Applied Mathematical Sciences, Vol. 4, 2010, no. 28, 1367 - 1386

A Comparison of Various Influential Points Diagnostic Methods and Robust Regression Approaches: Reanalysis of Interstitial Lung Disease Data

Arezoo Bagheri [1,a], Habshah Midi [a], Mojtaba Ganjali [b] and Samaneh Eftekhari [b]

[a] Laboratory of Applied and Computational Statistics, Institute for Mathematical Research,University Putra Malaysia, 43400 Serdang, Selangor, Malaysia
abagheri_000@yahoo.com

[b] Department of Statistics, Faculty of Mathematical Sciences, Shaheed Beheshti University, Evin, Tehran, Iran

Abstract

In a linear regression model, the estimation of regression parameters by ordinary least squares method is affected by some anomalous points in the data set. Thus, detection of these abnormal points is one of the essential steps in regression analysis. There are many classical single deletion diagnostic measures which may fail to detect strange points due to masking effect. Local influence is an alternative method to evaluate the influence of local departures from assumptions in a proposed model. The main objective of this paper is to obtain the best resistant regression method which is robust to outliers in both the response and the explanatory variables. To achieve this objective, the weight vectors of the most commonly used robust regression techniques, such as the M- and the Generalized M-regressions are studied. A new measure based on the normal curvatures of the likelihood displacement is also proposed for comparing different robust regression methods. A medical data set is reanalyzed to underline that the use of only one alternative detection method or robust regression approach may not be sufficient to detect all influential points or to conclude the best robust method. A Monte Carlo simulation study is performed to confirm the results.

Mathematics Subject Classification: 62H12, 62J99

Keywords: robust regression methods; local influence; likelihood displacement

1 Introduction

Many quantitative models are utilized to diagnose and evaluate the response to a therapy in medical studies. Of them regression analysis is an important statistical tool that is routinely applied in most applied sciences. To ease the model formulation and computation, some desired assumptions such as normality of the response variable are made on the regression structure. Out of many possible regression techniques for fitting the model, the ordinary least squares (OLS) method has been traditionally adopted due to the ease of computation. However, there is presently a widespread awareness of the dangers posed by the occurrence of outliers in the OLS estimates (Rousseuw and Leroy, 2003). Outliers occur very frequently in real data, and they often go unnoticed because currently much data are processed by computers without careful monitoring. An OLS analysis can be totally spoiled not only by outliers in the response variable but also by outliers in the explanatory variables. Such influential points do not always show up in the usual OLS residual plots thus they remain hidden to the users.

Two useful procedures can be employed to protect against outliers in the X- or the Y-direction. These are the regression diagnostics (Cook , 1977; Cook , 1979 and Belsley et al., 1980) and the robust regression (Huber ,1973; Huber , 2003; Rousseeuw ,1984; Rousseeuw. and Yohai ,1984 and Andersen, 2008). Both of these have the same approach but, they proceed in the opposite direction. In regression diagnostics context, at first a regression model is fitted to the data set and then potential outliers are investigated. Subsequently, a model with the clean data (the data set without outliers) will be refitted. On the other hand, the procedures of robust regression follow two stages whereby in the first stage a model which is appropriate for the majority of the data is fitted, and then those observations with large robust residuals are detected as outliers.

The local influence can be considered as a practical procedure for assessing the validity of a fitted model. This is carried out by omitting or down weighing suspicious outliers through using influence graph to confirm whether they are outliers or influential points. Cook's local influence approach (see Cook, 1986) based on normal curvature is an important diagnostic tool for assessing local influence of minor perturbations to a statistical model. Assessing the influence of local perturbation of a statistical model has been an active area of statistical research in the past twenty years. The first measure which will be used in this paper is Cook's measure (Cook, 1977) which is based on a scaled distance of the Y and $\hat{Y}_{(i)}$, $n \times 1$ vectors of fitted values based on full data and the data without i-th case, respectively. The Cook's D_i Measure (hereafter, CDM) is used to detect cases that should be carefully investigated for gross error. The Normal Curvature Measure is defined by Cook (1986) (hereafter NCM) by extending the CDM measure. By this measure, instead of deleting the i-th individual, one can see the effect of that individual by giving it different weights. Another practical measure for detecting the influential measure has been defined by Hadi (1992). The Hadi's measure (hereafter HM) investigates the prediction of errors and presence of the outliers in X-direction. A similar in expression measure to HM, L-Measure, has been

introduced by Rancel and Sierra (2000) (hereafter, LM). Poon and Poon(2002) also define the standardized arc-length measure, P-Measure, for perturbing the weight of *i*-th case (hereafter PM). The last reviewed measure, S-Measure, is Pena's measure (2005) which determines the effect of deletion of each observation on the forecast (hereafter SM). We are going to compare the performance of these measures on detecting the influential points of a real data set.

Robust methods have been defined to deal with the influential points in regression analysis. One of the oldest robust approaches which predates OLS by 50 years is L_1-norm (Armstrong and Kung, 1978). Least median of squares (LMS) and Least Trimmed Squares (LTS) are introduced by Rousseeuw (1983 and 1984). M-estimators are suggested by Huber (1973). Generalized M-estimators are introduced by Schweppe (which is given in Hill, 1977) and Coakley and Hettmansperger (1993) (their algorithms are available in Wilcox, 2005) and MM-estimators are presented by Yohai (1987). In this paper, we will propose a new measure, for assessing different robust regressions methods. This measure is based on normal curvature of the likelihood displacement.

The performance of our proposed measure and all robust regression methods will be examined on a simulation study and will be applied to a real data set. A modified method of Coakley and Hettmansperger 's (GM-estimators) will be also presented and applied.

This paper is organized as follows. In Section 2 we will discuss the local influence approach of Cook (1986) and a general overview of the above mentioned diagnostic methods of influential points will be presented. In Section 3 we will review the different types of robust regression methods and their properties such as breakdown point, efficiency and bounded influence (the comprehensive definition of these properties can be found in Andersen, 2008). Interstitial lung disease (ILD) data set, taken from Narula et al.(1999) (who only used a $\boldsymbol{L_1}$-norm regression to analyze the data and they did not use any detection method to find the influential points) will be reexamined more carefully for detecting influential points and will be reanalyzed in Section 4. A Monte Carlo simulation study will be employed in Section 5 to confirm the results of numerical example, and to evaluate the performance of our introduced measure and our modified GM-regression method. The conclusion of the study will be given in the last Section.

2 Local Influence and Diagnostic Methods

Influence emerges from the interaction between the model and the bad elements of the data for which valid conclusion from a fitted model cannot be drawn. By definition, local influence is the minor perturbations of a model. Hence, assessment of local influence is necessary for the best fit of a model. Several diagnostics have been developed for assessing the local influence for the perturbations of case-weights, explanatory variables and for assessing effect of specific perturbations on the parameter estimates. In this section at first we review the local influence and then reexamine some of the detection methods.

2.1 Local Influence

The method of local influence was introduced by Cook (1986) and modified by Billor and Loynes (1993) as a general tool for assessing the influence of local departures from the assumptions underlying the statistical models. Consider the following standard linear regression model:

$$y = X\beta + \epsilon \tag{1}$$

where y is a $n \times 1$ vector of dependent or response variable, X is a $n \times p$ $(p = k + 1)$ design matrix (the number of independent variables predicting y is equal to k), β is a $p \times 1$ vector of unknown parameters and ϵ is the $n \times 1$ vector of error with distribution$N[0, \sigma^2 I_n]$. Chatterjee and Hadi (1986) gave an excellent review of several measures assessing influence of observations in regression modeling. Cook (1986) considered a generalized version of Cook's Distance Measure (CDM):

$$D_i = \frac{\|\hat{Y} - \hat{Y}_{(i)}\|^2}{p\sigma^2} = \frac{h_{ii}}{(1-h_{ii})} \frac{{r_i}^2}{p} \tag{2}$$

where $\hat{Y}$, $\hat{Y}_{(i)}$ are the $n \times 1$ vectors of fitted values based on the full data and the data without the i-th case, respectively, and p is the dimension of β. He introduced the use of:

$$D_i(w) = \frac{\|\hat{Y} - \hat{Y}_{(w)}\|^2}{p\sigma^2}$$

where, $\hat{Y}_{(w)}$ is the vector of fitted values obtained when the i-th case has weight w and the remaining cases have weight 1. This idea has been extended to general models. This extension is partially motivated by the following relationship between $D_i(w)$ and the log-likelihood $L(\beta)$ for model (1),

$$pD_i(w) = \frac{\|Y - \hat{Y}_{(w)}\|^2 - \|Y - \hat{Y}\|^2}{\sigma^2} = 2[L(\hat{\beta}) - L(\hat{\beta}_W)] \tag{3}$$

where $\hat{\beta}$ $=\hat{\beta}_W$ when $w = 1$ and $\hat{\beta}_W$ is the maximum likelihood estimator of β when the i-th case has weight w. The form of this relationship is a consequence of the statistical structure assumed for the errors in model (1).

In general, consider $L(\theta)$ as log likelihood corresponding to the postulated model where θ is a $p \times 1$ vector of unknown parameters. Perturbations into the model may be defined through the $q \times 1$ vector w restricting to some open subset Ω of R^q. The log-likelihood for the unperturbed and perturbed models are denoted by $L(\theta)$ and $L(\theta|w)$, respectively. Then the likelihood displacement $LD(w)$ is defined by:

$$LD(w) = 2[L(\hat{\theta}) - L(\hat{\theta}_W)] \tag{4}$$

where $\hat{\theta}$ and $\hat{\theta}_w$ are the maximum likelihood estimators of θ under the unperturbed and perturbed models, respectively. The vector of the values w and $LD(w)$ forms the surface of interest as w varies over certain space. The direction of maximum curvature of the likelihood displacement surface in the postulated model (where $w = w_0$) indicates the greatest local sensitivity against perturbations. The direction of maximum curvature is used as the main diagnostic tool in the local influence method.

An obvious way to see if perturbations of the model influence key results of the analysis is to compare the results derived from the original and perturbed models using an influence graph which is a geometric surface formed by the values of the $(q + 1) \times 1$ vector;

$\alpha(w) = (w, LD(w))'$ where w varies through Ω which is some open subset of R^q. To characterize the behavior of an influence graph around w_0 in Ω, geometric normal curvature is used. Some direction $l_{q\times 1} \in R^q$ ($\|l\| = 1$) is chosen to see the normal curvature C_l at w_0 in the direction of l. The expression for C_l will reduce to:

$$C_l = 2\left|l^T \Delta^T (L^{"})^{-1} \Delta\, l\right| \tag{5}$$

where Δ is a $p \times q$ matrix with elements:

$$\Delta_{ij} = \frac{\partial^2 \mathrm{L}(\theta|\mathrm{w})}{\partial \theta_i\, \partial \mathrm{w_j}} \Big|_{\{\theta=\widehat{\theta}, \mathrm{w}=\mathrm{w_0}\}}$$

and - L" is the observed information matrix for the postulated model ($\mathrm{w} = \mathrm{w_0}$).

Let make (5) desirable for our purpose. Consider *w* denote the $n \times 1$ vector of case-weights for the regression model (1) and σ^2 is known. The relevant part of the log-likelihood for the perturbed model is:

$$L(\beta|w) = -\frac{1}{2\sigma^2}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{w_i}(\mathrm{y_i} - \mathrm{x_i^T}\beta)^2 \tag{6}$$

where w_i and y_i are the *i*-th components of w and Y, respectively, and $\mathrm{x_i^T}$ is the *i*-th row of *X*. Differentiating (6) with respect to β and *w*, and evaluating at $\hat{\beta}$ and $w_0 = 1$, we find;

$$\Delta = X^T D(r)/\,\sigma^2$$

where $r = (r_i)$ is the n × 1 vector of ordinary residuals when $w = 1$ and D(r) = diag(r_1,. . ., r_n). Since $l\,"(\widehat{\beta}) = -\mathrm{X^T X}/\,\sigma^2$, normal curvature can be defined as:

$$C_l = 2l^T D(r) H D(r) l/\sigma^2 \tag{7}$$

where *H* is the hat matrix and $\|l\| = 1$. Moreover, $C_{max} = max_l C_l$ in the direction of $l_{max}$where l_{max} is the eigenvector corresponding to maximum eigenvalue of $D(r)HD(r)$ matrix.

2.2 Diagnostics Measures

It is important to point out that, in real situation, many data sets for which one often makes normal assumption, has heavy-tailed distribution which may arise as a result of outliers. Consequently, the outliers will have an unduly effect on the regression results (see Pearson, 1931 and Box, 1953). There are three different groups of strange points that may occur and need further attentions because their presence will have a great influence on the OLS estimates. The first group is regression outliers which sometimes are called the group of vertical outliers (Rousseeuw and van Zomeren, 1990). These outliers stand apart from the general pattern for the bulk of the data. Specifically, they are observations which are discrepant in terms of their y values. Regression outliers are characterized by relatively large residuals. It is essential to explain the leverage concept here. The farther the observation is from the mean of X (either in a positive or negative direction), the greater is its leverage. There are two types of leverage points which play different roles in regression, good leverage points and bad leverage points. Leverage points are referred as good or bad depending on whether they are reasonably consistent with the true regression line. If they are consistent with the true regression line then they are referred as good leverage points, otherwise they are bad leverage points. These leverage points are also regression outliers and fall on the second

group of strange points. Finally, the last group of strange points are the good leverage points which are not regression outliers. Points in this group can reduce the standard errors of the OLS estimates. On the other hand, a bad leverage point can result in a poor fit to the bulk of the data. Additionally, observations which have unduly influence on the regression results are identified as influential observations. Thus, vertical outliers and bad leverage points are influential observations which should be carefully investigated in regression analysis.

Diagnostic measures are certain quantities computed for the purpose of revealing influential observations. Much work has been accomplished on different methods to detect these unusual points beginning with the key methods of Cook and Weisberg (1982).

In view of the fact that the structure of the local influence concept is quite useful for identifying influential subset, and providing a further justification for local influence analysis, some of the established diagnostics methods based on the local influence will be reviewed.

A practical measure, NCM, is the curvature for the influence graph obtained by modifying the weight attached to a single case, suppose the *i*-th, this is:

$$C_i = 2r_i{}^2 h_{ii}/\sigma^2 = 2P(1-h_{ii})^2 D_i$$

where D_i is defined in (2).

Another method is a new measure which shows that the local influence analysis of perturbations of the variance is similar to the usual regression diagnostic based on the CDM for detecting influential subset. Hadi (1992) proposed HM for detecting influential subset of observations which is resistant to masking and swamping effects. HM is based on the simple fact that potentially influential observations are outliers in the X-space, the Y-space, or both, which yields HM to be:

$$HM_i{}^2 = \frac{p}{(1-h_{ii})}\frac{d_i{}^2}{(1-d_i{}^2)} + \frac{h_{ii}}{(1-h_{ii})} \qquad i = 1, \dots, n \tag{8}$$

where $d_i{}^2 = r_i{}^2/r'r$, is the square of the *i*-th normalized residual and h_{ii} is the *i*-th diagonal element of the hat matrix. HM is the sum of two components with different interpretation. A large value of the first term on the right hand side of (8) can be resulted in a poor fit or large prediction error while a large value of the second term indicates the presence of high leverage point. Hadi (1992) introduced a robust cutoff point for HM as $median(HM_i{}^2) + \mathrm{C} \times \mathrm{mad}(HM_i{}^2)$ where:

$$Mad(HM) = \frac{median|HM_i{}^2 - median(HM_i{}^2)|}{0.6745} \qquad for\ i = 1, \dots, n$$

where C can be taken as constant values of 2 or 3.

Rancel and Sierra (2000) introduced a quasi likelihood displacement to consider the influence of the high-leverage observations in likelihood displacement. The quasi likelihood displacement defined as:

$$LD_{(i)}(w_i) = 2[L(\hat{\theta}) - L_{(i)}(\hat{\theta}_{W_i}|w_i) + [var(\hat{Y}_i) - var(\hat{Y}_{w_i})].$$

where $L_{(i)}(\hat{\theta}_{W_i})$ is the log-likelihood displacement under the perturbed model when the *i*-th observation is deleted. So that, the slope of the maximum increment direction of $LD_{(i)}(w_i)$ which has an expression similar to HM indicates an existing relation between local and deletion diagnostic. This can be defined as LM which is:

$$l_{i(i)} = 1 - r'r\left[\frac{p}{(1-h_{ii})}\frac{{d_i}^2}{(1-{d_i}^2)} + \frac{h_{ii}}{(1-h_{ii})^2}\right]. \quad (9)$$

Another measure to characterize the behavior of the influence graph of the likelihood displacement over the entire perturbation range is PM which was developed by Poon and Poon (2002). They developed a relation between the normal curvature of basis r_i and all influential eigenvectors at w_0 under some assumptions, and concluded that this measure which is called the standardized arc-length is an effective measure for assessing local influence. This measure can be defined for perturbing the weight of the *i*-th case as:

$$P_i = \int_0^1 \sqrt{1 + \frac{4}{\sigma^4}\frac{t^2 {r_i}^4 {h_{ii}}^2}{(1-th_{ii})^6}}\ dt. \quad (10)$$

Formula (10) implies that $P_i > 1$; and $P_i = 1$ if and only if the leverage h_{ii} or the residual r_i is equal to zero.

Pena (2005) has introduced a new statistics which present how sensitive the forecast of the *i*-th observation is to the deletion of each observation in the sample. In contrast with CDM which measures the influence of an observation through deleting the observation from the sample and computes the changes in the vector of forecast, SM measures the changes of the forecast of one observation after deleting each of the sample points one by one. So, SM is defined for the *i*-th observation as:

$$S_i = \frac{1}{ps^2 h_{ii}}\sum_{j=1}^{n}\frac{{h_{ji}}^2 {r_j}^2}{(1-h_{jj})^2} \quad (11)$$

where $s^2 = r'r/(n-p)$ and h_{ji} 's are the off diagonal elements of hat matrix. The most important difference between this measure and CDM is that the distribution of SM for large sample sizes with many explanatory variables will be approximately normal. However, this measure will not be useful in situations that the outliers in Y-direction has low leverages. Hence, it will be very efficient when the data set contains high leverage outliers or bad leverage points.

3 Robust Regression Methods

Utilizing the Ordinary Least Squares (OLS) method, the estimator of β are found by minimizing the sum of squared residuals, $min_{\hat{\beta}} \sum_{i=1}^{n} {r_i}^2$ where $r_i = y_i - \hat{y}_i$. This gives the OLS estimator for β as:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

The OLS estimate is optimal when the error distribution is assumed to be normal (Hampel, 1974 and Mosteller and Tukey, 1977). In the presence of influential observations, robust regression is a suitable alternative to the OLS. Robust procedures have been the focus of many studies recently, all of which triggered by the ideas of Hampel (1974). These methods are mainly aimed to provide stable results in the presence of outliers. One of the first robust methods is called L_1-norm, least absolute values (LAV) or minimum sum of absolute errors (MSAE) regression estimator. This estimator obtains a higher efficiency than OLS through minimizing the sum of the absolute errors ($\min\sum_{i=1}^{n}|r_i|$) introduced by Armstrong and Kung

(1978). The combination of the low breakdown point (1/n) and non-sensitivity to outliers in X-direction (see Mosteller and Tukey, 1977) makes LAV less attractive than most of the other existing robust regression methods. Rousseeuw (1983 and 1984) introduced two other robust methods: least median of squares (LMS) and Least Trimmed Squares (LTS). LMS attempts to minimize The median of $r_i{}^2$. LMS estimator has deficiencies, which limit its use, such as a relative efficiency of 37% (see Rousseeuw and Croux, 1993) and low convergence rate for its influence function (Rousseeuw, 1983). Despite these limitations, LMS estimators can highly influence the calculation of the much more efficient MM-estimators by providing initial estimates of the residuals which will be explained later. The other method, LTS minimizes the sum of the trimmed squared residuals, $min\sum_{i=1}^{h}(r^2{}_{(i)})$ where $h=[n(1-\alpha)+1]$ is the number of the observations including in the calculation of the estimator, and α is the proportion of trimming. Using $h=\left({}^{n}/_{2}\right)+1$ ensures a high breakdown point for the estimator. Both the LMS and the LTS methods have high breakdown as 50%. However, they produce unbounded influence estimators (Rousseeuw, 1983 and 1984). Moreover, the highly resistant LTS estimator suffers badly in terms of relative efficiency at about 8% (see Stromberg et.al., 2000). Although, the LTS has low efficiency, it has an important role in the calculation of some other robust estimators such as the GM-estimators.

Huber (1973) suggested M-estimators of β as another robust estimator. These obtained by solving:

$$\sum_{i=1}^{n}\psi\left(\frac{y_i-x_i'\widehat{\beta}}{s}\right)x_i=0 \qquad (12)$$

where ψ-function may be a monotonic ψ-function such as Huber's ψ-function which is defined as:

$$\psi(t)=\begin{Bmatrix} t & if\ |t|<b \\ bsgn(t) & if\ |t|\geq b \end{Bmatrix}.$$

The M-estimators are the simplest high efficient robust estimators having both computationally and theoretically desirable asymptotic properties. It is worth to mention that the M-estimators are not robust in the X-direction and have low break down point (1/n) (Simpson, 1995). Schweppe introduced a class of robust methods which is called the Generalized M-estimators (GM-estimators) (see Hill, 1977). Simpson (1995) made an extensive comparison between different types of GM-estimators. The major aim of these methods is to down weight those high leverage points which have large residuals or bad leverage points. Simpson (1995) has reported that these estimators have high efficiency and bounded influence properties which achieved a moderate break down point equal to $1/p$. The GM-estimator is the solutions of the normal equations:

$$\sum_{i=1}^{n}\pi_i\psi\left(\frac{y_i-x_i'\widehat{\beta}}{s\pi_i}\right)x_i=0 \qquad (13)$$

where, π_i's are defined to down weight high leverage points with high residuals and S is a robust scale estimate. Different methods may be used to solve (13). The most common method is the one-step or fully Iteratively Reweighted Least Squares (IRLS). At convergence, the GM-estimator may be written as:

$$\widehat{\beta}_{GM} = (X'WX)^{-1}\ X'Wy \tag{14}$$

where in this case the diagonal elements of W are the weights w_i defined as:

$$w_i = \frac{\psi[(y_i - x_i'\widehat{\beta}_{GM})/\pi_i s]}{(y_i - x_i'\widehat{\beta}_{GM})/\pi_i s} \tag{15}$$

GM-estimators may have high breakdown point if we obtain appropriate initial estimators. Birch (1980) pointed out that a good starting value is always important in an iterative scheme. Two of the main existing GM-estimators which are called GM1 and GM6 according to Wilcox (2005) will be discussed in this paper. The first one, GM1, is due to Schweppe introduced in Handshin et al. (1975). This estimator has defined the π-weight function which contains a square root function of the diagonal elements of the hat matrix; $H = X(X^TX)^{-1}X^T$. The initial estimator was obtained from the least squares method and $\hat{\tau} = 1.48\,[median\ of\ the\ largest(n-p) of\ the\ |r_i|]$ is recommended as the scale estimator. The final estimate is obtained using fully iterated reweighted least squares where the weight function comes from the following equation:

$$w_i = \frac{\sqrt{1-h_{ii}}}{e_i}\psi\left(\frac{e_i}{\sqrt{1-h_{ii}}}\right) \tag{16}$$

where $e_i = r_i/\hat{\tau}$ and ψ is the Huber function with tuning constant equal to $2\sqrt{(p+1)/n}$. In fact, this GM-estimator has bounded influence while its finite-sample break down point is only *2/n*. Hence, it can handle one outlier, but two outliers might destroy it. The first GM-estimator (GM6) with high asymptotic efficiency for the normal model, high breakdown nearly equal to 50% and bounded influence was proposed by Coakley and Hettmansperger (1993). To overcome the limitation of Schweppe's method in using OLS as initial estimator, Coakley and Hettmansperger (1993) postulated another method, GM6. The method consists of the high breakdown LTS estimator as initial estimator and LMS scale estimate as $\hat{\tau} = 1.4826(1 + 5/(n-p))Median|r_i|$ where r_i is residual of LTS method. A one step Newton Raphson has been used as convergence approach. More formally, letting $\hat{\beta}_0$ be the LTS estimator, this estimator can be derived from:

$$\hat{\beta}_{ch} = \hat{\beta}_0 + (x'Bx)^{-1}x'W\psi(r_i/(w_i\hat{\tau})\,\hat{\tau}) \tag{17}$$

where $w = diag(w_i)$ and $w_i = \min\{1, [x^2_{0.95,p}/{RD_i}^2]\}$ where

$$RD^2 = (x - m_x\acute{)}C^{-1}(x - m_x) \tag{18}$$

and the quantities m_x and C are the minimum-volume ellipsoid (MVE) estimators of location and scale (Rousseeuw and van Zomeren ,1990). Moreover, $\psi'(x)$ is the derivative of Huber's ψ and $B = diag\ \psi'(r_i/\hat{\tau}\,w_i))$. They suggested using the tuning constant equal to 1.345 in Huber's ψ- function.

We made a slight modification on GM6, by utilizing the S-estimator instead of the LTS estimator which has more asymptotic efficiency (Andersen, 2008 and Campbell et al., 1998) as initial estimator. Thus the $\hat{\boldsymbol{\tau}}$ for S-estimator residuals will be defined and the same convergence method of one step Newton Raphson will be employed with the same weight function as what is used in (18). We will present the merit of this new modified robust method by real data set and simulation study.

One of the most practical robust methods is the MM-estimators which was first proposed by Yohai (1987). These estimators combine high breakdown value estimators (50%) and M–

estimators which have high efficiency (approximately 95% relative to OLS under the Gauss-Markov assumptions). The MM-estimators in the name refers to the fact that more than one M-estimation procedure is used to calculate the final estimates. For more information about the procedure which is employed in this estimator, one can refer to Yohai (1987) and Andersen (2008).

To compare the mentioned robust regression methods, a new indicator based on the distance of robust weight vectors to normal curvature of the likelihood displacement direction can be defined. Two simple distance measures that can be used in this situation are the angle and chord distance. The chord distance is defined as the length of the chord between two vectors of unit length which is the same for normalized and original vectors. Moreover, for normalized vectors the chord and Euclidian distances are the same (Causton, 2003 and Timm, 2002). Eigenvector l_{max} associated with C_{max} indicates how to perturb the postulated model to obtain the greatest change in likelihood displacement. This is the most important diagnostics which comes from this approach (Cook, 1986). The larger the absolute value of the i-th element of l_{max}, the more influential the i-th observation. The most important aim of any robust method is to down weight the influential points. Hence, the vector $u = u_i$ with components $u_i = w_i/\|w\|$ of any robust method can be used to determine the influential points according to that specific robust method. Any method which has given less weight to the i-th influential point will have smaller value of corresponding u_i. Hence, distance of two normalized vectors of l_{max} and u_i can be computed by the Euclidean distance between these two normalized vectors (Timm, 2002). We define Distance to Maximum Normal Curvature Direction (DMNCD) as an Euclidean distance:

$$DMNCD = \sqrt{\sum_i^n (u_i - l_{\max i})^2} \tag{19}$$

where u_i is the i-th component of u. Consequently, any of the robust methods with larger DMNCD will make more perturbation in parameter estimation. Larger DMNCD value suggests that the robust estimates are further from the OLS estimates and this may not be desirable. Hence, DMNCD will let researcher know how far one is from OLS method.

4 Application to a Published Clinical Trial

Narula et al. (1999) used ILD data set which designed to verify the association between objective indicators of lung damage and severity of functional impairment in Interstitial Lung Disease (ILD) patients where ILD refers to a diffuse inflammatory process that occurs predominantly within the interstitial spaces and supporting structures of a lung. In these data, specimens were obtained by retrospective review of the medical and pathological records and biopsies of 24 patients with ILD whom were selected from the file cases of open chest lung biopsies of Surgical Pathology Service of the teaching hospital of Faculdade de Medicina da Universidade de Sao Paulo. For this set of patients, the pulmonary function measurements were gathered within 30 days before the biopsy. This data set consists 14 explanatory variables (see appendix of Narula et al. (1999) for the data set and the description of all explanatory variables) while response variable is FVC (forced vital

capacity). The model and all the regression coefficients with all explanatory variables were not significant at the 5% level which may have been the result of multicollinearity. To select a significant model, a stepwise least squares regression (see Draper and Smith, 1998 and Montgomery and Peck, 1982) has been employed. The resulting significant model contains AGE (in years), EPIT (epithelial cells), that is area fraction of epithelial cells/10000 μm2 of alveolar tissue), CELL (cellular infiltration, which is total cellularity/10000 μm2 of alveolar tissue) and HONEY (honeycombing, which is a score of zero to four honeycombing). This data set has two outliers (observations 11 and 15) and one high leverage point according to Narula et al. (1999). It is noticeable that Narula et al. (1999) didn't mention which case is the leverage point.

Statistical Package S-PLUS (Version 8) is used in all different stages of analyzing this data set. Robust estimators can be obtained from Robust library of this package while, GM-estimators are not yet available However, easy-to-use S-PLUS functions are supplied by Wilcox (2005); as an example bmreg (for Schewppe's GM-estimator) and chreg functions (for Coakley and Hettmansperger, 1993).

Index plot of OLS standardized residuals and hat matrix are presented in Figure 1. The OLS standardized residuals diagnoses two outlier cases 11 and 15. Moreover, hat matrix can detect cases 3 and 23 as leverage points. While these two high leverage points doesn't have large residuals and it seems that they are good leverage points.

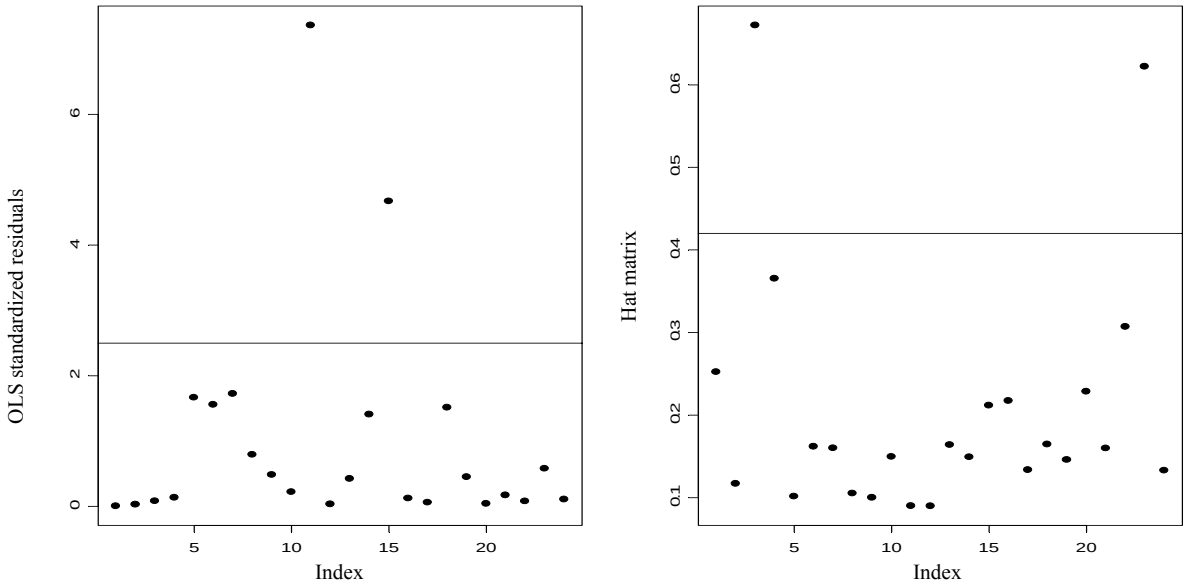

Figure 1. Index plot of OLS standardized residuals and hat matrix

Index plot of local influence measures; CDM, HM, LM, PM, NCM and SM are illustrated in figure 2. It is worth to mention that the cutoff point of CDM is set to be $4/(n-p-1)$ (Anderson, 2008) while a robust non-parametric cutoff point of $Median\ (\theta_i) + 3Mad(\theta_i)$ can be defined for all the other measures in order to make these measures more comparable.

According to the Figure 2, the CDM can recognize the case 15 as influential point while two other influential points of 11 and 23 are somehow far away from the other observations in the data set whilst the NCM could identify only outlier cases of 11 and 15. Both HM and LM can identify 4 influential cases 3, 11, 15and 23 thus these points needs more consideration. Moreover PM is able to diagnose three cases 11,
15 and 23 similar to the claim of Narula et. al. (1999). More importantly, SM couldn't find any influential points due to its sensitivity to large high leverage with large residuals (Pena, 2005). Consequently, according to these classical influential diagnostics measures, this data set contains four influential observations. Since all these diagnostics are affected by outliers, thus, robust diagnostics should be applied in this data set.

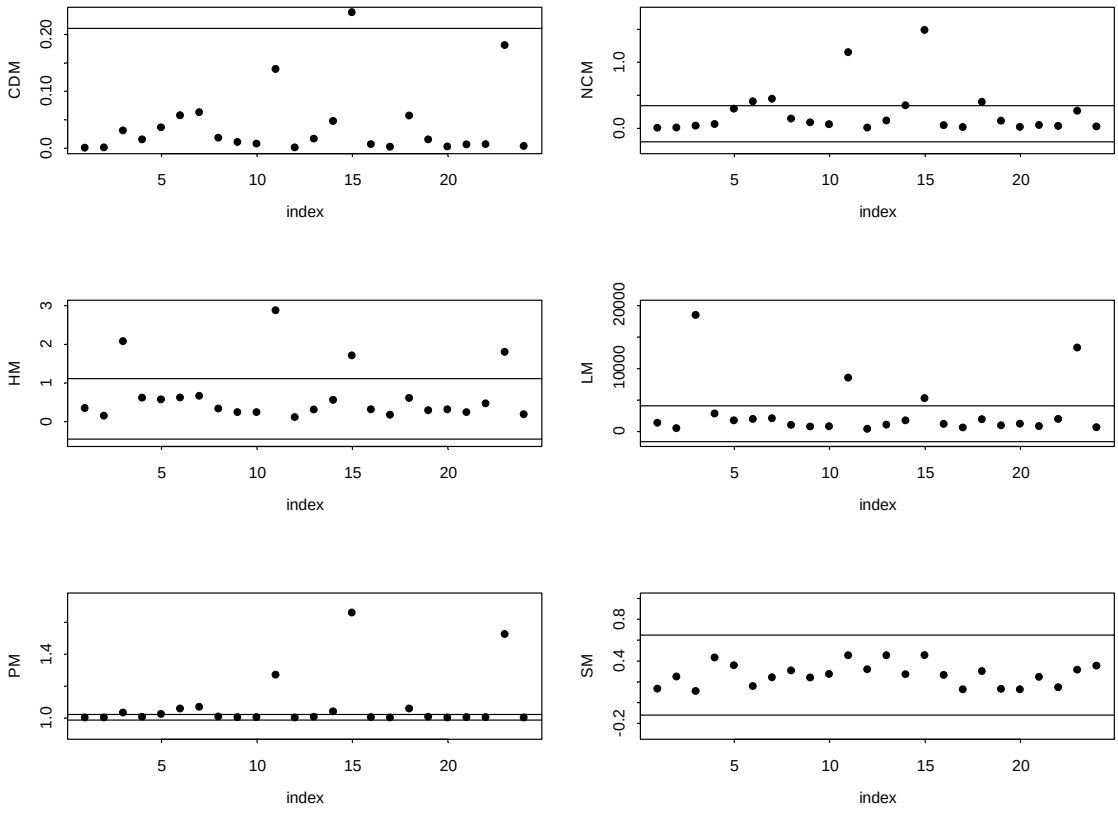

Figure 2. Index plot of Cook Distance (CDM), Hadi's diagnostics Measure (HM) and four diagnostics measures of $l_{i(i)}$ (quasi likelihood displacement measure, LM), p_i (standardized arc-length measure, PM), C_i (normal curvature measure, NCM) and S_i measure (SM)

To explore more about the outliers in this data set, the robust diagnostic methods should be considered. For example, Rousseeuw and van Zomeren (1990) proposed an influence plot of the robust residuals against robust distances which detects multiple outliers more accurately than the traditional methods (see also Cook and Hawkins, 1990, Ruppert and Simpson,1990 and Kempthorne and Mendel,1990 for debate about this topic). Robust residuals which come from the highly resistant LMS or LTS regressions are often employed. The absolute robust residuals which are more than 2.5 will be considered as vertical outliers. One of the vital diagnostics methods of high leverage points is robust Mahalanobis distance which is introduced in equation (18). The points with RD^2 more than $x^2_{(0.95,p)}$ may consider as high leverage points ($x^2_{0.95,4} = 9.49$ for our application). Rousseeuw and van Zomeren's regression diagnostic plot for ILD data is shown in Figure 3. Plotted against the square root of robust Mahalanobis distance based on MVE is the standardized residuals from a LTS regression (Willcox, 2005). Robust distance indicates that cases 3, 4, 22 and 23 are good high leverage points. Moreover, the robust residuals suggest that two cases (11 and 15) are outliers.

To investigate the effect of these two outliers in the parameter estimates of the model, these estimates for the whole data set and without outliers 11 and 15 are presented in Table I. For data without outliers, the AGE effect decreases after deleting these two outliers from the data set (p-value increases from 0.018 to 0.028).

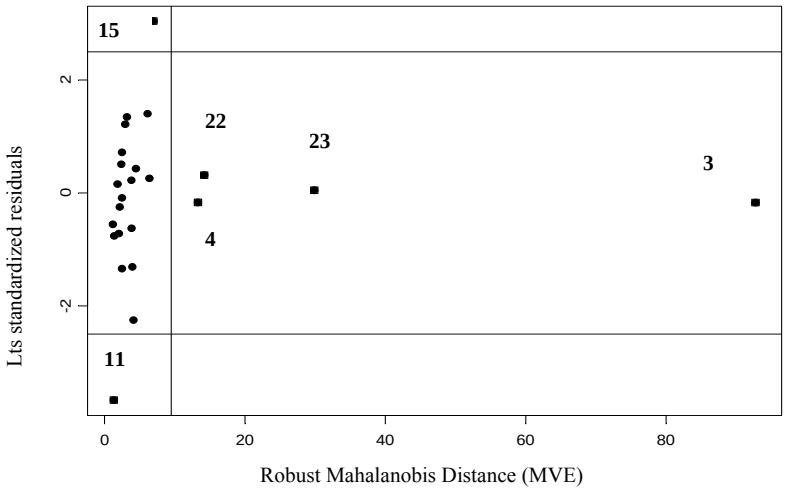

Figure 3. Plot of standardized Robust Residuals (from LTS fit) against Robust Mahalanobis Distance based on MVE for ILD data set

In view of the fact that this data set has two outliers, utilizing robust methods for estimating the parameter of the fitted model is necessary. Thus, different robust methods such as L_1-norm, LTS, LMS, MM, GM6 and modified GM6 have been fitted to this data set. Since in this data set, the sample size, *n*, is small, the Asymptotic Standard Errors (ASEs) which has been defined by Draper and Smith, 1998 are not reliable (see Huber, 2004). Thus bootstrapping is an alternative way to obtain the standard errors for robust regression estimators. There are two different bootstrapping ways which are random-X and fixed-X bootstrapping (for more details, one can refer to Anderson, 2008). Since the explanatory variables are assumed to be fixed in this data set, fixed-X bootstrapping has been used and the results are presented in table II. The coefficient estimation and fixed-X bootstrapping standard error of the residuals for all of the selected robust regression techniques are listed in Table II. Each of these methods has made some changes in the coefficients through down weighting some of the influential points. It is interesting to note AGE and EPIT are insignificant according to LMS parameter estimates. Furthermore, AGE is not significant by L_1-norm, LMS and LTS. Thus these three robust methods are going so far from the results of OLS estimates and also OLS without outliers which is not desirable. However, AGE is significant (on 5% level) for the other robust methods. Hence, L_1- norm, LMS and LTS estimators will be omitted in our further analysis.

Table I. Parameter estimates (standard errors) of resulting significant model for ILD data set with and without outliers

Method	INTERCEPT	AGE	EPIT	CELL	HONEY
OLS with all observations	46.654 (11.280)	0.614 (0.236)	-0.061 (0.017)	107.733 (37.919)	-10.639 (1.936)
OLS without outliers 11 and 15	54.841 (8.196)	0.438 (0.182)	-0.064 (0.012)	112.576 (27.294)	-10.485 (1.459)

Table II. Parameter estimates and bootstrap standard deviations of different robust methods for ILD data set

Parameter	Robust Methods							
	L_1-norm		LTS		LMS		M	
	Par. Est.	Boot.SE	Par. Est.	Boot.SE	Par. Est.	Boot.SE	Par. Est.	Boot.SE
INTERCEPT	56.555	9.803	60.735	11.607	24.651	21.0161	53.301	9.965
AGE	0.423	0.279	0.294	0.231	0.634	0.361	0.500	0.209
EPIT	-0.069	0.016	-0.065	0.020	-0.044	0.040	-0.065	0.015
CELL	116.673	35.907	124.266	42.510	181.415	90.040	110.468	33.715
HONEY	-10.392	1.670	-10.887	1.897	-9.794	3.759	-10.885	1.602

Robust Methods(continued)

Parameter	MM		GM1		GM6		Modified GM6	
	Par. Est.	Boot.SE	Par. Est.	Boot.SE	Par. Est.	Boot.SE	Par. Est.	Boot.SE
INTERCEPT	51.681	12.746	54.845	9.504	52.650	10.927	52.268	12.742
AGE	0.619	0.227	0.446	0.199	0.485	0.230	0.478	0.232
EPIT	-0.067	0.022	-0.066	0.015	-0.062	0.020	-0.062	0.025
CELL	101.491	48.972	112.809	33.484	111.747	50.093	113.980	50.322
HONEY	-11.547	1.838	-10.293	1.635	-10.676	2.072	-10.576	1.817

To explore more about the robust method which outperforms the others, the use of DMNCD may be an alternative approach. This is calculated from equation (19). The values of this statistics are given in Table III for different robust methods. As already been mentioned, this data set doesn't consist of any high leverage points with considerably large residuals. The results of Table III reveal that the MM-estimators have the largest DMNCD value compared to the other robust methods and one may wrongly conclude that GM-estimators do not perform better than MM-estimators and specially M-estimator. However, GM-estimators may consider all high leverage points and outliers by down weighting them without going far from the null model (having weight for all individuals equal to one) and so, for this data set, the use of GM6 or modified GM6 is suggested.

Table III. Distance to Maximum Normal Curvature Direction (DMNCD) for different robust regression methods for ILD data set

Robust Methods	DMNCD
M-estimator	1.124
MM-estimator	1.247
GM1	1.119
GM6	1.054
Modified GM6	1.067

5 Simulation Study

A Monte Carlo simulation study is carried out to confirm the result of the numerical example. Consider a linear regression model $y = 1 + 2 \times X_1 - 1 \times X_2 + \varepsilon$ where $X_i \sim N(0,1)$ $for\ i = 1,2$ and ε comes from a standard normal, i.e N(0,1) and two heavy

tailed distribution, the Standard skew-normal distribution when shape parameter is equal to 4 (SN(4), see Azzalini, 1985) and T-distribution with 3 degrees of freedom. The moderate sample size n=100 with 10000 replications has been used. The range of contamination is fixed to be 10 % of sample size in both X_1-and Y-direction. The contamination has been obtained by substituting the clean data with 2 times the maximum value of the generated data from the standard normal distribution.

Table IV consists of DMNCD of different robust methods. The same pattern as the analysis of our data set can be seen in the Table IV. Since the M-estimators are not robust to high leverage points, we are not going to discuss them, here. The GM6 and the modified GM6 have less DMNCD than those of MM and GM1. Thus, as it mentioned, the GM6, and the modified GM6 estimator doesn't go so far from the OLS estimates while they take into consideration the effects of outliers and high leverage points.

To see the efficiency of our new proposed method comparing to the other robust methods, the Mean Square Error (MSE) for the estimation of parameters by different robust regression methods are computed. The MSE of the estimator $\hat{\beta}$ of parameter β is defined as:

$$MSE_{\hat{\beta}} = \frac{\sum_{i=1}^{m}(\beta - \hat{\beta})^2}{m}$$

where m is the number of simulation replications and β is the true value of the simulated model. Table V exhibits the MSE of the three parameter estimates where 10% contamination exists in X_1-and Y-directions. According to this table, comparing the M-estimator, MM-estimators, GM1 and GM6 gives the conclusion that M-estimators give the largest MSE for estimating β_1 which makes these estimators unacceptable in our situations. The MM-estimator, the GM6 and the modified GM6 have acceptable MSEs for estimation of the parameters in all three different error distributions, normal and heavy-tailed distributions. Whereas, GM1 cannot be trusted when data generating process has heavy tail or is skewed. The simulation results reveal that the GM6 and the modified GM6 are reasonably and gives MSEs close to each other.

Table IV. Distance to Maximum Normal Curvature Direction (DMNCD) of different robust methods when 10% contamination exists *in* X_1- *and* Y- direction

Robust methods	N(0,1)	T(3)	SN(4)
	DMNCD	DMNCD	DMNCD
M	1.000	1.073	1.091
MM	1.198	1.216	1.189
GM1	1.181	1.220	1.235
GM6	1.020	1.095	1.166
Modified GM6	1.022	1.106	1.168

Table V. Mean Square Error (MSE) for estimation of the parameter in different robust regression methods where 10% contamination exists in X_1- and Y- direction

Coefficient Estimation	Error Distribution				
	N(0,1)				
	M	MM	GM1	GM6	Modified GM6
β_0	0.062	0.016	0.045	0.011	0.019
β_1	0.426	0.06	0.295	0.068	0.086
β_2	0.045	0.06	0.043	0.015	0.03
Coefficient Estimation	T(3)				
	M	MM	GM1	GM6	Modified GM6
β_0	0.299	0.058	0.142	0.055	0.047
β_1	2.157	0.157	0.694	0.352	0.169
β_2	0.068	0.087	0.043	0.034	0.041
Coefficient Estimation	SN(4)				
	M	MM	GM1	GM6	Modified GM6
β_0	0.830	0.457	0.718	0.601	0.507
β_1	0.428	0.021	0.336	0.008	0.020
β_2	0.010	0.005	0.007	0.003	0.005

6 Concluding Remarks

Least squares estimation is the predominant technique for regression analysis in most of different fields such as medical studies due to its universal acceptance, elegant statistical properties, and computational simplicity. Unfortunately, the statistical properties that make least squares so powerful depend on several assumptions that are often violated using real data. The normally distributed errors assumption, which enables tests of regressor significance, is invalid if only a single outlying observation occurs in the data. Not only is detection of these influential points important but also utilizing regression methods which are less sensitive to these points is more important in regression analysis. Among different diagnostics of influential points such as local influence (which is a method to evaluate the influence of local departures from assumptions in the model) and robust based measures, robust diagnostics methods are more powerful to detect outliers accurately. Moreover, robust regression methods are more resistant to outliers than the method of least squares. The GM-estimators are robust estimators which guard against leverage points which are the outliers in the X-direction, as well as the outliers in the Y-direction. Several proposed GM-estimation methods exists in the literature. In this study, two of the most common methods of Schweppe (GM1) and Coakley and Hettmansperger (GM6) were compared. A modified version of the GM6 was also proposed. Normal curvature of

likelihood displacement is a very important concept which provides a general approach to study the problem of influence in this paper. To compare the robust regression methods, Distance to Maximum Normal Curvature Direction (DMNCD) was also introduced. The results of the conducted simulation indicate that our modified GM-estimator has the minimum DMNCD through down weighting outliers especially high leverage points. To verify our results, a clinical trial data set and a simulation study were performed. The results of the real data agree reasonably well with the results of the simulation study that the performance of the GM6 and modified GM6 are equally good, robust in both x and y directions and outperform other robust estimators. Hence, the modified GM6 should provide a robust alternative to the well known GM6 estimator.

References

[1] R. Andersen, Modern Methods for Robust Regression, Sara Miller McCune, SAGE publications, The United States of America , 2008.
[2] R. D. Armstrong and M. T. Kung, Least absolute values estimates for a simple linear regression, problem, Applied Statistics, 27 (1978), 363-366.
[3] A. Azzalini, A class of distribution which includes the normal ones, Scandinavian Journal of Statistics, 12 **(**1985), 171-178.
[4] D. A. Belsley, E. Kuh and R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, New York ,1980.
[5] N. Billor, R. M. Loynes, Local Influence: A new approach, communications in statistics.-theory and methodology, 22 (1993), 1595-1611.
[6] J. B. Birch, Some convergence properties of iterated reweighted least squares in the location model, Communications in Statistics - Simulation and Computers, B9 (1980), 359-369.
[7] G. E. P. Box, Non-normality and tests on variances. Biometrika, 40 (1953), 318-335.
[8] N. A. Campbell, H. P. Lopuhaä and P. J. Rousseeuw, On the calculation of a robust S-estimator of a covariance matrix, Statistics in Medicine, 17(23) (1998), 2685-2695.
[9] H. C. Causton, J. Quackenbush and A. Brazma, Microarray Gene Expression Data Analysis: A Beginner's Guide, Blackwell publishing, 2003.
[10] S. Chatterjee, A. S. Hadi, Influential observations, high leverage points, and outliers in linear regression, Statistical Science, 1 (3) (1986), 379-416.
[11] C. W. Coakley, T. P. Hettmansperger, A bounded-influence, high-breakdown, efficient regression estimator, Journal of the American Statistical Association, 88 (1993), 872–880.
[12] R. D. Cook, Assessment of local influence (with discussion), Journal of the Royal Statistical Society, Series. B: 48 (1986), 133-169.
[13] R. D. Cook, Detection of influential observation in linear regression. Technometrics, 19 (1977), 15-18.
[14] R. D. Cook, Influential observations in linear regression, Journal of American Statistical Association, 74 **(**1979), 169-174.

[15] R. D. Cook, and D. M. Hawkins, Unmasking multivariate outliers and leverage points: comment, Journal of the American Statistical Association, 85 (1990), 640-44.
[16] R. D. Cook, and Weisberg S., Residuals and Influence in Regression. London, Champan Hall, 1982.
[17] N. R. Draper, and Smith H., Applied Regression Analysis, Wiley, New York, 1998.
[18] A. S. Hadi, A new measure of overall potential influence in linear regression, Computational and Statistical Data Analysis, 14 (1992), 1-27.
[19] F. R. Hampel, The influence curve and its role in robust estimation, Journal of the American Statistical Association, 69 (1974), 383-393.
[20] E. Handschin, F. C. Schweppe, J. Kohlas, and A. Fiechter, Bad data analysis for power system state estimation, IEEE Transactions of Power Apparatus and Systems, PAS-94 (1975), 329–337.
[21] R. W. Hill, Robust Regression When There Are Outliers in the Carriers, unpublished Ph.D, dissertation, Harvard University, Boston, MA, 1977.
[22] P. J. Huber, Robust regression: asymptotic, conjectures, and Monte Carlo, The Annals of Statistics, 1 (1973), 799-821.
[23] P. J. Huber, Robust statistics: A review. annual of mathematics and statistics, 43 (1972), 1041-1067.
[24] P.J. Huber, Robust Statistics. John Wiley & Sons, New York, 2004.
[25] P. J. Kempthorne, and M. B. Mendel, Unmasking multivariate outliers and leverage points: comment, Journal of the American Statistical Association, 85 (1990), 647-48.
[26] D. C. Montgomery and E. A. Peck, Linear Regression Analysis, Wiley, New York, 1982.
[27] F. Mosteller and J. W. Tukey , Data Analysis and Regression. Reading, MA: Addison-Wesley, 1977.
[28] S. C. Narula, P. N. Saldiva, C. D. S. Andre, S. N. Elian, A. F. Ferreira, and V. Capelozzi, The minimum sum of absolute errors regression: a robust alternative to the least squares regression. Statistics in Medicine, 18 (1999), 1401-1417.
[29] E. S. Pearson, The analysis of variance in cases of non-normal variation. Biometrika, 23 **(**1931), 114-133.
[30] D. Pena, A new statistics for influence in linear regression. Technometrics, 47(1) (2005), 1-12.
[31] W. Y. Poon, and Y. S. Poon, Total behavior of likelihood displacement. Statistica Sinica, 12 (2002), 599-607.
[32] M. M. S. Rancel, and M. A. G. Sierra, A connection between local and deletion influence. Sankhya: The Indian Journal of Statistics, Series A, 62 (2000), 144-149.
[33] P. J. Rousseeuw and C. Croux, Alternatives to the median absolute values, Journal of the American statistical association, 88 (1993), 1273-83.
[34] P. J. Rousseeuw, and B. C. van Zomeren, Unmasking multivariate outliers and leverage points, Journal of the American Statistical Association, 85 (1990), 633–639.
[35] P. J. Rousseeuw, and V. Yohai, Robust regression by means of S-estimators, Robust and Nonlinear Time series Analysis. Lecture Notes in Statistics, Springer Verlag, New York, 26 (1984), 256-272.

[36] P. J. Rousseeuw, Least median of squares regression. Journal of the American Statistical Association, 79 **(**1984), 871-880.
[37] P. J. Rousseeuw, Multivariate estimation with high breakdown point. Mathematical Statistics and Applications, Vol. B (1983), 283-297.
[38] P. J. Rousseuw, and A. M. Leroy, Robust Regression and Outlier Detection, New York, John Willy, 2003.
[39] D. Ruppert, and D. G. Simpson, Unmasking multivariate outliers and leverage points: comment, Journal of the American Statistical Association, 85 (1990), 644-646.
[40] J. R. Simpson, New Methods and Comparative Evaluations for Robust and Biased-Robust Regression Estimation. PhD thesis, Arizona State University, 1995.
[41] A. J. Stromberg, O. Hossjer, and D. M. Hawkins, The least trimmed differences regression estimator and alternatives, Journal of the American Statistical Association, 95 **(**2000), 853-864.
[42] N. H. Timm, Applied Multivariate Analysis, Springer, 2002.
[43] R. R. Wilcox, Introduction to Robust Estimation and Hypothesis Testing, 2th edition, Elsevier academic press, USA, 2005.
[44] V.J. Yohai, High breakdown point and high efficiency robust estimates for regression. The Annals of Statistics, 15 (1987), 642–656.

Received: November, 2009