Evaluating Imputation Methods using ImpACT

First Case Study

UNECE Workshop on Statistical Data Editing
August 31, 2020

Delivering insight through data for a better Canada

Darren Gray
Statistics Canada
Introduction

• ImpACT: the Imputation Assessment and Comparison Tool
• Builds on similar work at Statistics Canada by Haziza and Stelmack
• Underlying framework presented at JSM (2019) using synthetic data
• Today’s presentation covers first case study using survey data
ImpACT

- Objective: provide survey methodologists with an easy-to-use, generalized tool to test and compare imputation methods in a controlled, simulation environment
- Consists of three modules: non-response, imputation and analysis
- Combining non-response patterns with specific analysis types can lead to a variety of useful insights
ImpACT Framework

Original data

User-provided
User-specified

Non-response
Non-response pattern

Imputation

Imputed data

Analysis
Three types of analysis:
• Distributional accuracy
• Estimation accuracy
• Prediction accuracy
Retail Commodity Survey (RCS)

- Monthly survey measuring commodity sales
- Estimates broken down by commodity using North American Product Classification System (NAPCS)
- For each record, commodity sales obey additive structure
Retail Commodity Survey (RCS)

- Multi-stage data editing approach
- Missing values not imputed in one stage pass on to subsequent stages
- Altered imputation strategy for simulation purposes
- Study plan included a general assessment and sub-method comparison

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deductive imputation</strong></td>
<td></td>
</tr>
<tr>
<td><strong>S1</strong></td>
<td>Historical imputation (previous year)</td>
</tr>
<tr>
<td><strong>S2</strong></td>
<td>Historical imputation (previous month)</td>
</tr>
<tr>
<td><strong>S3</strong></td>
<td>Donor imputation <strong>(modified)</strong></td>
</tr>
<tr>
<td><strong>S4</strong></td>
<td>Ratio imputation (first set of imputation classes)</td>
</tr>
<tr>
<td><strong>S5</strong></td>
<td>Ratio imputation (second set of imputation classes)</td>
</tr>
<tr>
<td><strong>S6</strong></td>
<td>Percentage imputation</td>
</tr>
<tr>
<td></td>
<td><strong>Pro-rating</strong></td>
</tr>
</tbody>
</table>
First study: general assessment

- **Goal**: Assess performance of overall strategy under simulated missing-completely-at-random (MCAR) non-response
- Three simulations (500 trials each) at non-response rates of 10%, 30% and 50%
- Within each trial, fixed non-response size and coordinated to ensure records were imputed the same number of times
- Analysis types presented: *distributional* and *estimation* accuracy
Distributional accuracy

- Compare distribution of missing values to imputed values, aggregated over all trials
- Options include kernel distribution, jitter plot, boxplot and mean comparison
- Combination of different elements gives a comprehensive picture
- (Note: values displayed on log scale)
Estimation accuracy

• Calculate mean from original values
• Compare to mean generated by each trial
• Results show negative bias that increases in magnitude as non-response increases
• Investigation of clustering showed sensitivity to outliers
Estimation accuracy

- Removed outliers, repeated simulation
- Drastically reduced bias and variance
First study: inferential limitations

• Imputation rate by imputation stage showed that the simulations did not accurately reflect the true imputation process
• This can indicate a non-representative training set, or inaccurate non-response mechanism
• Rate of missingness for historical data differed between training set and the actual target of imputation

<table>
<thead>
<tr>
<th>Stage</th>
<th>Simulation non-response rate</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10%</td>
<td>30%</td>
</tr>
<tr>
<td>S1</td>
<td>72.7</td>
<td>72.7</td>
</tr>
<tr>
<td>S2</td>
<td>25.5</td>
<td>25.5</td>
</tr>
<tr>
<td>S3</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>S4</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>S5</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>S6</td>
<td>0.3</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Second study: sub-method comparison

- **Goal**: Investigate sub-methods using a leave-one-out cross-validation approach
- One trial for each record
- Analysis type presented: **predictive accuracy**
- Challenge: comparing methods that only impute a subset of all records
Predictive accuracy

- Compare imputation errors against original values (log scale) across all methods
- Imputation rates vary by method
- Only stage raising concerns is the final one (percentage imputation)
- Should we change the order?
Predictive accuracy

- Comparison of historical ratio methods
- Use of previous month data outperforms previous year (could be specific to our reference period)
Predictive accuracy

- Compare donor imputation to first run of ratio imputation
- Ratio imputation method outperforms donor when limited to records imputed by both methods
- Donor imputation has other benefits
Conclusions

• ImpACT can provide insights into behaviour of survey imputation process
  • Insights into accuracy of imputation strategy
  • Lead to further investigation
  • Users should be aware of inferential limitations

• Challenges posed by survey data and complex imputation design:
  • Missingness in auxiliary variables
  • Multi-stage imputation strategy required new assessment tools
  • Multivariate sub-method (donor imputation) required modification
THANK YOU!

Contact:
darren.gray@canada.ca

*The content of this presentation represents the position of the author and may not necessarily represent that of Statistics Canada.*