Use of Contamination Models for Selective Editing

M.Teresa Buglielli, Marco Di Zio, Ugo Guarnera

Istat, Italian National Institute of Statistics

Abstract

The aim of selective editing is to prioritize for reviewing the observations with the potential highest impact on the target estimates. The ordering is obtained by assigning to each unit a “score" based on a “risk component" and an “influence component". In this paper we assume a lognormal model for the correct data and an “intermittent" error mechanism such that a proportion of data is contaminated by an additive Gaussian error. The resulting distribution for the observed data is a mixture of two Gaussian distributions with the same mean vector but proportional covariance matrices, where the “largest" one corresponds to contaminated data. For each unit, the probability of belonging to the mixture component that corresponds to contaminated data is the risk component, while the influence component for a given variable is obtained as expected difference between true and observed value of that variable conditioned on the observed value and on the event that the observation is contaminated. Thus, the scores can be interpreted as expected values of the errors conditional on the observed data. Consequently, a set of units can be selected such that the expected residual error in data is below a prefixed threshold. The model that we propose in this paper allows also to deal with the case where non contaminated variables are available.

Keywords Mixture Models, Score Function, Influential Errors.

1 Introduction

In the last years selective editing is becoming more and more important in the editing and imputation phase of business surveys. Selective editing is based on the appealing idea of looking for important errors (errors with a harmful impact on estimates), in order to focus the treatment (follow up, recontact) only on this subset of units to reduce the cost of E&I, mantaining at the same time a certain 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 error on the estimates of interest (Latouche and Berthelot, 1992), and then all the units above a given threshold are selected. The definition of the *score function* involves of two elements: the probability of the observation to be in error (“risk component"), and the magnitude of the error, or “influence component" (see for example Jäder and Norberg, 2005; Hedlin, 2003).

The most commonly used methods to determine these elements are based on the comparison of the observed value of a certain variable for a given unit with a value predicted by an implicitly or explicitly assumed model. The differences between observed and predicted values are then used to calculate appropriate score functions which allow to identify the observations having the largest impact on the target estimates. However, the difference between observed and predicted values is determined by the potential error as well as by the natural variability of the analysed quantity. In the usual setting, there is no possibility to distinguish these two elements, and the score of an observation is not directly related to the expected error of that unit. As a consequence we will not be able to relate the selective editing threshold to the desired degree of accuracy in the final estimates.

In order to deal with this limitation, Di Zio et al. (2008) have proposed to use a latent variable model allowing, under certain assumptions, to estimate the expected error associated to each unit. The method is based on the use of contamination normal models (see Little, 1988), where it is assumed that the distribution of the erroneous data can be obtained from the distribution of the error free data by inflating the variance (see Ghosh-Dastidar and Schafer, 2006).

In this paper, we study the selective editing based on contamination normal models in a regression context, generalizing the one described in Bellisai et al., (2009) to the multivariate case. An experimental application to the 2005 Istat survey on *Small and Medium Enterprises* (SME) is carried out.

The paper is structured as follows. In Sections 2 and 3 we descibe the contamination model and the selective editing approach respectively. In particular in Section 3 there is also a proposal for a new score function with certain desired properties. In Section 4 we describe the application to the *SME* data and discuss the results.

2 The model

The idea is to consider the observed data as coming from a mixture of two distributions corresponding to “true” (i.e. correct) data and contaminated data respectively. This is generally referred to as “contamination model” and often used for automatic outlier identification. True data are represented as a $n \times p$ matrix $\mathbf{Z}^*$ of n independent realizations from a random p-vector assumed to follow a log-normal distribution whose parameters may depend on some set of q covariates not affected by error. If $\mathbf{Y}^* = \ln \mathbf{Z}^*$, we have the regression model:

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

where $\mathbf{X}$ is a $n \times q$ matrix whose rows are the measures of the q covariates on the n units, $\mathbf{B}$ is the $q \times p$ matrix of the coefficients, and $\mathbf{U}$ is the $n \times p$ matrix of normal residuals:

$$\mathbf{U} \sim N(\cdot; 0, \mathbf{\Sigma})$$

We assume that observed (log)data $\mathbf{Y}$ are correct or erroneous according to a Bernoulli r.v. I with parameter π, where $I = 1$ if an error occurs and $I = 0$ otherwise. Further, given that $I = 1$, errors affect data through an additive mechanism represented by a Gaussian r.v. ϵ with zero mean and covariance matrix $\mathbf{\Sigma}_\epsilon$ proportional to $\mathbf{\Sigma}$, i.e., given $\{I = 1\}$:

$$\mathbf{Y} = \mathbf{Y}^* + \epsilon, \quad \epsilon \sim N(0, \mathbf{\Sigma}_\epsilon), \;\; \mathbf{\Sigma}_\epsilon = (\alpha - 1)\mathbf{\Sigma}, \;\; \alpha > 1.$$

The error model can be formally expressed through the conditional distribution:

$$f_{\mathbf{Y}|\mathbf{Y}^*}(\mathbf{y}|\mathbf{y}^*) = (1-\pi)\delta(\mathbf{y} - \mathbf{y}^*) + \pi N(\mathbf{y}; \mathbf{y}^*, \mathbf{\Sigma}_\epsilon). \tag{2}$$

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

As a particular case, the set of $\mathbf{X}$-variates may be empty, so that variables $\mathbf{Y}_i$, $(i = 1 \ldots, n)$ are normally distributed with the same mean vector μ. In the previous model, it is crucial the "intermittent" nature of the error implied by the introduction of the Bernoullian variable I. Due to this assumption, it is conceptually possible to think of data as partitioned into correct and erroneous, and to estimate, for each observation, the probability of being correct or corrupted.

The distribution of the observed data is easily derived multiplying the true data density which leads to formula (1) and the error density (2), and integrating over $\mathbf{Y}^*$:

$$f_{\mathbf{Y}}(\mathbf{y}) = (1-\pi)N(\mathbf{y}; \mathbf{B}'\mathbf{x}, \mathbf{\Sigma}) + \pi N(\mathbf{y}; \mathbf{B}'\mathbf{x}, \alpha\mathbf{\Sigma}) \tag{3}$$

Expression (3) represents a mixture of two regression models having the same coefficient matrix $\mathbf{B}$ but different (though proportional) residual variance-covariance matrices. The last distribution relates to observed data and can be estimated by maximizing the likelihood based on n sample units via an ECM algorithm (see Meng and Rubin, 1993).

3 The selective editing approach

In this section we illustrate how to use the contamination model in the context of selective editing. As we have anticipated, the selective editing method is based on the distribution of the true data conditional on the observed data. Starting from the error model (2) and the assumed (conditional) distribution of the error-free data Y^*, for each observation $(\mathbf{x}_i, \mathbf{y}_i)$, it can be easily derived, via Bayes formula, the following conditional distribution of $\mathbf{y}^*$ given $\mathbf{y}, \mathbf{x}$:

$$f_{\mathbf{Y}^*|\mathbf{X},\mathbf{Y}}(\mathbf{y}^*|\mathbf{x}, \mathbf{y}) = \tau_1(\mathbf{x}, \mathbf{y})\delta(\mathbf{y}^* - \mathbf{y}) + \tau_2(\mathbf{x}, \mathbf{y})N(\mathbf{y}^*; \tilde{\mu}_{x,y}, \tilde{\mathbf{\Sigma}}) \tag{4}$$

where τ_1 and τ_2 are the posterior probabilities of belonging to correct and erroneous data respectively:

$$\begin{aligned}
\tau_1(\mathbf{x}_i, \mathbf{y}_i) &= Pr(\mathbf{y}_i = \mathbf{y}_i^*|\mathbf{x}_i, \mathbf{y}_i) = \frac{(1-\pi)N(\mathbf{y}_i; \mathbf{B}'\mathbf{x}_i, \mathbf{\Sigma})}{(1-\pi)N(\mathbf{y}_i; \mu, \mathbf{\Sigma}) + \pi N(\mathbf{y}_i; \mathbf{B}'\mathbf{x}_i, \alpha\mathbf{\Sigma})}, \\
\tau_2(\mathbf{x}_i, \mathbf{y}_i) &= Pr(\mathbf{y}_i \neq \mathbf{y}_i^*|\mathbf{x}_i, \mathbf{y}_i) = 1 - \tau_1(\mathbf{x}_i, \mathbf{y}_i), \\
i = 1, \ldots & \quad n,
\end{aligned}$$

and

$$\tilde{\mu}_{x,y} = \frac{(\mathbf{y} + (\alpha - 1)\mathbf{B}'\mathbf{x})}{\alpha}; \quad \tilde{\mathbf{\Sigma}} = \left(1 - \frac{1}{\alpha}\right)\mathbf{\Sigma}.$$

It is worthwhile to note that τ_2 can be interpreted as a " risk component" to be considered in the score function definition.

From formula (4) the distribution $f_{Z^*|Z}$ in the original scale can also be derived. In fact, as it can easily verified, the conditional distribution of $\mathbf{z}_*$ given $\mathbf{z}$ is a mixture of a delta function with mass in $\mathbf{z}$ and a log-normal distribution $LN(\mathbf{z}; \tilde{\mu}_{x,\ln z}, \tilde{\Sigma})$ with parameters $\tilde{\mu}_{x,\ln z}, \tilde{\Sigma}$. The mixing weights are the same as in the mixture (4). Thus, we can express the target conditional distribution as:

$$f_{\mathbf{Z}^*|\mathbf{Z}}(\mathbf{z}^*|\mathbf{z}) = \tau_1(\ln(\mathbf{z}))\delta(\mathbf{z}^* - \mathbf{z}) + \tau_2(\ln(\mathbf{z}))LN(\mathbf{z}^*; \tilde{\mu}_{x,\ln z}, \tilde{\mathbf{\Sigma}}) \tag{5}$$

where for the sake of simplicity, we have suppressed the X variables in the notation whenever they appear as conditioning variables, by incorporating them in the parameters of the relevant distributions. Hereafter, this notation will be used. The parameters of the distribution (5) are the same as those of the distribution (3). Thus the MLE estimates obtained via ECM on the basis of the observed data can be used as plug in estimates for the distribution (5). Thus, using these estimates, ”prediction” of ”true” values $\hat{\mathbf{z}}_i$, conditional on observed values $\mathbf{z}_i$, can be obtained for all the observations $i = 1 \ldots, n$ as:

$$\hat{\mathbf{z}}_i = E(\mathbf{z}_i^* | \mathbf{z}_i) = \int \mathbf{z}_i^* f_{\mathbf{Z}^*|\mathbf{Z}}(\mathbf{z}_i^* | \mathbf{z}_i) d\mathbf{z}_i^*.$$

Thus, for $i = 1 \ldots, n,$ the *expected error* $_i$ can be defined as:

$$\mathcal{E}_i = \hat{\mathbf{z}}_i - \mathbf{z}_i.$$

It is instructive to re-write the expected error as $\tau_2(\ln(\mathbf{z})) \times E(\mathbf{Z}^* - \mathbf{Z}|(\mathbf{Z} = \mathbf{z}, I = 1))$. In this way, if $E(\mathbf{Z}^* - \mathbf{Z}|(\mathbf{Z} = \mathbf{z}, I = 1))$ (possibly multiplied by a sample weight) is interpreted as an “influence component", the expected error is explicitly expressed as product of a “risk component" and an “influence component". More in general, we can obtain ”robust estimates” of any function $g(y_i)$:

$$\hat{g}(\mathbf{z}_i) = \int g(\mathbf{z}_i^*) f_{\mathbf{Z}^*|\mathbf{Z}}(\mathbf{z}_i^* | \mathbf{z}_i) d\mathbf{z}_i^*.$$

The micro-level predictions can be used for the estimation of some finite population quantities in randomization inference. For instance, if target estimate is a total $T_z = \sum_i g(z_i)$ we can obtain the ”robust estimate”:

$$T_z^* = \sum_i w_i \hat{g}(\mathbf{z}_i)$$

where w_i is a sampling weight. In the following, we assume that the target estimate is given by the total T_z of the variable Z, *i.e.* $T_z = \sum_i z_i$, so that $T_z^* = \sum_i w_i \hat{\mathbf{z}}_i$. Actually, in the present context, we are interested in using predictions for selective editing more than for robust estimation. To this aim, note that the absolute value of the estimated expected errors can be used to define a score function. Interpreting the score function as expected error is particularly useful in that it makes it possible to estimate the residual error remaining in the data after the interactive editing of the units with the highest expected error. Thus, a number of units to be interactively reviewed can be chosen such that the residual error is below a prefixed threshold. It is natural to define the threshold in terms of ratio between expected residual error and a (robust) reference estimate. Let us introduce the relative individual error r_i as the ratio between the (weighted) expected error and the reference estimate T_z^*:

$$r_i = \frac{w_i(\hat{z}_i - z_i)}{T_z^*}.$$

Moreover, let us define the score function as $S_i = |r_i|$, and R_M as the absolute value of the expected residual percentage error remaining in data after removing errors in the units belonging to the set M :
$R_M = \Big| \sum_{i \in \bar{M}} r_i \Big|$, where $\bar{M}$ denotes the complement of M in $(1, \ldots, n)$.

Once an “accuracy" threshold η is chosen, the selective editing procedure consists of:

1. sorting the observations in descending order according to the value of S_i;
2. selecting the first $\bar{k}$ units for reviewing, where:
   $\bar{k} = min\left\{k \in (1, \ldots, n) \mid R_{M_j} < \eta \;\; \forall j > k\right\}$, and M_m is the set composed of the first m units.

As last remark, we note that the ordering and stopping criterion ensure that expected error on each not revised unit is kept *under control*, since $S_k \leq 2\eta \;\; \forall k > \bar{k}$, (see Di Zio et al., 2008).

4 Application to real data

In this section an experimental application to the Istat survey on *Small and Medium Enterprises* (SME), year 2005, is illustrated. The variables considered in the application are the *Labour Cost (LC)*, that is the variable with errors, and the *Number of Employees (NE)*, that is considered correct and used as covariate in the regression model. The analysis is restricted to the subset of data of some Manufacturing subsections.

The procedure described in Sections 2 and 3 is applied to the *official* SME data that are artificially contaminated. The official data are our target (“true") data.

More in detail:

1. given the *true* data, the values of *LC* and *NE* are transformed in log scale and a percentage $p = 0.05$ of contaminated data is simulated by adding an error component drawn from a normal distribution $N(0, (\alpha - 1)\Sigma)$, where $\alpha = 10$;
2. the model parameters are estimated on the contaminated data obtained in step (1), and for each unit the score function S_i described in Section 3 is computed;
3. the units are ordered by descending values of S_i;
4. the units are selected with respect to an accuracy threshold η referring to the target estimates T that is the total of *LC*.

In Figure (1), observations classifed as influential errors (grey triangles) and as outliers (black dots) are depicted. The selection is made respect to a threshold $\eta = 0.005$. All the inluential errors are outliers, but we notice also that there are some outliers that are not influential error. This is an important peculiarity of selective editing that allows to save resources for data revision, in fact even if observations are classified as errors their impact on the estimates is negligible with respect to the chosen level η.

Figure (2) shows the line plot of the true (dashed line) and the estimated residual error (dotted line) for LC on the subset of the first 40 observations. The two curves are close each other, and this is an important evidence for the quality of the estimation methods.

A Monte Carlo study is performed by repeating the Steps (1) to (4) 100 times. Over those repetitions, the average of the number of selected units ($Av.n.sel$), the simulated relative bias ($Rbias$) , and the simulated relative root mean square error ($RRMSE$) are computed. This is carried out for different thresholds η. The results are shown in Table 1. We notice that the $RRMSE$ are quite close to the threshold η. We notice that for $\eta = 5\%$, the procedure suggests that no units should be selected, this is due to the fact that the error introduced has an impact on data, i.e. $RRMSE$ on observed data, equal to 3.9%.

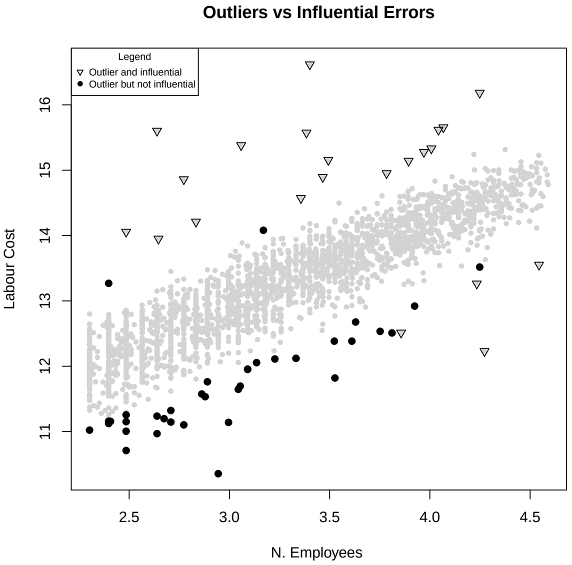

Figure 1: Outliers and Influential errors when $\eta = 0.005$

Table 1: $Av.n.sel$, $Rbias$, and $RRMSE$ according to different thresholds η

$\eta(\%)$	$Av.n.sel$	$Rbias(\%)$	$RRMSE(\%)$
5	0.1	3.3	3.6
2	1,8	2.2	2.3
1	4.4	1.4	1.4
0.5	9.3	0.7	0.8

The experiments show that selective editing based on contamination model and on the sorting and stopping selection criterion could be usefully applied to business surveys. The advantage is that the score funtion can be seen as the expected error of the unit, this implies that once the researcher decides what is the error that can be left in data the procedure automatically selects the influential units. Finally, we remark that the contamination model gives also, without any other effort, a robust estimate that can be used in the score function, making the selection procedure feasible also in the case we have no guess of the estimates to be produced.

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*. (http://www.unece.org/stats/documents/2009.10.sde.htm).

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* (http://www.unece.org/stats/documents/2008.04.sde.htm).

Ghosh-Dastidar B., Schafer J.L. (2006). Outlier Detection and Editing Procedures for

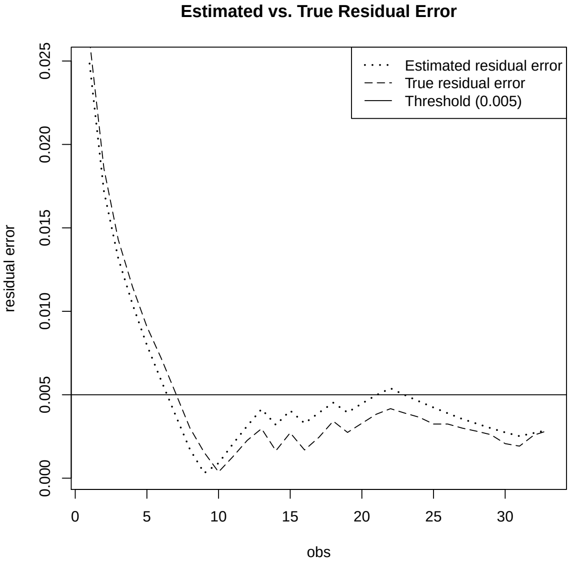

Figure 2: Estimated and true residual error functions for SME data with a threshold $\eta = 0.005$.

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* (http://www.unece.org/stats/documents/2005.05.sde.htm).

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*.

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*.

Lee H. (1995). *Outliers in Business Surveys, in: Business Survey Methods*, Cox B.G., Binder D.A., Chinappa B.N., Christanson A., Colledge M.J. and Kott P.S. (Eds), John Wiley and Sons, Inc. 503-526.

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*.

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