Replication package README

Subjective Life Expectancies, Time Preference Heterogeneity, and Wealth Inequality

Richard Foltyn Jonna Olsson

March 19, 2024

This replication package generates the empirical part of the analysis for the paper "Subjective Life Expectancies, Time Preference Heterogeneity, and Wealth Inequality", conditionally accepted at Quantitative Economics (Feb 21, 2024). It also includes the intermediate analysis data sets generated. For the full replication package, including the quantitative model, see the published version of the paper in Quantitative Economics.

Contents

1 Data Availability Statement

1.1 HRS data

The main data set used for the empirical analysis is the Health and Retirement Study (HRS). The raw data is NOT included in the replication package, since HRS Conditions of Use prohibit redistribution of any HRS data product. However, the data is easily downloadable from the HRS website:

1. Go to `https://hrsdata.isr.umich.edu/data-products/rand` (you need to be logged in, registration is free and easy).
2. Click on the link "RAND HRS Archived Data Products".
3. Download the "RAND HRS 1992-2018 (V2)" dataset for Stata (`randhrs1992_2018v2_archive_STATA.zip`).
4. The zip file contains two .dta files. Place those in the `input/` folder.

1.2 PSID data

The PSID data is used for estimating the health process for ages below 50 and the age-health dependent labor income profile. The raw data is NOT included in the replication package, since the PSID Conditions of Use prohibit redistribution of the data (including user-created extracts). However, the data is easily downloadable from the PSID Public Data Extract Repository:

1. Go to `https://doi.org/10.3886/E198564V1`.
2. Download the file `PSID_raw.txt`.
3. Place the file in the `input/` folder.

1.3 Data retrieved from FRED

The data series for CPI is included in the replication package. It was originally downloaded from the FRED website at `https://fred.stlouisfed.org/series/CPIAUCSL`.

1.4 Data retrieved from BLS

The data series for CPI for Medical Care Services is included in the replication package. It was originally downloaded from the BLS website at `https://beta.bls.gov/dataViewer/view/timeseries/CUUR0000SAM`.

2 Content of the replication package

The following describes all sub-directories and their content:

- `input/`

  Input data from FRED (included), RAND HRS (needs to be downloaded separately) and PSID (needs to be downloaded separately).

- `output/`

  All output data created by the replication package.

- `01-data/`

  This directory includes all data processing of the HRS, PSID and FRED input data to create the final data sets for estimation.

  Additionally, the folder contains do-files to run the regressions for wealth or wealth changes reported in the paper.

  - `env.do`

    Environment setup used to set paths, in particular to the RAND HRS files and PSID extract.

  - `main.do`

    Main do-file to run the files `01...` to `09...` described below all at once.

  - `01_hrs_import.do`

    Imports original RAND HRS longitudinal and imputation and creates a processed version stored in `hrs_processed.dta`.

  - `02_hrs_nonresp.do`

    Tabulates non-response patterns in HRS and writes these to `hrs_nonresp_stats.csv`.

  - `03_hrs_structure.do`

    Tabulates the HRS survey structure and writes it to `hrs_survey_structure.csv`.

  - `04_hrs_estim_sample.do`

    Creates the final estimation sample for health transitions, survival and medical expenditures and writes these to `hrs_trans.dta` and `hrs_medex.dta`.

  - `05_regress_wealth_beliefs.do`

    Runs regression of wealth on subjective survival beliefs.

  - `06_regress_wealth_hshock.do`

    Runs regressions of wealth changes on health shocks.

- 07_psid_import.do

  Processes raw PSID data and stores it in psid_processed.dta.

- 08_psid_labincome.do

  Computes life cycle of earnings by health and stores results in earn_profile_health.csv.

- 09_psid_estim_sample.do

  Creates final PSID estimation sample for health transitions below the age of 50 and stores it in psid_trans.dta.

- main_bias.do

  Runs survival bias by age/race/sex/health regressions and stores the outputs in surv_bias_pred_MLE_NL1.csv and surv_bias_pred_by_health_MLE_NL1.csv. This file is called automatically from 02-estimation/02_post_estimation.py as described below.

- main_flatness_bias.do

  Runs survival bias by age/sex regressions and stores the output in hrs_subj_vs_obj_surv_ref_w6_MLE_NL1.csv. This file is called automatically from 02-estimation/02_post_estimation.py as described below.

- include/

  Directory for additional include files.

- lib/

  Directory for additional *.ado and Mata files.

- 02-estimation/

  This directory contains the implementation of the MLE, NLS and GMM estimators to estimate health transitions, survival probabilities and medical expenditure risk.

  - 01_run_estimation.sh

    Unix shell script to run all MLE, NLS and GMM estimations.

    This script calls the Python scripts estim_obj.py, estim_subj.py, estim_medex.py and estim_medex_annual.py to perform these tasks

  - 02_post_estimation.py

    Python script to run all post-estimation analysis and create input data for the OLG model.

  - estim_obj.py

    Runs the MLE estimation for objective health and survival probabilities.

  - estim_subj.py

    Runs the NLS estimation for subjective survival probabilities.

- `estim_medex.py`

  Runs the GMM estimation for medical expenditure risk at biennial frequency.
- `estim_medex_annual.py`

  Runs the GMM estimation for medical expenditure risk at annual frequency.
- `env.py`

  Environment setup script. Does not need to be modified.
- `fo/`

  Directory contains the implementation of the MLE, NLS and GMM estimators and the post-estimation analysis.
- `external/`

  Directory contains the following bundled external Python libraries:
  - pydynopt: https://github.com/richardfoltyn/pydynopt
  - footable: https://github.com/richardfoltyn/footable

3 Instructions to run

3.1 Detailed instructions

Replicating the empirical part of the analysis of the paper involves the following steps:

1. Download the RAND HRS files and the PSID extract as described in Section 1 and place these in the `input/` folder.

   Alternatively, if the HRS files are already present on your system you can change the variables `HRS_INPUT_FILE` and `RAND_IMPUTATION_FILE` in the file `01-data/env.do` to point to these:

   ```
   // RAND HRS longitudinal file
   global HRS_INPUT_FILE = "/path/to/randhrs1992_2018v2.dta"
   // RAND HRS detailed imputation file
   global RAND_IMPUTATION_FILE = "/path/to/randhrsimp1992_2018v2.dta"
   ```

2. Change to the directory `01-data/` and run the do-file `main.do` with Stata. This will also install the packages `estout` and `winsor`.

3. Go back to the directory `sle-replication`. Create a Python environment using the environment definition `environment.yml`. With the Anaconda command prompt, this can be achieved using

   ```
   conda env create -f environment.yml
   ```

   This creates an environment called `SLE` which needs to be activated using

   ```
   conda activate SLE
   ```

4. Change to the directory `02-estimation/` and run the commands in `01_run_estimation.sh` in a Unix shell:

   ```
   cd 02-estimation
   bash 01_run_estimation.sh
   ```

   Due to the bootstrapping performed in this step, it takes around 4.5 hours to run on a 32-core workstation (see section 5.2). The run-time can be reduced by turning off bootstrapping by setting

   ```
   BOOTSTRAP=0
   ```

   in `01_run_estimation.sh`. However, then some of the generated figures will not display confidence intervals.

   The number of processes used to bootstrap in parallel is governed by the variable `NUM_THREADS` in `01_run_estimation.sh` and should be adapted to your environment, e.g.,

   ```
   # Bootstrap using 16 parallel processes
   NUM_THREADS=16
   ```

5. Run the post-estimation analysis after the estimation is complete by executing

   ```
   python 02_post_estimation.py
   ```

   This Python scripts needs to call Stata to perform some of the steps. This will fail if Stata is installed in a non-standard location, in which case the following do-files have to be run manually from the directory `01-data/`:

   a) `main_bias.do`

   b) `main_flatness_bias.do`

   After running these, the Python script `02_post_estimation.py` needs to be run one more time to create the missing figures from the Stata output.

3.2 Summary

To summarize, in a compatible computing environment (see section 5.1), the following steps run the entire project and create all figures and tables:

```
# Prepare data
cd 01-data
stata-se main.do

# Run MLE, NLS and GMM estimation
cd ..
conda env create -f environment.yml
conda activate SLE
cd 02-estimation
bash 01_run_estimation.sh
python 02_post_estimation.py
```

4 Output generated

All generated output is stored in the folder `output/` which is structured as follows:

- `output/`

  All intermediate and final data are stored in this folder. The final outputs generated by the MLE, NLS and GMM estimators in step 4 are included as part of the replication package so the steps in 5 described above can be run even without re-estimating and recomputing everything.

  The Stata do-files run from `01-data/main.do` generate the following intermediate data:

  - `hrs_processed.dta`, `psid_processed.dta`

    Processed HRS and PSID data files.

  - `hrs_trans.dta`, `psid_trans.dta`

    Estimation sample for health and survival transitions for the HRS and PSID.

  - `hrs_medex.dta`

    Estimation sample for the medical expenditures estimation.

  - `hrs_nonresp_stats.csv`, `hrs_survey_structure.csv`

    CSV files summarizing HRS sample sizes and survey structure used to generate the figures in online appendix section A.

  - `earn_profile_health.csv`

    Health-dependent life cycle profile of earnings used in OLG model.

  The estimation outputs generated by running `02-estimation/01_run_estimation.sh` (including 1001 bootstraps, where applicable) are stored in the following files:

  - `MLE_NL1.pkl.xz`

    Results for ML estimate of objective health and survival probabilities.

  - `MLE_NL1_subj_sample.pkl.xz`

    Results for ML estimate of objective health and survival probabilities restricted to the sub-sample of respondents who also report subjective survival beliefs (for results reported in the appendix).

  - `SLE1.pkl.xz`

    Results for NLS estimate of subjective survival probabilities.

  - `MLE_NL1e_H3.pkl.xz`

    Results for ML estimate of objective health and survival probabilities with education as additional covariate (for results reported in the appendix).

- `MLE_MNL1_PSID_f0b0.pkl.xz`

  Results for ML estimate of objective health transition probabilities for ages below 50 using PSID data.

- `MedEx_SLE_f0b0.pkl.xz`

  Results for the GMM estimate of medical expenditure risk at biennial frequency.

- `MedEx_SLE_f0b0_1Y.pkl.xz`

  Results for the GMM estimate of medical expenditure risk at annual frequency.

The Python script `02-estimation/02_post_estimation.py` generates the following data intermediate data files:

- `hrs_subj_vs_obj_surv_ref_w6_MLE_NL1.csv`

  Intermediate data used to generate flatness bias Figure 2.

- `surv_bias_pred_MLE_NL1.csv`

  Intermediate data used to generate survival bias Figure 4.

- `surv_bias_pred_by_health_MLE_NL1.csv`

  Intermediate data used to generate survival bias Figure 5.

The health transition and survival probabilities created by the script `02-estimation/02_post_estimation.py` which are used as inputs into the OLG models are stored in the following files:

- `health_surv_prob_obj.txt`

  Health transitions and objective survival probabilities for ages 20–110.

- `health_surv_prob_subj.txt`

  Health transitions and subjective survival probabilities for ages 20–110.

5 Software and hardware requirements

5.1 Software

The replication package was tested on Ubuntu Linux versions 22.04 and 23.10 using the following software versions:

- Stata 18

  The code automatically installs the required packages `winsor` and `estout`. This can be disabled in `01-data/env.do` by commeting the lines

```
capture ssc install winsor
capture ssc install estout
```

- Python 3.10.x and the Python libraries specified in `environment.yml`. We use the Anaconda distribution instead of the Python version shipped with Ubuntu 22.04 which can be downloaded for free at https://www.anaconda.com/download/.

5.2 Hardware

The replication package was run on a 32-core / 64-thread AMD Ryzen Threadripper PRO 5975WX processor with 128 GB RAM. The code will work with 64 or possibly even 32 GB of RAM, however, the bootstrapping is heavily parallelized and will take a long time to run on standard laptop-class or even desktop-class hardware available in 2024.

For reference, we also ran the replication package on an AMD Ryzen 7840U CPU with 8 cores / 16 threads which was a mid-range laptop CPU in 2023. The following table lists the run-times for the performance-critical parts of the replication package:

	Time	
	Threadripper PRO 5975WX	Ryzen 7840U
Step	32 threads	16 threads
`02-estimation/01_run_estimation.sh`		
With boostrap	4h 25min	8h 12min
Without boostrap	54min	1h 25min

The other steps have negligible run-times of less than 10 minutes each on the same hardware.

6 License

The main source code (see exception below) and outputs shipped with this replication package are licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, see https://creativecommons.org/licenses/by/4.0/ and the full license text in `LICENSE.txt`.

When using any of these materials, a citation referencing the original paper must be included.

The source code from external projects contained in the folders

```
02-estimation/external/*
```

is included only for convenience and distributed under various other (permissive) licenses. See the documentation in these subfolders for details.

7 Data citations

CPI

U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: All Items in U.S. City Average [CPIAUCSL], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CPIAUCSL, March 3, 2024.

CPI for Medical Care Services

U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: Medical Care Services in U.S. City Average [CUUR0000SAM], retrieved from BLS; https://beta.bls.gov/dataViewer/view/timeseries/CUUR0000SAM, March 4, 2024.

HRS

Health and Retirement Study, RAND HRS 1992-2018 (V2) public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, 2023.

PSID

Panel Study of Income Dynamics, public use dataset. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI (2024).

Extract: Foltyn, Richard, and Olsson, Jonna. Foltyn and Olsson (2024). Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2024-02-23. https://doi.org/10.3886/E198564V1