SeleMix: an R Package for Selective Editing

Ugo Guarnera, Maria Teresa Buglielli

December 12, 2013

1 Introduction

Selective editing is the art of finding influential errors in survey data, i.e., errors having the potential highest impact on the target estimates. In practice, units are prioritized according to a *score function* based on a "risk component" and an "influence component" (Lawrence and McDavitt, 1994; Lawrence and McKenzie, 2000; Latouche and Berthelot, 1992). SeleMix is an R package R Core Team (2012) for selective editing based on expliciltly modeling both true data and error mechanism. True data are modeled through a normal or log-normal distribution, and the " intermittent nature" of the error mechanism is captured through a Bernoullian random variable associated with the error occurrence. Given the event that some values are not correctly reported in a unit, the error is supposed to be additive and Gaussian with zero mean and covariance matrix proportional to the covariance matrix of the true data. 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 so that the expected residual error in data is below a prefixed threshold (Buglielli et al., 2010; Di Zio and Guarnera, 2011, 2013).

2 The contamination model

True data, possibly in log-scale are represented as a $n \times p$ matrix $\mathbf{Y}^*$ of n independent realizations from a random p-vector assumed to follow a normal distribution whose parameters may depend on some set of q covariates not affected by error. The resulting regression model is:

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

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:

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

As a particular case, the set of X-variates may be empty, so that variables Y_i, $(i = 1 \ldots, n)$ are normally distributed with the same mean vector μ. In the previous model, it is We assume that the vector Y_i of observed items for unit i is error-free or erroneous according to a Bernoulli r.v. I_i with parameter π, where $I_i = 1$ if an error occurs and $I_i = 0$ otherwise $(i = 1, \ldots, n)$. Further, given that $I_i = 1$, the error follows 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 $\{I_i = 1\}$:

$$Y_i = Y_i^* + \epsilon_i, \quad \epsilon_i \sim N(0, \Sigma_\epsilon), \;\; \Sigma_\epsilon = (\alpha - 1)\Sigma, \;\; \alpha > 1.$$

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

$$f(y_i|y_i^*) = (1 - \pi)\delta(y_i - y_i^*) + \pi N(y_i; y_i^*, \Sigma_\epsilon). \tag{2}$$

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

It is crucial the intermittent nature of the error implied by the introduction of the Bernoullian variables. 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 Y^* :

$$f(y_i) = (1 - \pi)N(y_i; \mu_i, \Sigma) + \pi N(y_i; \mu_i, \alpha\Sigma), \tag{3}$$

where $\mu_i = \mathbf{B}' x_i$.

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 (Meng and Rubin, 1993).

3 Selective editing

Selective editing is based on comparison between observed values and predictions or 'anticipated values' for true unobserved data. In SeleMix, predictions are obtained from the distribution $f(y_i^*|y_i)$ of the true data conditional on the observed data (possibly including values of error-free covariates X not appearing in the notation). A straightforward application of the Bayes formula provides:

$$f(y_i^*|y_i) = \tau_1(y_i)\delta(y_i^* - y_i) + \tau_2(y_i)N(y_i^*; \tilde{\mu}_i, \tilde{\Sigma}) \tag{4}$$

where:

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

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

$$\begin{aligned} \tau_1(y_i) &= Pr(y_i = y_i^*|y_i) = \frac{(1-\pi)N(y_i;\mu_i,\Sigma)}{(1-\pi)N(y_i;\mu_i,\Sigma) + \pi N(y_i;\mu_i,\alpha\Sigma)}, \\ \tau_2(y_i) &= Pr(y_i \neq y_i^*|y_i) = 1 - \tau_1(y_i), \\ i &= 1,\ldots,n. \end{aligned}$$

Predictions are defined in terms of the conditional expected values $\tilde{y}_i = E(y_i^*|y_i)$. From (4) it follows:

$$\tilde{y}_i = \tau_1(y_i)y_i + \tau_2(y_i)\tilde{\mu}_i, \quad i = 1,\ldots,n. \tag{5}$$

Correspondingly, we can define the "expected error" as

$$y_i - \tilde{y}_i = \tau_2(y_i)(y_i - \tilde{\mu}_i). \tag{6}$$

The last expression makes it natural to interpret τ_2 and $y_i - \tilde{\mu}_i$ as " risk component" and "influence component" respectively to be considered in the score function definition. In practice, the method uses the previous formulas with the MLEs of the involved parameters in place of the true parameters.

The methodology can be easily adapted to the lognormal distribution, which is most frequently used in case of business surveys. In fact, for $i = 1,\ldots,n$, let $Y_i^* = \ln Z_i^*$, $Y_i = \ln Z_i$, where Z_i^*, Z_i represent the variables associated to true and contaminated data respectively. Then, it follows that the distribution of Z_i^* given z_i is:

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

where $LN(\cdot;\mu,\Sigma)$ denotes the lognormal density with parameters μ and Σ.

In the following, the MLE of the model parameters will be denoted by $\hat{\pi}, \hat{B}, \hat{\Sigma}, \hat{\lambda}$, analogously $\hat{\tau}_1(y_i)$ and $\hat{\tilde{\mu}}_i$ will denote the corresponding estimates of the posterior probabilities and of the mean vectors $\tilde{\mu}_i$ respectively.

In SeleMix the score function is defined in terms of the estimated expected error (see formula 6), so that the threshold value can be directly linked to the level of accuracy of the estimates of interest. The units will be selected in a such a way that the estimated residual error is below a prefixed level of accuracy η that is actually the threshold value.

Let us suppose the target aggregate to estimate is the total of the variable Y_j, i.e., $T_j^* = \sum_{i=1}^n w_i y_{ij}^*$. The relative individual error for the ith unit with respect to the variable Y_j is defined as the ratio between the (weighted) expected error and an estimate of the target paprameter $\hat{T}_j$, that is

$$r_{ij} = \frac{w_i(y_{ij} - \hat{y}_{ij})}{\hat{T}_j}. \tag{7}$$

Note that, the estimated expected error is $y_i - \hat{y}_i = \hat{\tau}_2(y_i)(y_i - \hat{\tilde{\mu}}_i)$, and $\hat{\tau}_2$ and $y_i - \hat{\tilde{\mu}}_i$ can be thought of as an estimate of the" risk component" and "influence component" respectively.

The local score function for the variable Y_j used in SeleMix is $S_{ij} = |r_{ij}|$. Different local scores are combined togheter in a single "global score" $GS_i = max_j S_{ij}$.

In order to illustrate the stopping criterion for selections of inlfluential errors, 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 units, that is $R_{ij} = |\sum_{k \geq i}^{n} r_{kj}|$.

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

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

This algorithm ensures that the expected error is below η for all the totals of all the variables Y_j. Moreover, it is easy to show that $S_{kj} \leq 2\eta \quad \forall k > \bar{k}, j = 1, \ldots, J$, so that the expected error on each not revised unit is kept under control.

The reference estimate T_j^* to be used in the score definition can be obtained by using the preditions $\hat{y}_{ij}$:

$$T_j^* = \sum_i w_i \hat{y}_{ij}.$$

This is in fact the default choice in SeleMix.

4 Practical steps in an application of selective editing

The operations required to individuate the influential errors using **SeleMix** can be summarised with these steps:

- analysis of data in order to choose the response variables Y and verify if auxiliary information is available;
- estimation of model parameters;
- identification of critical units corresponding to the most influential errors;
- interactive editing of critical units and automatic editing of npn-critical ones.

4.1 Example Data

These examples refer to the data frame `Labour` contained in the package **Ecdat** (Croissant, 2012).
By typing the following statements in the R environment

```
> library(Ecdat)
> data (Labour)
```

data frame is loaded.

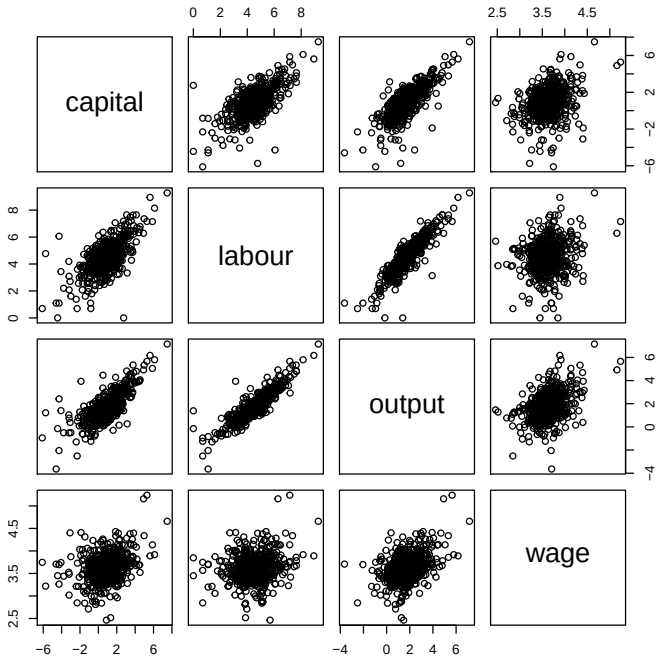
It contains 569

observations on Belgian firms for 1996 (see `Labour` help pages for details). The variables are:

- capital: total fixed assets, end of 1995 (million).
- labor: number of workers (employment).
- output: value added (million).
- wage: wage costs per worker (thousand).

The following Scatterplot shows that log-normality assumption seems to be plausible.

```
> pairs (log(Labour))
```



In the next paragraph, SeleMix is applied on variables `capital` and `output` to detect possible influential errors.

4.2 Two Y variables

As a first example, consider variables `capital` and `output` assuming that both are subject to measurement error. Thus, both variables are considered as Y variables, and what is to be modeled is the their joint distribution. For simplicity, variables `capital` and `output` will be denoted by Y_1 and Y_2 respectively.
The first step is to estimate the parameters of the contamination model using `ml.est` function, that in this case are:

B: the mean vector of the Gaussian distribution of (Y_1, Y_2) (returned in matrix form);

Σ: the covariance matrix;

λ: the variance inflaction factor;

`w`: the mixing proportion of the contamination model (a priori probability of being erroneous).

Input data are obtained by subsetting the two columns of data frame `Labour` corresponding to `capital` and `output`. The input parameters of `ml.est` are set to their default values.

```
> library (SeleMix)                          # Load the library
> y.names<- c("capital", "output" )          # vector of y variables
> est1 <- ml.est(y=Labour[,y.names])         # model estimation
```

The `ml.est` function returns, for each unit, a prediction for each Y variable.

```
> head(est1$ypred)     # predicted values
```

```
  capital.p  output.p
1  2.661581  9.223937
2  1.386256  3.742292
3 20.678225 27.072279
4 10.126059  4.513060
5  1.231076  2.986418
6 10.414401 16.810403
```

Outlier analysis can be performed based on the vector of posterior probabilities:

```
> head(est1$tau,5)       # vector of posterior probability to be contaminated
```

```
[1] 0.02750480 0.02238539 0.09053905 0.12298131 0.02410082
```

units with posterior probability greater than the input parameter `t.outl` are flagged as outliers. Default value is 0.5.

```
> head(est1$outlier)       # vector of flag: 1=outlier
```

```
[1] 0 0 0 0 0 1
```

```
> n.outlier <-sum(est1$outlier)   # numbers of outliers
> n.outlier
```

```
[1] 37
```

Two control parameters for checking the convergence of EM algorithm are also returned. Convergence is not attained if after `max.iter` iterations change in log-likelihood is greater than the input parameter `eps`.

```
> est1$is.conv       # TRUE convergence is reached
```

```
[1] TRUE
```

```
> est1$n.iter        # number of iterations
```

```
[1] 50
```

In order to evaluate goodness of fit, BIC and CAIC are computed both for the contamination model and for the simple Gaussian model ($\lambda = 0$). This parameter should be used to prevent overfitting.

```
> est1$bic.aic       # bic and aic
```

```
BIC.norm  BIC.mix AIC.norm  AIC.mix
3497.839 2329.403 1743.060 1156.498
```

Both the scores show that the contamination model fits data better than the simple Gaussian model.

The second step is to identify influential errors using function `sel.edit`. As a reference estimate to evaluate relative residual error in data after selective editing, the sum of (possibly weighted) predicted values is used (default option). Influential errors are detected by computing differences between observed (`y=Labour[,y.names]`) and predicted (`ypred=est1$ypred`) values. It is possible to take into account unequal sample weigths in the differences and in the reference estimate using parameter `wgt`. The stopping criterion for the units to be selected as influential is determined by parameter `t.sel` (0.02 in the example).

```
>  sel1 <- sel.edit(y=Labour[,y.names], ypred=est1$ypred, t.sel=0.02)
>  (n.sel <-sum(sel1[,"sel"]))   # number of influential observations
```

```
[1] 22
```

```
>  head(sel1,3)                # first lines of result matrix
```

```
    capital    output capital.p  output.p weights capital.score
1  2.606563  9.250759  2.661581  9.223937       1  1.845072e-05
2  1.323237  3.664310  1.386256  3.742292       1  2.113381e-05
3 22.093692 28.781516 20.678225 27.072279       1  4.746874e-04
  output.score global.score capital.reserr output.reserr
1 5.527027e-06 1.845072e-05  -0.0008634218 -0.0006099649
2 1.606899e-05 2.113381e-05  -0.0014026767 -0.0010770438
3 3.522043e-04 4.746874e-04  -0.0084922430 -0.0057397998
  capital.sel output.sel rank sel
1           0          0  469   0
2           0          0  439   0
3           0          0   74   0
```

	non influential	influential errors	Sum
non outliers	525	7	532
n.outl	22	15	37
Sum	547	22	569

Table 1: Outliers vs Influential Errors

```
> sel.pairs (Labour[,y.names], est1$outlier, sel1[,"sel"])
```

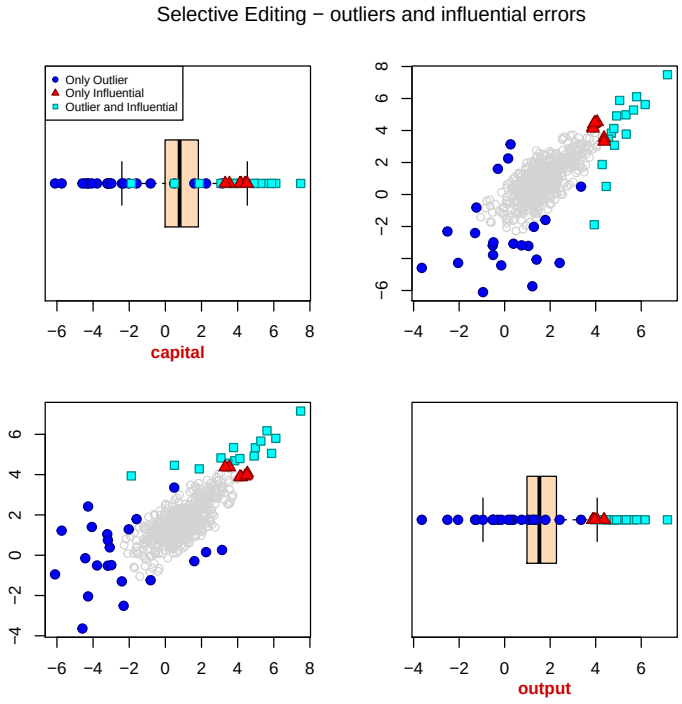


Figure 1: Outliers vs Influential Errors

Table 1 shows the number of outliers units versus the number of units that are flagged as influential.
The scatterplot in Figure 1 shows the same results in a graphical way.

References

Buglielli, M., M. Di Zio, and U. Guarnera (2010). Use of contamination models for selective editing. In *Q2010, European Conference on Quality in Survey Statistics*, Helsinki. http://q2010.stat.fi/sessions/session-19/.

Croissant, Y. (2012). *Ecdat: Data sets for econometrics*. R package version 0.1-6.1.

Di Zio, M. and U. Guarnera (2011). Selemix: an r package for selective editing via contamination models. In *Proceedings of Statistics Canada Symposium 2011*, Ottawa.

Di Zio, M. and U. Guarnera (2013). A contamination model for selective editing. *Journal of Official Statistics 29*(4), 539–555.

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

Lawrence, D. and C. McDavitt (1994). Significance editing in the australian survey of average weekly earnings. *Journal of Official Statistics 10*(4), 437–447.

Lawrence, D. and R. McKenzie (2000). The general application of significance editing. *Journal of Official Statistics 16*(3), 243–253.

Meng, X.-L. and D. B. Rubin (1993). Maximum likelihood estimation via the ecm algorithm: a general framework. *Biometrika 80*, 267–278.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.