contributors: - Peter Arcidiacono - Esteban Aucejo - Arnaud Maurel - Tyler Ransom

README

This repository contains all computer code and data files to completely reproduce the results in "College Attrition and the Dynamics of Information Revelation" by Arcidiacono, Aucejo, Maurel and Ransom.

Overview

The code in this replication package constructs the analysis file from the three data sources (NLSY, CPS, SIPP) using Stata. Our five-stage structural estimation procedure uses Matlab. We assess model fit and counterfactual simulations using Matlab and Julia. Six main files run all of the code to generate the data for the 1 figure and 35 tables in the paper. The file `src/master.bash` contains all computational steps in a single bash script. The replicator should expect the code to take more than two weeks to run. The vast majority of this time is spent on the counterfactuals; estimation takes about 1 day.

The source code contains approximately 80,000 lines of code.

Data Availability and Provenance Statements

Statement about Rights

- [x] I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- [x] I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permission are documented in the `data/LICENSE.txt` file.

License for Data

The data are licensed under a GPL license where available. See `data/LICENSE.txt` for details.

Summary of Availability

- [x] All data **are** publicly available.
- [ ] Some data **cannot be made** publicly available.
- [ ] **No data can be made** publicly available.

Details on each Data Source

NPSAS 2008

Data from the 2008 National Postsecondary Student Aid Study were used to obtain auxiliary regression coefficients. Since these data are confidential, we accessed them through the National Center for Education Statistics (NCES) DataLab platform at https://nces.ed.gov/datalab/.

We used this platform to obtain estimates of grant and loan receipts by individual characteristics and type of college attended (2-year vs. 4-year), as well as overall average tuition by type of college.

NLSY97

Data from the National Longitudinal Survey of Youth 1997 (NLSY97) were downloaded from the NLS Investigator website hosted by the U.S. Bureau of Labor Statistics. The URL is https://www.nlsinfo.org/investigator/pages/login. We provide the variable extracts required to recreate the data used in this archive. The data are in the public domain.

Extract files: `data/nlsy97/raw/*.NLSY97` Datafiles: `data/nlsy97/raw/*.dct` (not provided)

NLSY97 ASVAB Equipercentile Data

Data on equipercentile mappings for ASVAB tests in the NLSY79 and NLSY97 were downloaded from Fabian Lange's website (http://www.fabianlange.ca/data.html). Lange's files provide raw data on AFQT scores from both the NLSY79 and NLSY97. We extend this approach to six components of the ASVAB, as well as math and verbal components. These additional data and programs are included here and are not publicly available anywhere else.

Datafiles: `data/nlsy97/raw/AFQT_MATCHING*/`

SIPP

Data from the Survey of Income and Program Participation (SIPP) were obtained via the National Bureau of Economic Research (NBER) at the following URL: https://www.nber.org/research/data/survey-income-and-program-participation-sipp. We provide a bash script for downloading the data; see directly below. The data and code used to prepare it are licensed by NBER under the GNU General Public license. Note that the SIPP data could in principle be obtained directly from the U.S. Census Bureau. In that case, the data would be in the public domain.

Bash script to download data files: `data/sipp/downloader.sh` Datafiles: `data/sipp/raw/*.dct` (not provided)

CPS

The paper uses Current Population Survey (CPS) data via IPUMS CPS (Flood et al., 2022). IPUMS-CPS data does not allow for redistribution, except for the purpose of replication archives. We have not obtained the requisite permissions, but the data can be downloaded by creating an account on https://cps.ipums.org/cps/. Login, click "Get Data" and then click "Select Data." Click "Select Samples" and then choose years 2007-2014 in the "ASEC" tab and then **deselect** all pre-made selections in the "Basic Montly" tab. Click "Submit Sample Selections" which will bring you back to themain screen. On the main screen, search for or select the following variables:

- `age`
- `sex`
- `race`
- `educ`
- `educ99`
- `occ2010`
- `ind`

Then click "View Cart" which will allow you to create the extract. Note that your extract may have a different number than the file we use (see directly below).

Note that the CPS data could in principle be obtained directly from the U.S. Census Bureau. In that case, the data would be in the public domain.

Datafile: `data/cps/raw/cps_00018.dat.gz` (not provided)

Dataset list

Data file	Source	Notes	Provided
`data/nlsy97/cleaned/y97_all_tscrGPA.dta`	All listed	Combines NLSY, SIPP, CPS and serves as input for code in `src/descriptives/`. File is 2 GB.	Yes
`data/nlsy97/cleaned/wide_data_male20220401_tscrGPA.mat`	All listed	Alternate form of data described in previous row. Serves as input for all structural model estimation code (`src/estimation-stage*/`)	Yes

Computational requirements

Software Requirements

- Stata (code was last run with version 17)
  - `sutex` (version 04 Sep 2001)
  - `tabout` (version 2.0.42)
  - `spell` (version 1.7.1)
  - `egenmore` (version 2019-01-24)
  - `estout` (version 3.23)
  - `texdoc` (version 2.4.0)
  - `grstyle` (version 1.1.1)
  - the program "`src/config_stata.do`" will install all dependencies locally, and should be run once.
- Matlab (code was run with Matlab Release 2022a)
- Julia (code was last run with version 1.6.1)
  - the program "`src/config_julia.jl`" will install all dependencies locally, and should be run once.

Portions of the code use bash scripting, which requires Linux. The code was executed on a high-performance computing (HPC) cluster running Red Hat Enterprise Linux 8, kernel `4.18.0-408.el8.x86_64`.

Controlled Randomness

The random seed is used in many different Matlab and Julia programs and is set at the beginning of all programs that require it.

Memory and Runtime Requirements

Summary

Approximate time needed to reproduce the analyses on a standard (CURRENT YEAR) desktop 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

Some steps of the code (data preparation, descriptive statistics, post-simulation analysis) can be run on a laptop machine with 32 GB of RAM.

Executing the estimation and simulation code requires an HPC cluster. These portions of the code were last run on a HPC with 175 available nodes each allowing 50 GB of RAM. The project requires about 200 GB of hard drive space.

A rough breakdown of resource requirements (note that computational time can be traded off with number of nodes used):

Computation step	Time	Memory per node	Cluster nodes
Data prep	2 hours	<10 GB	single
Descriptive analysis	minutes	<10 GB	single
Flexible first-stage estimation	2 hours	20 GB	1
Structural learning model	4 hours	20 GB	1
Structural choice model	30 hours	50 GB	20
Forward simulation of choice model	16 hours	50 GB	single
Counterfactual simulations of choice model	2-3 weeks	15 GB	175
Parametric bootstrap of model parameters	2.5 days	30 GB	151
Post-simulation analysis	minutes	<10 GB	single

Description of programs/code

The file `src/master.bash` can be run from start-to-finish to reproduce all computation.

- Programs in `src/data-cleaning` will extract and reformat all datasets referenced above. See the first 93 lines of `src/master.bash`. These programs create some of the tables in Appendix A.
- The lone program in `src/descriptives` (`descriptives.do`) will generate the first five tables of the paper. This program creates `.tex` or `.csv` files that are stored in `exhibits/tables`.
- Programs in `estimation-stage*` will estimate the structural model. No tables are generated.
- Programs in `simulations-forward-fit` will simulate the model to assess model fit. No tables are generated at this stage.
- Programs in `simulations-counterfactual` will simulate the model under various counterfactuals. No tables are generated at this stage.
- Programs in `parametric-bootstrap` will compute standard errors of all estimates by parameteric bootstrap. The file `estimates_tables.m` in this folder will generate all of the tables in the paper that contain parameter estimates of the structural model. This includes many of the tables in Appendix B as well as many of the tables in the body of the paper.
- Programs in `sim-results` conduct post-simulation analysis of the actual data and its simulated counterparts. `model-fit.do` assesses the fit of the model and produces Figure 1 and model fit tables in Appendix B. `cfl.do` assesses the impact of the counterfactuals and produces the last set of tables in the paper as well as some additional tables in Appendix B.
- Custom Matlab functions are stored in a `functions` subfolder of each `estimation-stage*` or `simulations-forward-fit` folder. Custom Julia functions are contained in a single `.jl` file, `allfuns.jl` in each subfolder corresponding to a separate counterfactual simulation.
- The program `src/config_stata.do` will install and update (if necessary) required custom Stata commands available from SSC. `src/config_julia.jl` does the same for Julia packages.
- The random seed is set at the top of each program that requires it. You should obtain identical estimates regardless of order of execution.

License for Code

The code is licensed under an MIT license. See LICENSE.txt for details.

Instructions to Replicators

- Run `src/master.bash` to run all steps in sequence.

Details

- `src/config_stata.do`: installs needed ado packages
- `src/config_julia.jl`: installs needed Julia packages
- Data cleaning:
  - `src/data-cleaning/cps/IPUMS_CPS.do`: cleans the CPS data
  - `src/data-cleaning/sipp/create_sipp.do`: cleans the SIPP data
  - `src/data-cleaning/nlsy97/y97_import_all.do`: reads in the *.dct files obtained from NLS investigator
  - `src/data-cleaning/nlsy97/y97_create_master.do`: manipulates the imported data to create many different variables of interest
  - `src/data-cleaning/nlsy97/y97_create_trim.do`: final file to clean the NLSY data and bring in information from CPS and SIPP
  - `src/data-cleaning/nlsy97/y97_create_matlab_data_tscrGPA.do`: reshapes the data and only include necessary NLSY variables, for ease of

use in Matlab
    - `src/data-cleaning/nlsy97/data_import_20220401_tscrGPA.m` : imports the reshaped data and saves as Matlab data format
- Descriptive analysis:
    - `src/descriptives/descriptives.do` : analyzes the Stata data (created at the end of `src/data-cleaning/nlsy97/y97_create_matlab_data_tscrGPA.do`) and generate tables of summary statistics used in the paper
- Structural estimation
    - `src/estimation-stage1/runmodel_all_stage1_interact_type.m` : estimates the flexible parametric model used to pin down the unobserved types as well as the missing GPA and missing college maojrs using an EM algorithm
    - `src/estimation-stage2-learning/runmodel_learning_only.m` : estimates parameters of the learning model (using an EM algorithm), as well as the parameters of the graduation logit (by maximum likelihood)
    - `src/estimation-stage4-gridsearch/runmodel_consump_jointsearchfrictions_WCabsorb.m` : estimates (by maximum likelihood) the white collar offer arrival rates, the wage AR(1) parameters, and the parameters used to construct the Conditional Choice Probabilities (CCPs) for the future value terms for many different values of the CRRA parameter
    - `src/estimation-stage3-5-structural/runmodel_consump_jointsearchfrictions_WCabsorb.m` : estimates (by maximum likelihood) the white collar offer arrival rates, the wage AR(1) parameters, and the parameters used to construct the Conditional Choice Probabilities (CCPs) for the future value terms
    - `src/estimation-stage3-5-structural/computeFVfastParallel.m` : computes the future value terms given the parameters estimated in the previous bullet
    - `src/estimation-stage3-5-structural/runmodel_consump_structural_FVfast.m` : estimates (by maximum likelihood) the structural flow utility parameters given the future value terms computed in the previous bullet
- Parametric bootstrap
    - `src/parametric-bootstrap/parboot.m` : re-estimates the entire model by generating the outcomes of the model using the initial conditions in the data and the estimated model parameters
- Postestimation forward simulation
    - `src/simulations-forward-fit/simulatorMCintEstData10D.m` : simulates the model forward given initial conditions and the parameter estimates
- Counterfactual simulations
    - `src/simulations-counterfactual/baseline/shellsimbatch.jl` : executes the counterfactual simulation of the model where individuals have perfect information about their abilities
    - `src/simulations-counterfactual/no-frictions/shellsimbatch.jl` : executes the counterfactual simulation of the model where individuals have perfect information about their abilities **and** can always choose the white collar occupation
    - `src/simulations-counterfactual/no-cred-cons/shellsimbatch.jl` : executes an unreported counterfactual simulation of the model where individuals have perfect information about their abilities **and** student loans are set to $0 and in-college consumption is set to the 75th percentile for all individuals
- Analysis of simulated data
    - `src/sim-results/model-fit.do` : produces figures and tables illustrating the fit of the model
    - `src/sim-results/cfl.do` : produces tables illustrating the effects of various counterfactual policies

List of tables and programs

The provided code reproduces:

- [x] All numbers provided in text in the paper
- [x] All tables and figures in the paper
- [ ] Selected tables and figures in the paper, as explained and justified below.

Figure/Table #	Program (`src/`)	Line Number	Output file (`exhibits/tables`)	Note
Table 1	`descriptives/descriptives.do`	374	`table_1.tex`	
Table 2	`descriptives/descriptives.do`	524	`table_2.tex`	
Table 3	`descriptives/descriptives.do`	865	`table_3*.tex`	Table 3 is generated from two separate `.tex` files
Table 4	`descriptives/descriptives.do`	982	`table_4*.tex`	Table 4 is generated from two separate `.tex` files
Table 5	`descriptives/descriptives.do`	677	`table_5.tex`	
Table 6	`parametric-bootstrap/estimates_tables.m`	134	`gpa_eqn_estimates.tex`	
Table 7	`parametric-bootstrap/estimates_tables.m`	178	`wage_eqn_estimates.tex`	
Table 8	`parametric-bootstrap/estimates_tables.m`	359	`correlation_matrix.tex`	
Table 9	`parametric-bootstrap/estimates_tables.m`	317	`idiosyncratic_vars.tex`	
Table 10	`parametric-bootstrap/estimates_tables.m`	463	`util_est_matrix.tex`	
Table 11	`sim-results/cfl.do`	1014	`table-sorting-abil-fwd.tex`	

Figure/Table #	Program (src/)	Line Number	Output file (exhibits/tables)	Note
Table 12	sim-results/cfl.do	715	table-comp-status-cfl.tex	
Table 13	sim-results/cfl.do	903	table-comp-status-cfl-inc	
Table 14	sim-results/cfl.do	1014	table-sorting-abil-cfl.tex	
Table 15	sim-results/cfl.do	1648	table_wage_decomp.tex	
Table A1				Created by hand
Table A2	descriptives/descriptives.do	686	None	
Table A3	data-cleaning/nlsy97/y97_create_trim_tscrGPA.do	356	wageAppendix_tscrGPA.tex	
Table A4	data-cleaning/nlsy97/y97_create_trim_tscrGPA.do	339	dataAppendix_tscrGPA.tex	
Table B1				Created by hand
Table B2				Created by hand
Table B3	parametric-bootstrap/estimates_tables.m	695	types_all_eqns.tex	
Table B4	parametric-bootstrap/estimates_tables.m	618	type_mass_probabilities.tex	
Table B5	parametric-bootstrap/estimates_tables.m	779	schabil_meas_sys.tex	
Table B6	parametric-bootstrap/estimates_tables.m	892	schabilpref_meas_sys.tex	
Table B7	parametric-bootstrap/estimates_tables.m	984	workabilpref_meas_sys.tex	
Table B8	parametric-bootstrap/estimates_tables.m	178	wage_AR1_estimates.tex	
Table B9	parametric-bootstrap/estimates_tables.m	282	grad_logit_estimates.tex	
Table B10	parametric-bootstrap/estimates_tables.m	241	offer_arrival_estimates.tex	
Table B11	parametric-bootstrap/estimates_tables.m	590	static_util_est_matrix.tex	
Table B12	sim-results/model-fit.do	558	table-fit-choice-shrs.tex	
Table B13	sim-results/model-fit.do	579	table-fit-choice-shrs-grad.tex	
Table B14	sim-results/cfl.do	1692	table_abil_sort_appendix.tex	
Table B15	sim-results/cfl.do	1117	table-sorting-postvar-fwd.tex	
Table B16	sim-results/cfl.do	1014	table-sorting-abil-cflnofric.tex	
Figure 1	sim-results/model-fit.do	397	exhibits/figures/modelFitByT.eps	

References

Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren and Michael Westberry. Integrated Public Use Microdata Series, Current Population Survey: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2022. https://doi.org/10.18128/D030.V10.0