Problem Set 1

A randomized evaluation of microfinance

August 2023

General objectives and description

The goal of this problem set is go through the typical steps involved in the analysis of data from a randomized trial. The data is from the study by Banerjee et al. (2015) that we discussed in class.[1] That paper uses household-level data for a randomized evaluation of a group-lending microcredit program in neighborhoods (referred to as "areas" below) of Hyderabad, India. In this problem set you will first produce basic descriptive statistics of these data, and then replicate some of the study's main results. Whenever something is unclear about the variables used or other aspects of the analysis, try to consult the original paper. The data used in the paper are available on Moodle. The databases are:

- The data file `neighborhood.dta` contains neighborhood level baseline characteristics and the treatment assignment indicator variable.
- The data files `household_endline1.dta` and `household_endline2.dta` contain outcomes at the household level from the first and second endline surveys, respectively.
- The data file `household_baseline.dta` contains data at the household level, collected at baseline. (This file is included for completeness, but is not required for this problem set.)

To use these data files in R (if you do not use Stata) they will need to be imported. Consult the internet if you need help with that.

Neighborhood-level data

i) Load and familiarize yourself with the `neighborhood.dta`. Answer the following questions:

a) How many neighborhoods (observations) are there in the data set? How many are in the treatment group and how many in the control group?

b) In terms of *Number of households (baseline)*, how large are the smallest and biggest neighborhood?

[1] Banerjee, A.V., Duflo, E., Glennerster, R. and Kinnan, C., 2015. The Miracle of Microfinance? Evidence from a Randomized Evaluation. *American Economic Journal: Applied Economics*, 7(1): 22-53

c) Create a variable that measures the number of businesses per household in each neighborhood. Briefly summarize descriptive statistics for this variables in a suitable plot.

ii) Create a table showing the means of all variables named `area_*` and the variable that you generated in **i.c)** for two sub-samples: the control group and the treatment group. Add a column to the table showing the results of individual t-tests for whether each of the variables differs between the control and the treatment group. Report the p-values for all the tests. Give a concise interpretation of the results in the table. What can we learn from them? Note that these are all pre-treatment measurements, so we do not learn about treatment effects here.(< 100 words)

Hint

Potentially useful Stata commands: `estpost`, `ttest`. Potentially useful R packages and commands: `stargazer`, `xtable`, `t.test`.

Merge the `neighborhood` dataset with `household_endline1` and `household_endline2`. This merged dataset will be used to analyze the treatment effects described in Banerjee et al. (2015). First, you will replicate a subset of the results presented in Table 2 of the paper, studying the treatment effect on borrowing behavior of households. After that, a subset of the results presented in Table 6 are replicated, which are concerned with the treatment effect on expenditures.

Treatment effect: Access to microcredit

iii) Run 4 OLS regression, using the variables `spandana_1`, `anyloan_1`, `spandana_2`, and `anyloan_2` as dependent variables.[2] Use `treatment` and these six area-level control variables as independent variables in all regressions: `area_pop_base`, `area_literate_base`, `area_debt_total_base`, `area_business_total_base`, `area_exp_pc_mean_base`, `area_literate_head_base`. Cluster your standard errors at the area level and weight the regressions to account for oversampling of Spandana borrowers, i.e., use the weights `[pweight=w1]` for endline 1 and `[pweight=w2]` for endline 2 in Stata.[3]

a) Show your 4 estimation results in a single table. Restrict your table to show only output that is relevant to discuss the effects of microfinance. Describe and interpret your results. What is the effect of access to microcredit in treated areas? Compare the estimated effect size against the mean of the dependent variable in the control group. (< 150 words)

Hint

Potentially useful Stata commands: `estpost`. Potentially useful R packages and commands: `stargazer`, `xtable`.

[2] This roughly corresponds to columns 1 and 6 Table 2. [3] The authors oversampled Spandana borrowers, because they expected a larger variance in those respondents' outcomes.

Treatment effect: Consumption

The same data as before are used to study treatment effects on household spending.

iv) Run 6 OLS regressions using the variables `total_exp_mo_pc_1`, `durables_exp_mo_pc_1`, `temptation_exp_mo_pc_1`, `total_exp_mo_pc_2`, `durables_exp_mo_pc_2`, `temptation_exp_mo_pc_2`, as dependent variables.[4] As before, use the treatment dummy and the area-level controls as right-hand-side variables, cluster your standard errors at the area level and weight your regressions to account for the oversampling of Spandana borrowers.

a) Show your 6 estimation results in a single table. Restrict your table to show only output that is relevant to discuss the effects of microfinance. Describe and interpret your results. What is the effect of access to microcredit in treated areas? (< 150 words)

b) Reproduce the same table as in part **a)**, this time running the regressions without the area-level control variables. Do the results change qualitatively? Interpret your observations. (< 100 words)

v) Assume someone suggests that the missing observations in `temptation_exp_mo_pc_2` may be due to selective attrition and the results are only insignificant for this reason. To investigate this, estimate treatment effect bounds that account for attrition. A very simple approach (you are free to use another one) would be to:

- Create a "worst case" outcome variable that has the same values `temptation_exp_mo_pc_2` for all non-missing values. Then impute a very high value for missing outcomes in the treatment group—say the 75th percentile of the distribution of the variable in the sample—and impute a very low value for missing outcomes in the control group—say the 25th percentile of the distribution of the variable in the sample.
- Create a "best case" outcome variable that does the same in revers (low values for missings in the treatment group and high values for the missings in the control group.)
- Run the treatment effect regression once with the worst-case and once with the best-case scenario as outcomes.

Interpret your result (< 300 words).

Bonus questions (hard)

vi) What kind of spillovers can you imagine in the context of an RCT related to microfinance? Either think of the situation as in the paper, i.e., treatment happening at the neighborhood level, or imagine an alternative treatment in which access to microfinance is randomly allocated at the household level. Explain how the spillover in your example effects the outcome in the treatment and/or control group and how this affects the conclusions drawn from the RCT. (< 100 words)

[4] This roughly corresponds to columns 1 and 2 and 7 of Table 6.