SELECTIVE EDITING USING CONTAMINATION MODEL

I. Burakauskaitė[1] and V. Nekrašaitė-Liegė[2]

[1] Statistics Lithuania
e-mail: ieva.burakauskaite@stat.gov.lt

[2] Vilnius Gediminas Technical University, Lithuania and Statistics Lithuania
e-mail: vilma.nekrasaite-liege@vilniustech.lt

Abstract

Selective editing was applied to the data editing process of the quarterly statistical survey on service enterprises (turnover indicator) of Statistics Lithuania. Predictions of the target variable were obtained using the contamination model. An impact of a potential error on a sample estimate was evaluated using a score function with a standard structure – a difference between the observed value of the target variable and its prediction multiplied by a sample weight and a suspicion component. A discrete and a continuous suspicion components were used and an impact of the suspicion component on the effectiveness of selective editing was investigated.

Keywords: contamination model; selective editing; data validation; statistical survey; official statistics.

Introduction

An appropriate accuracy of sample estimates is one of the most important results to be achieved using sampling methods in official statistics. Accuracy of sample estimates depends not only on sampling strategy (a sampling plan and an estimator) but on the quality of statistical data as well. Commonly, an unknown part of statistical data contains errors. According to various studies, in order to achieve a desired accuracy of a sample estimate, it is unnecessary to edit all of the detected errors. The main idea of selective editing is to identify and sort errors according to the influence they have on the sample estimate (Lawrence and McDavitt 1994; Lawrence and McKenzie 2000). It is also worth noting that error detection is usually carried out before the calculation of sample estimates. Therefore, it is important to identify only the part of erroneous data that must be edited. Selective editing remains an important, uncommon topic for research in Lithuania.

The first part of the paper introduces the contamination model and the selective editing method that form the base for the practical study of the outlier detection. The second part of the paper shortly presents a study that was carried out using statistical data. During the study some randomly selected values of statistical data were replaced with errors. The detection of randomly introduced errors were then carried out using a few versions of selective editing. The comparison of results as well as its summary are presented in the Conclusions. Calculations were carried out with the statistical programming language R and its package `SeleMix` that has been designed to execute the selective editing method (RDocumentation 2020).

1 Methodology on Selective Editing

1.1 Contamination Model

Suppose that true (unobserved) data are independent realizations of p-variate random vectors $\mathbf{Y}_i^* = (\mathbf{Y}_{i1}^*, \ldots, \mathbf{Y}_{ip}^*)'$, $i = 1, \ldots, n$, with a Gaussian distribution with mean vectors $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ip})'$ and common covariance matrix $\boldsymbol{\Sigma}$. Also, a set of q covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iq})'$ exists for every sampled unit i and $\boldsymbol{\mu}_i = \mathbf{B}'\boldsymbol{x}_i$ where $\mathbf{B}$ is a $q \times p$ matrix of unknown coefficients (Di Zio and Guarnera 2013). The corresponding true data model can be expressed as

$$\mathbf{Y}^* = \mathbf{XB} + \mathbf{U} \tag{1}$$

where $\mathbf{Y}^*$ is the $n \times p$ true data matrix, $\mathbf{X}$ – $n \times q$ covariate matrix and $\mathbf{U}$ – $n \times p$ matrix of normal residuals. Rows of the $\mathbf{U}$ matrix are independent realizations of Gaussian random vectors with mean equal to $\mathbf{0}$ and a covariance matrix $\boldsymbol{\Sigma}$.
The generic marginal probability distributions of the ith sampled unit of matrices $\mathbf{Y}^*$ (true data) and $\mathbf{U}$ (residuals) are denoted as

$$f(\boldsymbol{y}_i^*) = N(\boldsymbol{y}_i^*; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \quad f(\boldsymbol{u}_i) = N(\boldsymbol{u}_i; \mathbf{0}, \boldsymbol{\Sigma}), \quad i = 1, \ldots, n. \tag{2}$$

In general form $N(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a marginal probability distribution of the p-variate random vector $\mathbf{Y}$ with mean equal to $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$.
It is assumed that the presence of errors in data is described by independent Bernoulli random variables. Therefore the observed (erroneous) data can be expressed as

$$\mathbf{Y} = \mathbf{Y}^* + \mathbf{I}\boldsymbol{\epsilon} \tag{3}$$

where $\mathbf{I}$ is a diagonal $n \times n$ matrix with its diagonal elements equal to Bernoullian variables $I_1, \ldots, I_n$ ($I_i = 1$ if the corresponding sampled unit is erroneous and $I_i = 0$ otherwise, $i = 1, \ldots, n$). A marginal probability distribution of the p-variate random vector $\boldsymbol{\epsilon}_i$ (random noise) can be expressed as

$$f(\boldsymbol{\epsilon}_i) = N(\boldsymbol{\epsilon}_i; \mathbf{0}, \boldsymbol{\Sigma}_\epsilon), \quad \boldsymbol{\Sigma}_\epsilon = (\alpha - 1)\boldsymbol{\Sigma}, \tag{4}$$

with a numeric constant $\alpha > 1$.
$f(\boldsymbol{y}|\boldsymbol{y}^*)$ denotes a conditional marginal probability distribution of random variables $\mathbf{Y}$ and $\mathbf{Y}^*$. Therefore, model (3) can be expressed equivalently:

$$f(\boldsymbol{y}|\boldsymbol{y}^*) = (1 - \pi)\delta(\boldsymbol{y} - \boldsymbol{y}^*) + \pi N(\boldsymbol{y}; \boldsymbol{y}^*, \boldsymbol{\Sigma}_\epsilon) \tag{5}$$

where π is "a priori" probability of contamination and $\delta(\boldsymbol{y} - \boldsymbol{y}^*)$ is the delta function with mass at $\boldsymbol{y}^*$.
Furthermore, a marginal probability distribution of the observed data can be expressed as

$$f(\boldsymbol{y}_i) = (1 - \pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \alpha\boldsymbol{\Sigma}). \tag{6}$$

Coefficients of the later observed data model can be obtained by the maximum likelihood estimation.

1.2 Selective Editing

Selective editing is based on the comparison between the observed data and predictions of the true (unobserved) data. The later can be obtained from a conditional marginal probability distribution $f(\boldsymbol{y}_i^*|\boldsymbol{y}_i)$ (Di Zio and Guarnera 2013). An application of the Bayes formula provides:

$$f(\boldsymbol{y}_i^*|\boldsymbol{y}_i) = \tau_1(\boldsymbol{y}_i)\delta(\boldsymbol{y}_i^* - \boldsymbol{y}_i) + \tau_2(\boldsymbol{y}_i)N(\boldsymbol{y}_i^*; \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}) \tag{7}$$

where $\tilde{\boldsymbol{\mu}}_i = \frac{\boldsymbol{y}_i+(\alpha-1)\boldsymbol{\mu}_i}{\alpha}$ and $\tilde{\boldsymbol{\Sigma}} = \left(1-\frac{1}{\alpha}\right)\boldsymbol{\Sigma}$, $\delta(\boldsymbol{y}_i^* - \boldsymbol{y}_i)$ is the delta function with mass at $\boldsymbol{y}_i$, $\tau_1(\boldsymbol{y}_i)$ and $\tau_2(\boldsymbol{y}_i)$ are posterior probabilities that the ith sampled unit with observed values $\boldsymbol{y}_i$, $i = 1, \ldots, n$, is not erroneous and that it is contaminated respectively:

$$\begin{aligned} \tau_1(\boldsymbol{y}_i) &= P(\boldsymbol{y}_i = \boldsymbol{y}_i^* | \boldsymbol{y}_i) = \frac{(1-\pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma})}{(1-\pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \alpha\boldsymbol{\Sigma})}, \\ \tau_2(\boldsymbol{y}_i) &= P(\boldsymbol{y}_i \neq \boldsymbol{y}_i^* | \boldsymbol{y}_i) = 1 - \tau_1(\boldsymbol{y}_i). \end{aligned} \tag{8}$$

Posterior probabilities (8) are defined in terms of the conditional expected value $\tilde{\boldsymbol{y}}_i = E(\boldsymbol{y}_i^* | \boldsymbol{y}_i)$, $i = 1, \ldots, n$. Therefore, the expected error can be defined as

$$\boldsymbol{y}_i - \tilde{\boldsymbol{y}}_i = \tau_2(\boldsymbol{y}_i)(\boldsymbol{y}_i - \tilde{\boldsymbol{\mu}}_i). \tag{9}$$

In practice, formula (9) is usually applied by using maximum likelihood estimates instead of the corresponding true data values.

1.2.1 Definition of Score Function

Hereinafter $\hat{p}$ denotes a maximum likelihood estimate of some parameter p.
Suppose one seeks to estimate a sum of the variable Y_j^*, $j = 1, \ldots, p$, with a sampling weight w_i of the ith sampled unit, – $T_j^* = \sum_{i=1}^n w_i y_{ij}^*$. A ratio between the expected error (9) with a sampling weight w_i multiplied by a suspicion component s_{ij} (probability that the ith sampled unit is erroneous) and target parameter estimate $\hat{T}_j = \sum_{i=1}^n w_i \hat{y}_{ij}$ denotes the conditional error of the ith sampled unit:

$$r_{ij} = \frac{s_{ij} w_i (y_{ij} - \hat{y}_{ij})}{\hat{T}_j}. \tag{10}$$

The local score function for the variable Y_j is denoted as $S_{ij} = |r_{ij}|$. Separate local scores can be combined into one global score GS_i in a few different ways: $GS_i = \max_j S_{ij}$ or $GS_i = \sum_j S_{ij}$. In order to identify an optimal number of observations to be edited, the corresponding sampled units are sorted descendingly according to the GS_i. First $\tilde{k}$ observations are then chosen for the editing procedure:

$$\tilde{k} = \min\{k^* \in 1, \ldots, n \mid \max_j R_{kj} < \eta, \ \forall\ k > k^*\} \tag{11}$$

where $R_{ij} = |\sum_{k \geq i}^n r_{kj}|$ with an accuracy level η.
The suspicion component s_{ij} can take on a discrete form ($s_{ij} \in \{0, 1\}$) and a continuous form ($s_{ij} \in [0, 1]$). In the paper the later continuous suspicion component is defined according to Norberg et al. (2010). An additional test variable should be defined prior to defining the suspicion component:

Definition 1 (Test variable) *Test variable* can be a combination of variables from a statistical survey and (or) additional information. Statistical errors can then be identified by checking whether a value of the test variable $\mathbf{t}_{j'}$, $j' = 1, \ldots, p'$, for the ith sampled unit falls into some chosen acceptance region $\left(\hat{t}_{ij'}^{(L)}, \hat{t}_{ij'}^{(U)}\right)$.

Definition 2 (Discrete suspicion component) *Discrete suspicion component* equals to 1 when a value of the jth survey variable of the ith sampled unit y_{ij} is a non-statistical error or a value of the jth test variable of the ith sampled unit $t_{ij'}$ is a statistical error ($t_{ij'} \notin \left(\hat{t}_{ij'}^{(L)}, \hat{t}_{ij'}^{(U)}\right)$). The later case gives $s_{ij} = 1$ for every survey variable y_{ij} that is a part of the combination $t_{ij'}$. Otherwise $s_{ij} = 0$.

Nonetheless, it is important to take into consideration different distances between observations that do not fall into the chosen acceptance region $\left(\hat{t}_{ij'}^{(L)}, \hat{t}_{ij'}^{(U)}\right)$ and the corresponding bound of the region. A continuous suspicion component should convey the information on the later distance more effectively.

Definition 3 (Continuous suspicion component) Hereinafter $\hat{t}_{ij'}$ denotes a prediction of the test variable $t_{ij'}$.

1) $s_{ij} = 1$ if a value of the jth survey variable of the ith sampled unit y_{ij} is a non-statistical error;

2) $\tilde{s}_{ij'} = \frac{\hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'} - \hat{t}_{ij'}^{(L)}\right) - t_{ij'}}{\max\left\{\left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}^{(L)}\right),\ \alpha \cdot \hat{t}_{ij'}\right\}}$ if $t_{ij'} < \hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'} - \hat{t}_{ij'}^{(L)}\right)$;

3) $\tilde{s}_{ij'} = \frac{t_{ij'} - \hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}\right)}{\max\left\{\left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}^{(L)}\right),\ \alpha \cdot \hat{t}_{ij'}\right\}}$ if $t_{ij'} > \hat{t}_{ij'} + \kappa \cdot \left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}\right)$;

4) $\tilde{s}_{ij'} = 0$ if $\hat{t}_{ij'} - \kappa \cdot \left(\hat{t}_{ij'} - \hat{t}_{ij'}^{(L)}\right) < t_{ij'} < \hat{t}_{ij'} + \kappa \cdot \left(\hat{t}_{ij'}^{(U)} - \hat{t}_{ij'}\right)$.

Continuous suspicion component then equals to $s_{ij'} = \frac{\tilde{s}_{ij'}}{\tau + \tilde{s}_{ij'}}$ with parameters $\kappa \geq 0$, $\alpha > 0$ and $\tau > 0$. $s_{ij} = \max_{j'} s_{ij'}$ for every survey variable y_{ij} that is a part of the combination $t_{ij'}$.

2 Selective Editing Application on Statistical Survey Data

The outlier detection study was carried out using statistical data from the quarterly statistical survey on service enterprises of Statistics Lithuania. Enterprise turnover[1] of the accounting period was the target variable of the study. Predictor variables and the corresponding number of observations in data sets are given in Table 1.

Table 1: Number of Observations in Statistical Data Sets

Predictor variable	Number of observations (n)
Turnover from VAT declarations	4085
Turnover from the quarterly F-01 questionnaire	574
Average number of employees	4867
Total hours worked	4931

Before applying selective editing on statistical data it was important to ensure that all items for the target and predictor variables are not missing and greater than 0. Therefore, a number of observations in data sets (primary populations) varies according to the chosen predictor variable. In order to control the data contamination process, detected outliers in primary populations were replaced with contamination model predictions. The following procedure was then applied to every primary population:

1. Data were contaminated in 3 different ways:
   (a) 1,5 percent of observations were multiplied by 100,
   (b) 2 percent of observations were trimmed leaving only the first and the last digits,

[1] *Enterprise turnover* – enterprise income gained during the accounting period for sold goods and granted services. It does not include value-added tax (hereinafter referred to as VAT), income for long-term material assets, income for financial and investment activities, dividends, etc. (Official Statistics Portal, 2015).

(c) 20000000 was added to $1,5$ percent of observations;

2. Estimation of model coefficients and outlier (potential error) detection were carried out using the statistical programming language R and its package `SeleMix` (function `ml.est`);

3. Values of the target variable were sorted descendingly according to estimates of the global score function. An estimate of the global score function is close to 0 when a value of the target variable is not identified as an outlier and therefore has no major impact on the accuracy of the sample estimate, and greater than 0 when a value of the target variable is identified as an outlier;

4. The part of outliers that have a major impact on the accuracy of the sample estimate (influential errors) were chosen for the editing procedure.

The later influential error detection procedure was repeated in two different ways – by calculating estimates of the score function (1) with a discrete suspicion component that is the same among all observations ($s_i = 1$), and (2) with a continuous suspicion component. The later suspicion component was designed using an acceptance region between the first and the third quartiles $\left(\hat{t}^{(L)}, \hat{t}^{(U)}\right)$ where $\hat{t}_i = \hat{y}_i$ $(i = 1, \ldots, n)$, parameters κ and τ varies, $\alpha = 0,05$.
Selective editing with different accuracy levels gives a different number of influential errors. If all of the detected influential errors were introduced by the data contamination procedure, the corresponding accuracy level was chosen for the following study (see Table 2).

Table 2: Levels of Accuracy (Threshold Values) for Statistical Data Sets

Predictor variable	Level of accuracy
Turnover from VAT declarations	0.011
Turnover from the quarterly F-01 questionnaire	0.004
Average number of employees	0.027
Total hours worked	0.026

The results of selective editing were then compared by estimating the relative absolute bias after every edit of an influential error. This way a number of influential errors to be edited in order to achieve the desired accuracy of sample estimates was determined (see Table 3).

Table 3: Number of Influential Errors in Statistical Data Sets

Predictor variable	Total number of influential errors	Number of influential errors to be edited
(1) Selective editing with a discrete suspicion component		
Turnover from VAT declarations	134	92
Turnover from the quarterly F-01 questionnaire	23	14
Average number of employees	90	> 90
Total hours worked	111	> 111
(2) Selective editing with a continuous suspicion component		
Turnover from VAT declarations	93	92
Turnover from the quarterly F-01 questionnaire	15	14
Average number of employees	136	121
Total hours worked	124	123

It is important to note that selective editing with predictor variables such as average number

of employees and total hours worked gives a lower number of influential errors with a dicrete suspicion component compared to the case when a continuous suspicion component is used. Nonetheless, the chosen accuracy level is not achieved even after editing all of the identified influential errors. The main reason is a weak dependency between the target variable of the study and the corresponding predictor variables (correlation coefficient estimates are lower than 0.6). The later aspect causes greater differences between true values of the target variable and its contamination model predictions. Applications of selective editing with different predictor variables have shown an effectiveness of a continuous suspicion component on the outlier detection procedure as this approach to selective editing lets to identify a lower number and more important influential errors.

Conclusions

After calculations of the relative absolute bias dependency on the number of edited influential errors, selective editing with a continuous suspicion component was determined to be an optimal method of the outlier detection procedure. The later version of selective editing prevents from the unnecessary statistical data editing.
Turnover from VAT declarations and turnover from the quarterly F-01 questionnaire were identified as the most suitable predictor variables for the outlier detection procedure. The main property of a suitable predictor variable turned out to be a high correlation between the later predictor variable and the target variable of the study.

References

Di Zio, M., Guarnera, U. (2013) A Contamination Model for Selective Editing. *Journal of Official Statistics*, **29**(4), 539-555.

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

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

Norberg, A., Adolfsson, C., Arvidson, G., Gidlund, P., Nordberg, L. (2010) *A General Methodology for Selective Data Editing*. Statistics Sweden, Stockholm.

Official Statistics Portal (2015). Paslaugų įmonių veiklos statistinio tyrimo metodika. Retrieved from `https://osp.stat.gov.lt/documents/10180/687662/Metodika_2012DI121.pdf`.

RDocumentation (2020). Functions in SeleMix (1.0.2). Retrieved from `https://rdocumentation.org/packages/SeleMix/versions/1.0.2`.