Global Income Dynamics Database Project

Instructions for the STATA Code

Serdar Ozkan[*] Sergio Salgado[†]

January, 2020

1 General Directions

This document discusses the codes used to generate the sample and the statistics for the core section of the Global Income Dynamics Database Project (GID). The code packet–available in GitHub at https://github.com/salga010/QE-MasterCode– contains six main do files that execute the initialization of the parameters, the sample creation, and produce the figures for the core section of the paper. The packet also contains two auxiliary files used for summary statistics (myprogs.do) and plotting (myplots.do). The codes are written in Stata 13, designed to produce the statistics listed in the GID Guidelines document, save the results in CSV files, and the figures in pdf. The basics steps to run these codes are the following (with additional details in section 2).

1. Create in your local machine the following sub-folders (all in lower cases) under the same folder:

   - do
   - dta
   - log
   - out
   - figs

   Next, download the provided do files in folder /do and copy the country-specific raw data file in folder /dta. The log files will be saved in the /log folder, the results will be saved under /out, and figures will be saved under /figs.[1]

2. Open 0_Initialize.do in Stata and assign country-specific parameters such as the starting and ending years of the sample, the name and location of the raw dataset, the country's CPI, etc., for which further instructions are given in Section 2.

[*] University of Toronto; serdar.ozkan@utoronto.ca

[†] The Wharton School, University of Pennsylvania; ssalgado@upenn.edu

[1] Notice the folder fig/ will contain several additional subfolders (created by the plotting code) which will orderly save the figures for each section of the core section of the paper.

3. Open 1_Gen_Base_Sample.do in Stata, specify the directory of the main folder that contains the above five sub-folders in your local machine and run. This do file renames the variables, does very basic sample selection, creates new variables (e.g., log and residual earnings), and generates the master_sample.dta, which is a wide form dataset which will be used in the rest of the do files. The main output of this do file (master_sample.dta) is saved in the folder /dta and contains the following variables:

    (a) personid: id of the individual used throughout the do files

    (b) male: indicator variable equal to 1 if male and 0 if female

    (c) yob: year of birth of the individual

    (d) yod: year of death of the individual

    (e) educ: indicator variable with education categories

    (f) labor: real labor earnings in levels

    (g) logearn: real labor earnings in log levels

    (h) permearn: permanent income defined as $P_{it-1} = \frac{\sum_{s=t-3}^{t-1} y_{i,s}}{3}$ as explained in the Guidelines

    (i) permearnalt: alternative measure of permanent income. See the Guidelines, section "Key statistics 4: Mobility" for additional details

    (j) researn: residual log earnings

    (k) researn1F: 1-year forward residualized log earnings change, g_{it}

    (l) researn5F: 5-year forward residualized log earnings change, g_{it}^5

4. Open 2_DescriptiveStats.do in Stata, specify the directory of the main folder in your local machine and run. This do file generates a folder under /out, whose name consists of the date the program is run and "Descriptive_Stat". This folder will contain a set of .csv files with the statistics for the section "Key statistics 1: Descriptive statistics" described in the Guidelines of the project.

5. Open 3_Inequality.do in Stata, specify the directory of the main folder in your local machine and run. This do file generates the moments described under the section "Key statistics 2: Inequality and Concentration" in the Guidelines document. These moments will be saved under the corresponding folder under /out.

6. Open 4_Volatility.do in Stata, specify the directory of the main folder in your local machine and run. This do file generates a set of .csv files with the statistics for the section "Key statistics 3: Volatility and Higher-Order Moments".

7. Open 5_Mobility.do in Stata, specify the directory of the main folder in your local machine and run. This do file generates a set of .csv files with the statistics for the section "Key statistics 4: Mobility".

8. Open 6_Core_Figs.do in Stata, specify the directory of the main folder in your local machine, the directories where the different results are saved (Inequality, Mobility, etc.) and where the figures will be saved. The default is the folder /figs and figures are saved in pdf format.[2]

In the next section, we provide some additional details for the 0_Initialize.do do-file. All programs are heavily commented and we have made the best of our efforts to make them run bug-free. If you find any bug, please let us know so we can update the code and share the information with the rest of the teams. Furthermore, there may be situations where some changes on the code will be necessary due to the idiosyncratic features of each country's dataset or you may need further instructions. To facilitate smooth communication and collaboration between us and the teams, we have created a "Workspace" in a widely used messaging app, *Slack*.[3] You'll receive invitations in your emails to join this workspace. Please make sure that at least one of the members of your team joins this messaging group for further communication.

2 The Initialize Do file

The 0_Initialize.do defines the variable names, time span, and vectors used throughout the codes and allows each team to select some options that best suit their dataset. Given its importance, here we discuss several key details (more comments can be found in the do-file).

Lines 5 to 18 of 0_Initialize.do define general variables that must be followed by the teams to generate the core statistics. Hence, no change is required in this section. These definitions ensure that the sample used for the core section of the paper is comparable across countries.

Lines 20 to 100 require the input of each team. Please read in detail.

1. **Unix vs. Windows**. Define whether the machine you are running your codes is Unix/Mac (unix=1) or Windows (unix=0).[4]

2. **Wide vs. Long Format**. Define whether the raw sample is in wide form (wide =1) or long form (wide=0). If it is in long form, the 1_Gen_Base_Sample.do file will convert it to wide form (one row per individual) when creating the dataset master_sample.dta. The rest of the codes are designed to work with this .dta.

    (a) By long format, we mean a dataset in which each observation (row) is an individual-year pair. In other words, workers' observations are stacked, there is one column that defines the unit of time (year) and one column for each variable defining the value of each variable within the year (one column for earnings, one for education, etc.).

[2] To plot additional figures that you might be interested but are not covered in the file 6_Core_Figs.do you might also need to modify the file myplots.do. If that is the case, we encourage you to contact us before making changes so all the plots maintain a similar format.

[3] For more information on this app, visit https://slack.com/features.

[4] Although STATA run on Windows machines corrects the folder separators, just to be on the safe side, we specify whether the separator is "/" or "\", which will then be used to locate the sub-folders.

(b) By wide format, we mean a dataset in which each observation (row) is an individual and different columns define different observations for the same individual. In other words, workers' observations are side-by-side, and there is one column per year defining each variable (one column is the earnings in 2000, a second column is the earnings in 2001, and so on).

3. **Missing values for labor income.** If there are genuine missing values for labor income please set global ${miss_earn} to 1 (lines 33 to 36). If it is set to zero (the default), the code will convert all missing earnings observations to zero. This is particularly important if your raw dataset is in long form and there are no observations for zero labor income in a given year.

4. **Variable Names**. Specify the names of the variables in your data set between lines 41 and 48. These variables are the minimum set necessary to generate all the statistics in the Guidelines, hence, each team must make sure the raw data contains these variables. The 1_Gen_Base_Sample.do file then will rename these variable to our choices in the master_sample.dta. This helps to simplify the code in the rest of the do-files.

5. **Variable Types.** The do files are written under certain assumptions about the type of variables available in each dataset. We did not attempt to change the format of the variables, hence, each team must make sure that the raw data contains the correct format (i.e. education must be a numerical categorial integer variable, gender must be binary, etc.). Here we describe in detail the variables used in the analysis

   (a) ${personid_var}: Numerical categorical variable. Teams must make sure an individual id appears only one time per year in the sample.

   (b) ${male_var}: Numerical categorial variables which is equal to 1 if the individual is male, 0 if female.

   (c) ${yob_var}: Numerical categorical variable that defines the year of birth of an individual. Teams must make sure this is not missing or changes across different observations of the same individual (if the raw data is in long form). Individuals with missing age will be dropped from the sample.

   (d) ${yod_var}: Numerical categorical variable that defines the year of death of an individual. Teams must make sure this does not change across different observations of the same individual. Individuals with missing ${yod_var} will be treated as they where still alive by the end of the sample.

   (e) ${educ_var}: Numerical categorical variable that defines the education group of an individual. This can change across different observations of an individual. There is no restriction on the number of categories this might contain. The GID Guidelines, however, defines certain commonly use groups.[5]

   (f) ${labor_var}: Numerical variable that defines the labor earnings of an individual. This variable might contain missing values. Recall that you also need to choose whether the missing observations are set to 0 by setting global ${miss_earn} to 1 or 0 in line 36.

[5] Some datasets do not contain information on education. Instead of modify the code, we would recommend to create a place holder of such variable which is equal to one for all individuals in the sample. This will not change any of the results but will allow the code to run smoothly.

(g) ${year_var}: Numerical variable that defines the year of the observation if the raw data is in long form.

6. **First and last year.** Specify the first and last year of the sample for which the statistics will be calculated. The sample is assumed to have no gaps in between (all years between ${yrfirst} and ${yrlast} are available). If that is not the case and your sample contains gaps, please contact us.[6]

7. **Density estimation.** Global ${kyear} defines for which years the densities will be calculated. By default, the code calculated the densities in years ending with 0 and 5 (i.e. 1995, 2000, 2005, etc.). In case you want to calculate densities every year, change ${kyear} = 1.

8. **Quantile estimates.** Quantile estimates are mainly used in the 5_Mobility.do do file. See the code for additional details. The global ${nquantiles} defines how many quantiles will be used to divide the distribution of permanent income. The default is 40, as suggested in the Guidelines. The global ${nquantilesalt} does the same for the quantiles of the distribution of alternative permanent income. The global ${nquantilestran} defines the number of quantiles used in the transition matrices.

9. **Heterogeneity groups.** The global ${hetgroup} specifies what heterogeneous characteristics are considered. By default the code follows the GID Guidelines, calculating the statistics by gender, education, age, and the cross groups. Additional levels of heterogeneity can be easily incorporated as long as the corresponding variables are passed to the sample.[7] Contact us in case you have issues incorporating additional degrees of heterogeneity.

10. **CPI, min income, and exchange rate.** The matrices cpimat, rmininc, and exrate contain the CPI, the min income threshold, and the exchange rate (nominal) that is used throughout the code. These need to be imputed from ${yrfirst} to ${yrlast} *without* gaps. All nominal variables must be deflated by 2018 prices. Hence, set the global ${cpi2018} equal to the corresponding value the CPI in 2018 for your country.

The rest of the code re defines the minimum income threshold as suggested in the Guidelines and several lists of years that will be used by other do-files for different calculations. Please do not change them. If you think you have a good reason to change them please contact us.

[6] The code is flexible enough to deal with samples with gaps but will require some small changes. In the case your sample has gaps, please contact us.

[7] Check the code myprogs.do for details.