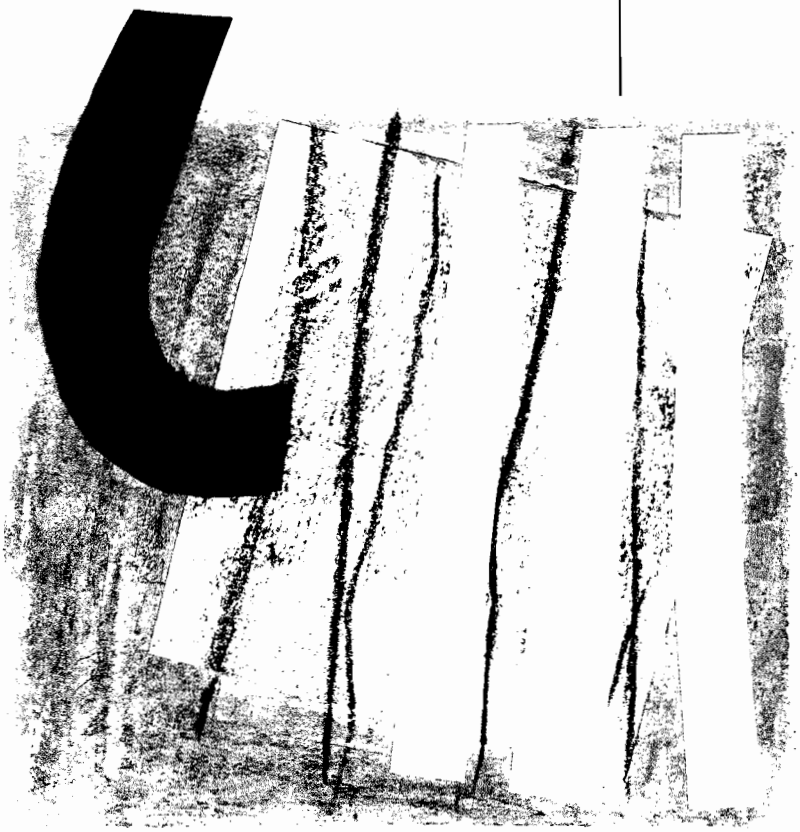
A PROCEDURE FOR ROBUST ESTIMATION AND DIAGNOSTICS IN REGRESSION

Daniel Peña and Victor Yohai

96-48



WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 96-48
Statistics and Econometrics Series 19
December 1996

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

A PROCEDURE FOR ROBUST ESTIMATION AND DIAGNOSTICS IN REGRESSION

Daniel Peña and Victor Yohai*

Abstract

We propose a new procedure for computing an 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 methods based on resampling require an unfeasible or extremely costly computer time. An important advantage of the procedure is that it can be incorporated in any high breakdown procedure and improve it with just a few seconds of computer time. The procedure minimizes the robust scale over a set of tentative parameter vectors. Each of these parameter vector is obtained as follows. We represent each data point by the vector of changes of the least squares forecasts of that observation, when each of the observations is deleted. Then the sets of possible outliers are obtained as the extreme points of the principal components of these vectors, or as the set of points with large residuals. The good performance of the procedure allows the identification of multiple outliers avoiding masking effects. The efficiency of the procedure for robust estimation and its power as an outlier detection tool are investigated in a simulation study and some examples.

Key Words
Masking; outliers; robust regression.

*Universidad Carlos III de Madrid, e-mail: dpena@est-econ.uc3m.es; Universidad de Buenos Aires.

A PROCEDURE FOR ROBUST ESTIMATION AND DIAGNOSTICS IN REGRESSION.

by

Daniel Peña and Victor Yohai
Universidad Carlos III de Madrid Universidad de Buenos Aires

SUMMARY

We propose a new procedure for computing an 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 methods based on resampling require an unfeasible or extremely costly computer time. An important advantage of the procedure is that it can be incorporated in any high breakdown procedure and improve it with just a few seconds of computer time. The procedure minimizes the robust scale over a set of tentative parameter vectors. Each of these parameter vector is obtained by least squares after eliminating a set of possible outliers, which are obtained as follows. We represent each data point by the vector of changes of the least squares forecasts of that observation, when each of the observations is deleted. Then the sets of possible outliers are obtained as the extreme points of the principal components of these vectors, or as the set of points with large residuals. The good performance of the procedure allows the identification of multiple outliers avoiding masking effects. The efficiency of the procedure for robust estimation and its power as an outlier detection tool are investigated in a simulation study and some examples.

Key Words: Masking; Outliers; Robust Regression.

January 1997

1 Introduction

Several robust estimates for regression with high breakdown point have been proposed. We may cite the least median of squares estimate (LMSE) proposed by Rousseeuw (1984), the scale (S) estimates proposed by Rousseeuw and Yohai (1984) the MM–estimates proposed by Yohai (1987) and the tau estimates proposed by Yohai and Zamar (1988). These estimates have a very high computational complexity and therefore the usual algorithms compute only approximate solutions. Rousseeuw (1984) proposed an approximate algorithm based on drawing random subsamples of the same size than the number of carriers. Ruppert (1991) proposed a refinement of this algorithm for S–estimates which 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 which eventually computes the exact the LMSE was proposed by Hawkins (1993). However all these algorithms require a computation time that increases exponentially with the number of independent variables. Therefore they can only be applied when this number is not too large.

In this paper we propose a different type of approximate solution to the high breakdown point estimates mentioned above which can be applied with a large number of independent variables. We do not claim that the approximate procedure we propose keeps the breakdown point of the original estimates. However the procedure succeeds in the detection of groups of outliers in many situations where due to a masking effect, the usual diagnostic procedures fail and the robust estimates require a prohibitive computer time. This is shown by means of a Monte Carlo study and with several classical examples. An important advantage of the procedure proposed in this paper is that it can be incorporated in any high breakdown procedure and improve 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 influence directions that will be used for finding outliers. In Section 3 we present the approximate procedure for the minimization of a robust scale. In Section 4 we prove that the procedure has a breakdown point close to 0.5 when a sample is contaminated with identical high leverage observations. In Section 5 we discuss the relationship of the present procedure and the previous one presented by Peña and Yohai (1995). In Section 6 we illustrate the proposed procedure with some well known examples in the literature and report the results of the Monte Carlo study. Finally, Section 7 contains some concluding remarks.

We assume a regression model with p independent variables (including the constant if there is intercept) and n observations $o_i = (y_i, x_{i,1}, \ldots, x_{1,p}) 1 \leq i \leq n$. Then

$$y_i = \boldsymbol{\beta}' \mathbf{x}_i + \epsilon_i \qquad i = 1, \ldots, n, \tag{1}$$

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 will use the following notation: $\mathbf{y} = (y_1, \ldots, y_n)'$, X is a full rank $n \times p$ matrix whose (i,j) element is $x_{i,j}$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$. Then (1) may be also written as

$$\mathbf{y} = X\boldsymbol{\beta} + \epsilon.$$

All the robust estimates mentioned above, with the exemption of the MM–estimates, are defined throughout the minimization of a certain scale S of the residuals, that is, they are defined by

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

where

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

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

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in A} S(e_1(\boldsymbol{\beta}) \ldots e_n(\boldsymbol{\beta})), \tag{3}$$

where $A = \{\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(N)}\}$ is a finite set. Rousseeuw (1984) proposed obtaining the elements of A by random subsampling. The procedure is as follows: Choose at random N subsamples of p different data points. Let the j–subsample be $\{o_{i_1(j)}, \ldots, o_{i_p(j)}\}$, then $\boldsymbol{\beta}^{(j)}$ is the vector of regression coefficients which fit the p data points, i.e.,

$$y_{i_h(j)} = \boldsymbol{\beta}^{(j)\prime} \mathbf{x}_{i_h(j)}, \quad 1 \leq h \leq p.$$

If p/n is small, it can be shown (see, Rousseeuw and Leroy, 1987) that the probability that the estimate defined by (3) can not break down when there is a fraction of outliers equal to ϵ is approximately given by

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

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

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

This number increases exponentially with p, and therefore the method based on random subsampling can be applied only when p is not very large. For example, when $\epsilon = .5$ and $\alpha = .05$ the method is prohibitively expensive for $p > 20$.

Atkinson (1994) proposed a fast method for the detection of multiple outliers in which a simple forward search from random starting points is shown to be useful to identify outliers. Instead of drawing m basic subsamples Atkinson suggested to draw $h < m$ random subsamples and use least squares estimate (LSE) to fit subsets of size $p, p+1, ..., n$, 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, therefore, the procedure will be not very effective when the number of variables p is large.

In this paper 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 which potentially may provoke 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 influence directions

The principal influence directions will be the directions in which the change on the vector of forecast when each observation is deleted is the largest. More, precisely, let

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

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

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

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

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

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

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

where h_{ij} is the ij–th element of H. 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 observation i: The first is by representing the point i by the vector

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

i.e., looking at the influence on the forecast vector provoked by the deletion of the i–th observation. This is the approach followed by Peña and Yohai (1995). In this paper we explore a second alternative: each data point is represented by the vector

$$\mathbf{r}_i = (\hat{y}_i - \hat{y}_{i(1)}, \ldots, \hat{y}_i - \hat{y}_{i(n)})',$$

i.e., we look at how sensitive is the forecast of the i–th observation to the changes of the estimate of $\boldsymbol{\beta}$ which are induced by the deletion of each of the observation in the sample. This sensitivity depends basically on the leverage of the points. Therefore, in order to look for high leverage outliers we may look at the projections of the $\mathbf{r}_i$'s on the directions v where these effects are the largest. Note that good high leverage points may also appear as extreme points on these directions. However, the key idea of the procedure is to identify high leverage outliers, and the fact that some good high leverage points appear in this stage will not be a difficulty as we shall see in section 3.

The first of these directions is given by

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

subject to $\|\mathbf{v}\| = 1$. The vector $\mathbf{v}_1$ is the eigenvector corresponding to the largest eigenvalue of the matrix $M = \sum_{i=1}^{n} \mathbf{r}_i\mathbf{r}_i'$. From (6), we get that the matrix whose rows are the $\mathbf{r}_i$'s is given by

$$T = HW, \tag{7}$$

where W is the diagonal matrix with terms $e_i/(1 - h_{ii})$. Then we have that

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

Note that the rank of M is p and its ij–th 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 $\mathbf{v}_1$, which is $\mathbf{z}_1 = T\mathbf{v}_1$. It is straightforward to show that $\mathbf{z}_1$ is an eigenvector corresponding to the largest eigenvalue of the matrix P defined by

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

with ij–th element

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

In a similar way, we can search for groups of outliers projecting the $\mathbf{r}_i$'s on the directions of the other eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_n$ of the matrix M, corresponding to the other non null eigenvalues $\lambda_2 \geq ... \geq \lambda_p$. The eigenvector $\mathbf{v}_i$ will have the following property

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

subject to

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

The corresponding projections

$$\mathbf{z}_h = T\mathbf{v}_h, \; h = 2, ...p \tag{12}$$

will be eigenvectors of P. Then, we call the vectors $\mathbf{z}_h$, $1 \leq h \leq p$, principal influence directions.

The principal influence directions form an orthogonal base of the p–dimensional subspace of the eigenvectors of the projection matrix H corresponding to the eigenvalue one. This can be shown by using the definition of the $\mathbf{z}_i$ as eigenvectors of P, i.e., $HW^2H\mathbf{z}_i = \lambda_i\mathbf{z}_i$ and multiplying this equation by H. Note that this base is selected taking into account information about the residuals e_i's. Put $\mathbf{z}_i = (z_{i,1}, ..., z_{i,n})$, since

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

looking for extreme coordinates of each vector $\mathbf{z}_i$ implies a finer analysis than looking at the leverages h_{jj}. Therefore, as mentioned before, it seems reasonable to expect that a group of masked high leverage observations will appear as extreme coordinates in the same projection for at least one of these p orthogonal principal influence directions. In particular, we will show in section 4 that the procedure will identify groups of outliers producing strong masking effect in some extreme cases.

These directions have another related interpretation. Instead of looking at the changes of the forecasts we can try to identify outliers by looking at the standardized changes on the parameters. For this purpose we define the standardized effects on the regression coefficients when deleting observation i by

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

Usually, the influence of observation i–th is summarized by the univariate Cook (1977) statistics

$$D_i = \frac{1}{ps^2}\|\boldsymbol{\gamma}_i\|^2.$$

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, for a recent analysis of this problem). In this situation, masked outliers will have similar effects on the estimated parameter $\boldsymbol{\beta}$ and there will be some directions in R^p where 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

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

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

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

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

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

and therefore

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

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

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

which represents the forecast change for each observation in the direction $\mathbf{u}_i$. Note that although the eigenvectors of Q are defined up to an orthogonal transformation (this property is inherited from the similar property of $(X'X)^{-1/2}$), 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.

It is expected that the forecasting of high leverage outlier observations will be sensitive to changes in the regression coefficients. These changes are especially large in the directions of the eigenvectors linked to large eigenvalues of the Q matrix. Therefore, the coordinates that correspond to these outliers are expected to appear as extreme ones in the vectors $\mathbf{Z}_i$'s. Let us now show that the $\mathbf{Z}_i$'s are also the eigenvectors of the P matrix defined in (9) and then they are equal (except by a scalar factor) to the $\mathbf{z}_i$'s .

Since $\mathbf{u}_i$ is an eigenvector of Q, using (15) we get,

$$(X'X)^{-1/2}(X'W^2X)(X'X)^{-1/2}\mathbf{u}_i = \lambda_i\mathbf{u}_i. \tag{16}$$

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

$$HW^2\mathbf{Z}_i = \lambda_i \mathbf{Z}_i. \tag{17}$$

and multiplying this last equation by H we obtain

$$HW^2\mathbf{Z}_i = \lambda_i H\mathbf{Z}_i. \tag{18}$$

Comparing (17) and (18) $\mathbf{Z}_i = H\mathbf{Z}_i$. Replacing this result in the left hand of (17) we obtain that $\mathbf{Z}_i$ is the eigenvector of HW^2H corresponding to the eigenvalue λ_i.

3 The Procedure

The procedure suggested here has two stages. In the first stage we find a robust estimate using the criterion of minimizing a robust scale of the residuals over a finite set A according to (3). In the second stage we find a more efficient estimate eliminating observations which large residuals and applying the LSE to the remaining data points. The points deleted are tested using the studentized residual for outlyingness and a final estimate is computed by LSE using the cleaned sample.

Stage 1: Initial estimate. We propose an iterative procedure to compute the initial estimate of $\boldsymbol{\beta}$. In the i–th iteration the estimate $\hat{\boldsymbol{\beta}}^{(i)}$ is defined by

$$\hat{\boldsymbol{\beta}}^{(i)} = \arg\min_{\boldsymbol{\beta}\in A_i} S(e_1(\boldsymbol{\beta}), \ldots, e_n(\boldsymbol{\beta})),$$

where the set A_i is modified at each iteration and has $3p+2$ elements, except for $i=1$ where A_1 has $3p+1$ elements. The estimator $\hat{\boldsymbol{\beta}}^{(i)}$ is used to identify outliers, and these outliers will be omitted in the construction of the set A_{i+1} for the next iteration. The procedure ends when $\hat{\boldsymbol{\beta}}^{(i+1)} = \hat{\boldsymbol{\beta}}^{(i)}$.

We now describe how to determine the set A_i for each iteration. Except for $i=1$, we start deleting outliers using $\hat{\boldsymbol{\beta}}^{(i-1)}$. Let $e^{(i)} = y - X\hat{\boldsymbol{\beta}}^{(i-1)}$ be the residuals using the whole data set, and let $s^{(i-1)}$ be its corresponding robust scale. Then we delete all the observations j such that

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

As the objective of this stage is to obtain a preliminary robust estimate the value of C_1 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.

Then, with the remaining observations (for $i=1$ with all the observations) we compute the LSE, which will also be an element of A_i, and the principal influence directions $\mathbf{z}_j$, $j = 1, ..., p$ that are the eigenvectors of the P matrix defined by (9). For each vector $\mathbf{z}_j$, $1 \leq j \leq p$ we compute three estimates by LS, 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 and the third eliminating the half corresponding to the largest absolute values. Finally, for $i > 1$ $\hat{\boldsymbol{\beta}}^{(i-1)}$ is also included in A_i. The estimate that minimizes the robust scale on this stage will be called $\hat{\boldsymbol{\beta}}_1$.

Note that this stage includes two mechanisms which allow the elimination of the effects of the outliers. The high leverage outliers will correspond to the extreme values of the projection of x variables on the set of orthogonal directions $\mathbf{u}_i$. The low leverage outliers which do not appear as extreme points in the $z_i's$ vectors will be deleted due to their large residuals. The iterations are similar to a reweighting algorithm to compute M estimators.

An interesting point is that the estimate computed in this way is affine, regression and scale equivariant. That is, consider a vector of responses $\mathbf{y}$ and a matrix of explanatory variables X, and suppose we transform these variables by $\mathbf{y}^* = a\mathbf{y} + X\gamma$ and $X^* = XA$, where a is a scalar, $\gamma \in R^p$ and A is an $p \times p$ non singular matrix. Let $\hat{\beta}$ the estimate based on $\mathbf{y}$ and X and $\hat{\beta}^*$ the one based on $\mathbf{y}^*$ and X^*, then $\hat{\beta} = aA^{-1}(\hat{\beta} + \gamma)$.

Stage 2: Final estimate. Following a suggestion by Rousseeuw (1984), in order 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 $e_j = y_j - \hat{\beta}_1' \mathbf{x}_j$, $1 \leq j \leq n$ and a robust scale s of the e_j's. Then we eliminate all the observations j such that $|e_j| > C_2 s$. Let n_1 be the number of observations eliminated and let $(\mathbf{y}_2, X_2)$ be the sample with the $n - n_1$ remaining observations. We compute the LSE, $\hat{\beta}_2 = (X_2'X_2)^{-1}X_2'\mathbf{y}_2$ and test the n_1 points previously eliminated by using the studentized out of sample residual $t_j = (y_j - \hat{\beta}_2' \mathbf{x}_j)/\hat{s}_2\sqrt{1 + h_j}$, where $\hat{s}_2^2 = \sum(y_j - \hat{\beta}_2' \mathbf{x}_j)^2/(n - n_1 - p)$ and $h_j = \mathbf{x}_j'(X_2'X_2)^{-1}\mathbf{x}_j$. Each observation in the set of n_1 points is finally eliminated and considered as an outliers if $|t_j| > C_3$. With the observations that are not deleted we compute the LSE, $\hat{\beta}$, that will be the final estimate. In our Monte Carlo study of section 5 we have used $C_2 = 2.5$ and $C_3 = 3$. We have also tried with $C_3 = 2.5$ which leads to a more powerful procedure but worsen its null behavior.

4 A high breakdown point property of the procedure

The following Lemma, 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 $\hat{\beta}$ is bounded or the proposed procedure will detect the outliers. In fact in this case at least for one eigenvector the coordinates corresponding to the outliers will have absolute value larger than the median. Then the breakdown point for identical high leverage outlier is at least $n/(2n - p + 1)$.

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

Consider a set of regression observations $\mathbf{z}_1 = (y_1, \mathbf{x}_1), ..., \mathbf{z}_n = (y_n, \mathbf{x}_n)$, where $\mathbf{x}_i = (x_{i,1}, ...x_{i,p})'$. Let X_0 be the $n \times p$ matrix whose $i-$th row is $\mathbf{x}_i'$ and $\mathbf{y}_0 = (y_1, ..., y_n)'$. Because of the equivariance of the procedure we can assume without loss of generality that $V_0 = X_0'X_0 = I_p$ and $X_0'\mathbf{y}_0 = 0$. The latter condition imply that the LSE using these n observations is $\mathbf{0}$. We are going to add to the sample m identical arbitrary data points $\mathbf{z}_{n+i} = (y_{n+i}, \mathbf{x}_{n+i}) = (y^*, \mathbf{x}^*), \mathbf{x}^* = (x_1^*, ..., x_p^*)'$, $i = 1, ..., m$. Let X be the $(n+m) \times p$ matrix whose i-th row is $\mathbf{x}_i'$, and $\mathbf{y} = (y_1, ..., y_n, y^*, ..., y^*)'$. We denote by $\mathbf{x}^i, 1 \leq i \leq p$ the i-th columns of X, and by $\mathcal{V}_{n,m}(\mathbf{x}^*)$ the subspace of R^{m+n} spanned by $\{\mathbf{x}^1, ..., \mathbf{x}^p\}$. Observe that the elements of $\mathcal{V}_{n,m}(\mathbf{x}^*)$ have the last m coordinates identical.

Lemma: Suppose that the observations $\mathbf{x}_1, ..., \mathbf{x}_n$ are in general position, i.e., any p arbitrary points $\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_p}$ are lineally independent. Then given $m < n - p + 1$, there exists M such that

$||\hat{\beta}|| > M$ and $||\mathbf{x}^*|| > M$, imply that for any set $V = \{\mathbf{v}_1, ..., \mathbf{v}_p\}$, $\mathbf{v}_i = (v_{i,1}, ..., v_{i,n}, v_i^*, ..., v_i^*)$ of orthogonal eigenvectors of HW^2 we have

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

5 Comparison with Other Procedures for Outlier Detection

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

$$\mathbf{v}_i = T'\mathbf{z}_i,$$

where $T = HW$. Since $H\mathbf{z}_i = \mathbf{z}_i$, we have that

$$\mathbf{v}_i = W\mathbf{z}_i. \tag{19}$$

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, we do the same analysis but using the vectors $\mathbf{t}_i$'s or $\boldsymbol{\gamma}_i$'s indistinctly, we will get the directions $\mathbf{v}_i$'s. This 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 verify equation (16). 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 equation (16) by Γ we get

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

and the $\mathbf{g}_i$'s are the eigenvectors of the matrix M, i.e., they are the same as the $\mathbf{v}_i$ except for a scalar factor. Therefore the procedure by Peña and Yohai (1995) can be interpreted as: (i) Finding the uncentered covariance matrix of the standardized effects on the regression coefficients $\boldsymbol{\gamma}_i$, (Q) or the corresponding for the $\mathbf{t}_i$ (M), indistinctly; (ii) Obtaining the eigenvectors of any of these covariance matrices ; (iii) Projecting the $\boldsymbol{\gamma}_i$ or the $\mathbf{t}_i$ on these principal directions; (iv) Searching for extreme coordinates on these projections.

The procedure proposed in this paper can be seen in two alternative ways. The first interpretation does the four steps (i) to (iv) above using the $\mathbf{r}_i$. The second one does steps (i) and (ii) with the $\boldsymbol{\gamma}_i$, but in (iii) the $\mathbf{x}_i$ vectors are projected over the directions $\mathbf{u}$ found in step (ii). 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 paper.

The relationship between the eigenvectors of M and P given by (19) indicates why the procedure of 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 shown by Peña and Yohai (1995) the individual leverage of each point may be small, whereas the residual may be very close to zero.

This implies that the absolute value of $W_i = e_i/(1-h_{ii})$ corresponding to these point may be very small. Then according to (19) 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 (19) we can conclude that the $\mathbf{z}_i$'s will also reveal the groups of outliers in this case.

Cook and Weisberg (1982) considered also the vector $\hat{\beta} - \hat{\beta}_{(i)}$ 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, page 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

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

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

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 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 idea for the success of the method is to obtain a clean initial subset of data. However, this assumption may fail when the sample contains a set of several high leverage outliers. In this case the usual diagnostic statistics are expected to be very small at these points,and the set of outliers will be included with high probability in the initial subset. Then, when this occurs, the procedure will fail. This is confirmed in our Monte Carlo study of section 5.

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

6 Examples and Monte Carlo results

The procedure proposed in this paper has been tested with many examples. We try with all the examples in Rousseeuw and Leroy, and in all the cases we get an estimat close to the LMSE. Here we present four examples, the first three are simple regression examples and the fourth has three explanatory variables plus the intercept. The first two examples are the Number of International Telephone Calls (NITC) data and the Hertzsprung–Rusell Diagram (HRD) and are found in Rousseeuw and Leroy (1987). The third and the fourth are set of simulated data proposed by Rousseeuw (1984), and by Hawkins, Bradu and Kass (1984) (HBK). In all of them

the procedure presented in Section 3 is able to identify the outliers. Figure 1, 2 and 3 present the data and the regression line fitted with our procedure for the NITC, the HRD and the Rousseeuw simulated data. These figures show that the final robust estimate is not affected by the set of outliers. Figure 4 presents the residuals standardized for the MAD scale from the robust fit for HBK data. The first ten observations, that correspond to the outliers, show up very clearly. Note that observations 11 to 14, that correspond to good leverage points, have a small standardized residuals.

(Figure 1,2,3,4 around here)

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

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \beta_{p+1} + \epsilon_i, \quad 1 \leq i \leq n$$

where for $1 \leq i \leq 40 - n_0$ the vectors $G_i = (y_i, x_{i1}, x_{i2}...., x_{ip}, \epsilon_i)$ are independent random samples from a $N((0,0,0,...,0), I)$ and, therefore, correspond to the case $\beta_1 = \beta_2 = ... = \beta_{p+1} = 0$. For $n - n_0 + 1 \leq i \leq n$, the G_i's are independent samples from a $N((y_0, x_0, 0, ..., 0), 0.01I)$. This design does not suppose any loss of generality due to the affine, regression and scale equivariance of the method and the sphericity of the distribution of the x variables.

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 as S a τ scale defined as follows. Let S_0 be the MAD scale, i.e.

$$S_0(e_1, ..., e_n) = \frac{1}{0.6745}\text{median}\{|e_1|, ..., |e_n|\}$$

Then

$$S_(e_1, ..., e_n) = S_0^2 \frac{1}{n} \sum_{i=1}^{n} \rho_k \left(\frac{e_i}{S_0}\right)$$

and

$$\rho_k(e) = \min(e^2, k^2).$$

The τ-scales were introduced by Yohai and Zamar to obtain estimates which combine high efficiency under normality and high breakdown point. The value of k used here is 2.5.

The second procedure (PR2) is the same as PR1 by replacing the z_i's by the v_i's and, according to the discussion given in Section 4, it is directly related to the procedure given in Peña and Yohai (1995).

The third procedure (PR3) is based on the LMSE computed by random subsampling as proposed by Rousseeuw (1984). A first estimate is computed by (3) with $S(e_1, ..., e_n) = \text{median}(|e_1|^2, ..., |e_n|^2)$ and the set generated by N subsamples. Then we apply to this estimate the stage 2 of PR1 as described in Section 3.

We also simulated the procedure M1 proposed by Hadi and Simonoff (1993) choosing the initial subset by means of the Cook statistics. In order to make it comparable to the other three procedures, the threshold for the t-statistics was set at 3. Therefore, the sets of rejected observations are larger than those obtained with thresholds based on the Bonferroni inequality as proposed by the authors. Finally we also simulated the LSE.

We consider first the case of $n = 40$ and $p = 3$. 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. For the PR3, the LMSE was computed using 500 subsamples.

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 square errors defined as follows. Let $\boldsymbol{\beta}^{(i)}$, $1 \leq i \leq m$ be the estimate corresponding to the replication i of one of the procedures. Then the MSE is given by

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

where $||\ ||$ denotes Euclidean norm. In Table 4 we show the median square errors defined by

$$\text{MNSE} = \text{median}\{||\boldsymbol{\beta}^{(i)}||^2,\ 1 \leq i \leq M\}$$

In Table 5 we show the null behavior of the different procedures, i.e., when the samples do not contain outliers

To determine the performance of the proposed procedure for large number of independent variables we consider also the case of $p = 30$ and $n = 200$. Due to the bad behavior for $p = 3$, the procedure M1 was not simulated for $p = 30$. The procedure based on the LMSE (PR3) was computed with 5000 subsamples, increasing around 40 times the computing time. This forced us to make a more limited Monte Carlo study. In this case we take $x_0 = 10$ and the contaminating slope, $m = y_0/x_0$, was fixed at 1, 1.5, 2 and 3. The number of outliers was taken as 20 and 30, corresponding to 10% and 15% contamination. The Monte Carlo study was done with 100 replications. Tables 6–10 show the results of the Monte Carlo for $p = 30$

Tables 1 and 2 shows that the procedure PR1 proposed in the paper works quite well and outperforms all the others when the fraction of outliers increases. In the case of low leverage outliers ($x_0 = 1$) all procedures have a very small power for small outlier size ($m = 1, 2$),and although procedure M1 has the highest power it also has the largest detection of false outliers (see Table 2). When the size of the low leverage outliers increases ($m = 3, 4$) PR1 is better than PR3 in seven out of eight cases. Also it always detects a smaller number of false outliers. For moderate or high leverage outliers ($x_0 = 5, 10$) M1 does not work. For instance, for $x_0 = 10, m = 4$,and 20% of outliers it has half the power of the other procedures and detects more than four times the average number of false outliers. In the 32 comparisons made in Table 1 for moderate or high outliers the procedure proposed in this paper is the best or among the best in 30 cases, and, in many cases the difference with the other ones is important. For instance, when m=2 and the fraction of outliers is 20% it can double the power of the LMSE computed by resampling with N=500. The two cases in which it is not the most powerful correspond to the situation in which m is relatively small, that is the outliers have a small influence. Then all the procedures have a small power. Besides, when m=1, it should be noted that the power of all the procedures computed with the robust scale in Table 1 are smaller than the power of the method presented in Table 1 of Peña and Yohai (1995). This is related to the asymptotic maximum bias of any estimate based on the minimization of a robust scale. When the outliers have a slope, m, smaller than the asymptotic maximum bias of the corresponding contamination and, at the same time, they have high leverage, the estimate based on a robust scale fits the outliers exactly, and

therefore it cannot identify them. For instance, with 15% of outliers the maximum bias of the LMSE is 1.07, and for 20% is 1.52. This means that the LMSE is not expected to detect 10% of outliers with $m = 1$ and large leverage. This results explain the low power of the PR1, PR2 and PR3 which are based on the minimization of a robust scale. On the other hand,it can be seen from Table 1 that PR1 is uniformly more powerful that PR2, which is based on the eigenvectors previously considered by Peña and Yohai (1995).

TABLE 1 ABOUT HERE

Table 2 shows that PR1 has the best performance as far as detecting false outliers in 44 out of the 46 cases considered (95,6% of the cases considered). The two cases in which it is not the best correspond to a small fraction of outliers (5%) of moderate leverage, and then it is slightly improved by M_1. However, when the fraction of outliers increases the overall performance of PR1 in clearly the best.

TABLE 2 ABOUT HERE

Tables 3 and 4 confirm in general terms the results of tables 1 and 2. In Table 3 it can be seen that for $x_0 = 1, m = 1, 2$ the most efficient estimate is LSE, followed by PR1. For $x_0 = 5, 10$, PR1 has the largest efficiency with two exemptions. When $m = 1$ and the fraction of outliers is large (20%) the most efficient estimate is LSE. These results may seem surprising but they are due to the bad performance of estimates based on a robust scale when the outliers have a small slope m, as previously discussed. Table 4 leads to similar conclusions. Again the procedure PR1 improves in this criterium when the fraction of outliers and leverage increase.

TABLES 3 and 4 ABOUT HERE

Table 5 presents the null behavior of the procedure. As it can be expected, all robust estimates are less efficient than the LSE in this case, but the most efficient robust procedure is PR1. Of course, one can improve the efficiency of the robust estimates in this case but at the cost of loosing power when outliers are present. Also the loss of efficiency with respect to LSE with any of the criteria considered is smaller than 10%.

TABLE 5 ABOUT HERE

The usefulness of the proposed procedure is especially clear when the number of explanatory variables is large. 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. For m=1, the power of all estimates is very low, as expected from the previous discussion for p=3. This is consistent with the results of Table 7, where the average of false outliers is large for m=1. Tables 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.

TABLE 6–10 ABOUT HERE

The mean computer time for the different methods in the case of 30 independent variables using a personal computer with a 60 MHz Pentium microprocessor and a MATLAB program were: 55 s. for PR1, 57 s. for PR2 and 3 m. 15 s. for PR3. We have run a small sample of the M1 procedure and obtained a mean of 1 m. 46 s.

7 Concluding Remarks

The robust estimate presented in this paper can be used as an alternative to resampling methods based on the minimization of a robust scale, as the LMSE or tau-estimates. It may be also used to improve them, by combining the solutions provided by resampling with those generated by our procedure. In this way we can apply the robust procedures to regression problems with a large number of explanatory variables where the pure resampling scheme is not feasible with the available computer power. To be specific, suppose that $\widehat{\beta}^{(1)}$ is an approximate solution to the minimization problem

$$\hat{\boldsymbol{\beta}} = \arg\min S(e_1(\boldsymbol{\beta}), \ldots, e_n(\boldsymbol{\beta})),$$

which has been computed using subsampling . Let $\widehat{\beta}^{(2)}$ the estimate we propose in the paper. Then define

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

This estimate will have at least the same breakdown point that $\widehat{\beta}^{(1)}$and, in some cases, will be much 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 the way to improve any resampling scheme.

Of course, a good robust estimate gives directly an useful diagnostic tool to identify multiple outliers. This has been shown in Tables 1 and 7 of our Monte Carlo study. However, the ideas presented in this paper can be used directly as a diagnostic method to identify multiple outliers. The method will be similar to the one described in Peña and Yohai (1995), but using the eigenvectors recommended in this paper that are shown, in the simulation study, to be more powerful to detect outliers than the ones previously suggested.

8 Appendix: Proof of the Lemma in Section 4

Proof: It is easy to prove that the LS-estimate is

$$\hat{\boldsymbol{\beta}} = \frac{m y^* \mathbf{x}^*}{1 + m||\mathbf{x}^*||^2}, \tag{20}$$

and then we derive that

$$u^* = y^* - \hat{\boldsymbol{\beta}}' \mathbf{x}^* = \frac{y^*}{1 + m||\mathbf{x}^*||^2}, \tag{21}$$

and

$$u_j = y_j - \hat{\boldsymbol{\beta}}'\mathbf{x} = y_j - \frac{m\mathbf{x}_j'\mathbf{x}^*y^*}{1+m||\mathbf{x}^*||^2},\ 1 \le j \le n. \tag{22}$$

Moreover it also holds

$$h_{ii} = h^* = \frac{||\mathbf{x}^*||^2}{1+m||\mathbf{x}^*||^2},\ n+1 \le i \le n+m, \tag{23}$$

and then

$$\lim_{||\mathbf{x}^*||\to\infty} h_{ii} = \lim_{||\mathbf{x}^*||\to\infty} h^* = \frac{1}{m},\ n+1 \le i \le n+m, \tag{24}$$

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

$$\lim_{||\mathbf{x}^*||\to\infty} h_{ii} = 0,\ 1 \le i \le n. \tag{25}$$

Put $r_j = \mathbf{x}_j'\mathbf{x}^*/||\mathbf{x}^*||$, since $X_0'X_0 = I_p$ it is clear that

$$|r_j| \le 1,\ 1 \le j \le n. \tag{26}$$

Since 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 $\mathbf{x}^*$

$$\#\{j : 1 \le j \le n,\ |r_j| > \gamma\} \ge n - p + 1. \tag{27}$$

In fact, suppose that (27) 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

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

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

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

therefore

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

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

Using (22) and (20) we get

$$u_j = y_j - r_j||\hat{\boldsymbol{\beta}}|| = ||\hat{\boldsymbol{\beta}}||\left(\frac{y_j}{||\hat{\boldsymbol{\beta}}||} - r_j\right) \tag{28}$$

and by (21) and (20)

$$|u^*| = \frac{||\hat{\boldsymbol{\beta}}||}{||\mathbf{x}^*||m} \tag{29}$$

Let F be the diagonal matrix defined by

$$F = \frac{W^2}{||\hat{\boldsymbol{\beta}}||^2}, \tag{30}$$

and we will denote the first n diagonal elements of F by f_j, and the last m by f^*.

Take

$$\epsilon = \min\left(\frac{\gamma^2}{48p^3n^2}, \frac{1}{2n^{1/2}}\right). \tag{31}$$

We will show that there exists M_1 such that if $||\mathbf{x}^*|| > M_1$, then there exists $\mathbf{v} = (v_1, ..., v_n, v^*, ..., v^*)' \in \mathcal{V}_{n,m}(\mathbf{x}^*)$ such that

$$||\mathbf{v}|| = 1, \tag{32}$$

and

$$|v_i| \leq \epsilon, \ i = 1, ..., n. \tag{33}$$

In fact take $M_1 = \sqrt{p}/\epsilon$, then if $\mathbf{x}^* = (x_1^*, ..., x_p^*)'$ and $||\mathbf{x}^*|| > M_1$, there exists i such that $|x_i^*| > 1/\epsilon$. Then $\mathbf{x}^i = (x_{1i}, ..., x_{ni}, x_i^*, ..., x_i^*) \in \mathcal{V}_{n,m}(\mathbf{x}^*)$ and since $|x_{ij}| \leq 1$, and $||\mathbf{x}^i|| \geq 1/\epsilon$ we obtain that $\mathbf{v} = \mathbf{x}^i/||\mathbf{x}^i|| \in \mathcal{V}_{n,m}(\mathbf{x}^*)$ and satisfies (32) and (33).

From , (32) and (33) we obtain that

$$1 = ||\mathbf{v}||^2 \leq n\epsilon^2 + mv^{*2} \tag{34}$$

and therefore using (31) we get

$$v^* \geq \left(\frac{1 - n\epsilon^2}{m}\right)^{1/2} > \frac{1}{2m^{1/2}} \geq \frac{1}{2n^{1/2}}. \tag{35}$$

Moreover using (24), (25), (28), (29), (30) and (33), if $m > 1$ there exists M_2 such that if $||\mathbf{x}^*|| > M_2$ and $||\hat{\boldsymbol{\beta}}|| > M_2$ then

$$\frac{r_j^2}{2} < f_j = \frac{u_j^2}{(1 - h_{ii}^2)||\hat{\boldsymbol{\beta}}||^2} = \frac{1}{(1 - h_{ii}^2)}\left(\frac{y_j}{||\hat{\boldsymbol{\beta}}||} - r_j\right)^2 < 2, \ 1 \leq j \leq n, \tag{36}$$

and

$$f^* = \frac{1}{(1 - h^{*2})m^2||\mathbf{x}^*||^2} < \epsilon. \tag{37}$$

Put $M = \max(M_1, M_2)$. In the rest of the proof we will assume that $||\hat{\boldsymbol{\beta}}|| > M$ and $||\mathbf{x}^*|| > M$. Since the eigenvalues of H are 0 or 1, then $||HF\mathbf{v}|| \leq ||F\mathbf{v}||$, and since by (33), (36) and (37) $||F\mathbf{v}|| < 3\sqrt{n}\epsilon$, we get

$$||HF\mathbf{v}|| < 3n^{1/2}\epsilon. \tag{38}$$

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

$$\mathbf{v} = \sum_{i=1}^{p} \theta_i \mathbf{v}_i, \tag{39}$$

and since by (32) $|\theta_i| \leq 1$, $1 \leq i \leq p$, using (35) we get that there is i_0 such that

$$|\theta_{i_0}| \geq \frac{v^*}{p} > \frac{1}{2pn^{1/2}}, \tag{40}$$

and

$$|v^*_{i_0}| \geq \frac{v^*}{p} > \frac{1}{2pn^{1/2}}. \tag{41}$$

Moreover applying HF in both sides of (39)

$$HF\mathbf{v} = \sum_{i=1}^{p} \theta_i \lambda_i \mathbf{v}_i, \tag{42}$$

and by (38) and (40) we get that $\lambda_{i_0} < 6pn\epsilon$. Using the fact that $\mathbf{v}_{i_0}$ is also an eigenvector of H corresponding to the eigenvalue 1, we get

$$|\mathbf{v}'_{i_0} F \mathbf{v}_{i_0}| = |\mathbf{v}'_{i_0} HF \mathbf{v}_{i_0}| = \lambda_{i_0} ||\mathbf{v}_{i_0}||^2 = \lambda_{i_0} < 6pn\epsilon. \tag{43}$$

Now, by (36) we get

$$\frac{|v_{i_0,j}|^2 r_j^2}{2} < 6pn\epsilon, \tag{44}$$

and by (27)

$$\# \left\{ j : 1 \leq j \leq n,\ |v_{i_0,j}|^2 < \frac{12pn\epsilon}{\gamma^2} \right\} \geq n - p + 1.$$

Therefore by (31) and (41) we get

$$\#\{j : 1 \leq j \leq n,\ |v_{i_0,j}| < |v^*_j|\} \geq n - p > \frac{n+m}{2},$$

and the Lemma is proved.

Acknowledgment This work has been supported by DGICYT, Spain with grants PB–93–0232 and PB94–0374, by the University of Buenos Aires with grant EX187 and by CONICET, Argentina, with grant PID–BID 0436.

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*, 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.

Hawkins, D.M., Bradu, D., Kass, G.V. (1984), " Location of several outliers in multiple regression data using elemental sets," *Technometrics*, **26**, 197–208.

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

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

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, B*, **57**, 145–156.

Rousseeuw, P.J. (1984), " Least median of squares regression," *Journal of the American Statistical Association*, **79**, 871–880.

Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier detection*, New York: John 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, 256–272.

Ruppert D.(1991), " Computing S–Estimates for regression and multivariate location/ dispersion," *J. Comp. Graph. Statist.*, **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" *Annals of Statistics*, **15**, 642–656.

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* A, **25**, 507–26

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

%outliers	Estimate	$x_0 = 1$				$x_0 = 5$				$x_0 = 10$			
		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	0.0	10.4	86.2	99.2	78.6	99.8	100.0	100.0	87.2	100.0	100.0	100.0
	PR2	0.0	18.0	87.8	99.2	73.6	99.6	100.0	100.0	77.2	99.4	100.0	100.0
	PR3	1.8	28.0	86.2	99.6	76.4	99.2	100.0	100.0	83.2	99.4	99.8	100.0
	M1	12.8	43.0	92.8	99.8	61.2	93.0	96.4	97.6	36.8	51.2	51.2	66.2
10%	PR1	0.2	9.4	81.4	99.0	63.4	98.2	100.0	100.0	67.4	98.2	100.0	100.0
	PR2	0.2	11.8	82.6	98.8	51.4	93.2	99.6	100.0	53.6	97.2	99.8	100.0
	PR3	0.2	19.2	80.4	98.8	45.2	92.4	99.0	100.0	52.4	97.0	99.6	100.0
	M1	6.4	39.4	90.2	99.2	39.4	63.0	63.0	70.8	26.6	44.6	49.4	61.2
15%	PR1	0.0	5.8	73.8	99.0	32.8	92.8	99.8	100.0	28.0	91.6	99.8	100.0
	PR2	0.0	7.0	68.6	96.6	23.0	83.4	98.8	99.8	16.8	81.6	96.6	99.6
	PR3	0.0	9.8	63.2	93.2	13.6	73.2	96.4	99.4	12.0	79.6	96.6	99.8
	M1	4.8	37.4	86.2	97.8	29.0	48.8	56.0	64.2	20.6	38.8	48.6	57.8
20%	PR1	0.0	3.6	60.8	97.4	10.6	71.0	97.2	99.6	7.8	68.4	96.4	100.0
	PR2	0.0	2.6	51.6	86.2	4.0	50.2	90.6	98.2	1.6	46.2	84.6	96.8
	PR3	0.0	3.6	31.8	72.0	1.4	33.4	80.8	95.2	1.2	36.6	81.2	95.0
	M1	5.0	31.8	84.8	93.2	21.6	41.0	52.2	56.6	10.0	29.2	34.4	42.4

Table 2. Average of false Outliers for p=3

%outliers	Estimate	$x_0 = 1$				$x_0 = 5$				$x_0 = 10$			
		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	1.18	1.07	1.10	1.01	0.64	0.57	0.47	0.54	0.64	0.54	0.47	0.53
	PR2	1.42	1.17	1.17	1.14	0.72	0.64	0.56	0.55	0.89	0.70	0.56	0.56
	PR3	2.67	1.93	1.88	1.79	1.43	1.20	1.13	0.98	1.41	1.09	1.04	1.04
	M1	5.43	4.28	4.64	4.44	0.59	0.86	0.50	0.53	0.82	0.62	0.53	0.65
10%	PR1	1.50	0.93	0.93	0.86	0.93	0.47	0.50	0.45	0.90	0.47	0.46	0.43
	PR2	1.86	1.24	1.10	1.06	1.34	0.76	0.58	0.54	1.24	0.57	0.53	0.50
	PR3	3.71	2.05	1.54	1.48	2.31	1.10	0.91	0.84	2.04	0.82	0.77	0.76
	M1	5.61	4.23	4.01	4.04	1.27	0.97	1.13	0.73	1.24	1.09	1.19	0.95
15%	PR1	2.20	1.41	0.87	0.70	1.68	0.65	0.31	0.39	2.01	0.74	0.41	0.30
	PR2	2.57	2.09	1.32	0.95	2.10	1.31	0.40	0.43	2.47	1.45	0.71	0.41
	PR3	5.37	3.67	1.75	1.10	4.07	2.23	0.68	0.57	4.47	1.73	0.79	0.43
	M1	8.15	4.77	3.91	3.50	2.73	1.65	1.65	1.42	2.80	1.97	1.90	1.36
20%	PR1	2.97	2.60	1.29	0.66	2.80	2.05	0.54	0.32	3.05	2.20	0.62	0.35
	PR2	3.58	4.18	2.80	1.74	3.41	3.54	1.12	0.54	3.75	3.63	1.67	0.71
	PR3	8.08	6.83	4.78	2.75	6.54	5.53	2.06	0.85	6.94	5.09	1.94	0.93
	M1	10.73	5.85	4.05	3.68	4.53	4.21	2.71	2.68	4.93	4.40	4.10	3.53

Table 3. Mean Squared Errors for p=3

%outliers	Estimate	$x_0 = 1$				$x_0 = 5$				$x_0 = 10$			
		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	0.18	0.21	0.18	0.16	0.24	0.15	0.14	0.13	0.25	0.13	0.14	0.13
	PR2	0.19	0.19	0.17	0.17	0.26	0.15	0.15	0.13	0.34	0.16	0.14	0.13
	PR3	0.24	0.23	0.20	0.19	0.29	0.18	0.16	0.15	0.31	0.17	0.17	0.15
	M1	0.31	0.28	0.26	0.28	0.31	0.29	0.31	0.36	0.65	1.73	3.70	4.56
	LSE	0.13	0.13	0.18	0.22	0.47	1.61	3.56	6.24	0.88	3.23	7.15	12.57
10%	PR1	0.21	0.27	0.23	0.16	0.43	0.21	0.14	0.14	0.47	0.22	0.15	0.14
	PR2	0.23	0.31	0.24	0.19	0.54	0.42	0.22	0.14	0.62	0.28	0.17	0.15
	PR3	0.32	0.37	0.26	0.19	0.66	0.45	0.25	0.15	0.68	0.29	0.20	0.15
	M1	0.38	0.39	0.31	0.36	0.57	1.18	2.45	3.34	0.86	2.35	4.65	6.31
	LSE	0.14	0.19	0.32	0.46	0.71	2.55	5.58	9.80	1.01	3.78	8.42	14.97
15%	PR1	0.30	0.47	0.36	0.19	0.78	0.43	0.16	0.16	0.97	0.53	0.16	0.14
	PR2	0.31	0.57	0.52	0.33	0.88	0.84	0.26	0.18	1.09	1.01	0.51	0.22
	PR3	0.51	0.82	0.65	0.40	1.15	1.33	0.52	0.26	1.29	1.13	0.51	0.18
	M1	0.57	0.65	0.57	0.43	0.81	1.96	3.72	5.19	1.05	2.78	5.23	7.49
	LSE	0.17	0.29	0.53	0.84	0.83	3.05	6.66	11.81	1.09	4.02	9.04	15.93
20%	PR1	0.40	0.89	0.77	0.31	1.12	1.59	0.47	0.25	1.27	1.81	0.54	0.18
	PR2	0.44	1.12	1.39	1.11	1.21	2.50	1.14	0.50	1.40	2.95	1.88	0.81
	PR3	0.78	1.67	2.30	2.07	1.53	3.56	2.30	1.07	1.63	3.65	2.42	1.41
	M1	0.72	1.02	0.79	0.94	1.10	2.84	4.76	7.58	1.32	3.68	7.28	11.22
	LSE	0.20	0.41	0.76	1.28	0.93	3.39	7.47	13.22	1.12	4.19	9.28	16.35

Table 4. Median Squared Errors for p=3

%outliers	Estimate	$x_0 = 1$				$x_0 = 5$				$x_0 = 10$			
		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	0.12	0.14	0.13	0.12	0.13	0.11	0.11	0.10	0.12	0.11	0.11	0.10
	PR2	0.13	0.15	0.13	0.12	0.13	0.11	0.11	0.10	0.15	0.11	0.11	0.10
	PR3	0.15	0.16	0.14	0.13	0.14	0.12	0.12	0.11	0.14	0.11	0.11	0.11
	M1	0.20	0.17	0.18	0.16	0.19	0.11	0.11	0.10	0.79	0.49	0.43	0.18
	LSE	0.10	0.10	0.15	0.19	0.46	1.56	3.56	6.14	0.85	3.16	7.06	12.51
10%	PR1	0.16	0.20	0.14	0.12	0.19	0.12	0.11	0.11	0.19	0.11	0.12	0.11
	PR2	0.17	0.22	0.15	0.13	0.42	0.13	0.11	0.12	0.35	0.12	0.12	0.11
	PR3	0.23	0.23	0.16	0.14	0.68	0.13	0.12	0.12	0.45	0.13	0.12	0.12
	M1	0.22	0.23	0.18	0.17	0.63	0.22	0.21	0.17	0.95	3.45	7.26	0.22
	LSE	0.11	0.17	0.27	0.41	0.68	2.51	5.48	9.66	0.96	3.70	8.25	14.68
15%	PR1	0.23	0.33	0.18	0.13	0.86	0.12	0.11	0.12	1.07	0.13	0.11	0.11
	PR2	0.25	0.39	0.20	0.13	0.90	0.14	0.11	0.12	1.10	0.16	0.11	0.12
	PR3	0.39	0.53	0.23	0.14	1.01	0.17	0.11	0.12	1.20	0.16	0.11	0.12
	M1	0.45	0.35	0.23	0.19	0.81	2.55	0.29	0.21	1.05	3.78	8.05	0.25
	LSE	0.15	0.27	0.47	0.77	0.80	2.99	6.55	11.65	1.04	3.90	8.80	15.61
20%	PR1	0.32	0.60	0.25	0.14	1.04	0.19	0.12	0.12	1.16	0.20	0.14	0.13
	PR2	0.35	0.89	0.43	0.16	1.10	0.76	0.14	0.13	1.21	3.99	0.16	0.13
	PR3	0.62	1.41	1.39	0.20	1.28	4.10	0.15	0.13	1.37	4.38	0.17	0.14
	M1	0.60	0.57	0.25	0.19	1.01	3.24	0.46	0.29	1.15	4.14	8.98	15.51
	LSE	0.17	0.36	0.72	1.25	0.89	3.33	7.30	13.04	1.07	4.07	9.07	16.05

Table 5. Null Behavior for p=3

Estimate	PR1	PR2	PR3	M1	LS
Average of false outliers	0.56	0.66	1.47	0.61	
Mean Squared error	0.13	0.13	0.16	0.13	0.12
Median Squared error	0.11	0.11	0.12	0.10	0.10

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

Estimate	%outliers=10				%outliers=15			
	m=1	m=1.5	m=2	m=3	m=1	m=1.5	m=2	m=3
PR1	1	59	99	100	5	60	100	100
PR2	0	5	72	100	0	7	61	100
PR3	0	0	6	48	0	1	10	49

Table 7. Average of False Outliers for p=30

Estimate	%outliers=10				%outliers=15			
	m=1	m=1.5	m=2	m=3	m=1	m=1.5	m=2	m=3
PR1	23.41	14.36	2.63	2.00	21.70	13.17	1.99	2.08
PR2	27.83	34.20	13.40	2.21	27.16	33.22	17.91	2.07
PR3	11.46	12.19	11.51	5.80	12.48	11.08	10.73	6.37

Table 8. Mean Squared Errors for p=30

Estimate	%outliers=10				%outliers=15			
	m=1	m=1.5	m=2	m=3	m=1	m=1.5	m=2	m=3
PR1	1.91	1.87	0.32	0.26	1.83	1.67	0.26	0.27
PR2	1.98	3.93	2.13	0.26	1.93	3.80	2.81	0.26
PR3	1.59	3.20	5.10	6.04	1.66	3.11	4.84	6.14
LSE	1.31	2.67	4.60	10.01	1.32	2.68	4.60	10.11

Table 9. Median Squared Errors for p=30

Estimate	%outliers=10				%outliers=15			
	m=1	m=1.5	m=2	m=3	m=1	m=1.5	m=2	m=3
PR1	1.84	0.31	0.25	0.24	1.84	0.32	0.24	0.25
PR2	1.91	3.90	0.27	0.24	1.86	3.90	0.33	0.25
PR3	1.53	3.12	5.13	9.76	1.58	3.09	5.07	9.50
LSE	1.31	2.67	4.56	9.98	1.30	2.67	4.60	9.96

Table 10. Null Behavior for p=30

	PR1	PR2	PR3	LS
Average of false outliers	3.61	4.74	1.47	
Mean Squared error	0.21	0.22	0.19	0.18
Median Squared error	0.20	0.21	0.19	0.18

Figure 1. NITC Data

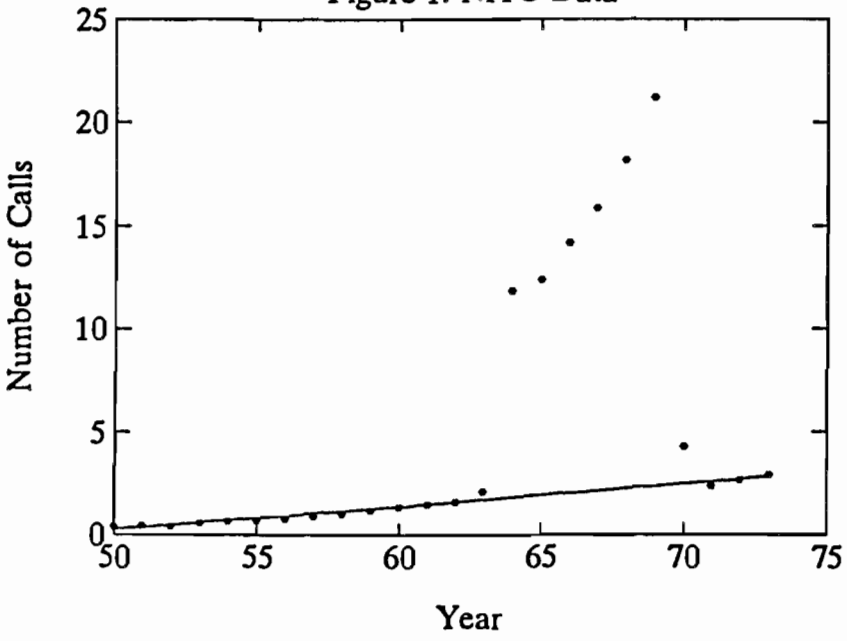


Figure 2. HRD Data

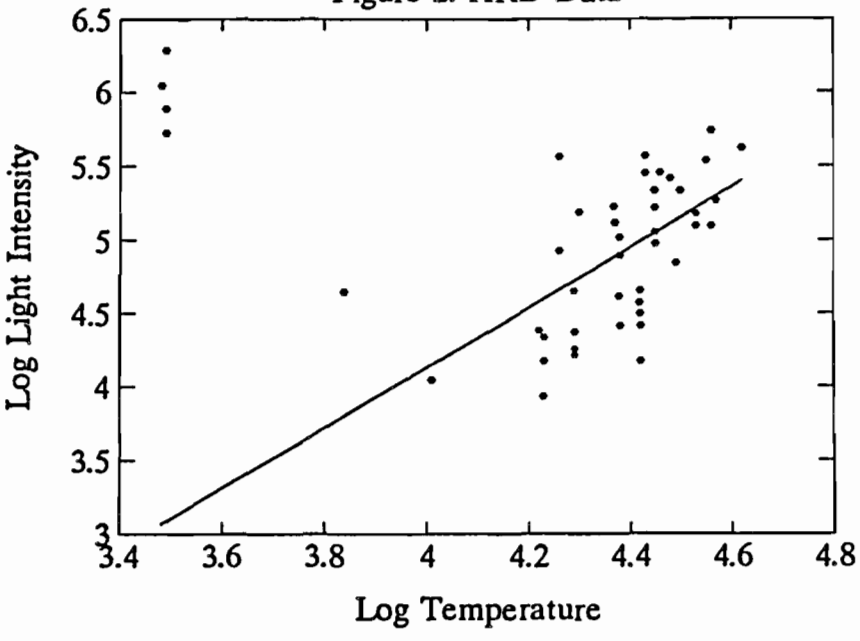


Figure 3. Rousseeuw Data

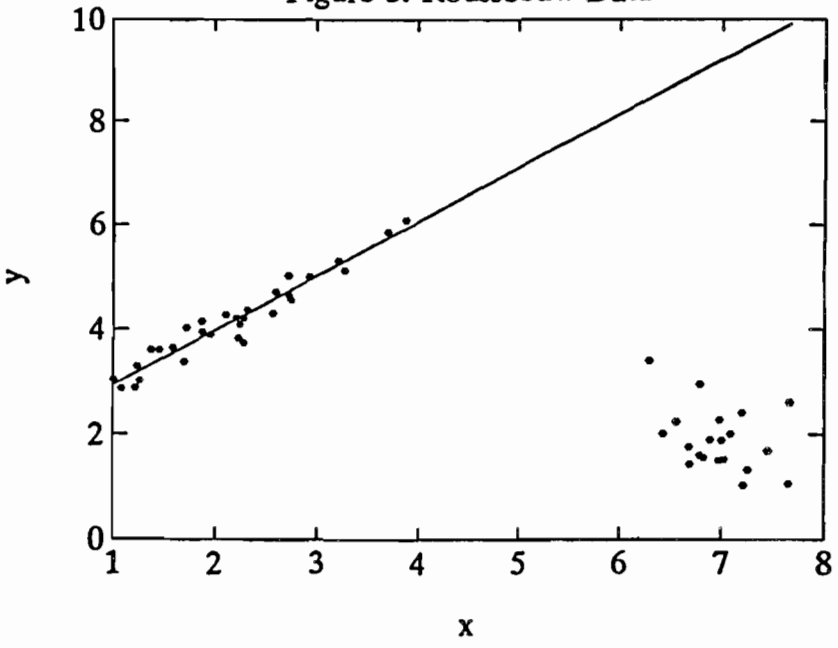


Figure 4. HDK Data