README: Replication Files for "Improving Precision through Design and Analysis in Experiments with Non-Compliance"

Erin Hartman Melody Huang

Overview

The code in this replication package replicates the analysis for Hartman and Huang, 2023+ "Improving Precision through Design and Analysis in Experiments with Non-Compliance." The replication files are split into two parts: (1) simulation, and (2) an empirical application, which re-analyzes a set of Get-Out-the-Vote experiments.

Instructions to Replicators

The replication files are split into two folders: `Simulation` and `GOTV`, with corresponding helper functions and log files located within each folder. Figures are outputted to the folder `Figures`.

Simulation. To run the simulations, replicators should start within the `Simulation` folder, and run `01_run_simulations.R`. This will run all the simulations discussed in the paper and export a set of `.Rdata` files to the folder `Simulation Results`. The results are analyzed in the subsequent files: `02_simulation_analysis.R`, which generates all the results in the main manuscript, and `03_simulation_appendix.R`, which generates all the results in the Appendix. Previously exported results from running `01_run_simulations.R` are provided in the subfolder `Simulation Results`. Thus, `02_simulation_analysis.R` and `03_simulation_appendix.R` may be run directly using the data files from `Simulation Results`. See section on Computational Requirements for details on running the simulations.

Empirical Application (i.e., GOTV). To run the empirical application, replicators should start within the `GOTV` folder, and start with `01_load_data.R`. This will call in the original experimental data and generate a cleaned data frame, stored in `Data/generated/gotv_cleaned.Rdata`. Then, to generate the matched data, replicators must run `02_load_matches.R`. The rest of the files will call in `Data/generated/gotv_cleaned.Rdata`, as well as `Data/results_matching.Rdata` and should be run sequentially:

- `03_empirical_app_manuscript.R`
- `04_empirical_app_appendix.R`
- `05_variable_importance.R`
- `06_model_performance.R`

`03_empirical_app_manuscript.R` generates all results that correspond to the empirical application in the main manuscript, while the other 3 files will generate results that appear in the Appendix. `04_empirical_app_appendix.R` *must* be run after running `03_empirical_app_manuscript.R`, while

05_variable_importance.R and 06_model_performance.R can be run standalone, as long as the file gotv_cleaned.Rdata has already been generated.

The generated log files that correspond to each of the R files are also provided in the subfolder log for reference. Log files were generated using the R package logr.

Data Availability and Provenance Statements

All data used in the analysis are publicly available. In particular, the analysis relies on a set of multi-site Get-Out-the-Vote experiments conducted in Gerber, and Nickerson (2003). The original experimental data is available from Yale's Institute for Social and Policy Studies (ISPS) (https://isps.yale.edu/research/data/d017). The raw data files are provided in GOTV/Data, with the following file name: GreenGerberNickerson_JP_2003-1_EDITED.csv. The file was downloaded on May 15, 2020.

Computational requirements

The simulations were run on a **Linux 64-bit server, running Ubuntu 20.04.6 LTS**, using R version 4.0.3 (2020-10-10) and 48 cores. The total run time for the simulations takes about 14 hours. The simulation analysis and empirical application were run on an **Intel-based laptop with macOS Big Sur 10.16**, running R Version 4.0.5 (2021-03-31). The total run time for all the analysis files should take under 10 minutes.

Description of programs/code

A brief description is provided below.

Simulations

- 01_run_simulation.R: This R file will run all the simulations. The file will call on a set of helper functions, found in the folder Code/helper_functions.R.
  - The script will generate the following files, located in the subfolder Simulation Results:
    - sims_results_all.Rdata: Results corresponding to the basic simulations
    - sims_er_violation.Rdata: Results corresponding to simulations related to violations in the exclusion restriction
    - sims_comp_prob.Rdata: Results corresponding to simulations related to varying the compliance probability
    - sims_pi_violation.Rdata: Results corresponding to simulations related to violations in principal ignorability

  Note: To replicate the results exactly, because a parallelization procedure was implemented to run the simulations, the simulations must be run on a 48 core machine. Currently, the code is set to detect total cores available and parallelize across the maximum number of cores. Thus, if the machine used to replicate the simulations has over 48 cores, to replicate exactly, replicators should set the number of cores to be equal to 48. If the number of cores available is less than 48, the results may vary slightly due to differences in the random number generation. The outputted .Rdata files used in a previous run of the simulation, which was used for the production of the results in the manuscript are provided for reference.

- `02_simulation_analysis.R`: This R file will run all code corresponding to analysis of the simulation results within the main text. The R file will call in the 4 `.Rdata` files generated by `01_run_simulation.R`.
  - This file outputs three figures, located in the folder `Figures`:
    - Figure 2 - `Figures/fig2_mse_breakdown_pi.pdf`
    - Figure 3 - `Figures/fig3_standard_error_compl.pdf`
  - This file also generates results for Table 1. Table output is generated using `xtable` within `R` and copied into the main manuscript.
- `03_simulation_analysis.R`: This R file will generate the output for all results in the Appendix related to the simulations. The file will call in the same 4 `R` files from `Simulation Results`, generated by `01_run_simulation.R` to generate the results.
  - This file outputs the following figure, located in the folder `Figures`:
    - Figure A-3.2 - `Figures/figA3_2_mse_breakdown_er.pdf`
  - This file also generates results for Table A-3.2 and A-3.2 using `xtable`.

GOTV

- `01_load_data.R`: This R file call in the raw data from the original GOTV experiments, located in `Data/raw/GreenGerberNickerson_JP_2003-1_EDITED.csv` and output a cleaned data file: `Data/generated/gotv_cleaned.Rdata`. The cleaned data file used in the analysis is provided for reference. Replicators may bypass this step and directly use the cleaned data file `gotv_cleaned.Rdata` if they wish.
- `02_load_matches.R`: This file will generate the matched data used in the analysis. The file will call on a set of helper functions, found at `Code/helper_functions.R`. Furthermore, the file will call on an R script found at `Code/run_matching.R` to run the matching procedure. The matching procedure in `Code/run_matching.R` will also generate a data file that contains the results, located in `Data/generated/results_matching.Rdata`. Replicators may bypass this step and directly use the provided `results_matching.Rdata` instead.

  Note: To replicate the matching results *exactly*, the same version of the Matching package must be used. All versions of the packages used to generate the results are provided in the corresponding log files. Using a different version of the Matching package will result in very similar results, though the specific estimates may differ slightly.
- `03_empirical_app_manuscript.R`: This R file will generate all the results corresponding to the empirical application in the main manuscript.
  - This file outputs one figure, located in the folder `Figures`:
    - Figure 4 - `Figures/fig4_gotv_estimates.pdf`
  - This file also generates results for Table 2. Table output is generated using `xtable` within `R` and copied into the main manuscript.
- `04_empirical_app_appendix.R`: This R file generates some supplemental results for the empirical application that appear in the Appendix. These results include illustrating the sensitivity analysis (Section A-2.2) and additional tables found in Section A-4. This file must be run after running `02_empirical_app_manuscript.R`, as it relies on objects created in the previous file.

  - This file outputs one figure, located in the folder `Figures`:
    * Figure A-2.1 - `Figures/figA2_1_contour.pdf`
  - This file also generates results for Table A-4.3 and Table A-4.5.

- `05_variable_importance.R`: This R file generates the results for the Appendix section related to variable selection (A-4.1).
  - This file outputs two figures used for Figure A-4.3, located in the folder `Figures`:
    * Figure A-4.3 (top) - `Figures/figA4_3_1_varImp_outcome.pdf`
    * Figure A-4.3 (bottom) - `Figures/figA4_3_2_varImp_compliance.pdf`

- `06_model_performance.R`: This R file generates the results for the Appendix section related to predicting compliance scores (A-4.2).
  - This file outputs the results for Table A-4.4.

Contact

Erin Hartman (ekhartman@berkeley.edu) and Melody Huang (melodyyhuang@berkeley.edu)