Robust Outlier Detection via Forward Search: Application to the ABI Dataset

EUREDIT Workpackage 4.2

EUREDIT Deliverable D4.2.1

Adão L. Hentges
Department of Social Statistics
University of Southampton
Highfield, Southampton, SO17 1BJ, U.K.

1. Introduction

We take a sample of size n from a finite population of size N where $\mathbf{y}$ are observed values of the multivariate variable $\mathbf{Y} = (Y_1, \ldots, Y_p)$ and $\mathbf{x}$ values of the multivariate covariate $\mathbf{X} = (X_1, \ldots, X_q)$. The covariate $\mathbf{X}$ will have a mix of categorical and continuous variables, for which we have further information about the mean of their components or the values of non sampled elements in the population. In addition, the values $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are supposed to be correctly measured. It will be assumed that the method of sampling is ignorable given $\mathbf{X}$, and that there is complete response.

Our goal is to identify possible unit outliers in the observed vector $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$, where each i–th component $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})$ has p observed values for Y components. A brief overview of the literature shows lots of different methods in order to identify outliers and influential points in a data set. Generally the identification has to be carried out relative to some assumed model for the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ in the sample.

Using the covariates available to specify a regression structure the standard linear model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ will be considered, where β is a q vector of unknown parameters and ϵ is a n vector of random errors. In order to locate possible outliers in the vector $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ the residuals are evaluated by fitting that model to a multivariate data set. Large residuals from the estimation of $\mathbf{Y}$ given $\mathbf{X}$ will be evidence that the unit where such deviation occur is a potential outlier.

A procedure for detection of multiple outliers in a sample may be insensitive when some suspected units form subgroups, creating the *masking* effect (Barnett and Lewis, 1994). Robust methods are then necessary in order to locate the true outliers.

We concentrate attention on the robust forward search approach (Hadi and Simonoff, 1993; Atkinson, 1994; Riani and Atkinson, 2000). The algorithm will start from a subset of observations intended to be outlier free. This subset is augmented at each step using the units which best fit the regression model based on the clean data. The search can stop when some significant outlier has joined the basic clean subset or it may be carried out up to the full sample size n and the behaviour of the residuals analyzed in order to detect suspicious units.

A real data set provided by the Office of National Statistics (ONS - United Kingdom) will be used for practical applications. In this study the detection of outliers will be initially explored on individual components of $\mathbf{Y}$ and later for the multivariate case ($p > 1$).

2. The forward search method

When masked multiple outliers are present in the data generally it is difficult to locate the true outliers. A single search by fitting a model to the full sample (a one stage search) may not reveal all the true outlying units. Starting with the full sample and removing sequentially all the suspected units until no more outliers are present in the data may be an appealing method. However, it is an expensive algorithm since the swamping problem may also be present in the sample, affecting discordancy test for blocks of two or more suspected units. A different option is the forward search method, which seems to overcome those problems.

We follow the approach described by Hadi and Simonoff(1993) and similarly by Riani and Atkinson(2000). The basic idea is to start with a relatively clean data set of size m and include observations until only outlying observations remain out.
Let

$$C_{(m)} \;=\; \{(\mathbf{y}_i, \mathbf{x}_i),\; i = 1, \ldots, m \;(m < n)\} \tag{1}$$

be the initial clean data, supposedly outlier–free, and $(\mathbf{y}, \mathbf{x})$ the sample values of the response multivariate variable $\mathbf{Y}_p$ and the covariate vector $\mathbf{X}_q$.

This starting subset of data may be defined in different ways. Hadi and Simonoff (1993) suggests two procedures: by fitting a regression model to the full data and then ordering the n observations by an appropriate diagnostic measure; or constructing a single linkage clustering tree and ordering the clusters from most to least extreme by the order of joining. Riani and Atkinson (2000) perform a robust analysis of the matrix of bivariate scatterplots and take as the initial subset those observations that are not outlying on any scatterplot.

The forward search then moves from m observations to $m+1$ by choosing the $m+1$ observations with the smallest residuals from the fit on data of $C_{(m)}$. Standardized residuals from the estimate $\hat{\beta}_{(m)}$ for the linear regression model $E(\mathbf{Y}) = \mathbf{X}\beta$ are computed by Hadi and Simonoff (1993) in the univariate case ($p = 1$) as

$$d_i \;=\; \begin{cases} \dfrac{y_i \;-\; \mathbf{x}_i^T \;\hat{\beta}_{(m)}}{\hat{\sigma}_{(m)} \;\sqrt{1 \;-\; \mathbf{x}_i^T (\mathbf{X}_{(m)}^T \mathbf{X}_{(m)})^{-1} \;\mathbf{x}_i}} \;, & \text{if} \;\; i \in C_{(m)} \\ \dfrac{y_i \;-\; \mathbf{x}_i^T \;\hat{\beta}_{(m)}}{\hat{\sigma}_{(m)} \;\sqrt{1 \;+\; \mathbf{x}_i^T (\mathbf{X}_{(m)}^T \mathbf{X}_{(m)})^{-1} \;\mathbf{x}_i}} \;, & \text{if } i \notin C_{(m)}, \end{cases} \tag{2}$$

where $\hat{\beta}_{(m)}$ are the estimated regression coefficients computed from fitting the linear model to $C_{(m)}$ and

$$\hat{\sigma}^2_{(m)} \;=\; \frac{\sum_{i=1}^{m} \;(y_i - \mathbf{x}_i^T \;\hat{\beta}_{(m)})^2}{m - q} \tag{3}$$

the corresponding residual mean square. When $i \in C_{(m)}$, d_i is then the internally studentized residual and when $i \notin C_{(m)}$, d_i is the scaled prediction error based on the subset $C_{(m)}$. Atkinson (1994) uses almost similar residuals, except that for $i \in C_{(m)}$, d_i is defined as the least squares residuals but in their comparison no evidence was found in favor of one type.

For $p \geq 1$, Riani and Atkinson (2000) uses the squared Mahalanobis distances

$$d_i^2 \;=\; \{(\mathbf{y}_i - \hat{\mathbf{y}}_{i(m)})^T \;\hat{S}^{-1}_{(m)} \;(\mathbf{y}_i - \hat{\mathbf{y}}_{i(m)})\}, \quad i = 1, \ldots, n, \} \tag{4}$$

to order observations for the forward search where $\hat{\mathbf{y}}_{i(m)}$ and $\hat{S}_{(m)}$ are obtained from regression of $\mathbf{Y}$ on $\mathbf{X}$ based on the m observations from the basic clean data $C_{(m)}$. Hadi (1994) creates a multi dimensional clean data by ordering the n observations according to robust measures, using

$$D_i(L_R, S_R) \;=\; \sqrt{\{(\mathbf{y}_i - L_R)^T \; S_R^{-1} \; (\mathbf{y}_i - L_R)\}}, \quad i = 1, \ldots, n, \tag{5}$$

where L_R and S_R are robust location and covariance matrix estimators from the fit in the full sample. The observations are rearranged in ascending order according to similar measure based on $L_{(m)}$ and $S_{(m)}$, the mean and covariance matrix of the basic subset. At the next step the clean data increases its size to $m+1$ using the n distances obtained from $C_{(m)}$.

A stopping criterion is used for Hadi and Simonoff (1993) in the forward search. For $p = 1$, $d_{(s+1)}$ is defined as the $(s+1)$–th order statistic of the n absolute residuals $|d_i|$, where s is the size of the current subset $C_{(m)}$. If

$$d_{(s+1)} \geq t_{(\alpha/2(s+1),s-q)} \tag{6}$$

then all observations satisfying $|d_i| > t_{(\alpha/2(s+1),s-q)}$ are declared outliers and the forward search finishes.

For $p \geq 1$, similarly, Hadi (1994) orders the n evaluated squared Mahalanobis measures D_i^2 and defines $D_{(s+1)}^2$ as the $(s+1)$–th order statistic of the D_i^2. In a regression model, residuals from fitting of $\mathbf{Y}$ on $\mathbf{X}$ are used to evaluate D_i^2. The multivariate search stops if

$$D_{(s+1)}^2 \geq \chi_{(p,\alpha/n)}^2, \tag{7}$$

and then all observations with $D_i^2 \geq \chi_{(p,\alpha/n)}^2$ are identified as outliers. If the basic data set increases to $C_{(m)} = C_{(n)}$, without the stopping criterion being met, then the data set is declared outlier free.

Atkinson (1994) and Riani and Atkinson (2000) perform forward searches but without a stopping rule. The emphasis there is analyzing plots of the residuals obtained from a full search, starting from the clean data and increasing up to the full sample size.
At each particular stage $m, m+1, m+2, \ldots, n$ each observation $\mathbf{y}_i$ is tested if it is an outlier according to the Mahalanobis distances from (4). The cuttof value used is the maximum expected value from a sample of n chi–squared random variables on p degrees of freedom, approximated by

$$E(max\ \chi_p^2) \ = \ \chi_p^2\ \{(n-0.5)/n\}. \tag{8}$$

Having then performed $n-m$ steps in the algorithm it is possible to analyze the behaviour of the sequence of the n residuals. Units with a clear outlying pattern could be detected through the analysis of those residuals, with graphical plots being a powerful aid.

When the full search ahead is performed, the units which have been identified as outliers in most of the steps can have a close examination. Empirically we define a set of outliers by taking the observations which are not on the current clean data when the relative "jump" on the residual variance on the fit on $C_{(m)}$ is maximum. Let

$$\tau_j \ = \ \frac{det(S_{(j)}) \ - det(S_{(j-1)})}{det(S_{(j-1)})}, \quad j=2,\ldots,n, \tag{9}$$

where $S_{(m)} \ = \ (m-q)^{-1}\ \sum_{i=1}^{m}\ (\mathbf{y}_i-\hat{\mathbf{y}}_i)^T\ (\mathbf{y}_i-\hat{\mathbf{y}}_i)$ is the estimated residual covariance matrix based on the clean data with current size m and $det(S)$ its determinant. Since the search is based in Mahalanobis distances in ascending order, an important outlier joining the clean data at some stage should cause a breakdown for $S_{(m)}$. At some step j where τ_j is maximum we declare the unit joining the clean data and all those not included yet as outliers. The distribution of τ_j is not available and is related to the sequence $S_{(m)}, S_{(m+1)}, \ldots, S_{(n)}$ with dependent components, since generally units included in the clean data at step m should be present at step $m+1$ too.

Also, considering the number of times each sample unit was declared outlier in the whole search, we apply a binomial test to define a set of outlying units. For example, suppose π_i is the true probability that sample unit i is an outlier in the population. Let

$$\delta_i \ = \ \sum_{k=1}^{n-m}\ I_{ik}$$

be the number of times the unit was identified as outlier based on the $n-m$ steps performed, where I_{ik} is equal 1 when residual d_i is outlying on the k–th step of the search and 0 otherwise.

Assume now that $\delta_i \sim B(n-m, \pi_i)$, at least approximately since the I_{ik} are not independent. Defining $\hat{p}_i = \delta_i/(n-m)$ we then declare unit i as a true outlier (by specifying $\pi_i = 1$) if

$$\frac{\sqrt{(n-m)-1}\,(1-\hat{p}_i)}{\sqrt{\hat{p}_i\,(1-\hat{p}_i)}} \;<\; c_{(\alpha)}, \tag{10}$$

where $c_{(\alpha)}$ is the cutoff given by the asymptotical normal $N(0,1)$ distribution.

These two empirical procedures will be performed just to have some comparison between the outliers defined by the precise stopping rule from Hadi, in a way to use results from the full search and check if the outlying sets agree.

3. Example for MOD

As an illustration we apply the MOD approach to a perturbed data set provided by the Office for National Statitics (ONS) both in the univariate and also the multivariate case ($p > 1$). Since the true data is also available it may be possible to compare the "cleaned" data set after the perturbed file has its outliers identified.

3.1. The data set in study

The UK Annual Business Inquiry (ABI) data set is a sample of 6099 enterprises carried out in the private economy in 1997/1998 and contains responses to selected questions for two sectors. Sector 1 is the base for five files, three of them about the year 1997, which are developed for training purpose: $sec197(true)$ has the true values, $sec197(y2)$ has missing values, $sec197(y3)$ has errors and missing values. Similarly, $sec198(y2)$ and $sec198(y3)$ refer to 1998.

The evaluation dataset to which the forward search will be applied is the file $sec197(y3)$, which had true values subjected to perturbation and some of them probably became outliers. (The exact mechanism for perturbation of the true values is unknown and is not our concern. However, it seems that basically part of the Y values are randomly multiplied by constant numbers like 10, 100 or 1000 and part had added constant values.) This working file is then a raw dataset including errors and outliers with information about 33 numerical variables from which 26 are independent and the other 7 can be derived from them.

Many variables in the list refer to expenditure on single components of business itens like, for example, payments, taxes and purchases. Due to the dimensionality of the data and to deal with some nonresponse and recorded zero values we have chosen to work only with the following seven numeric response variables:

Y_1 (*turnover*): Total turnover,
Y_2 (*taxtot*): Total taxes paid,
Y_3 (*purtot*): Total purchases of goods and services,
Y_4 (*emptotc*): Total employment costs,
Y_5 (*employ*): Total number of employees,
Y_6 (*assdisp*): Total proceeds from capital asset disposal,
Y_7 (*assacq*): Total costs of all capital assets acquired,

which refer to totals and perhaps may have been recorded with better precision than individual expenditures.

Under the assumption of complete response, we impute the missing observations with the true values provided by $sec197(true)$ and actually define our working dataset as only containing the original values plus the perturbed ones.

Two variables listed in the dataset were considered to be used as covariates to perform linear models fittings, namely *turnreg* (registered turnover) and *empreg* (registered number of employees), the last one available in a categorical type in the range {0,1,...,5}. Only *turnreg* was used as covariate because in general it has a very good linear relationship with the seven response variables, specially in log scale and then we denote it by X. When we tried to fit $Y_i|X,\ i = 1, \ldots, 7,$ within the six classes of *empreg* no major changes were detected to justify including it as a second covariate. Some classification variables such as, for example, *class*, are also available and will be discussed later for stratification purposes when looking for appropriate models for $Y|X$.

For two non–negative variables X and Y four partitions are possible: $(X = 0, Y = 0), (X = 0, Y > 0), (X > 0, Y = 0)$ and $(X > 0, Y > 0)$. In the real data the first case would not be surely the situation of an enterprise "open for business" and should not be present in the data. $(X = 0, Y > 0)$ could only be some kind of activity where the owner is the only employee and is not regarded as so when filling up the survey. However, it is possible that $(X > 0, Y = 0)$ has genuine units in the sample although it suggests a wrongly recorded zero value for Y. In terms of estimation, in order to find some estimate for μ_Y it would be essential to distinguish the observed $y = 0$ values according to if it is a true zero or a wrongly recorded zero. Ren (2001)–b considers this

issues, where logistic models can be used to find estimates of the evaluation of the conditional probability $Prob(y = 0|x)$ and appropriate weights to the recorded y values can then be applied to reach a good estimate of μ_Y or for the population total $T_Y = \sum_{i=1}^{N} Y_i$.

In some editing process, initially applied to the data before searching for outliers, the units related to the first three cases would probably be removed from the dataset. Therefore, due to the nature of the business data only the last case $(X > 0, Y > 0)$ will be studied here, where we will search for outliers possibly present into it. (A similar dataset will be defined for a search in the multivariate vector $\mathbf{Y}$.)

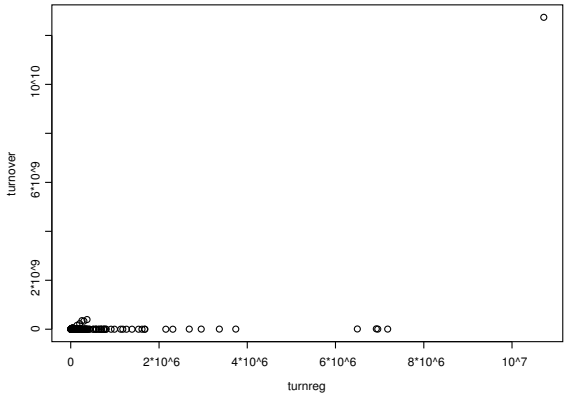
The analysis of the sample, when the study is restricted to $p = 1$, non–missing Y and $(X > 0, Y > 0$ leads to reduced individual data sets obtained from the ABI database. (The multivariate case $(p > 1)$ will also be considered but the study concentrates more on $p = 1$.) Different proportions of units for each variable were perturbed as Table I shows for those seven individual reduced datasets, from an initial analysis on the data.

Table I - Sample sizes(n_i) and number of perturbed units (np_i) for each individual response variable $Y_i,\ i = 1, \ldots, 7$, on reduced datasets defined by $(X > 0, Y_i > 0)$

Variable	n_i	np_i
Y_1 (turnover)	6082	239
Y_2 (taxtot)	5694	474
Y_3 (purtot)	6080	624
Y_4 (emptotc)	5423	330
Y_5 (employ)	5363	46
Y_6 (assdisp)	1451	219
Y_7 (assacq)	3048	242

Figure 1 - Scatterplot for data about *turnover* and *turnreg*, in the raw scale (a) and in the transformed log scale (b), with perturbed units marked with "×" symbol

(a) raw data



(b) transformed data

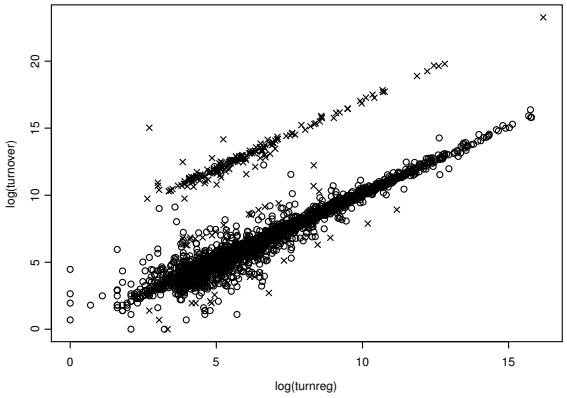
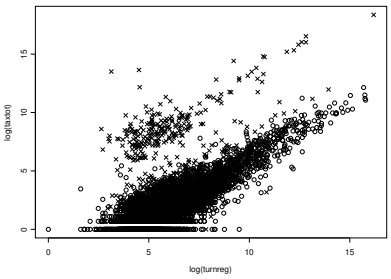
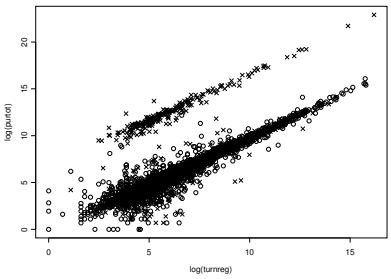


Figure 2 - Scatterplot for data about response variables $Y_i,\ i = 2, \ldots, 7$ and X in the transformed log scale, with perturbed units marked with "×" symbol

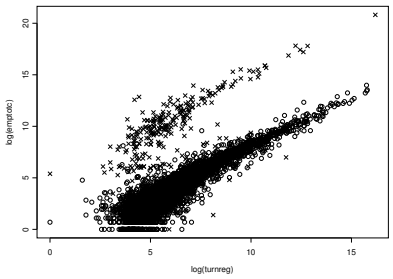
(a) Taxtot



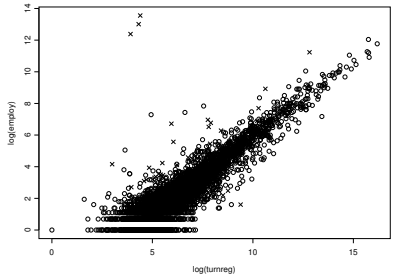
(b) Purtot



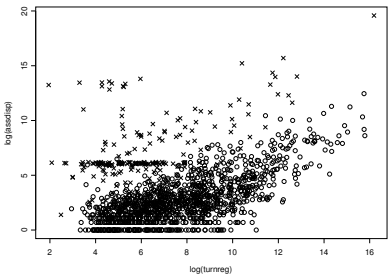
(c) Emptotc



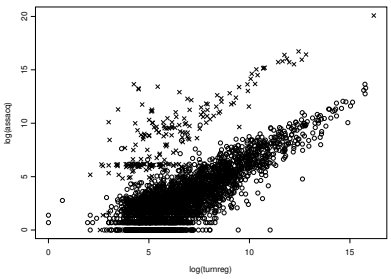
(d) Employ



(e) Assdisp



(f) Assacq



From an exploratory data analysis for each particular response variable $Y_i,\ i = 1, \ldots, 7$, it is clear that some transformation may be useful in order to provide a better linear fit for $Y|X$ as Figure 1a shows, for example, for Y_1. There is a huge spread for the sample y_1 data and a linear fit could suffer from the presence of influential or outlying values. After a simple log transformation applied to both variables Figure 1b shows that it may be easier now to search for outliers when the scale factor was kept in control and a good linear fit holds for $Y_1|X$. Units in the plot of Figure 1b marked with a "$\times$" symbol are those perturbed. It is clear that the most suspected values to be potential outliers were artificially generated by the perturbation mechanism. However, not all perturbed cases become outlying values, since lots of perturbed units are just inliers and cannot be seen on the plot. Similarly, some non perturbed units have an outlying pattern but cannot be seen on the plot because they are hidden in the main cloud of perturbed values.

The forward search to find outliers would require a good explanatory covariate X to predict Y and then identify outlying observations from the conditional distribution $Y_1|X$. Figure 2 also shows that a linear model looks appropriate for the other six response variables and the covariate chosen when they are studied in log scale, except for Y_6. Different power and log–log transformations were applied to the data but they were unable to create two new better correlated variables. We thus concentrate our search for outliers in the new variables defined by $Z_i = log(Y_i),\ i = 1, \ldots, 7,$ and use $V = log(X)$ as the available covariate. Some adjustments may be needed for correcting the bias in estimates in the raw scale for parameters of Y_i if transformed data Z_i is used both for outlier detection as for estimation; see Ren (2001)–b.

3.2. The choice of the starting subset

A clean subset from the data is required for performing the forward search for outliers. In order to define that subset the linear regression model is fitted to the data and the units with smallest residuals from (2) chosen to create it. Many different proportions of the sample were tried as the initial size for the clean data, such as 10%, 25%, 50% and 75%, alternatively to $m = q + 1$. No difference was found in the subset of units declared as outliers according to the different size for the starting clean data when checking the results. The proportion of detected outliers in the sample will generally be small and the clean subset increases at each step with the units which best agrees to the model fitted. Only at the last steps outlying units will then join it and so in practical applications probably there is no need to start with a very small clean subset to perform the search. Therefore, to save computational time the initial size was fixed in 75% of the sample size $n_i, i = 1, \ldots, 7$.

An important issue, however, concerns the way the regression model is fitted to the data since the basic starting clean data will be created from it. Figure 3 shows just for illustration part of the sample data for $Z_1 : log(turnover)$, where an influential point is present (actually an observation with $turnreg = 0$, but excluded from the data according to discussion in §3.1). If we use a least square fit both to define the clean subset and also to identify the outliers, it will arrive at units quite difficult to be accepted as so, seen on the plot on Figure 3*a* with a cross symbol "+". If a robust fit is used (here via the RREG Splus routine) to define the starting clean data then a much more reasonable set of outliers is identified even if at the next steps the ordinary least squares fit is used. Most importantly, now the introduced influential point, highlighted on both plots with a bigger symbol size, is identified as outlier.

Automatic programs that analyze large data sets may identify unreasonable units as outliers and it seems that the basic clean data set $C_{(m)}$ must be initially defined from a robust fit. In general we found that as long as $C_{(m)}$ is defined by the m units with smallest residuals from a robust fit in the whole sample there is no difference in the final set of outliers, regardless of the type of fit, least squares or robust, used in the subsequent steps. To be more specific, the n residuals at step m

$$\mathbf{e}_{(m)} = \{(e_1, \ldots, e_n),\ e_i = (y_i - x_i^{'}\hat{\beta}_{(m)})\}$$

depend on the estimate $\hat{\beta}_{(m)}$, which is found from the fit using only the clean data $C_{(m)}$. At the next step, the forward search moves to fitting $m+1$ observations after $C_{(m+1)}$ was defined by choosing the $m+1$ units with smallest residuals from $\mathbf{e}_{(m)}$.
When a robust fit is applied to the whole sample to choose the best m observations, the starting data $C_{(m)}$ can be quite clean and free of outliers. At the stage $m+1$ the estimates $\hat{\beta}_{(m+1)}$ will not present big changes since the just introduced $(m+1)$–th observation has the best agreement with the last fit. The most important is that the existing outliers (which will have the largest residuals) will not have any chance of being selected at an early stage. Since they are not chosen to take part on $C_{(m+1)}$ they will not influence $\hat{\beta}_{(m+1)}$ nor will be able to change the ordering in the next set of n residuals $\mathbf{e}_{(m+1)}$, which are used to define $C_{(m+2)}$. It thus may be an unnecessary sophistication to continue to robustify the fit on $C_{(m+1)}, C_{(m+2)}, \ldots$, if $C_{(m)}$ was already created by a robust fit.

Atkinson (1994) defines a random starting clean data $C_{(m)}$ and performs a series of different forward searches. When defining $C_{(m)}$ randomly we found that very soon distinct forward searches have exactly the same working clean data $C_{(m+s)}$, after a few s steps ahead. Hadi and Simonoff (1993) and Riani and Atkinson (2000) define a clean data to start from a robustly chosen subset and that is the way we prefer to do here.

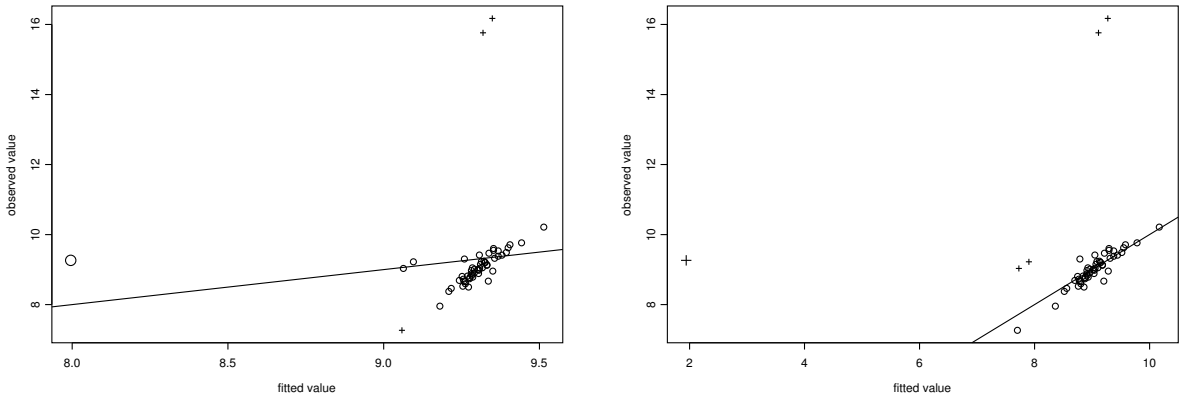
The kind of fit (ordinary least squares or robust regression) for the next steps seems to be equivalent as long as the initial subset is defined by a robust fit.

3.3 Implementation of the forward search

For each data set analyzed about univariate response variables $Z_i, i = 1, \ldots, 7$, a forward search was carried out starting with a clean subset defined by size $0.75n$ from the robust fit. Hadi and Simonoff (1993) stopping rule (6) was used and the resulting outliers identified. However, even after the stopping criterion had been met, the forward search was performed up to the whole sample to record the number of times each sample unit was declared as outlier and to have the sequence of residuals for all steps.

Figure 3 - Set of outliers ("+" symbol) identified in part of the Z_1 data, which has one influential observation included, according to type of fit used, (a)LS and (b) Robust

(a) LS (b) Robust



We define three sets of outliers:

O_{hs}: identified by Hadi and Simonoff stopping rule (6);
O_{mj}: identified by the maximum relative jump on τ (9);
O_b: identified by the binomial test on the proportion of steps that units were declared outliers (10).

The multivariate vector $\mathbf{Z}$ was defined for only the five first components, $\mathbf{Z} = (Z_1, \ldots, Z_5)$, since Y_6 and Y_7 had a large proportion of zeros. Both two components had also lots of missing values even in the true data. We assumed data fully observed and so missing values on those two variables cannot be recovered. The starting subset had a size $m = p + 1 = 6$ units, chosen from the Mahalanobis distances (5) from a robust fit in the full sample.

Here for $p > 1$ we identify outliers based on Hadi rule (7), comparing those results with the binomial test and on the maximum relative jump on τ, the determinant of the $p \times p$ clean residual covariance matrix.

The plot of residuals for the univariate search and the plot of the Mahalanobis distances for the multivariate case will be used as an aid to identify graphically the outlying and suspicious units.

Throughout his study any test performed will have a significance fixed at $\alpha = 0.01$, as we prefer to locate the most potential outliers, that is, our aim is to identify outliers that are really important.

4. Results

4.1. Outlier detection for univariate data $Z_1, \ldots Z_7$

Initially Table II displays the proportion of steps out of the 1521 steps performed in the forward search each unit in data set for variable Z_1 were declared outlier using rule (8). At least all the 197 units which were identified as outliers in all the 1521 steps performed must be declared outliers and some more units are also suspicious according to the high proportion of steps they did not fit the clean data to join it.

Table II - Frequency of p_i, the proportion number of steps
out of the 1521 steps performed,
units of response variable Z_1:log(turnover) were declared outlier

p_i	0	(0,10]	(10,20]	(20,30]	(30,40]	(40,50]
frequency	5166	55	49	53	68	45

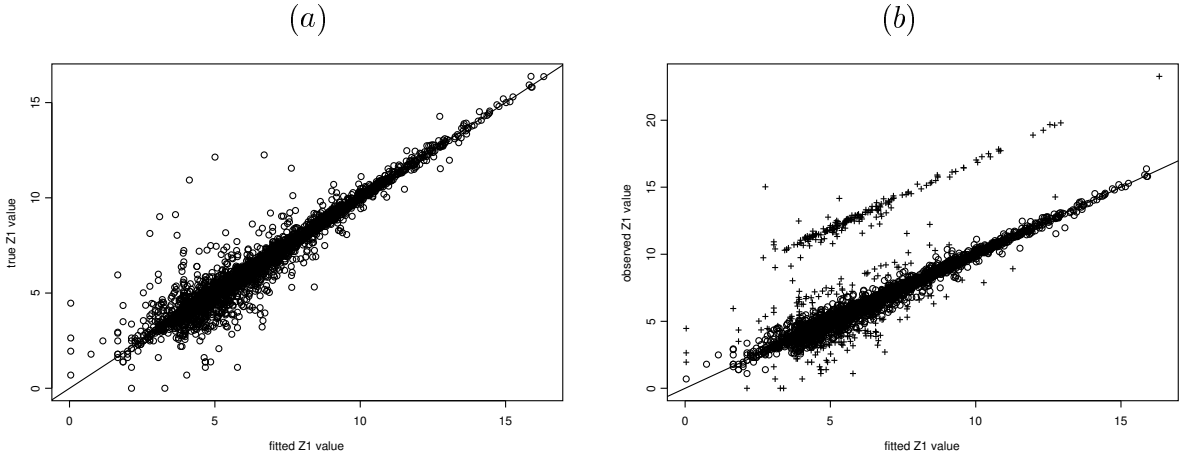
p_i	(50,60]	(60,70]	(70,80]	(80,90]	(90,100)	100
frequency	57	97	108	165	22	197

From this search the set of outliers O_{hs} identified by Hadi and Simonoff stopping rule has 349 units, which can be seen on Figure 4*b*. The figure displays the identified outliers with a cross symbol, in a plot with observed z_1 data against fitted data at the last step of the search (a full sample fit).

Figure 4*a* displays the fit on the true data about variable Z_1, which looks quite clean with only a few potential outliers whereas the perturbed data had the most important outlying units identified from the search, as Figure 4*b* shows. In real problems the true data would not be available and here it is only shown to visualize whether O_{hs} is a reasonable set of outliers or not.

The other two sets of outliers, denoted by O_{mj} and O_b, are also shown on Figure 5. Comparing Figure 4*b* and Figure 5 it appears that the HS approach locates much more outliers than the MJ and Binomial decision rule. A close examination on those three sets shows that all 198 units on O_b are included in the 268 components of O_{mj}, which are also part of O_{hs}, that is, $O_b \subset O_{mj} \subset O_{hs}$.

Figure 4 - (*a*) Fit on the true data for Z_1:log(turnover) and (*b*) Set of outliers O_{hs} identified (with "+" symbol) for Z_1:log(turnover) on the scatterplot for perturbed data after 1521 steps performed



The choice of the best set to represent the true outliers present on the data depend on how conservative we are about rejecting observed values when they look suspicious. Riani and Atkinson (2000) do not try to identify the outliers by a formal test but instead study the residuals through "fan" plots, monitoring changes associated with the fitting as the clean subset $C_{(m)}$ increases.

Figure 5 - Set of outliers $(a)O_{mj}$ and (b) O_b
identified for Z_1:log(turnover) on the scatterplot ("+" symbol)
for perturbed data after 1521 steps performed

(a) (b)

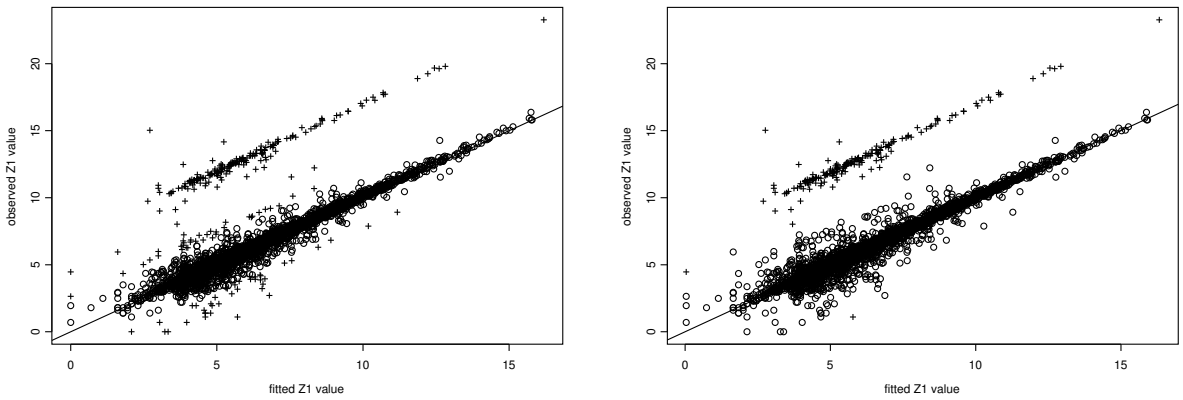
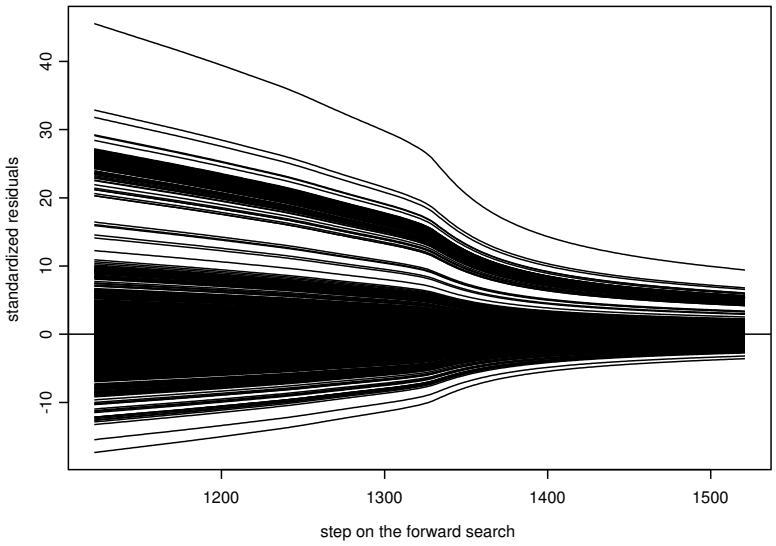


Figure 6 displays the standardized residuals for all the 6082 observed values for Z_1, for simplicity just for the last 400 steps of the search. It is clear that outliers are present on the data and at last steps important variations happen on the residuals as those outlying units are finally being included in the clean data. Not only the units with largest residuals are outliers but also some units with residuals on the border of the main bulk look very suspicious and therefore it may be the reason why O_{hs} includes those observations. A clear cluster of outlying units can be seen on this plot. The block of units with big positive standardized residuals corresponds exactly to the block of perturbed values, far away from the linear fit of Z_1 on V, according to Figure 5.

Figure 6 - Sequence of standardized residuals (2) for sample units of data Z_1:log(turnover) for the last 400 steps of the full forward search



For this data set it seems that there was no need to perform an expensive forward search starting with 3/4 of the sample size and going 1521 steps ahead. The analysis of the residuals clearly shows the outlying behaviour of some units as the clean data is augmented and this pattern is quite monotonic. However, an important change on the residuals is seen close to the step 1330, as important outliers start joining the clean data at that stage. This number suggests about 190 outliers, which agrees with the 198 units detected on O_b.

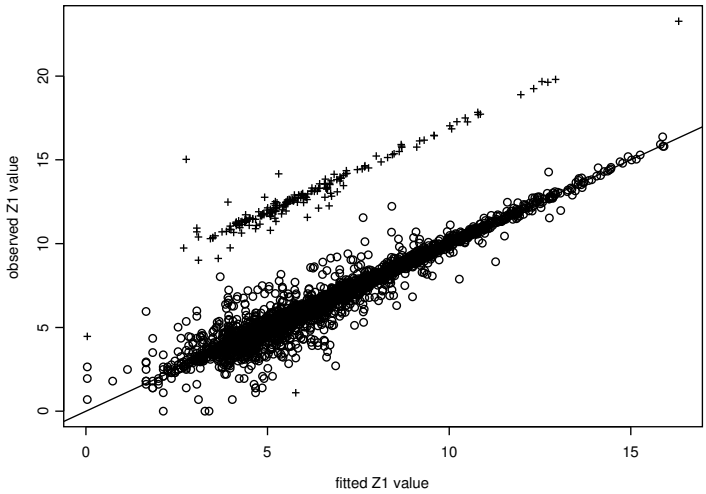
4.2. Importance of the forward search for outlier detection

Despite Figure 6 suggests no need to perform an extensive forward search there is a difference on results found from that approach and a simple outlier detection applied to the whole sample. Figure 7 displays the outliers detected on a single search on data about Z_1 when $C_{(m)} = C_{(n)}$.

Comparing the set of outliers from Figure 7 against those defined by O_{hs} on Figure 4*b*, 151 more units are detected as outlying on the forward search than on the single search. Probably, some backward search starting with the full sample n and deleting the outliers at each step until only clean data remains would reach the same 349 units

as those on O_{hs}. However, sophisticated actions should be taken to prevent the masking problem. At the second "trimming" step, we would need to insert back all the possible combinations of $A_{(1)}$ units previously detected as outliers at stage 1 (when $m_{(1)} = n$). It would be then possible to check the impact some k units from the first set of outliers $A_{(1)}$ would have on the construction of the second set $A_{(2)}$ (when analyzing the new dataset with size $m_{(2)} = n - A_{(1)} + k$). In performing the backward search that way only the true outliers should be identified.

Figure 7 - Set of outliers identified for Z_1:log(turnover) on the scatterplot (with "+" symbol) for perturbed data on a single search ($m = n$)



The exhaustive enumeration of all $A_{(j)}$ distinct tuples at current $j+1$ step may not be feasible specially for the sample sizes of the datasets studied here. Therefore, it seems that the forward search is less expensive and more reliable than the backward option. Table III shows the number of ouliers identified at the last step (a single search with the full sample) and those detected by the other rules considered in the forward search for all the seven datasets analyzed.

Table III - Number of outliers detected in the last step($m = n$) and by the forward methods, for datasets about $Z_i, i = 1, \ldots, 7$.

search	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7
$m = n$	198	231	242	225	51	37	121
O_b	198	232	243	225	62	43	128
O_{hs}	349	224	361	226	11	14	8
O_{mj}	268	231	295	225	55	38	123

Particularly for variable Z_7 there is an extreme disagreement between the results from the HS forward rule and from the single search ($m = n$). Outliers detected by the HS rule seem to be important and reasonable in the six other datasets as Figure 8 shows the outlying points present on O_{hs}. However, on data about Z_7 that rule has a breakdown for only 8 units and fails in locating quite similar outlying units on the same dataset. The main reason could be the nature of the data, where there is a huge dispersion between the fitted and observed values due to the bad linear relationship. Table IV follows the results from the outlier identification approach O_{hs} applied on the seven individual perturbed data sets, according to the nature of the sample values (perturbed or not). Perturbation has transformed some original values to ouliers but from previous plots on Figure 1*b* and Figure 2 it could be seen that not all identified outliers are perturbed observations, since outlying units were already present on the true data sets.

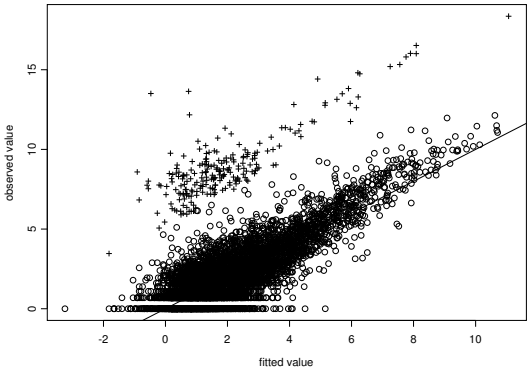
Table IV - Frequency of the number of sample units on perturbed data sets were identified as outliers by O_{hs}, according to their status (perturbed or not)

Status	Z_1 outlier		Z_2 outlier		Z_3 outlier	
	no	yes	no	yes	no	yes
Non pert.	5719	124	5215	5	5376	80
Perturbed	14	225	255	219	343	281

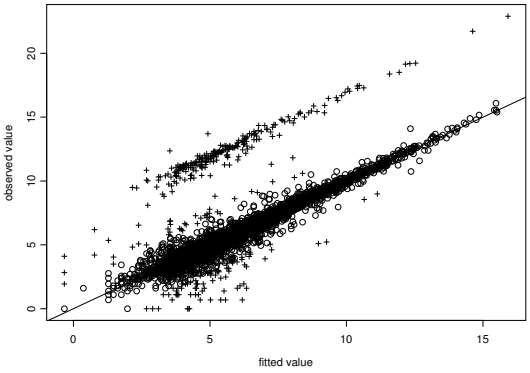
Status	Z_4 outlier		Z_5 outlier		Z_6 outlier		Z_7 outlier	
	no	yes	no	yes	no	yes	no	yes
Non pert.	5086	7	5311	6	1232	0	2806	0
Perturbed	111	219	41	5	205	14	234	8

Figure 8 - Set of O_{hs} outliers identified for $Z_2, \ldots, Z_7$, on the scatterplot (with "+" symbol) for perturbed data

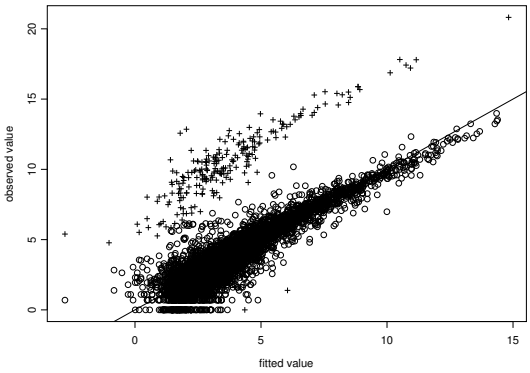
(a) $Z_2 : log(taxtot)$



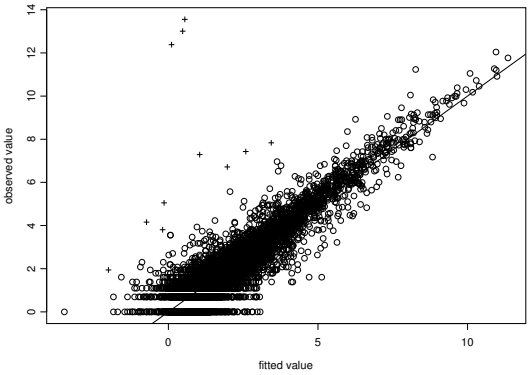
(b) $Z_3 : log(purtot)$



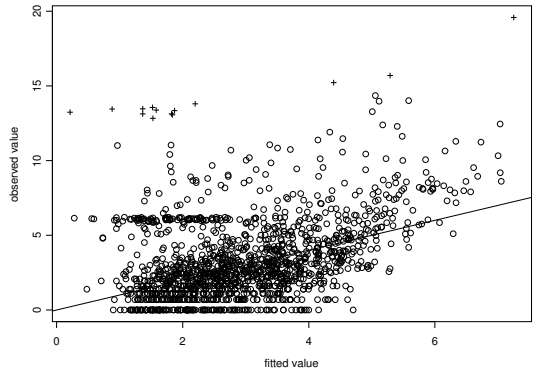
(c) $Z_4 : log(emptotc)$



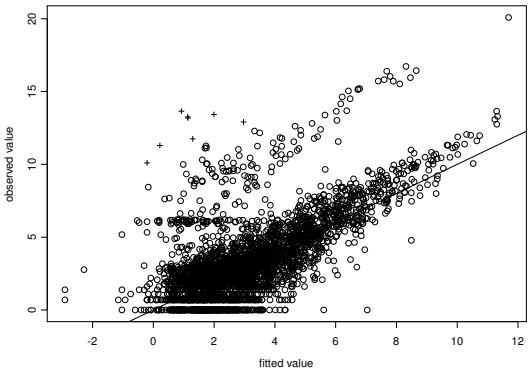
(d) $Z_5 : log(employ)$



(e) $Z_6 : log(assdisp)$



(f) $Z_7 : log(assacq)$



Some idea of the performance for the outlier detection method HS can be perhaps better obtained from the information on Table V, where units are investigated according to their classification status (outlier or not), on both the true and the perturbed individual data sets.

Table V - Frequency of the number of sample units on perturbed data sets were identified as outliers by O_{hs}, according to their classification status (outlier or not) on the original true data sets

Status on true data	Z_1 outlier		Z_2 outlier		Z_3 outlier	
	no	yes	no	yes	no	yes
Non outlier	5733	220 $^{(220)}$	5449	204 $^{(204)}$	5697	276 $^{(276)}$
Outlier	0	129 $^{(8)}$	3	6 $^{(1)}$	21	84 $^{(4)}$

Status on true data	Z_4 outlier		Z_5 outlier		Z_6 outlier		Z_7 outlier	
	no	yes	no	yes	no	yes	no	yes
Non outlier	5173	202 $^{(202)}$	5349	2 $^{(2)}$	1304	3 $^{(3)}$	2943	3 $^{(3)}$
Outlier	4	7 $^{(0)}$	3	6 $^{(0)}$	0	0	2	0

Small numbers on parentheses, on the columns for units related to identified outliers on the perturbed data sets, denote the number of those observations that have been subjected to perturbation. It can be seen, for example for variable Z_1:log(turnover), that 220 cases were detected as outliers in the perturbed data set although they had not been declared as outlying previously in the original true data set. The small number in parentheses, (220), just indicates that all those 220 units had been perturbed, which is the reason they are outliers in the new data set, that is, they are artificially generated outliers. Figure 9 shows the situation about those perturbed 8 units, out of the 129 cases, declared as outliers in both data sets, the original true data and the perturbed version. It is clear that the perturbation mechanism just made their degree of outlyingness worse.

Similarly, just for illustration, Figure 10 examines the 21 cases of Z_3:log(purtot), which were detected as outliers on the true data but not on the perturbed sample. The introduction of huge outliers has made the 21 units no longer declared as outliers on the perturbed data set.

Figure 9 - Position of 8 particular observations of Z_1:log(turnover)
a) declared as outliers in the original scale in the true data, and
(b) declared as outliers in the perturbed data after being subjected to perturbation

(a) True data (b) Perturbed data

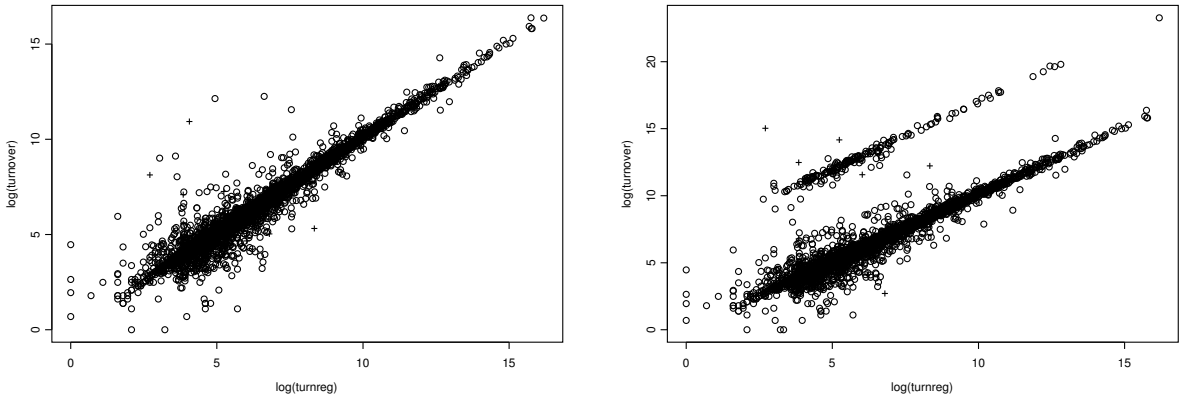
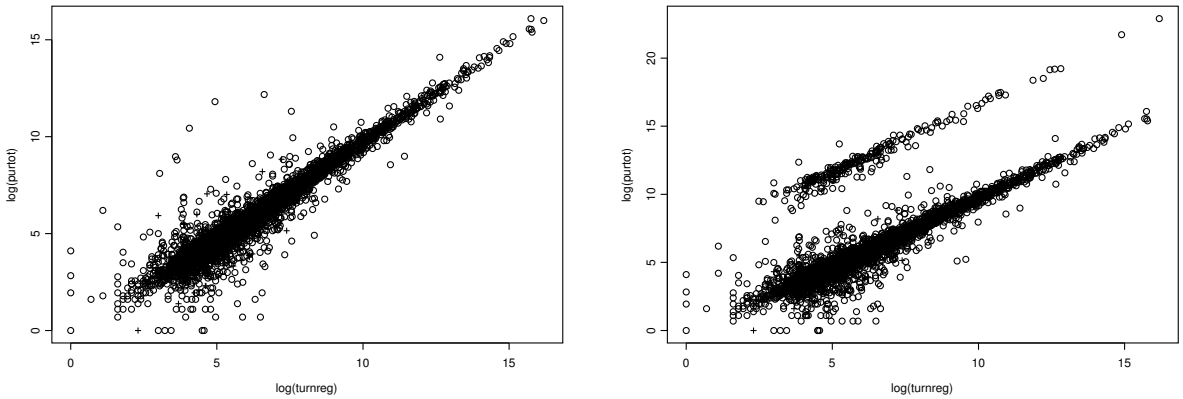


Figure 10 - Position of 21 particular observations of Z_5:log(purtot)
(a) declared as outliers in the true data and (b) not declared outliers in the perturbed data

(a) True data (b) Perturbed data



Results checked in different ways were found to be generally reasonable and so the outlier detection via the forward search approach seems to perform well.

4.3. Outlier detection for multivariate data $\mathbf{Z}_5 = (Z_1, \ldots, Z_5)$

For a five dimension vector $\mathbf{Z}_{(5)}$ we could make the partition of the associated data into different subsets $D_{(j)}$, according to the number of components $Z_1, \ldots, Z_5$ with non zero values in their sample units. For $D_{(2)}$, for example, only two individual components would be non zero, possibly reducing the outlier detection method to a two dimension problem. In the ABI file under study with $n = 6099$ units, the partition with the highest frequency is $D_{(5)}$, where 5118 observations are non zero for all the five considered components ($Z_1 > 0, \ldots, Z_5 > 0$). (Also, all those units have a non zero covariate value, like for the univariate definition of the data sets.)

We then also illustrate the forward search for this particular subset of data. The forward search starts now with a multivariate clean data defined by the units with smallest Mahalanobis distances from the robust fit on the whole sample, according to (5). Results are related to the starting basic subset with size fixed at $p + 1 = 6$ and do not disagree with those provided by another forward search, which started with sample size $m = \frac{n}{2}$.

By performing rotations on different combinations of three axes on the whole sample it is possible to identify some blocks of observations that are far from the main cloud of points. The forward search moves ahead at some stage by including a 5–dimensional sample unit and then probably on the subsequent steps incorporates a block of neighbouring units on that space. The sequence of the Mahalanobis distances seems to reflect this kind of structure of the data. From the last 600 steps performed on the search with Mahalanobis distances displayed on Figure 11, it is possible to visualize some kind of a multi dimensional cluster, far from the main block.

Table VI also provides the frequency of the proportion of number of steps in which the sample units were declared outlier in out of 5112 steps performed. The O_b set has 359 units declared as outlier when applying the binomial test to the relative frequency of times an observation is identified as outlying in the full search. Such number agrees with information provided by Table VI and shows that the forward search is perhaps worthwhile since the single outlier detection (for $m = n$) pointed out 345 units. By using the information from the maximum relative jump on the determinant of the clean residual variance, τ, this criterion identifies 351 outliers joining the clean data.

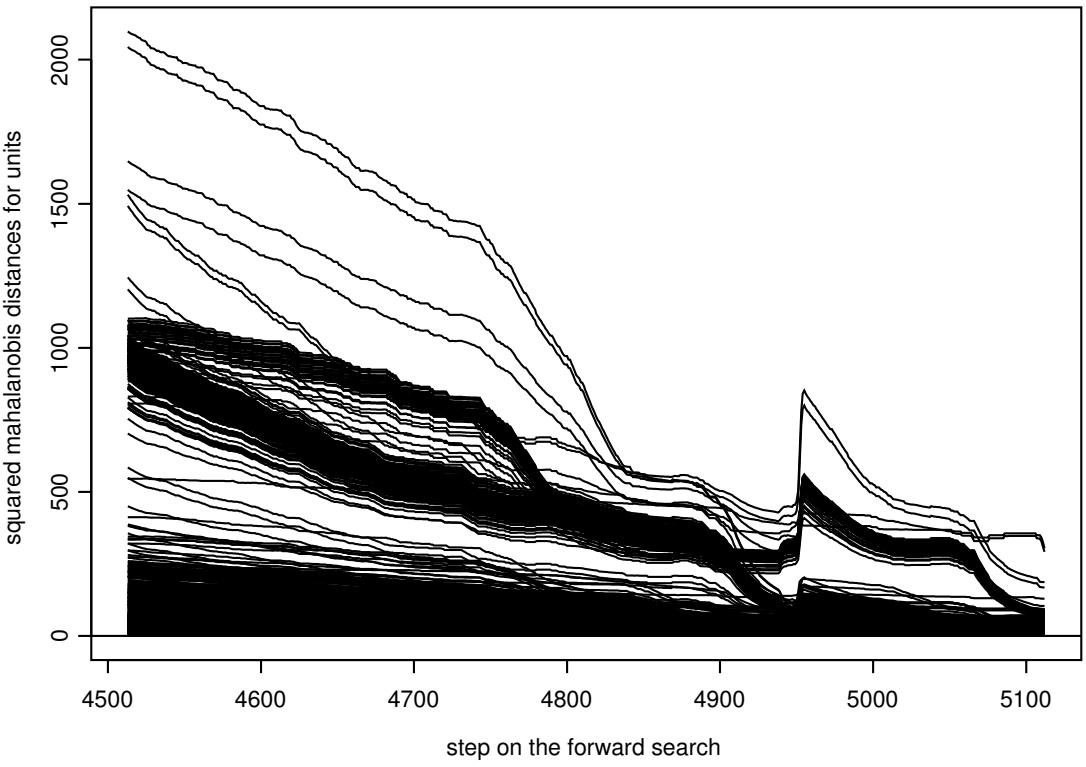
The Hadi rule has a breakdown for 481 outliers, according to the stopping criterion (6), again identifying more outlying units than the other two rules.

Table VI - Frequency of p_i, the proportion number of steps out of the 5112 steps performed, units of multivariate vector $D_{(5)}$ were declared outlier according to (8)

p_i	0	(0,10]	(10,20]	(20,30]	(30,40]	(40,50]
frequency	24	1443	918	531	370	289

p_i	(50,60]	(60,70]	(70,80]	(80,90]	(90,100)	100
frequency	242	194	184	251	327	345

Figure 11 - Sequence of Mahalanobis distances (4) for sample units of $D_{(5)}$ for the last 600 steps of the forward search



Concentrating only on the units detected in the O_{hs} set, we follow the number of multivariate units detected as outliers according to the number of their individual components with a particular characteristic. Table VII displays the possible impact that perturbation in the individual components may have in the multivariate sample unit, in order to cause it to be declared as a multivariate outlier.

Table VII - Frequency of the number of multivariate outliers from O_{hs}, according to $n_{(per)}$, the number of individual components of $D_{(5)}$ which have been perturbed

$n_{(per)}$	Multivariate outlier		Total
	No	Yes	
0	4198	96	4294
1	295	194	489
2	134	31	165
3	10	3	13
4	0	154	154
5	0	3	3
Total	4637	481	5118

Since perturbation does not necessarily creates outliers, the declaration of a sample unit as a multivariate outlier is considered according to the number of individual components that were declared outlier in the univariate search. Again only outliers identified by the HS rule are considered for results presented on Table VIII. It can be seen, for example, that any unit with more than just one individual component $Z_1, \ldots, Z_5$, identified as outlier in the univariate search is also declared as a multivariate outlier.

The most important finding, perhaps is related to the fact that 240 units out of the 254 units which have been detected as outlying in just one component have also been identified as a multivariate outlier. It means that a univariate search in individual components of the 5–dimension response vector will eventually identify almost the same units as outliers, when comparing to the multivariate search.

Table VIII - Frequency of the number of multivariate outliers from O_{hs}, according to $n_{(ind)}$, the number of individual components of $D_{(5)}$ detected as outliers on the univariate search

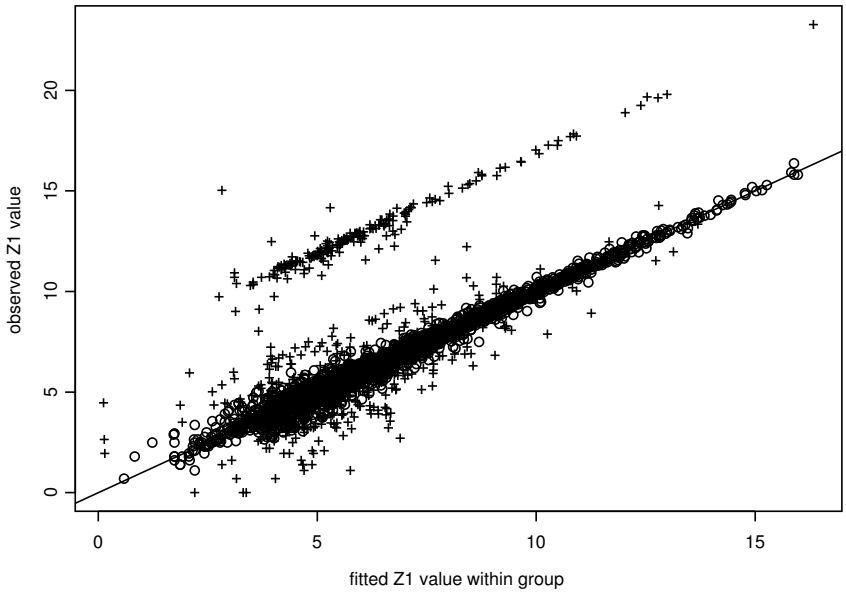
$n_{(ind)}$	Multivariate outlier		Total
	No	Yes	
0	4623	36	4659
1	14	240	254
2	0	39	39
3	0	16	16
4	0	146	146
5	0	4	4
Total	4637	481	5118

4.4. Outlier detection within stratum

We also performed forward searches within groups defined by classification variables. From the different options available, the best fitting for the selected response variables $Z_1, \ldots, Z_7$, against the covariate *turnreg*, all in log scale, was within the stratification defined by the variable *class*. Small sample sizes for some classes forced to create then 27 groups, where the biggest had 2167 cases. By defining individual datasets with only non-zero values there was a variation on the sample sizes of the 27 stratum. Generally the linear relationship between Z_i and V still holds but in some stratum the fit could be poor.

Figure 12, for example, presents the location of the outliers in the univariate search on the dataset about variable Z_1 when we apply the HS rule. The identification of 467 observations as outliers clearly outnumbers the 349 HS cases detected in the across–stratum search since inliers on the overall data can be detected as outliers within particular groups. Comparing Figure 12 against Figure 4 is not straightforward since the plots are presented for fitted and observed values. All the 27 groups have sample values plotted together in the same figure and so fitted values within strata differ from those found by a overall across–stratum fit.

Figure 12 - Outliers located for variable Z_1:log(turnover) according to the HS rule in the within stratum forward search (with "+" symbol)



Similarly to Table III, results for the within strata search are presented on Table IX, according to the rule used to detect outliers.

Table IX - Number of outliers detected in the last step($m = n$) and by different rules, for datasets about $Z_i,\ i = 1, \ldots, 7$, in the within stratum forward search

search	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7
$m = n$	201	223	243	222	56	34	116
O_b	233	271	270	245	103	103	163
O_{hs}	467	245	441	279	77	56	102
O_{mj}	286	276	309	250	82	89	155

For all variables studied a within stratum search locates more outliers than the across–stratum study except on variable Z_6, where $Z_6|V$ had not a good linear fitting. The identification of multivariate outliers was not performed within stratum since most of the 27 groups have a small sample size to justify a good fit in five dimension. However, accordingly with the results above for the univariate search, it should be expected more outlying multivariate units than the number located in the across stratum search.

4.5. Performance evaluation of the HS rule

We follow the evaluation approach from and Xinqiang and Chambers (2002), where the measures suggested to evaluate the efficiency of an outlier detection method are defined as:

$$R_1 = \frac{N_{error}}{N_{pert}},$$

$$R_{sig} = \frac{N_{sig}}{N_{pert}},$$

$$R_2 = 1 - \frac{N_{error}}{N_{out}},$$

and N_{error} refers to the number of true errors identified by the outlier detection rule, N_{pert} is the number of perturbed units in the data set , N_{out} is the number of detected outliers and N_{sig} is the number of identified significant outliers, which have been subjected to a strong perturbation (the observed value differs from the true value in more than 100% of the true value, in absolute terms).
It is desirable that both R_{sig} and $R_1(1-R_2)$ values are maximum. For the first measure, a high value for R_{sig} means that the majority of the important outliers would have been detected. For $R_1(1-R_2)$ we want to identify as many outliers as possible but keeping the number of non–errors identified as outliers as small as possible.
Table X–1 that follow and Table X–2 to Table X–7 in the appendix present those evaluated measures for the Hadi/Simonoff approach, for the two kind of forward searches performed, across and within stratum, both for significance levels 1% and 5% in the outlier detection procedure.
It is also possible to have a preliminary comparison about the importance of the forward search, taking a look at the evaluated measures for the performance of the single search for outliers ($m = n$). In that latter simplistic approach, we just try to identify the outliers present in the data looking just once at the whole sample. Usually this rule identifies most of the important perturbed errors and then has a high value for R_1 and R_{sig}.
It seems that the single search performs quite well in terms of the measures R_{sig} and $R_1(1-R_2)$ as this conservative rule prefers to identify only the most important outliers. As a consequence it results in a small value for the proportion R_2 of non–errors identified as outliers. More precise outlier detection methods, like the forward search algorithm, can have a better performance in finding the true outliers (not subjected to the perturbation mechanism) present in the data sets but would then assume a higher value for R_2.
At this stage we do not feel that the forward search is not worthwhile in terms of its

computational cost by just comparing the tables for its evaluated performance to results from the single search. Important issues like the presence of true outliers and masking should be better explored in a further study in order to have a more comprehensive evaluation of this more expensive Hadi/Simonoff algorithm.

Table X–1 - HS outlier identification performance on **TURNOVER** data

($n = 6082$, 5843 true cases, 239 perturbed, 206 significant)

Number of accepted and rejected sample units in the editing process
by the type of search (across or within stratum),
according to their status (true or perturbed value)

Editing at significance 1%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	5719	124	5603	240
Perturbed value	14	225	12	227
Total	5733	349	5615	467

Editing at significance 5%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	5641	202	5483	360
Perturbed value	13	226	9	230
Total	5654	428	5492	590

Evaluated performance for the Hadi/Simonoff rule

Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	349	225	206	0.9414	1	0.3553	0.6069
Across 5%	428	226	206	0.9456	1	0.4720	0.4993
Within 1%	467	227	206	0.9498	1	0.5139	0.4617
Within 5%	590	230	206	0.9623	1	0.6102	0.3752

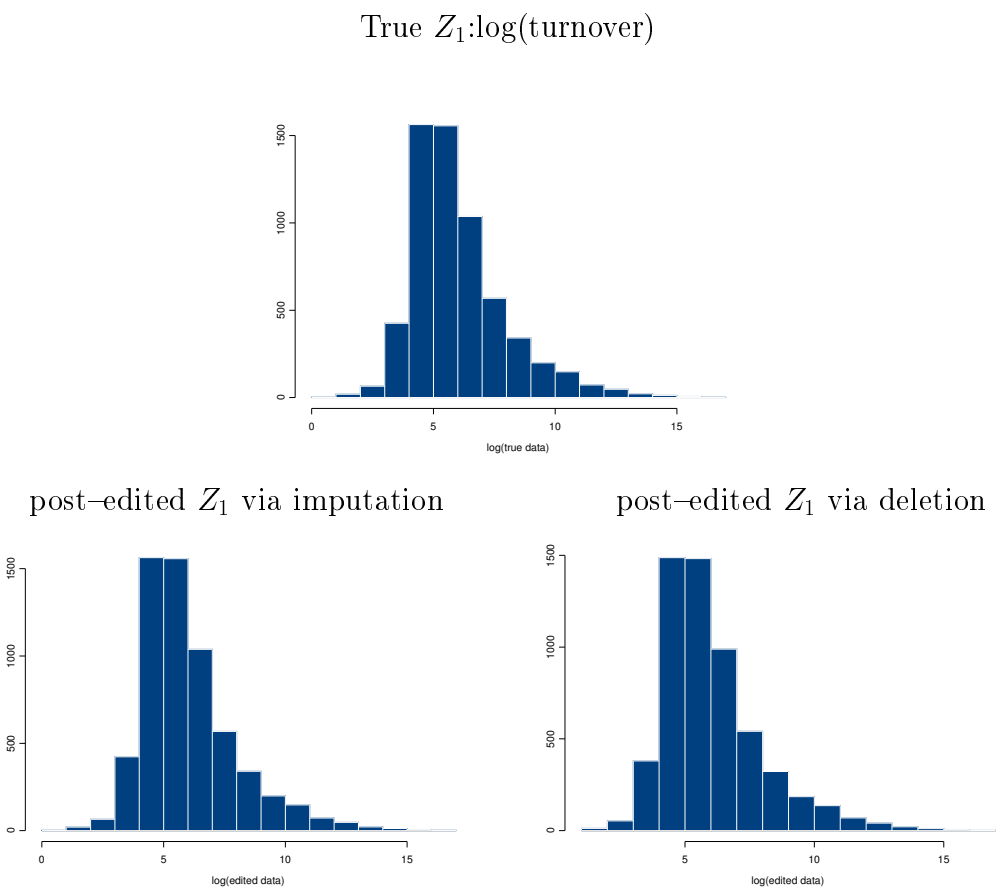
Evaluated performance for the "single" search ($m = n$ or last step only)

Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	198	190	189	0.7950	0.9175	0.0404	0.7629
Across 5%	209	192	190	0.8033	0.9223	0.0813	0.7380
Within 1%	201	191	189	0.7992	0.9175	0.0498	0.7594
Within 5%	211	195	193	0.8159	0.9369	0.0758	0.7540

4.6. Post–editing for the HS outlier detection rule

Suppose once the most important outliers have been detected an editing process is carried out with two possible options: impute the rejected observations or perform a follow–up survey to recover/to confirm the true data, or just remove those suspicious values from the data set being analyzed.
When an efficient imputation method is applied to replace the identified outliers hopefully all the rejected observations would have a donor value very close to the true value. If such an efficient recovery for the identified observations at error is not feasible, suppose we then delete those suspicious values from our data and just work with the remaining observations.
Figure 13 displays the histograms for the univariate data set about variable TURNOVER in log scale, for the true observations and for the post–edited data, by assuming the imputed value perfectly recovered the actual true value or by removing the rejected values.
It is clear that the editing has been successful in detecting the majority of the important outliers for the variable TURNOVER in log scale. These histograms reflect the good performance for the HS outlier detection rule confirming the measures evaluated on Table X–1.

Figure 13 - True and post–edited data for Z_1
after the O_{hs} outliers (identified in the across stratum forward search)
have been corrected by appropriate imputed values or removed



For the next four survey variables (TAXTOT, PURTOT, EMPTOTC and EMPLOY) both options (a good imputation scheme or the deletion of identified errors) generate clean data sets close to the true data as it can be seen from Figure 14 to Figure 17, presented in the appendix with the corresponding tables for the evaluated performance for the HS rule.

From Figure 18, however, it is clear that the Hadi/Simonoff rule fails in identifying important perturbed values for the variable ASSDISP and the replacement of the few outliers detected by the forward search, even when recovered perfectly by the true value, is not enough to generate a clean distribution close to the true data. The simplistic option of removing suspicious values from the sample still presents a post–edited distribution far from the distribution of the true data. Figure 19 for the ASSACQ data shows again the bad performance for the HS method as Table X–7 had already suggested; see the appendix.

5. General conclusions

Different rules for outlier detection have been subject of our investigation. For the particular datasets and response variables used in the application of those techniques some remarks could be mentioned:

(a) It appears that the forward search is important to be performed for outlier detection purpose since sets of outlying observations from different rules consistently differ from the set located on a single search on the whole sample data. Probably in the perturbed datasets under study the most important outliers, detected on a single search, were masking the presence of some more outlying observations. Those extra values should also be identified and generally the forward search locates them. Despite the computational cost involved in the forward search we believe it is a better procedure than backward methods.

(b) For the three rules considered results suggest that the binomial test locates only the most extreme outliers under the assumption that δ, the probability that some unit value is a true outlier in the population, is equal to 1. If that assumption is relaxed to accept slightly less suspicious observations, by setting δ to, say, 0.9, more outliers should be identified.

The empirical rule based on the maximum relative increase on the residual variance from the clean data (as it increases at each step) generally provides less outliers than the HS rule. Although in some occasions the sets O_{mj} and O_{hs} have almost similar observations the rule from Hadi and Simonoff (1993) seems to depend not exactly on the maximum relative increase on the clean residual variance. The first local maximum on that relative increase, caused by the introduction of the first important outlier, seems to cause the breakdown of the HS rule. Since an exact statistic is not available to test the *maximum jump* criterion and because its results should generally be similar to the HS rule, we prefer to use the precise Hadi/Simonoff approach.

(c) The kind of search, within or across–stratum, depends on the type of inference the outlier identification is performed for. If estimation is desirable within stratum then outliers need to be identified in each group to have the estimates more precise.
In our study results concentrate on the across–stratum search because some variables present a bad fit with the covariate within some particular groups. In general the linear relationship between the response variables and the covariate is quite good in the whole sample data. Therefore the outliers can be reasonably well detected and we prefer to refer on results from the across–stratum study.

(d) The multivariate search for outliers was performed here only in a across–stratum study mainly due to a good general fit for the whole sample and also due to small sample sizes in some groups.

From the search on the individual components of the response vector it can be seen that as long as an unit is detected as outlier in more than one component it will be a multivariate outlier too. There is also a high relative frequency of units that are declared as a multivariate outlier when they have just one single component identified as outlying in an univariate search.

The last fact emphasizes the importance of the univariate search on individual components of a response vector. For practical applications in real data problems, particularly for the type of data this study explores, it would have the advantage of locating suspicious units in a survey when one single component is not acceptable. If the outlier detection method is applied in a univariate way, the whole file obtained from some sample unit could be rejected or at least subject to further investigation when just one of its values recorded for those most important survey response variables is considered outlier.

References

Atkinson, A.C. (1994). "Fast very robust methods for the detection of multiple outliers". *JASA*, **89**, 428, 1329–1339.

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data*, Wiley.

Hadi, A.S. (1994). "A modification of a method for the detection of outliers in multivariate samples", *J.R. Statist. Soc. B*, **56**, No. 2, 393–396.

Hadi, A.S. and Simonoff, J.S. (1993). "Procedures for the identification of multiple outliers in linear models", *JASA*, **88**, 1264–1272.

Riani, M. and Atkinson, A.C. (2000). "Robust diagnostic data analysis: transformations in regression", *Technometrics*, **42**, 4, 384–398.

Xinquiang, Z. and Chambers, R. (2002). "Outlier identification and imputation using robust regression trees", *Euredit* deliverables, D4.2.1, University of Southampton, U.K.

APPENDIX

Table X–2 - HS outlier identification performance on **TAXTOT** data

to

Table X–7 - HS outlier identification performance on **ASSACQ** data

Figure 14 - True and post–edited data for Z_2

to

Figure 19 - True and post–edited data for Z_7

Table X–2 - HS outlier identification performance on **TAXTOT** data

($n = 5694$, 5220 true cases, 474 perturbed, 439 significant)

Number of accepted and rejected sample units in the editing process
by the type of search (across or within stratum),
according to their status (true or perturbed value)

Editing at significance 1%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	5215	5	5210	10
Perturbed value	255	219	239	235
Total	5470	224	5449	245

Editing at significance 5%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	5214	6	5192	28
Perturbed value	251	223	231	243
Total	5465	229	5423	271

Evaluated performance for the Hadi/Simonoff rule

Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	224	219	219	0.4620	0.4989	0.0223	0.4517
Across 5%	229	223	223	0.4705	0.5080	0.0262	0.4581
Within 1%	245	235	235	0.4958	0.5353	0.0408	0.4755
Within 5%	271	243	242	0.5127	0.5513	0.1033	0.4597

Evaluated performance for the "single" search ($m = n$ or last step only)

Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	231	225	225	0.4717	0.5125	0.0260	0.4594
Across 5%	247	238	238	0.5021	0.5421	0.0364	0.4838
Within 1%	223	219	219	0.4620	0.4989	0.0179	0.4537
Within 5%	254	246	246	0.5190	0.5604	0.0315	0.5026

Table X–3 - HS outlier identification performance on **PURTOT** data

($n = 6080$, 5456 true cases, 624 perturbed, 252 significant)

Number of accepted and rejected sample units in the editing process
by the type of search (across or within stratum),
according to their status (true or perturbed value)

Editing at significance 1%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	5376	80	5309	147
Perturbed value	343	281	330	294
Total	5719	361	5639	441

Editing at significance 5%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	5344	112	5251	205
Perturbed value	337	287	321	303
Total	5681	399	5572	508

Evaluated performance for the Hadi/Simonoff rule

Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	361	281	247	0.4503	0.9802	0.2216	0.3505
Across 5%	399	287	248	0.4599	0.9841	0.2807	0.3308
Within 1%	441	294	250	0.4712	0.9921	0.3333	0.3141
Within 5%	508	303	250	0.4856	0.9921	0.4035	0.2896

Evaluated performance for the "single" search ($m = n$ or last step only)

Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	242	232	232	0.3718	0.9206	0.0413	0.3564
Across 5%	249	235	235	0.3766	0.9325	0.0562	0.3554
Within 1%	243	233	233	0.3734	0.9246	0.0412	0.3580
Within 5%	253	238	237	0.3814	0.9405	0.0593	0.3588

Table X–4 - HS outlier identification performance on **EMPTOTC** data

($n = 5423$, 5093 true cases, 330 perturbed, 265 significant)

Number of accepted and rejected sample units in the editing process
by the type of search (across or within stratum),
according to their status (true or perturbed value)

Editing at significance 1%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	5086	7	5051	42
Perturbed value	111	219	93	237
Total	5197	226	5144	279

Editing at significance 5%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	5083	10	5034	59
Perturbed value	106	224	84	246
Total	5189	234	5121	305

Evaluated performance for the Hadi/Simonoff rule

Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	226	219	218	0.6636	0.8226	0.0310	0.6431
Across 5%	234	224	223	0.6788	0.8415	0.0427	0.6498
Within 1%	279	237	229	0.7182	0.8642	0.1505	0.5766
Within 5%	305	246	236	0.7455	0.8906	0.1934	0.5475

Evaluated performance for the "single" search ($m = n$ or last step only)

Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	225	219	219	0.6636	0.8264	0.0267	0.6459
Across 5%	245	233	233	0.7061	0.8792	0.0490	0.6715
Within 1%	222	218	218	0.6606	0.8226	0.0180	0.6487
Within 5%	242	234	234	0.7091	0.8830	0.0331	0.6856

Table X–5 - HS outlier identification performance on **EMPLOY** data

($n = 5363$, 5317 true cases, 46 perturbed, 30 significant)

Number of accepted and rejected sample units in the editing process
by the type of search (across or within stratum),
according to their status (true or perturbed value)

Editing at significance 1%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	5311	6	5264	53
Perturbed value	41	5	22	24
Total	5352	11	5286	77

Editing at significance 5%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	5308	9	5245	72
Perturbed value	38	8	22	24
Total	5346	17	5267	96

Evaluated performance for the Hadi/Simonoff rule

Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	11	5	5	0.1087	0.1667	0.5455	0.0494
Across 5%	17	8	7	0.1739	0.2333	0.5294	0.0818
Within 1%	77	24	19	0.5217	0.6333	0.6883	0.1626
Within 5%	96	24	19	0.5217	0.6333	0.7500	0.1304

Evaluated performance for the "single" search ($m = n$ or last step only)

Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	51	17	17	0.3696	0.5667	0.6667	0.1232
Across 5%	113	28	28	0.6087	0.9333	0.7522	0.1508
Within 1%	56	27	27	0.5870	0.9000	0.5119	0.2830
Within 5%	108	28	28	0.6087	0.9333	0.7407	0.1578

Table X–6 - HS outlier identification performance on **ASSDISP** data

($n = 1451$, 1232 true cases, 219 perturbed, 213 significant)

Number of accepted and rejected sample units in the editing process
by the type of search (across or within stratum),
according to their status (true or perturbed value)

Editing at significance 1%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	1232	0	1228	4
Perturbed value	205	14	167	52
Total	1437	14	1395	56

Editing at significance 5%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	1232	0	1218	14
Perturbed value	204	15	122	97
Total	1436	15	1340	111

Evaluated performance for the Hadi/Simonoff rule

Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	14	14	14	0.0639	0.0657	0	0.0639
Across 5%	15	15	15	0.0685	0.0704	0	0.0685
Within 1%	56	52	52	0.2374	0.2441	0.0714	0.2205
Within 5%	111	97	97	0.4429	0.4554	0.1261	0.3871

Evaluated performance for the "single" search ($m = n$ or last step only)

Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	37	37	37	0.1689	0.1737	0	0.1689
Across 5%	73	71	71	0.3242	0.3333	0.0274	0.3153
Within 1%	34	34	34	0.1553	0.1596	0	0.1553
Within 5%	73	68	68	0.3105	0.3192	0.0685	0.2892

Table X–7 - HS outlier identification performance on **ASSACQ** data

($n = 3048$, 2806 true cases, 242 perturbed, 231 significant)

Number of accepted and rejected sample units in the editing process
by the type of search (across or within stratum),
according to their status (true or perturbed value)

Editing at significance 1%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	2806	0	2794	12
Perturbed value	234	8	152	90
Total	3040	8	2946	102

Editing at significance 5%

Status of Y_{ij}	Across stratum		Within stratum	
	accepted	rejected	accepted	rejected
True value	2806	0	2792	14
Perturbed value	224	18	119	123
Total	3030	18	2911	137

Evaluated performance for the Hadi/Simonoff rule

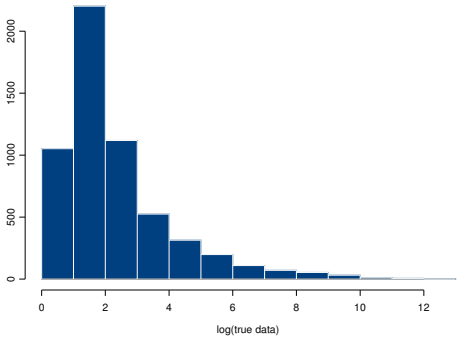
Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	8	8	8	0.0331	0.0346	0	0.0331
Across 5%	18	18	18	0.0744	0.0779	0	0.0744
Within 1%	102	90	90	0.3719	0.3896	0.1176	0.3281
Within 5%	137	123	123	0.5083	0.5235	0.1022	0.4564

Evaluated performance for the "single" search ($m = n$ or last step only)

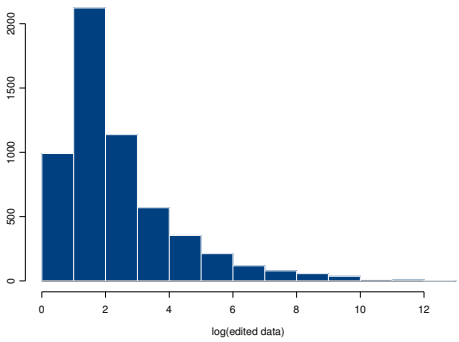
Search	N_{out}	N_{error}	N_{sig}	R_1	R_{sig}	R_2	$R_1(1-R_2)$
Across 1%	121	119	119	0.4917	0.5152	0.0165	0.4836
Across 5%	180	174	174	0.7190	0.7532	0.0333	0.6950
Within 1%	116	114	114	0.4711	0.4935	0.0172	0.4630
Within 5%	168	163	163	0.6736	0.7056	0.0298	0.6535

Figure 14 - True and post–edited data for Z_2
after the O_{hs} outliers (identified in the across stratum forward search)
have been corrected by appropriate imputed values or removed

True Z_2:log(taxtot)



post–edited Z_2 via imputation



post–edited Z_2 via deletion

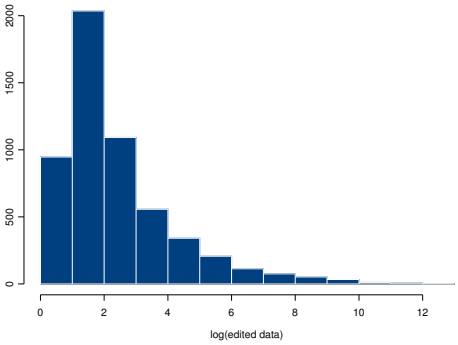
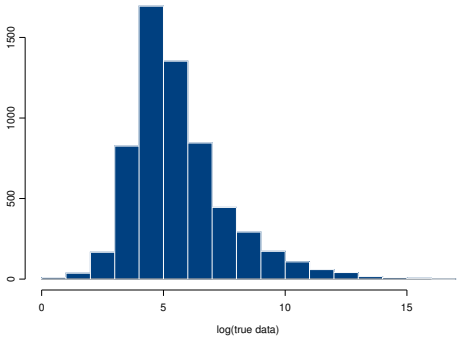
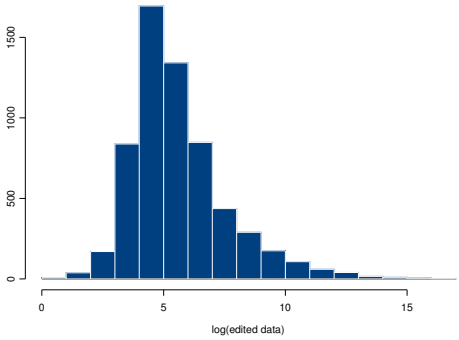


Figure 15 - True and post–edited data for Z_3
after the O_{hs} outliers (identified in the across stratum forward search)
have been corrected by appropriate imputed values or removed

True Z_3:log(purtot)



post–edited Z_3 via imputation



post–edited Z_3 via deletion

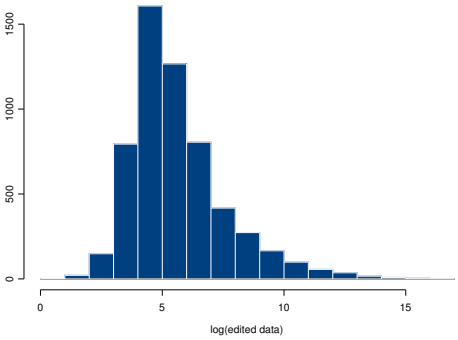
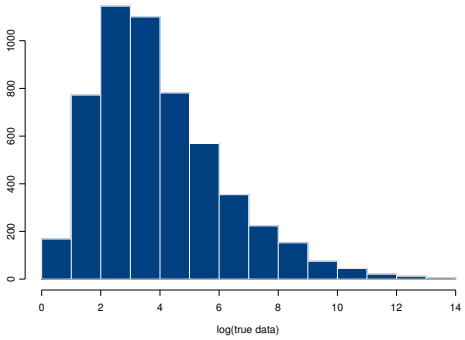
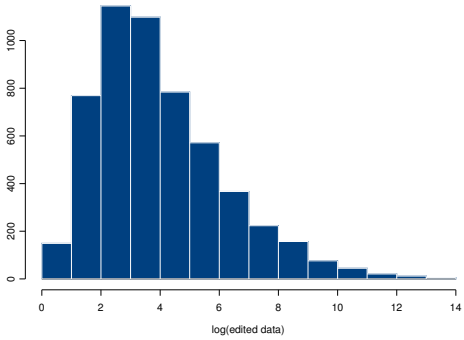


Figure 16 - True and post–edited data for Z_4
after the O_{hs} outliers (identified in the across stratum forward search)
have been corrected by appropriate imputed values or removed

True Z_4:log(emptotc)



post–edited Z_4 via imputation



post–edited Z_4 via deletion

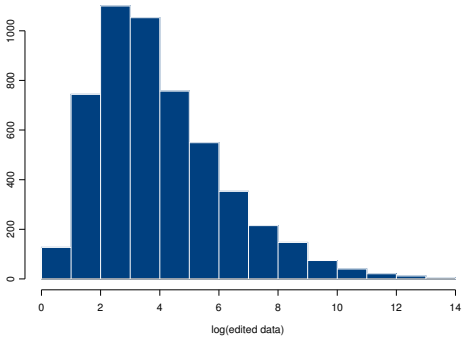
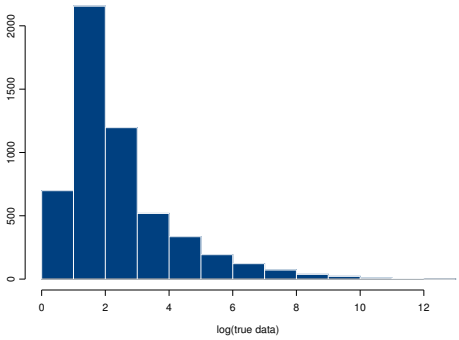
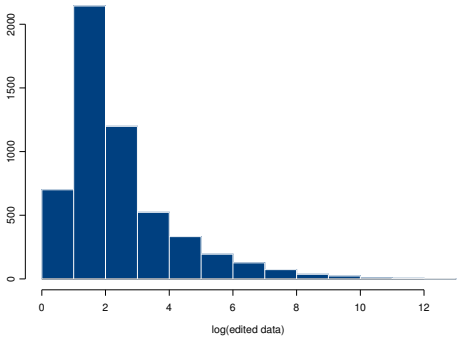


Figure 17 - True and post–edited data for Z_5
after the O_{hs} outliers (identified in the across stratum forward search)
have been corrected by appropriate imputed values or removed

True Z_5:log(employ)



post–edited Z_5 via imputation



post–edited Z_5 via deletion

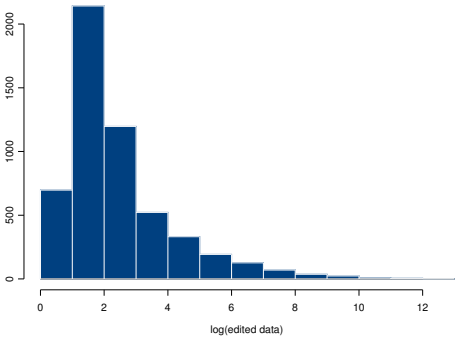
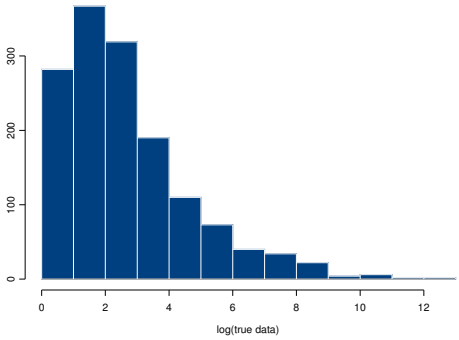
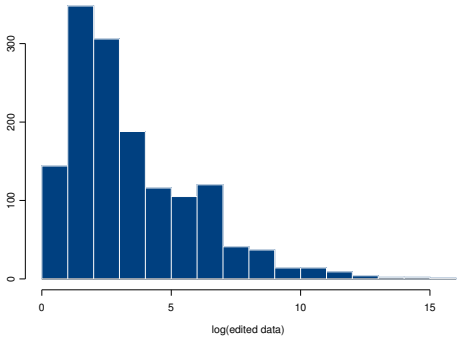


Figure 18 - True and post–edited data for Z_6
after the O_{hs} outliers (identified in the across stratum forward search)
have been corrected by appropriate imputed values or removed

True Z_6:log(assdisp)



post–edited Z_6 via imputation



post–edited Z_6 via deletion

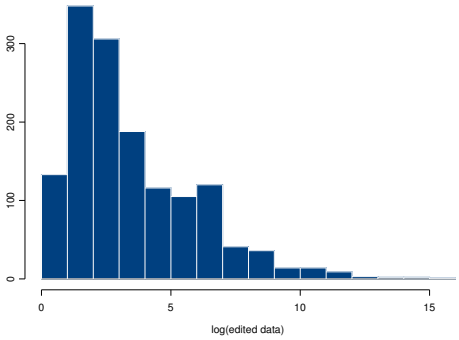
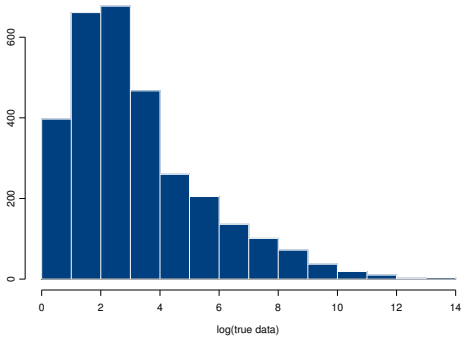
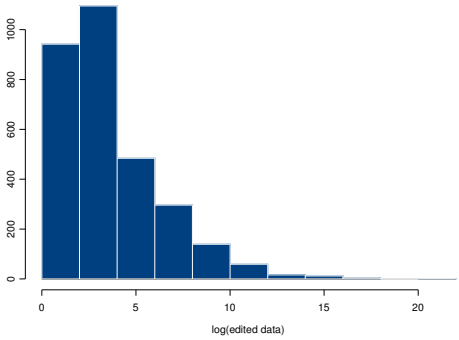


Figure 19 - True and post–edited data for Z_7
after the O_{hs} outliers (identified in the across stratum forward search)
have been corrected by appropriate imputed values or removed

True Z_7:log(assacq)



post–edited Z_7 via imputation



post–edited Z_7 via deletion