Overview of Files for .\FINAL_Replication_CLMP_JPE

This folder and its contents include (almost) everything needed for replicating the analysis of "Child skill production: Accounting for parental and market-based time and goods investments" by Caucutt, Lochner, Mullins, and Park. Due to PSID Terms of Use, all PSID data (and any data files derived from PSID) are hosted at the ICPSR PSID Repository:

DOI: https://doi.org/10.3886/E236003V1

These data will need to be downloaded separately and saved in .\data\data_PSID or .\data\data_derived as instructed below. We maintain the same file structure within .\data in the ICPSR PSID Repository and in this JPE Harvard Dataverse Replication package .\data folder.

Folder Structure

This folder contains the following subfolders:

- .\code – contains all code for the project in the following subfolders:
  - .\counterfactual – all code used for counterfactual analysis
    - .\data – Stata code to organize PSID data files for MATLAB
    - .\constrained – MATLAB code for the case where all households are borrowing constrained
    - .\unconstrained – MATLAB code for the case where no households are borrowing constrained
  - .\data_cleaning – all code used to clean and organize data for analysis
    - .\pred_wage_instrument – Stata code used to create predicted log wage instrument from Census 2000 data
    - .\prices – Stata code used to clean, merge, and organize data in childcare and goods/services investment input prices by state and year
    - .\psid – code to clean raw PSID data and create a child-level panel (ChildPanelCDS) and mother-level panel (MotherPanelCDS).
  - .\estimation – all code used to create summary statistics and for estimation
    - .\basics_rel_demand – Stata code used to create summary figures and tables, as well as perform relative demand estimation
    - .\gmm_estimation – Julia code to perform gmm estimation and perform monte-carlo simulations. The subdirectory .\gmm_estimation\scripts contains the scripts to reproduce estimation results in the paper, while .\gmm_estimation\src contains all source code.
- .\data – contains all data needed for the project (temporary data files stored elsewhere)
  - .\data_Census – extracted IPUMS 2000 Census data used to create predicted log wage instrument
  - .\data_ChildCareAware – Child Care Aware data on childcare costs
  - .\data_CPI_RPP – data from CPI and RPP on prices of goods and services by state and year

- .\data_CPS – extracted monthly CPS data used to calculate average wages for childcare workers by state-year

- .\data_PSID – all extracted data from PSID and PSID-CDS (can be downloaded from PSID Repository and saved in this folder)

- .\data_derived – contains data files used in main empirical analysis (this folder contains all the data required to execute Steps 3-5 below)
    - p_extendedcosts.dta/csv

    - predicted_lnwage3.dta

    - ChildPanelCDS.dta – can be downloaded from PSID Repository or produced by executing all code in .\code\data_cleaning (see Step 2 below)

    - MotherPanelCDS.dta – can be downloaded from PSID Repository or produced by executing all code in .\code\data_cleaning (see Step 2 below)

    - psid_fam.dta/csv – can be downloaded from PSID Repository or produced by executing all code in .\code\data_cleaning (see Step 2 below)

    - wage_types.csv – can be downloaded from PSID Repository or produced by executing all code in .\code\data_cleaning (see Step 2 below)

- .\output – contains all output tables and figures for the paper and appendix
    - .\figures – contains all figures

    - .\tables – contains all tables

    - OrganizeFigsTexTables.tex – a LaTex file the can be used to organize all the .tex tables in .\tables and figures in .\figures so they are in order as presented in the paper and appendix. (Note: the .tex files were heavily edited for the paper.)

Step-by-step instructions for executing code

1. Download .\FINAL_Replication_CLMP_JPE replication package from JPE's Harvard Dataverse. Also, download any needed PSID data from the ICPSR PSID Repository https://doi.org/10.3886/E236003V1

2. Data cleaning (steps a-c are independent and can be done in any order):
    a. Execute Stata code to create predicted log wage instrumental variable from 2000 Census: .\code\data_cleaning\pred_wage_instrument\main.do
        i. See README_pred_wage_instrument.pdf in that same folder for additional details.

ii. This produces Stata data set .\data\data_derived\predicted_lnwage3.dta, which is used in empirical analysis. If the researcher is not interested in re-creating this file (provided in replication package), this step can be skipped.

b. Execute Stata code to combine and impute childcare and goods/services input prices from data provided by Child Care Aware and CPI/RPP, respectively: .\code\data_cleaning\prices\main.do
   i. See README_Prices.pdf in that same folder for additional details.
   ii. This produces Stata/Excel data set .\data\data_derived\p_extendedcosts.dta/csv, which is used in empirical analysis. If the researcher is not interested in re-creating this file (provided in replication package), this step can be skipped.

c. Execute .\code\data_cleaning\psid\clean_psid_data.R to clean and arrange raw PSID data and combine it into a panel of PSID mothers (MotherPanelCDS) and children (ChildPanelCDS).
   i. First ensure that PSID Repository data (in folder .\data\data_PSID) is downloaded and saved in .\data\data_PSID.
   ii. This script assumes that .\code\data_cleaning\psid is the working directory.
   iii. For additional information, clean_psid_data.qmd combines code with descriptions of what the code is doing in notebook format, and clean_psid_data.html formats this notebook in html with a table of contents for easier navigation. If the researcher is not interested in creating these files, they are also provided in the PSID data repository.

3. Create basic summary tables/figures and perform relative demand estimation using Stata:
   a. Execute Stata code .\code\estimation\main.do
      i. See README_basics_rel_demand.pdf in that same folder for additional details.
      ii. This uses data sets created in steps 1a, 1b, and 1c. Data produced in steps 1a and 1b are provided in the replication package (in .\data\data_derived), but data used and produced in step 1c must be downloaded from the associated PSID Data Repository. If step 1c is to be skipped, the researcher must download ChildPanelCDS.dta and MotherPanelCDS.dta from the PSID Data Repository (in .\data\data_derived), saving these data files in .\data\data_derived.
      iii. This code produces Stata/.csv data set psid_fam.dta/csv, which is used in both GMM estimation (Step 4) and counterfactual analysis (Step 5)
      iv. This code produces all figures for the paper and appendix, saving them in .\output\figures.
      v. This code produces results for Tables 1-4 in the paper, as well as Tables E-1 and Tables G-1 to G-16. These results are saved in (crudely formatted) LaTex .tex files within .\output\tables. Other related results also can be found in the associated Stata log files in subfolder .\code\estimation\basics_rel_demand\logs

4. Perform the GMM estimation, direct estimation, and Monte Carlo exercises in Julia.
   a. Set .\code\estimation\gmm_estimation as the working directory and execute .\code\estimation\gmm_estimation\scripts\run_estimation.jl
      i. See README_gmm_estimation.pdf for additional details.

ii. This script first performs the clustering algorithm using the panel MotherPanelCDS.csv, and saves the classifications to .\data\data_derived\wage_types.csv.
iii. Next, it runs all gmm specifications described in the paper and creates Table 5, along with Tables G17-G23 in the paper.

b. In the same directory, run scripts\direct_estimation.jl to perform the direct estimation exercise and create the output used in Table G-24.
c. In the same directory, run scripts\monte_carlo.jl to conduct the Monte Carlo exercise that evaluates direct estimation and create output for Tables F-1 and F-2. Note: on line 145, monte_carlo.jl fixes the seed for random number generation, which ensures that the same answer will produced every time it is run on the same machine, but does not ensure that the answer remains the same across different machines.
d. The file scripts\run_estimation_alt_clustering.jl does not produce any output reported in the main paper, but verifies the claim in Appendix E, footnote 10 that results are not sensitive to the inclusion of additional controls in the clustering routine.

5. Counterfactual analysis
    a. Execute Stata code .\code\counterfactual\data\main.do in that directory
        i. This code reads the PSID datasets psid_fam.csv and wage_types.csv from .\data\data_derived, processes the variables, and saves them in a format suitable for use in MATLAB
    b. Execute MATLAB code .\code\counterfactual\constrained\source\main.m
        i. This code produces results for Tables 6, 7, D1, D2, D3, and D7. These results are saved in .csv files within .\output\tables.
    c. Execute MATLAB code .\code\counterfactual\unconstrained\source\main.m
        i. This code produces results for Tables D4 to D6. These results are saved in .csv files within .\output\tables.

6. Additional files
    a. The LaTex file .\output\OrganizeFigsTexTables.tex can be run to produce pdf document .\output\OrganizeFigsTex\Tables.pdf, which organizes all figures and .tex tables produced in Step 3 to match the order in the paper and appendix. (Note that many .tex files require considerable additional "cleaning" to look like those in the paper/appendix.)

Additional information

1. Software requirements:
    a. Stata – code last run on Stata/MP 16.1
    b. Matlab – code last run on Matlab Release 2015b
    c. R – code last run in R 4.5.1
    d. Julia – code last run in Julia 1.11

2. Storage requirements:
    a. Final data for analysis in Steps 3-5 above requires less than 250 MB.
    b. Preliminary data used in Steps 1 and 2 above requires 7-10 GB.