Multivariate Outlier Detection and Treatment in Business Surveys

Beat Hulliger
University of Applied Sciences Northwestern Switzerland
School of Business FHNW
4600 Olten, Switzerland
e-mail: beat.hulliger@fhnw.ch

Abstract

Multivariate outlier detection based on the Mahalanobis distance with the BACON-EEM algorithm, the TRC algorithm and the ER algorithm is presented and imputation of outliers and further missing values is discussed. The methods are illustrated with a data set on Swedish municipalities. The relation between outliers, influential observations and selective editing is explored. Finally robust multivariate imputation with survey data is discussed.

Keywords: Sampling, robustness, selective editing, imputation, sensitivity function, winsorization.

1 Introduction

Outliers in business surveys are common due to the structure of the economy: many small enterprises, some medium enterprises, and few large enterprises. This is reflected in the common sample designs for business surveys where large enterprises are often in take-all strata while small enterprises have low sampling fractions. For a single variable which is well correlated with the size of the business, often measured by number of employees, outlier problems are attenuated and tractable with the basic model of a ratio estimator which is robustified against outliers in residuals (See (Gwet and Rivest 1992) and (Hulliger 1999)). However, a coherent approach over all variables of interest must be considered from a multivariate point of view.

When enterprises have to be analyzed more in depth, multivariate outliers become even more important. For example, when the relation between investments into different types of measures against pollution is investigated, the correlation between these investments and the size of the enterprise may become rather weak and the correlation between different types of investment may be negative.

Multivariate robust estimators and multivariate outlier detection (MOD) have been developed in statistics for years now but the problems are still formidable. In public statistical agencies the focus has been traditionally more on detection of multivariate outliers because the separation from treatment permits the use of clerical work for verification of outliers and subsequent imputation. The particular problems of survey data have not been addressed until recently.

Among the first applications, Statistics Canada uses a projection pursuit method for MOD (Franklin, Thomas, and Brodeur 2000). The EUREDIT project of the EU has brought forward several MOD methods which are capable of dealing with data from sample surveys and missing values (EUREDIT 2003), (Charlton 2004). The transfer of these methods into applications is slow but first experiences have been made. Among others the Transformed Rank Correlations algorithm has been tested with data from the Environment Protection Expenditure Survey and from the Survey on Hospitals in Switzerland.

Survey data is often incomplete. An outlier detection method which cannot cope with incomplete data is not appropriate when missing values occur. Few methods have been developed that cope with missing values.

Variance estimation for robust estimators which can cope with incomplete survey data is still in its infancy. Some experience with linearized variance estimators for univariate robust estimators exist (Hulliger 1999). For MOD with Transformed Rank Correlations followed by regression imputation variance estimation by multiple imputation has shown good performance (Hulliger and Münnich 2006).

2 Multivariate Outlier Detection

The majority of methods for multivariate outlier detection is based on the Mahalanobis distance with robust estimates for the mean and covariance. The combination of the EM-algorithm with a robust estimation in the maximisation step seems to be the first method which copes with missing data (Little and Smith 1987). The authors called their algorithm ER-algorithm. The robust estimator starts from a non-robust mean and covariance estimator and adds one weighting step, i.e. does the first iteration which would lead to an M-estimator. However, the non-robust start implies that the resulting estimator is not robust. Furthermore the breakdown point of multivariate M-estimators is relatively low. This may be the reason that other researchers have used Minimum Covariance Determinant estimation in the maximisation step of the EM-algorithm (Cheng and Victoria-Feser 2000). Their algorithm seems computationally so heavy that it was not applied in practice.

The projection pursuit method of Statistics Canada is built on a robust version of principal component analysis. A limitation of this approach is that for the projection step a full data matrix is needed and therefore it seems difficult to adapt the method for missing values.

The BACON-EEM algorithm is based on the forward search algorithm BACON (Billor, Hadi, and Vellemann 2000). BACON starts with an outlier-free sub-sample and adds good points as long as possible. To cope with missing values

in survey samples the EM-algorithm was extended to the EEM-algorithm, which works with the estimate of the sufficient statistics in the expectation step. The combination, called BACON-EEM is robust since the outer loop of the BACON algorithm protects the non-robust EEM inner loop. The BACON-EEM is comparably fast, tolerates a considerable amount of missingness and, at least empirically, showed a remarkably high breakdown point (Béguin and Hulliger 2003).

Transformed Rank Correlations (TRC) is a non-iterative algorithm which is based on the bivariate Spearman Rank Correlations that are assembled into a preliminary covariance matrix (Béguin and Hulliger 2004). To ensure positive definiteness an orthogonal transformation into the eigen-space and reestimation of the center and covariance by the median and median absolute deviation is used. To cope with missing values a simple robust regression imputation with the best available regressor yields an ad hoc imputation.

Imputation under the multivariate normal model based on the non-robust mean and covariance calculated with the original EM-algorithm, followed by the application of the Minimum Covariance Determinant method (Rousseeuw and van Driessen 1999) without consideration of the sampling weights was used as a comparison base line. We call this algorithm GIMCD (Gaussian Imputation with Minimum Covariance Determinant).

A few algorithms have been adapted to survey data which are not based on the Mahalanobis distance. One is a multivariate version of a weighted robust tree algorithm, called WAID (Chambers, Hentges, and Xinqiang 2004). Outliers can be detected because they obtain a low overall weight over the tree-nodes in which they are located. Another non-Mahalanobis distance algorithm is based on an epidemic in a point cloud where the epidemic infects the outliers late (Béguin and Hulliger 2004).

2.1 Data and results for MOD

The MU284 data set from (Särndal, Swensson, and Wretman 1992) contains data about Swedish municipalities. We use the variables *population in 1975* and *population in 1985* (P75 and P85), *revenue from municipal taxes 1985* (RMT85), *number of municipal employees 1984* (ME84) and *real estate value 1984* (REV84). We assume the data is a stratified sample where municipalities with P75<20 have been sampled with rate 10% while the larger municipalities form a take all stratum. A preliminary analysis shows that a good model can be seen for the per capita figures, i.e. we use RMT85/P85, ME84/P85 and REV84/P85 and denote these per capita variables by lower case letters rmt85, me84 and rev84. A second look at the data suggests a logarithmic transformation for the auxiliary variable P75 and for rev84. We will call these log-transformed variables lp75 and lre84 respectively.

The three largest cities of Sweden are outliers compared with the other municipalities in any sense. We exclude them from the data set in order to see finer details. The resulting data set is called MU281.

We include the auxiliary variable lp75 when doing outlier detection and treatment. Of course we raise the dimensionality of the problem by doing so. But population size is the variable on which the design is built and it is an important explanatory variable even after the log transformation.

To determine representative outliers the data set without missing values is augmented by adding 9 replicates of the stratum of smaller enterprises. Then a MCD algorithm is run on this artificial population and the 10% of the artificial population with largest Mahalanobis distance are declared representative outliers. In the original MU281 sample there are 34 representative outliers.

Missingness at random is created with probabilities that are decreasing for increasing P75. The probabilities vary according to a logistic model. For each value an independent bernoulli trial with the missingness rate of the observation is carried out to determine whether a value is set to missing.

Non-representative outliers are determined with a probability depending on P75 again. For P75<10 the non-representative outlier rate is set to $p_{nr,1} = 0.10$ and for P75$\geq$ 10 to $p_{nr,2} = 0.20$ (low amount of non-representative outliers) or to $p_{nr,1} = 0.25$ and $p_{nr,2} = 0.35$ or $p_{nr,1} = 0.3$ and $p_{nr,2} = 0.4$ respectively (middle and high amount of non-representative outliers). Representative outliers which are determined non-representative outliers, too, remain representative outliers.

The outlier values are set by a simple linear regression model $y' = a + by$ with $a = 8, 50, 2.4$ and $b = 0.2, 0.1, 0.2$ for rmt85, me84 and lre84 respectively. This is a contamination which is a concentrated point cloud close to the bulk of the data.

With 24.8% missingness rate the parameter for the BACON-EEM algorithm must be set to $\alpha = 0.001$ or smaller if no non-representative outliers are present. Otherwise the detection of representative outliers is not good (9 of the 34 detected with $\alpha = 0.01$.) On the other hand for missingness rate 10.7% $\alpha = 0.001$ yields only 18 detected outliers compared with 24 when $\alpha = 0.01$. Thus the tuning of BACON-EEM is relatively important.

Table (1) shows the number of representative and non-representative outliers detected by the methods ER, BACON-EEM, TRC and GIMCD. The number of outliers is equal for the methods (sum of number of representative plus non-representative outliers). The best method for each situation in terms of number of outliers detected is bold-faced. There is no method best in all situations. ER seems to cope relatively well with the representative outliers but not with the non-representative outliers. GIMCD is relatively good for low amount of non-representative outliers. TRC is remarkably good in many situations. BEM is usually close to TRC and better for high missingness rate and middle amount of non-representative outliers. In the most difficult situation in the last line with high missingness rate and high amount of non-representative outliers all methods have problems, the best method, TRC, detecting a moderate 55% of the outliers.

3 Influential Observations and Outliers

In a business survey we would like to know and control the impact of an observation on the results of the survey. The main

Table 1: Number of (rep, non-rep) outliers detected

miss-rate	n. of non-reps	ER	BEM	TRC	GIMCD
10.7	0	18	24	**27**	20
10.7	51: low	22,21	19,47	21,48	**20, 51**
24.8	0	**19**	17	13	14
24.8	51: low	22,20	**23,45**	**20,48**	16,48
30.1	0	21	20	**27**	16
30.1	51: low	21,21	16,45	32,12	**17,47**
30.1	84: middle	24,32	**24,61**	33,32	28,15
30.1	98: high	27,31	27,31	**33,39**	26,27

For 0 non-rep. outliers only representative outliers can be detected.

results of the survey are estimators or, more generally, statistics. However, there are many possible statistics, including variance estimators, ratios, correlation coefficients, quantiles, test statistics and derived measures like p-values and confidence intervals. There are so many possible statistics that we cannot check for every statistic what the impact of every observation is. This is the reason why we try to solve the seemingly simpler problem to check whether an observation is an outlier. An observation will be declared as an outlier when it is compared with a model. The implicit trick is to assume that whatever observation is compatible with the model will not have a high influence for any statistic. Under this assumption it is simpler to detect and treat outliers than to check influential observations for every statistic. However, the outlier-model reflects a choice of statistics and it may well be that for statistics which are not reflected well in the outlier-model there are influential observations which are not detected. This is particularly the case when the outlier-model is defined in transformed variables, e.g. log-transformed as with lrev84 in the Swedish Municipality Data MU281. Statistics based on the untransformed variables may not be protected from influential observations then. Therefore we advocate to look for influential observations in addition to detection and treatment of outliers. In the following we discuss a simple measure of influence.

The influence function (Hampel 1974) focusses on a statistic $T(F)$ as a functional at the distribution F and its behavior under an infinitesimal change in F. The statistic may be multi-parameter and the distribution may be multivariate, i.e. $T \in R^d$ and $Y \sim F, Y \in R^p$. For finite populations we may look at the population distribution function F_U which is estimated by the (sampling weighted) empirical distribution function F_S. A sampling analogue of the empirical influence function is the sensitivity curve

$$SC(x; T, y_S, i) = n\left(T(y_{S\setminus i}, x) - T(y_{S\setminus i})\right). \qquad (1)$$

Here $T(y_{S\setminus i}, x)$ is the statistic T evaluated at the full sample when the value y_i of observation i is replaced by the function argument x. The statistic T evaluated at the sample when observation i is considered a missing unit is $T(y_{S\setminus i})$. Of course, in practice the definition and calculation of $T(y_{S\setminus i})$ may be far from simple, e.g. involving calibration or response propensity modeling. We may approximate $T(y_{S\setminus i})$ by a simpler expression involving re-weighting but some care is needed to ensure that the influence is still described correctly.

For finite population sampling the sensitivity curve depends not only on the value of x but also on the particular observation i. For the Horvitz-Thompson estimator this dependence from the sample is expressed with the sampling weight (see below). However, for other estimators the dependence on i may be more complicated, for example when we deal with a rotational panel and each unit has its own inclusion history.

The sensitivity curve of the Horvitz-Thompson estimator $T = \sum_{i\in S} w_i y_i$ with the inverse of the inclusion probabilities as weight is

$$\begin{aligned} & SC(x; HT, y_S, i) \\ &= n\left(\sum_{S\setminus i} w_j y_j + w_i x - \frac{\sum_S w_j}{\sum_{S\setminus i} w_j}\sum_{S\setminus i} w_j y_j\right) \\ &= n w_i (x - \hat{y}_i), \end{aligned} \qquad (2)$$

where $\hat{y}_i = \frac{\sum_{S\setminus i} w_j y_j}{\sum_{S\setminus i} w_j}$ is the Hajek-estimator of y_i based on the reduced sample $S \setminus i$. In other words $\hat{y}_i$ is an imputed mean.

The basic expression for the sample sensitivity curve (1) will be different for other estimators and it may be much more complicated. For example the Yates-Grundy-Sen variance estimator will involve the double inclusion probabilities. When multivariate characteristics are involved like the correlation coefficient the sensitivity curve will be a function of all the variables. It may be difficult to find a closed form expression for the sensitivity curve of non-linear statistic like the Spearman rank-correlation though its value may be quite simple to calculate numerically.

We call the value of $SC(y_i; T, y_S, i)$ the impact of observation i on a statistic T. Of course the impact is a "leave-one-out" score like Cook's distance in regression.

Outlier measures and impact measures often correlate highly. This is the reason why outlier detection and treatment is used as a substitute for limiting the influence of observations on particular statistics. However, outlier detection and treatment is no guarantee for limiting the impact on all possible statistics!

The impact of the observations in MU281 on the Horvitz-Thompson estimator for the total of variable rev84 of MU281 is shown in Figure 1. The weights of the two strata determine

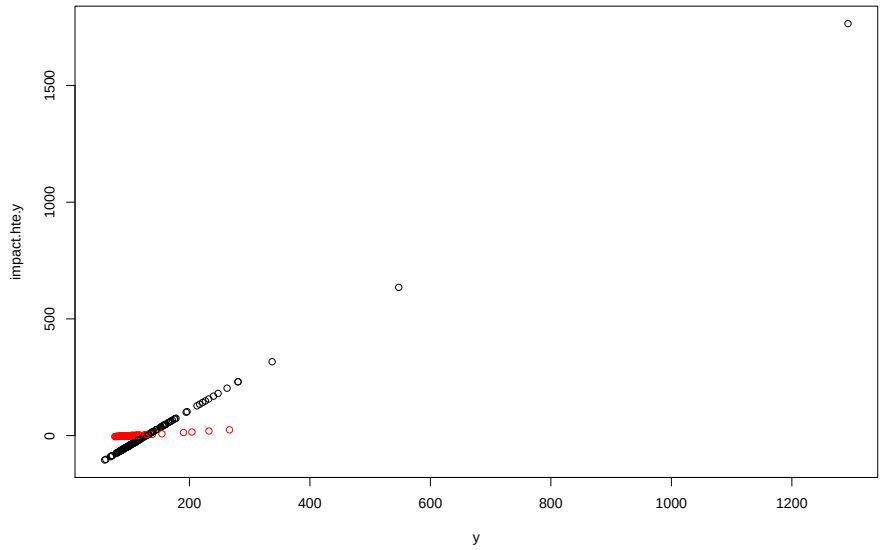

Figure 1: MU281: Impact on Horvitz-Thompson estimator for rev84

the slope of the underlying sensitivity curves, which is linear in our case.

A different picture can be seen for the impact on a robust version of the Horvitz-Thompson estimator (see Figure 2). The robustification is based on a univariate robustness weight based on the median and median absolute deviation of a specific variable. The impact of the extreme observations is bounded above and below. However, if the model to determine the outliers is fitted on a logarithmic scale (lre84) and robustness weights are derived accordingly, the final (robustly) weighted mean has impacts as shown in Figure 3. Obviously the impact is no more bounded though it is much smaller than for the non-robust Horvitz-Thompson estimator (Figure 1). This shows that influential observations and outliers are not the same because outliers depend on a model, while influential observations depend on a statistic.

4 Selective Editing

The impact function (2) at $x = y_i$ of the Horvitz-Thompson estimator contains the sample based estimate $\hat{y}_i$. If $\hat{y}_i$ can be replaced by a value derived from previous surveys or from other external data, with $\tilde{y}_i$ say, then the sensitivity curve and the impact can be calculated individually. In other words when replacing $\hat{y}_i$ with $\tilde{y}_i$ the impact becomes a score function for selective editing (Lawrence and McKenzie 2000) and can be applied in micro-editing. Of course the score functions of selective editing have been developed with several criteria in mind (Latouche and Berthelot 1992) and consider combinations of score functions or impacts. Nevertheless, a basic ingredient of the score functions of selective editing often are approximations to the impact of the Horvitz-Thompson estimator or ratios and linear combinations of it. The strength of selective editing is that it can be applied at the micro level, i.e. for each observation independently. We do not need the complete sample to calculate the score function. The sensitivity score of a statistic is clearly a macro level approach since we need the complete sample to calculate it.

An obvious drawback of selective editing is that the score function reflects a particular choice of one or a few statistics. This may be the most important statistic like the Horvitz-Thompson estimator. Nevertheless, selective editing with a particular score function will not prevent a high impact on other statistics than the ones represented in the score function.

In practice the theoretical goal of knowing and controlling the impact of any observation on any possible statistics is not attainable. The way to proceed is a good and affordable combination of the three actions

- Selective editing with score function detection.
- Outlier detection and treatment (univariate, multivariate, with appropriate models)
- Investigating the impact scores of important statistics.

Note that the latter two actions are macro-editing and imputation tasks while selective editing is a micro-editing task.

5 Robust multivariate imputation

The multivariate outlier detection methods that are based on a robust covariance matrix and the Mahalanobis distance lead to a direct model based imputation, using the same robust covariance matrix.

Observations are declared outliers if their Mahalanobis distance is larger than a constant c. The imputation for an observation x_i with Mahalanobis distance $d_i > c$ is

$$\hat{x}_i = m + (x_i - m)\frac{c}{d_i}. \tag{3}$$

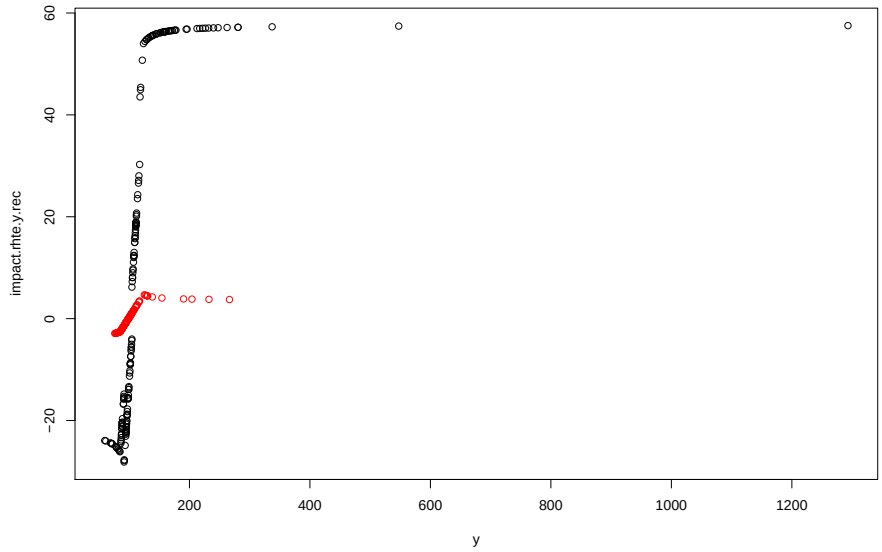

Figure 2: MU281: Impact on Horvitz-Thompson estimator for rev84, robustified on rev84

The squared Mahalanobis distance of the winsorized value $\hat{x}_i$ to the robust mean m using a robust covariance C is

$$\begin{aligned} \hat{d}_i^2 &= (\hat{x}_i - m)^\top C^{-1} (\hat{x}_i - m) \\ &= \frac{c}{d_i}(x_i - m)^\top C^{-1}(x_i - m)\frac{c}{d_i} \\ &= c^2. \end{aligned} \tag{4}$$

Thus the above imputation correponds to winsorizing the Mahalanobis distance of the vector $x_i - m$ to c while leaving its direction unchanged.

If there are missing values in the outliers then the observed variables of the outlier may be winsorized in the same way as above but with the mean and covariance only referring to the sub-space of observed values. Note that when calculating the Mahalanobis distance with missing values a factor p/q is applied to compensate at least partially for the number of missing dimensions $p - q$.

Once the outliers are winsorized the missing values may be imputed under the multivariate normal model with mean m and covariance C. A missing value is imputed by a fitted value under a regression model with all variables with present values as predictors. We may, of course, add a random error term to this imputation. It seems more natural to winsorize before the imputation to avoid the imputation of outliers. If the same robust mean and covariance are used for imputation and winsorization the imputation may be carried out first. The two versions may lead to different results because the Mahalanobis distance with missing values usually differs from the Mahalanobis distance with imputed values.

A nearest neighbor algorithm based on the Mahalanobis distance may be applied whether there are missing values or not. The outliers should be excluded from being donors to prevent outlier imputation. The robust covariance matrix and mean of the outlier detection phase may be used directly. Alternatively the mean and covariance may be re-estimated using only the robustness weights of the preceding outlier detection. This may be useful when further editing or even call-backs to respondents have clarified outliers or when the outlier threshold should be raised to change less observations than suggested by the outlier detection.

Re-estimation of m and C has been used in the EUREDIT project for the POEM algorithm (EUREDIT 2003). However, it turned out that estimating a positive definite covariance matrix is problematic. A nearest neighbor algorithm implemented at Swiss Federal Statistical Office uses the robust covariance matrix of the TRC algorithm for outlier detection.

To show the effect of imputation when outliers are present estimations of means and correlations for the MU281 data are presented in Table 2 and 3. Missingness was introduced at random (See Section 2). The TRC algorithm was used for outlier detection and imputation. For the three variables with missing values and outliers, rmt85, me84 and lre84, the means after imputation lie closer to the means of the complete and winsorised data than the raw and raw winsorised data. For the imputed data the correlation of rmt85 with me84 is higher than for the raw winsorised data and differs more from it than for the raw and raw winsorised data. The correlation between rmt85 and lre84 with imputed data is roughly half the correlation with raw winsorised data. Winsorisation and imputation move the correlation in the same direction. The correlation between me84 and lre84 is reduced to nearly 0 by the winsorisation of the complete data. The correlation of the raw winsorised and the imputed data follow this move to zero and even a slightly negative value.

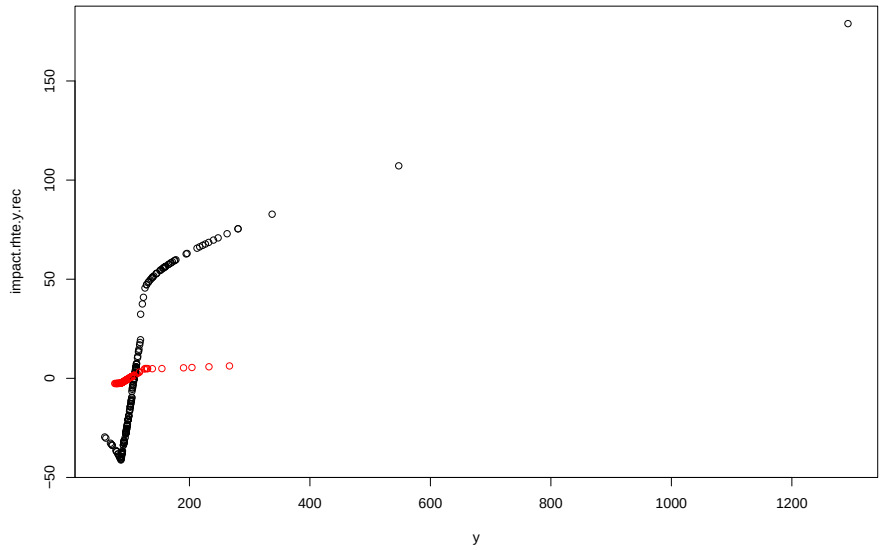

Figure 3: MU281: Impact on Horvitz-Thompson estimator for rev84, robustified on lrev84

Table 2: MU281 data: Estimation of population mean

data	rmt85	me84	lre84
complete	6.92	49.86	2.061
complete winsorised	6.90	49.53	2.044
raw	6.94	49.97	2.062
raw winsorised	6.93	49.80	2.049
imputed	6.91	49.70	2.047

Table 3: MU281 data: Estimation of correlations

data	rmt85,me84	rmt85,lre84	me84,lre84
complete	0.630	0.151	0.182
compl. wins.	0.624	0.159	0.005
raw	0.625	0.120	0.130
raw wins.	0.627	0.098	0.022
imputed	0.671	0.083	-0.036

6 Conclusions

BACON-EEM and TRC are multivariate outlier detection algorithms which can cope with incomplete survey data. Gaussian Imputation followed by Minimum Covariance Determinant outlier detection performs remarkably well and deserve more research. There is a limit of missingness and outlyingness where the methods cannot cope anymore.

The scores of selective editing often are particular instances of impacts: Selective editing cannot protect all possible statistics. Outliers and influential observations do not necessarily coincide, in particular not, when the model involves transformations. It is necessary to check the impacts on the statistic of interest during macro-editing, even if selective editing was applied in micro-editing and outlier detection in macro-editing.

More experience is necessary with multivariate imputation in the presence of outliers. Preliminary results show that the behavior is not always as expected. In addition, variance estimation after outlier detection and imputation must be investigated further.

References

Béguin, C. and B. Hulliger (2003). Robust multivariate outlier detection and imputation with incomplete survey data. Deliverable D4/5.2.1/2 Part C, EUREDIT.

Béguin, C. and B. Hulliger (2004). Multivariate oulier detection in incomplete survey data: The epidemic algorithm and transformed rank correlations. *J.R.Statist.Soc.A 167*(Part 2.), 275–294.

Billor, N., A. S. Hadi, and P. F. Vellemann (2000). BACON: Blocked Adaptative Computationally-efficient Outlier Nominators. *Computational Statistics and Data Analysis 34*(3), 279–298.

Chambers, R., A. Hentges, and Z. Xinqiang (2004). Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society, Series A: Statistics in Society 167*(2), 323–339.

Charlton, J. (2004). Editorial: Evaluating automatic edit and imputation methods, and the euredit project. *Journal of the Royal Statistical Society: Series A Volume 167, Issue 2*, 199–207.

Cheng, T.-C. and M.-P. Victoria-Feser (2000, June). Robust correlation estimation with missing data. Technical Report 2000.05, Université de Genève.

EUREDIT (2003). *Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project*, Volume 1 and 2. EUREDIT consortium. http://www.cs.york.ac.uk/euredit/results/results.html.

Franklin, S., S. Thomas, and M. Brodeur (2000). Robust multivariate outlier detection using Mahalanobis' distance and a modified Stahel-Donoho estimator. Technical report, Statistics Canada.

Gwet, J.-P. and L.-P. Rivest (1992). Outlier resistant alternatives to the ratio estimator. *J. Amer. Statist. Assoc. 87*(420), 1174–1182.

Hampel, F. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association 69*, 383–393.

Hulliger, B. (1999). Simple and robust estimators for sampling. In *Proceedings of the Section on Survey Research Methods*, pp. 54–63. American Statistical Association.

Hulliger, B. and R. Münnich (2006). Variance estimation for complex surveys in the presence of outliers. In *Proceedings of the Survey Research Methods Section*. American Statistical Association.

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

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

Little, R. and P. Smith (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association 82*, 58–68.

Rousseeuw, P. and K. van Driessen (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41*, 212–223.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer.