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

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
• 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

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

• Each item receives an estimated probability of being a “field to impute” (FTI), $p_{ij} = E[Y_{ij}]$

• 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

<table>
<thead>
<tr>
<th>Field Selected as FTI</th>
<th>Observed Average Percent Share</th>
<th>Simulated Average Percent Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assets</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Liabilities</td>
<td>1.6</td>
<td>1.8</td>
</tr>
<tr>
<td>Sales</td>
<td>23.2</td>
<td>23.6</td>
</tr>
<tr>
<td>Net Income</td>
<td>4.5</td>
<td>5.5</td>
</tr>
<tr>
<td>Employee Compensation</td>
<td>23.9</td>
<td>25.2</td>
</tr>
<tr>
<td>Gross PP&amp;E</td>
<td>18.2</td>
<td>16.2</td>
</tr>
<tr>
<td>R&amp;D</td>
<td>9.2</td>
<td>8.2</td>
</tr>
<tr>
<td>Employees</td>
<td>19.1</td>
<td>19.2</td>
</tr>
</tbody>
</table>
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:

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

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

\[
\bar{x}_j = \frac{5,000 \left[ \frac{\sum_{l=1}^{m} \left| S_{l(jk)} - o_{l(j)} \right|}{m_{jk}} \right]}{\sum_{i=1}^{n} o_{ij}} \times 100
\]
## Summary of Test Results

### Accuracy of Imputations by Field and Test

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

- **Sales**
  - Graph showing the average absolute % difference over runs.

- **Net Income**
  - Graph showing the average absolute % difference over runs.

- **R&D**
  - Graph showing the average absolute % difference over runs.
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!