Replication package for:

Evaluating the accuracy of counterfactuals – Heterogeneous survival expectations in a life cycle model

This README explains how to reproduce the results in the paper, from the manipulation of raw data to all figures in the main text and online appendices. Its structure is as follows:

1. **Data availability statement**
2. **Description of programs/code**
   - 2.1 **Data manipulation:** takes the raw DHS data as input and produces the analysis data. This includes both the data used to estimate and simulate the measurement model of subjective survival and that for the retirement model.
   - 2.2 **Subjective survival model:** the code that implements the likelihood, runs the optimization (estimation), and carries out simulations regarding subjective survival. Optimization of the likelihood is computationally intensive and took more than a week on a desktop computer.
   - 2.3 **Retirement model:** this solves the retirement model, runs the optimization (estimation), and carries out simulations for retirement models that utilize all four specifications for survival expectations discussed in the paper. Optimization of the Method of Moments objective function is extremely computationally intensive. The models without heterogeneity in survival took about a week on a high-performance cluster, with parallel computations on 512 CPU cores. The model with individual-specific heterogeneity in subjective expectations took about six weeks of computer time on the cluster.
   - 2.4 **Figures and tables:** these scripts take the analysis data and model simulations as inputs and produce the figures and tables presented in the paper.
3. **Instructions to replicators**
4. **List of tables and programs**

Throughout this README `monospace font` is used to indicate scripts and underline indicates data files.

1 Data availability statement

The analysis is based on the 1993-2001 and 2006-2016 waves of the DNB Household Survey (DHS). While the author has legitimate access and permission to use the data used in the manuscript, he does not have permission to redistribute or publish the data in this replication package. Only Centerdata has the permission to publish and distribute DHS data. Raw data can be downloaded from https://www.dhsdata.nl/site/users/login after completion of a free registration procedure.

The data repository that contains the DHS is accredited with the CoreTrustSeal that ensures long-term accessibility. For more information see section 6.6 of *Data Management and Preservation DHS 3.0*, which is included in this replication package.

2 Description of programs/code

2.1 Data manipulation

All data manipulation is carried out in Stata version 15. The transformation from raw data to analysis data consists of the following steps:

1. `ConstructData.do`
   Uses raw data as input, combines the different waves into a single dataset, recodes variables, and outputs the following data:
   (a) DHScombined_1993_2016.dta: combined DHS data for period 1993-2016.
   (b) rawdata.dta: main dataset used as input for model of subjective survival expectations and for the retirement model.
   (c) aux_spinc1.dta and aux_spinc2.dta: data used to estimate the auxiliary process for income of the spouse in the pre-reform (1) and post-reform (2) periods.
   (d) wagedata.dta: data for the auxiliary wage process. This includes men aged 40-60, which is a younger age bracket than used for the other analyses.

2. `AuxiliaryProcesses.do`
   Takes rawdata.dta as input and produces estimates of the auxiliary processes for the retirement model, including survival based on life tables (but not subjective survival):
   (a) health_estimation.csv and health_simulation.csv: health transitions in pre-reform and post-reform periods.
   (b) death_estimation.csv and death_simulation.csv: mortality process based on actuarial forecasts. Uses mortalityDB.csv, life tables downloaded from mortality.org, as additional input.
   (c) eqscale_estimation.csv and eqscale_simulation.csv: equivalence scales by age.
   (d) OOPexp_estimation.csv and OOPexp_simulation.csv: Out-Of-Pocket medical expenditures.

(e) sp_inc_estimation_zeroinc_sp.csv and sp_inc_simulation_zeroinc_sp.csv: exogenous income of the spouse.

3. `Auxiliary_subjective_longevity.do`
   Takes rawdata.dta as input and generates the data to be used for the estimation of subjective survival. It also constructs Halton sequences used to simulate the likelihood. It produces:

   (a) data_mortality_1.csv and data_mortality_2.csv: the datasets used to estimate subjective survival processes in the pre-reform (1) and post-reform (2) samples.

   (b) nobs_round_1.csv and nobs_round_2.csv: meta-data for the datasets on which subjective survival is estimated.

   (c) draws100round_ind1_[1,2].csv, draws100round_ind2_[1,2].csv, draws100round_seq1_[1,2].csv, draws100round_seq2_[1,2].csv: Halton draws used to simulate the likelihood across the distributions of both levels of unobserved heterogeneity (individual and question-sequence effects) in the pre- and post-reform periods.

4. `Wage_process.do`
   Takes wagedata.dta as input and estimates the autoregressive process for wage risk.

5. `Constructmom_estsample.do` and `Constructmom_simsample.do`
   Use rawdata.dta as input. `Constructmom_estsample.do` produces the dataset from which moments are computed that are targeted in the estimation of preference parameters in the retirement model (net of household size and birth cohort). It also calculates the diagonal weighting for those moments. `Constructmom_simsample.do` constructs analogous datasets for the post-reform period. The output of `Constructmom_estsample.do` is:

   (a) estimationMOM_coh1940_LS1940.dta: dataset with variables from which cohort and household size effects are removed.

   (b) weight_zeros_coh1940_LS1940.csv: diagonal weighting matrix for moments targeted in estimation.

   (c) data_estimation_coh1940_LS1940.csv: used in the estimation of the retirement models to compute data moments to be targeted.

   (d) dnobs_coh1940_LS1940.csv: meta-data from data_estimation_coh1940_LS1940.csv, e.g., the number of variables and observations.

6. `Combine_adjustedmom.dta`
   Takes estimationMOM_coh1940_LS1940.dta and simulationMOM_incLSall_conditionalLS.dta as inputs and generates actual_data_1940_LS1940_INC_allobs.dta, a dataset with moments to be matched in estimation for the pre-reform sample and their equivalents for the post-reform sample. Note that this dataset is not used in the estimation routine,

which computes moments directly from the individual-level data. It is only used to draw the figures for target moments, in-sample fit, and behavior in the post-reform pension regime.

7. `Moments_conditional_on_LE.do`
   Takes data_estimation_coh1940_LS1940.csv and posterior_means_estimation.dta and the corresponding files for the post-reform sample as inputs and generates data_estimation_hetexp_LS1940.dta and data_simulation_hetexp.dta. These files are used to draw figures for behavior conditional on subjective survival. It also generates data_estimation_exp_LS1940.csv and data_simulation_exp.csv: the data files used to estimate and simulate the retirement model with heterogeneous survival expectations.

8. `Bootstrap_CIs_datamoments.do`
   Constructs 95% confidence intervals for data moments in the pre- and post-reform periods using a bootstrap clustered by individual. It takes estimationMOM_coh1940_LS1940.dta, simulationMOM_incLSall_conditionalLS.dta, data_estimation_hetexp_LS1940.dta, and data_simulation_hetexp.dta as inputs and produces the following:

   (a) CI_all_estimationsample.dta and CI_all_simulationsample_conditionalLS.dta: confidence bounds on moments targeted in estimation and their analogs in the post-reform data.

   (b) CI_all_estimationsample_hetexp.dta and CI_all_simulationsample_hetexp.dta: confidence bounds on data moments conditional on subjective survival expectations.

9. `Initial_conditions_LS1940.do`, `Initial_conditions_LS1940_hetexp.do`, and `Initial_conditions_healthbehaviors.do`
   Take rawdata.dta and the relevant files with cohort- and family size-adjusted wealth as inputs (e.g., estimationMOM_coh1940_LS1940.dta) and produce initial conditions to be used to simulate behavior for the retirement models:

   (a) estimationINIT_coh1940_LS1940.csv, simulationINIT.csv, estimationINIT_hetexp_LS1940.csv, simulationINIT_hetexp.csv.

   (b) estimationINIT_coh1940_LS1940_healthbehaviors.csv and simulationINIT_healthbehaviors.csv, these require regressions_expectations.dta as additional inputs. (That file contains health behaviors: smoking and drinking.)

10. `LT_survival_HealthBehavior.do`
    Takes rawdata.dta and regressions_expectations.dta as inputs and constructs auxiliary process for survival conditional on health behaviors (smoking and drinking) based on published estimates of their effects on mortality. It produces death_hbehavior_estimation.csv and death_hbehavior_simulation.csv.

11. `quantiles_expectations_forgrids.do`
    Input: estimationINIT_hetexp_LS1940.csv and simulationINIT_hetexp.csv. Computes

quantiles of posterior means that capture variation in survival in the initial conditions, to be used in the discretization of subjective survival when solving the retirement model.

2.2 Subjective survival model

This folder contains the code required to estimate the measurement model of subjective survival. The procedure is the same for the pre-reform (estimation) and post-reform (simulation) samples, so the information for the pre-reform sample below also applies to the post-reform sample. The simulated likelihood is maximized in Fortran by means of Simulated Annealing. Standard errors are calculated in Stata in a single iteration of its Maximum Likelihood routine.

2.2.1 Estimation

`compile.sh` uses `setup_types_round_est.f95`, `objective_function_round_est.f95`, and `simann_round_est.f90` to generate `run`: the program that carries out the optimization.

1. `setup_types_round_est.f95`
   Initiates the arrays that will hold the data and the parameters that describe the dimensions of those arrays. It also sets the path to the inputs for the problem, notably the data and initial parameter vector. IMPORTANT: while pointing to inputs by means of this path works on macOS, it does not compile on Linux systems. For Linux, adapt `simann_round_est.f90` and use relative paths from the current directory.

2. `objective_function_round_est.f95`
   Computes the value of the likelihood at given parameters. It uses openMP to run computations in parallel across simulation draws.

3. `simann_round_est.f90`
   Loads the data and initial values, both stored in the 'Params'-folder, and optimizes the likelihood by means of Simulated Annealing. Uses Fortran library 'dcdflib_mac.a' to compute probabilities from normal distributions. Use 'dcdflib_lykle.a' on a Linux system, or compile the library anew.

4. `Run_model_rounding_est.do`
   Loads the data and initial values. Runs `Model_rounding_est.do`, which computes the likelihood at given parameter values. Computes standard errors in a single iteration of Stata's Maximum Likelihood routine.

2.2.2 Other scripts

1. `SEs_for_covariance_matrix.do`
   Takes the estimates for the measurement model of survival as input and calculates the elements of the covariance and correlation matrices of unobserved heterogeneity and the associated standard errors from the estimated Cholesky factors.

2. `posterior_means_gamma.do` and `posterior_means_gamma_simsample.do`
   Take as inputs data_mortality_[1,2].dta and the estimates of the measurement model and simulate posterior means of unobserved heterogeneity given the survival probabilities reported by respondents. The main outputs are posterior_means_estimation.dta and posterior_means_simulation.dta, which contain the posterior means. The scripts also draw figures of the kernel densities of those posterior means and the distributions of survival probabilities that they generate.

3. `Simulate_survival_probabilities_at_sample_covariates.do`
   Simulates survival probabilities across the distribution of unobserved heterogeneity and calculates average survival probabilities. These probabilities are saved in death_subj_estimation.csv and death_subj_simulation.csv to be used as auxiliary processes in the retirement model with average subjective expectations. The script also draws figures that plot survival probabilities against age.

4. `Regressions_indeffects_backgroundinfo.do`
   Takes DHS_combined1993_2016.dta, rawdata.dta, posterior_means_estimation.dta, and posterior_means_simulation.dta as inputs. It constructs proxy variables for preferences and regresses the posterior means for subjective survival expectations on background characteristics and preferences.

2.3 Retirement model

The estimation of preference parameters and simulation of behavior took place on LISA, a high-performance cluster managed by SURF. More information can be found on the website. The following environment applied at the time of the analysis:

- Operating system: Debian GNU/Linux 10 (buster)
- Kernel: Linux 4.19.0-24-amd64
- Architecture: x86-64

Estimation used the following version of Fortran: mpif90 version GNU Fortran (GCC) 9.3.0. Slight differences in function values for the Method of Moments objective function can occur when one runs the code in a different environment.

One estimation routine is included for each of three versions of the retirement model that correspond to different specifications of survival: fully homogeneous expectations based on life tables or average subjective expectations, fully heterogeneous subjective expectations, and life tables adjusted for smoking and drinking. Since estimation is carried out in exactly the same way for all three models, the description for the model with homogeneous expectations below also applies to the other two versions.

The models are compiled using `Makefile`, which

1. Specifies the Fortran compiler to be used.

2. Specifies the flags to be switched on (i.e., the types of errors Fortran should look for and report).

3. Compiles `quantile_function.f95`, `simann_dp.f90`, and `estimation_bootstrapS.f95`: the programs that govern estimation.

   (a) `quantile_function.f95`
   Calculates percentiles of a vector. Used to compute quantiles of wealth from simulated data.

   (b) `simann_dp.f90`
   Implements the Simulated Annealing optimization algorithm described in "Global Optimization of Statistical Functions with Simulated Annealing," Goffe, Ferrier and Rogers, Journal of Econometrics, vol. 60, no. 1/2, Jan./Feb. 1994, pp.65-100. This is a slight adaptation of the example code provided with that publication.

   (c) `estimation_bootstrapS.f95`
   The main program that loads data (subroutine 'initdata') and initial parameter values (subroutine 'initpar'), initiates the state space (subroutine 'initstatespace'), solves the dynamic programming retirement model at given values of parameters (subroutines 'termination', 'optimal', and 'recursion'), simulates data for 5000 individuals (subroutine 'simulate'), computes the Method of Moments objective function (subroutine 'fcn'), and runs the estimation (subroutine 'doestimation'). This list of subroutines is not exhaustive: most subroutines call other programs to carry out partial tasks. Based on user input 'todo', the program does one of the following:

      i. *todo = sa*: optimize objective function by means of Simulated Annealing.
      ii. *todo = once*: evaluate the objective function once at parameter vector specified in 'initpar' and save simulated data to output-folder.
      iii. *todo = SEs*: calls subroutine 'docovariance' to calculate standard errors at parameter vector specified in 'initpar'.

   The initial parameter vector is set to the estimates for the model with life table expectations reported in the paper.

Section 8 in the paper and Appendix J present simulations for the pre- and post-reform periods based on the model with heterogeneous survival expectations in which one or more preference parameters are set to their estimates based on the model with homogeneous, average, subjective survival expectations. These simulations are obtained by setting preference parameters in the model with heterogeneous survival to the estimates obtained for the model with average subjective survival.

Appendix H reports a decomposition of the total difference between simulations for the pre- and post-reform periods by changing one and only one aspect at a time (initial conditions and survival, auxiliary processes, taxes, or occupational pensions). This is implemented by substituting the corresponding element from the code for the post-reform period into that for the pre-reform period.

2.4 Figures and tables

1. `Decisions_age.do`
   Draws Figure 1 of the paper, which shows the decisions available in the model to agents at different ages.

2. `Figure2.do`
   Draws Figure 2 of the paper, which illustrates the logic of the measurement model for subjective survival.

3. `Figure_sim_data.do`
   Takes simulations from retirement models and data moments as inputs and draws Figure 3 of the paper, which shows model fit.

4. `Figure_simulations.do`
   Takes simulations from retirement models and data moments as inputs and draws Figures 4 in the main text and G1 in Appendix G.

5. `sample_selection.do`
   Takes rawdata.dta as input and produces the descriptive statistics regarding the sample selection procedure reported in Table F1 in Appendix F that describes first-step estimation results.

6. `Retage.do`
   Takes DHScombined1993_2016.dta and rawdata.dta as inputs and constructs retirement ages for the pre- and post-reform samples. It computes the summary statistics for observed data reported in Table 3.

7. `Retage_in_simulations.do`
   Takes simulated data for pre- and post-reform periods for the three retirement models (e.g., simpop_estdata_postupdate_LT.dat) and calculates retirement ages. It computes the summary statistics for simulated data reported in Table 3.

8. `Simulated_moments_by_tertile_of_expectations.do`
   Takes simulated datasets as inputs and computes moments for labor supply, benefit claiming, and wealth conditional on subjective survival. Draws Figure 5 in the paper.

9. `Simulations_by_tertiles_expectations_diff_prefs.do`
   Takes simulations for the post-reform period based on the model with heterogeneous survival expectations at different sets of preference estimates as inputs and draws Figure 6 in the paper. That figure shows differences in labor supply, pension claiming, and wealth between those who expect to die at young and old age.

10. `Figure_all_survivalexpectations.do`
    Combines all datafiles with survival expectations (life tables, average subjective expectations, and heterogeneous subjective expectations), merges them, and draws figure with all survival processes (Figure F1 in Appendix F of the paper).

11. `Descriptives_initial_conditions.do`
    Loads files with initial conditions for the simulation of life courses, e.g., estimationINIT_coh1940_LS1940.csv, and computes descriptive statistics such as those in Table F3 in Appendix F of the paper.

12. `Figures_auxiliary.do`
    Takes all estimated auxiliary processes, except survival, as input and draws Figure F2 in Appendix F.

13. `Raw_data_conditional.do`
    Takes rawdata.dta and posterior_means_estimation.dta as inputs and draws Figure C2 in Appendix C.

14. `Regressions_indeffects_backgroundinfo.do`
    Takes DHScombined1993_2016.dta, rawdata.dta, posterior_means_estimation.dta, and posterior_means_simulation.dta as inputs and runs the regressions of the posterior means of the measurement model of subjective survival on demographics reported in Table C2 in Appendix C.

15. `FigureG2_simulations.do`
    Takes simulations from retirement models and data moments as inputs and draws Figure G2 in Appendix G.

16. `Figure_sim_data_incl_HB.do`
    Takes simulations from retirement models and data moments as inputs and draws Figure I1 in Appendix I.

17. `Figure_sim_data_differentprefs_estsample.do`
    Takes simulations from retirement models and data moments as inputs and draws Figure J1 in Appendix J.

18. `Simulated_moments_by_tertiles_different_prefs.do`
    Takes simulations from estimated retirement models and data moments as inputs and draws Figure J2 in Appendix J.

19. `dutchtax.f95`
    Sets up a grid for earnings before-tax and calculates taxes due both before and after the tax reform. This is done separately for workers and pensioners, since they face different contribution rates for social insurance. It generates dutchtax98_worker_2.dat, dutchtax98_pensioner_2.dat (pre-reform), dutchtax04_worker_2.dat, and dutchtax04_pensioner_2.dat (post-reform).

20. `Figure_taxes.do`
    Takes the tax rates calculated by `dutchtax.f95` as inputs and draws Figures OA1 and OA2 in the online supplementary material.

3 Instructions to replicators

3.1 Directories

Each Stata do-file starts with a declaration of the path to the working directory and possibly other paths to the data and figures. The structure of folders is as follows:

- **.../research/** main folder (working directory)
  - **.../research/Data/DHS/** raw DHS data
  - **.../research/Structural2/** folder that contains all processed datafiles and the code and figures for the paper
    - **.../research/Structural2/Temp/** temporary data files, e.g., bootstrap iterations for the confidence intervals on target moments
    - **.../research/Structural2/Paper/Figures** all figures used in the paper go here
    - **.../research/Structural2/Paper/Params/** holds the files that are used as inputs by the retirement model (e.g., data and auxiliary processes)
      - **.../research/Structural2/Paper/Params/Estimation**
      - **.../research/Structural2/Paper/Params/Simulation**
    - **.../research/Structural2/Data/**
      - **.../research/Structural2/Data/Moments/**
      - **.../research/Structural2/Data/Subj_mortality/**
      - **.../research/Structural2/Data/Simulations/**
      - **.../research/Structural2/Data/Simulations/Simulations changing prefs to subj exp – pre-reform sample/**
      - **.../research/Structural2/Data/Simulations/Simulations changing prefs to subj exp – post-reform sample/**

3.2 Running the code

The overview in Section 2 organizes programs according to their functionality: data manipulation, estimation of models, and generation of tables and figures. However, dependencies between programs mean that they cannot be run in that order. In particular, some programs that manipulate data require the posterior means generated by the measurement model of subjective longevity. The code can be run in the following order:

1. `ConstructData.do`
2. `AuxiliaryProcesses.do`
3. `Auxiliary_subjective_longevity.do`

4. `Wage_process.do`
5. `Constructmom_estsample.do`
6. `Constructmom_simsample.do`
7. `Combine_adjustedmom.do`
8. Estimate measurement model of subjective survival (see section 2.2.1)
9. Generate posterior means of subjective survival (see section 2.2.2)
10. `quantiles_expectations_forgrids`
11. `Moments_conditional_on_LE.do`
12. `Bootstrap_CIs_datamoments.do`
13. `Initial_conditions_LS1940.do`, `Initial_conditions_LS1940_hetexp.do`, `Initial_conditions_healthbehaviors.do`
14. `LT_survival_HealthBehavior.do`
15. Estimate the four versions of the retirement model and run simulations for the post-reform period (see section 2.3). Note that this requires one to copy the inputs saved in the "Params"-folder to the relevant "input" folder for each version of the model.
16. Draw figures and calculate table entries by running the programs in section 2.4

4 List of tables and programs

Figure/Table #	Program	Output file
a. Main text		
Figure 1	Decisions_age.do	decisions.eps
Figure 2	Figure_2.do	illustrate_model_exp.eps
Figure 3	Figure_sim_data.do	data_simulations_final_DHSwage.eps
Figure 4	Figure_simulations.do	impactreform_detailed_onlyHRS_DHSwages.pdf
Figure 5	Simulated_moments_by_tertile_of_expectations.do	exp_tertiles_estsample_DHSwages.pdf
Figure 6	Simulations_by_tertiles_expectations_diff_prefs.do	exp_tertiles_diff_simsample_DHSwages.pdf
Figure 7	Retage_under_different_prefs.do	prefs_hetexp_subj_DHSwages.eps
Table 1	Auxiliary_subjective_longevity.do	
Table 2	See section 2.3: Retirement model	
Table 3	Retage.do and Retage_in_simulations.do	
b. Appendices		
Figure C1	posterior_means_gamma_simsample.do	variation_exp.eps
Figure C2	raw_data_conditional.do	simple_indeff_complex_indeff.pdf
Figure E1	Bootstrap_CIs_datamoments.do	moments_est_LS1940.eps
Figure F1	Figure_all_survivalexpectations.do	surv_exp_all.eps
Figure F2	Figures_auxiliary.do	auxiliary2_zeroinc_sp.eps
Figure G1	Figure_simulations.do	impactreform_detailed_onlypart_DHSwages.pdf
Figure G2	FigureG2_simulations.do	data_simulations_simulationsample_allmoments.eps
Figure I1	LT_survival_HealthBehavior.do	mortality_healthbehavior_literature_both.eps
Figure I2	Figure_sim_data_incl_HB.do	data_simulations_incl_HB.eps
Figure J1	Figure_sim_data_differentprefs_estsample.do	model_fit_postupdate_hetexp_subjexp.eps
Figure J2	Simulated_moments_by_tertiles_different_prefs.do	exp_tertiles_estsample_postupdate_twoprefs1.pdf, exp_tertiles_estsample_postupdate_twoprefs2.pdf
Table C1	SEs_for_covariance_matrix.do	
Table C2	Regressions_indeffects_backgroundinfo.do	
Table F1	sample_selection.do	
Table F2	Wage_process.do	
Table F3	Descriptives_initial_conditions.do	
Table H1	See section 2.3: Retirement model	
Table I1	See section 2.3: Retirement model	
Table I2	Retage_in_simulations.do	
c. Online Appendices		
Figure OA1	Figure_taxes.do	taxes_estsample.eps
Figure OA2	Figure_taxes.do	taxes2.eps