README for Replication of
The Earnings and Labor Supply of U.S. Physicians

Joshua D. Gottlieb Maria Polyakova
Kevin Rinz Hugh Shiplett
Victoria Udalova

December 2024

Overview

The code for this replication package produces all exhibits in the published paper. The entire replication package can be run in a single click by typing `run_project_from_start_to_finish` into the terminal from within the `/replication/code` folder. This command submits a PBS batch job that runs the do file `docinc-run_project.do`, which in turn executes all the other Stata and Python scripts for the project. As of December 2024, this batch job takes approximately five hours to run with the current server setup.

Data

Replicating our work requires access to restricted versions of the Centers for Medicare and Medicaid Services' (CMS) National Plan and Provider Enumeration System (NPPES) data, federal income tax data, ACS data, and administrative records from the Social Security Administration (SSA). Access to these datasets is governed by strict data use agreements that specify allowed uses of these data. These data are available only to researchers with Special Sworn Status (SSS) working on approved internal Census Bureau projects authorized under Titles 13 and 26. All non-Census-held data used in this project are publicly available. For additional data details, see Appendix B.1.

Project Structure

All code is contained within the `/replication/code` folder. Subfolders of `/replication/code` are listed below. The order corresponds to the order in which the code is run.

- `/cleaning`
- `/descriptives`
- `/twfe`
- `/govt_policy`
- `/spec_choice`
- `/pdv`
- `/tuition`
- `/exhibits`

Code in each subfolder generates results corresponding to that subfolder's name. For instance, the code in `/twfe` produces our two-way fixed-effects estimates. Folders have corresponding `do` files named after them in the `/exhibits` folder. Taking the previous example, exhibits containing results from `/twfe` are produced in `docinc-exhibits_twfe.do`.

The folders `/descriptives` through `/tuition` above produce outputs which, after rounding according to Census disclosure rules, are saved in `/replication/intermediate_csv`. Intermediate outputs such as unrounded estimates and log files are saved to `/replication/output` or `/replication/temp`. The `/exhibits`

folder produces exhibits using the rounded estimates in `/intermediate_csv`, saving the figures in `/figures` and the tables in `/tables`.

Adjusting the code

Our code assumes the pre-existence of two data files, `panel_lawyers.dta` and `panel_physicians.dta`, that contain income and demographic data for our panels of lawyers and physicians. These files were created by Census employees, who processed the raw tax data using SAS. Variable names in our code have been redacted and replaced with intuitive placeholder names (e.g., `w2wgs`, `social_security`, etc.). Replicators with access to the original raw tax data will need to map these placeholder names back to their original counterparts.

Replicators may need to adjust file paths based on the location of their data. These adjustments should be made in `docinc-run_project.do`, which defines file paths using global macros relative to the `/replication/code` directory.

Computational requirements

Our code was run December 2024 using the following software:

- Stata 18
- Python 3.9.12
- Bash shell
- LaTeX

All dependencies above are available by default on the FSRDC server. For all Stata and Python packages, we used the server's default versions. Custom `ado` functions to round estimates according to Census disclosure rules are stored in the `/replication/code/ado` folder.