Computational estimates of data editing variance

Mark van der Loo, Jeroen Pannekoek, Lisanne Rijnveld
Data editing process

- Data editing

- Estimation

- 'Naive' estimator $\hat{V}(\hat{\theta})$

- Full variance estimator $\hat{V}(\hat{\theta})$
Estimation of variance

Example: naive estimation of variance of the mean

\[ \hat{V}'(\hat{\theta}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2 \]

Total variance includes effects of data editing

\[ V(\hat{\theta}) = E_\delta V(\hat{\theta}|\delta) + V_\delta E(\hat{\theta}|\delta) \]
Bootstrap approach

\[
\hat{\theta} = \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_j
\]

\[
\hat{\sigma}^2 (\hat{\theta}^*) = \frac{1}{B} \sum_{j=1}^{B} (\hat{\theta}_j - \hat{\theta}^*)^2
\]
The applied data editing procedure

Unit of measure error correction
↓
Correct typing errors
↓
Correct rounding errors
↓
Error localization
↓
Deductive imputation
↓
Ratio imputation
↓
Median imputation
↓
Adjust to edits

In bootstrap
Convergence of the mean

Bootstrap convergence of the mean

- net turnover
- other income
- total income
- opr expenses
- opr income
- taxresult

nr of bootstrap samples

0.95
0.96
0.97
0.98
0.99
1.00
1.01
1.02
### Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>bootstrap mean</th>
<th>percent bias</th>
<th>bootstrap stdev</th>
<th>naive stdev</th>
<th>percent diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>net turnover</td>
<td>27181.0</td>
<td>0.03</td>
<td>2371.5</td>
<td>2364.3</td>
<td>0.6</td>
</tr>
<tr>
<td>other income</td>
<td>130.8</td>
<td>2.83</td>
<td>75.1</td>
<td>75.4</td>
<td>-0.7</td>
</tr>
<tr>
<td>total income</td>
<td>27328.0</td>
<td>0.05</td>
<td>2388.8</td>
<td>2378.1</td>
<td>0.9</td>
</tr>
<tr>
<td>opr expenses</td>
<td>21938.0</td>
<td>-0.09</td>
<td>2058.4</td>
<td>2096.7</td>
<td>-3.8</td>
</tr>
<tr>
<td>opr income</td>
<td>758.4</td>
<td>1.66</td>
<td>144.7</td>
<td>134.2</td>
<td>13.9</td>
</tr>
<tr>
<td>taxresult</td>
<td>725.8</td>
<td>2.35</td>
<td>176.2</td>
<td>162.5</td>
<td>14.9</td>
</tr>
</tbody>
</table>
Convergence of variance

Bootstrap convergence of standard deviation

nr of bootstrap
net turnover
other income
total income
opr expenses
opr income
taxresult

nr of bootstrap taps
<table>
<thead>
<tr>
<th>Variable</th>
<th>bootstrap mean</th>
<th>percent bias</th>
<th>bootstrap stdev</th>
<th>naive stdev</th>
<th>percent diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>net turnover</td>
<td>27181.0</td>
<td>0.03</td>
<td>2371.5</td>
<td>2364.3</td>
<td>0.6</td>
</tr>
<tr>
<td>other income</td>
<td>130.8</td>
<td>2.83</td>
<td>75.1</td>
<td>75.4</td>
<td>-0.7</td>
</tr>
<tr>
<td>total income</td>
<td>27328.0</td>
<td>0.05</td>
<td>2388.8</td>
<td>2378.1</td>
<td>0.9</td>
</tr>
<tr>
<td>opr expenses</td>
<td>21938.0</td>
<td>-0.09</td>
<td>2058.4</td>
<td>2096.7</td>
<td>-3.8</td>
</tr>
<tr>
<td>opr income</td>
<td>758.4</td>
<td>1.66</td>
<td>144.7</td>
<td>134.2</td>
<td>13.9</td>
</tr>
<tr>
<td>taxresult</td>
<td>725.8</td>
<td>2.35</td>
<td>176.2</td>
<td>162.5</td>
<td>14.9</td>
</tr>
</tbody>
</table>

blue = ‘significant’ considering bootstrap convergence.
Summary / lookout

Conclusions

- Effect of variance including DE can be estimated with bootstrap
- For skewly distributed data, $B \gg n$ but by tracking convergence, conclusions can be drawn.
- We found larger DE variance for variables of lower quality

To improve

- Simple bootstrap not accurate for complex survey design