

Readme

Andreas Olden and Jarle Møen

14 11 2021

Overview

The code in this replication package reproduces the rejection rates simulations of our paper on the triple difference estimator (Olden & Møen, 2021). It takes raw data from the Population Survey in their fourth interview month, in the Merged Outgoing Rotation Group, from 1979 to 1999, merge it, and simulate treatment effects of varying sizes and differing number of treatment clusters as described in appendix B. This produces all results and table of our paper. This can take several days on a standard desktop.

To reproduce our results, open and run the following scripts, located in the folder `r_scripts`, in sequential order as administrator:

- `simulate_triple_difference.Rproj` (In main folder)
- `0_merge.rmd`
- `1_wrangle.rmd`
- `2_simulate.R`
- `3_results.rmd`

Make especially sure that you always open the Rproj file first to ensure that the relative file paths work. Since running the code can take several days, we have for your convenience included intermediate data and results (in folder `r_scripts`) from all steps of the procedure:

- `0_merge.rmd` produces `df_merged.rds` (needs to be unzipped, included as `df_merged.7z` in `r_scripts/data/cleaned_data`)
- `1_wrangle.rmd` produces `df_strip.rds` (needs to be unzipped, included as `df_strip.7z` in `r_scripts/data/cleaned_data`)
- `2.simulate.R` creates 4 files, also included in the folder `r_scripts/sim_res`
- The raw data is available in `r_scripts//data/raw_data`.

To ensure reproducibility of the packages/environment we use the package ‘renv’. Documentation is available at: <https://cran.r-project.org/web/packages/renv/vignettes/renv.html> and <https://rstudio.github.io/renv/articles/renv.html>

Note:

- The first time you run `renv` it might take some time and you need to give `renv` certain permissions
- Each script/rmd-file calls `renv` to ensure reproducibility also if you only reproduce parts of the process.
- DoRNG is used to ensure reproducibility with parallel computing

Reproducibility has been checked across Windows platforms. We have also taken steps to ensure across platform reproducibility, but due to different backends under parallelization between windows and linux/ios user might still encounter issues. Please contact the authors at andreasolden@gmail.com if you face such issues.

Trouble shooting

Running this code might require admin privileges due to parallelization and renv. If renv does not bootstrap correctly, please install it independent of this replication to ensure all permissions are granted in the installation process. If this still does not work, try running Rstudio as an administrator.

Folder structure

- simulate_triple_difference is the main folder
- r_scripts contains a folder 'data' and a folder 'sim_res' and R/RMD files for analysis.
- The data folder contains two folders, namely 'cleaned_data' and 'raw_data'
- The sim_res folder contains raw simulations results

Data Availability and Provenance Statements

The data is the Population Survey in their fourth interview month, in the Merged Outgoing Rotation Group, from 1979 to 1999. The data can be accessed, with descriptions, from <https://www.nber.org/research/data/current-population-survey-cps-data-nber>. The data was downloaded 19 November 2020 from: <https://data.nber.org/morg/annual/>. A copy of the raw data is provided as part of this archive in r_scripts/data/raw_data

Datafiles: morg79.dta-morg99.dta

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.
- ☐ Some data **cannot be made** publicly available.
- ☐ **No data can be made** publicly available.

Controlled Randomness

- [registerDoRNG(seed = 2211)] Random seed is set at line 26 of 2_simulate.R
- [registerDoRNG(seed = 2211)] Random seed is set at line 195 of 2_simulate.R
- [registerDoRNG(seed = 2211)] Random seed is set at line 364 of 2_simulate.R
- [registerDoRNG(seed = 2211)] Random seed is set at line 533 of 2_simulate.R

Memory and Runtime Requirements

Approximate time needed to reproduce the analyses on a standard (CURRENT YEAR) desktop machine:

- ☐ <10 minutes
- ☐ 10-60 minutes
- ☐ 1-8 hours

- ☐ 8-24 hours
- ☒ 1-3 days
- ☐ 3-14 days
- ☐ > 14 days
- ☐ Not feasible to run on a desktop machine, as described below.

Details The simulations ran for about 4 hours using 24 cores (max by default in the code) on:

Microsoft Windows Server 2016 Standard Intel(R) Xeon(R) CPU E5-2660 v2 @ 2.60GHz, 64 RAM R version 4.0.3 (2020-10-10) – “Bunny-Wunnies Freak Out” x86_64-w64-mingw32/x64 (64-bit)

Rstudio Version 1.3.1093 “Apricot Nasturtium” (aee44535, 2020-09-17) for Windows

```
sessionInfo() R version 4.0.3 (2020-10-10) Platform: x86_64-w64-mingw32/x64 (64-bit) Running
under: Windows 10 x64 (build 19043)
```

Matrix products: default

```
locale: [1] LC_COLLATE=Norwegian Bokmål_Norway.1252 LC_CTYPE=Norwegian Bokmål_Norway.1252
[3] LC_MONETARY=Norwegian Bokmål_Norway.1252 LC_NUMERIC=C
[5] LC_TIME=Norwegian Bokmål_Norway.1252
```

attached base packages: [1] parallel stats graphics grDevices datasets utils methods base

```
other attached packages: [1] kableExtra_1.3.4 fixest_0.8.4 doRNG_1.8.2 rngtools_1.5 doParallel_1.0.16
iterators_1.0.13 [7] foreach_1.5.1 haven_2.4.1 here_1.0.1 forcats_0.5.1 stringr_1.4.0 dplyr_1.0.5
[13] purrr_0.3.4 readr_1.4.0 tidyr_1.1.3 tibble_3.1.1 ggplot2_3.3.3 tidyverse_1.3.1
```

Description of programs/code

To reproduce our results, open and run the following scripts (in folder `r_scripts`), in sequential order:

- ‘`simulate_triple_difference.Rproj`’ is the project file and ensures a consistent path structure
- ‘`0_merge.rmd`’ takes the original `morgXX.dta` files and combines them into a single file. It reads data from `r_scripts/data/raw` and saves it to `r_scripts/data/cleaned_data/df_merged.rds`
- ‘`1_wrangle.rmd`’ recodes variables. It reads the `r_scripts/data/cleaned_data/df_merged.rds` transforms the data to a more stripped and simulation friendly format and saves it to `df_strip.rds` at the same location.
- ‘`2_simulate.R`’ runs the simulations by first loading the `df_strip.rds` data frame.
- ‘`3_results.rmd`’ creates the tables of the paper in both html format that can be viewed in the rmd file itself and raw latex code.

All rmd files have accompanying html files to be viewed in any browser.

List of tables and programs

The provided code reproduces:

- ☒ All numbers provided in text in the paper
- ☐ All tables and figures in the paper
- ☒ Selected tables and figures in the paper, as explained and justified below.

It contains all numeric tables. The remaining tables are bibliographies and lists over articles referencing the triple difference estimator.

- Table 4 of the paper is reproduced in 3_results.rmd at line 38
- Table 5 of the paper is reproduced in 3_results.rmd at line 76
- Table 6 of the paper is reproduced in 3_results.rmd at line 115

References

National Bureau of Economic Research (NBER), n.d. “Current Population Survey Merged Outgoing Rotation Groups repository 1979-1999”. Accessed 19 November 2020. <https://data.nber.org/morg/annual/>