Replication package readme for Automated Tax Filing: Simulating a Pre-Populated Form 1040

June 2023

Lucas Goodman*
Office of Tax Analysis,
U.S. Department of the Treasury
1500 Pennsylvania Avenue NW
Washington, DC 20220
lucas.goodman@treasury.gov

Katherine Lim*
U.S. Department of Agriculture,
Economic Research Service
katherine.lim@usda.gov

Bruce Sacerdote*
Department of Economics
Dartmouth College
6106 Rockefeller
Hanover, NH 03755
bruce.i.sacerdote@dartmouth.edu

Andrew Whitten*
Office of Tax Analysis,
U.S. Department of the Treasury
1500 Pennsylvania Avenue NW
Washington, DC 20220
andrew.whitten@treasury.gov

I Overview

Code for data cleaning and analysis is provided as part of this replication package. It is available at `https://dataverse.harvard.edu/dataverse/ntj` for review.

The vast majority of the data used in this paper is derived from administrative tax data. The tax data was accessed and handled solely by Lucas Goodman and Andrew Whitten, employees of the United States Department of the Treasury. Any taxpayer data used in this research was kept in a secured Treasury or IRS data repository, and all results have been reviewed to ensure that no confidential information is disclosed. This document describes the construction and use of the administrative tax datasets.

II Data Availability

The tax data (Internal Revenue Service, 2023, 2021) were accessed by Lucas Goodman and Andrew Whitten, employees of the United States Department of the Treasury. Under section 6103 of the Internal Revenue Code, this data cannot be shared. For this reason, we are unable to include this data in the replication package. Additionally, some elements of the code itself must be redacted so as not to reveal protected information about variable names and database structure. Researchers can apply for access

*This research was conducted while Goodman and Whitten were employees at the U.S. Department of the Treasury, and the research was supported in part by the U.S. Department of Agriculture, Economic Research Service. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors and do not necessarily reflect the views or the official positions of the U.S. Department of the Treasury and should not be construed to represent any official USDA or U.S. Government determination or policy. Any taxpayer data used in this research was kept in a secured Treasury or IRS data repository, and all results have been reviewed to ensure that no confidential information is disclosed. The authors declare that they have no relevant financial interests that relate to the research described in this paper.

to the IRS tax data via the Statistics of Income Joint Statistical Research Program: `https://www.irs.gov/statistics/soi-tax-stats-joint-statistical-research-program`.

Researchers who have access to the restricted tax data also have access to a restricted git repository. That git repository contains the unredacted code. The code is stored under the title "Pre-populated returns."

III Computational requirements

The analysis was performed on IRS servers. This readme presumes access to these specific servers.

The initial data construction was performed in SAS (9.4) and Stata 17.0. The data analysis was performed in Stata 17.0.

For Stata, the following commands are required and can be installed from `SSC` and `netsearch`: reghdfe (Correia, 2017), gtools (Caceres, 2018), niceloglabels (Cox, 2017), and grc1leg (Wiggins, 2003).

IV Description of data construction

IV.A Directory structure

The directory structure is as follows.

- SAS
  - codeSAS
  - dataSAS
  - logSAS
  - out
  - toSAS
- Stata
  - code
    - get_data
      - nrp
      - post_sas
      - pre_sas
    - analysis
  - dataNRP
  - dataSTATA
  - fromSAS
  - log
    - analysis
    - get_data
      - nrp
      - post_sas
      - pre_sas
  - out
    - figures
    - tables

IV.B Data construction

The main data is constructed as follows.

1. First, in Stata, `code/get_data/pre_sas/insole_pull.do` is run. This produces a file, `dataSTATA/soi_2019.dta`, which is then transferred to SAS in `toSAS`.
2. Second, in SAS, we retrieve information from the database of tax and information returns. The code is structured as a set of macros, called by a master file `master_autofiling.sas`. This portion of the code extracts the data from the underlying database; this portion of the code is partially redacted. The code also performs additional cleaning steps. This code produces `out/deps.dta` and `out/result.dta`, which is transferred to Stata into the `fromSAS` directory.
3. Third, we continue with additional data preparation in Stata, in `code/get_data/post_sas.do`. The do files can be called (in the appropriate order) by the master do file, `code/get_data/master_post_sas.do` in that folder.

The data from the National Research Program is constructed under a separate stream, using the two do files in Stata, `code/get_data/nrp`. The do file `get_nrp.do` should be called before `get_nrp_w2.do`. This portion of the code extracts the data from the underlying database; this portion of the code is partially redacted.

V Instructions

Replicators should first gain access the IRS servers. Members of the public will not be able to do this. Then, they should replicate the folder structure described above. Next, they should make sure that they have installed all user-written Stata commands described above.

Users must modify the following files in order to set their home directory:

- `SAS/codeSAS/master_autofiling.sas`. Additionally, users should input their user name and password in this file.
- `Stata/code/globals.do`

After running these preliminary steps, users should construct the data as described in section IV.B. Then, they should run the do-files indicated in Table 1 to create the figures and tables, all of which are located in `Stata/code/analysis`.

In addition, certain statistics cited in the paper that are not found in tables are generated by the `summary_stats.do` and `nonfilers.do` code files.

References

Caceres, Mauricio. 2018. "GTOOLS: module to provide a fast implementation of common group commands." http://fmwww.bc.edu/RePEc/bocode/g.

Correia, Sergio. 2017. "REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects." *Statistical Software Components, Boston College Department of Economics*, https://ideas.repec.org/c/boc/bocode/s457874.html.

Cox, Nicholas J. 2017. "NICELOGLABELS: module to produce nice axis labels for logarithmic scales." *Statistical Software Components S458416, Boston College Department of Economics*, https://ideas.repec.org/c/boc/bocode/s458416.html.

Internal Revenue Service. 2021. *INSOLE file 2019*, Department of Treasury, Washington DC, accessed 2023-02-28.

Internal Revenue Service. 2023. *Population Files 2014-2019*, Department of Treasury, Washington DC, accessed 2023-02-28.

Wiggins, Vince. 2003. "GRC1LEG: Combine graphs into one graph with a common legend." *Statalist distribution*.

Table 1: Concordance between do files and tables/figures

Figure/Table	**Analysis do file**
Table 1	overall_success.do
Table 2	common_bad_situations.do
Table 3	table_who_benefits.do
Table 4	line_by_line.do
Table 5	subsets.do
Table 6	nrp_table.do
Table 7	nrp_table.do
Table 8	nonfilers.do
Table 9	table_mvpf.do
Table A1	summary_stats.do
Table A2	subsets.do
Figure 1	overall_success.do
Figure 2	overall_success.do
Figure 3	progressivity.do
Figure A1	f1099misc_counts.do
Figure A2	graph_bounds_by_age.do
Figure A3	graph_over_underestimates_by_agi.do

Notes: All analysis do files are located in `Stata/code/analysis`. Note that `f1099misc_counts.do` is partially redacted, as that do file extracts some data from the underlying database, whose metadata is not public information.