Journal of Official Statistics, Vol. 29, No. 4, 2013, pp. 539–555, http://dx.doi.org/10.2478/jos-2013-0039

A Contamination Model for Selective Editing

Marco Di Zio[1] *and Ugo Guarnera*[1]

The aim of selective editing is to identify observations affected by influential errors. A score function based on the impact of the potential error on target estimates is useful to prioritize observations for accurate reviewing. We assume a Gaussian model for true data and an "intermittent" error mechanism such that a proportion of data is contaminated by an additive Gaussian error. In this setting, scores can be related to the expected value of errors affecting data. Consequently, a set of units can be selected such that the expected residual error in data is below a prefixed threshold. In the context of economic surveys when positive variables are analyzed, the method is more realistically applied to logarithms of data instead of data in their original scale. The method is illustrated through an experimental study on real business survey data where contamination is simulated according to error mechanisms frequently encountered in the practical context of economic surveys.

Key words: Statistical data editing; influential errors; finite mixture models; score function.

1. Introduction

Selective editing is based on the idea of looking for observations containing important errors in order to focus the treatment only on them thus reducing the cost of the editing and imputation phase (E&I), while maintaining a desired level of quality of estimates (Granquist 1997; Lawrence and McKenzie 2000; Lawrence and McDavitt 1994). The underlying assumption is that the true values for the selected units can be obtained through follow-up or interactive treatment. In practice, observations are prioritized according to the importance of errors expressed by the values of a score function (Latouche and Berthelot 1992; Hedlin 2003), and units having a value of the score function above a given threshold, are selected for a careful treatment.

The most commonly used methods to determine the scores are based on the difference between observed and predicted values. This difference is composed of the possible measurement error and the prediction error. When only raw data are available, traditional methods do not allow the estimation of these two elements separately, hence scores are not directly related to the expected errors. The consequence is that the value of the selective editing threshold will not be directly interpretable as a level of accuracy of estimates of interest and it will be difficult to find a stopping rule related to the expected level of quality of estimates.

The introduction of a contamination model naturally leads to building a score function as defined in Jäder and Norberg (2005). It is defined in terms of a risk component

[1] ISTAT, Italian National Institute of Statistics, Via Cesare Balbo 16, 00184 Rome, Italy. Emails: dizio@istat.it and guarnera@istat.it

(the probability of being in error) and an influence component (the magnitude of error), and allows the estimation of the expected error associated with each unit. In particular, the contamination normal model is characterized by peculiarities that make it useful for the problems generally treated by selective editing. In fact, it is usually applied to deal with gross errors (see Ghosh-Dastidar and Schafer 2006) and is based on a latent variable addressing the status of error for each observation. The latent variable describes the intermittent nature of the errors generally affecting surveys carried out by National Statical Institutes (NSIs) where in fact only a part of the observations are affected by errors. In order to make the model useful in practice, it is extended to deal with lognormal variables, to manage the presence of auxiliary variables not affected by errors (for instance in the case of administrative variables), and to cope with missing values. As far as incomplete observations are concerned, usual methods may lack of possibility of computing a set of consistent and comparable scores. In our setting, the score is coherently computed by taking into account the relevant marginal distribution obtained from the estimated multivariate distribution. In the proposed approach, the scores can be interpreted as expected errors, and a threshold can be determined such that the expected error of the target estimates due to residual errors left in data is below a predefined value. An algorithm to select the units to be edited is also proposed. Although the contamination model, the score function and the selection algorithm are presented as parts of a unique procedure, they can be used separately in different selective editing strategies.

Some experiments showed that the procedure can be usefully applied even when there are some departures from the model assumptions (Buglielli et al. 2011). It is currently used in some Istat surveys such as the *Building permits survey*, the *Structure of earning surveys*, and the *Information and communication technology survey*.

The selective editing procedure is modularly implemented in an R package named SeleMix (Buglielli and Guarnera 2011) that is available on the R website (http://cran.r-project.org/).

The article is structured as follows. The contamination model is described in Section 2 where, in particular, it is explained how to obtain predictions for each single observation (Subsection 2.1). The algorithm to estimate the model parameters is illustrated in Section 3. The use of the model in presence of missing data is presented in Section 4. Section 5 describes the application of the contamination model in the selective editing setting, and in particular a proposal for a score function and for a selection criterion is given. In Section 6 we present an experimental application to illustrate the approach and to empirically evaluate its properties. Concluding remarks are given in Section 7.

2. True Data Model and Error Mechanism

An important feature of the proposed model is that it explicitly takes into account the fact that only a proportion of survey data are affected by errors, that is, the error mechanism has an intermittent nature. Data may be partitioned in two groups: error-free data and contaminated data, the membership of each unit being unknown. This naturally leads to modelling the observed data through a latent class model, where the latent variable is a binary variable to be interpreted as an error indicator variable. When the interest is focused on the identification of gross errors, one possible approach consists in specifying a distribution for the observed data as a mixture of two probability distributions corresponding to error-free and contaminated data respectively. This is the approach followed, for instance, by Ghosh-Dastidar and Schafer (2006), that uses the membership posterior probabilities to asses the degree of outlyingness of each observation. In the context of selective editing however, one is mostly interested in identifying errors having high impact on some estimate of interest, rather than in identifying implausible observations. Thus there is the need to estimate the error magnitude. This can be done if the distribution of the "true" unobserved data and the error mechanism are specified separately. In particular, the error mechanism is specified via the conditional distribution of observed data given true data. In the following, the true data model and the error mechanism are described in detail.

We suppose that true unobserved data are independent realizations of p-variate random vectors $\boldsymbol{Y}_i^* = (\boldsymbol{Y}_{i1}^*, \ldots \boldsymbol{Y}_{ip}^*)'$, $i = 1, \ldots, n$, whose distributions are Gaussian with mean vectors $\boldsymbol{\mu}_i$ and common covariance matrix $\boldsymbol{\Sigma}$. Furthermore, it is assumed that on each sampled unit i a (possibly empty) set of q covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iq})'$ is also available and that $\boldsymbol{\mu}_i = \boldsymbol{B}'\boldsymbol{x}_i$, where $\boldsymbol{B}$ is a $q \times p$ matrix of unknown coefficients. The previous hypotheses can be expressed in matrix form as

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

where $\boldsymbol{Y}^*$ is the $n \times p$ true data matrix, $\boldsymbol{X}$ is the $n \times q$ covariate matrix, and $\boldsymbol{U}$ is the $n \times p$ matrix of normal residuals whose rows are independent realizations of Gaussian random vectors with zero mean and covariance matrix $\boldsymbol{\Sigma}$.

Hereafter, the notation $f(\boldsymbol{v})$ will denote the generic marginal probability distribution (or density) for the random variable $\boldsymbol{V}$. Analogously, $f(\boldsymbol{v}, \boldsymbol{w})$ and $f(\boldsymbol{v}|\boldsymbol{w})$ will denote joint and conditional distributions involving variables $\boldsymbol{V}$ and $\boldsymbol{W}$. Thus, for instance, for the ith unit, $f(\boldsymbol{y}_i^*)$ and $f(\boldsymbol{u}_i)$ are the marginal distributions of the true value and of the residual respectively. From the previous assumptions:

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

where, as usual, $N(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

We assume that presence of errors in data is governed by n independent Bernoulli random variables $I_i, (i = 1, \ldots, n)$ with parameter π, that is, $I_i = 1$ if an error occurs on unit i and $I_i = 0$ otherwise. Furthermore, given that an error is present on the ith unit (i.e., given the event $\{I_i = 1\}$), its action is described through an additive random noise represented by a p-variate random Gaussian variable $\boldsymbol{\epsilon}_i$ with zero mean and covariance matrix $\boldsymbol{\Sigma}_\epsilon$ proportional to $\boldsymbol{\Sigma}$. If $\boldsymbol{Y}$ denotes the data matrix associated with the observed (possibly contaminated) data and $\boldsymbol{\epsilon}$ the error matrix whose ith row is $\boldsymbol{\epsilon}_i'$, we can formally express the error mechanism as:

$$\boldsymbol{Y} = \boldsymbol{Y}^* + \boldsymbol{I\epsilon}, \; f(\boldsymbol{\epsilon}_i) = N(\boldsymbol{\epsilon}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon), \; \boldsymbol{\Sigma}_\epsilon = (\alpha - 1)\boldsymbol{\Sigma}, \tag{3}$$

where α is a numeric constant greater than 1, and $\boldsymbol{I}$ is a diagonal $n \times n$ matrix whose diagonal elements are the Bernoullian variables $I_1, \ldots, I_n$. Equivalently, we can specify the error model through the conditional distribution:

$$f(\boldsymbol{y}|\boldsymbol{y}^*) = (1-\pi)\delta(\boldsymbol{y}-\boldsymbol{y}^*) + \pi N(\boldsymbol{y};\boldsymbol{y}^*,\boldsymbol{\Sigma}_\epsilon). \tag{4}$$

where π (mixing weight) represents the "a priori" probability of contamination and $\delta(\boldsymbol{t}'-\boldsymbol{t})$ is the delta-function with mass at $\boldsymbol{t}$.

In the previous model, the crucial aspect is the intermittent nature of the error implied by the introduction of the Bernoullian variables I_i. Due to this assumption, it is conceptually possible to think of data as partitioned into the two groups of error-free and contaminated data, and to estimate, for each observation, the posterior probability of group membership, i.e., the probability of being error-free or contaminated. This is the key aspect of the proposed approach to selective editing. In fact, as we will see, differently from most selective editing methods, the "suspiciousness" of each observation is naturally incorporated in the model through the posterior probabilities.

Once the true data distribution and the error mechanism have been specified, the distribution of the observed data can also be easily derived through multiplying the true data density by the error density (4), and integrating over $\boldsymbol{y}^*$. The resulting distribution is:

$$f(\boldsymbol{y}_i) = (1-\pi)N(\boldsymbol{y}_i;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i;\boldsymbol{\mu}_i,\alpha\boldsymbol{\Sigma}). \tag{5}$$

Expression (5) represents a mixture of two regression models having the same coefficient matrix $\boldsymbol{B}$ and proportional residual variance-covariance matrices. This distribution can be estimated by maximizing the likelihood based on n sample units via an ECM algorithm (see Meng and Rubin 1993). Details are provided in Section 3.

2.1. Predictions

The contamination model can be used to obtain predictions or "anticipated values" for true unobserved data. The separate specification of true data model and error model allows, contrarily to the direct specification of the observed data distribution, to derive, for $i = 1,\ldots,n$, the distribution $f(\boldsymbol{y}_i^*|\boldsymbol{y}_i)$ of the true data conditional on the observed data, where we have suppressed the reference to the $\boldsymbol{X}$ variates in the notation. A straightforward application of the Bayes formula provides:

$$f(\boldsymbol{y}_i^*|\boldsymbol{y}_i) = \tau_1(\boldsymbol{y}_i)\delta(\boldsymbol{y}_i^*-\boldsymbol{y}_i) + \tau_2(\boldsymbol{y}_i)N(\boldsymbol{y}_i^*;\tilde{\boldsymbol{\mu}}_i,\tilde{\boldsymbol{\Sigma}}) \tag{6}$$

where

$$\tilde{\boldsymbol{\mu}}_i = \frac{(\boldsymbol{y}_i + (\alpha-1)\boldsymbol{\mu}_i)}{\alpha}; \quad \tilde{\boldsymbol{\Sigma}} = \left(1-\frac{1}{\alpha}\right)\boldsymbol{\Sigma},$$

$\delta(\boldsymbol{y}_i^*-\boldsymbol{y}_i)$ is the delta function with mass at $\boldsymbol{y}_i$, and $\tau_1(\boldsymbol{y}_i)$, $\tau_2(\boldsymbol{y}_i)$ are the posterior probabilities that a unit with observed values $\boldsymbol{y}_i$ belongs to the correct or erroneous data group respectively:

$$\tau_1(\boldsymbol{y}_i) = Pr(\boldsymbol{y}_i = \boldsymbol{y}_i^*|\boldsymbol{y}_i) = \frac{(1-\pi)N(\boldsymbol{y}_i;\boldsymbol{\mu}_i,\boldsymbol{\Sigma})}{(1-\pi)N(\boldsymbol{y}_i;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i;\boldsymbol{\mu}_i,\alpha\boldsymbol{\Sigma})},$$

$$\tau_2(\boldsymbol{y}_i) = Pr(\boldsymbol{y}_i \neq \boldsymbol{y}_i^*|\boldsymbol{y}_i) = 1-\tau_1(\boldsymbol{y}_i),$$

$$i = 1,\ldots,n.$$

In order to make the meaning of Formula (6) clear, let us consider the univariate case in absence of covariates ($E(Y^*) = \mu$). Let σ^2, σ_ϵ^2 denote the variances of true data and errors respectively, and define $\alpha = \left(\sigma^2 + \sigma_\epsilon^2\right)/\sigma^2$. Then it is easily seen that the mean $\tilde{\mu}_y$ of the second component of the mixture (6) is given by $\left(\sigma_\epsilon^{-2} y + \sigma^{-2}\mu\right)/\left(\sigma^{-2} + \sigma_\epsilon^{-2}\right)$. In other words, given that the observed value $\boldsymbol{y}$ is not correct, the expectation of the corresponding true value is a weighted mean of the observed value y and the unconditioned mean μ with weights proportional to the inverse of the variances σ^2 and σ_ϵ^2 respectively. Moreover, the variance $\tilde{\sigma}^2 = \left(\sigma^{-2} + \sigma_\epsilon^{-2}\right)^{-1}$ is lower than both σ^2 and σ_ϵ^2, that is, the knowledge of the error mechanism reduces the uncertainty about y^* and the knowledge of the true data model reduces the uncertainty about the evaluation of the error $y - y^*$ that actually occurred.

It is natural to define predictions in terms of the conditional expected value $\tilde{\boldsymbol{y}}_i = E(\boldsymbol{y}_i^* | \boldsymbol{y}_i)$. From (6) it follows:

$$\tilde{\boldsymbol{y}}_i = \tau_1(\boldsymbol{y}_i)\boldsymbol{y}_i + \tau_2(\boldsymbol{y}_i)\tilde{\boldsymbol{\mu}}_i, \quad i = 1, \ldots, n. \tag{7}$$

Correspondingly, we can define the expected error as

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

The last expression makes it natural to interpret τ_2 and $\boldsymbol{y}_i - \tilde{\boldsymbol{\mu}}_i$ as "risk component" and "influence component" respectively to be considered in the score function definition. In practice, parameters involved in expected errors are unknown, and have to be estimated. The algorithm to obtain maximum likelihood estimates (MLE) of the parameters is described in Section 3, and their use in a score function is illustrated in Section 5.

We remark that in the context of economic surveys, when positive variables are analyzed, logarithms of data instead of data in their original scale are often modeled through a Gaussian distribution. The above methodology can be easily adapted to the lognormal case. In this case the error model assumed for data in original scale is multiplicative; more precisely, contaminated data are related to true data by means of the relation

$$\boldsymbol{Z} = \boldsymbol{Z}^* \boldsymbol{e}^\epsilon$$

where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon)$.

For $i = 1, \ldots, n$, let $\boldsymbol{Y}_i^* = \ln \boldsymbol{Z}_i^*$, $\boldsymbol{Y}_i = \ln \boldsymbol{Z}_i$, where $\boldsymbol{Z}_i^*$ and $\boldsymbol{Z}_i$ represent the variables associated with true and contaminated data respectively, and $\boldsymbol{Y}_i^*, \boldsymbol{Y}_i$ are modeled as previously illustrated (Formulas 2–6). The distribution of $\boldsymbol{Z}_i^*$ given z_i is:

$$f(z_i^* | z_i) = \tau_1(\ln(z_i))\delta(z_i^* - z_i) + \tau_2(\ln(z_i))LN(z_i^*; \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}), \tag{8}$$

where $LN(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the lognormal density with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

3. Estimation

In this section, the algorithm to obtain MLEs of the model parameters is described. The log-likelihood to be maximized is:

$$\sum_{i=1}^{n} \log f_i(\boldsymbol{y}_i),$$

where

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

and $\boldsymbol{\mu}_i = \boldsymbol{B}'\boldsymbol{x}_i$. An ECM algorithm is used and it consists in repeatedly applying, until convergence, the following steps:

E-step

$$\begin{aligned} \tau_1(\boldsymbol{y}_i) &= \frac{(1-\pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma})}{(1-\pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \alpha\boldsymbol{\Sigma})} \\ \tau_2(\boldsymbol{y}_i) &= 1 - \tau_1(\boldsymbol{y}_i) \\ i &= 1, \ldots, n. \end{aligned}$$

CM-step

(M1) *update the mixing weight* (π)

$$\pi = \frac{1}{n}\sum_{i=1}^{n} \tau_2(\boldsymbol{y}_i)$$

(M2) *update regression parameters* (**B**)

$$\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{Y}$$

(M3) *update covariance matrix* ($\boldsymbol{\Sigma}$)

$$\boldsymbol{\Sigma} = \frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})'\boldsymbol{\Omega}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})}{n}$$

(M4) *update variance inflation parameter* (α)

$$\alpha = \frac{trace\left\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})'\boldsymbol{\tau}_2^D(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})\boldsymbol{\Sigma}^{-1}\right\}}{q\pi}$$

where:

$$\boldsymbol{\Omega} \doteq \boldsymbol{\tau}_1^D + \frac{\boldsymbol{\tau}_2^D}{\alpha},$$

and $\boldsymbol{\tau}_j^D$ denotes the diagonal matrix of which the ith diagonal element is $\tau_j(\boldsymbol{y}_i)$, $j = 1, 2$. Note that, in (M1)–(M4), maximization with respect to model parameters is not simultaneous but conditional on the other parameters remaining fixed. This make the convergence of the algorithm, convergence slower than it would be in a genuine EM algorithm. In order to initialize the algorithm we use as starting points for $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ the estimates of the corresponding parameters obtained through ordinary linear squares (OLS) based on all data. A random value for π in the interval [0.6, 1] is chosen, and α is initialized with some reasonable value, for instance $\alpha \in [5, 10]$.

In case of log-normal data, the ECM algorithm has to be applied to logarithms of data.

In the following, the MLEs will be denoted by $\hat{\pi}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\Sigma}}, \hat{\alpha}$. Analogously, $\hat{\tau}_1(\boldsymbol{y}_i)$, $\hat{\tau}_2(\boldsymbol{y}_i)$ and $\hat{\tilde{\boldsymbol{\mu}}}_i$ will denote the estimates of $\tau_1(\boldsymbol{y}_i)$, $\tau_2(\boldsymbol{y}_i)$ and $\tilde{\boldsymbol{\mu}}_i$.

4. Incomplete Data

The previous methodology can be easily extended to situations where observed data are incomplete and the nonresponse mechanism is assumed to be MAR. According to the usual notation for incomplete data, the equality $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i,o}, \boldsymbol{Y}_{i,m})$ means that the random vector $\boldsymbol{Y}_i$ can be partitioned in two subvectors $\boldsymbol{Y}_{i,o}$, $\boldsymbol{Y}_{i,m}$ corresponding to the observed and missing items respectively for the ith unit. The partition induces a similar decomposition for the starred variables: $\boldsymbol{Y}_i^* = (\boldsymbol{Y}_{i,o}^*, \boldsymbol{Y}_{i,m}^*)$. Note that by definition, the $\boldsymbol{Y}^*$ variables are never observed, so that partitioning is determined only by the missing pattern of the contaminated variables. According to the partition of $\boldsymbol{Y}$ and $\boldsymbol{Y}^*$ vectors, we obtain the partition of all relevant vectors and matrices. The matrix $\boldsymbol{\Sigma}$ can be partitioned as:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{oo} & \boldsymbol{\Sigma}_{om} \\ \boldsymbol{\Sigma}_{mo} & \boldsymbol{\Sigma}_{mm} \end{pmatrix}$$

so that, analogously to the complete data case, we can define matrices $\tilde{\boldsymbol{\Sigma}}_{oo}$ and $\tilde{\boldsymbol{\Sigma}}_{mm}$ as $(1 - 1/\alpha)\boldsymbol{\Sigma}_{oo}$ and $(1 - 1/\alpha)\boldsymbol{\Sigma}_{mm}$ respectively.

In the same manner, for each missing pattern, we can partition the matrix $\boldsymbol{B}$ as $\boldsymbol{B} = [\boldsymbol{B}_o, \boldsymbol{B}_m]$, where the columns of matrices $\boldsymbol{B}_o$ and $\boldsymbol{B}_m$ correspond to observed and missing variables respectively. Furthermore, for $i = 1, \ldots, n$, let:

$$\boldsymbol{\mu}_{i,o} = \boldsymbol{B}_o'\boldsymbol{x}_i; \quad \boldsymbol{\mu}_{i,m} = \boldsymbol{B}_m'\boldsymbol{x}_i; \quad \tilde{\boldsymbol{\mu}}_{i,o} = \frac{(\boldsymbol{y}_{i,o} + (\alpha - 1)\boldsymbol{\mu}_{i,o})}{\alpha}; \quad \tilde{\boldsymbol{\mu}}_{i,m} = \frac{(\boldsymbol{y}_{i,m} + (\alpha - 1)\boldsymbol{\mu}_{i,m})}{\alpha}.$$

Our goal is to estimate, for $i = 1, \ldots, n$, the conditional distribution of $\boldsymbol{Y}_i^*$ given $\boldsymbol{Y}_{i,o}$. We have:

$$f(\boldsymbol{y}_i^* | \boldsymbol{y}_{i,o}) = f(\boldsymbol{y}_{i,o}^*, \boldsymbol{y}_{i,m}^* | \boldsymbol{y}_{i,o}) = \frac{f(\boldsymbol{y}_{i,o} | \boldsymbol{y}_{i,o}^*, \boldsymbol{y}_{i,m}^*) f(\boldsymbol{y}_{i,m}^* | \boldsymbol{y}_{i,o}^*) f(\boldsymbol{y}_{i,o}^*)}{f(\boldsymbol{y}_{i,o})}. \tag{9}$$

From the assumed error model in Formula (3), each observed variable, conditionally on the corresponding true variable, is independent of all other true variables, thus we can rewrite (9) as

$$f(\boldsymbol{y}_i^* | \boldsymbol{y}_{i,o}) = \frac{f(\boldsymbol{y}_{i,o} | \boldsymbol{y}_{i,o}^*) f(\boldsymbol{y}_{i,o}^*)}{f(\boldsymbol{y}_{i,o})} f(\boldsymbol{y}_{i,m}^* | \boldsymbol{y}_{i,o}^*). \tag{10}$$

The fraction in (10) is the conditional density of $\boldsymbol{Y}^*_{i,o}$ given $\boldsymbol{y}_{i,o}$ and can be obtained from Formula (6) of Subsection 2.1:

$$\frac{f(\boldsymbol{y}_{i,o}|\boldsymbol{y}^*_{i,o})f(\boldsymbol{y}^*_{i,o})}{f(\boldsymbol{y}_{i,o})} = f(\boldsymbol{y}^*_{i,o}|\boldsymbol{y}_{i,o}) = \tau_1(\boldsymbol{y}_{i,o})\delta(\boldsymbol{y}^*_{i,o} - \boldsymbol{y}_{i,o}) + \tau_2(\boldsymbol{y}_{i,o})N(\boldsymbol{y}^*_{i,o}; \tilde{\boldsymbol{\mu}}_{i,o}, \tilde{\boldsymbol{\Sigma}}_{oo}).$$

Thus, we can write:

$$f(\boldsymbol{y}^*_i|\boldsymbol{y}_{i,o}) = \tau_1(\boldsymbol{y}_{i,o})f_1(\boldsymbol{y}^*_i|\boldsymbol{y}_{i,o}) + \tau_2(\boldsymbol{y}_{i,o})f_2(\boldsymbol{y}^*_i|\boldsymbol{y}_{i,o}),$$

where

$$f_1(\boldsymbol{y}^*_i|\boldsymbol{y}_{i,o}) = \delta(\boldsymbol{y}^*_{i,o} - \boldsymbol{y}_{i,o})f(\boldsymbol{y}^*_{i,m}|\boldsymbol{y}^*_{i,o}) = \delta(\boldsymbol{y}^*_{i,o} - \boldsymbol{y}_{i,o})f(\boldsymbol{y}^*_{i,m}|\boldsymbol{y}_{i,o}) \tag{11}$$

$$f_2(\boldsymbol{y}^*_i|\boldsymbol{y}_{i,o}) = N(\boldsymbol{y}^*_{i,o}; \tilde{\boldsymbol{\mu}}_{i,o}, \tilde{\boldsymbol{\Sigma}}_{oo})f(\boldsymbol{y}^*_{i,m}|\boldsymbol{y}^*_{i,o}). \tag{12}$$

Both conditional densities in (11) and (12) can be obtained from that of a bipartitioned multivariate normal distribution. The density (11) can be directly derived from the true-data distribution (2). The density $f_2(\boldsymbol{y}^*_i|\boldsymbol{y}_{i,o})$ is normal, but the derivation is somewhat more involved. It is thus possible to obtain closed expressions of the expected true values given the observed ones. The adaption of these results to the log-normal distribution is straightforward. All the details are given in the Appendix.

As far as parameter estimation is concerned, we have used the ECM algorithm described in Section 3 on completely observed data. This approach is a suboptimal and could be properly modified in order to take into account also incomplete observations. The adaption of the ECM algorithm is a topic for a future study.

5. Selective Editing and Score Function

The score function is the main tool to prioritize observations according to the impact of errors on target estimates. 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 with predicted values, taking into account the probability of being in error (suspiciousness). The latter element arises from the implicit assumption that only a certain proportion of data is affected by error, or, from a probabilistic perspective, that each measured value has a certain probability of being erroneous. When the degree of suspiciousness is not taken into account a large proportion of false alarms is expected, as noted in several case studies by Norberg et al. (2010).

Prediction and suspiciousness are usually combined to form a score for a single variable, named local score. An example of local score for the unit i with respect to the variable Y_j when the target quantity to be estimated is the total $t^*_j = \sum_{i=1}^N y^*_{ij}$ in a population $\mathcal{P}$ of size N is:

$$S_{ij} = \frac{p_i w_i |y_{ij} - \hat{y}_{ij}|}{t^{ref}_j}$$

where p_i is a degree of suspiciousness, y_{ij} is the observed value of the variable Y_j on the ith unit, $\hat{y}_{ij}$ is the corresponding prediction, w_i is the sampling weight, and t^{ref}_j is a reference estimate of the target parameter t^*_j. A review can be found in De Waal et al. (2011).

When the interest is on more than one variable, the local scores can be combined to form a global score GS_i, examples of global scores are $GS_i = \sum_j S_{ij}$, or $GS_i = \max_j S_{ij}$, see Hedlin (2008).

The global score is used to evaluate the impact on the target estimates of the errors remaining in the unedited observations. To this aim, observations are ordered by their global score and all the units with a score above a threshold value are selected. The threshold should be chosen so that the impact on the target estimates of the errors remaining in the unedited observations is negligible.

The evaluation of the impact of errors remaining in data and so of the threshold is generally done through a simulation study based on raw and edited data from a previous occasion of the same survey (De Waal et al. 2011). This approach is based on the assumption that the edited data can be considered as true data and that the error mechanism and the data distribution are the same in the two survey occasions. Moreover it cannot be applied when raw and edited data from previous occasions of the survey are not available.

In our setting, the introduction of a model allows to define a score function that can be interpreted as an estimate of the expected error of the observation, and consequently the threshold value η can be directly linked to the level of accuracy of the estimates of interest.

The proposed score function for the total t_j^* is based on the relative individual error for the ith unit with respect to the variable Y_j. The latter is defined as the ratio between the (weighted) expected error and the reference estimate t_j^{ref} of the target parameter t_j^*, that is

$$r_{ij} = \frac{w_i(y_{ij} - \hat{y}_{ij})}{t_j^{ref}}, \tag{13}$$

where the prediction $\hat{y}_{ij}$ for the variable Y_j on the ith unit is obtained plugging in the MLE of the parameters in the conditional expectation as expressed in Formula (7). The local score function is defined as

$$S_{ij} = |r_{ij}|. \tag{14}$$

Note that, the estimated expected error is $\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i = \hat{\tau}_2(\boldsymbol{y}_i)(\boldsymbol{y}_i - \hat{\tilde{\boldsymbol{\mu}}}_i)$, that is the product of the probability of being in error, $\hat{\tau}_2$, and the difference $(\boldsymbol{y}_i - \hat{\tilde{\boldsymbol{\mu}}}_i)$ between the observed value and the expectation of the true value conditional on the event that $\boldsymbol{y}_i$ is contaminated. Hence, S_{ij} can be seen as composed of a "risk component" $\hat{\tau}_2(\boldsymbol{y}_i)$ and an "influence component" $w_i(\boldsymbol{y}_i - \hat{\tilde{\boldsymbol{\mu}}}_i)$.

In the next paragraph, an algorithm for the selection of units to be accurately edited is described. For $i = 0, 1, \ldots, n$, let us define R_{ij} as the absolute value of the expected residual relative error for the variable Y_j remaining in data after removing errors in the first i ordered units (when $i = 0$ no observations are selected), that is $R_{ij} = \left|\sum_{k>i}^{n} r_{kj}\right|$. Once an accuracy level (threshold) η is chosen, the selective editing procedure consists of:

1. sorting the observations in descending order according to the value of S_{ij};
2. finding $n_e \equiv n_e(\eta)$ such that $n_e = min\{k^* \in (0, 1, \ldots, n) | R_{kj} < \eta, \ \forall k \geq k^*\}$, that is, selecting the first n_e units such that all the residual errors R_{kj} (for a given j) computed from the $(n_e + 1)$th to the last observation are below η.

This procedure implies that the absolute value of the expected difference between the estimator $\hat{t}_j^e$ computed on edited data and the estimator $\hat{t}_j^*$ computed on true data is below the accuracy level ηt_j^{ref}. Furthermore, $S_{kj} < 2\eta, \ \forall k > n_e$ for each unit not revised, implying that also the error at micro level is bounded.

The algorithm described so far is easily extended to the multivariate case by defining a global score function $GS_i = max_j S_{ij}$. The two-step algorithm is:

1. order the observations with respect to GS_i (decreasing order);
2. find n_e such that $n_e = min\{k^* \in (0, 1, \ldots, n) | max_j R_{kj} < \eta, \ \forall k \geq k^*\}$, that is, select the first n_e units such that all the residual errors R_{kj} computed from the $(n_e + 1)$th to the last observation are below η.

The above accuracy properties are still valid for all the variables. In fact,

$$\left| E\left(\hat{t}_j^e - \hat{t}_j^*\right) \right| < \eta t_j^{ref}, \ j =, 1, \ldots, J$$

and $S_{kj} < 2\eta, \ \forall k > n_e, j = 1, \ldots, J$.

We remark that different values of the parameter η can be set for the analyzed variables in order to take into account their possible different importance.

The reference estimate t_j^{ref} in Formula (13) can be computed by using the predictions $\hat{y}_{ij}$ obtained by the contamination model,

$$\hat{t}_j^{ref} = \sum_{i=1}^{n} w_i \hat{y}_{ij}.$$

As an alternative, reference estimates can be obtained by using data from a previous survey occasion.

6. Application to Real Data

In this section we describe an experimental evaluation based on data from the 2008 Istat *Survey on small and medium enterprises*. The application refers to the subset of enterprises in the Nace Rev2 sections B, C, D and E corresponding to aggregation of economic activities in *Manufacturing, mining and quarrying and other industry*. This group of units ($N = 5{,}399$) has been used as the reference population (U) and for this population the variables *turnover* (X) and *labor cost* (Y) have been used, assuming that the available data are error-free. Errors are artificially introduced into the Y variable according to some error mechanisms frequently met in the context of NSI surveys; they are explicitly described in the next paragraphs. We suppose that the population parameter to be estimated is the total of the variable Y. The variable X is used as a covariate in the contamination model to obtain predictions for Y. The Gaussian contamination model is assumed for log-transformed data, according to Formula (8).

A Monte Carlo study based on 1,000 iterations has been carried out to study the performance of the proposed selective editing strategy. Each iteration of the Monte Carlo experiment consists of the following steps:

1. ***Sampling***
   Draw a simple random sample without replacement (srswor) s_a of $n_a = 1{,}000$ observations from the target population U.

2. ***Data contamination***
   - Multiply Y values by 1,000 in 1% of data.
   - Swap the first two digits of Y values in 2% of data.
   - Swap the last two digits of Y values in 2% of data.
   - Replace the Y value with the value "1" in 2% of data.

3. ***Model estimation and score computation***
   Compute on the logarithm of data the MLEs of the model parameters and use them to calculate the score function (14) for each unit. Order observations accordingly. In order to assess the impact of the risk component τ_2, a score function based only on the influence component $y_i - \hat{\tilde{\mu}}_i$ is also computed.

4. ***Selective editing***
   Given the threshold η, the most influential observations n_e are selected according to the procedure described in Section 5. In an alternative experiment, we have selected $n_{\tilde{e}}$ observations according to an analogous procedure where the score function is based only on the influence component, as described in the previous step. The selected units are replaced with the corresponding true values.

5. ***Target estimates***
   Compute the Horvitz-Thompson estimates of the variable Y on the true data ($\hat{t}_y^{*}$), on the corrupted data ($\hat{t}_y$), and on the two sets of edited data, that is, $\hat{t}_y^{e}$ and $\hat{t}_y^{\tilde{e}}$.

The results are summarized through the empirical relative root mean squared error (RRMSE) and the empirical relative bias (RB) based on the 1,000 Monte Carlo realizations $\hat{t}_y^{*(i)}$, $\hat{t}_y^{(i)}$, $\hat{t}_y^{e(i)}$ and $\hat{t}_y^{\tilde{e}(i)}$ $(i = 1, \ldots, 1{,}000)$ of the three estimators in Step 5. The error is to be intended as deviation from the estimate based on true data $\left(\hat{t}_y^{*(i)}\right)$, because we are interested in evaluating the effectiveness of the methods regardless of the sampling error. Thus, for instance, for the estimator $\hat{t}_y^{(i)}$ RRMSE and RB are defined respectively as:

$$RRMSE = \sqrt{\frac{1}{1{,}000}\sum_{i=1}^{1{,}000}\left(\frac{\hat{t}_y^{(i)} - \hat{t}_y^{*(i)}}{\hat{t}_y^{*(i)}}\right)^2}$$

and

$$RB = \frac{1}{1{,}000}\sum_{i=1}^{1{,}000}\frac{\hat{t}_y^{(i)} - \hat{t}_y^{*(i)}}{\hat{t}_y^{*(i)}}.$$

Empirical RRMSE and empirical RB are reported in the 3rd and 4th column of Table (1) according to different threshold levels η. The efficiency of the procedure is measured by comparing the percentage of selected units $n_e\%$ with $n_{\tilde{e}}\%$, and with the percentage of observations $n_{e*}\%$ we would obtain by using true values as predictions, that is, by replacing the expression in Formula (13) with $(y_i - y_i^{*})/\hat{t}_y^{*}$. The average percentage of selected units ($n_e\%$) is also reported in the last column of the table.

Table 1. Empirical RRMSE, empirical RB of the estimates computed on contaminated and edited data, and average percentage of edited units according to the threshold η

η		$\hat{t}_y$	$\hat{t}_y^e$	$\hat{t}_y^{\tilde{e}}$	$n_{e*}\%$	$n_e\%$	$n_{\tilde{e}}\%$
0.05	RRMSE	10.882	0.016	0.011	1.0	1.0	12.1
	RB	9.893	−0.006	0.005	–	–	–
0.01	RRMSE	10.542	0.006	0.001	1.7	2.4	61.0
	RB	9.641	0.002	0.000	–	–	–
0.005	RRMSE	10.910	0.005	0.000	2.4	3.1	71.0
	RB	9.984	0.002	0.000	–	–	–

The impact of errors on the estimates is particularly harmful; in fact the RRMSE computed on observed data ranges from 10.54 to 10.91. After the selective editing procedure, the RRMSE dramatically decreases, and its value is (on average) below the accuracy level required and expressed by η. As far as the efficiency is concerned, the results show that n_e is close to the number of selected observations n_{e*} that would be selected in the ideal situation in which true data were known. Based on the comparison of n_e with $n_{\tilde{e}}$, we can note that not taking into account the risk component leads to the selection of a much higher number of observations.

These results are important because they show that the editing procedure performs satisfactorily even though data clearly do not satisfy the assumptions of the model; in particular the error mechanism is clearly far from the normality assumption.

In order to obtain a picture of some important parameters of the procedure, a single Monte Carlo realization is described in Figure 1 and Figure 2.

In Figure 1(a) outliers and selected observations according to $\eta = 0.01$ are reported on the scatter plot of contaminated log data. An observation is considered an outlier if

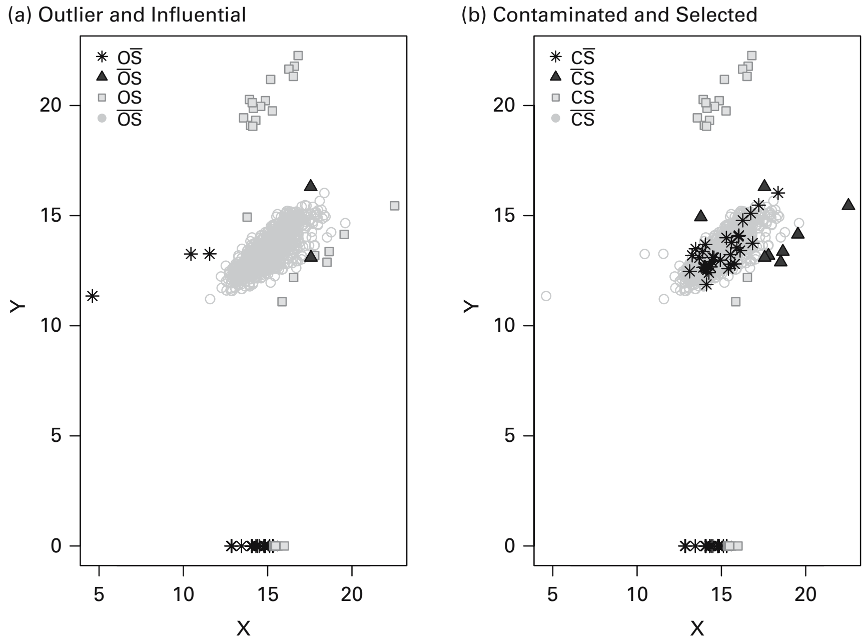

Fig. 1. Outliers, contaminated and selected observations in logdata

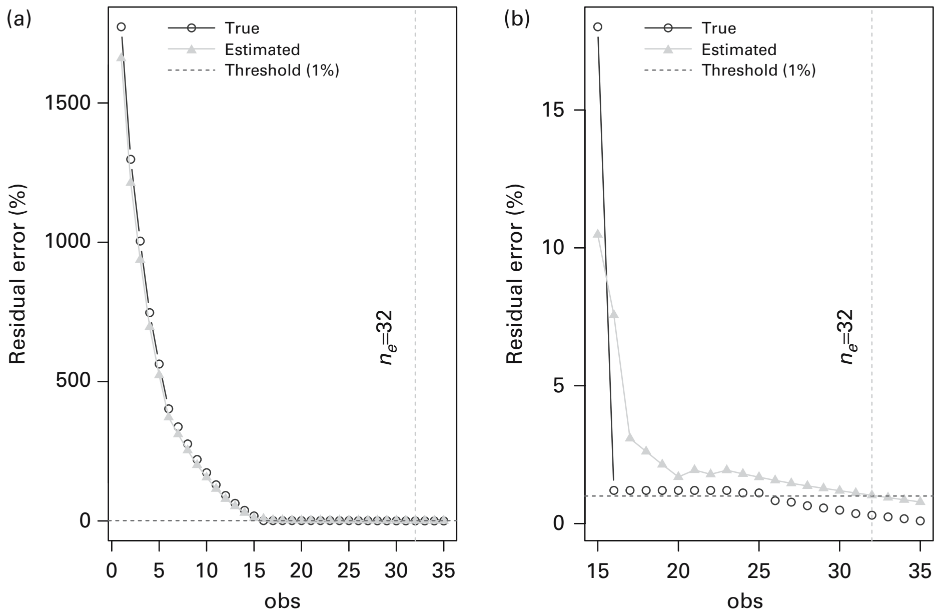

Fig. 2. Estimated vs. true residual error

the estimated conditional probability of being in error $\hat{\tau}_2(\boldsymbol{y}_i)$ is greater than 0.5. Observations classified as outliers and not selected are denoted by $O\bar{S}$, selected and not outliers by $\bar{O}S$, as both outliers and selected by OS. The remaining units that are not selected and not outliers are denoted by $\overline{OS}$.

Figure 1(b) shows contaminated units and selected observations. Observations that are contaminated but not selected are denoted by $C\bar{S}$, not contaminated units that are selected by $\bar{C}S$, contaminated and selected observations by CS. The remaining units that are not selected and not contaminated are denoted by $\overline{CS}$.

The estimated and true residual errors are reported in Figure 2(a) for the first 35 observations, while in Figure 2(b) the same residual errors are reported from the 15th observations onward in order to zoom in and avoid masking scale effects. The horizontal dashed line is the threshold and the vertical dashed line corresponds to the number of selected units in this experiment ($n_e = 32$).

Figure 2(a) and Figure 2(b) show that the estimated residual errors are close to the true residual errors. It is worth noting that the accuracy of the estimate is below the threshold even though many errors are left in the data (see Figure 2), as it is required from a selective editing procedure. As far as the outliers are concerned, it is interesting to note that not all the outliers are considered influential by the procedure, and on the other hand some selected units are not outliers. The distinction is due to the impact of the estimated error on the estimates.

7. Conclusions

In this article a model-based approach to selective editing is proposed. The considered model is referred to in the literature as a contamination model and it is typically used to detect gross errors. The introduction of a model for both true data and error mechanism makes it possible to define a score function that can be interpreted as an estimate of the error affecting data. This allows the relation between the choice of a threshold for selection of the units to be reviewed and the level of accuracy required for the estimates to be made explicit. According to this peculiarity, an algorithm to select influential errors is proposed.

Since the remaining uncertainty due to the unedited data can be properly estimated under the model-based approach based on latent classes, it is possible to determine a threshold for the score function conditional on the actual sample observations of the current survey. By contrast, traditional methods do not assume an explicit measurement-error model and the threshold value for the score function is usually set based on edited data from previous surveys. Since the error mechanism and the data distribution do not remain exactly the same over time, the remaining uncertainty of the current unedited data can only be heuristically controlled.

The main advantages of the proposed approach are due to the introduction of an explicit model for true data and error mechanism, and of course the limits lie in the validity of the hypothesis on which the model is based. Nevertheless, the experimental studies carried out in this paper suggest that the use of a Gaussian contamination model can be usefully applied also when data and error mechanism deviate from the model assumptions, especially when data are contaminated by gross errors.

An implication of the error model described is that errors on different items are not independent of each other; this means that the intermittence nature of the error is at record level and not at variable level. Further studies should be devoted to study more general models able to encompass this assumption.

The use of edits in such a procedure is an open issue. However, some remarks are needed in this respect. Soft edits such as ratio edits are implicitly taken into account by the procedure, since the analysis of anomalous relationships between variables is the core of the proposed approach. By contrast, it is not easy to treat hard edits consistently in the model, and further analysis should be devoted to this aspect.

The editing described in the article can be classified as “output editing”, meaning that a certain amount of data from the current survey is needed to estimate the model. However, it can also be used from an “input editing” perspective, in situations where the model is applied to a previous survey occasion, and the estimated parameters are used to select influential units in the current survey.

Finally, even though the article describes a strategy composed of a latent class model for predicting data and an algorithm to select influential units, they can be used independently of each other. In fact, parameter estimation, computation of predicted values and selection of influential errors are separately implemented in the R package SeleMix.

Appendix

The density in (11),

$$f_1(\boldsymbol{y}_i^*|\boldsymbol{y}_{i,o}) = \delta(\boldsymbol{y}_{i,o}^* - \boldsymbol{y}_{i,o})f(\boldsymbol{y}_{i,m}^*|\boldsymbol{y}_{i,o}^*) = \delta(\boldsymbol{y}_{i,o}^* - \boldsymbol{y}_{i,o})f(\boldsymbol{y}_{i,m}^*|\boldsymbol{y}_{i,o}),$$

is:

$$f_1(\boldsymbol{y}_i^*|\boldsymbol{y}_{i,o}) = \delta(\boldsymbol{y}_{i,o}^* - \boldsymbol{y}_{i,o})N(\boldsymbol{y}_{i,m}^*; \boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\boldsymbol{y}_{i,o}, \boldsymbol{\Sigma}_{m|o}).$$

In the density (12),

$$f_2(\boldsymbol{y}_i^*|\boldsymbol{y}_{i,o}) = N(\boldsymbol{y}_{i,o}^*; \tilde{\boldsymbol{\mu}}_{i,o}, \tilde{\boldsymbol{\Sigma}}_{oo})f(\boldsymbol{y}_{i,m}^*|\boldsymbol{y}_{i,o}^*),$$

the factor $f(\boldsymbol{y}_{i,m}^*|\boldsymbol{y}_{i,o}^*)$ is the true-data conditional distribution corresponding to the missing pattern being considered, and can be derived from the true-data multivariate Gaussian distribution in Formula (2):

$$f(\boldsymbol{y}_{i,m}^*|\boldsymbol{y}_{i,o}^*) = N(\boldsymbol{y}_{i,m}^*; \boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\boldsymbol{y}_{i,o}^*, \boldsymbol{\Sigma}_{m|o}),$$

where

$$\boldsymbol{\alpha}_{m,i|o} = \boldsymbol{\mu}_{i,m} - \boldsymbol{\beta}_{m|o}\boldsymbol{\mu}_{i,o}, \quad \boldsymbol{\beta}_{m|o} = \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1} \quad \boldsymbol{\Sigma}_{m|o} = \boldsymbol{\Sigma}_{mm} - \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{om}.$$

In order to obtain an explicit expression for the second density $f_2(\boldsymbol{y}_i^*|\boldsymbol{y}_{i,o})$, it suffices to observe that $N(\boldsymbol{y}_{i,o}^*; \tilde{\boldsymbol{\mu}}_{i,o}, \tilde{\boldsymbol{\Sigma}}_{oo})N(\boldsymbol{y}_{i,m}^*; \boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\boldsymbol{y}_{i,o}^*, \boldsymbol{\Sigma}_{m|o})$ is the factorisation of a multivariate Gaussian density $N(\boldsymbol{y}_{i,o}^*, \boldsymbol{y}_{i,m}^*; \bar{\boldsymbol{\mu}}_i, \bar{\boldsymbol{\Sigma}})$ of which the parameters are:

$$\bar{\boldsymbol{\mu}}_i = [\tilde{\boldsymbol{\mu}}'_{i,o}, (\boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\tilde{\boldsymbol{\mu}}_{i,o})']', \quad \bar{\boldsymbol{\Sigma}} = \begin{pmatrix} \bar{\boldsymbol{\Sigma}}_{oo} & \bar{\boldsymbol{\Sigma}}_{om} \\ \bar{\boldsymbol{\Sigma}}_{mo} & \bar{\boldsymbol{\Sigma}}_{mm} \end{pmatrix},$$

where:

$$\begin{aligned}
\bar{\boldsymbol{\Sigma}}_{oo} &= \tilde{\boldsymbol{\Sigma}}_{oo} = \frac{\alpha - 1}{\alpha}\boldsymbol{\Sigma}_{oo} \\
\bar{\boldsymbol{\Sigma}}_{mo} &= \bar{\boldsymbol{\Sigma}}'_{om} = \boldsymbol{\beta}_{m|o}\bar{\boldsymbol{\Sigma}}_{oo} = \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\tilde{\boldsymbol{\Sigma}}_{oo} = \frac{\alpha - 1}{\alpha}\boldsymbol{\Sigma}_{mo} \\
\bar{\boldsymbol{\Sigma}}_{mm} &= \boldsymbol{\Sigma}_{m|o} + \bar{\boldsymbol{\Sigma}}_{mo}\tilde{\boldsymbol{\Sigma}}_{oo}^{-1}\bar{\boldsymbol{\Sigma}}_{om} = \\
&= \boldsymbol{\Sigma}_{mm} - \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{om} + \frac{\alpha - 1}{\alpha}\boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{om} = \\
&= \boldsymbol{\Sigma}_{mm} - \frac{1}{\alpha}\boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{om}.
\end{aligned}$$

From the previous formulas it follows that the expected value of Y_i^* conditional on the observed value $y_{i,o}$ is:

$$E(\boldsymbol{Y}_i^*|\boldsymbol{y}_{i,o}) = \tau_1(\boldsymbol{y}_{i,o})\boldsymbol{E}_{1i} + \tau_2(\boldsymbol{y}_{i,o})\boldsymbol{E}_{2i},$$

where

$$\begin{aligned}
\boldsymbol{E}_{1i} &= [\boldsymbol{y}'_{i,o}, (\boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\boldsymbol{y}_{i,o})']' = [\boldsymbol{y}'_{i,o}, (\boldsymbol{\mu}_{i,m} + \boldsymbol{\beta}_{m|o}(\boldsymbol{y}_{i,o} - \boldsymbol{\mu}_{i,o}))']', \\
\boldsymbol{E}_{2i} &= [\tilde{\boldsymbol{\mu}}'_{i,o}, (\boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\tilde{\boldsymbol{\mu}}_{i,o})'] = [\tilde{\boldsymbol{\mu}}'_{i,o}, (\boldsymbol{\mu}_{i,m} + \boldsymbol{\beta}_{m|o}(\tilde{\boldsymbol{\mu}}_{i,o} - \boldsymbol{\mu}_{i,o}))']'.
\end{aligned}$$

The case of incomplete log-normal data can also be easily treated, in fact with a slight shift of the notation and letting $\boldsymbol{y}_{i,0} = ln(z_{i,o})$ we have:

$$\boldsymbol{E}(\boldsymbol{Z}_i^* | z_{i,o}) = \tau_1(\ln(z_{i,o}))\boldsymbol{E}_{1i}^L + \tau_2(\ln(z_{i,o}))\boldsymbol{E}_{2i}^L,$$

where:

$$\boldsymbol{E}_{1i}^L = \left[\exp\left(\boldsymbol{y}_{i,o} + \frac{1}{2}\boldsymbol{\Sigma}_{00}^d\right)', \quad \exp\left(\boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\boldsymbol{y}_{i,o} + \frac{1}{2}\boldsymbol{\Sigma}_{m|0}^d\right)'\right]$$

$$\boldsymbol{E}_{2i}^L = \left[\exp\left(\tilde{\boldsymbol{\mu}}_{i,o} + \frac{1}{2}\bar{\boldsymbol{\Sigma}}_{00}^d\right)', \quad \exp\left(\boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\tilde{\boldsymbol{\mu}}_{i,o} + \frac{1}{2}\bar{\boldsymbol{\Sigma}}_{m|0}^d\right)'\right],$$

and $\boldsymbol{\Sigma}^d$ denotes the vector of the diagonal elements of the matrix $\boldsymbol{\Sigma}$.

8. References

Buglielli, M.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, Brussels, 22–24 February 2011.

Buglielli, T. and Guarnera, U. (2011). SeleMix: Selective Editing via Mixture models. R package version 0.8.1. Available at: http://cran.r-project.org/web/packages/SeleMix/index.html (accessed October 9, 2013).

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

Ghosh-Dastidar, B. and Schafer, J.L. (2006). Outlier Detection and Editing Procedures for Continuous Multivariate Data. Journal of Official Statistics, 22, 487–506.

Granquist, L. (1997). The New View on Editing. International Statistical Review, 65, 381–387.

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. In Proceedings of UN/ECE Work Session on Statistical Data Editing, 21–23 April, Vienna. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/2008/04/sde/wp.31.e.pdf

Jäder, A. and Norberg, A. (2005). A Selective Editing Method Considering both Suspicion and Potential Impact, Developed and Applied to the Swedish Foreign Trade Statistics. In Proceedings of UN/ECE Work Session on Statistical Data Editing, 16–18 May, Ottawa. Available at: http://www.unece.org/stats/documents/2005.05.sde.htm (accessed October 9, 2013).

Latouche, M. and Berthelot, J.M. (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing 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.

Meng, X.L. and Rubin, D.B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: a General Framework. Biometrika, 80, 267–278.

Norberg, A., Adolfsson, C., Arvidson G., Gidlund, P., and Nordberg, L., (2010). A General Methodology for Selective Data Editing. Stockholm: Statistics Sweden. Available at: http://gauss.stat.su.se/master/statdatabaser/HT10/Literature/SwedishEditingMethods.pdf (accessed October 9, 2013).

Received January 2013
Accepted September 2013