Overview of Data Files and Programs for "Effects of COVID-19 on the Nonprofit Sector"

Stephanie Karol[1] and Jennifer Mayo[2]

[1]Department of the Treasury
[2]University of Missouri

April 8, 2024

1 Guide to Programs

The programs included in the replication files include three Stata do-files, one R script, and one TeX file. They are listed as follows:

1. **Replication.do**: This Stata do-file first compiles and cleans the data for the analyses in Sections 1 - 5 and Online Appendices A-D. It then carries out these analyses and outputs the results in one of two ways. Either regression results are output to LaTeX-formatted tables, or coefficients and standard errors are output to CSV files. Next, this program calls "ThreeYearHorizon.do" and "PretrendsData.do," both of which produce output as either LaTeX-formatted tables or CSV files.

2. **ThreeYearHorizon.do**: This Stata do-file first compiles and cleans the data for the analyses in Section 6. It then carries out these analyses and outputs the results either as LaTeX-formatted tables or to CSV files.

3. **PretrendsData.do**: This Stata do-file first compiles and cleans the data for the analyses represented in Figure 2. It then carries out these analyses and outputs the results to CSV files.

4. **all-figures.R**: This R script takes in CSV output and uses it to build all figures in the paper except for Figure 1.

5. **fig1tikz.tex** This TeX file creates Figure 1 as a TikZ picture.

To begin the replication, ensure that Stata, RStudio, and LaTeX are installed, and edit all programs to point to the same local directory. This directory should likewise include all datasets listed in Section 2. Run the do-file "Replication.do", then the R script "all-figures.R", and finally the TeX file "fig1tikz.tex."

2 Datasets for Analysis

The programs listed in Section 1 rely on the following datasets:

1. **core.csv**: This CSV file contains cleaned electronic filings of Form 990 data. These data were scraped from the Amazon Web Services repository of e-filed Form 990 data that was maintained by the IRS until 2021. This data repository is no longer available; see https://registry.opendata.aws/irs990/ for details. However, the underlying data can still be

accessed publicly, either by requesting the data from the IRS or by processing the XML files published at https://www.irs.gov/charities-non-profits/tax-exempt-organization-search-bulk-data-downloads. This file contains select data on Section 501(c)3 charities for fiscal years 2011 - 2020.

2. **cleaned-990.dta**: This dta file is created from the e-filed Form 990 data, scraped from the now-deprecated Amazon Web Services repository described above. Observations are dropped from the raw dataset for four reasons. First, observations are dropped if the charity's mailing address is outside the 50 US states plus the District of Columbia. Second, an observation is dropped if the charity reports negative contributions (from public or private sources), revenues, assets, salaries, or expenditures (for fundraising, program services, or overall). Third, observations are dropped if the charity appears to have submitted a revised return for the same fiscal year: for each organization-fiscal year pair, only the most recent return is retained. Finally, observations are dropped if they represent short years.

3. **cleaned-990-diff3.dta**: This dta file is created from the e-filed Form 990 data, downloaded from the XML archives hosted on the IRS website (https://www.irs.gov/charities-non-profits/tax-exempt-organization-search-bulk-data-downloads). The data processed from these XML files is cleaned in the same way as the data included in cleaned-990.dta. Data for fiscal year 2021 is then appended to the data compiled in cleaned-990.dta.

4. **cpifisyr.dta**: This dataset contains 12-month moving averages of monthly CPI-U data, sourced from the Bureau of Labor Statistics. It also contains a variable called "fisyrendmo," which is a year-month pair, formatted as *yyyy-mm*. This variable is used to link the average CPI-U observed during the time period covered by the charity's fiscal year to the corresponding Form 990. This variable is used to deflate all charity financial variables. The base period is 1982-1984.

5. **censuspop.dta**: This dataset contains state-level population estimates from the Census for years 2018 through 2020.

6. **ntee.dta**: The National Taxonomy of Exempt Entities codes for all charities in the sample come from archived versions of the IRS' Business Master File of tax-exempt entities. These files are archived by the Urban Institute's National Center for Charitable Statistics, and can be found at https://nccs.urban.org/nccs/catalogs/catalog-bmf.html. This project relies on NTEE codes assigned in Business Master Files archived between December 2018 and April 2020.

7. **FootTraffic-Clean.dta**: This dataset contains data on foot traffic by state, based on Google's Community Mobility Reports. The calculation of cumulative differences in foot traffic in each month of the year 2020, relative to the five-week period beginning January 3, 2020, is described in Appendix B.1.1. The resulting dataset includes state FIPS code and month identifiers.

8. **matched-f990-ppp.dta**: This dataset contains matched data on charities' Paycheck Protection Program (PPP) loan approvals and forgiveness. The data on PPP loans comes from the Small Business Association (https://www.sba.gov/funding-programs/loans/covid-19-relief-options/paycheck-protection-program/ppp-data), and includes organization names and the specific dates on which loans were approved or forgiven. The organization names listed in the PPP data are matched to the names of organizations included in *cleaned-990.dta* using the procedure outlined in Online Appendix C. This procedure results in a

crosswalk between the PPP data and the Form 990 data. As the Form 990 data includes information on the timing of the nonprofit's fiscal year, the loan amounts approved or forgiven can then be allocated to the appropriate fiscal year based on whether the approval or forgiveness date falls within the corresponding date range.

3 Assembling Datasets

The primary dataset used for analysis is *cleaned-990.dta*. It is unique at the EIN-fiscal year level. The datasets *matched-f990-ppp.dta* is matched to the primary dataset by EIN and fiscal year. The dataset *ntee.dta* is matched to the primary dataset by EIN. The dataset *censuspop.dta* is matched to the primary dataset by fiscal year. The datasets *cpifisyr.dta* and *FootTraffic-Clean.dta* are matched to the primary dataset by a combination of the fiscal year and the ending month of the fiscal year.

The analyses carried out in Section 6 rely on data from *cleaned-990-diff3.dta* instead of *cleaned-990.dta*, but the steps required to assemble the dataset used for these analyses are identical to the steps described above.