README

Overview of Programs

The code in this replication package utilizes R and Matlab to produce all tables and figures presented in the manuscript and online appendix. There are four steps in the production of the tables and figures:

- **Step 1: Data Preparation.** This step performs preparation of the raw confidential micro-data in R. This step requires access to the confidential micro-data discussed below. The replication code for this step is confidential and available to qualified researchers internally from the Statistics of Income division of the IRS.
- **Step 2: Micro-data Estimation.** This step performs estimation using the confidential micro-data to produce intermediate statistical extracts in R. This step requires access to the confidential micro-data produced by step 1. The replication code for this step is available in subdirectory `2_MicrodataEstimation/` of our replication package, and is executed by sourcing the `main.R` script. We provide a Code Ocean replication capsule for this step applied to simulated data at CodeOcean.
- **Step 3: Permanent-Transitory Process (Table A3 only).** Using certain confidential statistical extracts produced by step 2, this step uses a polynomial optimization approach to estimate a permanent-transitory process in Matlab. The results from this step are only relevant to Online Appendix Table A3 of our manuscript, so this step may not be of primary interest to replicators. The replication code for this step is available in subdirectory `3_PTprocess/` of our replication package, and is executed by sourcing the `main.m` script. We provide a Code Ocean replication capsule for this step applied to simulated data at CodeOcean.
- **Step 4: Analyses of Statistical Extracts.** Given the statistical extracts and process estimates produced by steps 2 and 3, this step performs final analyses and produces all tables and figures in R. The replication code for this step is available in subdirectory `4_StatisticalAnalyses/code/` of our replication package, and is executed by sourcing the `main.R` script. We provide Code Ocean replication capsules for this step applied to both the real data at CodeOcean and to the simulated data at CodeOcean.

Step 1 -- Data Preparation

Overview of this Step

This step performs preparation of the raw confidential micro-data in R. This step requires access to the confidential micro-data discussed below. The replication code for this step is confidential and available to qualified researchers internally from the Statistics of Income division of the IRS.

Data Provinence and Availability

The project relied on US Treasury confidential de-identified micro-data to which access is restricted. We accessed this data through the Joint Statistical Research Program, a program administered by IRS Statistics of Income (SOI) to allow government-academic research partnerships. All data analyses occurred onsite at a secure IRS facility and were supervised and reviewed by SOI staff.

We obtained access to the data by submitting a project proposal in response to a call for proposals posted by the SOI on this website: https://www.irs.gov/statistics/soi-tax-stats-joint-statistical-research-program. The data may be obtained by submitting a proposal in response to the next call for propsals. Note that it can take some months to gain access to the data once a proposal is approved. The authors will assist with any reasonable replication attempts for two years following publication.

Statement about Rights and Availability

We certify that the authors of the manuscript have legitimate access to and permission to use the data used in this manuscript. No data can be made publicly available. However, the statistical extracts are included in the public replication package and allow for replication of all tables and figures presented in the paper and its online appendix.

Instructions to Replicators

The data preparation is performed by the "LMS_data" data replication pack available to internal researchers on the SOI server. It includes a README file which explains how to prepare the software environment and source the data construction script, which is written in R. See the Data Availability Statement above for further details on accessing this data. Once access is obtained, email the SOI administrator to request the replication pack. Run time for this step is approximately 10 hours.

Step 2 -- Micro-data Estimation

Overview of this Step

This step performs estimation using the confidential micro-data to produce intermediate statistical extracts in R. This step requires access to the confidential micro-data produced by step 1. The replication code for this step is available in subdirectory `2_MicrodataEstimation/` of our replication package, and is executed by sourcing the `main.R` script. We provide a Code Ocean replication capsule for this step applied to simulated data at CodeOcean.

Software Requirements

All estimation in this step is performed using R on the SOI server; R and its packages are installed by server administrators (it was most recently tested with R 4.0.1). At the time of most recent testing, all required packages were installed on the SOI server. However, if some are no longer available, request that they be installed by the server administrator. Package versions are listed based on most recent testing of the code.

Packages availabe from CRAN:

- `AER` (1.2-9)
- `checkmate` (2.0.0)
- `coop` (0.6-2)
- `data.table` (1.14.0)
- `futile.logger` (1.4.3)
- `ivpack` (1.2)
- `lfe` (2.8-6)
- `magrittr` (2.0.1)
- `modi` (0.1.0)
- `R.matlab` (3.6.2)
- `ShiftShareSE` (1.0.1)

Packages available from GitHub:

- `setzler/eventStudy/eventStudy`
- `setzler/LMS/LMS`
- `setzler/ShiftShareIV`
- `tlamadon/rblm`

Instructions to Replicators

In order to execute step 2, complete the following tasks in order:

- copy the R scripts from subdirectory `2_MicrodataEstimation/` onto the SOI server (with appropriate permission from the server administration);
- ensure that line 7 of `main.R` is uncommented so that `runtype = "IRS-real"`, and edit the directory paths on lines 47-50 to point to your local directory on the SOI server;

- source `main.R`. Run time is approximately 20 hours.

Replication Capsule

We have prepared a replication capsule on Code Ocean which demonstrates all of step 2 using simulated data that has the same data structure as the confidential micro-data. The capsule is available at CodeOcean. This capsule also makes clear the software required to execute `main.R`. Make sure to set `runtype = "CodeOcean-simulation"` by uncommenting line 6 and commenting line 7 in `main.R` and copy the contents of `2_MicrodataEstimation/` into the `code/` subdirectory of the Code Ocean capsule. Code Ocean run time is approximately 25 minutes.

Note: The purpose of the simulated data is to illustrate code execution. The simulated data is not meant to replicate statistical characteristics of the real data, and estimates based on the simulated data should not be the same as estimates based on the real data.

Step 3 -- Permanent-Transitory Process Estimation

Overview of this Step

Using certain confidential statistical extracts produced by step 2, this step uses a polynomial optimization approach to estimate a permanent-transitory process in Matlab. The results from this step are only relevant to Online Appendix Table A3 of our manuscript, so this step may not be of primary interest to replicators. The replication code for this step is available in subdirectory `3_PTprocess/` of our replication package, and is executed by sourcing the `main.m` script. We provide a Code Ocean replication capsule for this step applied to simulated data at CodeOcean.

Software Requirements

All estimation in this step is performed using Matlab (it was most recently tested with Matlab 2019a). Two add-ons must be saved to Matlab's general path:

- **SeDuMi:** This can be downloaded from https://github.com/sqlp/sedumi/archive/master.tar.gz. Note that it must be unpacked and saved to the Matlab general path before use.
- **gloptipoly3:** This can be downloaded from http://homepages.laas.fr/henrion/software/gloptipoly3/gloptipoly3.zip. Note that it must be unpacked and saved to the Matlab general path before use.

See the `postInstall` script of our Code Ocean capsule (discussed below) for an automated approach to download these two add-ons, unpack them, and add them to the Matlab general path.

Instructions to Replicators

In order to execute step 3, complete the following tasks in order:

- in the environment in which the Matlab software has been prepared, create three subdirectories called `code/`, `data/`, and `results/`;
- copy the contents of subdirectory `3_PTprocess/` from this replication pack into the `code/` subdirectory in the Matlab environment;
- locate the 6 files with extension `.dat` produced by step 2, which were saved to the output subdirectory `matlab/` on the SOI server. Copy these 6 files into the `data/` subdirectory of the Matlab environment, after receiving approval from SOI administration to remove these files from the SOI server;
- source `main.m`. Run time is approximately 10 minutes.

Replication Capsule

We have prepared a replication capsule on Code Ocean which demonstrates all of step 3 using simulated data that has the same data structure as the confidential input data. The capsule is available at CodeOcean. For the 6 `.dat` files, it uses the results produced by the Code Ocean capsule described under step 2, which were based on simulated data. The `postInstall` file available in the Code Ocean capsule demonstrates how to automatically prepare the software. Run time is approximately 2 minutes.

Step 4 -- Analyses of Statistical Extracts

Overview of this Step

Given the statistical extracts and process estimates produced by steps 2 and 3, this step performs final analyses and produces all tables and figures in R. The replication code for this step is available in subdirectory `4_StatisticalAnalyses/code/` of our replication package, and is executed by sourcing the `main.R` script. We provide Code Ocean replication capsules for this step applied to both the real data at CodeOcean and to the simulated data at CodeOcean.

Software Requirements

All estimation in this step is performed using R on a local machine (it was most recently tested with R 4.0.5). Package versions are listed based on most recent testing of the code.

Packages availabe from CRAN:

- `data.table` (1.14.0)
- `futile.logger` (1.4.3)
- `ggplot2` (3.35)
- `magrittr` (2.0.1)
- `R.matlab` (3.6.2)
- `scales` (1.1.1)
- `stringr` (1.4.0)

Packages available from GitHub:

- `setzler/textables`

Note: One must also have `pdflatex` available in the local environment, as it is utilized to produce the PDF versions of the tables.

Instructions to Replicators

In order to execute step 4, complete the following tasks in order:

- copy the entire output directory from step 2 into the `4_StatisticalAnalyses/data/` subdirectory of this replication pack (maintaining the same subdirectory structure), after receiving approval from SOI administration to remove these files from the SOI server. The `matlab/` subdirectory from step 2 can be omitted;
- copy the 3 output files ending in `.mat` format from step 3 into the `4_StatisticalAnalyses/data/matlab/` subdirectory of this replication pack;
- ensure that line 7 of `main.R` is uncommented so that `runtype = "local-real"`, and edit the directory path on lines 16-18 to point to your local directory;
- source `main.R`. Run time is approximately 1 minute.

Replication Capsule

We have prepared a replication capsule on Code Ocean which demonstrates all of step 4 using the statistical extracts from the *real IRS data*. The capsule is available at CodeOcean. Make sure to comment line 7 and uncomment line 8 so that `runtype = "CodeOcean-real"`. The capsule reproduces all tables and figures in the paper and its online appendix. Run time is approximately 3 minutes.

We have also prepared a replication capsule on Code Ocean which demonstrates all of step 4 using the statistical extracts from the *simulated data*. The capsule is available at CodeOcean. Make sure to comment line 7 and uncomment line 6 so that `runtype = "CodeOcean-simulation"`. Run time is approximately 3 minutes.

List of Output and Associated Programs

All tables and figures in the main text and online appendix of the manuscript are produced by step 4 (see details above). Input files are located in the `4_StatisticalAnalyses/data/` subdirectory of this replication pack. The list of tables and figures and their associated programs is as follows:

Main Text:

Figure/Table #	Function Call	Function Source	Output File(s)	Input File(s)	Input(s) Produced by
Table 1	table.1()	model_results.R	table1.pdf	model/LMS_model_descriptives_naics2_cz_broadmarket.csv, params/LMS_params_naics2_cz_broadmarket.csv	Step 2
Table 2	table.2_A4()	passthrough_results.R	table2.pdf	params/LMS_params_naics2_cz_broadmarket.csv, params/LMS_params_naics2_cz_broadmarket_bootstraps.csv, params/LMS_shiftshare_naics2_cz_broadmarket.csv, params/results_forLMS_LSelasticity.csv	Step 2
Table 3	table.3()	passthrough_results.R	table3.pdf	params/LMS_params_naics2_cz_broadmarket.csv, params/LMS_params_naics2_cz_broadmarket_bootstraps.csv	Step 2
Table 4	table.4()	model_results.R	table4.pdf	model/LMS_psidecomp_naics2_cz_broadmarket.csv, model/LMS_psidecomp_AKM_naics2_cz_broadmarket.csv	Step 2
Table 5	table.5()	model_results.R	table5.pdf	model/LMS_welfare_naics2_cz_broadmarket.csv	Step 2
Figure 1	figure.1()	passthrough_results.R	figure1.pdf	params/LMS_DiD_naics2_cz.csv	Step 2
Figure 2	figure.2()	FE_results.R	figure2.pdf	FE/blm_adjusted_naics2_cz_broadmarket.rds	Step 2
Figure 3	figure.3abcd()	model_results.R	figure3a.pdf, figure3b.pdf, figure3c.pdf, figure3d.pdf	FE/blm_adjusted_naics2_cz_broadmarket.rds, model/LMS_shrinksort_naics2_cz_broadmarket.csv	Step 2

Online Appendix:

Figure/Table #	Function Call	Function Source	Output File(s)	Input File(s)	Input(s) Produced by
Table A1	table.A1_A2()	descriptives_results.R	tableA1.pdf	descriptives/LMS_descriptives_naics2_cz.csv, descriptives/LMS_descriptives_naics2_cz_broadmarket.csv, descriptives/LMS_descriptives_naics2_cz_movers.csv, descriptives/LMS_descriptives_naics2_cz_broadmarket_movers.csv, descriptives/LMS_descriptives_naics2_cz_stayersteps.csv, descriptives/LMS_descriptives_naics2_cz_broadmarket_stayersteps.csv	Step 2
Table A2	table.A1_A2()	descriptives_results.R	tableA2.pdf	descriptives/LMS_descriptives_naics2_cz.csv, descriptives/LMS_descriptives_naics2_cz_broadmarket.csv, descriptives/LMS_descriptives_naics2_cz_movers.csv, descriptives/LMS_descriptives_naics2_cz_broadmarket_movers.csv, descriptives/LMS_descriptives_naics2_cz_stayersteps.csv, descriptives/LMS_descriptives_naics2_cz_broadmarket_stayersteps.csv	Step 2
Table A3	table.A3()	passthrough_results.R	tableA3.pdf	matlab/overall_passthrough_matlab_results.mat, matlab/bootstrap_passthrough_matlab_results.mat, matlab/bootstrap_passthrough_matlab_results_MA2.mat	Step 3
Table A4	table.2_A4()	passthrough_results.R	tableA4.pdf	params/LMS_params_naics2_cz_broadmarket.csv, params/LMS_params_naics2_cz_broadmarket_bootstraps.csv, params/LMS_shiftshare_naics2_cz_broadmarket.csv, params/results_forLMS_LSelasticity.csv	Step 2
Table A5	table.A5()	passthrough_results.R	tableA5.pdf	params/LMS_params_naics2_cz.csv, params/LMS_params_naics2_cz_broadmarket.csv, params/LMS_params_naics2_county_broadmarket.csv, params/LMS_params_naics2_state_broadmarket.csv, params/LMS_params_naics3_cz_broadmarket.csv, params/LMS_params_supersector_cz_broadmarket.csv	Step 2
Table A6	table.A6()	FE_results.R	tableA6.pdf	FE/blm_unadjusted_naics2_cz_broadmarket.rds, FE/blm_adjusted_naics2_cz_broadmarket.rds	Step 2
Table A7	table.A7()	passthrough_results.R	tableA7.pdf	params/LMS_params_naics2_cz_broadmarket.csv, params/LMS_params_naics2_cz_broadmarket_bootstraps.csv	Step 2
Figure A1	figure.A1ab()	passthrough_results.R	figureA1a.pdf, figureA1b.pdf	params/LMS_params_naics2_cz_broadmarket.csv, params/LMS_params_naics2_cz_subsamples.csv	Step 2
Figure A2	figure.A2()	descriptives_results.R	figureA2.pdf	descriptives/LMS_taxfit.csv	Step 2

Figure/Table #	Function Call	Function Source	Output File(s)	Input File(s)	Input(s) Produced by
Figure A3	figure.A3ab()	passthrough_results.R	figureA3a.pdf, figureA3b.pdf	params/LMS_params_naics2_cz_broadmarket.csv	Step 2
Figure A4	figure.A4abcd_A5()	model_results.R	figureA4a.pdf, figureA4b.pdf, figureA4c.pdf, figureA4d.pdf	model/LMS_fitgrid_naics2_cz_broadmarket.csv	Step 2
Figure A5	figure.A4abcd_A5()	model_results.R	figureA5.pdf	model/LMS_fitgrid_naics2_cz_broadmarket.csv	Step 2
Figure A6	figure.A6()	model_results.R	figureA6.pdf	model/LMS_compdiffs_naics2_cz_broadmarket.csv	Step 2
Figure A7	figure.A7ab()	model_results.R	figureA7a.pdf, figureA7b.pdf	model/LMS_shrinks_naics2_cz_broadmarket.csv	Step 2

This output can be viewed in the `results/` subdirectory of our Code Ocean replication capsule at CodeOcean.

Licenses