README and Guidance

This README guides the interested reader through the available material and a route to replicating the empirical results in the research paper *Quantile-based Test for Heterogeneous Treatment Effects* by EunYi Chung and Mauricio Olivares.

Overview

The code in this replication package constructs the empirical analysis from one data source using Stata and R. The data source corresponds to MDRC's Jobs First (Bloom, 2022) as in Bitler, Gelbach, and Hoynes (2006, 2017). The replicator should expect the empirical application code to run for about 11 hours under the conditions specified below.

Data Availability and Provenance Statements

☐ This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

This paper uses MDRC's Jobs First data sets, housed at the Inter-University Consortium for Political and Social Research (ICPSR). Several files in this data collection have special restrictions to protect respondent privacy. Thus, these data files are restricted from general dissemination but may be obtained with Restricted Data Use Agreements with the MDRC/ICPSR. Permissions as per `www.mdrc.org/public-use-data-files` have been obtained. Researchers interested in obtaining access to the data employed in this paper are required to submit a written application to gain approval and must agree to the terms and conditions of a Restricted Data Use Agreement.

We initiated the data access requirement on October 12th, 2018. The confidentiality agreements were received and signed on October 31st, 2018, and the raw data files were obtained in June 2019. The restricted files were distributed by MDRC in CD format, and the raw data sets were recorded in SAS format. To honor our commitment to protect data confidentiality, we will use the generic placeholder `rawdata.SAS` whenever we make references to these files.

Statement about Rights

☒ We certify that the authors of the manuscript have legitimate access to and permission to use the data used in this manuscript.
☐ We certify that the authors 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.

Details on Raw Data Source

Data.Name	Data.Files	Location	Provided
"Jobs First"	rawdata.SAS	data/acquired	FALSE

where `rawdata.SAS` corresponds to the data sets as provided by the MDRC (not provided).

Computational requirements

Software Requirements

- Stata (code has been run with versions 15 and 18)
- R (code was last run with version R 4.3.3)
  - `rprojroot` (2.0.4)
  - `quantreg` (5.98)
  - `readstata13` (0.10.1)
  - `dplyr` (1.1.4)
  - `mutiwayvcov` (1.2.3)
  - `RATest` (0.1.10)
  - the script `programs/global_libraries.r` will install all dependencies (latest version).

Controlled Randomness

☐ Random seed is set at line _____ of program _____

Memory and Runtime Requirements

Summary Approximate time needed to reproduce the empirical analyses in Section 7 and Supplementary Appendix V on a standard MacBook Pro 2021 machine:

☐ <10 minutes
☐ 10-60 minutes
☐ 1-2 hours
☐ 2-8 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 code that generates the main results in the empirical application (Tables 3 in the main text) was last run on an **Apple M1 Pro laptop with MacOS version 14.5** on July 7th 2024. The R code was last tested on an R terminal on **Visual Studio Code version 1.91.1** (Universal), and the Stata .do files were last tested on Stata 18. Both of these tests were conducted on August 8th 2024.

Description of programs/code

- The program `programs/01_processdata.do` loads Jobs First data set (restricted use) and performs some data manipulations. For example, it labels data, creates new variables by interacting covariate levels, reshapes the data set, etc.
- The program `programs/02_processdata.do` loads data processed by `programs/01_processdata.do` and performs some additional data carpentry.
- The program `programs/00_processdata.do` sets up the local directories and runs programs `programs/01_processdata.do` and `programs/02_processdata.do`. After completion, it delivers the final data set we will use in the empirical application.
- The program `programs/global_libraries.r` installs the R packages needed to replicate the empirical analysis.
- The program `programs/PT.qp.r` calculates the p-value based on the permutation test proposed by the main text.
- The program `programs/report.pvalues.r` adjusts the p-value to account for multiple testing using Bonferroni and Holm methods.
- The program `analysis/main.r` will execute the entire data analysis and export the results needed for the empirical analyses in the main text, i.e., Section 7 and Supplementary Appendix V.

License for Code

Instructions to Replicators

- Request the restricted-use data file referenced above (`rawdata.SAS`). Upon acquisition, convert this data file to Stata format, e.g. using Stat/Transfer. Call this file `rawdata.dta` and store it in `data/acquired`.
- Modify lines 27, 31, and 32 of `programs/00_processdata.do` to reflect the user-specific working directories in local machine. No further action is needed on the replicator's part.
- Run `programs/00_processdata.do`. This is the Stata do file that creates and prepares the final data from the original source.
- Run `analysis/main.r`. This is the R file that calculates the p-values as reported in the main text.

Details

- `programs/00_processdata.do`.
  - This script requires restricted-use data. See details above.
  - Before running the script, the replicator must change the working directory to reflect the directories in their local machine. These changes take place in lines 27, 31, and 32. It is important to recognize that different operating systems call for different path conventions. Thus, the user might need to modify lines 35, 40, and 41 as well.
  - Once the data is acquired, this Stata .do file must be run only once in order to obtained the final version of the data set we use to replicate the empirical results in the paper.
  - This script calls Stata .do files `programs/01_processdata.do` and `programs/02_processdata.do`.
  - The output of this script is the final data set used in the empirical application. It is called `pre-processed.dta` and is stored in directory `data/processed`.
  - The program was last run on August 8th, 2024.
- `programs/01_processdata.do`.
  - This script requires restricted-use data from Jobs First, previously converted to Stata format as `rawdata.dta` in `data/acquired`.
  - This .do file is a minimal modification of a .do file from Bitler et al. (2017). Importantly, their Stata .do file is licensed under BSD-3 license (see LICENSE.txt for details).
  - The output of this file is a data set in Stata format named `data-final.dta` and stores it in `data/processed`.
  - Interested parties can get the original code in Bitler at al. (2017) replication file (`createdata.do`).
- `programs/02_processdata.do`.
  - This script requires the output of `programs/01_processdata.do`, i.e., `data-final.dta`, stored in `data/processed`.

  - This .do file is a minimal modification of a .do file from Bitler et al. (2017). Importantly, their Stata .do file is licensed under BSD-3 license (see LICENSE.txt for details).
  - Interested parties can get the original code in Bitler at al. (2017) replication file (`bgh_table2.do`).

- `analysis/main.r`:
  - The script uses R package `rprojroot` to define robust and flexible paths to files in the replication directory relative to a project root.
  - The script sources another R script to install all the R packages and dependencies; see `programs/global_libraries.r`.
  - The script calculates the p-values for all subgroups defined by different combinations of covariate levels, and stores the output in the same directory as **raw p-values.RData**. The program was last run on July 7th 2024 and it took approximately 11 hrs under the conditions stated above. The replicator can calculate separate p-values for particular covariates without running the entire analysis (e.g., individual rows of Table 3 in main text).

- `programs/PT.qp.r`:
  - The function calculates the permutation test based on the Khmaladze transformation of the quantile process. This is function relies on R packages `quantreg` and `RATest`. While this function is self-contained within this project, interested parties can obtain more details, e.g., see the documentation in R package `RATest` on CRAN, or read the description in main text and supplementary appendix.
  - The calculation of the permutation distribution relies on a stochastic approximation. Specifically, we draw permutations from the set of all permutations of indices $\{1, \ldots, N\}$ uniformly at random. Sampling of permutations of indices uses R's internal PRNG. We did not explicitly set a seed in the R function that samples the indexes for the permutations. Therefore, while the discrepancies are negligible and never reverse the conclusions, the replicator cannot get numerically identical p-values as reported in the text.
  - The number of permutations used in our analysis was 1000 (999 permutations plus the identity permutation).
  - The R script is run by `analysis/main.r`.

- `programs/report.pvalues.r`:
  - The function uses the p-values calculated by the proposed test and returns the adjusted p-values using Bonferroni and Holm methods to account for multiple testing. The number of comparisons is given by the user, and corresponds to the number of subgroups formed out of the different covariate levels.
  - The R script is run by `analysis/main.r`.

- The processed data set that is used in `analysis/main.r` contains several economic and demographic variables to form the subgroups. We followed the conventions stated in Bitler et al. (2017) up to minor modifications (these modifications derive from the fact that we import a Stata .dta file into R, and this process creates some changes in name conventions). In our empirical application, these variables are coded as follows:
  - `e`: Treatment indicator.
  - `quarter`: quarter.
  - `earnq`: Quarterly earnings.
  - `ed`: Education
  - `ykid`: Age of youngest kid.
  - `mh`: Marital Status.
  - `pq7e`: Earnings level seven quarters pre-random assignment (pre-RA).
  - `npqe`: Number quarters with positive earnings pre-RA.
  - `wh`: Welfare receipt seven quarters pre-RA.

  Whenever there were interactions between covariate levels, those are indicated by a dot in their name. For example, when the education level subgroups (`ed`) are interacted with earnings level seven quarters pre-RA (`pq7e`), we name that variable `pq7e.ed` for simplicity.

- Table 3: The table can be reproduced using the output of R script `analysis/main.r` (a RData file). The output contains a data frame with three elements: the raw p-values and the adjusted p-values using Bonferroni and Holm method. The other table that depends on the p-values calculated from the data set, Table 2 in the appendix, uses the p-values as reported in Bitler et al. (2017)

References

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. Replication data for: What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments. Nashville, TN: American Economic Association [publisher], 2006. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-12-07. https://doi.org/10.3886/E116239V1

Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. (2017). "Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? Evidence from a social experiment." Review of Economics and Statistics, 99(4), 683-697.

Bloom, D., and MDRC. Connecticut's Jobs First Program Analysis Data, 1996-2001. Inter-university Consortium for Political and Social Research [distributor], 2022-07-26. https://doi.org/10.3886/ICPSR38126.v1

Koren, M., M. Connolly, J. Llull, and L. Vilhuber. 2022. "Data and Code Availability Standard [Version 1.0]." Website. https://datacodestandard.org https://doi.org/10.5281/zenodo.7436134

Acknowledgements

This document was written following the Data and Code Availability Standard v1.0. Please visit Social Science Data Editors website for more details.

We are grateful to MDRC for granting access to the experimental data we use in this paper. Mauricio Olivares acknowledges support from the European Research Council (Starting Grant No. 852332).