Replication package for "Delivering aid to businesses through the payroll tax system: the case of the Employee Retention Credit"

Lucas Goodman

January 2023

1 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, an employee 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.

2 Data Availability

The tax data (Internal Revenue Service, 2022) were accessed by Lucas Goodman, an employee 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 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 "Employee Retention Credit."

In constructing the tax data, we make use of the following tax forms.

- Form 941, filed quarterly by employers to report payroll tax liability, income tax withheld on behalf of employees, tax deposits made throughout

the quarter, and to claim certain tax credits including the Employee Retention Credit (ERC).

- Form 941-X, used by employers to amend previously filed Forms 941, including to claim the ERC.
- Form 1040, used by individuals to file their individual income tax returns.
- Form 1040, Schedule C, used by sole proprietors to report the Employer Identification Numbers (EINs) of their businesses, as well as information about their businesses such as their gross receipts.
- Forms 1065, 1120S, and 1120, which report information about partnerships, S corporations, and C corporations, respectively.
- Schedule K-1 of Forms 1065 and Form 1120S, which we use to link partnerships and S corporations, respectively, to their owners.
- Form 1120, Schedule G, which we use to link certain C corporations with their owners.
- Form 851, which we use to link C corporations with their subsidiaries.
- Form W-2, which we use to identify employees and their wages.
- Internal entity-level files for businesses, which contain information on industry and other features of employers.
- An internal mapping from state and county to zip codes.
- Information on Paycheck Protection Program loans, as released by the Small Business Administration and stored on internal IRS servers. This dataset includes a pre-existing fuzzy match to the tax data based on name and geography.

3 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). The data analysis was performed in Stata 17.0; the result of this analysis is a set of small output datasets containing results (e.g., regression coefficients). These output datasets are converted to tables (.tex code) and figures in a separate Stata 15.0 instance.

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).

4 Description of data construction

4.1 Directory structure

The directory structure is as follows.

- SAS
  - codeSAS
  - dataSAS
  - logSAS
  - out
- Stata_analysis
  - codeSTATA
  - dataSTATA
  - fromSAS
  - out
  - outLOG
- Stata_output
  - codeSTATA
  - dta
  - out

4.2 Data construction

The initial data preparation is performed in SAS using the `.sas` files contained in `codeSAS`. The code is structured as a set of macros, called by a master file `master_ertc.sas`. This portion of the code extracts the data from the underlying database; this portion of the code is partially redacted. The code performs additional cleaning steps. In particular, `solve_owners.sas` traces through firm-to-firm links to find as many individual owners of businesses as possible. `do_merge.sas` performs additional calculations and cleaning. Ultimately, this SAS code exports a variety of `.dta` files, the most important of which is f941.dta. These `.dta` files are ready for immediate analysis in Stata.

5 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_ertc.sas`. Additionally, users should input their user name and password in this file.
- `Stata_analysis/codeSTATA/mg_ertc.do`
- `Stata_output/codeSTATA/make_globals_ertc.do`

After running these preliminary steps, users should run `SAS/codeSAS/master_ertc.sas` to retrieve the tax data. Next, they should transfer the `.dta` files in `SAS/out` to `Stata_analysis/fromSAS`.

Next, replicators should run `Stata_analysis/codeSTATA/mg_ertc.do` and then `Stata_analysis/codeSTATA/master_results.do` (which calls the other do files in that folder) to perform the analysis.

Next, replications should transfer the `.dta` files from `Stata_analysis/out` to `Stata_output/dta`. Then, replicators should run `Stata_output/make_globals_ertc.do`, followed by all of the other programs in that folder (in any order). This step turns the analysis output into figures and tables.

6 References

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. 2022. *Population Files 1999-2020*, Department of Treasury, Washington DC, accessed 2022-05-01.

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**	**Output do file**
Figure 1	time_series	time_series
Figure 2	by_size ertc_histogram	2020_2021
Figure 3	takeup_by_ppp	takeup_by_ppp
Figure 4	amendment_timing	amendment_timing
Figure 5	refund_by_size	refund_by_size
Table 1	rest_by_city	rest_by_city
Table 2	explain_all	explain_all
Table 3	preparers	preparers
Table 4	refund_share	refund_share
Table 5	delta_gr	delta_gr
Table 6	distributions	distributions

Notes: All analysis do files are located in `Stata_analysis/codeSTATA`. All output do files are located in `Stata_output/codeSTATA`.