UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 09-88 (27)
Statistics and Econometrics Series
December 2009

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

ROBUST ESTIMATION IN LINEAR REGRESSION MODELS WITH FIXED EFFECTS

MOLINA I., PEÑA D., AND PÉREZ B.[1]

Abstract

In this work we extend the procedure proposed by Peña and Yohai (1999) for computing robust regression estimates in linear models with fixed effects. We propose to calculate the principal sensitivity components associated to each cluster and delete the set of possible outliers based on an appropriate robust scale of the residuals. Some advantage of our robust procedure are: (a) it is computationally low demanding, (b) it is able to avoid the swamping effect often present in similar methods, (c) it is appropriate for contamination in the error term (*vertical outliers*) and possibly masked high leverage points (*horizontal outliers*). The performance of the robust procedure is investigated through several simulation studies.

Keywords: Fixed Effects Models; Outlier Detection; Principal Sensitivity Vectors.

[1] **Corresponding authors:** Universidad Carlos III de Madrid, c/ Madrid 126, 28903 Getafe (Madrid). **E-mail adresses:** isabel.molina@uc3m.es (Isabel Molina Peralta), daniel.pena@uc3m.es (Daniel Peña Sánchez de Rivera), betsabe.perez@uc3m.es (Betsabé Pérez Garrido). The research of Betsabé Pérez was supported by the contract from the Community of Madrid (Contrato de Personal Investigador de Apoyo). The research has been partially supported by grant MICINN SEJ2007-64500.

1. INTRODUCTION

Linear regression models are widely used in many fields of science. Probably the most popular fitting method for linear regression models is the least squares (LS) method. The great popularity of this method might be attributed to the fact that the idea behind this method, the minimization of the sum of squared residuals, is simple and comprehensive. However, it is also well known that in the presence of outliers, the LS estimators can be strongly affected. There are two main approaches to address the problem of atypical data in linear regression models. The first one consists in the use of a robust regression method which tries to devise estimators that are not so strongly affected by outliers. A second approach consists in the use of a method to detect outliers and then obtaining a robust fit by fitting the data discarding these outliers. Outliers can be of two types: high leverage points (*horizontal outliers*) or observations with large residuals (*vertical outliers*).

In the literature a lot of effort has been done in the development of robust estimation methods. Examples of these methods include the M-estimators (Huber, 1981), the least median of squares (Rousseeuw, 1984) and the S-estimators (Rousseeuw and Yohai, 1984). However, when the model includes continuous and categorical predictors, these robust estimation methods present some problems. For example, the M estimate becomes non robust while the S estimates become computationally very expensive (Maronna and Yohai, 2000). Solely a small body of the literature on robust methods has been focused on the problem of robust fitting of linear models when continuous and categorical variables are present.

In this work we follow the second approach to address the problem of atypical data. We concentrate on linear regression models with one categorical variable which divides observations in (many) clusters. The proposed robust procedure is based on the principal sensitivity components introduced by Peña and Yohai (1999). Some advantages of this robust procedure are: (a) it is computationally low demanding, (b) it is able to avoid the swamping effect often present in similar methods, (c) it is appropriate for contamination in the error term (*vertical outliers*) and high leverage points (*horizontal outliers*).

The work is organized as follows. Section 2 introduces the ideas of the principal sensitivity components. Section 3 describes the adapted procedure for fixed effects models. Section 4 describes robust procedures appearing in the literature for fitting a linear regression model with categorical variables. Section 5 presents the results of a Monte Carlo simulation study and finally, Section 6 concludes with a discussion.

2. THE PRINCIPAL SENSITIVITY COMPONENTS

Consider the lineal regression model with p continuous variables including the intercept if is the case,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2) \tag{2.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations with i-th element y_i, $\mathbf{X}$ is a full rank $n \times p$ matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters associated to $\mathbf{X}$, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of the random error term.

Let us consider the vector of estimated parameters, $\hat{\boldsymbol{\beta}}$ of model (2.1) defined by

$$\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.2}$$

and the vector of fitted values

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}, \tag{2.3}$$

whose elements are $\hat{y}_1, \ldots, \hat{y}_n$, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$ is the hat matrix with element in position (i, j) denoted h_{ij}, and $\mathbf{e}$ is the vector of the LS residuals $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ with i-th element e_i, where $\mathbf{I}$ represents the identity matrix of conformable size.

To measure the outlyingness of the i-th observation, it seems appropriate to calculate the sensitivity of the forecast of the i-th observation when each of the sample elements is deleted. This intuitive idea brings us the definition of the i-th sensitivity vector given by

$$\mathbf{r}_i = \left(\hat{y}_i - \hat{y}_{i(1)}, \hat{y}_i - \hat{y}_{i(2)}, \ldots, \hat{y}_i - \hat{y}_{i(n)}\right)^T, \tag{2.4}$$

where $\hat{y}_{i(j)}$ is the forecast of y_i when the j-th observation is deleted.

Taking into account that

$$\hat{y}_i - \hat{y}_{i(j)} = \frac{h_{ij}e_j}{1 - h_{jj}}, \tag{2.5}$$

the i-th sensitivity vector becomes

$$\mathbf{r}_i = \left(\frac{h_{i1}e_1}{1-h_{11}}, \frac{h_{i2}e_2}{1-h_{22}}, \ldots, \frac{h_{in}e_n}{1-h_{nn}}\right)^T, \tag{2.6}$$

with all sensitivity vectors we define the Sensitivity Matrix

$$\mathbf{R} = \begin{bmatrix} \mathbf{r}_1^T \\ \vdots \\ \mathbf{r}_n^T \end{bmatrix} \tag{2.7}$$

This matrix can be obtained as $\mathbf{R} = \mathbf{HW}$, where $\mathbf{W}$ is a diagonal matrix with diagonal elements equal to $e_j/(1 - h_{jj})$.

Observe that the $\mathbf{r}_i$'s belong to the p-dimensional subspace generated by the columns of $\mathbf{X}$. This suggests to search for the directions in which the maximum sensitivity change occurs, and then, to project the $\mathbf{r}_i$'s over these directions. But the directions of maximum sensitivity are the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$ associated to the nonnull eigenvalues $\lambda_1, \ldots, \lambda_p$ of the matrix $\mathbf{M} = \mathbf{R}^T\mathbf{R}$. Then, we just need to compute the projections

$$\mathbf{z}_q = \mathbf{R}\mathbf{v}_q, \quad q = 1, \ldots, p. \tag{2.8}$$

These projections are called the principal sensitivity components.

Note that the projections $\mathbf{z}_q$ inherit the properties of the principal components, which means that they are orthogonal vectors and that the variance associated to each projection $\mathbf{z}_q$ is given by its corresponding eigenvalue λ_q. For purposes of outlier detection, there are two relevant properties that these projections satisfy:

1. The extreme coordinates of the projections $\mathbf{z}_q$ correspond to high leverage points *(horizontal outliers)*, see Theorem 1 on page 438 in Peña and Yohai (1999).

2. The projections $\mathbf{z}_q$ represent the directions of maximum standardized change in the regression parameters.

The full robust procedure proposed by Peña and Yohai (1999) for detecting *horizontal* and *vertical outliers* is formalized in two stages:

Stage 1 This stage is iterative and we search for a preliminary robust estimator of $\boldsymbol{\beta}$. In the first iteration (r=1) we construct a set A_1 of candidates $\boldsymbol{\beta}$ with $3p+1$ elements. The first element corresponds to the LS estimator. The following $3p$ elements are obtained by calculating the p projections $\mathbf{z}_q$, $q=1,\ldots,p$, and deleting: (1) the half of the smallest coordinates of $\mathbf{z}_q$, (2) the half of the largest coordinates of $\mathbf{z}_q$ and (3) the half of the larges coordinates of $\mathbf{z}_q$ in absolute value. Then, from the set A_1 of $3p+1$ candidates, we select the estimate $\hat{\boldsymbol{\beta}}^{(1)}$ which minimizes of a certain scale s of the residuals, that is

$$\hat{\boldsymbol{\beta}}^{(1)} = argmin_{\boldsymbol{\beta}\epsilon A_1} s(e_i(\boldsymbol{\beta}),\ldots,e_n(\boldsymbol{\beta})). \tag{2.9}$$

In the next iterations (r≥2), we compute the vector of residuals, $\mathbf{e}^{(r)} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(r-1)}$ and their robust scale $s^{(r-1)}$, and eliminate the observations such that $|e_j^{(r)}| \geq C_1 \cdot s^{(r-1)}$ where C_1 is a constant. With the remaining observations, a LS estimator is computed and again we calculate the principal sensitivity components. We construct a set A_r with $3p+2$ candidates $\boldsymbol{\beta}$. The first $3p+1$ candidates are obtained identically as in the first iteration, and the last element is the previous estimator $\boldsymbol{\beta}^{(r-1)}$. The iterations end when $\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)}$ and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}^{(r)}$ is called the preliminary robust estimator, which is robust against possibly masked groups of high leverage points.

Stage 2 Compute the residuals $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}_1$, for all elements of the sample and let s be their robust scale. Find a set n_1 of observations such that $|e_j| > C_2 \cdot s$ where C_2 is a constant. With the remaining $n - n_1$ observations, compute $\widetilde{\boldsymbol{\beta}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{y}}$ where $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{y}}$ correspond to the elements of the sample after delating the n_1 observations. Then test whether each of the n_1 elements are outliers by using the test statistic

$$t_j = \frac{y_j - \mathbf{x}_j^T\widetilde{\boldsymbol{\beta}}}{\widetilde{s}_2\sqrt{1+h_{jj}}}, \tag{2.10}$$

where $\mathbf{x}_j^T$ represents the *j*-th row of $\mathbf{X}$,

$$\widetilde{s}_2^2 = \frac{\sum_{j=1}^{n-n_1-1}(y_j - \mathbf{x}_j^T\widetilde{\boldsymbol{\beta}})^2}{n-n_1-p} \quad \text{and} \quad h_{jj} = \mathbf{x}_j^T(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\mathbf{x}_j. \tag{2.11}$$

The observations of the set n_1 are finally eliminated if $|t_j| > C_3$ where C_3 is a constant. Based on simulation studies, Peña and Yohai (1999) proposed the use of the constants $C_1 = 2$ and $C_2 = C_3 = 2.5$, but for large sample size they recommend to increase them.

3. ROBUST PROCEDURE FOR LINEAL MODELS WITH FIXED EFFECTS

Consider the linear regression model with fixed cluster effects given by

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \alpha_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim \text{iid } N(0, \sigma_\varepsilon^2), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, I. \tag{3.1}$$

Now we assume that the data are clustered according to the categories of a qualitative variable. Here i represents the cluster index and j the index of an observation within a cluster. There are I clusters and each cluster contains n_i elements, so that the total sample size is $n = n_1 + \cdots + n_I$.

Let us define the following vectors and matrices obtained by stacking the elements as

$$\mathbf{y} = \underset{1\le i\le I}{\text{col}}(\underset{1\le j\le n_i}{\text{col}}(y_{ij})), \quad \mathbf{X} = \underset{1\le i\le I}{\text{col}}(\underset{1\le j\le n_i}{\text{col}}(\mathbf{x}_{ij}^T), \quad \boldsymbol{\varepsilon} = \underset{1\le i\le I}{\text{col}}(\underset{1\le j\le n_i}{\text{col}}(\varepsilon_{ij}))$$

and let $\mathbf{Z} = \underset{1\le i\le I}{\text{diag}}(\mathbf{1}_{n_i})$ be a block-diagonal matrix.

In matrix form, the model (3.1) can be written as

$$\mathbf{y} = \mathbf{X}^\star \boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2) \tag{3.2}$$

where $\mathbf{X}^\star = [\mathbf{X}\ \mathbf{Z}]$ is a matrix with $rank(\mathbf{X}^\star) = p + I$ and $\boldsymbol{\beta}^\star = (\boldsymbol{\beta}^T, \alpha_1, \alpha_2, \ldots, \alpha_I)^T$.

The hat matrix is $\mathbf{H}^\star = (\mathbf{X}^{\star T}\mathbf{X}^\star)^{-1}\mathbf{X}^{\star T}\mathbf{y}$ with element in position (ij, kl) denoted $h_{ij,kl}$.

Let y_{ij} be the *j*-th element in *i*-th cluster and $\hat{y}_{ij(kl)}$ the forecast of the observation y_{ij} when observation y_{kl} is deleted. Then, the Sensitivity Matrix takes the form

$$\mathbf{R} = \begin{pmatrix} \hat{y}_{11} - \hat{y}_{11(11)} & \cdots & \hat{y}_{11} - \hat{y}_{11(1n_1)} & \cdots & \hat{y}_{11} - \hat{y}_{11(I_1)} & \cdots & \hat{y}_{11} - \hat{y}_{11(In_I)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{y}_{1n_1} - \hat{y}_{1n_1(11)} & \cdots & \hat{y}_{1n_1} - \hat{y}_{1n_1(1n_1)} & \cdots & \hat{y}_{1n_1} - \hat{y}_{1n_1(I1)} & \cdots & \hat{y}_{1n_1} - \hat{y}_{1n_1(In_I)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{y}_{I1} - \hat{y}_{I1(11)} & \cdots & \hat{y}_{I1} - \hat{y}_{I1(1n_1)} & \cdots & \hat{y}_{I1} - \hat{y}_{I1(I1)} & \cdots & \hat{y}_{I1} - \hat{y}_{I1(In_I)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{y}_{In_I} - \hat{y}_{In_I(11)} & \cdots & \hat{y}_{In_I} - \hat{y}_{In_I(1n_1)} & \cdots & \hat{y}_{In_I} - \hat{y}_{In_I(I_1)} & \cdots & \hat{y}_{In_I} - \hat{y}_{In_I(In_I)} \end{pmatrix}, \tag{3.3}$$

From here, if j and l are two observations from the same cluster, the forecast of observation j when observation l is deleted is given by

$$\hat{y}_{ij(il)} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{(il)} + \hat{\alpha}_{i(il)}. \tag{3.4}$$

It holds that

$$\hat{y}_{ij} - \hat{y}_{ij(il)} = \mathbf{x}_{ij}^T (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(il)}) + (\hat{\alpha}_i - \hat{\alpha}_{i(il)}) = \frac{h_{ij,il}\, e_{il}}{1 - h_{il,il}}, \tag{3.5}$$

where $e_{il} = y_{il} - x_{il}^T \hat{\boldsymbol{\beta}} - \hat{\alpha}_i$ is the residual of observation l within cluster i. Observe that (3.5) is similar to (2.5).

Now let us partition the matrix $\mathbf{R}$ in $I \times I$ submatrices according to the clustered structure of the data,

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} & \ldots & \mathbf{R}_{1I} \\ \mathbf{R}_{21} & \mathbf{R}_{22} & \ldots & \mathbf{R}_{2I} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{R}_{I1} & \mathbf{R}_{I2} & \ldots & \mathbf{R}_{II} \end{pmatrix}, \tag{3.6}$$

where $\mathbf{R}_{ij}$ represents the matrix containing the sensitivity of the forecast of the observations of cluster i when each observation in cluster j is deleted.

Consider the set $\{\mathbf{R}_{11}, \ldots, \mathbf{R}_{II}\}$ of submatrices in the diagonal of $\mathbf{R}$. The elements of submatrix $\mathbf{R}_{ii}$ represent the sensitivity of the forecast of the observations belonging to cluster i when each observation in the same cluster is deleted. For each submatrix i, we can obtain the principal sensitivity components by computing the eigenvectors and the nonnull eigenvalues associated to $\mathbf{M}_i = \mathbf{R}_{ii}^T \mathbf{R}_{ii}$. The maximum eigenvalue of $\mathbf{M}_i$, λ_{i1} can be interpreted as the measure of the global effect of the observations of cluster i on the forecast of the observations in cluster i. The eigenvector $\mathbf{v}_{i1}$ associated with the greatest eigenvalue λ_{i1} can be interpreted as the direction of maximum change on the forecast of the observations in cluster i when the observations in cluster i are deleted. Thus, we can use the projections $\mathbf{z}_q$ on the directions $\mathbf{v}_q$ to detect high leverage points (*horizontal outliers*) in cluster i.

The new robust procedure for detecting *horizontal* and *vertical outliers* for a linear regression model with clustered observations is formalized in two stages:

Stage 1 Construct a set $A_1 = \{\boldsymbol{\beta}_1^\star, \boldsymbol{\beta}_2^\star, \boldsymbol{\beta}_3^\star, \boldsymbol{\beta}_4^\star\}$ of candidates $\boldsymbol{\beta}^\star$. The first element, $\boldsymbol{\beta}_1^\star$, is the LS estimator using all elements of the sample and the rest are constructed by eliminating given percentages of outliers as follows. First construct the Sensitivity Matrix (3.3) using the model (3.2). For each cluster i, $i = 1, \ldots, I$, consider its corresponding submatrix $\mathbf{R}_{ii}$ and compute the principal sensitivity components $\mathbf{z}_{iq}$, $q = 1, \ldots, p$. For each component q, compute the difference $\mathbf{d}_{iq} = |\mathbf{z}_{jq} - median(\mathbf{z}_{iq})|$ and delete the set of observations whose corresponding element of $\mathbf{d}_{iq}$ exceeds $C_1 \cdot MAD(\mathbf{d}_{iq})$, where MAD stands for the median absolute deviation. We add the restriction that the maximum number of observations eliminated in each cluster can not exceed 50%.

The last step is applied to every cluster i, $i = 1, \ldots, I$. Then we delete all the sets of possible outliers, and with the remaining observations we compute a LS estimator. In our simulation studies we used three different options for C_1: (a) the 90th percentile of a normal distribution for computing $\boldsymbol{\beta}_2^\star$; the 95th percentile for $\boldsymbol{\beta}_3^\star$; and the 99th percentile for $\boldsymbol{\beta}_4^\star$.

Then, we select the preliminary robust estimator $\hat{\boldsymbol{\beta}}^{\star(1)}$ under the criterion:

$$\hat{\boldsymbol{\beta}}^{\star(1)} = \underset{\boldsymbol{\beta}^\star \epsilon A_1}{argmin} \, s(e_{11}(\boldsymbol{\beta}^\star), \ldots, e_{In_I}(\boldsymbol{\beta}^\star)) \tag{3.7}$$

Stage 2 Compute the residuals $e_{ij} = y_{ij} - \mathbf{x}_{ij}^{\star T} \hat{\boldsymbol{\beta}}^{\star(1)}$ for all elements of the sample.

For each cluster i, $i = 1, \ldots, I$, compute a robust scale of the residuals s_i defined by

$$s_i = 1.481 \cdot \text{Med}(|e_{ij}|, e_{ij} \neq 0), \quad j = 1, \ldots, n_i \tag{3.8}$$

Delete the observations such that $|e_{ij}| > C_2 \cdot s_i$, where C_2 is a constant.

Let n_1 be the number of observations eliminated in the last step. With the remaining $n - n_1$ observations, compute $\hat{\tilde{\tilde{\boldsymbol{\beta}}}} = (\tilde{\tilde{\mathbf{X}}}^T \tilde{\tilde{\mathbf{X}}})^{-1} \tilde{\tilde{\mathbf{X}}}^T \tilde{\tilde{\mathbf{y}}}$ where $\tilde{\tilde{\mathbf{X}}}$ and $\tilde{\tilde{\mathbf{y}}}$ correspond to the elements of the sample after delating the n_1 observations. Then, we test each of the n_1 elements by using the test statistic

$$t_{ij} = \frac{y_{ij} - \mathbf{x}_{ij}^T \hat{\tilde{\tilde{\boldsymbol{\beta}}}}}{\hat{\tilde{\tilde{s}}}_2^2 \sqrt{1 + h_{ij,ij}}} \tag{3.9}$$

where

$$\hat{\tilde{\tilde{s}}}_2^2 = \frac{\sum_{j=1}^{n-n_1-1} (y_{ij} - \mathbf{x}_{ij}^T \hat{\tilde{\tilde{\boldsymbol{\beta}}}})^2}{n - n_1 - (p + I)} \quad \text{and} \quad h_{ij,ij} = \mathbf{x}_{ij}^T (\tilde{\tilde{\mathbf{X}}}^T \tilde{\tilde{\mathbf{X}}})^{-1} \mathbf{x}_{ij}$$

The observations of set n_1 are finally eliminated if $|t_{ij}| > C_3$ where C_3 is a constant. In the simulation studies we found that C_2 = 2.5 and $C_3 = 3.5$ work well.

4. OTHER ROBUST PROCEDURES

The RDL_1 estimator was proposed by Hubert and Rousseeuw (1997) and it uses a robust distance and L_1 regression. The RDL_1 estimator is defined by using a three stage procedure:

1. First, search for leverage points over the set of continuous variables applying the minimum volume ellipsoid (MVE) estimator (Rousseeuw, 1985) and then, based on it, compute robust distances.

2. Based on the robust distances, construct strictly positive weights for each observation. Then, regression parameters are estimated by a weighted L_1 procedure.

3. Compute a robust scale of residuals using the median absolute deviation (MAD) over the vector of residuals coming from the weighted L_1 regression.

4. An observation is considered as atypical if the absolute value of the corresponding standardized residual exceeds 2.5.

A possible disadvantage of the RDL_1 method is that it suffers of the swamping effect. This problem will be discussed and illustrated in Section 5.

Maronna and Yohai (2000) proposed two other classes of robust fitting methods when categorical variables are present in the model. They proposed the M-GM estimator, which is a weighted L_1 estimator, and an alternating M and S estimator, where a M-estimator is used for the categorical predictors and the S-estimator for the continuous ones. Two versions of the M-S estimator were proposed. Maronna and Yohai (2000) suggested that, as the number of continuous predictors increases, the advantages of the M-S method over the M-GM one also increase.

5. MONTE CARLO SIMULATIONS

In this section we present two simulation studies to compare the performance of our robust procedure base on the principal sensitivity components (PSC) against the RDL_1, M-GM or M-S methods. Two main performance criteria were used to compare the different robust methods. The first one is the mean percentage of correct detection defined as follows. Let L be the number of simulations, $l = 1, \ldots, L$. Then, the mean percentage of correct detection is defined as

$$\text{MPCD} = \frac{1}{L}\sum_{l=1}^{L} 100 \cdot \frac{\text{number of true outliers detected in simulation } l}{\text{number of true outliers}}. \tag{5.1}$$

The second criterion is the total incorrect detection defined as:

$$\text{TID} = \sum_{l=1}^{L} \text{number of false outliers detected in simulation } l. \tag{5.2}$$

In fact, this last criterion attempts to summarize a measure of the swamping effect. The swamping effect occurs when non-outliers are wrongly identified due to the effect of some hidden outliers, see Lawrence (1995).

5.1. Simulation 1

We simulated data imitating a data set concerning 1652 Australian farms from the Australian Agricultural and Grazing Industries Survey (AAGIS). The data set contains various variables among which we selected four of them: hectares, crops, beef and sheep. We simulated 10 clusters with a total sample size of 400 observations. The 10 clusters were divided into groups with the same cluster sample size each consisting of 2 clusters. The cluster sample sizes in the five groups were respectively 20, 30, 40, 50 and 60. Based on the distribution of the original variables we simulated four continuous variables from $X_1 \sim N(3.31, 0.68)$, $X_2 \sim N(1.74, 1.23)$, $X_3 \sim N(1.70, 1.65)$, $X_4 \sim N(2.41, 2.61)$, were the mean and standard deviations are those of hectares, crops, beef and sheep respectively. The L=1000 iterations were carried out as follows. We simulated I=10 values from a normal variable with zero mean and standard deviation $\sigma_\alpha = 0.05$ to generate the fixed effects α_i associated to the clusters; n = 400 values from a normal variable with zero mean and standard deviation $\sigma_\varepsilon = 0.05$ to generate the random error terms. In the simulation process we held the fixed effects and continuous variables invariant. In each iteration we calculated y_{ij} from model (3.2). Then, we considered three different scenarios:

1. No atypical data are present.

2. Type I (*vertical outliers*). We introduced contamination in three clusters i, specifically $i = \{5, 7, 9\}$. For each cluster, the contamination was created by calculating the mean of cluster i, say $\overline{y}_i$, and its corresponding standard deviation, γ_i. Then, we substitute some observations y_{ij} by a constant $c_1 = \overline{y}_i + k \cdot \gamma_i$ and other few by $c_2 = \overline{y}_i - k \cdot \gamma_i$, where k = 5.

3. Type 2 (*horizontal and vertical outliers*). Again we introduced contamination in the same three clusters $\{5, 7, 9\}$. The contamination over the set of continuous variables X_a, $a = \{1, 2, 3, 4\}$ was created by calculating the mean of the cluster, say $\overline{X_{l,i}}$, and their corresponding standard deviation, $\gamma_{X_{a,i}}$. Then, we substitute some observations $x_{l,ij}$ by a constant $c_3 = \overline{X_{l,i}} + k \cdot \gamma_{X_{a,i}}$ and then we replace their corresponding observations y_{ij} by a constant $c_4 = \overline{y}_i - k \cdot \gamma_i$, where k = 5.

To illustrate graphically the kind of contamination. Figure 1 shows the observations of one simulation under the type of contamination 1 using a level of contamination of 15%.

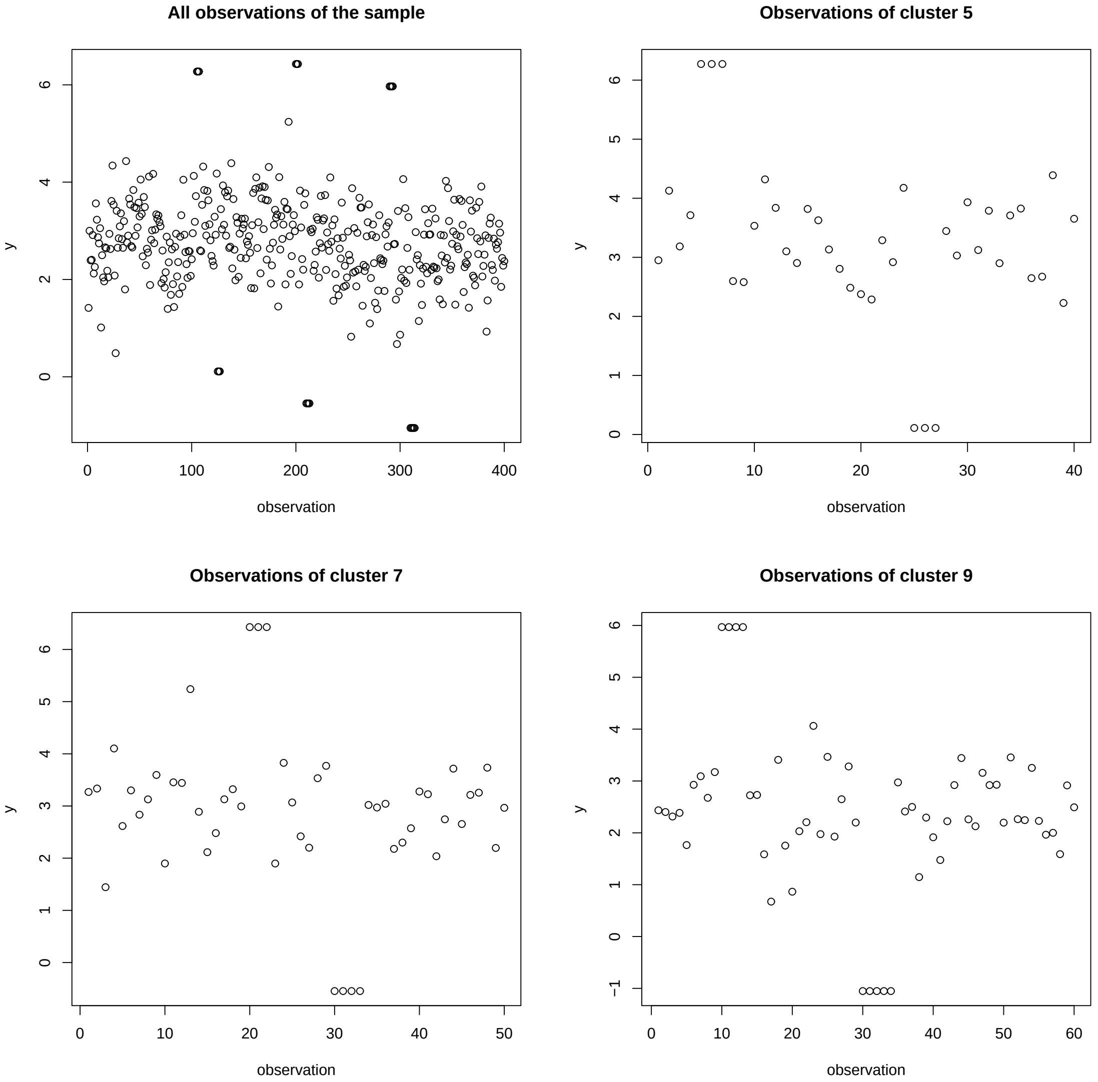

Figure 1: Scatterplot of y versus observation index for all observations of the sample (top left), for observations of cluster 5 (top right), for observations of cluster 7 (bottom left) and for observations of cluster 9 (bottom right).

Figure 2 shows graphically the type of contamination 2 using a level of contamination of 15%.

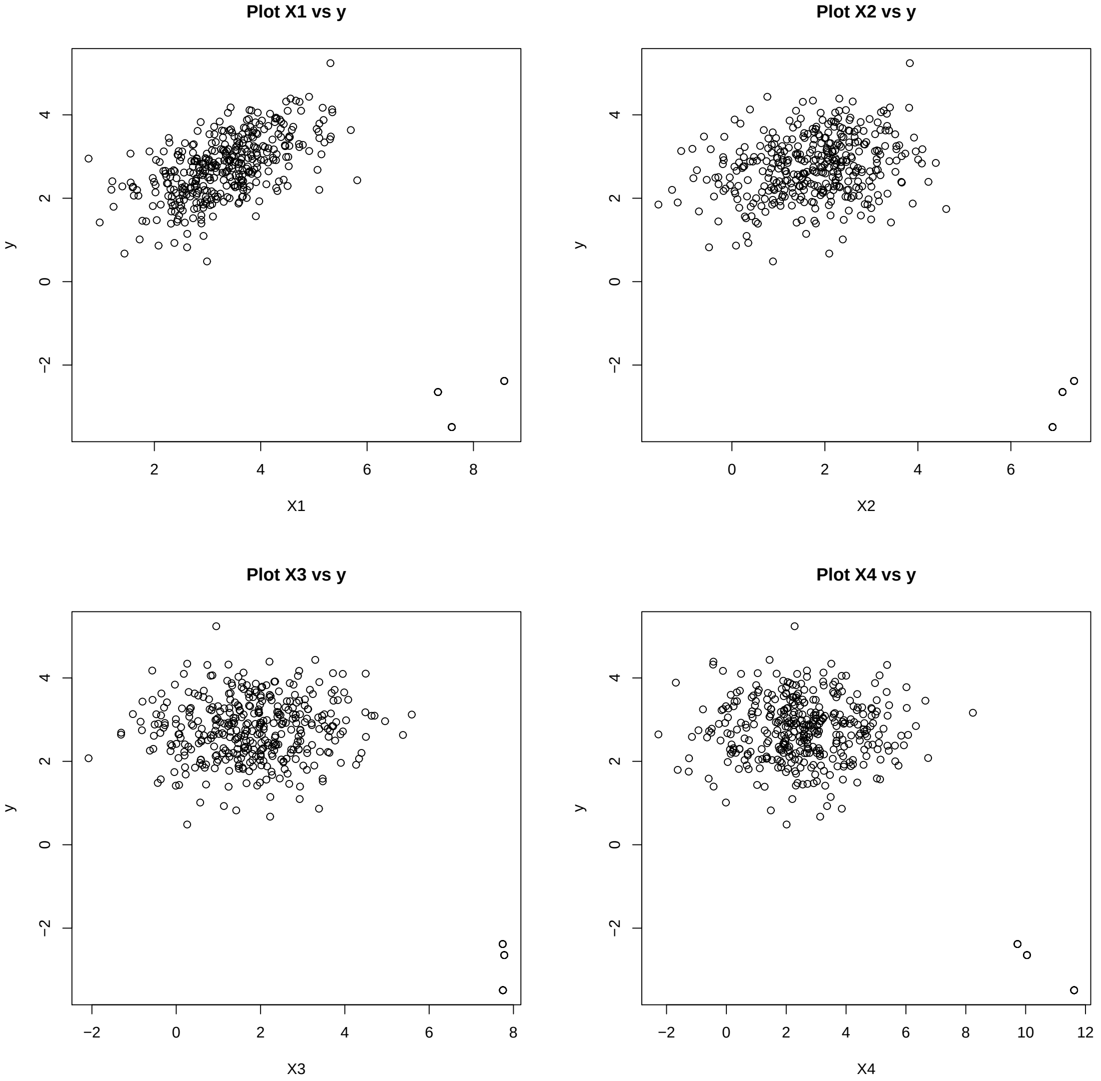

Figure 2: Scatterplot of y versus $X1$ (top left), versus X_2 (top right), versus X_3 (bottom left), versus X_4 (bottom right).

The results of the simulation study is reported in Table 1. The table summarizes the results of the two performance criteria MPCD and TID under levels of contamination 5%, 10% and 15%.

Table 1. Contamination 5%, 10% and 15%.

	No atypical	Contamination 5%				Contamination 10%				Contamination 15%			
	data	Type 1		Type 2		Type 1		Type 2		Type 1		Type 2	
Method	TID	MPCD	TID	MPCD	TID	MPCD	TID	MPCD	TID	MPCD	TID	MPCD	TID
PSC	409	100,00%	359	100,00%	354	99,93%	314	100,00%	318	99,95%	300	99,96%	284
RDL_1	7462	100,00%	6397	100,00%	6325	100,00%	5312	100,00%	5349	100,00%	4511	100,00%	4511
M-S	6112	100,00%	5197	100,00%	167	100,00%	4307	100,00%	123	100,00%	3576	100,00%	93

Observe that the PSC method presents a high percentage of correct detection while keeping small the number of observations wrongly identified as outliers. Furthermore, when the sample is not contaminated by outliers, the PSC method presents the lowest TID as compared with the RDL_1 and M-S methods. On the other hand, when contamination type 1 is present it seems that the RDL_1 and M-S methods suffer of the swamping effect because several non-outliers are wrongly identified as outliers.

The RDL_1 estimator was generated by using the code provided in the article by Hubert, M. and Rousseeuw, P. J. (1997). The M-S estimator was generated by using the $lmRob$ function implemented in S-PLUS version 8.0. Following the suggestions of Rousseeuw and Zomeren (1990) we plot robust distances (mahalanobis distances based on a robust covariance matrix) versus standarized residuals (using the MAD). Then, we considered an observation as a *vertical outlier* if the absolute value of the standarized residual exceeds 2.5. On the other hand, we considered an observation as a *horizontal and vertical outlier* if the observation is a *vertical outlier* and is on the right of the vertical line located at the upper 0.975 percent point of a chi-squared distribution with *p* degree of freedom.

5.2. Simulation 2

In this simulation we considered larger variability in the true fixed effects $\sigma_\alpha = 1$ and smaller in the errors $\sigma_\varepsilon = 0.1$. Tabla 2 summarizes the results of MPCD and TID under levels of contamination 5%, 10% and 15%.

Table 2. Contamination 5%, 10% and 15%.

	No atypical	Contamination 5%				Contamination 10%				Contamination 15%			
	data	Type 1		Type 2		Type 1		Type 2		Type 1		Type 2	
Method	TID	MPCD	TID	MPCD	TID	MPCD	TID	MPCD	TID	MPCD	TID	MPCD	TID
PSC	409	100,00%	353	100,00%	362	100,00%	272	100,00%	307	100,00%	280	100,00%	287
RDL_1	7462	100,00%	6397	100,00%	6325	100,00%	5312	100,00%	5348	100,00%	4512	100,00%	4511
M-S	6112	100,00%	5147	100,00%	170	100,00%	4279	100,00%	120	100,00%	3546	100,00%	91

Table 2 shows that the three robust methods correctly identify 100% of outliers. However, again when contamination type 1 is present the number of incorrectly identified outliers is large for RDL_1 and M-S methods.

6. RESULTS AND DISCUSSION

In this work we studied the detection of atypical data in linear regression models with fixed effects. Since the data are clustered into (many) clusters, we proposed to calculate the principal sensitivity components in each cluster to detect possibly masked high leverage points (*horizontal outliers*). Then, we fit the data and discard the observations with large residuals (*vertical outliers*). The simulation studies show that our robust procedure present a high mean percentage of correct detection (MPCD) and a small number of observations were wrongly detected as outliers (TID). Particulary, when contamination type 1 is present, the level of the swamping effect in our robust procedure is the lowest among the three robust methods. In this work we used the criterion of the minimization of a certain scale of the residuals and then we discard the observations with large residuals with respect to that scale. However, an other alternative is to approximate the quantiles of the $max|e_{ij}|$ by a resampling procedure, and then to examine each possible candidate and to decide whether it is atypical or not by comparing with a selected quantile. This last option might be computationally much more intensive.

References

[1] Davies, L. and Gather, U. (1993). The identification of multiple outliers, *Journal of the American Statistical Association*. **88**, 782–792.

[2] Huber, P. J. (1981). Robust Statistics, *Wiley*, New York.

[3] Hubert, M. and Rousseeuw, P. J. (1997). Robust regression with both continuous and binary regressors, *Journal of Statistical Planning and Inference*. **57**, 153–163.

[4] Lawrence, A.J. (1995). Deletion Influence and Masking in Regression, *Journal of the Royal Statistical Society*. B. **57**, 181–189.

[5] Maronna, R.A. and Yohai, V. J. (2000). Robust regression with both continuous and categorical predictors, *Journal of Statistical Planning and Inference*. **89**, 197–214.

[6] Maronna, R.A., Martin, D. y Yohai, V. J. (2006). Robust Statistics. Theory and Methods. *Wiley*.

[7] Peña, D. and Yohai, V. J. (1999). A fast procedure for outlier diagnostics in large regression problems, *Journal of the American Statistical Association. Theory and Methods*. **94**, 434–445.

[8] Peña, D. and Yohai, V. J. (1995). The detection of influential substets in Linear Regression using an Influence Matrix, *Journal of the Royal Statistical Society*. B. **57**, 145–156.

[9] Rousseeuw, P. J. (1984). Last Median of Squares Regression, *Journal of the American Statistical Association.* **79**, 871–880.

[10] Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, Pflug G., Vincze T. and Wertz W. Eds., *Mathematical Statistics and Applications*, **B**. Reidel, Dordrecht. The Netherlands. 283–297.

[11] Rousseeuw, P. J. and Leroy, A.M. (1987). Robust regression and outlier detection. *Wiley* , New York.

[12] Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by Means of S-estimators. *Robust and Nonlinear Time Series Analysis*. Lectures Notes in Statistics, **26**, *Springer* , New York. 256–272.

[13] Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*. **85**, 633–639.