README for "Improving Public Works Targeting for Short-Term and Medium-Term Impacts: Experimental Evidence from Cote d'Ivoire"

Overview

This replication package contains the data and code necessary to reproduce the results in the paper "Improving Public Works Targeting for Short-Term and Medium-Term Impacts: Experimental Evidence from Cote d'Ivoire". The code is written in Stata and R.

The process involves two main master files:

1. `Global_Master.do`: A Stata script that handles data preparation and generates the main tables and figures for the impact evaluation.

2. `MASTER_2_ML.R`: An R script that performs the machine learning analysis to assess impact heterogeneity and targeting improvements.

The replicator should expect the Stata code to run for approximately 11 hours and the R code for over 20 hours on a high-performance machine. Due to the computational intensity of the machine learning component, we recommend using a machine with multiple cores (the code is set to use 15 by default).

Data Availability and Provenance Statements

Statement about Rights

- ☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

- ☒ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.

Summary of Availability

- ☒ All data **are** publicly available.

  The data used in this study were collected via surveys conducted by the authors as part of the experiment. The data have been de-identified. All data will be made publicly available via the World Bank Microdata Library upon publication. Until then, the necessary data files are provided within this replication package.

Details on each Data Source

The data was collected through baseline, midline, and endline surveys with the study participants. The following table provides details on the primary data files.

Data Name	Data Files	Location	Provided	Citation

Data Name	Data Files	Location	Provided	Citation
Baseline Survey Data	`Baseline_i_hh_clean_var.dta`, `ENSETE_2013_17Septembre2014_nonmissing.dta`	`data/deidentified/`	Yes	See below
Midline Survey Data	`Midline_hh_clean.dta`, `Midline_i_clean_var_est.dta`	`data/deidentified/`	Yes	See beloe
Endline Survey Data	`Endline_i_hh_clean_est.dta`	`data/deidentified/`	Yes	See below

Data Publicly Available

All data included in this package is publicly available in the following links.

- Projet Emploi Jeune et Développement des Compétences (PEJEDEC) Impact Evaluation - Public Works Baseline Survey, 2013. Citation: Marianne Bertrand, Bruno Crépon, Alicia Marguerie, Patrick Premand. Côte d'Ivoire - Projet Emploi Jeune et Développement des Compétences (PEJEDEC) Public Works Impact Evaluation - Baseline Survey, 2013 (PEJEDEC PWP BL 2013). Ref: CIV_2013_PEJEDEC-PWP-BL_v01_M. Downloaded from https://microdata.worldbank.org/index.php/catalog/6774

- Projet Emploi Jeune et Développement des Compétences (PEJEDEC) Impact Evaluation - Public Works Midline Survey, 2013. Citations; Marianne Bertrand, Bruno Crépon, Alicia Marguerie, Patrick Premand. Côte d'Ivoire - Projet Emploi Jeune et Développement des Compétences (PEJEDEC) Public Works Impact Evaluation - Midline Survey, 2013 (PEJEDEC PWP ML 2013). Ref: CIV_2013_PEJEDEC-PWP-ML_v01_M. Downloaded from https://microdata.worldbank.org/index.php/catalog/6775

- Projet Emploi Jeune et Développement des Compétences (PEJEDEC) Public Works Impact Evaluation - Endline Survey, 2015. Citation: Source: Marianne Bertrand, Bruno Crépon, Alicia Marguerie, Patrick Premand. Côte d'Ivoire - Projet Emploi Jeune et Développement des Compétences (PEJEDEC) Public Works Impact Evaluation - Endline Survey, 2015 (PEJEDEC PWP EL 2015). Ref: CIV_2015_PEJEDEC-PWP-EL_v01_M. Downloaded from https://microdata.worldbank.org/index.php/catalog/6776

Dataset list

The replication package includes raw (de-identified), intermediate, and analysis-specific datasets.

Data File	Source	Notes	Provided
`data/deidentified/*.dta`	Authors' survey	De-identified raw data from baseline, midline, and endline surveys.	Yes
`data/intermediate/*.dta`	`MASTER_1_DATASETS.do`	Intermediate files created by merging survey rounds. Used as input for analysis.	Yes
`data/ML_reproduce/*.dta`	`MASTER_2_ML.R`	Output files from the Machine Learning analysis, containing predictions and estimation results.	Yes

Computational requirements

Software Requirements

- ☒ The replication package contains two master programs to run all analyses.
- **Stata version 15**
- **R version 3.6.4+**
- The R code relies on packages and functions from the following sources, which are handled within the scripts:
- MLInference
- Bob_Emploi_RCT
- Helper scripts `util.R`, `helper_functions.R`, and `methods_functions.R` manage packages and define necessary functions.

Controlled Randomness

- ☐ Random seed is set at line ____ of program _____
- ☒ No Pseudo random generator is used in the analysis described here.

Memory, Runtime, Storage Requirements

Summary

Approximate time needed to reproduce the analyses on a high-performance machine (e.g., Apple M2 Max with 96GB RAM):

- ☐ <10 minutes
- ☐ 10-60 minutes
- ☐ 1-2 hours
- ☐ 2-8 hours
- ☐ 8-24 hours
- ☒ 1-3 days
- ☐ 3-14 days
- ☐ > 14 days

Details:

- **Stata analysis (`Global_Master.do`)**: ~11 hours.
- **R analysis (`MASTER_2_ML.R`)**: 20+ hours.
- **WARNING**: The ML code is computationally intensive. We strongly recommend using a computer with multiple cores (default is set to 15). If this is not feasible, intermediate data from the ML process can be provided upon request to allow for the reproduction of the final tables and figures (Parts 3-5 of the R script).

Approximate storage space needed:

- ☐ < 25 MBytes
- ☐ 25 MB - 250 MB
- ☒ 250 MB - 2 GB
- ☐ 2 GB - 25 GB
- ☐ 25 GB - 250 GB
- ☐ > 250 GB

Details

The code was last run on an **Apple M2 Max with 96GB of RAM**. The R code is configured to take advantage of multi-core processing to reduce runtime.

Description of programs/code

The code is organized into two main workflows, controlled by master files.

- **`Global_Master.do`**: This is the main Stata master file. It calls other Stata scripts to perform data preparation and generate all tables and figures related to the primary impact evaluation.
- `MASTER_1_DATASETS.do`: Prepares and merges datasets.
- `MASTER_3_TABLES_AND_FIGURES.do`: Generates final exhibits.
- **`MASTER_2_ML.R`**: This is the main R master file for the machine learning analysis. It should be run after the initial data preparation in Stata. It produces the heterogeneity analysis, predicted impacts, and targeting simulations.

License for Code

The code is licensed under a MIT license. See LICENSE.txt for details.

Instructions to Replicators

To reproduce all results, follow these steps in sequence:

1. Set up Directory Paths

- In `Global_Master.do` (Stata), update the global macro for your main directory path.
- In `MASTER_2_ML.R` (R), update the directory paths on lines 53 and 57 to point to your working directory and the data folder.

2. Run Stata Data Preparation

- Open `Global_Master.do` in Stata.
- Run the first part of the script which calls `MASTER_1_DATASETS.do`. This will create the intermediate datasets that are required for the R script.

3. Run R Machine Learning Analysis

This step uses the `MASTER_2_ML.R` script.

WARNING: The code is quite demanding in computer power and will probably take weeks to run on a normal PC. We strongly recommend using computers with multiple cores (default set for 15) to run the proxy predictions. In case this is not possible, we can provide the intermediate outcomes, so parts 3-5 can be reproduced.

ML Folder Structure

Ensure your working directory has the following structure. The scripts will create the `results` directory and its sub-folders if they don't exist.

```
├──codes
├──Auxiliary
| ├── util.R
| ├── helper_functions.R
| └── methods_functions.R
└──Main
└── MASTER_2_ML.R
└──data
└──results
└──experiment_name
└──outcome
└──winsor_num
└──treatment_arm
├──B_and_S
├──settings
├──stata
└──samples
```

ML Input Data

The following datasets, generated in Step 2, are required in your `data` folder:

- `Baseline_Mid_End_200522.dta`
- `other_outcome_data.dta`

Reproducing ML Results

1. Open `MASTER_2_ML.R` file located in the `codes/Main` folder.
2. Set the workflow you will use with Dropbox on line 51 (Through the API (set `dropbox_API=TRUE`) or local folder (set `dropbox_API=FALSE`)).

3. Confirm the working directory and data/output directory paths are set correctly (lines 53 and 57).

4. Run the entire script. This will generate the necessary `.dta` files with ML predictions in the `data/ML_reproduce` folder.

5. **Specify the set of parameters:** Upon running the file, you need to enter a set of estimation parameters from the 18 settings outlined in the table below to execute via a command line prompt. You need to either specify one set of parameters at a time, or to reproduce all results, you must enter 0.

ML Prediction Settings

For example, choosing `chosen_setting = 1` will produce predictions for `y0_m` (log monthly earnings at midline) under the specification detailed in the first row.

Setting #	Target output	treatment arm	folder name	winsorization quantile	Table in paper that uses the predictions	Set of covariates
1.	y0_m	c(1,2,3)	ALL	97	Table 6, 7, 8, 9, 10, A6; Figure 2, 3, B4, B5	SMALL
2.	y0_e	c(1,2,3)	ALL	97	Table 6, 10, A6; Figure 2, 3, B4	SMALL
3.	y1_m	c(1,2,3)	ALL	97	Table 6, 10, A6, A10; Figure 4, B4, B5, B6	SMALL
4.	y1_e	c(1,2,3)	ALL	97	Table 6, 10, A6; Figure B4, B6	SMALL
5.	s0_m	c(1,2,3)	ALL	97	Table A11, Table A12	SMALL
6.	s0_e	c(1,2,3)	ALL	97	Table A11	SMALL
7.	s1_m	c(1,2,3)	ALL	97	Table A11	SMALL
8.	s1_e	c(1,2,3)	ALL	97	Table A11	SMALL
9.	y0_e	c(1)	WET	97	Table A8	SMALL
10.	y0_e	c(2)	SET	97	Table A8	SMALL
11.	y0_e	c(3)	PW	97	Table A8	SMALL
12.	y1_e	c(1)	WET	97	Table A9	SMALL
13.	y1_e	c(2)	SET	97	Table A9	SMALL
14.	y1_e	c(3)	PW	97	Table A9	SMALL
15.	y0_m	c(1,2,3)	ALL	97	Table A7	LARGE
16.	y0_e	c(1,2,3)	ALL	97	Table A7	LARGE

Setting #	Target output	treatment arm	folder name	winsorization quantile	Table in paper that uses the predictions	Set of covariates
17.	y1_m	c(1,2,3)	ALL	97	Table A7	LARGE
18.	y1_e	c(1,2,3)	ALL	97	Table A7	LARGE

Optional ML Changes

- You can change the number of cores used for proxy predictions (line 474) and for Chernozhukov's estimators (line 567) to match your machine's capabilities.

4. Generate Final Tables and Figures

- Once the R script has been run for all necessary settings and the prediction files are generated, run the final part of `Global_Master.do` (which calls `MASTER_3_TABLES_AND_FIGURES.do`). This script uses the outputs from both the Stata and R analyses to generate all the final tables and figures in the `Output` folder.

List of tables and programs

- ☒ All tables and figures in the paper are reproduced by the code.

Figure/Table #	Program	Output File	Note
Table 1: Baseline Summary Statistics and Balance Checks	`MASTER_3_TABLES_AND_FIGURES.do`	`Tables/Table_1_balance_weighted`	-
Table 2: Characteristics of Applicants and National Population of Urban Youths	`MASTER_3_TABLES_AND_FIGURES.do`	`Tables/Table_2_ENSETE_comp`	-
Table 3: Impacts during and post program, economic outcomes	`MASTER_3_TABLES_AND_FIGURES.do`	`Tables/Table_3_main_itt`	-
Table 4: Impacts during and post program, alternative definitions of earnings and savings outcomes	`MASTER_3_TABLES_AND_FIGURES.do`	`Tables/Table_4_rob_itt`	-
Table 5: Impacts during and post program, well-being, behaviors and work habits	`MASTER_3_TABLES_AND_FIGURES.do`	`Tables/Table_5_main_itt_behavior_wellbeing`	-
Table 6: Impacts post program on skills, investments in self-employed activities and search for wage jobs	`MASTER_3_TABLES_AND_FIGURES.do`	`Tables/Table_A5_itt_bhv_skills`	-

Figure/Table #	Program	Output File	Note
Table 7: Heterogeneity in impacts on earnings during and post program, machine learning results	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_6_ml_v3	Requires ML script output
Table 8: Baseline characteristics of the bottom and top quartiles of the distribution of predicted impacts on (ln) earnings during program	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_7_ml_charac_v3	Requires ML script output
Table 9: Impacts during and post program on main outcomes, by quartile of predicted impacts on (ln) earnings during the program	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_8_during_post_v3	Requires ML script output
Table 10: Impacts post program on intermediate outcomes, by quartile of predicted impacts on (ln) earnings during the program	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_9_ml_post_bis_v3	Requires ML script output
Table 11: Impacts on (ln) earnings and cost-benefit ratios under alternative targeting approaches	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_10_target_MLv3	Requires ML script output
Figure 1: Quantile treatment effects for (ln) earnings during and post program	MASTER_3_TABLES_AND_FIGURES.do	Figures/figure_1_qte_all.png	-
Figure 2: Quantile treatment effects for (ln) earnings post program, by treatment arm	MASTER_3_TABLES_AND_FIGURES.do	Figures/figure_B3_qte_Tarm.png	
Figure 3: Group average treatment effects (GATES) for (ln) earnings	MASTER_3_TABLES_AND_FIGURES.do	Figures/figure_2_gates_MLv3.png	Requires ML script output
Figure 4: Predicted	MASTER_3_TABLES_AND	Figures/figure_3_	Requires ML

Figure/Table #	Program	Output File	Note
impact on (ln) earnings during vs post program	_FIGURES.do	scatter_cate_v3.png	script output
Figure 5: CATE at Midline by predicted BL earnings	MASTER_3_TABLES_AND_FIGURES.do	Figures/graph_predbase_vs_predcate_mid.png	-
Table A1: Overview of Public Works Programs	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A1_programs	-
Table A2: Estimated impacts during and post program on economic outcomes, with baseline controls	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A1_bcontrol	-
Table A3: Estimated impacts during and post program on economic outcomes, without clustering standard errors	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A2a_noCluster	-
Table A4: Impacts during and post program, with full baseline sample at follow-up	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A2b_itt_140kept	-
Table A5: Impacts during and post program, LATE accounting for compliance in control between midline and endline	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A5	-
Table A6: Estimated impacts during program on economic outcomes for household members other than the beneficiary	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_hh_i_itt	-
Table A7: Estimated impacts during and post program on well-being index components	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A2_itt_wb_comp	-
Table A8: Estimated impacts during and post program on behavior index components	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A3_itt_bhv_comp	-

Figure/Table #	Program	Output File	Note
Table A9: Estimated impacts during and post program on risky behaviors	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A4_list_linear	-
Table A10: Estimated impacts post program on aspirations and reservation wage	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_aspiration_itt	-
Table A11: Comparison of Machine Learning algorithms to predict impacts on earnings during and post program	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A6_MLmethods_v3	Requires ML script output
Table A12: Heterogeneity in impacts on earnings during and post program, machine learning results for an extended set of covariates	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A7_ml_extcov_v3	Requires ML script output
Table A13: Estimated impacts on (ln) earnings post program, by treatment arms	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A8_ml_tarms_v3	Requires ML script output
Table A14: Estimated impacts on earnings (in levels) post program, by treatment arms	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A9_ml_tarms_lvl_v3	Requires ML script output
Table A15: Baseline characteristics of the bottom and top quartiles of predicted impacts on earnings (in levels) during program	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A10_ml_charac_lvl_v3	Requires ML script output
Table A16: Estimated impacts during and post program on savings	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A11_ml_svg_v3	Requires ML script output
Table A17: Baseline characteristics of the bottom and top quartiles of predicted impacts on (ln) savings during	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A12_ml_charac_svg_v3	Requires ML script output

Figure/Table #	Program	Output File	Note
program			
Figure B1: Cumulative distribution of earnings during and post program	MASTER_3_TABLES_AND_FIGURES.do	Figures/graph_cum_dist_income.png	-
Figure B2: Reported baseline earnings vs Predicted impacts during and post program	MASTER_3_TABLES_AND_FIGURES.do	Figures/graph_base_vs_pred.png	-
Figure B3: Cost-effectiveness ratios over time under alternative targeting rules, depending on the sustainability of post-program impacts	MASTER_3_TABLES_AND_FIGURES.do	Figures/figure_B5_CE_v3.png	Requires ML script output
Table D1: Baseline variables used in Machine Learning algorithms	MASTER_3_TABLES_AND_FIGURES.do	Tables/Table_A13_mlcov	-
Figure D1: Relation between predictions and actual earnings (in level) (Random forest)	MASTER_3_TABLES_AND_FIGURES.do	Figures/figure_B6_act_vs_pred_rf_lvl.pdf	Requires ML script output
Table G1: Summary of weights used with midline data	n/a	(Directly in LaTeX code)	Not produced by scripts
Table G2: Summary of weights used with endline data	n/a	(Directly in LaTeX code)	Not produced by scripts