Documentation on data and codes for the paper: The Impact of Soft-Skills Training for Entrepreneurs in Jamaica

1 Introduction

In this document we provide information on the files used to produce the results for the study "The Impact of Soft-Skills Training for Entrepreneurs in Jamaica". We include comments on the questionnaires, the datasets, the codes, the user-written commands, and the Stata version, as well as replication instructions.

All the data in this study was gathered by the research team through original data collection. There were three survey waves, denoted as "baseline", "follow-up 1" and "follow-up 2".[1]

2 Questionnaires

There is one questionnaire per survey wave, which we are including here. The baseline data was collected using two different platforms, Qualtrics and SurveyMonkey, but the questionnaire was the same. Here is the list of files:

- **Baseline_questionnaire.pdf**
- **Followup1_questionnaire.pdf**
- **Followup2_questionnaire.pdf**

Some of these files contain additional instructions for the programmers, but they are all presented in a human-readable format.

3 Dataset

The baseline data was collected using two different platforms, Qualtrics and SurveyMonkey. 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 field operations, course evaluations, and data on randomization.

All data are included in the file **Jamaica_data_final.dta**. This file contains the raw data, the clean variables, and auxiliary variables used in the analysis.[2] To preserve respondents' anonymity, we dropped variables that would allow identifying individuals.

[1] For additional information on the data collection, please refer to the paper.

[2] This is an example of each type of variable:

- Raw data variables: `m2q13_a` and `m2q14_a`, which correspond to the same questions in the questionnaire, asking for full-time PAID employees and full-time UNPAID employees, respectively.
- Clean variables: `employees_fulltime`, the total number of employees, generated as `m2q13_a+m2q14_a`.
- Auxiliary variables: `m_employees_fulltime`, which is the same as `employees_fulltime`, but with missing values replaced with 0's; or `d_employees_fulltime`, which is a dummy variable taking value 1 if `employees_fulltime` is missing.

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 variable collected using Qualtrics;
- `m_` prefix: raw baseline variable collected using SurveyMonkey;
- `fu1_` prefix: follow-up 1 variable (raw or clean);
- `fu2_` prefix: follow-up 2 variable (raw or clean);
- `z_` prefix: variable transformed into a z-score;
- `win_` prefix: variable winsorized at the 99% percentile or both the 1% and the 99% percentiles;
- `win5_` prefix: variable winsorized at the 95% percentile or both the 5% and the 95% percentiles;
- `ihs_` prefix: variable transformed using Inverse Hyperbolic Sine;
- `m_` prefix: variable with missing values replaced with 0's;
- `d_` prefix: dummy variable indicating if the original (i.e. non-`d_`) variable is missing.

4 Codes

4.1 Do-files

The do-file **Jamaica_data_analysis.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 file **Jamaica_data_final.dta**, as mentioned in the Dataset section.

In addition, we include the do files used to generate the final variables in **Jamaica_data_final.dta**. These files should not be run separately, since we already compiled the resulting variables in the final dataset.

For each survey wave (baseline, follow-up 1, and follow-up 2), the final data was created in two steps: (1) we cleaned the raw data, and (2) we merged it with auxiliary data, and generated the variables for the analysis. Finally, we merged the three survey waves together. The do-file **Jamaica_data_preparation.do** is simply a master-file that calls the different data preparation do-files, according to the steps described. More specifically:

- **1_bas_data_cleaning.do**: code to clean baseline data (both Qualtrics and SurveyMonkey);
- **2_bas_data_creation.do**: code to merge Qualtrics data, SurveyMonkey data and auxiliary data, and create new variables;
- **3_fu1_data_cleaning.do**: code to clean follow-up 1 data;
- **4_fu1_data_creation.do**: code to merge follow-up 1 data with auxiliary data and create new variables;
- **5_fu2_data_cleaning.do**: code to clean follow-up 2 data;
- **6_fu2_data_creation.do**: code to merge follow-up 2 data with auxiliary data and create new variables;
- **7_data_merge.do**: code to merge data from baseline, follow-up 1 and follow-up 2.

4.2 User-written commands

For the analysis, we used several user-written commands. They are either available on the SSC archive or included in the folder with the do-files.

- `balancetable`: Command available on the SSC archive.
- `balancetable_custom`: Variation of the `balancetable` command, including a column with the test for equality of means across three arms (the original command allows for the comparison of only two arms), and the possibility to add headers (panels) in the output table. The ado-file is included.
- `estout`: Command available on the SSC archive (the package includes `eststo`, `esttab` and other commands).

- `mhtexp`: Command available on the SSC archive. It requires the `lmhtexp.mlib` library in the current directory (see help-file).
- `qreg2`: Command avaible on the SSC archive.

4.3 Stata version

All the codes are compatible with Stata 14.

Multiple hypothesis testing constitutes an exception, because the `mhtexp` command requires Stata 15. The do-files account for this: when they are run on a lower Stata version, the multiple hypothesis testing is skipped and p-values are left blank.

5 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. Set up your directory according to the suggested folder structure (see Folder structure in the Appendix).
3. In the **Jamaica_data_analysis.do** file, change the content of the `directory` global to match your directory. You may also need to adjust the other global macros (`data_fin`, `results_fig`, `results_tab`) if you did not follow the suggested folder structure.
4. Make sure you install all the commands listed in the User-written commands section.
5. Run the code.

Appendices

A Folder structure

The data was processed and analyzed with the folder structure presented below. We suggest to follow it when replicating the paper, even if this is not strictly necessary.

- *Your main directory*
  - Codes
    - balancetable_custom.ado
    - Jamaica_data_analysis.do
  - Data
    - Final data
      - Jamaica_data_final.dta
  - Results
    - Figures
    - Tables
  - Questionnaires
    - Baseline_questionnaire.pdf
    - Followup1_questionnaire.pdf
    - Followup2_questionnaire.pdf