SELECTIVE EDITING AS PART OF THE
ESTIMATION PROCEDURE

1 Marco Di Zio 1 Ugo Guarnera 1

1Istituto Nazionale di Statistica
guarnera@istat.it

Work Session on Statistical Data Editing
Oslo, 23-26 September 2012
OUTLINE

1. SELECTIVE EDITING
2. PROBABILISTIC APPROACH TO EDITING
3. MONTE CARLO STUDY
4. RESULTS
5. CONCLUSIONS
OUTLINE

1. SELECTIVE EDITING
2. PROBABILISTIC APPROACH TO EDITING
3. MONTE CARLO STUDY
4. RESULTS
5. CONCLUSIONS
selective editing as part of an estimation procedure

OUTLINE

1. SELECTIVE EDITING
2. PROBABILISTIC APPROACH TO EDITING
3. MONTE CARLO STUDY
4. RESULTS
5. CONCLUSIONS
OUTLINE

1 SELECTIVE EDITING

2 PROBABILISTIC APPROACH TO EDITING

3 MONTE CARLO STUDY

4 RESULTS

5 CONCLUSIONS
OUTLINE

1. SELECTIVE EDITING
2. PROBABILISTIC APPROACH TO EDITING
3. MONTE CARLO STUDY
4. RESULTS
5. CONCLUSIONS
Techniques for identification of influential errors

The goal is to split survey data into 2 sets:

- **critical set** containing potentially influential errors to be interactively analyzed
- **non-critical set** composed of records that are correct or affected by non-influential errors. These cases are treated through automatic procedures.
SELECTIVE EDITING 2

Underline assumptions:

- estimation error is mainly due to a few gross errors in data (influential errors)

- influential errors can be detected by analyzing discrepancies between observed and "predicted" values

- interactive editing is able to recover "true values" (perfect editing)
DEPARTURE FROM ASSUMPTIONS

- The (possibly implicit) model used to predict plausible values may be inadequate.

- High number of non-influential errors (inliers) may remain in data also after selective editing and deteriorate the accuracy of the target estimates.
BIAS ESTIMATION


Their approach is based on a two-step procedure where selective editing is performed at first step and unedited observations are sampled at second step.
**TWO-PHASE ESTIMATOR**

\[ y^*_k = \text{true data}; \quad y_k = \text{observed data}; \quad I_{k}^{ed} = \text{editing indicator}; \]

\[ \tilde{y}_k = I_{k}^{ed} y^*_k + (1 - I_{k}^{ed}) y_k \]

population total:

\[ t(y^*) = \sum_{k \in U} y^*_k; \]

estimator:

\[ \hat{t}_y = \sum_{k \in s_a} \frac{\tilde{y}_k}{\pi_{ak}} - \sum_{k \in s_b} \frac{e_k}{\pi_{ak} \pi_{b|ak}} \]

\( \pi_{ak}, \pi_{b|ak} \) are the first order inclusion probability for unit \( k \) in the first and second phase samples \( (s_a, s_b) \) respectively and \( e_k = y_k - y^*_k \) is the observed error on sampled unit \( k \).
BIAS VS VARIANCE

- The two step estimator is unbiased but the bias correction term produces increase in variance. We need to analyze the trade off between bias and variance.

- Variance can be reduced if an efficient sampling strategy is used at second phase.

- For instance, (sub)sampling design can take advantage of the scores used for selective editing.
selective editing as part of an estimation procedure
probabilistic approach to editing

PPS SAMPLING USING SCORE FUNCTION

- a possible strategy is to use the score function as auxiliary variable for pps sampling

- in case of SE based on contamination model (as implemented in R-package SeleMix) scores are strictly related to expectation of errors

- not edited units with largest errors are (hopefully) sampled with higher probability. This should result in an efficient estimation of the total error remaining in data after editing
SCORE FUNCTION IN SELEMIX

\( y_k^* \sim N(\mu; \Sigma), \quad y_k|y_k^* \sim \pi \delta(y_k - y_k^*) + (1 - \pi) N(y^*; \alpha \Sigma) \)

The score \( S_k \) is based on the estimate of the expected error \( E_k = E(y_k^*|y_k) - y_k \) w.r.t. the conditional distribution

\[
 f_{Y^*|Y}(y^*|y) = \tau_1(y) \delta(y^* - y) + \tau_2(y) N(y^*; \tilde{\mu}_y, \tilde{\Sigma})
\]

where \( \tau_1 \) and \( \tau_2 \) are the posterior probabilities of being erroneous

\[
 \tau_1(y_k) = Pr(y_k = y_k^*|y_k), \quad \tau_2(y_i) = Pr(y_k \neq y_k^*|y_k)
\]

and

\[
 \tilde{\mu}_y = \frac{(y + (\alpha - 1)\mu)}{\alpha}; \quad \tilde{\Sigma} = \left(1 - \frac{1}{\alpha}\right) \Sigma
\]
**EXPERIMENTAL SET UP**

- **reference population** \(U\): data from Istat survey on small and medium enterprises \(N = 8500\)
- **target estimate**: total of *labour cost* \(Y\). *turnover* is used as auxiliary variable \(X\)
- **error mechanisms**: only \(Y\) is contaminated with multiplicative error \((\times 10, 100, 1000)\), first or last two digit permutation, “1” error (values of variable \(Y\) set to 1)
- **sample** \(s_a\): SRSWOR \(n = 1000\)
- **selective editing**: scores are computed with SeleMix and \(n_{ed}\) units with the highest score are edited and corrected when \(y_k^* \neq y_k\)
- **subsample** \(s_b\): \(n_b\) units are subsampled from unedited obs of \(s_a\) (both SRSWOR and pps) for bias correction
ESTIMATORS

\[ \hat{t}_{y^*} = \frac{N}{n_a} \sum_{k \in s_a} y^*_k; \quad \hat{t}_y = \frac{N}{n_a} \sum_{k \in s_a} y_k \]

editing + bias correction:

\[ \hat{t}^{2\text{step}} = \frac{N}{n_a} \sum_{k \in E} y^*_k + \frac{N}{n_a} \sum_{k \in s_a \setminus E} y_k + \frac{N}{n_a} \sum_{k \in s_b} (y^*_k - y_k) \pi_{b|a} \]

where \( E \) is the set of the \( n_{ed} \) edited units.

\[ \hat{t}^{2\text{step}} \rightarrow \begin{cases} \hat{t}^{se} & s_b = \emptyset \\ \hat{t}^{pps}, \hat{t}^{srs} & E = \emptyset \\ \hat{t}_2^{pps}, \hat{t}_2^{srs} & n_{ed} > 0, \ n_b > 0. \end{cases} \]
selective editing as part of an estimation procedure

RESULTS (1000 MC runs)

\( n_{rec} = n_{ed} + n_b \) is fixed

errors: \( \times 1000(0.5\%), \times 100(1\%), \) digit perm.(1\%), "1" error (1\%). \( \times 10 \)
error varies across experiments. In \( \hat{t}_2 \) estimators \( n_{ed} \) is chosen by SeleMix based on accuracy parameter \( \eta \)

<table>
<thead>
<tr>
<th>experiment</th>
<th>( \hat{t}_y )</th>
<th>( \hat{t}^{se} )</th>
<th>( \hat{t}^{srs} )</th>
<th>( \hat{t}^{pps} )</th>
<th>( \hat{t}_2 )</th>
<th>( \hat{t}^{pps}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_{rec} = 30 )</td>
<td>RRMSE%</td>
<td>4</td>
<td>611</td>
<td>4</td>
<td>1649</td>
<td>5</td>
</tr>
<tr>
<td>( \times 10(0%) )</td>
<td>RB%</td>
<td>0</td>
<td>545</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( n_{rec} = 150, \eta = 1 )</td>
<td>RRMSE%</td>
<td>4</td>
<td>837</td>
<td>13</td>
<td>999</td>
<td>10</td>
</tr>
<tr>
<td>( \times 10(15%) )</td>
<td>RB%</td>
<td>0</td>
<td>743</td>
<td>11</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( n_{rec} = 80, \eta = 2 )</td>
<td>RRMSE%</td>
<td>4</td>
<td>634</td>
<td>36</td>
<td>1000</td>
<td>17</td>
</tr>
<tr>
<td>( \times 10(15%) )</td>
<td>RB%</td>
<td>0</td>
<td>576</td>
<td>35</td>
<td>-1</td>
<td>0</td>
</tr>
</tbody>
</table>
CONCLUSIONS

- When data contain a few gross errors, selective editing with SeleMix performs better than a two-step approach.

- In presence of many small errors (inliers), bias correction can improve the estimate provided that we can resample a sufficient number of units.

- PPS sampling using SeleMix score as size-variable is much more efficient than SRSWOR.

- For a fixed number $n_{rec} = n_{ed} + n_b$ of observations to be interactively reviewed, the problem of trade-off between bias and variance can be viewed as the problem of optimally choosing $n_{ed}$ and $n_b$ (minimum MSE).


selective editing as part of an estimation procedure

conclusions

THE END

Thank you