A Fast Procedure for Outlier Diagnostics in Large Regression Problems

Daniel PEÑA and Victor YOHAI

We propose a procedure for computing a fast approximation to regression estimates based on the minimization of a robust scale. The procedure can be applied with a large number of independent variables where the usual algorithms require an unfeasible or extremely costly computer time. Also, it can be incorporated in any high-breakdown estimation method and may improve it with just little additional computer time. The procedure minimizes the robust scale over a set of tentative parameter vectors estimated by least squares after eliminating a set of possible outliers, which are obtained as follows. We represent each observation by the vector of changes of the least squares forecasts of the observation when each of the data points is deleted. Then we obtain the sets of possible outliers as the extreme points in the principal components of these vectors, or as the set of points with large residuals. The good performance of the procedure allows identification of multiple outliers, avoiding masking effects. We investigate the procedure's efficiency for robust estimation and power as an outlier detection tool in a large real dataset and in a simulation study.

KEY WORDS: Masking; Outliers; Robust regression.

1. INTRODUCTION

Several robust estimates for regression with high breakdown point have been proposed. These include the least median of squares estimate (LMSE) and the least trimmed squares estimate (LTSE) proposed by Rousseeuw (1984), the scale (S) estimates proposed by Rousseeuw and Yohai (1984) the MM estimates proposed by Yohai (1987), and the τ estimates proposed by Yohai and Zamar (1988). These estimates have a very high computational complexity, and thus the usual algorithms compute only approximate solutions. Rousseeuw (1984) proposed an approximate algorithm based on drawing random subsamples of the same size as the number of independent variables. Ruppert (1991) proposed a refinement of this algorithm for S estimates that seems to be more efficient than Rousseeuw's. Stromberg (1991) gave an exact algorithm for computing the LMSE, but it requires generating all possible subsamples of size $p + 1$. A more efficient algorithm that eventually computes the exact LMSE or the LTSE, the feasible solution algorithm (FSA), was proposed by Hawkins (1993, 1994). However, all of these algorithms require computation time that increases exponentially with the number of independent variables, and thus can be applied only when this number is not too large.

In this article we propose a different type of approximate solution to the high–breakdown point estimates just mentioned that can be applied with a large number of independent variables. We do not claim that our proposed approximate procedure keeps the breakdown point of the original estimates. However, the procedure succeeds in detecting groups of outliers in many situations where, due to a masking effect, the usual diagnostic procedures fail and robust estimates require prohibitive computer time. This is shown by means of a Monte Carlo study. For small or moderate datasets and when the computation speed is not a problem, we recommend combining our procedure with the solution obtained by a high–breakdown point estimation method. We can always compare the scales of the residuals obtained with both estimates and choose the one giving the smallest scale. In this way we may improve the solution of the high–breakdown procedure (never worsening it) with just a few seconds of additional time.

In the rest of this section we introduce notation and describe the usual approximations to the high-breakdown estimates based on resampling. In Section 2 we define the principal sensitivity components used for finding outliers. In Section 3 we present the approximate procedure for the minimization of a robust scale. In Section 4 we study the properties of the procedure and prove that when a sample is contaminated with less than 50% of any identical high-leverage observations, the solution remains bounded. In Section 5 we compare the procedure to other previous procedures for outlier detection. In Section 6 we report the results of a Monte Carlo study and present an example for a large dataset. In Section 7 we present some concluding remarks, and in an Appendix give the proof of the main result of Section 4.

We assume a regression model with p independent variables (including the constant if there is intercept) and n observations $(y_i, x_{i,1}, \ldots, x_{1,p}), 1 \leq i \leq n$; that is, $y_i = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i$, for $i = 1, \ldots, n$, where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})', \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, and ε_i is the error of observation i. We call $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{X}$ is a full rank $n \times p$ matrix whose (i, j) element is $x_{i,j}$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$. Then the model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon. \tag{1}$$

All of the robust estimates discussed earlier, with the exemption of the MM estimates, are defined throughout the minimization of a certain scale S of the residuals; that is,

Daniel Peña is Professor of Statistics, Department of Statistics and Econometrics, Universidad Carlos III de Madrid, 28903 Gestafe, Spain (E-mail: *dpena@est-econ.uc3m.es*). Victor Yohai is Professor of Statistics, Universidad de Buenos Aires, Argentina (E-mail: *vyohai@mate.dm.uba.ar*). This work has been partially supported by grant PB96-0111 from Ministerio de Educacion y Cultura, Spain; by European Project CHRCT94 0514; by grant TW92 from the University of Buenas Aires; by project 4186/96 from Consejo Nacional de Investigaciones Científicas y Tecnológicas, Argentina; and by project 03-00000-00576 from Agencia Nacional de Promocion Cientifica y Tecnologica, Argentina.

they are defined by

$$\hat{\boldsymbol{\beta}} = \text{ argmin } S(e_1(\boldsymbol{\beta}), \ldots, e_n(\boldsymbol{\beta})), \tag{2}$$

where

$$e_i(\hat{\boldsymbol{\beta}}) = y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i, \qquad 1 \le i \le n.$$

The usual approximate solutions to the estimates defined by (2) are of the form

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in A}{\text{argmin}}\, S((e_1(\boldsymbol{\beta}) \cdots e_n(\boldsymbol{\beta})), \tag{3}$$

where $A = \{\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(N)}\}$ is a finite set. Rousseeuw (1984) proposed obtaining the elements of A by choosing at random N subsamples of p different data points. If p/n is small, then it can be shown (see Rousseeuw and Leroy 1987) that the probability of getting a clean subset when there is a fraction of outliers equal to ε is approximately given by

$$1 - (1 - (1 - \varepsilon)^p)^n,$$

and the number of subsamples required to make this probability equal to $1 - \alpha$ is given by

$$\mathrm{N}(\varepsilon, \alpha, p) = \frac{\log \alpha}{\log(1 - (1 - \varepsilon)^p)} \simeq \frac{-\log \alpha}{(1 - \varepsilon)^p}. \tag{4}$$

This number increases exponentially with p, and thus the method based on random subsampling can be applied only when p is not very large. Rousseeuw (1993) proposed a modification of the algorithm based on resampling with a deterministic breakdown point. However, the number of subsamples required for this algorithm also increases exponentially with p.

Hadi and Simonoff (1993) presented two procedures for the identification of multiple outliers in linear models and compared them in a Monte Carlo study. The winner of their study, $M1$, is obtained as follows. Starting with the least squares estimate (LSE) fit to the full data, the n observations are ordered by an appropriate diagnostic measure, like the absolute value of the adjusted residual $e_i/\sqrt{(1 - h_{ii})}$, or Cook distance. Then the first p observations form the initial basic subset. A model is fitted to the basic subset, and the residuals are standardized and ordered. The basic set is increased one by one, by ordering the standardized residuals and fitting a model to the basic subset. When the basic subset reaches a size equal to the integer part of $(n + p - 1)/2$, the residuals are tested for outlyingness using the t statistics. The key to the success of the method is to obtain a clean initial subset of data. According to a Monte Carlo study reported by Peña and Yohai (1996), the procedure works well for low-leverage outliers but may fail when the sample contains a set of several high-leverage outliers.

Atkinson (1994) proposed a fast method for the detection of multiple outliers by using a simple forward search from random starting points. Instead of drawing N basic subsamples, Atkinson suggested drawing $h < N$ random subsamples and using the LSE to fit subsets of size $p, p + 1, \ldots, n$, from each subsample. Then outliers are identified as the points having large residuals from the fit that minimizes the least median of squares criterion. This procedure requires again that at least one of the h subsamples does not contain a high-leverage outlier. Then the number of subsamples required to guarantee that this occurs with probability α is given by (4), and thus the procedure will be not very effective when the number of variables p is large.

In this article we propose a fast iterative procedure to estimate $\boldsymbol{\beta}$. In each iteration an estimate is defined by (3) using a suitable set A. Each element of this set is obtained by using the LSE applied to a subsample. These subsamples are obtained by eliminating blocks of observations that potentially can produce a masking effect. The procedure is computationally feasible for very large values of p and seems to be able to avoid the masking problem in many situations where other diagnostic procedures fail.

2. PRINCIPAL SENSITIVITY COMPONENTS

In this section we show how to obtain a set of directions in which masking outliers are expected to appear as extreme values. The set $A = \{\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(N)}\}$ used in (3) is built by deleting extreme observations on these directions and computing the regression LSE of the remaining data. To show this, let

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

be the LSE and let $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)'$ be the vector of fitted values given by

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

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix and $\mathbf{e} = (e_1, \ldots, e_n)'$ is the vector of least squares residuals given by

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (I - \mathbf{H})\mathbf{y}.$$

We let $\hat{\boldsymbol{\beta}}_{(i)}$ denote the LSE when the ith data point is deleted. Then the corresponding change in the LSE is given by (see Cook and Weisberg 1982, p. 110)

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{e_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}{1 - h_{ii}}, \tag{5}$$

where h_{ij} is the ijth element of $\mathbf{H}$, which is the derivative of the prediction $\hat{y}_i$ with respect to y_j. Note that as $h_{ii} = \sum h_{ij}^2$, the leverage can be interpreted as the sum of squares of these derivatives. Call $\hat{y}_{j(i)}$ the forecast corresponding to observation j when observation i is deleted. Then, from (5) it is easily derived that

$$\hat{y}_j - \hat{y}_{j(i)} = \frac{h_{ij}e_i}{1 - h_{ii}}. \tag{6}$$

There are two ways to look at the outlyingness of the ith observation. The first way is by measuring its influence on the forecast of each of the sample points when the observation is deleted. This leads to the influence vectors

$$\mathbf{t}_i = (\hat{y}_1 - \hat{y}_{1(i)}, \ldots, \hat{y}_n - \hat{y}_{n(i)})' = \mathbf{h}_i e_i/(1 - h_{ii}),$$

where $\mathbf{h}_i$ is the ith column of the hat matrix. The components of this vector are proportional to the derivatives h_{ij}, and its norm is proportional to the square root of the

Cook (1979) statistic. An analysis based on these vectors was presented in earlier work (Peña and Yohai 1995) and is discussed in Section 5.

In this article we discuss a second way to measure the outlyingness of the ith observation. We consider the sensitivity of the forecast of the ith observation when each of the sample points is deleted. This leads to the sensitivity vectors

$$\begin{aligned}\mathbf{r}_i &= (\hat{y}_i - \hat{y}_{i(1)}, \ldots, \hat{y}_i - \hat{y}_{i(n)})' \\ &= (h_{i1}e_1/(1-h_{11}), \ldots, h_{in}e_n/(1-h_{nn})),\end{aligned}$$

in which the h_{ij} are weighted by the predicted (out-of-sample) residuals $e_j/(1-h_{jj})$. We define the sensitivity matrix as

$$\mathbf{R} = \begin{bmatrix} \mathbf{r}_1' \\ \cdots \\ \mathbf{r}_n' \end{bmatrix}.$$

The matrix $\mathbf{R}$ can be considered to be a data matrix in which each observation corresponds to a row $\mathbf{r}_i'$ and each variable corresponds to a column $\mathbf{t}_i$. From (6), we get

$$\mathbf{R} = \mathbf{HW}, \tag{7}$$

where $\mathbf{W}$ is the diagonal matrix with terms $e_i/(1-h_{ii})$.

The vectors $\mathbf{r}_i$ belong to the p-dimensional subspace generated by the columns of $\mathbf{X}$, and thus we may summarize their information by choosing an appropriate basis on this space and projecting them over this basis. The first vector of this basis can be obtained by the condition that the projection of the $\mathbf{r}_i$'s on it have maximum sensitivity. Then it is given by

$$\mathbf{v}_1 = \text{ argmax } \sum_{i=1}^{n} (\mathbf{v}'\mathbf{r}_i)^2,$$

subject to $\|\mathbf{v}\| = 1$. The vector $\mathbf{v}_1$ is the eigenvector corresponding to the largest eigenvalue of the matrix $\mathbf{M} = \sum_{i=1}^{n} \mathbf{r}_i\mathbf{r}_i'$, where, according to (7), this matrix is

$$\mathbf{M} = \mathbf{WHW} \tag{8}$$

and has rank p and its ijth element is

$$m_{ij} = \frac{e_i e_j h_{ij}}{(1-h_{ii})(1-h_{jj})}.$$

Let $\mathbf{z}_1$ be the vector whose coordinates are the projections of the $\mathbf{r}_i$'s on $\mathbf{v}_1$, given by $\mathbf{z}_1 = \mathbf{R}\mathbf{v}_1$. It is straightforward to show that $\mathbf{z}_1$ is an eigenvector corresponding to the largest eigenvalue of the matrix $\mathbf{P}$ defined by

$$\mathbf{P} = \mathbf{HW}^2\mathbf{H}, \tag{9}$$

with ijth element

$$p_{ij} = \sum_{k=1}^{n} \frac{e_k^2}{(1-h_{kk})^2}\ h_{ik}h_{jk}.$$

Observe that this expression for p_{ij} is similar to the decomposition $h_{ij} = \sum h_{ik}h_{kj}$. The difference is that each term of the sum is now weighted by the corresponding predicted residuals.

In a similar way, we can project the sensitivity vectors, $\mathbf{r}_i'$s, on the directions of the other eigenvectors, $\mathbf{v}_1, \ldots, \mathbf{v}_n$, of the matrix $\mathbf{M}$, corresponding to the other nonnull eigenvalues $\lambda_2 \geq \cdots \geq \lambda_p$. The eigenvector $\mathbf{v}_i$ will have the following property

$$\mathbf{v}_i = \underset{\|\mathbf{v}\|=1}{\text{argmax}} \sum_{i=1}^{n} (\mathbf{r}_i'\mathbf{v})^2 \tag{10}$$

subject to

$$\mathbf{v}_i'\mathbf{v}_h = 0 \qquad 1 \leq h \leq i-1. \tag{11}$$

The corresponding projections

$$\mathbf{z}_i = R\mathbf{v}_i, \qquad i = 2, \ldots p \tag{12}$$

will be eigenvectors of $\mathbf{P}$.

The vectors $\mathbf{z}_i, 1 \leq i \leq p$, which form an orthogonal base of the p-dimensional subspace generated by the columns of $\mathbf{X}$, are the principal components of the sensitivity observations $\mathbf{r}_i, i = 1, \ldots, n$. We call them principal sensitivity components.

The principal sensitivity components have two additional interesting properties. First, they are eigenvectors of the projection matrix $\mathbf{H}$ corresponding to the eigenvalue 1. This can be shown by using the definition of the $\mathbf{z}_i$ as eigenvectors of $\mathbf{P}$(i.e., $\mathbf{HW}^2\mathbf{H}\mathbf{z}_i = \lambda_i\mathbf{z}_i$) and multiplying this equation by $\mathbf{H}$. Note that this basis is selected taking into account information about the predicted residuals $e_i/(1-h_{ii})$. Let $\mathbf{B}$ be the $n \times p$ matrix with columns $\mathbf{z}_i/\sqrt{\lambda_i}$ so that $\mathbf{H} = \mathbf{BB}'$. Put $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,n})$; then

$$h_{jj} = \sum_{i=1}^{p} \frac{z_{i,j}^2}{\lambda_i},$$

and looking for extreme coordinates of each vector $\mathbf{z}_i$ implies a finer analysis than looking at the leverages h_{jj}.

The second relevant property of the principal sensitivity components is that they represent directions of maximum standardized change on the regression parameters. To show this, suppose that instead the forecast changes, we look at the standardized regression parameter changes. For this purpose, we define the standardized effect on the regression coefficients due to the ith observation by

$$\boldsymbol{\gamma}_i = (\mathbf{X}'\mathbf{X})^{1/2}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}). \tag{13}$$

Usually, this influence is summarized by the univariate Cook (1977) statistics $D_i = \|\boldsymbol{\gamma}_i\|^2/ps^2$, where $s^2 = (n-p)^{-1}\sum e_i^2$ is the residual variance. It is well known that the statistic D_i may fail to detect outliers when masking is present (see Lawrance 1995). Masked outliers will have similar effects on the estimated parameter $\boldsymbol{\beta}$, and in some directions in R^p these similarities will appear more strongly. Therefore, it seems natural to make a finer analysis by considering directions where the $\boldsymbol{\gamma}_i$'s are the largest. The first of these directions may be defined by

$$\mathbf{u}_1 = \underset{\|\mathbf{u}\|=1}{\text{argmax}} \sum_{i=1}^{n} (\boldsymbol{\gamma}_i'\mathbf{u})^2.$$

Then $\mathbf{u}_1$ is the eigenvector corresponding to the maximum eigenvalue λ_1 of the $p \times p$ uncentered covariance matrix $\mathbf{Q}$ of the $\boldsymbol{\gamma}_i$'s:

$$\mathbf{Q} = \sum \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i'.$$

From (5) and (13), we have that the matrix whose rows are the $\boldsymbol{\gamma}_i$'s is given by

$$\Gamma = \mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}, \tag{14}$$

and thus

$$\mathbf{Q} = (\mathbf{X}'\mathbf{X})^{-1/2}(\mathbf{X}'\mathbf{W}^2\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1/2}. \tag{15}$$

We can also define directions $\mathbf{u}_2, \ldots, \mathbf{u}_p$ by the eigenvectors corresponding to the other eigenvalues $\lambda_2 \geq \cdots \geq \lambda_p$ of the matrix $\mathbf{Q}$. These directions will also have a property analogous to (10) and (11). The eigenvectors of $\mathbf{Q}$ represent the directions of maximum variability of the standardized effects $\boldsymbol{\gamma}_i$. To transform the effects $\boldsymbol{\gamma}_i$ into changes of forecast, we must multiply the $\boldsymbol{\gamma}_i$ by the standardized matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}$. Therefore, the directions of maximum forecast change are obtained by multiplying the $\mathbf{u}_i$ by $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}$. Then, let us define

$$\mathbf{Z}_i = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{u}_i, \tag{16}$$

which represents the forecast change for each observation in the direction $\mathbf{u}_i$. Note that although the eigenvectors of $\mathbf{Q}$ are defined up to an orthogonal transformation (this property is inherited from the similar property of $(\mathbf{X}'\mathbf{X})^{-1/2}$), the vectors $\mathbf{Z}_i$ are uniquely determined (except for a scalar factor), and, moreover, they are invariant for affine transformations of the $\mathbf{x}_i$'s.

We now show that the $\mathbf{Z}_i$'s are also the eigenvectors of the $\mathbf{P}$ matrix defined in (9) and then are equal (except by a scalar factor) to the principal sensitivity components $\mathbf{z}_i$'s. Because $\mathbf{u}_i$ is an eigenvector of $\mathbf{Q}$, using (15) we get

$$(\mathbf{X}'\mathbf{X})^{-1/2}(\mathbf{X}'\mathbf{W}^2\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{u}_i = \lambda_i \mathbf{u}_i. \tag{17}$$

Multiplying this equation by $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}$, we get that $\mathbf{Z}_i$ is an eigenvector of $\mathbf{H}\mathbf{W}^2$. Therefore,

$$\mathbf{H}\mathbf{W}^2\mathbf{Z}_i = \lambda_i \mathbf{Z}_i, \tag{18}$$

and multiplying this last equation by $\mathbf{H}$, we obtain

$$\mathbf{H}\mathbf{W}^2\mathbf{Z}_i = \lambda_i \mathbf{H}\mathbf{Z}_i. \tag{19}$$

Comparing (18) and (19), $\mathbf{Z}_i = \mathbf{H}\mathbf{Z}_i$. Replacing this result on the left side of (18), we obtain that $\mathbf{Z}_i$ is the eigenvector of $\mathbf{H}\mathbf{W}^2\mathbf{H}$ corresponding to the eigenvalue λ_i.

This relationship provides a convenient way to compute the principal sensitivity components. Instead of computing the eigenvectors $\mathbf{z}_j$'s of the $n \times n$ matrix $\mathbf{P}$, we can compute the eigenvectors $\mathbf{u}_j$'s of the $p \times p$ matrix $\mathbf{Q}$. Then the $\mathbf{Z}_j$'s will be computed by (16). This makes the method faster for large numbers of observations.

3. THE PROCEDURE

The main idea of the procedure is to use a robust scale for evaluating a set of possible solutions. These solutions are determined by applying LSE to subsamples in which sets of potentially outlier observations have been deleted.

The proposed procedure has two stages. The first stage is iterative, and in each iteration a robust estimate is found using the criterion of minimizing a robust scale of the residuals over a finite set A according to (3). The elements of A are obtained in each iteration as follows. We start by deleting all of the observations with large residuals according to the current best estimate and computing the principal sensitivity components for the remaining sample. For each of these components, extreme points are deleted, and an element of A is obtained as the LSE on the remaining sample. The iterations continue until convergence.

In the second stage we improve the efficiency of the robust estimate obtained in the first stage. The residuals based on the estimate found in the first stage are computed, all observations with large residuals are eliminated, and the points deleted are tested one by one by using the studentized residual for outlyingness. The final estimate is computed by LSE using the cleaned sample.

According to the previous section, high-leverage outliers are expected to appear as extreme coordinates in at least one of the principal sensitivity components. The fact that good high-leverage points may also appear as extreme points on these directions can only affect the efficiency of the solution of the first stage. However, the efficiency of the final estimate is improved by testing each potential outlier one by one in the second stage of the procedure. Low-leverage outliers are detected by their large residuals. We present the details of the procedure in following sections.

Stage 1. In this stage we find a robust estimate of $\boldsymbol{\beta}$ by an iterative procedure. In each iteration, an estimate $\hat{\boldsymbol{\beta}}^{(i)}$ is defined by

$$\hat{\boldsymbol{\beta}}^{(i)} = \underset{\boldsymbol{\beta} \in A_i}{\operatorname{argmin}} S(e_1(\boldsymbol{\beta}), \ldots, e_n(\boldsymbol{\beta})). \tag{20}$$

In the first iteration, the set A_1 has $3p+1$ elements. One of these elements is the LSE, and for each principal sensitivity component $\mathbf{z}_j, j = 1, \ldots, p$ we compute three estimates by LS as follows: the first eliminating the half of observations corresponding to the smallest coordinates of $\mathbf{z}_j$, the second eliminating the half corresponding to the largest coordinates of $\mathbf{z}_j$, and the third eliminating the half corresponding to the largest absolute values.

For the next iterations, $i > 1$, we start computing the residuals $e^{(i)} = y - \mathbf{X}\hat{\boldsymbol{\beta}}^{(i-1)}$ and let $s^{(i-1)}$ be its corresponding robust scale. Then we delete all of the observations j such that

$$|e_j^{(i)}| \geq C_1 s^{(i-1)}. \tag{21}$$

Then, with the remaining observations we compute the LSE, $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}^{(i)}$, and the principal sensitivity components. The set A_i will contain $3p+2$ elements: the $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}^{(i)}$, $\hat{\boldsymbol{\beta}}^{(i-1)}$, and $3p$ estimates obtained by deleting extreme values in the principal sensitivity components as in the first iteration.

The procedure ends when $\hat{\boldsymbol{\beta}}^{(i+1)} = \hat{\boldsymbol{\beta}}^{(i)}$, and the estimate that minimizes the robust scale on this stage is called $\hat{\boldsymbol{\beta}}_1$.

Comments on Stage 1. As the objective of this stage is to obtain a preliminary robust estimate, the value of C_1 in (21) is taken relatively low to increase the power of the procedure. We have found that $C_1 = 2$ works well, and this value has been used in the simulations and the examples.

This stage includes two mechanisms for the elimination of the outlier effects. The low-leverage outliers, which may not appear as extreme points in the $\mathbf{z}_i$'s vectors, will be deleted due to their large residuals. The high-leverage outliers will correspond to the extreme values of principal sensitivity components $\mathbf{z}_i$'s. The iterations are similar to a reweighting algorithm for computing M estimators.

The estimate at the ith iteration, $\hat{\boldsymbol{\beta}}^{(i)}$, is used to identify outliers that will be omitted in construction of the set A_{i+1} for the next iteration. Because in the first iteration we do not have yet an estimate with some degree of robustness, this step is not carried out in the first iteration, to avoid swamping.

The fraction of extreme observations on the principal sensitivity components to be deleted may be different from .5. For instance, if the model contains many dummies variables, then deleting half of the observation may easily produce a singular matrix. In this case we recommend deleting a smaller fraction of the data, although of course this will affect the robustness of the procedure.

It is also possible to use this first-stage estimate as the starting value in the reweighted least squares algorithm to compute an S estimate (see, e.g., the algorithm proposed in Yohai, Stahel, and Zamar 1991). The resulting estimate will have the asymptotic normal distribution found by Rousseeuw and Yohai (1984) for S estimators, assuming regression errors with finite variance, and it will be the one to use in the second stage.

Stage 2. Following a suggestion by Rousseeuw (1984), to gain efficiency we define a new estimator as a one-step iteration of the initial one computed in stage 1. We compute the residuals $e_j = y_j - \hat{\boldsymbol{\beta}}_1' \mathbf{x}_j, 1 \le j \le n$, and a robust scale s of the e_j's. Then we eliminate all of the observations j such that $|e_j| > C_2 s$. Let n_1 be the number of observations eliminated and let $(\mathbf{y}_2, \mathbf{X}_2)$ be the sample with the $n - n_1$ remaining observations. We compute the LSE, $\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}_2$, and test the n_1 points previously eliminated by using the studentized out-of-sample residual $t_j = (y_j - \hat{\boldsymbol{\beta}}_2' \mathbf{x}_j)/\hat{s}_2\sqrt{(1 + h_j)}$, where $\hat{s}_2^2 = \sum (y_j - \hat{\boldsymbol{\beta}}_2' \mathbf{x}_j)^2/(n - n_1 - p)$ and $h_j = \mathbf{x}_j'(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{x}_j$. Each observation in the set of n_1 points is finally eliminated and considered as an outlier if $|t_j| > C_3$. With the observations that are not deleted, we compute the LSE, $\hat{\boldsymbol{\beta}}$, that will be the final estimate.

Comments on Stage 2. In our Monte Carlo study of Section 5, in which the sample size is small, we have used $C_2 = C_3 = 2.5$. In general, for large sample sizes we recommend increasing these constants. Note that the results of He and Portnoy (1990) show that one-step reweighting does not change the order of convergence of the initial estimate. In our case the initial estimate is an S estimate with order of convergence $n^{-1/2}$ and normal asymptotic distribution. Therefore, the one-step reweighted distribution also converges to a normal distribution with order of convergence $n^{-1/2}$. However, the asymptotic variances of both estimates may be very different. In our case the asymptotic efficiency of the initial estimate (with respect to the least squares estimate) is .24. The asymptotic covariance matrix of the one-step reweighted estimate can be computed using straightforward Taylor expansions. It was found that the asymptotic relative efficiency of this estimate for normal errors is .88.

4. PROPERTIES OF THE PROCEDURE

The estimate computed by the procedure is affine, regression, and scale equivariant. That is, consider a vector of responses $\mathbf{y}$ and a matrix of explanatory variables $\mathbf{X}$, and suppose that we transform these variables by $\mathbf{y}^* = a\mathbf{y} + \mathbf{X}\boldsymbol{\gamma}$ and $\mathbf{X}^* = \mathbf{X}\mathbf{E}$, where a is a scalar, $\boldsymbol{\gamma} \in R^p$, and $\mathbf{E}$ is an $p \times p$ nonsingular matrix. Let $\hat{\boldsymbol{\beta}}$ be the estimate based on $\mathbf{y}$ and $\mathbf{X}$ and $\hat{\boldsymbol{\beta}}^*$ be the estimate based on $\mathbf{y}^*$ and $\mathbf{X}^*$; then $\hat{\boldsymbol{\beta}} = a\mathbf{E}^{-1}(\hat{\boldsymbol{\beta}} + \boldsymbol{\gamma})$.

The following theorem, proved in the Appendix, establishes that if $m < n - p + 1$ high-leverage identical outliers are added to the good n data points, then either the LSE $\hat{\boldsymbol{\beta}}$ is bounded or the proposed procedure will detect the outliers. In fact, according to the theorem, either the LSE is bounded or, at least for one eigenvector, the coordinates corresponding to the outliers will have absolute value larger than the median. Then a fraction of any high-leverage identical outliers smaller than $(n - p + 1)/(2n - p + 1)$ will keep the estimate uniformly bounded.

Theorem 1. Consider a set of regression observations $\mathbf{z}_1 = (y_1, \mathbf{x}_1), \ldots, \mathbf{z}_n = (y_n, \mathbf{x}_n)$, where $\mathbf{x}_i = (x_{i,1}, \ldots x_{i,p})', 1 \le i \le n$, are in general position; that is, any p arbitrary points $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_p}$ are linearly independent. Suppose that we add to the sample m identical arbitrary data points $\mathbf{z}_{n+i} = (y_{n+i}, \mathbf{x}_{n+i}) = (y^*, \mathbf{x}^*), \mathbf{x}^* = (x_1^*, \ldots, x_p^*)', i = 1, \ldots, m$. Then, given $m < n - p + 1$, there exists M such that $\|\hat{\boldsymbol{\beta}}\| > M$ and $\|\mathbf{x}^*\| > M$ imply that for any set $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_p\}, \mathbf{v}_i = (v_{i,1}, \ldots, v_{i,n}, v_i^*, \ldots, v_i^*)$ of orthogonal eigenvectors of $\mathbf{HW}^2$, we have

$$\max_{1 \le i \le p} \#\{j\colon\ 1 \le j \le n, |v_{i,j}| < |v_j^*|\} > \frac{m + n}{2}.$$

We could not prove a similar result for moderate- or low-leverage outliers. However, the results of the simulations in Section 6 indicate that the procedure is able to cope with these types of outliers as well.

5. COMPARISON WITH RELATED PROCEDURES

Peña and Yohai (1995) proposed a procedure to identify outliers in regression based on the eigenvectors of the matrix $\mathbf{M}$ (using a matrix that includes an scalar that does not affect the analysis based on eigenvectors) defined by (8). The eigenvectors $\mathbf{z}_i, 1 \le i \le n$, of $\mathbf{P}$ proposed in this article are related to the eigenvectors $\mathbf{v}_i, 1 \le i \le n$ of $\mathbf{M}$ used

in the our earlier procedure (Peña and Yohai 1995) by

$$\mathbf{v}_i = \mathbf{R}'\mathbf{z}_i,$$

where $\mathbf{R} = \mathbf{HW}$. Because $\mathbf{Hz}_i = \mathbf{z}_i$, we have that

$$\mathbf{v}_i = \mathbf{Wz}_i. \tag{22}$$

It can be shown that if instead of looking for projections where the $\mathbf{r}_i$'s are largest (as we proposed in Section 2), when we do the same analysis but using the vectors $\mathbf{t}_i$'s or $\boldsymbol{\gamma}_i$'s, we will get the directions $\mathbf{v}_i$'s. This result is immediate for the $\mathbf{t}_i$'s. To show this result for the $\boldsymbol{\gamma}_i$'s, consider the eigenvectors $\mathbf{u}_i$ that satisfy (17). The projections of the $\boldsymbol{\gamma}_j$'s on $\mathbf{u}_i$ give the vector $\mathbf{g}_i = \Gamma\mathbf{u}_i$, where Γ is given by (14). Multiplying (17) by Γ, we get

$$\mathbf{WHW}\mathbf{g}_i = \lambda_i\mathbf{g}_i,$$

and the $\mathbf{g}_i$'s are the eigenvectors of the matrix $\mathbf{M}$; that is, they are the same as the $\mathbf{v}_i$ except for a scalar factor. Therefore, the our earlier procedure (Peña and Yohai 1995) can be interpreted as (a) finding the uncentered covariance matrix of the standardized effects on the regression coefficients $\boldsymbol{\gamma}_i, (\mathbf{Q})$ or the corresponding one for the $\mathbf{t}_i(\mathbf{M})$; (b) obtaining the eigenvectors of any of these covariance matrices; (c) projecting the $\boldsymbol{\gamma}_i$ or the $\mathbf{t}_i$ on these principal directions; and (d) searching for extreme coordinates on these projections.

The procedure proposed in this article can be seen in two alternative ways. The first interpretation includes the four steps (a)–(d) above using the $\mathbf{r}_i$. The second involves steps (a) and (b) with the $\boldsymbol{\gamma}_i$, but in step (c) the $\mathbf{x}_i$ vectors are projected over the directions $\mathbf{u}$ found in step (b). By projecting the X variables over the directions of maximum change on the regression coefficients, we analyze observations whose forecasts are more sensitive to changes in the parameters. As masking is especially produced by high-leverage observations, this may explain the better results obtained in the simulations and the examples with the procedure proposed in this article.

The relationship between the eigenvectors of $\mathbf{M}$ and $\mathbf{P}$ given by (22) indicates why our procedure (Peña and Yohai 1995) may fail when the number of outliers is high. Suppose that we have a set of identical high leverage outliers. Then, as we showed (Peña and Yohai 1995), the individual leverage of each point may be small, whereas the residual may be very close to 0. This implies that the absolute value of $\mathbf{W}_i = e_i/(1 - h_{ii})$ corresponding to these points may be very small. Then, according to (22), they may not appear as extremes in the $\mathbf{v}_i$ vectors, whereas they can be clearly extreme points in the principal directions $\mathbf{z}_i$. Peña and Yohai (1995) showed that inspection of the $\mathbf{v}_i$'s allows the detection of outliers in a case of extreme masking. By (22), we can conclude that the $\mathbf{z}_i$'s will also reveal the groups of outliers in this case.

Cook and Weisberg (1982) also considered the vector $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$ as a sample of p-dimensional vectors and suggested using Wilk's (1963) criterion for detecting a single outlier in a multivariate sample. They found that according to this criterion, the observations can be ordered by (Cook and Weisberg 1982, p. 130)

$$\delta_i^2 = \frac{e_i^2}{(1 - h_{ii})^2}\ \mathbf{x}_i' \left[\sum \frac{e_j^2}{(1 - h_{jj})^2}\ \mathbf{x}_j\mathbf{x}_j'\right]^{-1} \mathbf{x}_i;$$

that is, their procedure is equivalent to finding the largest element in the vector

$$\boldsymbol{\delta} = \mathbf{W}^2\ \mathrm{diag}(\mathbf{X}(\mathbf{X}'\mathbf{W}^2\mathbf{X})^{-1}\mathbf{X}'),$$

where diag (A) is a vector with the diagonal elements of A as components.

Finally, Jorgensen (1992) has studied a related problem using the eigenvectors of a modified $\mathbf{H}$ matrix. He proposed finding rank leverage subsets by looking at the eigenvectors of the matrix $\mathbf{L} = \mathbf{H}S^{-1}\mathbf{H}$, where $S = \mathrm{diag}(h_{11}, \ldots, h_{nn})$. The method is exploratory, and Jorgensen did not intend to present a procedure for detecting outliers.

6. EXAMPLES AND MONTE CARLO RESULTS

6.1 Examples

The procedure proposed in this article has been tested with many examples. We tried it with all of the examples of Rousseeuw and Leroy (1987); in all the cases we got an estimate very close to the LMSE. We have presented some of these examples in earlier work (Peña and Yohai 1996).

In this section we apply the procedure to a large dataset from the Spanish household budget survey, Encuesta de Presupuestos Familiares (EPF) collected from the Spanish Statistical Office (INE). This illustrates the procedure's performance with a large number of explanatory variables. The household members included in the sample are supposed to record all expenditures during a sample week. In the last EPF, April 1990–March 1991, the INE collected information about bulk food purchases on or during the three previous weeks. For some households no bulk purchases were observed, whereas for others bulk purchases were observed in the sample week and/or during the three previous weeks. The INE did not take into account the information about bulk purchases on the three previous weeks to the sample week, to estimate the household's annual food expenditures. Therefore, the food expenditure will be underestimated for some groups of households and overestimated for others, and this effect may produce groups of masked outliers. Peña and Ruiz-Castillo (1998) applied the procedure of Peña and Yohai (1995) to identify groups of outliers in a regression in which the proportion of expenditure allocated to food is explained by a set of 55 household variables (many of them dummy variables that take into account the age structure, education, location of the household, and so on), and they proposed better estimation methods for annual food expenditure. The analysis was made in the whole EPF sample data, which includes 21,067 households.

To provide a more manageable set of data that can be used easily on a PC to check our procedure and compare it with other methods, we have taken a random sample of 4,000 households from the original dataset. (This dataset, S_1, is available on request from the authors.) Then we have

Table 1. Percentage of Samples With All of the Outliers Detected for p = 3

		$x_0 = 1$				$x_0 = 5$				$x_0 = 10$			
Outliers (%)	*Estimate*	*m = 1*	*m = 2*	*m = 3*	*m = 4*	*m = 1*	*m = 2*	*m = 3*	*m = 4*	*m = 1*	*m = 2*	*m = 3*	*m = 4*
5	PR1	.20	21.20	87.20	98.40	86.20	100.00	100.0	100.0	90.80	100.00	100.00	100.00
	PR2	0	25.20	88.80	99.40	78.00	99.60	100.0	100.0	81.40	99.40	100.00	100.00
	PR3	2.40	45.80	92.80	99.40	81.20	99.40	100.0	100.0	85.60	99.40	100.00	100.00
	PR4	1.40	28.20	84.60	97.40	77.00	99.00	99.60	100.0	81.40	98.60	100.0	100.00
10	PR1	0	12.60	76.80	97.60	70.60	99.40	100.0	100.0	63.20	99.20	100.00	100.00
	PR2	0	12.80	77.60	97.60	58.00	96.20	99.80	100.0	51.40	97.80	100.00	100.00
	PR3	1.40	32.20	86.60	98.80	47.00	94.00	99.40	100.0	49.40	96.20	99.60	100.00
	PR4	.40	17.40	76.00	95.80	38.00	92.60	99.00	99.80	39.80	93.40	98.60	100.00
15	PR1	0	4.80	62.00	95.60	35.00	95.60	99.80	100.0	25.60	95.00	99.40	100.00
	PR2	0	6.80	61.80	94.60	21.20	88.00	98.60	100.0	15.00	84.60	98.20	100.00
	PR3	.40	16.20	70.40	94.80	11.00	71.60	95.20	99.20	12.00	71.20	95.20	98.80
	PR4	0	8.60	57.20	91.00	7.60	60.00	94.20	97.80	7.60	60.40	91.40	98.60
20	PR1	0	2.80	39.60	89.80	8.40	77.80	97.60	99.80	7.40	75.80	97.00	99.80
	PR2	0	2.60	39.20	87.00	3.00	54.60	92.60	98.80	4.60	50.80	90.80	99.60
	PR3	0	3.80	31.00	67.20	.60	25.80	66.40	91.40	1.20	27.40	73.40	93.40
	PR4	0	2.00	21.00	57.00	.20	19.20	60.00	88.60	.60	18.20	63.60	90.40

fitted the regression model

$$SF = \beta_0 + \beta_1 \ln \text{ PC} + \theta' z + e$$

where SF is the food expenditure share, PC is the per capita household total expenditure, and z is a vector of 53 additional explanatory variables (described in Peña and Ruiz-Castillo 1998). The LSE of β_1 is $-.10$, and if a search for outliers is carried out using the predicted univariate residuals from this regression, 77 points are found with values of this statistic larger than 2.5. However, when we used the estimation procedure proposed in this article, the number of possible outliers almost multiply by 2, because now 151 points have predicted residuals from the robust fit larger than 2.5.

To check the performance of the procedure presented in this article in large datasets with many explanatory variables, we have now modified 3% of the observations as follows. The first 120 data points have been transformed into outliers by changing in these observations the response, SF, and the first explanatory variable, $\ln$ PC. In the initial dataset S_1, the response variable, SF, has a mean of .31 and a standard deviation of .14, and its .0025 percentile is .02. To change these 120 points as outliers, the value of the response has been modified in all of them to .02. A similar modification has been applied to the explanatory variable $\ln$ PC. This variable has in the set S_1 a mean of 13.35 and a standard deviation of .587, and its .0025 percentile is roughly 11.5, which is the value chosen for this variable in the 120 modified points. As the relationship between both variables is negative, changing the variables in the same direction will generate a set of outliers. Note that the rest of the 53 explanatory variables in the regression have been kept as the original ones, and thus the outlier sizes generally will be different. The application of the standard univariate outlier detection techniques to this contaminated dataset leads to the detection of 75 points with predicted residuals larger than 2.5. Neither of them

Table 2. Average of False Outliers for p = 3

		$x_0 = 1$				$x_0 = 5$				$x_0 = 10$			
Outliers (%)	*Estimate*	*m = 1*	*m = 2*	*m = 3*	*m = 4*	*m = 1*	*m = 2*	*m = 3*	*m = 4*	*m = 1*	*m = 2*	*m = 3*	*m = 4*
5	PR1	1.72	1.19	1.46	1.26	1.45	1.30	1.27	1.30	1.55	1.29	1.41	1.30
	PR2	1.98	1.39	1.51	1.41	1.83	1.47	1.40	1.40	1.83	1.47	1.57	1.41
	PR3	3.96	3.20	3.25	3.21	3.78	3.22	3.14	3.04	3.54	3.12	3.31	3.21
	PR4	2.29	1.87	1.94	1.79	2.33	1.96	1.91	1.79	2.20	1.91	1.96	1.81
10	PR1	2.06	.90	.99	.83	1.47	.94	.85	.79	1.98	.95	.88	1.03
	PR2	2.48	1.25	1.16	.99	1.90	1.23	.98	.92	2.57	1.14	.97	1.10
	PR3	5.08	3.43	2.55	2.32	4.46	2.64	2.24	2.19	4.89	2.61	2.35	2.38
	PR4	3.25	2.20	1.56	1.28	2.91	1.64	1.32	1.25	3.57	1.72	1.44	1.37
15	PR1	3.22	1.44	.65	.59	2.95	.81	.58	.49	3.55	.82	.50	.62
	PR2	3.89	2.06	1.05	.68	3.93	1.34	.71	.56	4.46	1.52	.66	.65
	PR3	7.19	5.39	3.29	1.98	7.35	4.02	1.95	1.48	7.45	4.02	1.94	1.84
	PR4	5.02	3.73	2.39	1.23	5.44	3.43	1.34	.91	5.63	3.43	1.46	1.17
20	PR1	5.14	2.70	.96	.34	5.22	1.64	.54	.35	5.56	1.99	.55	.33
	PR2	6.20	4.25	2.29	.82	6.61	3.48	.96	.49	6.42	3.89	1.20	.34
	PR3	9.76	8.50	6.76	4.22	10.06	8.06	4.52	1.97	9.97	7.82	3.93	1.78
	PR4	7.61	6.48	5.36	3.60	8.07	6.73	3.98	1.60	8.11	7.15	3.81	1.56

Table 3. Mean Squared Errors for p = 3

		$x_0 = 1$				$x_0 = 5$				$x_0 = 10$			
Outliers (%)	*Estimate*	*m = 1*	*m = 2*	*m = 3*	*m = 4*	*m = 1*	*m = 2*	*m = 3*	*m = 4*	*m = 1*	*m = 2*	*m = 3*	*m = 4*
5	PR1	.17	.18	.17	.17	.24	.16	.16	.15	.25	.15	.16	.17
	PR2	.19	.20	.18	.17	.30	.17	.17	.15	.33	.18	.17	.18
	PR3	.27	.28	.24	.23	.37	.26	.23	.21	.36	.25	.23	.23
	PR4	.20	.23	.20	.19	.33	.22	.21	.17	.34	.24	.18	.18
	LSE	.12	.14	.16	.22	.48	1.63	3.53	6.14	.86	3.21	7.12	12.65
10	PR1	.23	.27	.23	.17	.40	.19	.16	.16	.55	.20	.16	.15
	PR2	.24	.32	.24	.19	.52	.32	.18	.16	.70	.28	.16	.16
	PR3	.39	.50	.34	.24	.77	.44	.28	.21	.86	.42	.26	.21
	PR4	.29	.43	.30	.22	.72	.45	.29	.20	.89	.50	.33	.17
	LSE	.14	.20	.30	.48	.71	2.56	5.60	9.87	1.02	3.80	8.40	14.83
15	PR1	.36	.52	.35	.24	.83	.35	.18	.16	1.04	.39	.23	.16
	PR2	.40	.61	.46	.26	.97	.68	.30	.17	1.17	.93	.35	.16
	PR3	.61	1.09	1.00	.51	1.36	1.55	.72	.33	1.46	1.78	.79	.50
	PR4	.46	.87	.89	.50	1.22	1.95	.78	.54	1.36	2.16	1.15	.50
	LSE	.16	.30	.51	.85	.84	3.08	6.74	11.88	1.07	4.03	9.02	15.89
20	PR1	.52	.86	.84	.30	1.32	1.20	.42	.19	1.39	1.45	.47	.20
	PR2	.61	1.13	1.34	.63	1.46	2.31	.95	.39	1.47	2.74	1.15	.24
	PR3	.88	1.99	3.01	2.76	1.77	4.32	4.05	1.93	1.81	4.49	3.32	1.71
	PR4	.75	1.53	2.49	2.71	1.60	4.15	4.55	2.45	1.65	4.81	4.29	2.27
	LSE	.18	.40	.74	1.28	.93	3.38	7.54	13.22	1.12	4.18	9.27	16.28

correspond to the true generated outliers, which are completely masked. When the procedure described in Section 3 is applied, a set of 256 outliers are detected that includes the 120 true generated outliers. Also, the value for the robust scale improves by 11% with respect to the one of the LSE fit.

The proportion 3% was chosen to represent a real situation, because in many datasets a proportion larger than this is not expected. However, to check the procedure in a more contaminated situation, we have also modified 10% of the data points. The observations for these 400 points have been changed as before; that is, by using the value .02 for the response variable and 11.5 for the first explanatory variable. Again, although the 400 outliers have the same value for the response variable and the first regressor, they have very different values for the rest of 53 explanatory variables, and in this way we can check the behavior of the procedure with a large number of outliers of different size, leverage, and distribution on the $\mathbf{X}$ space. As before, the univariate procedures fail to identify any member of the set of 400 outliers, and only 51 points (several of them good data points) have a predicted residual larger than 2.5. The 400 generated outliers are completely masked. When the proposed procedure is applied, the 400 outliers are clearly identified, and the improvement on the robust scale is now 17.6%.

Table 4. Median Squared Errors for p = 3

		$x_0 = 1$				$x_0 = 5$				$x_0 = 10$			
Outliers (%)	*Estimate*	*m = 1*	*m = 2*	*m = 3*	*m = 4*	*m = 1*	*m = 2*	*m = 3*	*m = 4*	*m = 1*	*m = 2*	*m = 3*	*m = 4*
5	PR1	.13	.14	.14	.13	.14	.11	.12	.12	.14	.12	.13	.13
	PR2	.14	.14	.13	.13	.17	.12	.13	.12	.16	.12	.14	.12
	PR3	.20	.18	.17	.17	.19	.15	.16	.15	.19	.15	.16	.17
	PR4	.15	.15	.15	.14	.18	.12	.14	.13	.17	.13	.14	.14
	LSE	.10	.11	.13	.19	.45	1.62	3.46	6.06	.85	3.18	7.03	12.56
10	PR1	.16	.20	.16	.12	.18	.13	.12	.12	.22	.12	.12	.11
	PR2	.18	.21	.16	.13	.28	.14	.12	.13	.46	.12	.12	.12
	PR3	.29	.27	.18	.15	.74	.16	.14	.16	.94	.16	.16	.15
	PR4	.22	.23	.18	.14	.76	.15	.12	.13	1.01	.15	.13	.13
	LSE	.11	.18	.27	.43	.69	2.51	5.52	9.82	.98	3.72	8.28	14.58
15	PR1	.28	.32	.21	.13	.92	.14	.12	.13	1.10	.13	.13	.12
	PR2	.31	.37	.22	.14	1.01	.14	.13	.12	1.16	.15	.14	.12
	PR3	.49	.73	.24	.16	1.26	.22	.14	.16	1.32	.23	.17	.15
	PR4	.36	.54	.30	.15	1.13	.29	.13	.13	1.27	.27	.16	.13
	LSE	.13	.28	.48	.78	.81	3.02	6.63	11.73	1.03	3.91	8.83	15.53
20	PR1	.42	.57	.50	.13	1.18	.16	.13	.12	1.24	.18	.14	.12
	PR2	.50	.93	.61	.14	1.28	.31	.13	.13	1.30	.52	.15	.12
	PR3	.68	1.65	2.43	.24	1.54	4.47	.23	.16	1.52	4.83	.23	.14
	PR4	.56	1.23	1.96	.32	1.35	4.29	.25	.15	1.42	4.90	.25	.14
	LSE	.16	.37	.68	1.20	.89	3.33	7.42	13.01	1.07	4.06	9.09	15.96

Table 5. Null Behavior for p = 3

Estimate	PR1	PR2	PR3	M1	LS
Average of false outliers	1.88	1.92	4.15	2.47	
Mean squared error	.16	.17	.24	.19	.11
Median squared error	.12	.12	.16	.13	.10

6.2 Simulation Results

The performance of the procedure was also investigated by Monte Carlo simulation. The model used to generate the data is

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \beta_{p+1} + \varepsilon_i, \qquad 1 \le i \le n,$$

where for $1 \le i \le 40 - n_0$, the vectors $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$ are independent random samples from an $\mathrm{N}((0, 0, 0, \ldots, 0), I)$ and thus correspond to the case $\beta_1 = \beta_2 = \cdots = \beta_{p+1} = 0$. For $n - n_0 + 1 \le i \le n$, the observations are independent samples from an $\mathrm{N}((y_0, x_0, 0, \ldots, 0), .01I)$. This design does not have any loss of generality due to the affine, regression, and scale equivariance of the method and the sphericity of the distribution of the regressors.

Three procedures based on the minimization of a robust scale were applied to estimate the parameters and detect outliers. The first procedure (PR1) is the one described in Section 3, using an M scale S. Given a sample $e_1, \ldots, e_n, S(e_1, \ldots, e_n)$ is defined by the solution of

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{e_i}{S}\right) = b,$$

where ρ is defined by

$$\rho(u) = \begin{cases} 3.048u^2 & \text{if } |u| < .810 \\ 2.763u^8 - 11.783u^6 & \\ \quad + 16.057u^4 - 5.926u^2 & \\ \quad + 1.792 & \text{if } .81 \le |u| \le 1.215 \\ 3.20 & \text{if } |u| > 1.215 \end{cases}$$

and $b = 3.2/2 = 1.6$. The function ρ so defined is twice differentiable, and b was chosen so that S has breakdown point .5. Moreover if u is $\mathrm{N}(0, 1)$, then $E(\rho(u)) = 1$. Therefore, S is Fisher consistent for the standard deviation of a normal variable of mean 0.

The second procedure (PR2) is the same as PR1 but replacing the $\mathbf{z}_i$'s by the $\mathbf{v}_i$'s and, according to the discussion given in Section 4, it is directly related to the our earlier procedure (Peña and Yohai 1995). The third and forth procedures (PR3 and PR4) are based on the LMSE and the LTSE, both computed using the FSA. Then we applied to these estimates the stage 2 of PR1 as described in Section 3. Finally, we also simulated the LSE.

We first consider the case of $n = 40$ and $p = 3$. Although, as we show, the advantage of the procedure appears for a larger number of independent variables, this case can illustrate that for the important case of similar, but not necessarily identical outliers, the procedure can help improve the high-breakdown methods. The values for x_0 were chosen to be 1, 5, and 10, and the contaminating slope, $m = y_0/x_0$, was fixed at 1, 2, 3, and 4. The number of outliers was taken as 2, 4, 6, or 8, corresponding to 5%, 10%, 15%, and 20% contamination. The Monte Carlo study was done with 500 replications. PR3 and PR4 compute the LMSE and the LTSE by using the FSA with 300 random starts.

In Table 1 we show the percentage of Monte Carlo replications where the procedures detect all the outliers. In Table 2 we indicate the average of false outliers found by these procedures. In Table 3 we present the mean squared errors (MSEs), defined as follows. Let $\boldsymbol{\beta}^{(i)}, 1 \le i \le m$ be the estimate corresponding to the replication i of one of the procedures. Then the MSE is given by

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m} \|\boldsymbol{\beta}^{(i)}\|^2,$$

where $\|\quad\|$ denotes Euclidean norm. In Table 4 we show the median square errors (MNSEs), defined by

$$\mathrm{MNSE} = \mathrm{median}\{\|\boldsymbol{\beta}^{(i)}\|^2, 1 \le i \le \mathbf{M}\}$$

In Table 5 we show the null behavior of the different procedures; that is, when the samples do not contain outliers.

Table 1 shows that in the case of low-leverage outliers $(x_0 = 1)$, the more powerful procedure is, in general, PR3. Instead, for higher-leverage outliers $(x_0 = 5, 10)$, PR1 is always better or equal than PR3, which outperforms PR4. Tables 2, 3, and 4 show better behavior of PR1 with respect to PR3.

Table 5 presents the null behavior of the procedure. As would be expected, all robust estimates are less efficient than the LSE in this case, but the most efficient robust pro-

Table 6. Percentage of Samples With All of the Outliers Detected for p = 30

	% Outliers = 10		% Outliers = 15	
Estimate	m = 2	m = 3	m = 2	m = 3
PR1	100	100	98	100
PR2	100	100	80	100
PR3	17	25	1	5

Table 7. Average of False Outliers for p = 30

	% Outliers = 10		% Outliers = 15	
Estimate	m = 2	m = 3	m = 2	m = 3
PR1	6.93	6.82	4.84	4.79
PR2	9.25	8.78	14.52	5.87
PR3	38.69	34.11	55.41	52.61

Table 8. Mean Squared Errors for p = 30

	% Outliers = 10		% Outliers = 15	
Estimate	m = 2	m = 3	m = 2	m = 3
PR1	.28	.28	.42	.28
PR2	.31	.31	1.64	.30
PR3	6.54	12.62	9.36	18.72
LSE	4.31	9.45	4.60	10.11

Table 9. Median Squared Errors for p = 30

	% Outliers = 10		% Outliers = 15	
Estimate	*m = 2*	*m = 3*	*m = 2*	*m = 3*
PR1	.28	.28	.27	.28
PR2	.30	.29	.33	.30
PR3	7.20	14.25	9.00	18.93
LSE	4.29	9.41	4.55	10.09

cedure is PR1. Of course, one can improve the efficiency of the robust estimates increasing C_3, but at the cost of losing robustness and outlier detection power.

To determine the performance of the proposed procedure for the most interesting case of a large number of independent variables, we consider the case of $p = 30$ and $n = 200$. Because the behavior for $p = 3$ of PR3 was superior to PR4, and given the heavy computational load involved in its simulation, this last method was deleted. The PR3 procedure was now computed with 15 random starts. This value may seem too low; however, the computation of each replication took about 5.5 minutes on a 266 MHz Pentium PC using a FORTRAN program, and this prevents us from increasing this number significantly. The computation of each replication of PR1 or PR2 took about 35 seconds using a MATLAB program. This increase in the computational time forced us to make a more limited Monte Carlo study. We made 100 replications for each of the three procedures. In this case we took $x_0 = 10$ and fixed the contaminating slope, $m = y_0/x_0$, at 2 and 3. The number of outliers was taken as 20 and 30, corresponding to 10% and 15% contamination. Tables 6–10 show the results of the Monte Carlo studies for $p = 30$.

Table 6 shows that PR1 is much more powerful than PR3, and the difference between these two procedures increases with the fraction of outliers. Tables 7, 8, and 9 confirm, in general terms, the result of Table 6. Finally, Table 10 shows roughly a loss of efficiency of the robust procedures with respect to least squares similar to the one found in Table 5.

7. CONCLUDING REMARKS

The robust estimate presented herein can be used successfully in regression problems with a large number of explanatory variables where high-breakdown estimates based on the minimization of a robust scale are not feasible with the available computer power. It may also be used to improve these estimates by combining solutions provided by approximate methods (e.g., subsampling) with those generated by our procedure. To be specific, suppose that $\hat{\beta}^{(1)}$ is a solution to the minimization problem

$$\hat{\boldsymbol{\beta}} = \text{ argmin } S(e_1(\boldsymbol{\beta}), \ldots, e_n(\boldsymbol{\beta})),$$

Table 10. Null Behavior for p = 30

	PR1	*PR2*	*PR3*	*LS*
Average of false outliers	14.93	19.95	24.27	
Mean squared error	.31	.37	.42	.19
Median squared error	.31	.37	.41	.18

which has been computed using an approximate procedure. Let $\hat{\beta}^{(2)}$ be the estimate that we propose in the article. Then define

$$\hat{\beta} = \begin{cases} \hat{\beta}^{(1)} & \text{if } S(e_1(\hat{\beta}^{(1)}), \ldots, e_n(\hat{\beta}^{(1)})) < S(e_1(\hat{\beta}^{(2)}), \ldots, \\ & \qquad e_n(\hat{\beta}^{(2)})) \\ \hat{\beta}^{(2)} & \text{if } S(e_1(\hat{\beta}^{(2)}), \ldots, e_n(\hat{\beta}^{(2)}) < S(e_1(\hat{\beta}^{(1)}), \ldots, \\ & \qquad e_n(\hat{\beta}^{(1)})). \end{cases}$$

This estimate will have at least the same breakdown point as $\hat{\beta}^{(1)}$ and, in some cases, will be better with almost no additional computational work. We believe that, in any case, the incorporation of solutions that use information about the structure of the points, as made by the proposed procedure, is a way to improve any resampling scheme.

As it is shown in Tables 2 and 8 of our Monte Carlo study, the estimate proposed in this paper gives directly an useful diagnostic tool to identify multiple outliers. However, the sensitivity components $\mathbf{z}_1, \ldots, \mathbf{z}_p$ can be used directly as a diagnostic method to identify multiple outliers. The procedure will be similar to the one that we described in earlier work (Peña and Yohai 1995), but using these vectors instead of the influence components $\mathbf{v}_1, \ldots, \mathbf{v}_p$. Because the results of our Monte Carlo shows that the $\mathbf{z}_i$'s are more powerful than the $\mathbf{v}_i$'s in detecting outliers, we can expect this change to improve the procedure.

APPENDIX: PROOF OF THEOREM 1

Let $\mathbf{X}_0$ be the $n \times p$ matrix whose ith row is $\mathbf{x}_i'$ and $\mathbf{y}_0 = (y_1, \ldots, y_n)'$. Because of the equivariance of the procedure, we can assume without loss of generality that $V_0 = \mathbf{X}_0'\mathbf{X}_0 = I_p$ and $\mathbf{X}_0'\mathbf{y}_0 = 0$. The latter condition implies that the LSE using these n observations is $\mathbf{0}$. Let $\mathbf{X}$ be the $(n + m) \times p$ matrix whose ith row is $\mathbf{x}_i'$, and $\mathbf{y} = (y_1, \ldots, y_n, y^*, \ldots, y^*)'$. Let $\mathbf{x}^i, 1 \le i \le p$ denote the ith columns of $\mathbf{X}$, and let $\mathcal{V}_{n.m}(\mathbf{x}^*)$ denote the subspace of $\mathbf{R}^{m+n}$ spanned by $\{\mathbf{x}^1, \ldots, \mathbf{x}^p\}$. Observe that the elements of $\mathcal{V}_{n.m}(\mathbf{x}^*)$ have the last m coordinates identical.

It is easy to prove that the LSE is

$$\hat{\boldsymbol{\beta}} = \frac{my^*\mathbf{x}^*}{1 + m\|\mathbf{x}^*\|^2}, \tag{A.1}$$

and we then derive that

$$e^* = y^* - \hat{\boldsymbol{\beta}}'\mathbf{x}^* = \frac{y^*}{1 + m\|\mathbf{x}^*\|^2} \tag{A.2}$$

and

$$e_j = y_j - \hat{\boldsymbol{\beta}}'\mathbf{x} = y_j - \frac{m\mathbf{x}_j'\mathbf{x}^*y^*}{1 + m\|\mathbf{x}^*\|^2}, \qquad 1 \le j \le n. \tag{A.3}$$

Moreover, it also holds that

$$h_{ii} = h^* = \frac{\|\mathbf{x}^*\|^2}{1 + m\|\mathbf{x}^*\|^2}, \qquad n + 1 \le i \le n + m, \tag{A.4}$$

and

$$\lim_{\|\mathbf{x}^*\| \to \infty} h_{ii} = \lim_{\|\mathbf{x}^*\| \to \infty} h^* = \frac{1}{m}, \qquad n + 1 \le i \le n + m, \tag{A.5}$$

and because $\sum_{i=1}^{n+m} h_{ii} = p$, we get

$$\lim_{\|\mathbf{x}^*\| \to \infty} h_{ii} = 0, \qquad 1 \le i \le n. \tag{A.6}$$

Put $r_j = \mathbf{x}_j'\mathbf{x}^*/\|\mathbf{x}^*\|$; because $\mathbf{X}_0'\mathbf{X}_0 = I_p$, it is clear that

$$|r_j| \le 1, \qquad 1 \le j \le n. \tag{A.7}$$

Because the observations $\mathbf{x}_j, 1 \le j \le n$, are in general position, it may be proved that there exists $\gamma > 0$ such that for all $\mathbf{x}^*$,

$$\#\{j\colon\ 1 \le j \le n, |r_j| > \gamma\} \ge n - p + 1. \tag{A.8}$$

In fact, suppose that (A.8) does not hold. Then there exists a sequence $\mathbf{x}_i^*$ such that if we call $\mathbf{a}_i = \mathbf{x}_i^*/\|\mathbf{x}_i^*\|$, then

$$\#\left\{j\colon\ 1 \le j \le n, |\mathbf{a}_i'\mathbf{x}_j| \le \frac{1}{i}\right\} \ge p, \quad \forall\, i.$$

Therefore, because $\|\mathbf{a}_i\| = 1$, and because there exists only a finite number of subsets of $\mathbf{x}_j$'s with p elements, there exists a subsequence i_h and $\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_p}$, such that $\lim_{h\to\infty} \mathbf{a}_{i_h} = \mathbf{a}$ and

$$|\mathbf{a}_{i_h}'\mathbf{x}_{j_k}| \le \frac{1}{h}, \qquad 1 \le k \le p, \quad \forall\, h.$$

Therefore,

$$\lim_{h\to\infty} |\mathbf{a}_{i_h}'\mathbf{x}_{j_k}| = |\mathbf{a}'\mathbf{x}_{j_k}| = 0, \qquad k = 1, \ldots, p,$$

contradicting the fact that the $\mathbf{x}_j$'s are in general position.

Using (A.3) and (A.1), we get

$$e_j = y_j - r_j\|\hat{\boldsymbol{\beta}}\| = \|\hat{\boldsymbol{\beta}}\|\left(\frac{y_j}{\|\hat{\boldsymbol{\beta}}\|} - r_j\right), \tag{A.9}$$

and by (A.2) and (A.1),

$$|e^*| = \frac{\|\hat{\boldsymbol{\beta}}\|}{\|\mathbf{x}^*\|m}. \tag{A.10}$$

Let $\mathbf{F}$ be the diagonal matrix defined by

$$\mathbf{F} = \frac{\mathbf{W}^2}{\|\hat{\boldsymbol{\beta}}\|^2}, \tag{A.11}$$

and denote the first n diagonal elements of $\mathbf{F}$ by f_j, and the last m by f^*. Take

$$\varepsilon = \min\left(\frac{\gamma^2}{48p^3n^2}, \frac{1}{2n^{1/2}}\right). \tag{A.12}$$

We show that there exists $\mathbf{M}_1$ such that if $\|\mathbf{x}^*\| > \mathbf{M}_1$, then there exists $\mathbf{v} = (v_1, \ldots, v_n, v^*, \ldots, v)\mathcal{V}_{n,m}(\mathbf{x}^*)$ such that

$$\|\mathbf{v}\| = 1 \tag{A.13}$$

and

$$|v_i| \le \varepsilon, \qquad i = 1, \ldots, n. \tag{A.14}$$

In fact, take $\mathbf{M}_1 = \sqrt{p}/\varepsilon$. Then, if $\mathbf{x}^* = (x_1^*, \ldots, x_p^*)'$ and $\|\mathbf{x}^*\| > \mathbf{M}_1$, there exists i such that $|x_i^*| > 1/\varepsilon$. Then $\mathbf{x}^i = (x_{1i}, \ldots, x_{ni}, x_i^*, \ldots, x_i^*) \in \mathcal{V}_{n,m}(\mathbf{x}^*)$, and because $|x_{ij}| \le 1$ and $\|\mathbf{x}^i\| \ge 1/\varepsilon$, we obtain that $\mathbf{v} = \mathbf{x}^i/\|\mathbf{x}^i\| \in \mathcal{V}_{n,m}(\mathbf{x}^*)$ and satisfies (A.13) and (A.14).

From (A.13) and (A.14), we obtain that

$$1 = \|\mathbf{v}\|^2 \le n\varepsilon^2 + mv^{*2}, \tag{A.15}$$

and thus, using (A.12), we get

$$v^* \ge \left(\frac{1 - n\varepsilon^2}{m}\right)^{1/2} > \frac{1}{2m^{1/2}} \ge \frac{1}{2n^{1/2}}. \tag{A.16}$$

Moreover, using (A.5), (A.6), (A.9), (A.10), (A.11), and (A.14), if $m > 1$, then there exists $\mathbf{M}_2$ such that if $\|\mathbf{x}^*\| > \mathbf{M}_2$ and $\|\hat{\beta}\| > \mathbf{M}_2$, then

$$\frac{r_j^2}{2} < f_j = \frac{e_j^2}{(1 - h_{ii}^2)\|\hat{\boldsymbol{\beta}}\|^2} = \frac{1}{(1 - h_{ii}^2)}\left(\frac{y_j}{\|\hat{\boldsymbol{\beta}}\|} - r_j\right)^2 < 2, \qquad 1 \le j \le n, \tag{A.17}$$

and

$$f^* = \frac{1}{(1 - h^{*2})m^2\|\mathbf{x}^*\|^2} < \varepsilon. \tag{A.18}$$

Put $\mathbf{M} = \max(\mathbf{M}_1, \mathbf{M}_2)$. In the rest of the proof we assume that $\|\hat{\boldsymbol{\beta}}\| > \mathbf{M}$ and $\|\mathbf{x}^*\| > \mathbf{M}$. Because the eigenvalues of $\mathbf{H}$ are 0 or 1, then $\|\mathbf{HFv}\| \le \|\mathbf{Fv}\|$, and because by (A.14), (A.17), and (A.18), $\|\mathbf{Fv}\| < 3\sqrt{n}\varepsilon$, we get

$$\|\mathbf{HFv}\| < 3n^{1/2}\varepsilon. \tag{A.19}$$

Now let $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ be a set of orthonormal eigenvectors corresponding to the nonnull eigenvalues of $\mathbf{HW}^2$. Then they are also eigenvectors of $\mathbf{HF}$, and the corresponding eigenvalues are denoted by $\lambda_1, \ldots, \lambda_p$. Because V is also a orthonormal base of the eigenvectors of $\mathbf{H}$ corresponding to the eigenvalue 1, and $\mathbf{v}$ belong to this subspace, we can write

$$\mathbf{v} = \sum_{i=1}^{p} \theta_i\mathbf{v}_i, \tag{A.20}$$

and because by (A.13) $|\theta_i| \le 1, 1 \le i \le p$, using (A.16), we get that there is i_0 such that

$$|\theta_{i_0}| \ge \frac{v^*}{p} > \frac{1}{2pn^{1/2}} \tag{A.21}$$

and

$$|v_{i_0}^*| \ge \frac{v^*}{p} > \frac{1}{2pn^{1/2}}. \tag{A.22}$$

Moreover, applying $\mathbf{HF}$ in both sides of (A.20),

$$\mathbf{HFv} = \sum_{i=1}^{p} \theta_i\lambda_i\mathbf{v}_i, \tag{A.23}$$

and by (A.19) and (A.21) we get that $\lambda_{i_0} < 6pn\varepsilon$. Using the fact that $\mathbf{v}_{i_0}$ is also an eigenvector of $\mathbf{H}$ corresponding to the eigenvalue 1, we get

$$|\mathbf{v}_{i_0}'\mathbf{F}\mathbf{v}_{i_0}| = |\mathbf{v}_{i_0}'\mathbf{HF}\mathbf{v}_{i_0}| = \lambda_{i_0}\|\mathbf{v}_{i_0}\|^2 = \lambda_{i_0} < 6pn\varepsilon. \tag{A.24}$$

Now, by (A.17) we get

$$\frac{|v_{i_0,j}|^2 r_j^2}{2} < 6pn\varepsilon, \tag{A.25}$$

and, by (A.8),

$$\#\left\{j\colon\ 1 \le j \le n, |v_{i_0,j}|^2 < \frac{12pn\varepsilon}{\gamma^2}\right\} \ge n - p + 1.$$

Therefore, by (A.12) and (A.22), we get

$$\#\{j\colon\ 1 \le j \le n, |v_{i_0,j}| < |v_j^*|\} \ge n - p > \frac{n+m}{2},$$

and the theorem is proved.

[Received May 1996. Revised October 1998.]

REFERENCES

Atkinson, A. C. (1994), "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, 89, 1329–1339.

Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.

Cook, R. D., and Weisberg, S. (1982), *Residuals and Influence in Regression*, London: Chapman and Hall.

Hadi, A. S., and Simonoff, J. S. (1993), "Procedures for the Identification of Multiple Outliers in Linear Models," *Journal of the American Statistical Association*, 88, 1264–1272.

Hawkins, D. M. (1993), "The Feasible Set Algorithm for Least Median of Squares Regression," *Computational Statistics and Data Analysis*, 16, 81–101.

——— (1994), "The Feasible Set Algorithm for Least Trimmed Squares Regression," *Computational Statistics and Data Analysis*, 17, 95–107.

Hawkins, D. M., Bradu, D., and Kass, G. V. (1984), "Location of Several Outliers in Multiple Regression Data Using Elemental Sets," *Technometrics*, 26, 197–208.

He, X., and Portnoy, S. (1990), "Reweighted LS Estimators Converge at the Same Rate as the Initial Estimator," *The Annals of Statistics*, 20, 2161–2167.

Jorgensen, B. (1992), "Finding Rank Leverage Subsets in Regression," *Scandinavian Journal of Statistics*, 19, 139–156.

Lawrance, J. (1995), "Deletion Influence and Masking in Regression," *Journal of the Royal Statistical Society*, Ser. B, 57, 181–189.

Peña, D., and Ruiz-Castillo, J. (1998), "The Estimation of Food Expenditure From Household Budget Data in the Presence of Bulk Purchases," *Journal of Business and Economic Statistics*, 16, 292–303.

Peña, D., and Yohai, V. J. (1995), "The Detection of Influential Subsets in Linear Regression Using an Influence Matrix," *Journal of the Royal Statistical Society*, Ser. B, 57, 145–156.

——— (1996), "A Procedure for Robust Estimation and Diagnostics in Regression," Working Paper 96-48, Universidad Carlos III de Madrid.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

——— (1993), "A Resampling Design for Computing High-Breakdown Regression," *Statistics and Probability Letters*, 18, 125–128.

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

Rousseeuw, P. J., and Yohai, V. J. (1984), "Robust Regression by Means of S-Estimators," in *Robust and Nonlinear Time Series* (Lectures Notes in Statistics No. 26), eds. J. Franke, W. Hardle, and R. D. Martin, New York: Springer-Verlag, pp. 256–272.

Ruppert, D. (1991), "Computing S-Estimates for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics*, 1, 253–270.

Stronberg, A. (1993), "Computing the Exact Value of the Least Median of Squares Estimate and Stability Diagnostic in Multiple Linear Regression," *Siam Journal of Scientific Computing*, 14, 1289–1299.

Yohai, V. J. (1987), "High Breakdown Point and High Efficiency Robust Estimates for Regression," *The Annals of Statistics*, 15, 642–656.

Yohai, V. J., Stahel, W., and Zamar, R. (1991), "A Procedure for Robust Estimation and Inference in Linear Regression," in *Directions in Robust Statistics and Diagnostics, Part II*, eds. W. Stahel and S. Weisberg, IMA Volumes in Mathematics and its Applications, Vol. 34, New York: Springer-Verlag, pp. 365–374.

Yohai, V. J., and Zamar, R. (1988), "High Breakdown–Point Estimates of Regression by Means of the Minimization of an Efficient Scale," *Journal of the American Statistical Association*, 83, 406–413.

Wilks, S. S. (1963), "Multivariate Statistical Outliers," *Sankhya*, Ser. A, 25, 507–526.