An Approach to Safely Expanding Access to IRS Data

Len Burman
Robert C. Pozen Director
Urban-Brookings Tax Policy Center

SOI Consultants Panel Meeting
June 10, 2016

Goals of the Project

- Produce synthetic data files with the same record layout as IRS Administrative Data Files that:
  - Protect the confidentiality of tax records
  - Can be used for statistically valid analysis for a variety of research purposes
  - Can be produced within the constraints of the SOI's tight budget
- Develop a safe procedure for selected researchers to remotely submit programs to perform statistical analysis on administrative data when the synthetic files are inadequate

Benefits

- Advance the state of public economics research
- Learn about the effects and effectiveness of current tax policies
  - State is a potentially enormously valuable source of (mostly) exogenous variation
- Also useful for other purposes
  - Chetty and Saez on economic mobility, for example
- Emerging bipartisan consensus on the importance of safely using administrative data to improve public policy
  - Bipartisan Evidence-Based Policymaking Commission Act

Joint Statistical Research Program

- Small number of researchers can access the administrative data through the Joint Statistical Research Program
  - Program is limited by SOI resource constraints—staff, equipment, space.
  - Approval for direct access to data takes a long time and requires significant investment of time by researchers and IRS staff

Public Use File (PUF)

- The Public Use File (PUF) is an invaluable resource, but procedures used to protect confidentiality limit the file's validity for many statistical purposes
  - High-income returns, defined as |AGI| > 250,000, have some data suppressed and other data altered
  - Extreme high or low values—highest or lowest 30 values—are aggregated into a single residual return
  - Some values are top-coded (e.g., number of dependents)
  - Much useful detail is excluded from the PUF

Problems with PUF Disclosure-Avoidance Procedures

- Much data that is available inside IRS must be imputed by PUF users
  - State of residence, wages and self-employment income of head and spouse, ages, contributions to and balances in retirement accounts, value of health insurance, information about nonfilers
- Data privacy procedures add noise to key variables which leads to inconsistent and inefficient econometric estimates
  - Weakens hypothesis testing and reduces the likelihood of finding statistically significant results, even where they exist in the undistorted administrative data
  - Impossible to derive the statistical properties of the modified PUF dataset

PUF may still not be disclosure-proof

- With explosion in availability of financial and other data online, the likelihood of a match between a unique tax record and independently available information is growing, raising the possibility of disclosure
- At some point, the IRS will need to find a replacement for the PUF that does not include any unique taxpayer information

Best to start work on a replacement now

Very Preliminary Outline of a New Approach

- Create synthetic datasets with same record layout as original administrative data designed to be statistically valid for particular kinds of research questions
- Create safe way to remotely submit statistical analyses to IRS
- Develop pricing mechanism
- Explore possibility of updating synthetic file based on statistical analyses that are submitted by users
- Explore possibility of automating the process of disclosure avoidance review

Creating Synthetic Datasets

- IRS would anonymize the data by adding random errors to independent variables
  - Identities, other relationships between variables, would be preserved
  - Issue about how to treat categorical data
  - Records that coincidentally are too close to unique records (or records that recur < *N* times) in administrative data would be deleted and replaced
- IRS would create a calibration dataset containing statistics derived from the underlying data
  - E.g., variance-covariance matrix; conditional means within income, filing status, state cells; etc.

Synthetic Dataset (continued)

- Outside researchers would find weights for the anonymized file so that its statistical properties match the calibration database (within defined tolerances) to create a statistically accurate replica of the original data
- This file would be a new kind of PUF, better in some ways than the original since it would be designed so that a predefined set of statistics (those that are functions of the statistics in the calibration database) could be measured without bias.
- Because the file will not contain any unaltered taxpayer records, it might also include data that are excluded from the PUF
- Multiple synthetic datasets could be drawn based on random subsamples of the master file
    - For example, one might contain detailed geographic information, but less detail about sources of income and deductions

Remote Access of Administrative Data

- Researchers would develop their programs using the synthetic dataset and submit the programs electronically to the IRS
- Procedure would be similar to the one under development by Vilhuber and Abowd (2015) for access to the confidential version of the SIPP
- Disclosure risk could be minimized by adding random errors to statistical estimates or basing estimates on subsamples of the restricted dataset (i.e., by drawing random samples with replacement from the original dataset)

Enhancements

- Explore whether it is possible to automate the disclosure review process.
- Explore whether it is possible to use administrative data runs to improve the synthetic data file over time

Financing Mechanism

- Pricing
  - Charging as a way to prevent data mining
  - Setting price based on the shadow cost of privacy draw
    - CS notion of a privacy budget
- Using IPAs to support SOI in providing disclosure review and improving synthetic data files

Issues

- Applying to small datasets or those with very high AGI
- Panel data
- Could there be a corporate tax model file?