

Data and Programs for “*Rising Inequality: Transitory or Persistent? New Evidence from a Panel of U.S. Tax Returns*” by Jason DeBacker, Bradley Heim, Vasia Panousi, Shanthi Ramnath, and Ivan Vidangos

This folder (and subfolders) contains data and programs used to generate the results in “*Rising Inequality: Transitory or Persistent? New Evidence from a Panel of U.S. Tax Returns*” by Jason DeBacker, Bradley Heim, Vasia Panousi, Shanthi Ramnath, and Ivan Vidangos.

The main folder contains the following subfolders (a more detailed description of the contents of each subfolder is provided subsequently):

- Data : Contains the covariance data constructed from the panel of tax returns and used in the estimation of error components models (ECMs) and computation of ECM-based variance decompositions. (The raw tax data cannot be shared due to confidentiality restrictions).
- DataSetup : Contains the programs that process, organize, and clean up the tax data.
- DescriptiveAnalysis : Contains the programs that compute descriptive statistics and perform the nonparametric analysis of the tax data.
- ErrorComponentsModels : Contains the programs that estimate the ECMs and the programs that compute the ECM-based variance decompositions.
- PSID : Contains data and programs for the analysis using PSID data presented in the online appendix.

Note: Programs will need to be edited by the user to set the correct paths and directory names to match the user’s computing system.

Note on file names:

Various programs and data files are named using shorthand names for the various income measures and samples used in the paper (see the paper for precise definitions of each variable and sample). The shorthand names are:

- labinc = male labor earnings
- hhinc = pre-tax household income (on sample of all households)
- hhincat = after-tax household income (on sample of all households)
- hhinc_labincsample = pre-tax household income on male-headed households sample
- hhincat_labincsample = after-tax household income on male-headed households sample
- income1 = household labor earnings
- income2 = income1 + transfers and retirement income
- income3 = income2 + investment income
- mar = married filers only
- fixedfamilycomp = treat each change in family composition as a different filing unit

- LT = use a minimum earnings threshold of half the baseline
- HT2 = use a minimum earnings threshold twice that of the baseline

Detailed description of folders and subfolders

The following provides a more detailed description of the contents of each of the folders listed above.

Data

This folder contains two subfolders:

- Subfolder *covariance* contains various Stata datasets, each containing the covariance data (i.e. the empirical moments) for the various income measures and samples. The covariance data are used in the estimation of the ECMs.
 - The key variables in each covariance dataset are:
 - yrvar = year
 - hvar = age (normalized so that 1 corresponds to age 25)
 - jvar = lead
 - covvar = empirical covariance
 - wgtvar = empirical weights (number of observations used to compute the corresponding empirical covariance)
 - As noted above, the raw tax data cannot be shared because of confidentiality restrictions. The covariance data included here contains no personal identifying information.
- Subfolder *weights* contains various Stata datasets, each containing weights representing the empirical age distribution in each calendar year for the corresponding income measure and sample. These weights are used in the construction of the ECM-based variance decompositions.

DataSetup

This folder contains the programs that set up the samples used in the analysis of the tax data.

- The files INSOLE79_86full.sas, INSOLE87_96full.sas, INSOLE97_09full1.sas, INSOLE97_09full2.sas, INSOLE97_09full3.sas, panel1979_2009full.sas pull the raw tax data (annual SOI individual sample files, corresponding W2 files, and SSA files for DOB and gender). Each year's INSOLE files are appended, variable names are changed, variables for analysis are created (e.g. labinc, hhinc – as described above), and merges between tax return data, W2s, and SSA files are carried out.

- `panel_BPEA_1987_2009.sas`: uses INSOLE files previously appended (in files described above) and keeps only years and variables of interest.
- `insole87_09_inequality.do`: gets a few additional variables from the INSOLE files missing in the above file.
- `inequality_data_setup.do`: creates the subsamples used for analysis in the paper, by making restrictions on age, gender, and income as described in the paper.

DescriptiveAnalysis

This folder contains the programs that generate all descriptive measures in the paper, including the simple descriptive statistics by year (number of observations, means, and standard deviations), the measures of inequality (cross-sectional variances and Gini coefficients), the measures of “volatility”, and the KSS and GM variance decompositions. (See “Note on file names” above for a description of the naming convention used in the program names.)

ErrorComponentsModels

This folder contains the Stata programs used to estimate the ECMs and compute the ECM-based variance decompositions.

- Subfolder *Baseline* contains programs corresponding to the baseline nonstationary ECM. Subfolder *LambdaPolynomial* contains the programs for the main version of the model, which restricts the ‘lambda’ parameters to a fourth-degree polynomial. Subfolder *LambdaUnrestricted* contains the programs for the unrestricted ‘lambda’ parameters, presented in the online appendix.
- Subfolder *Alternative* contains programs used for the estimation and analysis of the alternative nonstationary ECM, which is presented in the appendix. The estimation programs are written in Fortran90, and are located in subfolder *Fortran* (the Unix script *compile* compiles and runs the program). The program that collects the results to produce the variance decompositions is written in Matlab and is located in subdirectory *Matlab*.
- Subfolder *Stationary* contains the Stata programs that are used in the estimation of the stationary ECMs.

PSID

This folder includes several subfolders containing data and programs used for the analysis of PSID data, which is presented in the online appendix.

- Subfolder *Data* contains the cleaned-up data on male earnings and family income (which are used to construct the PSID covariance data), as well as the corresponding covariance data.

- Subfolder *DataCleanup* contains the programs used to define and rename variables, and clean up and process the raw PSID data to create the samples of male earnings and family income to be used in estimation (and included in subfolder *Data*).
- Subfolder *DataSetup* contains programs that further treat the PSID data on male earnings and family income, define the samples used in the analysis, and create the PSID covariance data.
- Subfolder *ECMs* contains the programs used for the estimation of the (baseline) ECM on PSID data and the corresponding variance decompositions.