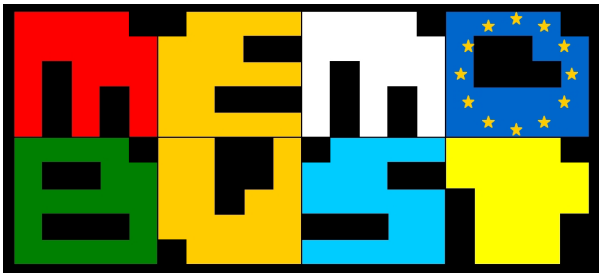

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Selective Editing

Contents

General section

1. Summary

The experience of NSIs in the field of correction of errors has led to assume that only a small subset of observations is affected by influential errors, i.e., errors with a high impact on the estimates, while the rest of the observations are not contaminated or contain errors having small impact on the estimates. Selective editing is a general approach to the detection of errors, and it is based on the idea of looking for important errors in order to focus the treatment on the corresponding subset of units to reduce the cost of the editing phase, while maintaining the desired level of quality of estimates. In this section a general description of the framework and the main elements of selective editing is given.

2. General description

2.1 Selective editing

The experience of NSIs in the field of correction of errors has led to assume that only a small subset of observations is affected by influential errors, i.e., errors with a high impact on the estimates, while the rest of the observations are not contaminated or contain errors having small impact on the estimates (Hedlin, 2003). This assumption and the fact that the interactive editing procedures, like for instance, recontact of respondents, are resource demanding, have motivated the idea at the basis of selective editing, that is to look for important errors (errors with an harmful impact on estimates) in order to focus the expensive interactive treatments (follow-up, recontact) only on this subset of units. This should reduce the cost of the editing phase maintaining at the same time an acceptable level of quality of estimates (Lawrence and McKenzie, 2000; Lawrence and McDavitt, 1994). In practice, observations are ranked according to the values of a *score function* expressing the impact of their potential errors on the target estimates (Latouche and Berthelot, 1992), and all the units with a score above a given threshold are selected.

2.2 Score function

The score function is an instrument to prioritise observations according to the expected benefit of their correction on the target estimates. According to this definition, it is natural to think of the score function as an estimate of the error affecting data. The estimate is generally based on comparing observed values with predictions (sometimes called *anticipated values*) obtained from some explicit or implicit model for the data. In the case of sample surveys, the comparison should also include the sampling weights in order to properly take into account the error impact on the estimates. An additional element often considered in the context of selective editing, is the *degree of suspiciousness,* that is an indicator measuring, loosely speaking, the probability of being in error. The necessity of this element arises from the implicit assumption of the intermittent nature of the error in survey data, i.e., the assumption that only a certain proportion of the data are affected by error, or, from a probabilistic perspective, that each measured value has a certain probability of being erroneous (Buglielli et al., 2011). Some authors do not introduce this element, others implicitly use it in their proposals. Norberg et al. (2010) state that several case studies indicate that procedures based only on the comparison of observed and predicted values without the use of a degree of suspiciousness tend to generate a large proportion of false alarm.

Several score functions are proposed in literature, the difference being mainly given by the kind of prediction and the use of 'degree of suspiciousness'.

Among the different methods used to obtain predictions it is worthwhile to mention the use of information coming from a previous occasion of the survey (Latouche and Berthelot, 1992), regression models (Norberg et al., 2010), contamination models (Buglielli et al., 2011). A detailed review can be found in De Waal et al. (2011).

As far as the degree of suspiciousness is concerned, a common drastic approach consists in introducing it in the score function through a zero-one indicator that multiplies the difference between observed and predicted values, where zero and one correspond to consistency or inconsistency respectively with respect to some edit rules. In this case it is assumed that errors appear only as edit failures and observations that pass the edits are considered error-free without uncertainty (Latouche and Berthelot, 1992). More refined methods to estimate the probability of being in error can be found in Norberg et al. (2010) and Buglielli et al. (2011). In the first case a nonparametric approach based on quantiles is used, while in the second a latent model based on a mixture of normal (or lognormal) distributions is proposed.

Prediction and suspiciousness can be combined to form a score for a single variable, named *local score*. A local score frequently used for the unit i with respect to the variable Y_j is

$$S_{ij} = \frac{p_i w_i \mid y_{ij} - \tilde{y}_{ij} \mid}{\hat{T}_{Y_j}}$$

where p_i is the degree of suspiciousness, y_{ij} is the observed value of the variable Y_j on the ith unit, $\tilde{y}_{ij}$ is the corresponding prediction, w_i is the sampling weight, and $\hat{T}_{Y_j}$ is an estimate of the target parameter.

Once the local scores for the variables of interest are computed, a global score to prioritise observations is needed.

Several functions can be used to obtain the global score (see Hedlin, 2008); an example is the sum of squares $GS_i^{(2)} = \sum_j S_{ij}^2$.

In some cases, some variables can be considered to be more important than others. Such situations can be dealt with by multiplying the local scores by weights stating their relative importance.

2.3 The selection rule

Once the observations have been ordered according to their global score, it is important to build a rule in order to determine the number of units to be reviewed.

A first rule can be suggested by budget constraints. In this case, it is obvious to choose the first n^* observations, in the given ordering, such that the budget constraints are satisfied.

A more interesting and complex approach is to select the subset of units such that the impact on the target estimates of the errors remaining in the unedited observations is negligible, that is in fact the core of selective editing. Since the true values are unknown, this bias cannot be evaluated and an approximation is used. This approximation can be expressed in terms of the weighted differences

between the raw values y_{ij} and the anticipated values $\tilde{y}_{ij}$ for the variable Y_j in the units i not selected for interactive treatment (EDIMBUS, 2007).

Let T_{Yj} be the target quantity related to the variable Y_j (for instance the total), the estimated bias is given by

$$EB_j(t) = \frac{|\sum_{i \notin E_t} w_i (y_{ij} - \tilde{y}_{ij})|}{\hat{T}_{Y_j}},$$

where w_i is the sampling weight of the ith unit, $\hat{T}_{Y_j}$ is an estimate of the target quantity T_{Y_j}, and E_t is the set of units to be selected. This set is composed of all the units having a global score $GS > t$, where t is a threshold value such that $EB_j(t)$ is below a predefined value.

An alternative measure known as the *estimated relative bias* is obtained by replacing the estimate of the total at the denominator of EB with the standard error of the estimate $\hat{T}_{Y_j}$. With this measure, the error due to the non-sampling error left in data is compared with the sampling error. The reasoning underlying is that there is no need to edit observations because the 'noise' due to their errors is overwhelmed by the sampling error.

We remark that when edited values are available, they can be used as anticipated values, in this case the estimated bias and the estimated relative bias are the absolute pseudo bias and the relative pseudo-bias introduced by Latouche and Berthelot (1992) and Lawrence and McDavitt (1994), respectively.

It is worthwhile to note the similarity between the terms appearing in the sum defining the estimated bias and the local score function. The main difference is in the parameter related to the suspiciousness. In fact in the estimated bias all differences between observed values and corresponding predictions are considered as they were determined by errors, while in the score functions, where the degree of suspiciousness is included, this is not assumed with certainty.

2.4 How to compute the threshold

There are two approaches: 1) through a simulation study, 2) by using a model.

2.4.1 Simulation approach

This approach is based on the availability of raw and edited data comparable with the data on which selective editing has to be applied. The idea is to simulate the selective editing procedure considering the edited data as if they were the 'true' data. Often data from a previous cycle of the same survey are used for this purpose.

The approach can be described by the following steps (De Waal et al., 2011).

- Compute the global scores for the raw data and order (decreasingly) the observations.
- Determine a subset E of units composed of the first p units and replace their raw values with the corresponding edited values.
- Compare the estimates computed using the completely edited data set and the raw data where the subset E is obtained according to step 2.

- Repeat steps 2 and 3 with different values of p until the difference between the two estimates is negligible. Let p^* be the first index such that this condition is fulfilled.
- The threshold t is the value of the GS corresponding to the p^*-th unit.

Remarks:

- The assumption of this approach is that the edited data can be considered as 'true' data. This is a limitation because it can be rarely assumed.
- The simulation approach is frequently applied to data of a previous survey occasion to obtain a threshold value to be used for the current survey. It is worthwhile to note that in this case we assume that the error mechanism and the data distribution are the same in the two occasions.
- The method cannot be applied when you deal with the first wave of a survey.

2.4.2 Model based approaches

In this context, some of the main elements of the problem are modelled through a probability distribution: the true data distribution, the error mechanism, the score functions.

The introduction of a model may be useful to give estimates of the error left in data after the revision of the selected units and thus to ease the determination of a threshold for the selection of units to be reviewed.

A first attempt can be found in Lawrence and McKenzie (2000). By denoting with a the threshold value, they assume that the difference between the observed and the predicted value for the non-selected observations follows a uniform distribution in the interval (–a,a), i.e., U(–a,a). The threshold a is determined so that the bias due to not editing a set of units is low if compared to the sampling error. A conservative solution is $a = \sqrt{\frac{3k}{n}} SE(\hat{Y})$, where $kSE(\hat{Y})$, $k < 1$ is the upper bound for the bias and n is the total number of observations.

The intermittent nature of the error is taken into account in Arbués et al. (2011). The search of a good selective editing strategy is stated as an optimisation problem in which the objective is to minimise the expected workload with the constraint that the expected error of the aggregates computed with the edited data is below a certain constant.

A model based approach is also adopted by Buglielli et al. (2011). They propose to consider (log)- true data y^*_i as realisations from a multivariate Gaussian distribution with mean vector possibly dependent on a set of error-free covariates: $\tilde{y}_i \sim N(\mu_i, \Sigma)$. Errors are supposed to act on a subset of data by inflating the variance, i.e., the covariance matrix of the contaminated data is $\lambda\Sigma$ where λ is a numerical factor greater than one. The intermittent nature of the error is reflected by a Bernoullian random variable with parameter π taking values zero or one depending on whether an error occurs in a unit or not, respectively. This approach naturally leads to a latent class model formulation, where observed data (y) can be viewed as realisation from a mixture of two Gaussian probability distributions associated to contaminated and error-free data:

$$f_Y(y) = (1-\pi)N(y;\mu,\Sigma) + \pi N(y;\mu,(\lambda+1)\Sigma).$$

In this context, the parameter π represents the mixing weight of the mixture and can be interpreted as the *a priori* probability of errors in data. The estimated conditional distribution of true data given observed ones is used to build an appropriate score function. More precisely, for a given variable of interest, a relative (local) score function is defined in terms of difference between the observed value and the expectation of the "true" value conditional on the observed one (the prediction). This approach allows to interpret the score function as the expected error, and to relate the threshold for interacting reviewing to the accuracy of the estimates of interest. A global score can be defined in many ways combining the different local score functions. In Buglielli et al. (2011) the global score is defined as the maximum of the single local scores. This ensures that the accuracy of the estimates is kept under control simultaneously for all the variables of interest.

In practice the steps to perform selective editing within this framework are similar to the ones detailed in the simulation approach, with the difference that the predicted value is obtained by using an explicit model, and that the score directly gives an estimate of the error contaminating each observation.

Remarks:

- The introduction of a model for the error mechanism allows to formalise the problem and hence to have a statistical interpretation of the elements characterising selective editing. Furthermore, using a latent class model implies the advantage that no edited data are required, and the bias of the simulation approach due to considering edited data as true data is avoided.
- The main drawback is that the validity of the conclusions depends on the validity of the model assumptions.

2.5 Dealing with errors remaining in data: a probability sampling approach to selective editing

Ilves and Laitila (2009) and Ilves (2010) propose a two-step procedure for selective editing. Their proposal is motivated by the fact that the non-selected observations may still be affected by errors resulting in a biased target parameter estimator $\hat{T}_Y$. To obtain an unbiased estimator a sub-sample is drawn from the unedited observations (below threshold for global scores), follow-up activities with recontacts are carried through and the bias due to remaining errors is estimated.

The estimated bias is used to make the target parameter estimator $\hat{T}_Y$ unbiased. If our target parameter is the total of the population, the bias-corrected estimator is obtained by subtracting the estimated bias from the HT estimator of the total computed on edited (selected by the selective editing procedure) and unedited (non-selected) observations. Formulas for the variance and a variance estimator are derived by using a two-phase sampling approach. The procedure is discussed in general without specifying a particular selective editing technique, but sampling with probabilities proportional to scores seems to be the obvious choice.

3. Design issues

In the following some important elements concerning the design of a selective editing procedure are reported.

- Selective editing can be applied only to numerical variables. This implies that selective editing is mainly applied to business surveys.
- Selective editing is useful when accurate interactive editing can be performed.
- Selective editing can be applied at the early stages of data collection. This kind of application is named *input editing*. The methods used in this context apply to each incoming record individually, classifying each record as critical or non-critical. The advantage of input editing is that time-consuming task procedures as interactive editing and follow-up are started as soon as possible, with positive effects on response burden and the timeliness of the results. The disadvantage is that the parameters needed for the selection of influential errors should be estimated before data are available. This can be performed only when data from previous survey occasions are available (or strong a priori knowledge is disposable), and the assumptions are that the situation is not changed from the previous surveys to the actual one. On the contrary, the approach consisting in applying selective editing when almost all the data are available is named *output editing*. The disadvantage is clearly related to the timeliness of the results because time consuming task as interactive editing or follow-up are moved to a later stage of the process. The advantage is that all the parameters needed for the selection of influential errors are estimated on the data at hand, so they refer to the actual distribution of data with a potential benefit effect on the precision of selection.
- It is advisable to apply selective editing after the process of detection and correction of systematic errors (see "Statistical Data Editing – Main Module"). Actually, also systematic errors can lead to significant bias but they can often be automatically detected and corrected easily and very reliably. It is highly efficient to correct these errors at an early stage.
- The application of selective editing should be limited to the subset composed of the most important target variables.
- Once one observation is selected, all the variables should possibly be revised, not only the ones considered in the score function.
- Sampling weights are important to estimate the impact of errors on the final estimates. When an input editing approach is chosen, initial sampling weights may be used.

4. Available software tools

- SeleMix is an R-package for selective editing based on contamination models (Di Zio and Guarnera, 2011) freely available on the website http://cran.r-project.org/.
- Selekt is a set of SAS-macros for selective editing, allowing "traditional" hard and soft edits as well as a nonparametric approach based on quantiles to produce measures of suspicion. Selekt works with one and two-stage samples and several sets of domains in output. (Norberg et al., 2010; Norberg, et al., 2011).

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Arbués, I., Revilla, P., and Saldaña, S. (2011), Selective Editing as a Stochastic Optimization Problem. UN/ECE Work Session on Statistical Data Editing, Ljubljana, Slovenia, 9-11 May 2011.

Buglielli, T., Di Zio, M., Guarnera, U., and Pogelli, F. R. (2011), Selective Editing of Business Survey Data Based on Contamination Models: an Experimental Application. NTTS 2011 New Techniques and Technologies for Statistics, Bruxelles, 22-24 February 2011.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011), *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons, New Jersey.

Di Zio, M. and Guarnera, U. (2011), SeleMix: an R Package for Selective Editing via Contamination Models. *Proceedings of the 2011 International Methodology Symposium, Statistics Canada. November 1-4, 2011, Ottawa, Canada.*

EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

Hedlin, D. (2003), Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. *Journal of Official Statistics* **19**, 177–199.

Hedlin, D. (2008), Local and Global Score Functions in Selective Editing. UN/ECE Work Session on Statistical Data Editing, Wien.

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

Ilves M. (2010), Probabilistic Approach to Editing. Workshop on Survey Sampling Theory and Methodology Vilnius, Lithuania, August 23-27, 2010.

Latouche, M. and Berthelot, J. M. (1992), Use of a Score Function To Prioritise and Limit Recontacts in Business Surveys. *Journal of Official Statistics* **8**, 389–400.

Lawrence, D. and McDavitt, C. (1994), Significance Editing in the Australian Survey of Average Weekly Earnings. *Journal of Official Statistics* **10**, 437–447.

Lawrence, D. and McKenzie, R. (2000), The General Application of Significance Editing. *Journal of Official Statistics* **16**, 243–253.

Norberg, A. et al. (2010), *A General Methodology for Selective Data Editing*. Statistics Sweden.

Norberg, A. et al. (2011), *User´s Guide to SELEKT 1.1, A Generic Toolbox for Selective Data Editing*. Statistics Sweden.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Data Editing – Main Module
2. Statistical Data Editing – Automatic Editing
3. Statistical Data Editing – Manual Editing
4. Statistical Data Editing – Macro-Editing
5. Imputation – Main Module

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. Phase 5 - Process

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1. GSBPM Sub-process 5.3: Review, validate and edit

Administrative section

14. Module code

Statistical Data Editing-T-Selective Editing

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	08-02-2012	first version	Di Zio Marco, Guarnera Ugo	Istat (Italy)
0.2	19-03-2012	second version	Di Zio Marco, Guarnera Ugo	Istat (Italy)
0.3	06-04-2012	third version	Di Zio Marco, Guarnera Ugo	Istat (Italy)
0.3.1	04-10-2013	preliminary release		
0.4	15-10-2013	changes according to the EB comments	Di Zio Marco, Guarnera Ugo	Istat (Italy)
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:11