UNECE Work Session on Statistical Data Editing, Oslo, 2012

PROBABILITY EDITING

Thomas Laitila\textsuperscript{1} and Maiki Ilves\textsuperscript{2}

\textsuperscript{1)Statistics Sweden and Örebro university}
\textsuperscript{2)Statistics Estonia}
PAPERS


IDEA

- Take a random sample of respondents for editing
  
  - Conditionally on the respondents, a random sample give information on the measurement errors among responses

  - Unconditionally, information on measurement errors among responses can be generalized for measurement bias adjustment of statistics (estimates)
ADVANTAGES

• Statistical properties of estimators can be studied with standard methodologies

• Measurement bias of estimators can be removed

• High flexibility
  o Prior information not necessary
  o Prior information can be used if available
  o Type of variable is no issue

• Connects estimation and editing

• Bias adjustment possible in secondary use of data
DISADVANTAGES

• May leave obviously erroneous observations unedited
  o Remedy: Combine selective and probability editing into the two-step procedure (Ilves and Laitila, 2009)

• New developments required
  o Theory (estimation and sampling designs)
  o Implementation, software
ILVES AND LAITILA (2009)

- HT-estimator, two-step procedure (= selective + probability editing)
- Theoretical results
- Simulation study
  - True values generated from a Po(5) distribution.
  - Erroneous observations generated from either a Po(2) or a Po(10) distribution
  - Rate of measurement errors = 0.4
    - Case I = erroneous observations are taken from Po(2) or Po(10) with equal probability (0.2).
    - Case II = all erroneous observations are generated from the Po(2) distribution.
Table 1: Relative bias and RMSE ratio of the HT estimator after selective editing (SE) and editing using a two-step procedure (TP), respectively. (Case I).

<table>
<thead>
<tr>
<th>Correlation between score values and observations</th>
<th>SE Edited</th>
<th>Rel. Bias</th>
<th>TP Edited</th>
<th>Rel. Bias</th>
<th>RMSE ratio (TP/SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>12%</td>
<td>2%</td>
<td>4%</td>
<td>0%</td>
<td>2.03</td>
</tr>
<tr>
<td>0.7</td>
<td>16%</td>
<td>2%</td>
<td>4%+12%</td>
<td>0%</td>
<td>1.88</td>
</tr>
<tr>
<td>0.5</td>
<td>14%</td>
<td>4%</td>
<td>6%+8%</td>
<td>0%</td>
<td>1.36</td>
</tr>
<tr>
<td>0.3</td>
<td>17%</td>
<td>6%</td>
<td>7%+10%</td>
<td>0%</td>
<td>1.02</td>
</tr>
</tbody>
</table>
ILVES AND LAITILA (2009) (cont’d)

Table 2: Relative bias and RMSE ratio of the HT estimator after selective editing (SE) and editing using a two-step procedure (TP), respectively. (Case II).

<table>
<thead>
<tr>
<th>Correlation between score values and observations</th>
<th>1.0</th>
<th>0.7</th>
<th>0.5</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE Edited</td>
<td>14%</td>
<td>14%</td>
<td>15%</td>
<td>17%</td>
</tr>
<tr>
<td>Rel. Bias</td>
<td>-18%</td>
<td>-21%</td>
<td>-20%</td>
<td>-20%</td>
</tr>
<tr>
<td>TP Edited</td>
<td>4% + 10%</td>
<td>4%+10%</td>
<td>4%+11%</td>
<td>5%+12%</td>
</tr>
<tr>
<td>Rel. Bias</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>RMSE ratio (TP/SE)</td>
<td>0.26</td>
<td>0.23</td>
<td>0.23</td>
<td>0.22</td>
</tr>
</tbody>
</table>
ILVES (2012a, 2012b)

- Theory extension
  - Probability editing and GREG estimation
  - Probability editing and calibration estimation for nonresponse
    - Adjustments for both measurement errors and nonresponse
    - (+ coverage errors)

- Theory developed for “pure” probability editing
  - Covers the two-step procedure as selective editing can be incorporated using selection probabilities = 1.

- Empirical illustrations
  - Overall good performance of suggested estimators under probability editing
  - Competitive variance=\text{MSE}
FURTHER TOPICS

• Poisson sampling using scores, how to scale selection probabilities?

• Monitoring the editing process for
  o Determining the (expected) number of observations edited
  o Score function construction

• Composite estimators

• Macro Editing
Thanks for Your time!
MACRO EDITING (Easy way)

Unbiased estimator under probability editing

\[ \hat{t}_z = \hat{t}_y - \hat{b}_y = \sum_r w_k y_k - \sum_{r2} w_{2k} (y_k - z_k) \]

\[ E(\hat{t}_z) = t_z \]

Set of observations under macro editing, determined non-randomly after probability editing

\[ r_{me} \subset r - r_2 \]

Estimator after macro editing

\[ \hat{t}_{zme} = \hat{t}_z - \sum_{r_{me}} w_k (y_k - z_k) \]
MACRO EDITING (Easy way, cont’d)

Since $\hat{t}_z$ is unbiased $\Rightarrow$ $\hat{t}_{zme}$ is biased

$\Rightarrow$ adjust the bias correction $\hat{b}_y$ with the same amount

$\Rightarrow$ $\hat{t}_z$ is not effected by macro editing
MACRO EDITING (Hard way)

- The set $r_{me}$ is random, determined by the random set $r_2$
  - Previous argument still holds, but opens an opportunity.

- A unit is edited either if it is drawn to the set $r_2$ or if it ends up in $r_{me}$.
  - Define new weights $w_{2mek}$ reflecting the probability of being in either $r_2$ or in $r_{me}$.
  - Redefine the estimator as

$$\hat{t}_{z, me} = \hat{t}_y - \hat{b}_{y, me} = \sum_r w_k y_k - \sum_{(r_2+r_{me})} w_{2mek} (y_k - z_k)$$

$$E(\hat{t}_{z, me}) = t_z$$

- Challenge:

Can $w_{2mek}$ be calculated?
Thanks for Your additional time!