Replication Package

Educational Ambition, Marital Sorting, and Inequality

Frederik Almar, Benjamin Friedrich, Ana Reynoso, Bastian Schulz and Rune Vejlin

The programs in this replication package generate all empirical results contained in the paper and the online appenix. They rely on restricted-access data provided by Statistics Denmark, which are accessible through ECONAU, the data resource for register-based research at the Department of Economics and Business Economics, Aarhus University. The procedure for gaining access to the data for replication purposes is outlined in the Data Availability section of the manuscript. All programs in the replication package use STATA code. The merging and exporting of the raw register data files is done in SAS.

The first part of the replication package (Data) generates datasets from the underlying register data. The code assumes that the folder /Data/Core_datasets contains eight merged *core* datasets in STATA format. The SAS code that generates them from the raw registers is archived in subfolders the ECONAU servers. The do-file /Data/master_data.do calls all do-files that are used in step one to transform the 8 core datasets into 18 derived analysis datasets, which are used to produce our results, see details below. The second part of the replication package (Results) contains the code that generates all results contained in the paper and the online appendix based on the analysis datasets. The code assumes that the analysis datasets, generated in step one, are saved in the folder /Data. The do-file /Results/master.do calls all do-files that use the analysis datasets to generate our results, see details below.

Details on Part 1 (Data)

The following core data sets, archived on the server, form the foundation of all data processing:

1. `sample_04_thesis` – Primary sample dataset (includes field of study information)
2. `sample_04_thesis_extended` – Extended version of primary sample
3. `ras_ftpt` – Full-time/part-time status data
4. `training_variable` – Manager training information
5. `main_careers_kids` – First child information
6. `wealth_from_SAS` – Wealth data
7. `aku_ras` – Labor force survey data
8. `disp_inc_data` – Disposable income data

Processing of the core data:

Based on these core data sets, the do-files included in the subfolder of the /Data folder combine and process the data to generate the analysis datasets as follows. All steps can be called from /Data/master_data.do.

Processing of Primary Samples

- `regions` ← `sample_04_thesis`
- `sample1_new` ← `sample_04_thesis` + `regions`
- `sample1_new_extended` ← `sample_04_thesis_extended`
- `field_of_study` ← `sample_04_thesis`

Ambition and Regional Data

- `kmeans ambition types` ← `sample1_new`
- `kmeans ambition types extended` ← `sample1_new_extended`

Ambition Measures (Life Cycle)

- `life_wage_ambition_final` ← `sample1_new`
- `life_earnings_ambition_final` ← `sample1_new`
- `wealth_ambition_final` ← `sample1_new` + `wealth_from_SAS`

Additional Variables (Part-time penalty, management training, first child)

- `Goldin_quotients_for_all` ← `sample1_new` + `ras_ftpt`
- `manager_training` ← `training_variable`
- `first_child` ← `main_careers_kids`

Datasets with Field of Study Information

- `dataset_with_singles` ← `sample1_new` + `field_of_study`
- `dataset_with_types` ← `sample1_new` + `field_of_study`
- `dataset_with_singles_extended` ← `sample1_new_extended`
- `post_secondary_fields_full_population` ← `kmeans ambition types` + `field_of_study`

Full List of Datasets

1. start_sample ← kmeans ambition types + post_secondary_fields_full_population
2. before_reduction ← multiple sources:
    - start_sample
    - Goldin_quotients_for_all
    - manager_training
    - first_child
    - life_wage_ambition_final
    - life_earnings_ambition_final
    - wealth_ambition_final
    - ras_ftpt
3. dataset_with_singles
4. dataset_with_singles_extended
5. dataset_with_types
6. field_of_study
7. first_child
8. Goldin_quotients_for_all
9. kmeans ambition types
10. kmeans ambition types extended
11. life_wage_ambition_final
12. life_earnings_ambition_final
13. manager_training
14. post_secondary_fields_full_population
15. regions
16. sample1_new
17. sample1_new_extended
18. wealth_ambition_final

Key Processing Steps

1. Sample Creation: Core sample extracted from `sample_04_thesis` to create `sample1_new`
2. Ambition Types: K-means clustering applied to create ambition type classifications
3. Field of Study: Post-secondary field information merged with ambition types
4. Variable Construction: Individual outcome variables (wages, earnings, wealth, first child, etc.) constructed separately

Notes for Replication

- Generation folders (`Gen_*`) contain the code necessary for creating each derived dataset.
- The dependency structure must be followed in sequence for successful replication.
- Some datasets appear in multiple dependency chains (e.g., `sample1_new`, `field_of_study`).

Details on Part 2 (Results)

Main Results

Step 1: Figure 1 & Figure A1a (run `Cloud plot full population.do`)

- Generates cloud plots.

Step 2: Figure 2 & Figure B4 (run `Create results.do`, `Robust sorting results.do`)

- Generates sorting trend plots.

Step 3: Figure 3 & Table 1 (run `Decompositions.do`, `Outputs.do`)

- Runs sorting and decomposition analysis.

Online Appendix

Step 4: Table A1 & Table A2 (run `Cross tables with full pop fields.do`)

- Creates descriptive tables.

Step 5: Figure A1 & Figure A2 (b+c+d) (run `LFS_clustering_prep_BF_dropmissing.do`, `Dataset with singles.do`, `Dataset with singles primeage.do`, `Cloud_new_primeage.do`)

- Prepares LFS clustering (drop missing values).
- Generates cloud plots for hours flexibility.

Step 6: Figure A2 (e+f) (run `Get life-time earnings updated.do`, `Get types 1990 2010 updated.do`, `Cloud plot with life earnings draft 50plus.do`)

- Generate clouds plot with lifetime earnings.

Step 7: Table A3 (run `Get full sample.do`, `Get life-time earnings.do`, `Run regressions final.do`)

▷ *Requires: Figure A1 code completed first*

- Regression analysis.

Step 8: Figure B1, B2 & B5 (run `Create results.do`, `Create figures.do`)

- Generates type-specific likelihood ratios and marginal distribution figures.

Step 9: Figure B3 (run `Single shares.do`)

▷ *Requires: Figure B1 code completed first*

- Generates single share figures.

Step 10: Table B1 (Robustness — 5 Scenarios)

▷ *Requires: Figure B1, Figure 3, and Figure A1 completed first*

- Runs sorting and decomposition analysis for Hours-based and Lifetime-Earnings-based Classifications, 5 different scenarios:
  - Scenario 1: Prime age long (run `Create results for fig 3 primeage long.do`, `Decompositions_eq_fig_4_primeage_long.do`)
  - Scenario 2: Prime age long baseline (run `Create results for fig 3 primeage long baseline.do`, `Decompositions_eq_fig_4_primeage_long_baseline.do`)
  - Scenario 3: Lifetime earnings 1990–2010 (run `Get life-time earnings.do`, `Get types 1990 2010.do`, `Create results lifetime 1990 2010.do`, `Decompositions lifetime 1990 2010_eq.do`)
  - Scenario 4: Ambition 1990–2010 (version 1) (run `Create results ambition 1990 2010.do`, `Decompositions ambition 1990 2010_eq.do`)
  - Scenario 5: Ambition 1990–2010 (version 2) (run `Create results ambition 1990 2010.do`, `Decompositions ambition 1990 2010_eq.do`)
- Collects and tabulates results (run `Collect results_eq_appendix.do`, `Table 5 lifetime addon_eq.do`)

Step 11: Table B2 (run `Decompositions.do`)

- Runs counterfactual Scenarios (ii) and (iii) based on Gutierrez (2020).

Step 12: Table B3 (Extensive Robustness — 14 Scenarios)

▷ *Requires: Figure B1 and Figure 3 completed first*

- Runs sorting and decomposition analysis for 14 different robustness scenarios:
  - Scenario 1 (run `Dataset with singles.do`, `Create results.do`, `Decompositions_eq.do`)
  - Scenario 2 (run `Men - sample 1.do`, `Men - sample 2.do`, `Women - sample 1.do`, `Women - sample 2.do`, `Joining couples.do`, `Create results.do`, `Decompositions_eq.do`)
  - Scenario 3 (run `First decade - sample 1.do`, `First decade - sample 2.do`, `Second decade - sample 1.do`, `Second decade - sample 2.do`, `Third decade - sample 1.do`, `Third decade - sample 2.do`, `Joining decades.do`, `Create results.do`, `Decompositions_eq.do`)
  - Scenario 4 (run `Create results.do`, `Decompositions_eq.do`)
  - Scenario 5 (run `Create_3_and_5_types.do`, `Create results.do`, `Decompositions_eq.do`)
  - Scenario 6 (run `Create results.do`, `Decompositions_eq.do`)
  - Scenario 7 (run `Dataset with singles.do`, `Create results.do`, `Decompositions_eq.do`)
  - Scenario 8 (run `Dataset with singles.do`, `Create results.do`, `Decompositions_eq.do`)
  - Scenario 9 (run `Create types.do`, `Create results.do`, `Decompositions_eq.do`)
  - Scenario 10 (run `Dataset with singles.do`, `Create results.do`, `Decompositions_eq.do`)
  - Scenario 11 (run `Create results.do`, `Decompositions_eq.do`)
  - Scenario 12 (run `Dataset with singles.do`, `Create results.do`, `Decompositions_eq.do`)
  - Scenario 13 (run `Programs in low low cluster.do`, `Create results.do`, `Decompositions_eq.do`)
  - Scenario 14 (run `Dataset with singles.do`, `Create results.do`, `Decompositions_eq.do`)
- Collect and tabulate all robustness results (run `Collect robustness_eq.do`, `Table 5_eq.do`, `Collect robustness_eq_2.do`, `Collect robustness_eq_3.do`, `Collect results.do`)

Step 13: Figure B6 (run `Cloud plot.do`)

▷ *Requires: Table B3 completed first*

- Generates cloud plots by gender.

Step 14: Figure B7 (run `Cloud plot.do`)

▷ *Requires: Table B3 completed first*

- Generates cloud plost by decade.

Step 15: Figure B8 (run `Cloud plot full population by institution.do`)

- Generates cloud plot for Main Degree Programs by University.