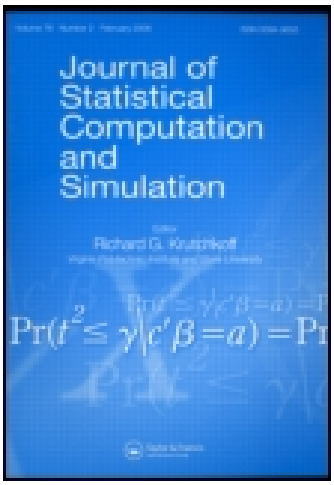

Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

Outlier detection and robust estimation in linear regression models with fixed group effects

Betsabé Pérez[a], Isabel Molina[a] & Daniel Peña[a]

[a] Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain

Published online: 26 Jun 2013.

To cite this article: Betsabé Pérez, Isabel Molina & Daniel Peña (2014) Outlier detection and robust estimation in linear regression models with fixed group effects, Journal of Statistical Computation and Simulation, 84:12, 2652-2669, DOI: 10.1080/00949655.2013.811669

To link to this article: http://dx.doi.org/10.1080/00949655.2013.811669

Journal of Statistical Computation and Simulation, 2014
Vol. 84, No. 12, 2652–2669, http://dx.doi.org/10.1080/00949655.2013.811669

Outlier detection and robust estimation in linear regression models with fixed group effects

Betsabé Pérez, Isabel Molina* and Daniel Peña

Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain

(*Received 29 September 2012; final version received 1 June 2013*)

This work studies outlier detection and robust estimation with data that are naturally distributed into groups and which follow approximately a linear regression model with fixed group effects. For this, several methods are considered. First, the robust fitting method of Peña and Yohai [A fast procedure for outlier diagnostics in large regression problems. J Am Stat Assoc. 1999;94:434–445], called principal sensitivity components (PSC) method, is adapted to the grouped data structure and the mentioned model. The robust methods RDL_1 of Hubert and Rousseeuw [Robust regression with both continuous and binary regressors. J Stat Plan Inference. 1997;57:153–163] and M-S of Maronna and Yohai [Robust regression with both continuous and categorical predictors. Journal of Statistical Planning and Inference 2000;89:197–214] are also considered. These three methods are compared in terms of their effectiveness in outlier detection and their robustness through simulations, considering several contamination scenarios and growing contamination levels. Results indicate that the adapted PSC procedure is able to detect a high percentage of true outliers and a small number of false outliers. It is appropriate when the contamination is in the error term or in the covariates, detecting also possibly masked high leverage points. Moreover, in simulations the final robust regression estimator preserved good efficiency under Normality while keeping good robustness properties.

Keywords: linear models with fixed effects; masking effect; outlier detection; principal sensitivity components; robust estimation

1. Introduction

Linear regression models are widely used in many fields of science. Fitting of these models is commonly done by least squares (LS), due to the simplicity of the idea of minimizing the sum of squared residuals and the interpretability of the final model parameter estimates. However, it is also well known that outliers, considered here as heterogeneous observations in comparison with the bulk of the data, might strongly affect these estimators. Robust estimation methods downweight observations with extreme residuals to provide an estimator that is less affected by these extreme values. On the other hand, diagnostic methods try to identify the outliers. Outlier detection is an important issue because singular observations might hide possibly relevant phenomena affecting our measurements. These outliers are typically pointed out using scaled residuals obtained from a previous model fit. However, both the scale and the previous fit used to obtain residuals might be also affected by the outliers unless they come from an initial robust fit. Thus, outlier detection and robust fitting are very related issues.

This paper deals with grouped data, where the groups might be socioeconomic population subgroups, geographical regions, strata used in the sampling scheme or, more generally, the

*Corresponding author. Email: isabel.molina@uc3m.es

levels of a categorical variable that is related with the outcome of interest. Assuming that the majority of the data in each group follow a linear regression model with fixed group effects, methods for outlier detection and robust estimation are studied. General robust fitting methods for linear regression are reviewed and their suitability to the grouped data structure is discussed. Then three methods that are appropriate for this setup are presented and compared in simulation studies, in terms of their performance in outlier detection and their robustness. The first method is an adaptation of the principal sensitivity components (PSC) method of Peña and Yohai [1] to the grouped data structure. The other two are particularizations of general methods designed to fit models with continuous and categorical variables, concretely the RDL_1 method of Hubert and Rousseeuw [2] and the M-S estimator of Maronna and Yohai.[3]

The work is organized as follows. Section 2 describes the data structure and the model with fixed group effects. Section 3 reviews general robust approaches for linear mixed models, discussing their potential applicability to deal with outliers under grouped data. Section 4 introduces the PSC method of Peña and Yohai [1] and describes its adaptation to the model with fixed group effects. Sections 5 and 6 particularize respectively the RDL_1 method of Hubert and Rousseeuw [2] and the M-S estimator of Maronna and Yohai [3] to the mentioned model. The results of a Monte Carlo simulation study are reported in Section 7. An application is included in Section 8 and finally, concluding remarks are given in Section 9.

2. Linear regression model with group effects

Let $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ be a vector of continuous auxiliary variables (also called covariates) related to the study variable (also called outcome) Y, with $p \geq 1$. Consider that there are n sample observations of X and Y coming from D different population groups of sizes $n_1, \ldots, n_D$ with $n_d \geq 2, d = 1, \ldots, D$, where the overall sample size is $n = \sum_{d=1}^{D} n_d$. These groups are defined by the categories of a categorical variable or by the crossings of the categories of several variables. We consider that the number of groups D is fixed, i.e., it does not grow with n. Let y_{dj} be the value of the study variable Y for jth sample unit from dth group and $\mathbf{x}_{dj} = (x_{dj1}, \ldots, x_{djp})^{\mathrm{T}}$ the vector with the values of the p covariates for the same unit. In absence of outliers, we assume that sample observations follow the linear regression model

$$y_{dj} = \mathbf{x}_{dj}^{\mathrm{T}}\boldsymbol{\beta} + \alpha_d + \varepsilon_{dj}, \quad j = 1, \ldots, n_d, \;\; d = 1, \ldots, D, \tag{1}$$

where α_d is the effect of dth group, assumed to be fixed, and ε_{dj} is the model error, satisfying the usual assumptions

$$\varepsilon_{dj} \sim \text{iid } N(0, \sigma^2), \quad j = 1, \ldots, n_d, \;\; d = 1, \ldots, D, \tag{2}$$

where $\sigma^2 > 0$ is unknown. Defining the vectors $\mathbf{y}_d = (y_{d1}, \ldots, y_{dn_d})^{\mathrm{T}}$ and $\boldsymbol{\varepsilon}_d = (\varepsilon_{d1}, \ldots, \varepsilon_{dn_d})^{\mathrm{T}}$ and the matrix $\mathbf{X}_d = (\mathbf{x}_{d1}, \ldots, \mathbf{x}_{dn_d})^{\mathrm{T}}$, the model can be expressed as

$$\mathbf{y}_d = \mathbf{X}_d\boldsymbol{\beta} + \alpha_d\mathbf{1}_{n_d} + \boldsymbol{\varepsilon}_d, \quad d = 1, \ldots, D,$$

where $\mathbf{1}_{n_d}$ denotes a vector of ones of size n_d. Here, $\boldsymbol{\varepsilon}_d \sim N(\mathbf{0}_{n_d}, \sigma^2 I_{n_d})$, where $\mathbf{0}_{n_d}$ is the zero vector of size n_d and I_{n_d} is the $n_d \times n_d$ identity matrix.

The LS estimators of $\boldsymbol{\beta}$ and α_d, $d = 1, \ldots, D$, are given by

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_X^{-1}\mathbf{s}_{XY}, \quad \hat{\alpha}_d = \bar{y}_d - \bar{\mathbf{x}}_d^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \;\; d = 1, \ldots, D, \tag{3}$$

where $\bar{\mathbf{x}}_d = (\bar{x}_{d1}, \ldots, \bar{x}_{dp})^{\mathrm{T}}$ with $\bar{x}_{dq}$ the mean of the qth auxiliary variable X_q within group d, for $q = 1, \ldots, p$, $\bar{y}_d = n_d^{-1} \sum_{j=1}^{n_d} y_{dj}$, $d = 1, \ldots, D$,

$$\mathbf{S}_X = \frac{1}{n} \sum_{d=1}^{D} \sum_{j=1}^{n_d} (\mathbf{x}_{dj} - \bar{\mathbf{x}}_d)(\mathbf{x}_{dj} - \bar{\mathbf{x}}_d)^{\mathrm{T}}, \quad \mathbf{s}_{XY} = \frac{1}{n} \sum_{d=1}^{D} \sum_{j=1}^{n_d} (\mathbf{x}_{dj} - \bar{\mathbf{x}}_d)(y_{dj} - \bar{y}_d).$$

Predicted values are given by

$$\hat{y}_{dj} = \mathbf{x}_{dj}^{\mathrm{T}} \hat{\boldsymbol{\beta}} + \hat{\alpha}_d, \quad j = 1, \ldots, n_d, \ \ d = 1, \ldots, D.$$

The vector of predicted values for group d, $\hat{\mathbf{y}}_d = \mathbf{X}_d \hat{\boldsymbol{\beta}} + \hat{\alpha}_d \mathbf{1}_d$, can be expressed as a linear combination of the outcome vectors for each group as $\hat{\mathbf{y}}_d = \sum_{\ell=1}^{D} \mathbf{H}_{d\ell} \mathbf{y}_\ell$, where

$$\mathbf{H}_{d\ell} = n_d^{-1} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^{\mathrm{T}} I(d = \ell) + (\mathbf{X}_d - \mathbf{1}_{n_d} \bar{\mathbf{x}}_d^{\mathrm{T}})(n\mathbf{S}_X)^{-1} (\mathbf{X}_\ell^{\mathrm{T}} - \bar{\mathbf{x}}_\ell \mathbf{1}_{n_\ell}^{\mathrm{T}}), \quad d, \ell = 1, \ldots, D.$$

Here, $I(d = \ell)$ denotes the indicator taking value 1 when $d = \ell$ and 0 otherwise. We define the hat matrix associated with dth group as $\mathbf{H}_{dd} = (h_{jk}^d)_{j,k=1,\ldots,n_d} = \partial \hat{\mathbf{y}}_d / \partial \mathbf{y}_d^{\mathrm{T}}$. The matrix $\mathbf{H}_{dd}$ is symmetric but not idempotent. The element (j, k) of this matrix measures the effect that an infinitesimal change in the outcome of kth observation from group d has on the predicted values of jth observation from that same group. The leverage effect of jth observation from group d is given by

$$h_{jj}^d = n_d^{-1} + (\mathbf{x}_{dj} - \bar{\mathbf{x}}_d)^{\mathrm{T}} (n\mathbf{S}_X)^{-1} (\mathbf{x}_{dj} - \bar{\mathbf{x}}_d), \quad d = 1, \ldots, D. \tag{4}$$

This indicates that observations in smaller groups have larger leverage effects than observations in larger groups, when keeping the values of the covariates the same.

In this paper, we assume that the model (1)–(2) holds for at least half of the data in each of the groups. Then, an outlier is an observation that does not follow the assumed model (1)–(2). Outliers are typically identified as the points with poor fit by means of residuals,

$$e_{dj} = y_{dj} - \hat{y}_{dj}, \quad j = 1, \ldots, n_d, \ \ d = 1, \ldots, D$$

after an appropriate scaling. To scale these residuals, a not necessarily very efficient but very robust estimator of the scale (in the sense of high breakdown point) is recommended. Still, outliers with high leverage effect are difficult to detect by standard procedures based on residuals. Specifically, several outliers with similar values on the variables can severely affect the final estimates, but these are exactly the ones that are more difficult to detect due to the masking effect.

3. General robust methods for linear models

It is well known that the LS estimators given in Equation (3) are very sensitive to outliers in the sense that a single outlier can have an arbitrarily large effect on the estimate. This means that their finite sample breakdown point (FBP), the proportion of observations that may drive the estimate to infinity, is $1/n$; their asymptotic breakdown point (ABP) is zero (see, e.g. [4]). Moreover, if we try to use a classical outlier detection procedure to find the outliers, delete them and maybe find a robust estimator based on the clean data, we might end up deleting a full group and this would prevent the estimation of the corresponding group effect. Thus, straightforward application of general outlier detection methods for linear models is not appropriate in this case.

Among robust fitting methods for linear models, we first find M estimation.[5] This method is designed to deal with outliers in the outcome and is based on minimizing a function ρ of

scaled residuals, where ρ is an even function with unique minimum at zero, and the scale can be a previous scale estimator or a simultaneous M scale estimator. However, again the ABP of M estimators is zero under the presence of extreme values in the covariates (or 'high leverage points'). Generalized M (GM) estimators [6] bound the influence of leverage points by solving a weighted version of the M estimating equations, in which leverage points are downweighted. Still, their ABP under the group effects model would be at most $1/(p + D + 1)$ [7] and simulation results indicate that reaching the maximum ABP is unlikely.

The least median of squares (LMS) method [8] minimizes the median of squared residuals. This method achieves a 50% ABP but has very low efficiency in absence of outliers, since its convergence rate is $n^{-1/3}$ instead of the usual $n^{-1/2}$ of other methods. Moreover, the LMS loss function is very nonsmooth, exhibiting many local minima. Since it is not differentiable, gradient methods cannot be applied. Stromberg [9] provided an exact algorithm that in the setup of this paper would be of order $C_{n,p+D+1}$, the number of combinations of $p + D + 1$ elements out of n. This is feasible only for small values of p, D and n. Approximate algorithms based on subsampling are also computationally expensive, and in the case of the grouped data, the subsamples are likely to be collinear yielding singular matrices.

Alternatively, least trimmed squares (LTS) minimizes the sum of trimmed squared residuals, in which a fraction of the largest absolute residuals are trimmed. The fraction can be chosen to achieve maximum breakdown point (BP) and their convergence rate remains to be $n^{-1/2}$. However, again the exact algorithm for this method is computationally very expensive whereas the methods based on subsampling procedures might produce singular matrices.[2]

Another method with high BP and fully efficient is the weighted likelihood estimator (WLE) of Agostinelli and Markatou.[10] This estimator solves a weighted likelihood equation with smaller weights for observations at which the kernel density estimator is far from the smoothed assumed density (large Pearson residual). Markatou et al. [11] developed a method that avoids the need for an initial high BP estimator, by searching for all the possible solutions using a bootstrap root search. The WLE method is not designed to deal with a grouped data structure and in our simulations the R function `wle.lm()` of the R library `wle` many times failed to give a solution.

Similarly, MM estimators have high BP and high efficiency.[12] They start with an initial consistent estimate with high BP but possibly low normal efficiency. Then a robust scale based on residuals from that estimate is obtained. Finally, an M estimate with bounded loss function ρ and using the preliminary scale estimator is computed. Unfortunately, MM methods need an initial high BP estimator such as the LMS or the LTS estimators, which have the already mentioned problems under the grouped data structure.

Thus, none of the mentioned methods are appropriate for grouped data and specific methods are needed under this situation.

Let $\boldsymbol{\gamma} = (\boldsymbol{\beta}^{\mathrm{T}}, \alpha_1, \ldots, \alpha_D)^{\mathrm{T}}$ denote the vector of regression parameters in model (1). Under this model, an S estimator [7] of $\boldsymbol{\gamma}$ is obtained by minimizing $s(e_{11}(\boldsymbol{\gamma}), \ldots, e_{Dn_D}(\boldsymbol{\gamma}))$, where s is an M scale estimator. Again, solving this problem exactly is usually too computationally demanding and, in practice, approximate S estimators are obtained by minimizing for $\boldsymbol{\gamma} \in A$, where $A = \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_N\}$ is a finite set of candidates. A method that pre-selects the candidates as those that are more likely to be the minimum can fasten considerably the algorithm. This is the basic idea of the groupwise principal sensitivity components (GPSC) algorithm introduced in Section 4.2. Unfortunately, it is known that S estimators cannot have both high BP and high normal efficiency. However, their efficiency can be improved without decreasing their BP by the one-step reweighting procedure proposed by Rousseeuw and Leroy.[13] Starting with a high BP estimate $\hat{\gamma}_0$ and a robust scale of residuals based on this estimate $s = s(e_{11}(\hat{\boldsymbol{\gamma}}_0), \ldots, e_{Dn_D}(\hat{\boldsymbol{\gamma}}_0))$, a new estimator is obtained by weighted LS with weights $w_{dj} = W(e_{dj}(\hat{\boldsymbol{\gamma}}_0)/s)$, where $W(t)$ is a decreasing function of $|t|$ such as the 'hard rejection' function $W(t) = I(|t| \leq k)$ with k equal to a quantile of the distribution of $|t|$. Under normality, this is equivalent to discarding observations

with large absolute standardized residuals and then obtain the LS estimator with the remaining ones. He and Portnoy [14] showed that this reweighting process preserves the order of consistency of the initial estimator $\hat{\boldsymbol{\gamma}}_0$. The GPSC procedure introduced in Section 4.2 leads to an approximate S estimator followed by a reweighting step.

The M-S method of Maronna and Yohai [3] described in Section 6 makes profit of the efficiency of M estimators and the high BP of S estimators. The RDL_1 of Hubert and Rousseeuw [2] described in Section 5 is a weighted L_1 estimator with small weights for high leverage points. The properties of these three methods will be compared under finite sample sizes. Another method that is suitable for the grouped data structure is the PROGROUP algorithm of Hubert and Rousseeuw.[4] This method is somehow similar to the GPSC procedure in that directions of maximum outlyingness are searched for each group. Then, these directions are used to find potentially clean subsets of data that help to approximate the LMS estimate. However, in our simulations, this method appeared to be very robust but its efficiency was very low in comparison with the other considered methods. At the same time, it was computationally much slower, so we decided not to include its results in Section 7 because we found it not competitive with the other methods.

4. Groupwise principal sensitivity components

4.1. *The PSC method*

Peña and Yohai [1] proposed a fast robust fitting procedure for a linear regression model called here PSC method. This method consists of two stages. In the first stage, an approximate S estimator is obtained by minimizing a robust scale of residuals. In the second stage, the efficiency of this estimator is improved by a kind of reweighting procedure based on robust *t* tests.

To obtain the approximate S estimator in Stage 1, the minimization over all possible values of the regression parameter is reduced to a finite set of candidate estimates. These candidates are obtained by LS fits to subsets of data that are potentially clean of low and high leverage outliers, including masked outliers. This stage is iterative. Each iteration (except for the first one which omits this step), starts by pointing out and deleting observations with extreme residuals according to a robust scale (low leverage outliers). The sensitivity vector associated with each observation is defined as the vector of changes in the predicted value of this observation when each of the data points is deleted. The PSC are simply the principal components of these sensitivity vectors. High leverage outliers are expected to appear with extreme coordinates in at least one of these PSC, see Theorem 1 in [1, p. 438]. Consider a LS estimator obtained by deleting the 50% of the points with most extreme coordinates in one of the PSC. This will be one of the candidate estimates of the regression parameter. Deletion of the 50% of the points with more extreme values in each component leads to a different estimate of the regression parameter. Each of these estimates has a set of residuals attached and the one with the minimum value of the robust scale of residuals is selected.

4.2. *The adapted PSC method*

The PSC method described above cannot be directly applied to model (1) because subsets of 50% of the observations might exclude some of the groups. Intuitively, a small group with a larger mean is likely to be fully discarded because the observations in that group might be considered as outliers in comparison with the rest of observations. In fact, since observations in smaller groups tend to have higher leverage, these small groups will be more likely to be fully discarded. We indeed experienced this problem in the simulation experiments described in Section 7.

Here we propose an adaptation of this method, in which subsets are selected for each group separately based on GPSC. Sensitivity vectors are defined for each group and the directions of maximum variability of these sensitivity vectors are computed for each group. Group specific PSC are more likely to point out to outliers within the groups. Besides, no more than 50% of data points of the same group can be discarded. Moreover, the procedure gives a large set of candidate estimates of the regression parameter. Minimization of a robust scale of residuals with respect to a larger set of candidate estimates makes it more likely to select an estimate that is based on an initial clean subset, which would lead to a final robust estimator.

Let $\hat{y}_{dj(dk)}$ be the predicted value of y_{dj} when observation $(y_{dk}, \mathbf{x}_{dk}^{\mathrm{T}})$ is deleted, that is

$$\hat{y}_{dj(dk)} = \mathbf{x}_{dj}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{(dk)} + \hat{\alpha}_{d(dk)}, \tag{5}$$

where $\hat{\boldsymbol{\beta}}_{(dk)}$ and $\hat{\alpha}_{d(dk)}$ denote respectively the LS estimates of $\boldsymbol{\beta}$ and α_d obtained exactly the same as in Equation (3), using all observations from all the groups except for $(y_{dk}, \mathbf{x}_{dk}^{\mathrm{T}})$. Similarly as in [1] but restricted to group d, for each observation y_{dj} within that group, we define the vector of changes in the predicted value when each data point from group d is eliminated, i.e.

$$\mathbf{r}_{dj} = (\hat{y}_{dj} - \hat{y}_{dj(d1)}, \ldots, \hat{y}_{dj} - \hat{y}_{dj(dn_d)})^{\mathrm{T}}.$$

The vectors $\mathbf{r}_{dj}$, $j = 1, \ldots, n_d$, are called sensitivity vectors of group d. Next, we define the sensitivity matrix $\mathbf{R}_d$ for dth group as the matrix with the sensitivity vectors of group d in the rows, i.e.

$$\mathbf{R}_d = \begin{pmatrix} \hat{y}_{d1} - \hat{y}_{d1(d1)} & \cdots & \hat{y}_{d1} - \hat{y}_{d1(dn_d)} \\ \vdots & \ddots & \vdots \\ \hat{y}_{dn_d} - \hat{y}_{dn_d(d1)} & \cdots & \hat{y}_{dn_d} - \hat{y}_{dn_d(dn_d)} \end{pmatrix} = \begin{pmatrix} \mathbf{r}_{d1} \\ \mathbf{r}_{d2} \\ \vdots \\ \mathbf{r}_{dn_d} \end{pmatrix} \tag{6}$$

It is easy to see that the elements of this matrix can be obtained from the leverages and the residuals of the LS fit as

$$\hat{y}_{dj} - \hat{y}_{dj(dk)} = \frac{h_{jk}^d \, e_{dk}}{1 - h_{kk}^d}, \tag{7}$$

avoiding to do n_d different fits. This matrix can be expressed as $\mathbf{R}_d = \mathbf{H}_{dd}\mathbf{W}_d$, where $\mathbf{W}_d = \mathrm{diag}_{1 \le j \le n_d}\{e_{dj}/(1 - h_{jj}^d)\}$. The matrix $\mathbf{R}_d$ has rank $p + 1$, which means that the sensitivity vectors for group d lie in a subspace of dimension $p + 1$. Then, similarly as in [1], we summarize their information by choosing an appropriate basis on this subspace and projecting them over this basis. The first vector $\mathbf{v}_1^d$ of this basis is chosen as the direction of maximum sensitivity, that is, the direction $\mathbf{v}$ that is solution of the problem

$$\max_{\|\mathbf{v}\|=1} \sum_{j=1}^{n_d} (\mathbf{v}^{\mathrm{T}} \mathbf{r}_{dj})^2.$$

Thus, our chosen basis is the set of eigenvectors $\{\mathbf{v}_q^d, q = 1, \ldots, p + 1\}$ associated with the non null eigenvalues of matrix $\mathbf{M}_d = \mathbf{R}_d^{\mathrm{T}}\mathbf{R}_d$. The maximum eigenvalue of $\mathbf{M}_d$, denoted λ_1^d, can be interpreted as a measure of global effect of the observations of dth group on the predicted values of the observations in that group. The eigenvector $\mathbf{v}_1^d$ associated with λ_1^d is the direction of maximum sensitivity of observations in dth group. Observe that $\{\mathbf{v}_q^d, q = 1, \ldots, p + 1\}$ are the orthogonal directions in which the joint effect of deleting several data points from group d in the predicted values is maximized. Also, note that since $\mathbf{W}_d$ is diagonal, the eigenvectors of $\mathbf{M}_d$ are the same as those of matrix $\mathbf{H}_{dd}\mathbf{H}_{dd}$, and also the same as those of the leverage matrix $\mathbf{H}_{dd}$. Thus, we can

use the projections $\mathbf{z}_q^d = \mathbf{R}_d \mathbf{v}_q^d$ on the directions $\mathbf{v}_q^d$, $q = 1, \ldots, p+1$, to detect the points that have a high joint effect within dth group. These projections are the principal components of the sensitivity vectors. As in [1], groups of points that jointly have a leverage effect within group d are expected to have extreme coordinates in at least one of the $p+1$ PSC $\{\mathbf{z}_q^d;\ q = 1, \ldots, p+1\}$.

As described in Section 4.3, selection of data subsets that are clean of low and high leverage points will lead to find an approximate S estimator in a computationally faster manner. The efficiency of this estimator will then be improved in a second stage by a one-step reweighting scheme based on robust testing for outlyingness of each potential outlier.

4.3. *The adapted robust fitting algorithm*

The GPSC procedure described above allows us to select subsets of data that are free of high leverage outliers. This procedure can be integrated in the following iterative algorithm that discards both high and low leverage outliers in each of the D groups and provides an approximate S estimator of the regression parameter $\boldsymbol{\gamma} = (\boldsymbol{\beta}^{\mathrm{T}}, \alpha_1, \ldots, \alpha_D)^{\mathrm{T}}$ under model (1):

STAGE 1. The first iteration, $r = 1$, starts by constructing a set A_1 of candidate estimates of $\boldsymbol{\gamma}$ as follows: Obtain the sensitivity matrix $\mathbf{R}_d$ using Equation (7) and compute its PSC $\mathbf{z}_q^d$, $q = 1, \ldots, p+1$ for each group $d = 1, \ldots, D$. Now, for each component q, construct different data sets as follows. Look at each group d and consider two different data sets from that group; in the first set include all observations from the group and in the second, delete the 50% of the observations with largest coordinates in the vector $\mathbf{d}_q^d = |\mathbf{z}_q^d - \mathrm{median}(\mathbf{z}_q^d)|$. Combining the two data sets from each of the D groups we have 2^D full samples. We do this for each of the components $q = 1, \ldots, p+1$, obtaining $2^D(p+1)$ samples. Some times the outliers appear spread in the extremes of different PSCs and therefore we cannot detect all of them at a time by looking at the 50% more extreme points in only one of the components. For this reason, another small but potentially clean data set was constructed as follows. For each observation from a given group, identify at which components this observation appears to be in the set of 50% more extreme points. In order to weight by the importance of each component (measured by the corresponding eigenvalue), we calculated the sum of eigenvalues of the components in which the observation appears in the set of the 50% more extreme points. The new data set was the $1/4$ of the observations in each group with smallest sum of eigenvalues. Thus, we have a total of $2^D(p+1)+1$ potentially clean samples. A much smaller set of samples which does not depend on D can be obtained by applying Remark 1 . Application of Remark 1 makes the algorithm considerably faster specially for D large. Then using each of these full samples, compute the LS estimators. The LS estimates obtained from each of these samples compose the set of candidate estimates A_1. For each candidate $\boldsymbol{\gamma} = (\boldsymbol{\beta}^{\mathrm{T}}, \alpha_1, \ldots, \alpha_D)^{\mathrm{T}}$, obtain residuals

$$e_{dj}(\boldsymbol{\gamma}) = y_{dj} - \mathbf{x}_{dj}^{\mathrm{T}}\boldsymbol{\beta} - \alpha_d, \quad j = 1, \ldots, n_d, \ \ d = 1, \ldots, D.$$

Then select the estimate $\boldsymbol{\gamma}^{(1)}$ satisfying

$$\boldsymbol{\gamma}^{(1)} = \underset{\boldsymbol{\gamma} \in A_1}{\mathrm{argmin}} \quad s(e_{11}(\boldsymbol{\gamma}), \ldots, e_{Dn_D}(\boldsymbol{\gamma})), \tag{8}$$

where s is an M scale estimator with high breakdown point such as the median absolute deviation (MAD). Let $\boldsymbol{\gamma}^{(r)} = ((\boldsymbol{\beta}^{(r)})^{\mathrm{T}}, \alpha_1^{(r)}, \ldots, \alpha_D^{(r)})^{\mathrm{T}}$ be the estimator obtained by minimizing the robust scale in iteration r. In iteration $r+1$, obtain the set of residuals associated with $\boldsymbol{\gamma}^{(r)}$,

$$e_{dj}^{(r+1)} = e_{dj}(\boldsymbol{\gamma}^{(r)}) = y_{dj} - \mathbf{x}_{dj}^{\mathrm{T}}\boldsymbol{\beta}^{(r)} - \alpha_d^{(r)}, \quad j = 1, \ldots, n_d, \ \ d = 1, \ldots, D,$$

and let $s_d^{(r+1)} = s(e_{d1}^{(r+1)}, \ldots, e_{dn_d}^{(r+1)})^{\mathrm{T}}$ be a robust scale for dth group such as the normalized MAD (NMAD). Then eliminate all observations with $|e_{dj}^{(r+1)}| \geq C_1 \cdot s_d^{(r+1)}$ where C_1 is a constant,

$d = 1, \ldots, D$. With all the remaining observations from the D groups, obtain the LS estimators as in Equation (3) and compute again the PSC. Construct the set A_{r+1} with the new set of candidate estimates $\boldsymbol{\gamma}$ exactly as described before, but include in the set also the estimator obtained in previous iteration $\boldsymbol{\gamma}^{(r)}$. The iterations end when $\boldsymbol{\gamma}^{(r+1)} = \boldsymbol{\gamma}^{(r)}$ and then, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^{(r+1)} = (\boldsymbol{\beta}^{*T}, \alpha_1^*, \ldots, \alpha_D^*)^{\mathrm{T}}$ is the preliminary robust estimator, which is an approximate S estimator. This preliminary robust estimator is obtained from a possibly clean subset of data points, in which potential outliers have been discarded. To improve the efficiency of this estimator, in Stage 2 each of these potential outliers is tested using a robust version of the t test that uses only the set of clean data points. Observations that are not rejected by this test are used to find the final estimator.

STAGE 2. Compute residuals from the preliminary robust estimator,

$$e_{dj}^* = e_{dj}(\boldsymbol{\gamma}^*) = y_{dj} - \mathbf{x}_{dj}^{\mathrm{T}}\boldsymbol{\beta}^* - \alpha_d^*, \quad j = 1, \ldots, n_d, \;\; d = 1, \ldots, D,$$

and let $s_d^* = s(e_{d1}^*, \ldots, e_{dn_d}^*)$ be a robust scale for dth group such as the NMAD. Delete the observations with $|e_{dj}^*| > C_2 \cdot s_d^*$, where C_2 is a constant, for $d = 1, \ldots, D$. Let n^* be the total number of deleted observations. With the remaining $n - n^*$ observations, compute the LS estimators as given in Equation (3) and denote them by $\tilde{\boldsymbol{\beta}}$ and $\tilde{\alpha}_d$, $d = 1, \ldots, D$. Compute also the standard error $\tilde{\sigma}$ using the residuals of these remaining observations and the corresponding leverages $\tilde{h}_{jj}^d$. Then, test the outlyingness of each of these n^* elements by using the robust t test statistic

$$t_{dj} = \frac{y_{dj} - \mathbf{x}_{dj}^{\mathrm{T}}\tilde{\boldsymbol{\beta}} - \tilde{\alpha}_d}{\tilde{\sigma}\sqrt{1 + \tilde{h}_{jj}^d}} \tag{9}$$

Each of the n^* observations is finally eliminated only if $|t_{dj}| > C_3$, where C_3 is a constant. The remaining observations are used to calculate the final LS estimator, denoted $\hat{\boldsymbol{\gamma}}^* = (\hat{\boldsymbol{\beta}}^{*T}, \hat{\alpha}_1^*, \ldots, \hat{\alpha}_D^*)^{\mathrm{T}}$.

This GPSC procedure detects individual outliers in the mean and gives a robust estimator of the vector of regression parameters with respect to this type of outliers. The final estimate could be used afterwards to detect outlying groups in the variance. This could be done easily by calculating residuals from the robust estimate, computing again robust group-specific scales s_d and then carrying a robust homogeneity test for the null hypothesis that all group-specific variances are equal.

As in any other robust procedure, there is always a trade off between robustness and efficiency. This trade off is controlled by the constants C_1, C_2 and C_3, with smaller values of these constants providing more robustness whereas larger values implying more efficiency. Since Stage 1 is meant to provide a high breakdown point estimator, it is convenient to choose a small value for C_1 in Stage 1. Since Stage 2 is more focused in achieving efficiency under absence of outliers, then constants C_2 and C_3 should be larger. By our experience obtained from several simulation studies, constants that seem to keep a good trade off between robustness and efficiency are $C_1 = 2$ and $C_2 = C_3 = 3$.

The procedure in Stage 1 gives an approximate S estimator. If the function ρ defining the M scale estimator in Equation (8) satisfies conditions (R1)–(R3) in [7], then the S estimator can achieve a high BP. Under model (1)–(2) with the grouped data structure, S estimators with a bounded ρ function achieve at least the maximum FBP of equivariant estimators, given by $[(\min(n_d) - 1)/2]/n$ where $[\cdot]$ denotes the integer part, see Section 5.6.1 of Maronna.[15] In addition, S estimators are asymptotically normally distributed with asymptotic covariance matrix given in [7], provided the extra regularity conditions in Theorem 3 of that paper hold. The final one-step reweighted estimator obtained from Stage 2 preserves the asymptotic normality and its asymptotic covariance matrix can be calculated by Taylor expansions.[16]

Remark 1 One way of speeding up considerably the GPSC fitting algorithm, specially for large D, is the following. In Stage 1, after computing the $p+1$ PSCs $\mathbf{z}_q^d$, $q=1,\ldots,p+1$, for each group d, instead of considering the two data sets obtained by deleting 0% and 50% of observations with largest coordinates in $\mathbf{d}_q^d$ within group d, we can consider only the data set obtained by deleting the 50% of observations with largest coordinates in $\mathbf{d}_q^d$. This would be done for each component $q=1,\ldots,p+1$. In this case, the number of candidate estimates A_1 does not depend on the number of groups D and the algorithm becomes much faster. Forcing the deletion of 50% of observations could in principle affect the efficiency of the algorithm, but Stage 2 then improves the estimator by returning to the sample the observations that are not really outliers. Simulation studies in Section 7 indicate that this faster version preserves similar finite sample properties as the original GPSC algorithm.

Remark 2 In Stage 1, it is necessary to compute the eigenvectors of matrix $\mathbf{M}_d$ of size $n_d \times n_d$. For groups d with $n_d > p+D$, this can be replaced by computing the eigenvectors of a $(p+D)\times(p+D)$ matrix. For this, define the matrices

$$\mathbf{X}=\begin{pmatrix}\mathbf{X}_1\\ \vdots\\ \mathbf{X}_D\end{pmatrix},\quad \mathbf{Z}=\operatorname{diag}(\mathbf{1}_{n_1},\ldots,\mathbf{1}_{n_D}),\quad \mathbf{X}^*=[\mathbf{X}|\mathbf{Z}]=\begin{pmatrix}\mathbf{X}_1^*\\ \vdots\\ \mathbf{X}_D^*\end{pmatrix}. \tag{10}$$

It can be seen that $\mathbf{M}_d=\boldsymbol{\Gamma}_d\boldsymbol{\Gamma}_d^{\mathrm{T}}$, where

$$\boldsymbol{\Gamma}_d=\mathbf{W}_d\mathbf{X}_d^*((\mathbf{X}^*)^{\mathrm{T}}\mathbf{X}^*)^{-1}((\mathbf{X}_d^*)^{\mathrm{T}}\mathbf{X}_d^*)^{1/2}.$$

Define now the $(p+D)\times(p+D)$ matrix $\mathbf{Q}_d=\boldsymbol{\Gamma}_d^{\mathrm{T}}\boldsymbol{\Gamma}_d$. For a group d with $n_d>p+D$, it is faster to compute eigenvalues and eigenvectors of matrix $\mathbf{Q}_d$. Consider an eigenvector $\mathbf{u}_k$ of matrix $\mathbf{Q}_d=\boldsymbol{\Gamma}_d^{\mathrm{T}}\boldsymbol{\Gamma}_d$ associated with eigenvalue λ_k. Then, the eigenvector of $\mathbf{M}_d=\boldsymbol{\Gamma}_d\boldsymbol{\Gamma}_d^{\mathrm{T}}$ associated with the same eigenvalue λ_k is equal to $\mathbf{v}_k=\boldsymbol{\Gamma}_d\mathbf{u}_k$. Then, the principal sensitivity component associated with $\mathbf{v}_k$ is the projection of the rows of $\mathbf{R}_d$ on $\mathbf{v}_k$, which is equal to

$$\mathbf{z}_k=\mathbf{R}_d\mathbf{v}_k=\mathbf{R}_d\boldsymbol{\Gamma}_d\mathbf{u}_k=\lambda_k\mathbf{X}_d^*((\mathbf{X}_d^*)^{\mathrm{T}}\mathbf{X}_d^*)^{-1/2}\mathbf{u}_k.$$

Remark 3 The final estimator $\hat{\boldsymbol{\gamma}}^*=\hat{\boldsymbol{\gamma}}^*(\mathbf{X}^*,\mathbf{y})$ obtained from Stage 2 is regression and scale equivariant, that is, if we transform $\mathbf{y}$ by $\lambda\mathbf{y}+\mathbf{X}^*\boldsymbol{\delta}$, where $\lambda\in I\!R$ and $\boldsymbol{\delta}\in I\!R^{p+D}$, then

$$\hat{\gamma}^*(\mathbf{X}^*,\lambda\mathbf{y}+\mathbf{X}^*\boldsymbol{\delta})=\lambda\hat{\gamma}^*(\mathbf{X}^*,\mathbf{y})+\boldsymbol{\delta}.$$

It is also affine equivariant when transforming the matrix of covariates $\mathbf{X}$ by $\mathbf{XA}$, where $\mathbf{A}$ is a nonsingular $p\times p$ matrix.

Remark 4 In this paper we consider that $n_d\to\infty$, $d=1,\ldots,D$, but $n_d/n\to\lambda\in(0,1)$, $d=1,\ldots,D$, whereas $D=O(1)$ as $n\to\infty$. This allows group effects to be estimated efficiently. Under this setup, the complexity of each iteration in Stage 1 of the algorithm is polynomial. It is difficult to find out the maximum number of iterations of this Stage, but in our simulations the procedure needed only two or three iterations most of the times and never more than eight iterations. The computational complexity of Stage 2 is also polynomial.

5. RDL_1 method

Hubert and Rousseeuw [2] proposed the RDL_1 method to find a robust regression estimator in linear models that include categorical variables. This method consists of using a robust distance to

downweight high leverage points, and then using those weights to obtain a weighted L_1 regression estimator. The method, particularized to model (1)–(2), proceeds as follows:

(1) First, search for high leverage points in the set $\mathcal{X} = \{\mathbf{x}_{dj},\ j = 1, \ldots, n_d,\ d = 1, \ldots, D\}$, by computing the minimum volume ellipsoid of Rousseeuw.[17] The idea is to consider all ellipsoids of approximately 50% of the observations and then select the one with smallest volume. The mean vector and the covariance matrix of that ellipsoid are considered as robust location and scatter matrix, $M(\mathcal{X})$ and $C(\mathcal{X})$, respectively, of the set of data points $\mathcal{X}$. Then, compute the robust distances of each observation to the location as

$$\mathrm{RD}(\mathbf{x}_{dj}) = (\mathbf{x}_{dj} - M(\mathcal{X}))C(\mathcal{X})^{-1}(\mathbf{x}_{dj} - M(\mathcal{X}))^{\mathrm{T}}, \quad j = 1, \ldots, n_d,\ d = 1, \ldots, D.$$

Observations with large robust distances are regarded as high leverage points.

(2) Let us define the vector of group effects $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)^{\mathrm{T}}$. Estimate the regression parameter $\boldsymbol{\gamma} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}$ by a weighted L_1 regression, that is, by solving the problem

$$\min_{\boldsymbol{\gamma}} \sum_{d=1}^{D} \sum_{j=1}^{n_d} w_{dj} |e_{dj}(\boldsymbol{\gamma})|,$$

where the weights are given by $w_{dj} = \min\{1, p/\mathrm{RD}(\mathbf{x}_{dj})\}$, $j = 1, \ldots, n_d$, $d = 1, \ldots, D$.

(3) Let $\hat{\boldsymbol{\gamma}}$ be the estimate obtained by the weighted L_1 regression in Step 2. Following the recommendation of Maronna and Yohai [3] we compute the NMAD of the non null residuals,

$$\hat{\sigma} = 1.4826 \cdot \mathrm{median}\{|e_{dj}(\hat{\boldsymbol{\gamma}})| \quad \text{with } e_{dj}(\hat{\boldsymbol{\gamma}}) \neq 0,\ j = 1, \ldots, n_d,\ d = 1, \ldots, D\}.$$

Then an observation is classified as an outlier if its corresponding absolute standardized residual, $|e_{dj}(\hat{\boldsymbol{\gamma}})/\hat{\sigma}|$, exceeds 2.5.

Hubert and Rousseeuw [2] state that an exact formula of the BP of the RDL_1 procedure is hard to find but by construction, the estimator RDL_1 protects against leverage points by giving them small weights, whereas vertical outliers have only a small effect on the L_1 stage.

6. M-S estimator

Maronna and Yohai [3] proposed an alternating M and S estimator for models that include categorical variables, where an M estimator is used for the vector of parameters of the categorical predictors and an S estimator is used for the parameters of the continuous ones. The particularization of this method to model (1) is defined as follows. Assume first that $\boldsymbol{\beta}$ is known. Then, obtain an M estimator of $\boldsymbol{\alpha}$ as

$$\boldsymbol{\alpha}(\boldsymbol{\beta}) = \operatorname*{argmin}_{\boldsymbol{\alpha}} \sum_{d=1}^{D} \sum_{j=1}^{n_d} \rho(y_{dj} - \mathbf{x}_{dj}^{\mathrm{T}}\boldsymbol{\beta} - \alpha_d), \tag{11}$$

where ρ is an even convex function. Consider the vectors of residuals

$$\mathbf{e}_d(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{y}_d - \mathbf{x}_{dj}^{\mathrm{T}}\boldsymbol{\beta} - \alpha_d \mathbf{1}_{n_d}, \quad d = 1, \ldots, D.$$

Then, the estimator of $\boldsymbol{\beta}$ is obtained by minimizing an M scale S estimator of the residuals obtained using the M estimator $\boldsymbol{\alpha}(\boldsymbol{\beta})$, that is

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} s(\mathbf{e}_1(\boldsymbol{\beta}, \boldsymbol{\alpha}(\boldsymbol{\beta})), \ldots, \mathbf{e}_D(\boldsymbol{\beta}, \boldsymbol{\alpha}(\boldsymbol{\beta}))).$$

Maronna and Yohai [3] introduced also an estimator called M-GM for models with categorical variables. This estimator is a weighted L_1 regression estimator similar to RDL_1, but in this case the

weights w_{dj} are function of a measure of the outlyingness of the previously centred data points. In a simulation experiment carried out by these authors, this estimator broke down when the number of continuous predictors was greater than 3, while the M-S estimator resisted. Thus, they recommended the latter for $p > 3$.

M-S estimators are regression and affine equivariant. They are also asymptotically normal. Moreover, under model (1)–(2), if $\min(n_d) < \min(n - p, n - D + 1)$, then the M-S estimator achieves also a FBP of at least $[(\min(n_d) - 1)/2]/n$, see Maronna and Yohai.[3]

7. Monte Carlo simulation experiment

Typically, when sample sizes grow, the effect on the final estimators of a limited number of finite outliers goes to zero. Thus, it seems convenient to study the performance of robust methods under limited sample sizes, which is also a much more realistic setup.

This section reports the results of a simulation experiment designed to compare the outlier detection performance and the robustness of LS and the three robust procedures introduced here, namely GPSC, RDL_1 and M-S methods, under finite group sample sizes. For this, we simulated data trying to imitate a data set from the Australian Agricultural and Grazing Industries Survey (AAGIS) and used in [18,19]. This data set contains several variables measured to 1652 Australian farms. Among these variables, we find the total cash receipts of the farm business over the surveyed year (*income*), the total area of the farm (*hectares*), the area of crops grown on the farm (*crops*), the number of beef cattle on the farm (*beef*) and the number of sheep (*sheep*).

Data corresponding to $D = 5$ groups with a total sample size of $n = 200$ were generated. The group sample sizes were respectively $n_k = 10k + 10$, $k = 1, 2, \ldots, 5$. We considered four covariates. For Scenarios A and B below, the values of the covariates were generated independently from the distributions $X_1 \sim N(3.31, 0.82)$, $X_2 \sim N(1.74, 1.10)$, $X_3 \sim N(1.70, 1.28)$ and $X_4 \sim N(2.41, 1.61)$, where the given means and standard deviations were taken as the sample means and standard deviations of the variables *hectares*, *crops*, *beef* and *sheep*, respectively, of the AAGIS data. In the case of Scenarios C and D, good data points were generated from a bivariate distribution for (X_1, X_2), with the same marginal for X_1 as before but $X_2 = 1 - 0.5X_1 + v$, with $v \sim N(0, 0.7)$. As true values of regression coefficients, we have taken the fitted values of the model to the AAGIS data, given by $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4) = (0.45, 0.14, 0.05, 0.005)$. The fixed effects $\alpha_d, d = 1, \ldots, 10$, were generated from a normal distribution with zero mean and standard deviation $\sigma_\alpha = 1$. The errors ε_{dj} were generated independently from a normal distribution with zero mean and standard deviation equal to $\sigma = 0.1$. Then, keeping the group effects α_d and the values of the covariates fixed, we carried out $L = 1000$ Monte Carlo replicates. In each replicate, we generated the model responses y_{dj} from model (1). Then, we considered four contamination scenarios:

A. *No contamination.*
B. *Vertical outliers*: A subset $\mathcal{D}_c \subseteq \{1, 2, \ldots, D\}$ of the groups was selected for contamination. Within these selected groups $\mathcal{D}_c$, a given percentage of the observations were contaminated only in the outcome. We considered two cases, asymmetric and symmetric contamination. In the symmetric case, for each selected group $d \in \mathcal{D}_c$, half of the contaminated observations were replaced by $c_{d1} = \bar{y}_d + k\ s_{Y,d}$ and the other half by $c_{d2} = \bar{y}_d - k\ s_{Y,d}$ with $k = 4, 5$, where $\bar{y}_d$ and $s_{Y,d}$ are respectively the mean and the standard deviation of the generated clean outcomes in dth group. Note that the outliers are generated taking into account also the slopes since $\bar{y}_d \approx \bar{\mathbf{x}}_d^{\mathrm{T}} \boldsymbol{\beta} + \alpha_d$. In this way, the contaminated observations are clear outliers as compared with the non-outliers. In the asymmetric case, all contaminated observations were replaced by c_{d1}.

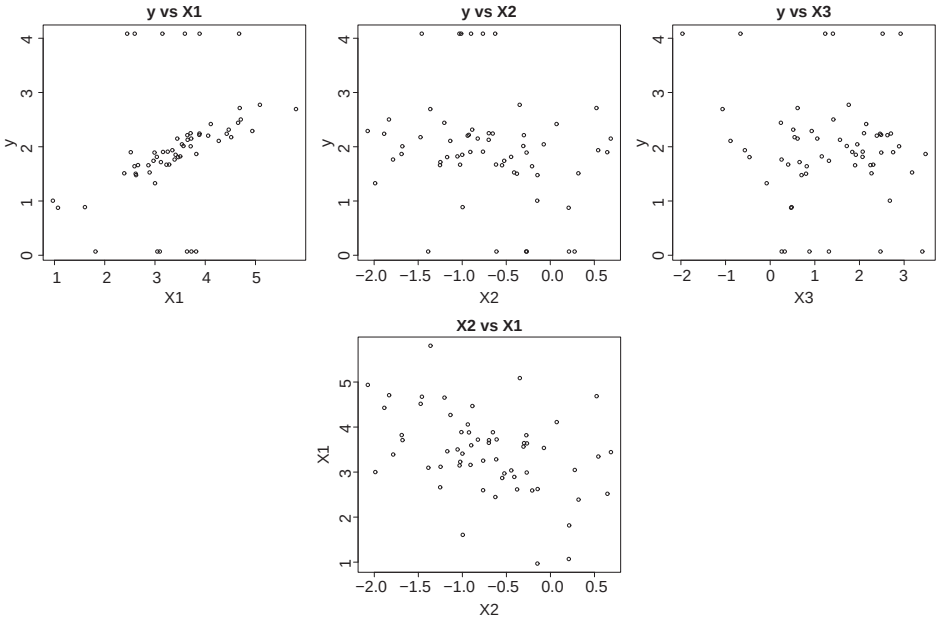

Figure 1. Scatterplots of Y vs. X_1, X_2 and X_3, and finally X_2 vs. X_1 for data with 20% of symmetric contamination type D with $k = 4$ within group $d = 5$.

C. *Good leverage points*: Here extreme values were introduced in the covariate values only. For X_3 and X_4, a percentage of observations was replaced by extreme cases. This was done marginally and similarly as before, that is, setting x_{djq}, for $q = 3, 4$, equal to $c_{d3} = \bar{x}_{dq} + k\, s_{X_q,d}$, where $\bar{x}_{dq}$ and $s_{X_q,d}$ are respectively the mean and standard deviation of the clean data of X_q in dth group and taking $k = 4, 5$. For X_1 and X_2, we considered atypical values in the joint distribution of (X_1, X_2). Note that these are not outliers according to our definition.
D. *High leverage outliers*: Again, extreme cases were created for the covariates in exactly the same way as in Scenario C. Finally, the responses y_{dj} corresponding to the extreme values of covariates were contaminated similarly as described in scenario B.

We selected for contamination three out of the $D = 5$ groups, concretely $\mathcal{D}_c = \{1, 3, 5\}$. Figure 1 shows graphically the data from group $d = 5$ containing $n_d = 60$ observations with 20% of symmetric contamination type D with $k = 4$. The four plots in this figure show respectively, from left to right and from top to bottom, the outcomes of all sample observations against their values in the covariates X_q, for $q = 1, 2, 3$, and the values of X_2 against those of X_1. Note that even if $k = 4$ seems a large value, since $k\, s_{Y,d}$ is added or subtracted to the mean value of the group $\bar{y}_d$ to create the outliers, in fact some of these outliers are rather close to some of the clean data points. Thus, some of them are actually very mild outliers.

Thus, for each Monte Carlo replicate $l = 1, \ldots, L$, we applied five different estimation procedures to fit model (1) to the simulated data: LS, the GPSC as described in Stages 1 and 2 (called GPSC1), the GPSC faster algorithm suggested in Remark 1 (GPSC2), the RDL_1 and the M-S methods. The original PSC algorithm of Peña and Yohai [1] was also applied, but in the case of high percentages of contamination (20% or more), the algorithm deleted either in Stages 1 or 2 all the observations in at least one of the groups in several Monte Carlo replicates. In particular, with 20% of symmetric contamination type B within selected groups, this occurred for six Monte Carlo replicates. Due to this, simulation results for the original PSC method cannot be shown. In

case of smaller groups, groups with more uneven means, larger percentages of contamination or outliers spread in more groups, the probability of deleting a whole group is likely to increase.

Four performance criteria were used to compare the results of the considered estimators. The first two are used to evaluate the outlier detection performance and the other two assess their robustness properties. The first one is the percentage of the Monte Carlo replications in which all outliers were detected (ALLD). The second criterion is the average over the Monte Carlo simulations of the number of false outliers (AFO) found by each of these procedures. The AFO summarizes the swamping effect, which occurs when non-outliers are wrongly identified due to the effect of some hidden outliers, see Lawrence.[20] The third criterion is the empirical mean squared error (MSE) of the final estimator $\hat{\boldsymbol{\gamma}}$ obtained by each of the three procedures, defined as

$$\mathrm{MSE}(\hat{\boldsymbol{\gamma}}) = \frac{1}{L}\sum_{l=1}^{L} \|(\hat{\boldsymbol{\gamma}}^{(l)} - \boldsymbol{\gamma})\|^2. \tag{12}$$

Finally, the fourth criterion is the empirical median squared error (MNSE), given by

$$\mathrm{MNSE}(\hat{\boldsymbol{\gamma}}) = \mathrm{median}\{\|(\hat{\boldsymbol{\gamma}}^{(l)} - \boldsymbol{\gamma})\|^2, 1 \le l \le L\}. \tag{13}$$

To get more insight, we computed also the separate MNSE for slopes and intercepts,

$$\mathrm{MNSE}(\hat{\boldsymbol{\beta}}) = \mathrm{median}\{\|(\hat{\boldsymbol{\beta}}^{(l)} - \boldsymbol{\beta})\|^2, 1 \le l \le L\},$$
$$\mathrm{MNSE}(\hat{\boldsymbol{\alpha}}) = \mathrm{median}\{\|(\hat{\boldsymbol{\alpha}}^{(l)} - \boldsymbol{\alpha})\|^2, 1 \le l \le L\}.$$

Hubert and Rousseeuw [2] provided the code for obtaining the RDL_1 estimator and the M-S estimator is implemented in the function *lmRob* of the R package *robust*. Now to detect outliers, since the M-S estimator is asymptotically normal, in this paper we follow the rule proposed by Rousseeuw and van Zomeren [21] for the LMS estimator, in which an observation is regarded as an outlier if the absolute value of the standarized residual (using the NMAD) exceeds 2.5. The analogous rule is considered using LS estimators with residuals standardized also with the NMAD. The GPSC automatically provides the labels of the outliers.

Tables 1–6 report the simulation results for three cases: symmetric contamination generated using $k = 5$, asymmetric contamination and $k = 5$, and symmetric contamination and $k = 4$. Table 1 lists the resulting values of the first performance criteria, ALLD, for the outlier detection rules based on LS, GPSC1, GPSC2, RDL_1 and M-S estimators, under contamination levels of 5%, 10%, 20% and 30%. Table 2 shows the values of the second performance criteria, AFO, for the same classification rules and contamination levels. Tables 3 and 4 report the resulting MSEs and MNSEs, respectively. Finally, Tables 5 and 6 show the separated MNSE values for the slopes and the intercepts respectively.

Tables 1 and 2 indicate that for the simulated data, the classifying rule based on the GPSC method achieves high percentages of correct detection while keeping very small the number of observations wrongly identified as outliers (swamping effect). However, under 30% of asymmetric contamination type B and with outliers generated using $k = 4$, the rules based on the two GPSC procedures have a slightly weaker power detection. Remember that for $k = 4$ not all the introduced outliers are clearly separated from the clean data points (Figure 1). Then, the observed weaker power detection of not so clear outliers might not be a sign of poor performance. See also that when the sample is not contaminated by outliers, the two GPSC rules present the lowest AFO as compared with the classifying rules based on the RDL_1 and M-S methods. Both the RDL_1 and M-S rules wrongly identify as outliers some non-outliers, see Table 2. This agrees with the fact that these methods have slightly larger measures ALLD. Thus, this methods are detecting practically all outliers together with few non outliers.

Table 1. ALLD for the outlier detection rules based on LS, GPSC1, GPSC2, RDL_1 and M-S fitting methods, under contamination scenarios B and D with 5%, 10%, 20% and 30% of symmetric or asymmetric contamination with $k = 4, 5$ within each group $d \in \{1, 3, 5\}$.

		5%		10%		20%		30%	
Case	Method	B	D	B	D	B	D	B	D
	LS	100.0	100.0	100.0	100.0	99.9	100.0	98.8	100.0
Sym	GPSC1	100.0	99.5	100.0	99.3	100.0	99.8	99.2	99.9
$k = 5$	GPSC2	100.0	99.6	100.0	99.3	100.0	99.9	99.1	99.8
	MS	100.0	99.9	100.0	99.9	100.0	100.0	100.0	100.0
	RDL_1	100.0	99.9	100.0	99.9	100.0	100.0	100.0	100.0
	LS	100.0	6.6	99.5	0.0	57.8	0.0	0.0	0.0
Asym	GPSC1	100.0	98.7	100.0	98.8	99.2	99.1	87.4	97.3
$k = 5$	GPSC2	100.0	98.5	100.0	98.1	99.3	99.3	82.4	96.3
	MS	100.0	99.8	100.0	99.8	100.0	100.0	100.0	100.0
	RDL_1	100.0	99.8	100.0	99.8	100.0	99.9	100.0	99.7
	LS	100.0	99.9	99.2	100.0	97.4	100.0	92.4	99.8
Sym	GPSC1	99.9	91.7	98.2	86.7	93.0	91.4	78.6	93.1
$k = 4$	GPSC2	99.6	91.6	97.6	85.0	93.1	89.9	78.9	92.3
	MS	100.0	97.1	99.7	92.9	98.7	96.8	94.9	98.5
	RDL_1	99.9	93.2	99.6	92.0	98.4	96.8	96.8	98.8

Table 2. AFO for the rules based on LS, GPSC1, GPSC2, RDL_1 and M-S methods, under contamination scenarios A, B and C with 5%, 10%, 20% and 30% of symmetric or asymmetric contamination with $k = 4, 5$ within each group $d \in \{1, 3, 5\}$.

			5%			10%			20%			30%		
Case	Method	A	B	C	D	B	C	D	B	C	D	B	C	D
	LS	2.84	2.28	2.78	2.34	1.78	2.57	1.73	1.15	2.45	0.98	0.61	2.31	0.49
Sym	GPSC1	1.16	0.45	0.52	0.48	0.34	0.50	0.39	0.28	0.47	0.27	0.20	0.45	0.19
$k = 5$	GPSC2	1.15	0.42	0.49	0.46	0.33	0.47	0.37	0.25	0.44	0.26	0.19	0.45	0.18
	MS	3.52	2.88	3.37	2.97	1.81	3.21	1.90	0.95	3.02	0.95	0.41	2.86	0.39
	RDL_1	2.86	2.33	2.70	2.43	1.56	2.64	1.51	0.79	2.30	0.82	0.28	2.24	0.31
	LS	2.84	3.85	2.78	2.26	2.86	2.57	1.98	0.33	2.45	1.59	0.03	2.31	1.25
Asym	GPSC1	1.16	0.47	0.52	0.43	0.37	0.50	0.36	0.28	0.47	0.23	0.71	0.45	0.17
$k = 5$	GPSC2	1.15	0.45	0.49	0.42	0.34	0.47	0.33	0.22	0.44	0.23	0.51	0.45	0.16
	MS	3.52	2.66	3.37	2.56	1.88	3.21	1.81	0.85	3.02	0.84	0.30	2.86	0.27
	RDL_1	2.86	2.17	2.70	2.22	1.56	2.64	1.55	0.78	2.30	0.74	0.27	2.24	0.29
	LS	2.84	2.27	2.78	2.31	1.67	2.57	1.72	1.08	2.45	0.92	0.57	2.31	0.45
Sym	GPSC1	1.16	0.44	0.52	0.48	0.36	0.50	0.36	0.29	0.47	0.27	0.19	0.45	0.19
$k = 4$	GPSC2	1.15	0.45	0.49	0.48	0.32	0.47	0.35	0.26	0.44	0.25	0.17	0.45	0.18
	MS	3.52	2.87	3.37	2.99	1.82	3.21	1.91	0.95	3.02	0.94	0.40	2.86	0.38
	RDL_1	2.86	2.33	2.70	2.43	1.56	2.64	1.51	0.79	2.30	0.82	0.28	2.24	0.31

Concerning now the robustness performance criteria MSE and MNSE, Tables 3 and 4 show that the two GPSC estimators are practically as efficient as LS under Scenario A without outliers. This agrees with the fact discussed above about the low percentage of false outliers detected. It also presents slightly better MSE and MNSE figures than the other estimators in most of the cases, except for the case of 30% of asymmetric contamination type B. In the rest of cases, the results of the M-S estimator are quite similar. See how the MSEs and MNSEs of LS estimators increase considerably even with 10% of outliers within selected groups. Tables 5 and 6 show similar conclusions when looking at the detailed MNSE results for slopes and intercepts.

Simulations were also performed by introducing contamination in several groups of the same size instead of groups of different sizes. Results suggested that the GPSC method works better

Table 3. MSE($\times 100$) of the LS, GPSC1, GPSC2, RDL_1 and M-S estimators, under contamination scenarios A, B and C with 5%, 10%, 20% and 30% of symmetric or asymmetric contamination with $k = 4, 5$ within each group $d \in \{1, 3, 5\}$.

			5%			10%			20%			30%		
Case	Method	A	B	C	D	B	C	D	B	C	D	B	C	D
	LS	0.73	8.51	0.68	10.70	31.57	0.66	36.02	71.47	0.66	72.46	130.60	0.71	117.99
Sym	GPSC1	0.75	0.70	0.71	0.84	0.78	0.68	0.80	0.84	0.67	0.85	0.90	0.73	0.98
$k = 5$	GPSC2	0.75	0.70	0.70	0.84	0.78	0.68	0.80	0.83	0.68	0.85	0.92	0.73	0.98
	MS	0.84	0.79	0.84	0.94	0.84	0.80	0.87	0.88	0.76	0.87	0.90	0.84	0.96
	RDL_1	1.27	1.38	1.26	1.36	1.44	1.37	1.44	1.75	1.37	1.66	2.02	1.38	1.72
	LS	0.73	21.13	0.68	14.57	59.61	0.66	27.38	193.54	0.66	42.49	388.35	0.71	47.65
Asym	GPSC1	0.75	0.72	0.71	0.80	0.79	0.68	0.85	0.83	0.67	0.84	6.36	0.73	1.75
$k = 5$	GPSC2	0.75	0.72	0.70	0.80	0.79	0.68	0.85	0.83	0.68	0.83	7.89	0.73	2.10
	MS	0.84	0.81	0.84	0.87	0.85	0.80	0.90	0.86	0.76	0.83	0.96	0.84	0.96
	RDL_1	1.27	1.45	1.26	1.42	1.71	1.37	1.64	2.21	1.37	1.94	3.40	1.38	3.01
	LS	0.73	6.21	0.68	9.27	24.01	0.66	31.14	58.07	0.66	61.37	109.26	0.71	98.94
Sym	GPSC1	0.75	0.70	0.71	0.88	0.79	0.68	0.88	0.87	0.67	0.94	1.08	0.73	1.17
$k = 4$	GPSC2	0.75	0.71	0.70	0.88	0.80	0.68	0.90	0.89	0.68	0.99	1.08	0.73	1.18
	MS	0.84	0.79	0.84	0.94	0.84	0.80	0.90	0.88	0.76	0.91	0.96	0.84	0.98
	RDL_1	1.27	1.38	1.26	1.36	1.44	1.37	1.44	1.75	1.37	1.66	2.02	1.38	1.72

Table 4. MNSE($\times 100$) of LS, GPSC1, GPSC2, RDL_1 and M-S estimators, under contamination scenarios A, B and C with 5%, 10%, 20% and 30% of symmetric or asymmetric contamination with $k = 4, 5$ within each group $d \in \{1, 3, 5\}$.

			5%			10%			20%			30%		
Case	Method	A	B	C	D	B	C	D	B	C	D	B	C	D
	LS	0.43	3.31	0.38	5.26	11.95	0.38	19.20	34.33	0.38	39.42	66.55	0.40	67.22
Sym	GPSC1	0.44	0.41	0.40	0.48	0.45	0.40	0.46	0.48	0.38	0.47	0.51	0.42	0.56
$k = 5$	GPSC2	0.44	0.41	0.40	0.49	0.45	0.40	0.46	0.48	0.37	0.47	0.52	0.41	0.56
	MS	0.47	0.44	0.44	0.50	0.47	0.43	0.48	0.49	0.40	0.49	0.52	0.44	0.54
	RDL_1	0.67	0.80	0.72	0.79	0.75	0.79	0.86	0.91	0.87	0.96	1.06	0.83	0.95
	LS	0.43	8.59	0.38	12.89	35.90	0.38	24.34	147.94	0.38	37.48	322.31	0.40	40.82
Asym	GPSC1	0.44	0.45	0.40	0.47	0.44	0.40	0.48	0.45	0.38	0.48	0.63	0.42	0.56
$k = 5$	GPSC2	0.44	0.45	0.40	0.47	0.44	0.40	0.49	0.44	0.37	0.47	0.67	0.41	0.57
	MS	0.47	0.45	0.44	0.49	0.47	0.43	0.48	0.45	0.40	0.47	0.52	0.44	0.56
	RDL_1	0.67	0.83	0.72	0.82	0.96	0.79	0.97	1.24	0.87	1.03	1.90	0.83	1.75
	LS	0.43	2.25	0.38	4.85	9.45	0.38	19.25	30.62	0.38	39.02	62.17	0.40	63.57
Sym	GPSC1	0.44	0.41	0.40	0.51	0.44	0.40	0.48	0.49	0.38	0.52	0.58	0.42	0.59
$k = 4$	GPSC2	0.44	0.40	0.40	0.50	0.45	0.40	0.50	0.51	0.37	0.53	0.59	0.41	0.59
	MS	0.47	0.44	0.44	0.51	0.48	0.43	0.49	0.50	0.40	0.51	0.54	0.44	0.54
	RDL_1	0.67	0.80	0.72	0.79	0.75	0.79	0.86	0.91	0.87	0.96	1.06	0.83	0.95

than the other methods for contaminated groups of medium or large size under Scenario B (low leverage outliers). Studies also showed that the GPSC method works better when the groups means are clearly different, i.e. when the variance of groups effects σ_α^2 is clearly greater than individual error variance σ^2.

8. Application

From the original AAGIS data set, we consider as outcome the variable *income*, as covariates the variables *hectares*, *crops*, *beef* and *sheep* and as grouping variable the variable *state*, which gives the state in which the farm is located, with 1 = New South Wales, 2 = Victoria,

Table 5. MNSE($\times 100$) for LS, GPSC1, GPSC2, RDL_1 and M-S estimators of the slopes vector, under contamination scenarios A, B and C with 5%, 10%, 20% and 30% of symmetric or asymmetric contamination with $k = 4, 5$ within each group $d \in \{1, 3, 5\}$.

			5%			10%			20%			30%		
Case	Method	A	B	C	D	B	C	D	B	C	D	B	C	D
	LS	0.01	0.17	0.02	0.71	0.54	0.01	2.08	1.17	0.01	2.96	2.20	0.01	3.38
Sym	GPSC1	0.01	0.01	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.03
$k = 5$	GPSC2	0.01	0.01	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.03
	MS	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.03
	RDL_1	0.02	0.03	0.03	0.04	0.03	0.03	0.04	0.03	0.03	0.04	0.04	0.03	0.04
	LS	0.01	0.22	0.02	2.82	0.48	0.01	6.32	1.08	0.01	12.39	1.72	0.01	17.35
Asym	GPSC1	0.01	0.01	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.03
$k = 5$	GPSC2	0.01	0.01	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.03
	MS	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.03
	RDL_1	0.02	0.03	0.03	0.04	0.03	0.03	0.05	0.04	0.03	0.10	0.04	0.03	0.22
	LS	0.01	0.12	0.02	0.68	0.37	0.01	1.98	0.86	0.01	2.74	1.67	0.01	3.09
Sym	GPSC1	0.01	0.01	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.03	0.02	0.01	0.03
$k = 4$	GPSC2	0.01	0.01	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.03	0.02	0.01	0.03
	MS	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.03
	RDL_1	0.02	0.03	0.03	0.04	0.03	0.03	0.04	0.03	0.03	0.04	0.04	0.03	0.04

Table 6. MNSE($\times 100$) of LS, GPSC1, GPSC2, RDL_1 and M-S estimators of the intercepts, under contamination scenarios A, B and C with 5%, 10%, 20% and 30% of symmetric or asymmetric contamination with $k = 4, 5$ within each group $d \in \{1, 3, 5\}$.

			5%			10%			20%			30%		
Case	Method	A	B	C	D	B	C	D	B	C	D	B	C	D
	LS	0.41	3.09	0.36	4.47	11.48	0.37	17.00	33.05	0.36	36.84	64.24	0.39	63.81
Sym	GPSC1	0.43	0.40	0.38	0.46	0.43	0.38	0.44	0.46	0.36	0.45	0.49	0.41	0.53
$k = 5$	GPSC2	0.42	0.40	0.38	0.47	0.43	0.38	0.44	0.46	0.35	0.45	0.50	0.40	0.53
	MS	0.46	0.42	0.43	0.47	0.45	0.42	0.45	0.47	0.39	0.46	0.50	0.43	0.51
	RDL_1	0.65	0.76	0.68	0.75	0.74	0.74	0.81	0.89	0.83	0.91	1.02	0.81	0.91
	LS	0.41	8.37	0.36	9.90	35.60	0.37	18.01	146.69	0.36	25.06	320.97	0.39	23.53
Asym	GPSC1	0.43	0.43	0.38	0.45	0.43	0.38	0.45	0.44	0.36	0.45	0.61	0.41	0.53
$k = 5$	GPSC2	0.42	0.43	0.38	0.44	0.43	0.38	0.46	0.43	0.35	0.45	0.64	0.40	0.54
	MS	0.46	0.43	0.43	0.46	0.46	0.42	0.45	0.44	0.39	0.45	0.50	0.43	0.54
	RDL_1	0.65	0.80	0.68	0.78	0.94	0.74	0.88	1.20	0.83	0.94	1.86	0.81	1.53
	LS	0.41	2.09	0.36	4.26	9.05	0.37	17.25	29.71	0.36	36.21	60.10	0.39	60.75
Sym	GPSC1	0.43	0.40	0.38	0.47	0.43	0.38	0.46	0.48	0.36	0.49	0.56	0.41	0.56
$k = 4$	GPSC2	0.42	0.39	0.38	0.47	0.44	0.38	0.47	0.50	0.35	0.50	0.57	0.40	0.55
	MS	0.46	0.42	0.43	0.48	0.45	0.42	0.46	0.48	0.39	0.48	0.53	0.43	0.51
	RDL_1	0.65	0.76	0.68	0.75	0.74	0.74	0.81	0.89	0.83	0.91	1.02	0.81	0.91

$3 =$ Queensland, $4 =$ South Australia, $5 =$ Western Australia, $6 =$ Tasmania and $7 =$ Northern Territory. If we fit model (1) using the raw variables, a histogram of residuals reveals a strongly skewed distribution. Taking logs of the outcome (adding a constant to make it always positive) and the covariates and fitting again the model, a histogram of residuals does not seem far from the normal density but still several outliers appear. Trying to identify the true outliers, we applied all the robust fitting methods considered in this paper apart from LS. Table 7 lists the number of farms remaining in each State after deleting the atypical farms pointed out by the classification rules based on each method. Observe that the rule based on RDL_1 method is the one that eliminates the most quantity of atypical farms over all States, with the largest difference in States 1 and 3. Finally,

Table 7. Number of farms remaining in each State after deletion of outliers based on LS, RDL_1, M-S, GPSC1 and GPSC2 methods.

State	Original	LS	RDL_1	M-S	GPSC1	GPSC2
1	451	438	432	443	446	446
2	265	258	258	262	260	260
3	382	358	355	361	372	372
4	241	235	235	238	239	239
5	221	214	210	210	215	215
6	62	60	61	61	62	62
7	30	28	26	28	29	29
Total	1652	1591	1577	1603	1623	1623

Table 8. Regression parameter estimates obtained by LS, RDL_1, M-S, GPSC1 and GPSC2 methods.

Parameters	LS	RDL_1	M-S	GPSC1	GPSC2
Hectares	0.335	0.339	0.379	0.367	0.367
Crops	0.169	0.144	0.165	0.165	0.165
Beef	0.079	0.060	0.065	0.068	0.068
Sheep	0.029	0.161	0.022	0.023	0.023
State 1	0.677	0.291	0.588	0.617	0.617
State 2	0.604	0.195	0.490	0.523	0.523
State 3	0.607	0.131	0.523	0.554	0.554
State 4	0.534	0.146	0.426	0.465	0.465
State 5	0.667	0.320	0.582	0.608	0.608
State 6	0.711	0.273	0.633	0.663	0.663
State 7	0.543	0.659	0.363	0.454	0.454

Table 8 reports the final regression parameter estimates provided by each method. Observe that the RDL_1 estimates of the group effects are quite different from the estimates obtained by the other methods. This might be due to the mentioned swamping effect that could be strongly affecting the RDL_1 estimates. The original GPSC method (GPSC1) and the faster version (GPSC2) deliver the same figures. Moreover, the results obtained by the M-S and the GPSC methods are somewhat similar. The observed similarity between the M-S and GPSC estimates gives some credibility to these two methods.

9. Concluding remarks

This work studies outlier detection and robust estimation for grouped data following a linear regression model with fixed group effects. We compare several robust methods and the corresponding outlier detection rules based on these methods. We introduce the GPSC method that provides an S estimation with a one-step reweighting procedure. The GPSC procedure as introduced here is recommended in practice when the swamping effect must be kept small, ensuring good efficiency under absence of outliers and at the same time keeping a good detection power.

We present two different implementations of the GPSC algorithm. The faster algorithm proposed in Remark 1 keeps similar properties as the original GPSC algorithm and at the same time reduces drastically the computation time, making it independent of the number of groups D. This faster implementation takes approximately 0.4 times the computation time of the M-S algorithm. R functions for the two versions of the GPSC procedure together with all necessary subroutines are available in the link http://halweb.uc3m.es/esp/Personal/personas/imolina/esp/perso.html.

Acknowledgements

We would like to thank all reviewers for their careful reading, their comments have led to a clear improvement of the paper. This work is supported by the grants MTM2009-09473 and SEJ2007-64500 from the Spanish Ministerio de Educación y Ciencia.

References

[1] Peña D, Yohai VJ. A fast procedure for outlier diagnostics in large regression problems. J Am Stat Assoc. 1999;94:434–445.
[2] Hubert M, Rousseeuw PJ. Robust regression with both continuous and binary regressors. J Stat Plan Inference. 1997;57:153–163.
[3] Maronna RA, Yohai VJ. Robust regression with both continuous and categorical predictors. J Stat Plan Inference. 2000;89:197–214.
[4] Hubert M, Rousseeuw PJ. Robust regression with a categorical covariable. Robust statistics, data analysis, and computer intensive methods, Lecture Notes in Statistics. Vol. 109, New York: Springer-Verlag; 1996. p. 215–224.
[5] Huber PJ. Robust statistics. New York: Wiley; 1981.
[6] Hampel F, Ronchetti E, Rousseeuw PJ, Stahel WA. Robust statistics: the approach based on influence functions. New York: Wiley; 1986.
[7] Rousseeuw P, Yohai V. Robust regression by means of S estimators. In: Frankle J, Härdle W, Martin D, editors. Robust and nonlinear time series analysis, Lecture Notes in Statistics No. 26. Berlin: Springer-Verlag; 1984. p. 256–272.
[8] Rousseeuw PJ. Leas median of squares regression. J Am Stat Assoc. 1984;79:871–880.
[9] Stromberg AJ. Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. SIAM Journal Sci Comp. 1993;14:1289–1299.
[10] Agostinelli C, Markatou M. A one-step robust estimator for regression based on the weighted likelihood reweighting scheme. Stat Probab Lett. 1998;37:341–350.
[11] Markatou M, Basu A, Lindsay BG. Weighted likelihood equations with bootstrap root search. J Am Stat Assoc. 1998;93:740–750.
[12] Yohai VJ. High breakdown-point and high efficiency estimates for regression. Ann Stat. 1987;15:642–656.
[13] Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. New York: Wiley; 1987.
[14] He X, Portnoy S. Reweighting LS estimators converge at the same rate as the initial estimator. Ann Stat. 1992;20:2161–2167.
[15] Maronna RA, Martin RD, Yohai VJ. Robust statistics. Theory and methods. New York: Wiley; 2006.
[16] Gervini D, Yohai VJ. A class of robust and fully efficient regression estimators. Ann Stat. 2002;30:583–616
[17] Rousseeuw PJ. Multivariate estimation with high breakdown point. In: Grossmann W, Pflug G, Vincze T, Wertz W, editors. Mathematical statistics and applications. Dordrecht: B Reidel; 1985. p. 283–297.
[18] Chambers R, Tzavidis N. M-quantile models for small area estimation. Biometrika. 2006;93:255–268.
[19] Chambers R, Pratesi M, Salvati N, Tzavidis N. M-quantile models with application to poverty mapping. Stat Methods Appl. 2008;17:393–411.
[20] Lawrence AJ. Deletion influence and masking in regression. J R Stat Soc B. 1995;57:181–189.
[21] Rousseeuw PJ, van Zomeren BC. Unmasking multivariate outliers and leverage points. J Am Stat Assoc. 1990;85:633–639.