CONTAMINATION MODELS FOR THE DETECTION OF OUTLIERS AND INFLUENTIAL ERRORS IN CONTINUOUS MULTIVARIATE DATA

Marco Di Zio - dizio@istat.it
Italian National Institute of Statistics

Ugo Guarnera, Orietta Luzi - Italian National Institute of Statistics

UN/ECE, Wien, 22 April 2008

Selective Editing

Key-elements for selective editing are:

- *score function*
- *cut-off* value determining the values to be recontacted

Score Function(1)

$\tilde{T}$: estimate of the total for Y computed after the selective editing,

T^*: estimate computed on data free of error.

Conditionally on the sample s, the goal is to have $|\frac{\tilde{T}-T^*}{T^*}| < \epsilon$.

Two main different random mechanisms should be taken into account:

1. the *error mechanism*;

2. the *probability of recovering the true value* (error free) for the unit selected for revision.

Score Function(2)

One way is to compute $\left|E\left(\frac{\tilde{T}-T^*}{T^*}\right)\right|$ where the expected value is with respect to the random mechanisms. HP: The prob. of recovering the true value is 1.

$$\left|E\left(\frac{\tilde{T}-T^*}{T^*}\right)\right| = \left|\sum_{i\notin M} E\left(\frac{Y_i-Y_i^*}{T^*}\right)\right| \leq \sum_{i\notin M}\left|E\left(\frac{Y_i-Y_i^*}{T^*}\right)\right|$$

a natural choice for the score function is $\left|E\left(\frac{Y_i-Y_i^*}{T^*}\right)\right|$.

but the sum (3rd term) is an approximation for the total error left in data (upper bound).

In the 2nd term, errors may compensate each other. It can happen that the difference between estimates is acceptable, but some important errors (at individual level) are still left in data.

Score Function(3)

To deal with these problems, we propose use both the elements and to select observations as

1. order the obs with respect to the $\left|E\left(\frac{Y_i - Y_i^*}{T^*}\right)\right|$ (decreasing order);

2. select for reviewing all the (ordered) obs until the $\bar{k}$th, where $\bar{k}$ is such that for any $k > \bar{k}$, $\left|\sum_{i=k}^{n} E\left(\frac{Y_i - Y_i^*}{T^*}\right)\right| < \epsilon$.

Note that $\left|E\left(\frac{Y_i - Y_i^*}{T^*}\right)\right| < 2\epsilon, \forall i > \bar{k}$.

Score Function(4)

According to the usual terminology,

1. score function: $\left| E\left(\frac{Y_i - Y_i^*}{T^*}\right)\right|$

2. the cut off value is computed by means of $\left|\sum_{i=k}^{n} E\left(\frac{Y_i - Y_i^*}{T^*}\right)\right| < \epsilon$.

Score Function(5)

The key problem now is to determine $E\left(\frac{Y_i - Y_i^*}{T^*}\right)$.

In general

$$E\left(\frac{Y_i - Y_i^*}{T^*}\right) = 0 \times P(Y_i = Y_i^*) + E\left(\frac{Y_i - Y_i^*}{T^*}\right) P(Y_i \neq Y_i^*)$$

Two quantities are involved:

1. the probability of being in error (*risk component*),

2. the magnitude of the error (*influence component*).

Contamination Model

Data free of errors are represented by a random p-vector $\mathbf{Y}^*$.

$$\mathbf{X}^* = log(\mathbf{Y}^*) \sim f_{\mathbf{X}^*}(\cdot) = N(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Our model assumes that the observed data are obtained by corrupting, with a certain probability, the r.v. $\mathbf{X}^*$ with an error that inflates the variance of the Gaussian distribution (Ghosh-Dastidar and Schafer, JOS, 2006).

$$\mathbf{x} = \mathbf{x}^* + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}_\epsilon)$ and $\boldsymbol{\Sigma}_\epsilon = (\alpha - 1)\boldsymbol{\Sigma}$, for $\alpha > 1$.

Intermittent Nature of the Error

An important feature is that errors do not affect all the data but only a portion of them.

This intermittent error mechanism implies that the distribution of the *observed data* $\mathbf{X}$, is a mixture of 2 pdf corresponding to erroneous and not erroneous data

$$f_{\mathbf{X}}(\mathbf{x}) = pN(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) + (1-p)N(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_{\epsilon})$$

$p \in [0, 1]$ can be interpreted as the proportion of non erroneous data.

The Conditional Model

The real goal is to compute the pdf of the free of error data given that we have observed some values, i.e. the conditional pdf

$$f_{\mathbf{x}^*|\mathbf{x}}(\mathbf{x}^*|\mathbf{x}) = \tau_1(\mathbf{x})\delta(\mathbf{x}^* - \mathbf{x}) + \tau_2(\mathbf{x})N(\mathbf{x}^*; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$$

where $\tau_1(\mathbf{x})$ is the posterior probability of being free of error (given the observed values $\mathbf{x}$), and $\tau_2(\mathbf{x})$ is the posterior probability of being in error: for each unit i, $\tau_1(\mathbf{x}_i) = Pr(\mathbf{x}_i = \mathbf{x}_i^*|\mathbf{x}_i)$ and $\tau_2(\mathbf{x}_i) = Pr(\mathbf{x}_i \neq \mathbf{x}_i^*|\mathbf{x}_i)$, $\delta(\mathbf{x} - \mathbf{x}^*)$ is the Dirac delta, assigning all the mass probability to the point $\mathbf{x}$ when $\mathbf{x} = \mathbf{x}^*$, and

$$\tilde{\boldsymbol{\mu}} = \frac{(\mathbf{x} + (\alpha - 1)\boldsymbol{\mu})}{\alpha};\ \tilde{\boldsymbol{\Sigma}} = \left(1 - \frac{1}{\alpha}\right)\boldsymbol{\Sigma}$$

Expected Loss - Original Data

In selective editing our target is not $\mathbf{X}$ but $\mathbf{Y}$, and thus

$$E(\mathbf{Y} - \mathbf{Y}^*|\mathbf{y}) = 0 \times \tau_1(\mathbf{y}) + \tau_2(\mathbf{y}) \int_0^\infty (\mathbf{y} - \mathbf{t}) f_{\mathbf{y}^*|\mathbf{y}}(\mathbf{t}) d\mathbf{t}.$$

it follows that, in the original scale, the estimated conditional bias for the jth variable in the ith unit Y_{ij} is:

$$Score(\mathbf{y}_{ij}) = \left| \widehat{\tau}_2(\mathbf{x}_i) \left\{ y_{ij} - exp\left[\left(1 - \frac{1}{\widehat{\alpha}}\right) \left(\frac{1}{2} \widehat{\Sigma}_{jj} + \widehat{\mu}_j \right) + \frac{log(y_{ij})}{\widehat{\alpha}} \right] \right\} \right|.$$

where $\widehat{\tau}_2(\mathbf{x}_i)$, $\widehat{\alpha}$, $\widehat{\Sigma}_{jj}$, $\widehat{\mu}_j$ are MLE.

Selection of Critical Units

In this case the two step algorithm is

1. order the obs with respect to the $Score(y_{ij})$ (decreasing order);

2. find the smallest $\bar{k}$ such that for any $k > \bar{k}$, $\left| \sum_{i=k}^{n} \frac{E(y_{ij} - Y_{ij}^*)}{\widehat{T}_j^*} \right| < \epsilon$

$\widehat{T}_j^*$ is a robust estimate obtained as $\sum_{i=1}^{n} E(Y_{ij}^* | y_{ij})$.

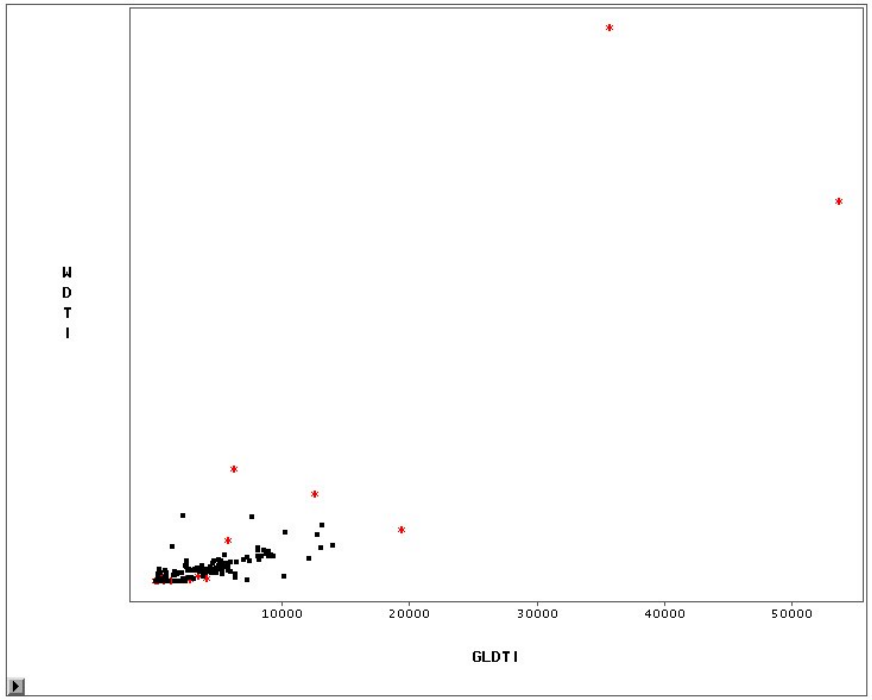
Experiments

1. given a sample s (*cleaned data from an Istat agricultural survey*), the values of *number of worked days (gldti)* and *wages and salaries (wdti)* are log-transformed, and a percentage p of contaminated data is simulated by adding to log-transformed data an error component drawn from $N(0, (\alpha - 1)\mathbf{\Sigma})$, ($\alpha = 2$ and $p = 0.05$);

2. the proposed *score function* and model are applied to these data.

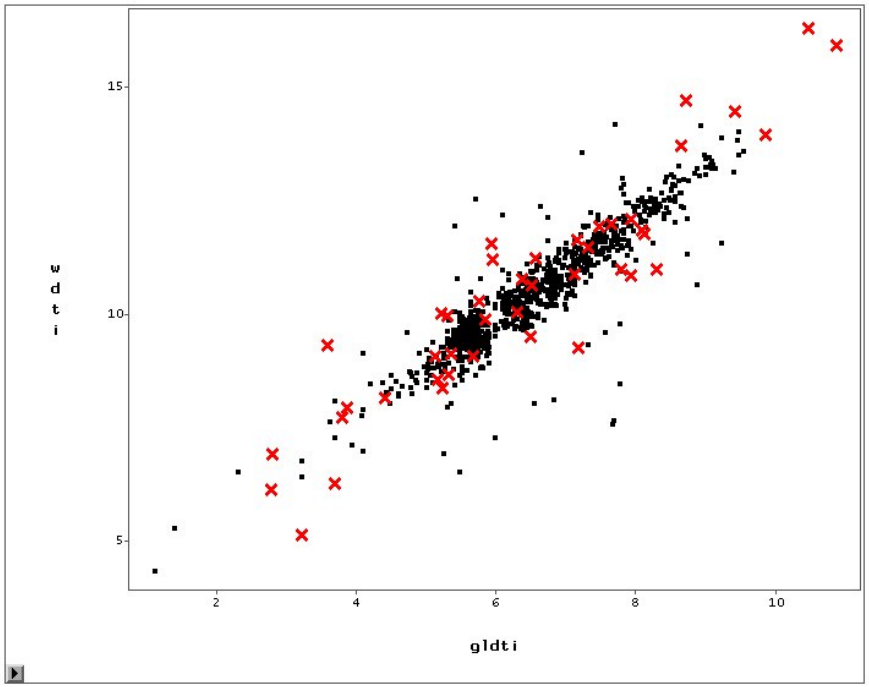
Data

Contaminated data in original scale. Errors are depicted with a red star.



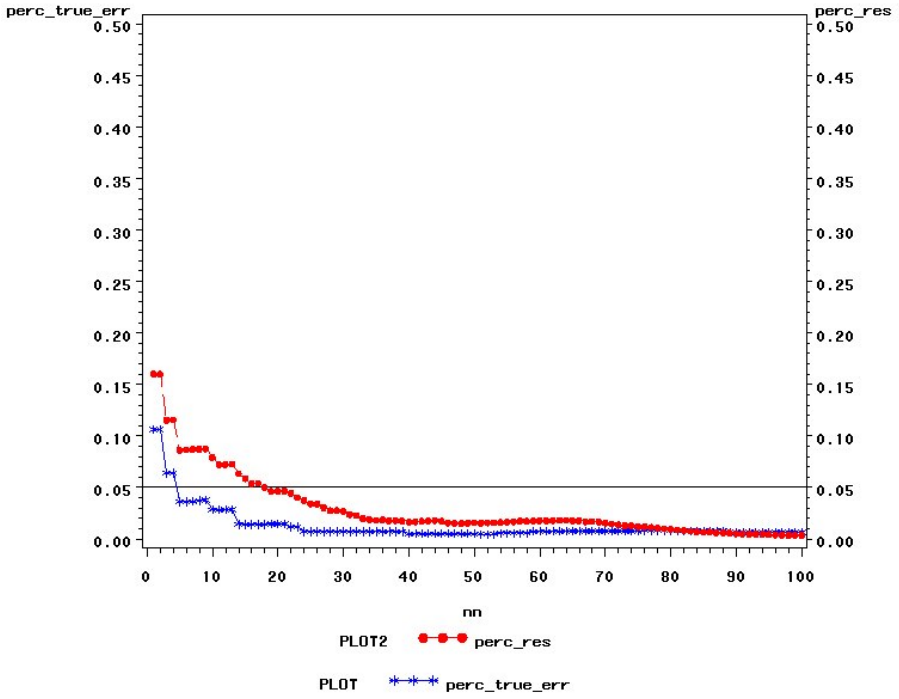
Log-data

Errors are depicted with a red star.



Data

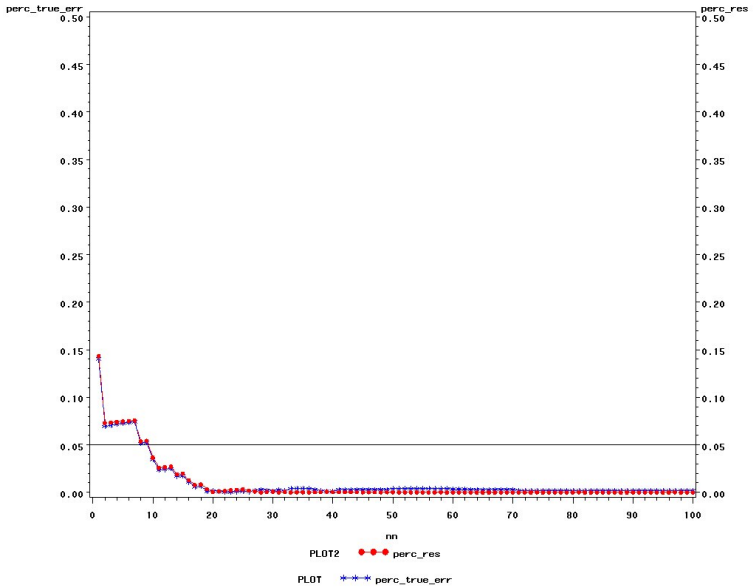
Estimated (perc_res - Red) and true (perc_true_err - Blue) residual error functions for $wdti$ with threshold 0.05.



Normal Data

A similar experiment is carried out on normal data.

Estimated (perc_res - Red) and true (perc_true_err - Blue) residual error functions for log(x) normally distributed and with threshold 0.05.



Conclusions and future works

Summarysing: once chosen the level of reliability of estimates, the proposed procedure coherently treats all the selective editing elements.

Note: observations are considered iid

In this first work, a cross-sectional survey without external information is considered.

When a longitudinal survey is analysed, historical data could be incorporated in the proposed procedure considering the outdated variables as covariates in the model. This aspect will be investigated in the future.

Potentially, through a modification of the EM algorithm, this model can treat the case when historical data are present only for a subset of data.

More experiments with real data are needed.