Synthetic PUF Project Update

SOI Consultants' Panel

September 8, 2017

Len Burman
Tax Policy Center/Urban Institute

TPC TAX POLICY CENTER
URBAN INSTITUTE & BROOKINGS INSTITUTION

Goals

- Produce synthetic data file with the same record layout as IRS administrative dataset (e.g., INSOLE) that:
  - Protects the confidentiality of tax records
  - May be used for statistically valid analysis for certain research purposes
  - May be used as a “training data set” to develop programs to run on confidential data
- Develop a validation server to run statistical analysis on the administrative data when the synthetic files are inadequate (for a fee)

Potential Benefits to IRS

- Reduce the cost of producing safe public use data files (and reduce lag time to release)
- Research might be useful for improving or automating the production of INSOLE file
- Facilitate collaboration between SOI analysts and outside researchers on issues related to tax administration

Synthesis process

1. Partition data
2. Synthesize the data within each partition
3. Refine the data within each partition by subsampling to hit certain statistical targets

1. Partition the data

- Divide the edited administrative data file into partitions to make the synthesis problem manageable
  - Could use the strata currently used for selecting INSOLE or some other partition
  - Group by state or relatively homogeneous regions
  - Partition sizes need to be small enough to make synthesis computationally feasible but not so small that it creates disclosure risks (from over-fitting)
  - There may be problems of discontinuities or inconsistencies at partition boundaries

2. Synthesize the data

- Use parametric or nonparametric or combination of methods
- Incorporate complex relationships between variables in synthesis process

Creating synthetic dataset: parametric methods

- Explicitly model the distribution of each variable to be synthesized
  - E.g., linear (or log-linear) regression for continuous variables; probit, logit, or ordered probit or logit for categorical variables; or other maximum-likelihood estimators
  - Advantage: relatively quick and easy
  - Disadvantages
    - very sensitive to model structure and assumed error distribution
    - may not capture complex nonlinear relationships
    - A sequence of conditional distributions (e.g., regressions) may be a poor approximation of the multivariate distribution
    - If errors are correlated, calculated variables (e.g., AGI) may be very inaccurate

Creating synthetic dataset: nonparametric methods

- Nonparametric methods make no prior assumptions about the underlying distribution or the process that generated the data.
- Example
  - CART (Classification and Regression Trees)
    - Divide data into relatively homogeneous partitions conditional on X. (These partitions are metaphorical branches of the data "tree.")
    - Randomly draw (using Bayesian bootstrap) one of the Y_1 observations from all the observations (leaves) on the branch. That is the synthesized value.
    - Repeat for Y_2 conditioning on the synthesized Y_1 and X; etc.
    - To protect against disclosure, you can smooth the leaves using kernel density and draw a random value from the empirical density.
    - Prune the tree so that the branches contain enough leaves to preserve confidentiality

Nonparametric methods (continued)

- Advantages
  - preserves complex relationships between variables
  - Software exists to do this without detailed model specification (SYNTHPOP)
  - Synthetic data have proven to be reliable for simple statistical analysis in simple applications
- Disadvantages
  - Computationally intensive; may not even be feasible on our giant dataset
  - Little experience using for fully synthetic data

Ultimate process could be a hybrid

- Parametric estimation for some variables and nonparametric for others
- For example, we might want to use CART for some key variables, parametric methods for others.
- We might want to use less precise methods for very sensitive variables. For example, linear regression imputations would trim extreme values, which might be desirable for some variables.

Issues/questions

- Should we target some calculated variables, like AGI?
  - Could synthesize AGI (because it is important) and shares
  - Or could synthesize all components, but scale so that they add up to synthesized AGI
- Must some variables be suppressed because of disclosure concerns?
  - The set of variables to synthesize will inform synthesis methodology
- How to deal with outliers?
- Must we eliminate synthesized records that are accidentally too close to actual records?

3. Refine the data

- Develop a procedure to extract the optimal sample of synthetic returns from the population of synthetic data records.
  - Based on the population, calculate a set of statistical targets (for example, means, $X'X$, counts of categorical variables, conditional on income, state, filing status, age).
  - Select the sample of size n_i within each partition i that minimizes the distance between sample statistics and population values.
  - Evaluate data quality; compare to a random synthetic file of the same size and to the PUF.
  - Test whether statistical inference in the refined sample matches inferences drawn from the population. We believe that this process would obviate the need to correct standard errors, as in multiple imputation.

Remote Access of Administrative Data

- Researchers would develop their programs using the synthetic dataset and submit the programs electronically to the IRS
  - Output would be reviewed by a contractor or IPA before returning to researcher
- Procedure would be similar to the one developed by Vilhuber and Abowd (2015) for access to the confidential version of the SIPP
- Disclosure risk could be reduced by basing estimates on random subsamples of the restricted dataset (i.e., by drawing random samples with replacement from the original dataset)
  - Programs could include generated random seed so that a particular analysis could be replicated, but any new analysis would start with a different seed

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