Documentation and Data Codes for the paper:

Estimating the Demand for Business Training: Evidence from Jamaica by Alessandro Maffioli, David Mckenzie and Diego Ubfal

1. Introduction

In this document we provide information on the files used to generate the final tables and figures in the paper. We include comments on the datasets, the codes, and the Stata version, as well as replication instructions.

2. Dataset

All the data in this study was gathered by the research team through original data collection. There were two survey waves, denoted as “baseline" and “follow-up”. In the Stata datasets, variables are named after the question code in the corresponding questionnaire. In addition, we used auxiliary data, such as information collected during the willingness to pay elicitation in the field, payment and attendance data.

All the data collected has been included in the file **data_anonymized.dta,** which contains the raw data, the clean variables, and auxiliary variables used in the analysis. This dataset contains all the raw data, except for variables that would allow identifying individuals, which we dropped to preserve respondents' anonymity. Variable labels and variable names should allow users to understand the content of each variable.

To help understand the variable names, here is a list of common prefixes:

- q_ prefix: raw baseline variables (coinciding with numbers in Baseline Questionnaire – see baseline_survey.pdf)
- f_ prefix: follow-up variables (raw or clean, raw variables use same numbers as in Follow-up_survey.pdf);
- Question_ prefix: variables obtained in the BDM willingness to pay elicitation (See WTP_elicitation_BDMexample.pdf)
- Week_ prefix: attendance data
- win_ prefix: variable winsorized at the 99% percentile or both the 1% and the 99% percentiles;
- m_ prefix: variable with missing values replaced with 0's;
- d_ prefix: dummy variable indicating if the original (i.e., non-d) observation is missing.

3. Codes

3.1 Do files

The dofile **replication.do** contains the code to generate all the results (tables and figures) of the paper. All the variables necessary for the analysis are already contained in the **data_anonymized.dta** file mentioned in the Dataset section.

The codes used for the data preparation are available upon request. For each survey wave (baseline, follow-up), the dataset was generated in two steps: (1) we cleaned the raw data, and (2) we merged it with auxiliary data and generated the variables for the analysis. At the end, we merged all data together.

3.2 User-written commands

For the analysis, we used some user-written commands (kmatch, estout, coefplot, wyoung). They are all available on the SSC archive.

3.3 Stata version

To generate the tables, the codes were run with Stata 16 and 17. But they are compatible with Stata 14 and newer versions.

4. Replication Instructions

In order to replicate the analysis, follow these steps:

1. Download the data and the code and save it in a local directory.
2. In the replication.do file, change the content of the directory to match your directory
3. Make sure you install all the commands listed in the User-written commands section.
4. Run the code.