WP.45
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (vii): New and emerging methods

An R package for selective editing based on a latent class model

Invited Paper

Prepared by M.T. Buglielli, M. Di Zio, U. Guarnera and F. R. Pogelli, Istat.

I. Introduction

1. Selective editing is based on the appealing idea of looking for units affected by important errors in order to limit accurate reviewing only to these units. The aim is to reduce the cost of the editing and imputation phase maintaining at the same time a certain level of quality of estimates (Lawrence and McKenzie, 2000; Lawrence and McDavitt, 1994).

2. Observations are prioritised according to the values of a score function that expresses the impact of their potential error on the estimates of interest (Latouche and Berthelot, 1992). All the units above a given threshold are selected since they potentially represent the observations affected by important errors.

3. The score function is generally based on the difference between observed and "anticipated" values; large differences correspond to influential errors (see Jäder and Norberg, 2005; Hedlin, 2003). The problem is that, roughly speaking, these differences are due to both measurement errors and to the natural variability of the phenomenon. The values of the scores so far obtained, cannot be interpreted as a direct evaluation of the accuracy of estimates. Hence, it is not possible to select, according to the score function values, the most influential units such that a prefixed level of accuracy for the target estimates is attained.

4. The use of a latent model, as the contamination models (see Little, 1988, Ghosh-Dastidar and Schafer 2006), allows to distinguish the error and the variability component of the residuals, and the score of an observation can be directly related to the expected error of that unit. In this framework, we can select the units by estimating the expected error left in data according to the prefixed level of quality of target estimates requested by the researcher (see Di Zio et al. 2008).

5. In this paper we present a package developed in R (*R Development Core Team*, 2008) that is mainly devoted to the selection of influential errors according to the contamination model described in Buglielli et al. (2011). The package implements the ECM-algorithm developed to estimate the parameters of the model, computes local and global scores and finally returns the set of observations affected by influential errors with respect to a certain prefixed level of accuracy of the target estimates. In addition, it provides anticipated values (predictions) for each unit for both observed and non observed variables. The latter characteristic makes the package useful also in the context of imputation. The imputation can be considered "robust" in that the model used to compute the "anticipated" values takes into account the presence of errors in data.

6. The package can be downloaded from the www.osor.eu that is the Open Source Observatory and Repository for European public administrations (OSOR), a platform for exchanging information, experiences and FLOSS-based code for use in public administrations.

7 The paper is structured as follows. In Section II, a short description of the model is introduced. Section III describes the functions of the package and takes advantage of the results obtained in experiments carried out on an Istat business surveys to graphically show some important outputs. Conclusions and future works are reported in Section IV.

II. Selective Editing via Contamination Models

8. The present approach is based on explicitly modelling both true (error-free) data and error mechanism. Details on model specification and parameter estimation can be found in Bellisai et al. (2009) and in Buglielli et al. (2010). The model assumptions can be summarized as follows. True data (possibly in log-scale) are thought of as n realizations from a random p-vector $\mathbf{Y}$ that, conditional on a set of q covariates $\mathbf{X}$, is normally distributed with mean vector $\mathbf{BX}$ and covariance matrix $\mathbf{\Sigma}$. The intermittent nature of the error, which is crucial to the present approach, is modelled through a Bernoullian r.v. $\mathbf{I}$, with parameter w, assuming value 1 or 0 depending on whether an error occurs in data or not respectively. The parameter w can be interpreted as the marginal probability of an observation of being affected by an error in at least one variable $\mathbf{Y}$. Conditional on $\mathbf{I}$=1 (presence of error), we assume a Gaussian additive error with zero mean and covariance matrix proportional to $\mathbf{\Sigma}$, the proportionality constant being some positive number λ. Thus the model parameters are $\theta = (\mathbf{B}, \mathbf{\Sigma}, w, \lambda)$.

9. The previous assumptions allow us to explicitly derive, via Bayes formula, the distribution of the true data conditional on the observed data. It is a mixture of a mass density corresponding to absence of error and a Gaussian distribution corresponding to presence of error. This mixture is the central object for the proposed selective editing method and is completely identified by the set of parameters θ. In order to estimate parameters θ we note that they also identify the (unconditional) distribution of the observed data which is another mixture whose components are non degenerate Gaussians. The model parameters can be estimated by maximizing the likelihood function based on the observed data using an EM-type algorithm

10. Once the model parameters have been estimated, they can be plugged into the functional form of the conditional distribution of true data given observed data. The selective editing strategy consists in using this estimated distribution to build up a score function. Specifically, for each unit we compute an "anticipated" value as expected "true value" conditional on the observed value. The anticipated value is obtained by means of a weighted average of the observed value and a synthetic value. The weights are given by the probability of being in error. The synthetic value is in turn the weighted average of the observed value and a robust estimate of the regressed value. The weights are the inverse of the estimated covariance matrices of the true and erroneous data respectively. Hence, a score function can be defined in terms of difference between observed and anticipated value (expected error), and the units to be interactively reviewed can be selected as those having higher score function. Once all the observations have been ordered according to this score function, we are able to estimate the residual error remaining in data after the correction of the first k units (k=1,..,n). The number of most critical units to be edited can be chosen so that the estimate of the residual error is below a prefixed threshold (see Section III). This feature is an important point in the proposed method. In fact, differently from most selective editing procedures, our approach allows to explicitly relate the efforts in editing activities (number of units to be manually checked) to the accuracy of the target estimates.

III. The package *SeleMix*

11. In order to implement the selective editing method based on contamination models, R functions have been developed and included in a package. We assume that the multivariate contaminated variables $\mathbf{Y}$ can be either with or without missing values. In the latter case, a prediction of missing values is automatically provided by the package. The software allows to include in the model also a set of

"cleaned" variables **X** to be used as explanatory variables. This characteristic is particularly useful when auxiliary information (e.g., administrative or historical data) is available.

12. The package is composed of three functions `ml.est`, `pred.y`, `sel.edit`. A further function providing graphical tools is under construction.

The main output of the package is the identification of critical units corresponding to the most influential errors given a prefixed threshold of accuracy of estimate of totals (or means). In the following we describe in detail the functions of the *SeleMix* package.

13. `ml.est`. This function estimates the parameters $\theta = (\mathbf{B}, \mathbf{\Sigma}, w, \lambda)$ by using an ECM-algorithm suitably developed for this modelling. Moreover, it returns the "anticipated" values for the **Y** variables for all the units.

The input of the `ml.est` function is the matrix of observed data and optionally the matrix of covariates **X**.

The user must specify whether true data are assumed to follow either normal or log-normal distribution. In the latter case, zeros are replaced by a small value (10E-8) and a warning is returned. By default the ECM-algorithm starts the iterations with $\lambda = 3$ and w=0.05, but the user can define different starting points.

The starting values of the regression coefficients **B** and the covariance matrix $\mathbf{\Sigma}$ have been computed on input data **Y** and **X** via OLS (i.e., as the data were error-free).

The ECM-algorithm stops either when convergence is achieved or when the user specified maximum number of iterations is reached.

Once the parameters are estimated, the function computes for each unit the posterior probability τ that it belongs to the mixture component corresponding to contaminated data. This probability is used to define a flag of outlyiness that is 1 if τ is greater than a specified threshold (by default equal to 0.5) and 0 otherwise.

The function returns the BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) scores in order to evaluate the goodness of fit of the mixture model versus the standard normal model. This information helps the user to assess the validity of the use of a mixture model for the data at hand.

The output of the `ml.est` function is provided as a list whose components are: the model parameters θ, the anticipated values, the BIC and the AIC scores, the outlier flags, and the posterior probabilities τ.

The function `ml.est` includes a call to another function `pred.y` that calculates the predictions for the variables **Y**.

14. `pred.y`. This function makes a prediction of the true values for the variables **Y** through their expected value conditional on all the available information, i.e., the observed **Y** variables and the **X** variables. It requires as input the parameters $\theta = (\mathbf{B}, \mathbf{\Sigma}, w, \lambda)$ and a set of observed data. Note that missing values are not allowed for the **X** variables.

It returns, for each unit, a "prediction" for both observed and missing items of each **Y** variable, the outlier flag and the posterior probability τ.

15. `sel.edit`. This function prioritises observations according to the score function values and flags the units to be edited so that the expected residual error is below a prefixed level of accuracy.
It is worth noting that `sel.edit` can be used independently of the other *SeleMix* functions. In fact, the identification of influential units can be performed regardless of the particular model used for the prediction.

As input the function receives: the matrix of observed data and the matrix of corresponding anticipated values, the reference estimate of the total of each **Y** variable, the sampling weights and the prefixed level of accuracy.

By default the reference total of **Y** is the weighted sum of the anticipated values, however the user can provide different reference totals. The sampling weights are assumed to be equal 1 if not differently specified, and the default threshold for the level of accuracy is 0.01.

Units affected by influential errors are selected according to the values of a global score computed as follows. First, a local score for a given variable is defined as the (possibly weighted) absolute difference between observed and anticipated values standardised with respect to the reference total estimate. Then, the global score is obtained by computing the maximum of the local scores, and the observations are ranked according to the descending values of the global score.
Finally, the function selects the first *k* units such that, for all the variables, the (expected) total residual error remaining in the other (*n*-*k*) units, is below the prefixed threshold.

The output of `sel.edit` is a matrix containing the flag of influential units, the observed and anticipated values ordered by the global score, the local scores.

16. The following two figures plot outliers versus influential errors, and estimated versus true residual error with respect to an experiment carried out on an Istat business survey. Details of the experiment are provided in Buglielli et al., (2010).
In Figure 1, the observations depicted with grey triangles are those classified as units affected by influential errors, while the black dots are outliers. In this example, the selection is made with respect to a threshold equal to 0.005. We notice that, all the influential errors are outliers, on the contrary some outliers are not influential errors. This is an important peculiarity of selective editing that allows to save resources for data revision. In fact, even if observations are classified as contaminated by errors, their impact on the target estimates is considered negligible with respect to the chosen level of accuracy.

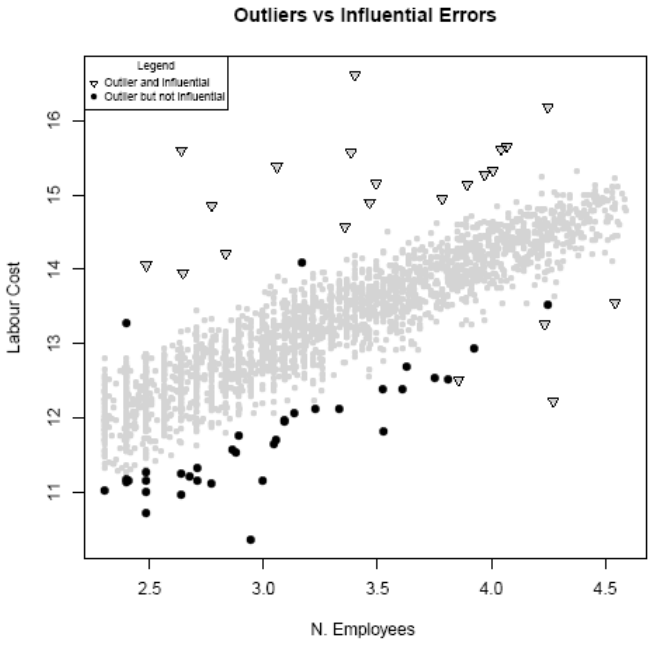

Figure 1: Outliers and Influential errors when the threshold is 0.005.

In Figure 2, dashed line shows the true residual error, while dotted line depicts the estimated residual error on the subset of the first 40 observations. All the units on the left of the vertical lines are the influential observations. We note that both true and estimated cumulative residual error curves are below the prefixed threshold also for some units before the last observation considered as influential. This is due to the fact that the cumulative error is computed on the difference between observed and anticipated

values, and the values can compensate each other. We remind that the stopping criterion ensures that the residual error is below a certain level of accuracy from the last selected unit. Details on the stopping criterion can be found in Di Zio et al., (2008).

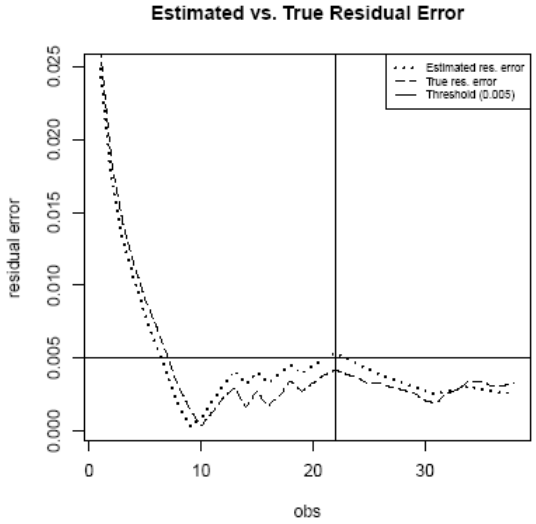

Figure 2: Estimated (dotted line) and true (dashed line) residual error with a threshold equal to 0.005.

III. Conclusions and future works

17. This paper presents a package developed in R to perform selective editing. The method, based on a latent class model, takes advantage of a probabilistic specification of the true data and of the error mechanism. The introduction of a latent class model naturally leads to the framework also discussed in Norberg et al., (2010) where the score is suggested to be composed of a part concerning the suspiciousness of the data and a part related to the impact of error. The method estimates the probability of being in error and the error impact, that suitably combined determine the conditional expected error.
The advantage with respect to traditional approaches is that the two components of the score function are explicitly identified and the values of the scores can be directly interpreted as the expected values of the errors.
In fact, in the traditional approaches, the score is computed by the difference between predicted and observed values without taking into account the probability of being in error, i.e., without considering the intermittent nature of the error. As a consequence, it is not possible to relate the score function values to the level of accuracy of the estimates.

18. The use of a contamination model leads also to a more automatic way of doing selective editing, since once the model assumptions are accepted the user must specify only the level of accuracy and afterwards the selection is done by means of the SeleMix package. There are of course advantages and limitations in the use of a statistical model. A limitation is that it is generally difficult to incorporate fatal edits in the model, and for these edits some ad-hoc and less formal solutions should be used, even if for the sake of truth this is the general approach adopted in practice. On the other hand, the methods generally used to do selective editing have limitations when they have to deal with the so-called soft edits (sometimes named query edits, statistical errors), i.e., all the cases when the values are anomalous but plausible. These edits are indeed implicitly considered in this approach since the units are classified as erroneous with a certain probability, and this probability is explicitly considered in the computation of the score that is in fact the expected error.

Future works concern the development of some graphical tools to be included in the package in order to facilitate the user in the analysis and in the evaluation of the application of the SeleMix procedure to the data at hand. Afterwards, the package will be made available also to the R community.

References

Bellisai D., Di Zio M., Guarnera U., Luzi O. (2009). A Selective Editing approach based on contamination models: an application to an Istat business survey, *UNECE Work Session on Statistical Data Editing.* Neuchatel, 5-7 Ottobre 2009.

Buglielli M.T., Di Zio M., Guarnera U., (2010). Use of Contamination Models for Selective Editing, *Q2010, European Conference on Quality in Survey Statistics,* 4-6 *May 2010, Helsinki*.

Buglielli T., Di Zio M., Guarnera U., (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

Di Zio M., Guarnera U., Luzi O. (2008). Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data. *UN/ECE Work Session on Statistical Data Editing, Vienna*.

Ghosh-Dastidar B., Schafer J.L. (2006). Outlier Detection and Editing Procedures for Continuous Multivariate Data. *Journal of Official Statistics*, Vol. 22, No. 3, 2006, pp. 487-506.

Hedlin D. (2003). Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics, *Journal of Official Statistics,* Vol. 19, No. 2, 177-199.

Jäder A., Norberg A. (2005). A Selective Editing Method considering both suspicion and potential impact, developed and applied to the Swedish Foreign Trade Statistics, *UN/ECE Work Session on Statistical Data Editing*, Ottawa.

Latouche M., Berthelot J.M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys, *Journal of Official Statistics*, 8, n.3, 389- 400.

Little, J.A. (1988). Robust estimation of the mean and covariance matrix from data with missing values, *J. R. Stat. Soc.*, Ser. C, Vol. 37, No. 1, 23-38.

Lawrence D., McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings, *Journal of Official Statistics*, Vol. 10, No. 4, pp. 437-447.

Lawrence D., McKenzie R. (2000). The General Application of Significance Editing. *Journal of Official Statistics*, 16, n. 3, 243-253.

Norberg A., Adolfsson C., Arvidson G., Gidlund P and Nordberg L. (2010) A General Methodology for Selective Data Editing. Technical report, SCB.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.