Assessing the Automated Imputation of Missing and Erroneous Survey Data: A Simulation-Based Approach

bea
Bureau of Economic Analysis
U.S. DEPARTMENT OF COMMERCE

Larkin Terrie

Federal Committee on Statistical Methodology Research and Policy Conference

March 7, 2018

Outline

- Introduction to auto-editing at BEA
- Proposal of simulation-based testing framework
- Results regarding how successfully the simulation mimics reality and the accuracy of auto-editing imputations
- Conclusions

Introduction: Auto-Editing at BEA

- Focused on annual direct investment surveys, which collect financial and operating data from:
  - U.S. multinational enterprises and their foreign affiliates
  - Foreign-owned U.S. companies
- Motivation: allow editors to spend more time on most complex/impactful responses, improve general efficiency of survey editing

Approach to Auto-Editing

- Implementation of Banff system for data editing and imputation
- Key procedures:
  - Error localization
  - Donor imputation
  - Estimator imputation

The Research Question

- How should auto-editing be evaluated?
  - BEA's current approach: compare to results of manual editing
  - Ideal approach: compare to true values

New Framework

- Find “clean” forms
- Simulate missing/erroneous data
- Impute
- Compare imputations to reported values

Testing of New Framework

- Data: 2015 BE-15C (8 numeric items)
- Key Issues:
  - Proximity of imputed values to reported values
  - Comparison of different versions of imputation procedures

Simulation Set-Up

- Problem: how to mimic actual distribution of missing/erroneous responses in simulated data?

- Solution: model likelihood of the *j*=1,…,8 numeric items on the *i*=1,…,*n* forms being missing/erroneous

$$E[Y_{ij}] = \frac{\exp(\mathbf{X}'_{ij}\boldsymbol{\beta})}{1 + \exp(\mathbf{X}'_{ij}\boldsymbol{\beta})}$$

Simulation

- Each item receives an estimated probability of being a “field to impute” (FTI), $p_{ij} = E\left[Y_{ij}\right]$
- In each simulation run, each item’s status is based on its p_{ij}
- 5,000 runs

How realistic are the simulated data?

- FTIs per form:
  - Actual data: 0.234
    Simulated data: 0.237

Distribution of FTIs among Survey Items in Actual vs. Simulated Data

Field Selected as FTI	Observed Average Percent Share	Simulated Average Percent Share
Assets	0.3	0.3
Liabilities	1.6	1.8
Sales	23.2	23.6
Net Income	4.5	5.5
Employee Compensation	23.9	25.2
Gross PP&E	18.2	16.2
R&D	9.2	8.2
Employees	19.1	19.2

The Tests

- Two versions of auto-editing system tested:

  1. Base settings

  2. Additional years of data used for donor and estimator imputation

Measuring the Accuracy of Imputations

- Average percent difference between actual and estimated aggregate value:

$$\bar{y}_j = \frac{\sum_{k=1}^{5,000}\left[\left(\left[\frac{\sum_{i=1}^{n} s_{ijk}}{\sum_{i=1}^{n} o_{ij}}\right]-1\right)\times 100\right]}{5,000}$$

- Average absolute percent difference between actual and estimated aggregate value:

$$\bar{x}_j = \frac{\left[\frac{\sum_{k=1}^{5,000}\left[\frac{\sum_{l=1}^{m_{jk}}\left|s_{l(jk)}-o_{l(j)}\right|}{m_{jk}}\right]}{5,000}\right]\times 100}{\sum_{i=1}^{n} o_{ij}}$$

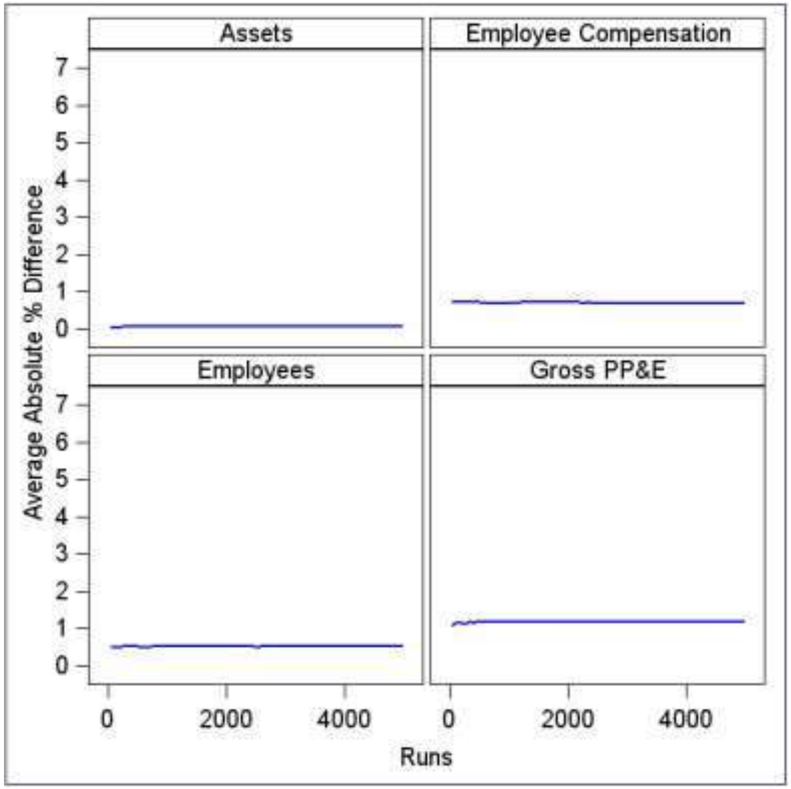
Summary of Test Results

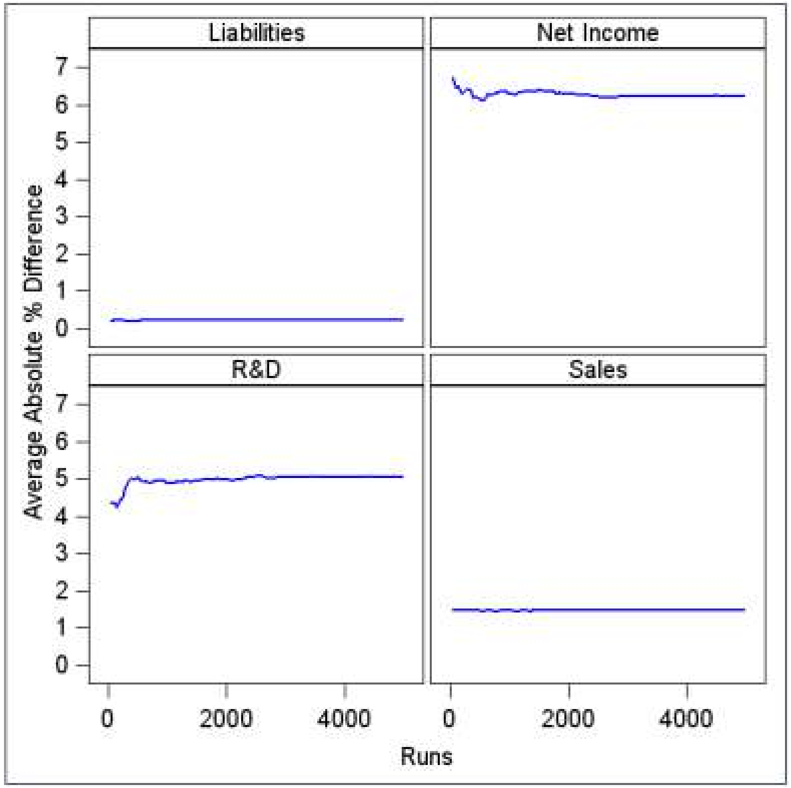
Accuracy of Imputations by Field and Test

	First Test		Second Test	
Field	Avg. % Diff. ($\bar{y}$)	Avg. Abs. % Diff. ($\bar{x}$)	Avg. % Diff. ($\bar{y}$)	Avg. Abs. % Diff. ($\bar{x}$)
Assets	-0.01	0.06	-0.01	0.05
Liabilities	0.01	0.23	-0.01	0.20
Sales	0.09	1.47	0.08	1.31
Net Income	-0.04	6.22	0.35	5.84
Employee Compensation	-0.08	0.70	-0.23	0.71
Gross PP&E	0.12	1.17	0.26	1.20
R&D	-1.81	5.04	-2.05	3.88
Employees	0.01	0.51	-0.03	0.52

Are 5,000 runs enough?

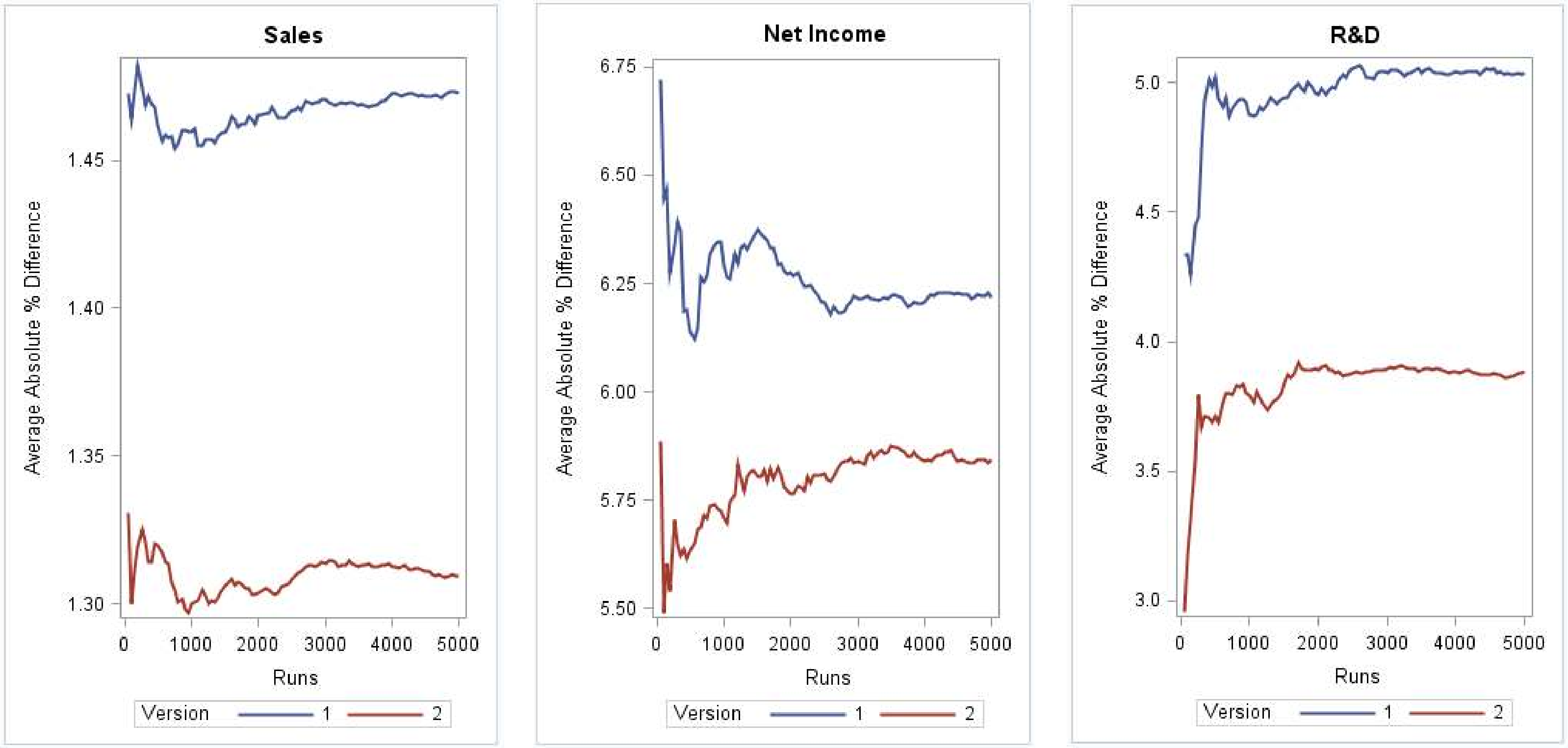
The Number of Runs and the Measurement of Imputations' Accuracy





Comparing Versions 1 and 2

Stability of Differences Between Versions 1 and 2 of Imputation Procedures



Summary and Conclusions

- Proposed new method for assessing BEA's auto-editing systems
- Found close agreement between imputed values and reported values
- Identified means of improving imputation procedures

Contact Information

- Questions on the presentation?
  - –Larkin Terrie: Larkin.Terrie@bea.gov

- Questions on BEA's direct investment statistics?
  - –Internationalaccounts@bea.gov

Thank You!