in all cases. Any 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, but you must each type your own answers and your own code. For more details, please see syllabus. [↑](#footnote-ref-1)
2. Rounded to the third decimal place. [↑](#footnote-ref-2)