UNECE Work Session on Statistical Data Editing, Oslo, 2012

PROBABILITY EDITING

Thomas Laitila[1] and Maiki Ilves[2]

[1)]Statistics Sweden and Örebro university
[2)]Statistics Estonia

PAPERS

Ilves, M. and T. Laitila (2009). Probability-Sampling Approach to Editing, *Austrian Journal of Statistics*, 38:3, 171-182.

Ilves, M. (2012a). GREG estimation and probabilistic editing, *Metron*, (To appear)

Ilves, M. (2012b). Estimation in the presence of nonresponse and measurement errors (Submitted).

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
  - Prior information not necessary
  - Prior information can be used if available
  - 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
  - Remedy: Combine selective and probability editing into the two-step procedure (Ilves and Laitila, 2009)

- New developments required
  - Theory (estimation and sampling designs)
  - 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.

ILVES AND LAITILA (2009) (cont'd)

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).

		Correlation between score values and observations			
		1.0	0.7	0.5	0.3
SE	Edited	12%	16%	14%	17%
	Rel. Bias	2%	2%	4%	6%
TP	Edited	4% + 8%	4%+12%	6%+8%	7%+10%
	Rel. Bias	0%	0%	0%	0%
	RMSE ratio (TP/SE)	2.03	1.88	1.36	1.02

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).

		Correlation between score values and observations			
		1.0	0.7	0.5	0.3
SE	Edited	14%	14%	15%	17%
	Rel. Bias	-18%	-21%	-20%	-20
TP	Edited	4% + 10%	4%+10%	4%+11%	5%+12%
	Rel. Bias	0%	0%	0%	0%
	RMSE ratio (TP/SE)	0.26	0.23	0.23	0.22

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=MSE

FURTHER TOPICS

- Poisson sampling using scores, how to scale selection probabilities?
- Monitoring the editing process for
  - Determining the (expected) number of observations edited
  - Score function construction
- Composite estimators
- Macro Editing

Thanks for Your time!

MACRO EDITING (Easy way)

Unbiased estimator under probability editing

$$\hat{t}_z = \hat{t}_y - \hat{b}_y = \sum_r w_k y_k - \sum_{r_2} w_{2k}(y_k - z_k)$$

$$E(\hat{t}_z) = t_z$$

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

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

$$E(\hat{t}_z^{me}) = t_z$$

- Challange:

Can w_{2mek} be calculated?

Thanks for Your additional time!