Sharing Federal Data Externally for Transparency and Reproducibility

Ruth Ann Killion
US Census Bureau
June 21, 2017

Acknowlegements and Disclaimer

Special thanks to Michael Cohen, John Abowd, John Eltinge, Ron Jarmin and many others I discussed these issues with!

This work was done upon request. All views are those of the author and do not represent the views or policies of the US Census Bureau.

Agenda

- Current Methods
- Transparency
- Census Bureau Research on Editing and Imputation
- Reproducibility
- What Question Do We Need to Answer?

Current Methods

- Public Use Microdata Sample Files
  - Provided for decades
  - Allow substantial additional analysis, though only on a sample
  - Re-identification becomes more of an issue as hackers, et. al., get better, craftier and have more access to Big Data sources
  - Census Bureau reviewed policy in 2013:
    - Identified knowledge of “participation in a survey” as biggest threat

Current Methods (cont)

- Special Sworn Status
  - Relatively easy approval process
  - No excessive cost
  - Research conducted on government machines, usually on site
  - No lasting access to data for researcher

Current Methods, cont

- Synthetic Files
  - Allows construction of a data set that preserves variance/covariance structure of the distribution
  - Quickly falls apart after second moment
  - Third moment (skew) is especially important with economic data
  - Disclosure techniques can alter underlying set
  - Not a perfect solution

Current Methods, cont

- Research Data Centers
  - More freedom for researchers – access to “real data”, including from multiple agencies
  - Extensive, clunky bureaucratic approval process that can involve multiple agencies
  - Expensive
  - Very tight IT Security controls

Transparency

- Documentation
  - Hardest product to produce
- Methodology
  - Often cannot disclose details (e.g., PSA)
- Disclosure Avoidance
  - Intentionally perturbs data
- Editing and Imputation
  - Not all automated or rule-based

Current Research

- Economic Statistics at Census Bureau
  - History of experts who do ad hoc editing based on previous responses and particular knowledge of firms
  - Not all data amenable (ACES)
  - Can we shorten the time to production
  - Can we make it more reproducible
  - Without changing the estimates and inferences of the data

Motivation for the Edit Reduction Effort

- Surveys need to increase efficiency of processes in editing data (reduce cost, improve timeliness, etc.)
- Decades of research show over-editing of data.
- Identifying and dealing with outliers is necessary, but detailed microdata cleanup may not be.
- Users desire more timely, relevant data.
- The surveys need accurate and repeatable editing practices.

ACES Edit Reduction – Experiments

Experiment	Purpose
Examining Quantities Over Time	To examine raw sums, estimates, standard errors, and the number of edit failures over time
Editing in the Absence of Edit Failures	To understand the nature of edits that are made to adjust data but not to correct for edit failures
Impact of Editing	To quantify the impact of editing on estimates by NAICS and edit type
Modeling Stopping Points	To model when the stop editing certain NAICS codes and switch resources to other NAICS codes

ACES Basic Results (1)

- Edits that do not address edit failures (i.e., expert edits) have very little impact on estimates
- Production of results could happen about two months earlier if did automated edits and dealt with outliers
  - Estimates of many (not all) variables are stabilized within the confidence intervals by that point in the processing cycle

ACES Basic Results (2)

- If use automated edits only, can increase ability to reproduce.
- New editing practices, including some related to Big Data, may increase the likelihood of estimates being stable and within the final confidence interval earlier.

Future Research

- Determine the impact of edits to create a hierarchical editing system.
- Automate certain types of edits.
- Build machine learning (AI) processes
- Research the use of Big Data editing techniques along with new data sources to increase data accuracy while decreasing analyst burden.
- Continue researching stopping point models so that editing can become more adaptive.

Reproducibility

- Many issues have to be addressed in order to make reproducibility practical
  - Automated editing (no two groups of researchers will have the same experts or expertise)
  - Use of random number generators in editing and imputation routines (very common)
  - Disclosure avoidance perturbs data

Basic Question

Should we expect to be able to reproduce publicly published federal data?

- Are our techniques robust and transparent enough to satisfy the goals of the scientific method?
- With many checks and balances built in and many people reviewing all aspects, can we claim what is produced is, indeed, reproduced?
- Can we improve our documentation enough?
- What do we need to change about how we operate?

In Other Words

What question should we be answering?

Reproducibility, *per se?*

Data/Process/Documentation Quality?

True transparency?

Contact information:

Ruth.Ann.Killion@Census.gov

301-763-2048