*Journal of Statistical Research* *ISSN 0256 - 422 X*
*2008, Vol. 42, No. 2, pp. 105-120*
*Bangladesh*

# DIAGNOSTIC ROBUST APPROACH OF OUTLIER DETECTION IN REGRESSION

A.H.M. Rahmatullah Imon

*Department of Mathematical Sciences, Ball State University, Muncie, IN 47306, U.S.A.*
*Email: imon_ru@yahoo.com*

## SUMMARY

The identification of outliers in data has been an area of a great deal of attention for many years. The outlier detection procedure is more cumbersome in regression where outliers may occur in the response variable or in the explanatory variables or both. A variety of diagnostic methods are now being used for the identification of different types of outliers in regression. These methods, however, are successful only if the data set contains a single outlier. In the presence of multiple outliers diagnostic methods often fail to detect the outliers. This is due to the well-known problems of masking and swamping effects. On the other hand the robust methods can identify the outliers correctly but they are too prone to declare observations to be outlier which is not also desired. In this paper we discuss an approach which is a compromise between these two approaches. We call this approach diagnostic-robust approach where the suspect outliers are identified first by robust methods and diagnostic methods are applied later to confirm the suspicion. We consider several well-known data sets to investigate the performance of the diagnostic-robust approach in the detection of outliers in regression.

*Keywords and phrases:* Outliers; High Leverage Points; Influential Observations; Regression Diagnostics; Masking; Swamping; Robust Regression; Generalized Studentized Residuals; Generalized Potentials; Generalized DFFITS.

*AMS Classification:* Primary 62 J20, Secondary 62G35

# 1 Introduction

The concept of outliers in a data set is considered to be as old as the subject of statistics. Barnett and Lewis (1994) point out that even before the formal development of statistical method, argument raged over whether, and on what basis, we should discard observations from a set of data on the grounds that they are unrepresentative. Roughly speaking an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data. Hampel et al. (1986) claim that a routine data set typically contains about 1-10% outliers in it and even the highest quality

© Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh.

data set can not be guaranteed free of outliers. Outliers are usually the extreme values in the sample. But this statement is not always true, especially in regression analysis.

In a regression problem, observations are judged as outliers on the basis of how unsuccessful the fitted regression equation is in accommodating them and that is why observations corresponding to excessively large residuals are treated as outliers. In linear regression it is almost conventional to use the ordinary least squares (OLS) method for estimating the parameters and fitting the model. Because the OLS technique minimizes squared deviations, it has a tendency to put a relatively heavy weight on outliers and parameter estimates are extremely sensitive to their presence. In OLS method, the residual mean square is generally used to estimate the variance of the errors. The residual mean sum of squares can be greatly inflated by outliers so that we may not be able to reliably estimate the variance of the errors and consequently the entire inferential procedure may be in faulty.

The unfortunate consequences of the presence of outliers in regression have been reported by many authors (see Hadi, Imon and Warner, 2008a; 2008b). Outliers could produce a wrong fitted line. The goodness-of-fit static like $R^2$ could be highly inflated or deflated in the presence of outliers. Outliers could induce nonnormality in the data and consequently all conventional tests based on $t$, chi-square and $F$ statistics become invalid in the presence of outliers. Chatterjee and Hadi (1988) pointed out that there may be some observations in the data whose presence could induce collinearity among the linearly independent regressors or whose presence could break the existing collinearity structure among the regressors. These observations are known as collinearity-influential observations. Imon and Khan (2003) point out that high leverage points are the prime source of collinearity-influential observations. The identification of variance heteroscedasticity could be very complicated in the presence of outliers (see Imon, 2008). Most of the commonly used variable selection techniques for model building are affected in the presence of outliers (see Imon and Alam, 2008). The detection and handling of outliers are really necessary to overcome these unfortunate consequences.

In the literature we find mainly two approaches of outlier detection in regression. The old but the most popular approach of outlier detection is the diagnostic approach where the regression model is fitted by the OLS method first and then diagnostic tools are used to find the observations that do not match with the fitted line and these observations are called outliers. But outliers can distort the OLS fit in a way that we may not identify the potential outliers unless an observation or a group of observations are omitted from the fit. This gives rise deletion diagnostics but the problem still remains how many and on what basis observations should be omitted to get a reliable fit. It is often observed that after the deletion of an outlier another observation may emerge as an outlier which was not understood at first. This type of problem is known as masking of outliers. The opposite effect of this distortion is known as swamping, which makes inliers appear outliers. Robust approach of outlier detection is introduced to fit the model where the effect of outliers is kept small. This approach gives better fit of the model overall and genuine outliers get better chance to be identified. Although the robust fit can handle the masking problem it has a tendency of swamping too many observations. Neither of these two situations is desirable and that is why we need an outlier detection technique which can identify the genuine outliers but not more than the necessary.

## 2 Diagnostic Approach of Outlier Detection

Diagnostics are designed to find problems with the assumptions of any statistical procedure. In diagnostic approach we estimate the parameters (in regression fit the model) by the classical method and then see whether there is any violation of assumptions and/or irregularity in the results. In regression the diagnostic approach of outlier detection is to fit the regression model first by the OLS method and then observations are judged as outliers on the basis of how unsuccessful they are to match with the OLS results. In the literature there are several versions of outliers for a regression problem. These are classified as

(i) $X$ Outlier: This is a point that is outlying in regard to the $x$-coordinate. In the literature an $X$ outlier is more popularly known as a high leverage point.

(ii) $Y$ Outlier: This is a point that is outlying only because its y-coordinate is extreme. This kind of outlier is also known as vertical outlier.

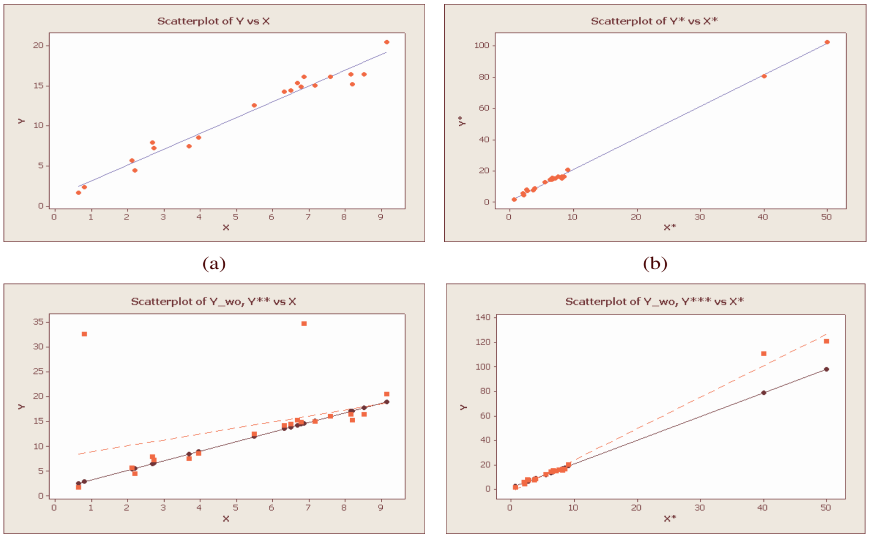
(iii) $X-Y$ Outlier: A point that is outlying in both $x$ and $y$ coordinates is known as $X-Y$ outlier. An $X-Y$ outlier is also known as a bad leverage point.

(iv) Regression Outlier: A regression outlier is a point that deviates from the linear relationship determined from the other points, or at least from the majority of those points. An observation which is an $X$ outlier not a regression outlier is known as a good leverage point.

(v) Residual Outlier: This is a point that has a large standardized residual. Most of the commonly used outlier detection methods are based on residual outliers.

In fitting a linear regression model by the OLS method we often observe that a variety of estimates can be substantially affected by few observations. To quote Chatterjee and Hadi (1988) "...not all the observations have an equal importance in least squares regression and, hence, in conclusions that result from an analysis". According to Belsley, Kuh and Welsch (1980), an influential observation is one which, either individually or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates than is the case for most of the other observations. In this situation parameter estimates or predictions may depend more on the influential observations than on the majority of the data and their omission from the data may result in substantial changes to important features of an analysis.

As we have already mentioned that in linear regression, outliers usually mean residual outliers and for this reason observations possessing large residuals are suspects, an immediate question comes to our mind, how large is large? Different versions of standardized residuals



(a) (b) (c) (d)

Figure 1: (a) no outlier (b) $X$ outliers (c) $Y$ outliers (d) $X-Y$ outliers

are used for the identification of outliers in linear regression (see Chatterjee and Hadi, 2006). Among them the deletion Studentized (externally Studentized or $R$-Student) residuals are most commonly used. These residuals are defined as

$$t_i = \frac{\hat{\in}_i}{\hat{\sigma}_{(i)}\sqrt{1-w_{ii}}}, \quad i = 1, \ldots, n \tag{2.1}$$

where $\hat{\in}_i$ is the $i$-th OLS residual, $\hat{\sigma}^2_{(i)}$ is the OLS estimates of the mean squared error (MSE) based on a data set with the $i-$th observation deleted and

$$w_{ii} = x_i^T (X^T X)^{-1} x_i, \quad i = 1, \ldots, n \tag{2.2}$$

is the $i$-th leverage value. As a thumb rule we call an observation outlier when its corresponding deletion Studentized residual value exceeds 3 in absolute term.

The diagnostic approach for the identification of high leverage points is to inspect the observation which does not match with the average leverage structure of the data. The $i$-th leverage value $w_{ii}$ as defined in (2.2) is in fact the $i$-th diagonal element of the matrix

$$W = X \left(X^T X\right)^{-1} X^T \tag{2.3}$$

which is generally known as a hat or weight or leverage matrix. Observations possessing excessively large $w_{ii}$ values are called high leverage points. The *twice-the-mean-rule* (Hoaglin and Welsch, 1978), the *thrice-the-mean-rule* (Vellman and Welsch, 1981), Huber's rule

(Huber, 1981) and Mahalanobis distances (see Rousseeuw and Leroy, 1987) are commonly used measures of leverages in the literature. Hadi (1992) introduced a new type of measure, where the leverage of the $i$-th point is based on a fit to the data with the $i$-th case deleted. He named this measure potentials and defined the $i$-th potential as

$$p_{ii} = x_i^T \left(X^T X\right)^{-1} x_i, \quad i = 1, \ldots, n \tag{2.4}$$

where $X_{(i)}$ is the data matrix $X$ with the $i$-th row deleted. Observations corresponding to excessively large potential values are considered as high leverage points.

A good deal of detection methods is now available in the literature for the identification of influential observations (see Chatterjee and Hadi, 2006). Among them the Cook's distance and the difference in fits (DFFITS) are commonly used. But DFFITS are more preferred (Imon, 2005) as they give better information about the scale estimate and retain signs which are extremely important in influence analysis. The $i$-th DFFITS is defined as

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_i^{(-i)}}{\hat{\sigma}_{(i)}\sqrt{w_{ii}}}, \quad i = 1, \ldots, n \tag{2.5}$$

where $\hat{y}_i^{(-i)}$ is the $i$-th fitted response and the estimated standard error with the $i$-th observation deleted. An observation is considered as influential if $|\text{DFFITS}i| > 3\sqrt{k/n}$

Most of the recent diagnostic approach of outlier detection is based on case deletion. According to Maronna, Martin and Yohai (2006) deleting an outlier, although better than doing nothing, still poses a number of problems: when is deletion justified? Deletion requires a subjective decision. When is an observation 'outlying enough' to be deleted? The user may think that 'an observation is an observation' and hence feel uneasy about deleting them. Since there is generally some uncertainty as to whether an observation is really an outlier, there is a risk of deleting 'good' observations, which results in underestimating data variability. Since the results depend on the user's subjective decisions, it is difficult to determine the statistical behaviour of the complete procedure. We also need to remember that outliers are not necessarily bad, on the contrary, they may be the most informative observations in the data. Diagnostic methods are not useful in the identification of multiple outliers. In the presence of outliers we cannot estimate models reliably so we need to identify outliers correctly. But we cannot identify outliers correctly unless the model is estimated correctly.

As an example, let us consider a data set given by Haith (1976) to study the relationship between water quality and land use on 20 river basins in New York State. A question of interest is how the land use around a basin contributes to the water pollution as measured by the mean nitrogen concentration. The scatter plot of nitrogen versus land use together with the corresponding fitted OLS line is given in Figure 2(a). It is clear from this plot that there are two clear outliers in the data. But the outliers pull the fitted OLS line to their directions in a way that one of the two outliers is completely masked. Masking of one outlier is clearly visible when we look at the index plot of the deleted Studentized residuals as shown in Figure 2(b). But it is more interesting to note that the numerical values of all deleted Studentized residuals are within range ±3 so we fail to identify even a single outlier by diagnostic approaches.

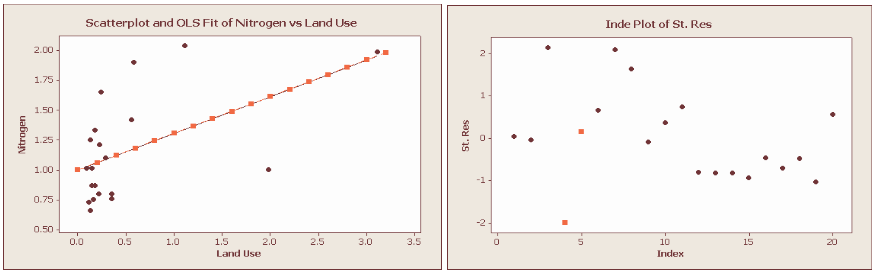


Figure 2: (a) Scatter plot of nitrogen versus land use (b) Index plot of deleted Studentized residual for the New York rivers data

Our next example is the well-known Hawkins-Bradu-Kass (1984) data. This three-predictor data set contains 75 observations with 10 outliers (cases 1-10) and 14 high leverage points (cases 1-14). The 3D plot of the predictors (Figure 3(a)) shows the existence of 14 high leverage points. But the index plot of leverages as shown in Figure 3(b) shows that only one of them is identified as a high leverage point while the other 13 get masked.

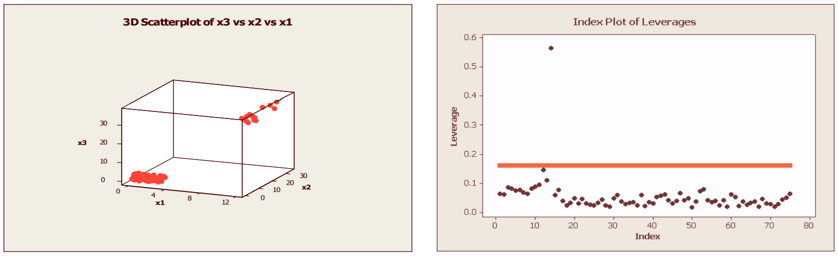


Figure 3: (a) 3D plot of predictors (b) Index plot of leverages for the Hawkins et al. (1984) data

We have already mentioned that the Hawkins-Bradu-Kass data contains 10 outliers (observations 1-10) which are also the points of high leverages. Consequently these 10 observations should be the most influential cases. This data also have a relatively less influential group (cases 11-14) which are high leverage points but not outliers. But we observe a different picture when we look at the index plot of DFFITS for the identification of influential observations for the Hawkins et al. (1984) data. We observe from Figure 4 that observations 11-14 appear to be more influential than observations 1-10 and none of the most influential 10 cases exceeds the cut-off value set by the DFFITS rule and thus get masked.

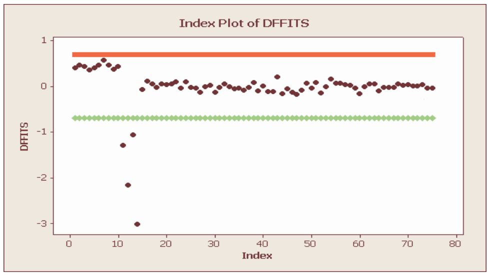


Figure 4: Index plot of DFFITS for Hawkins et al. (1984) data

## 3 Robust Approach of Outlier Detection

Robust statistics have been in use for hundreds of years (see Stigler , 1973) but not seriously until quite recently. According to Huber (1981) the term robustness signifies insensitivity to small deviations from the assumption. That means a robust procedure is nearly as efficient as the classical procedure when classical assumptions hold strictly but is considerably more efficient over all when there is a small departure from them. Much of the work in robust regression has been motivated by the Princeton Robustness Study (see Andrews et al., 1972), in which it was learned that the OLS estimator can be inferior to other estimation approaches when the distribution of the error term is not normal. To quote Hampel et al. (1986), "Unfortunate consequences of departure from the simple OLS model have long been suspected by statisticians Despite this fact, the OLS method has retained its popularity over the years in a hope that slight departure from standard assumptions would not affect inferences too much." It is now evident that this type of departure may have drastic consequences on both estimation of parameters and testing of hypotheses.

The main application of robust techniques in a regression problem is to try to devise estimators that are not strongly affected by outliers. In linear regression, robust techniques grew up in parallel to diagnostics and initially they were used to estimate parameters and to construct confidence intervals in such a way that outliers or departures from the assumptions do not affect them. But in recent years, a rationale for this technique has been mainly the identification of outliers. For this reason diagnostics and robust regression are considered to be complementary to each other. Diagnostic and robust regression have the same goals, but in the opposite order. To quote Rousseeuw and Leroy (1987) "When using diagnostic tools, one first tries to delete the outliers and then to fit the good data by the least squares, whereas a robust analysis first wants to fit a regression to the majority of the data and then to discover the outliers as those points which possess large residuals from that robust solution." According to Barnett and Lewis (1994) attitudes towards outliers have varied from one extreme to another: from the view that we should never sully the sanctity of the data by daring to adjugate its propriety, to an ultimate pragmatism expressing if in doubt, throw it out. This change in attitude towards outliers is also reflected in robust techniques in the way of handling them. Early techniques were based on trying to keep small the effects of the presence of outliers in the data set. Later techniques are developed on the idea that we should only use that part of a data set that 'fits together.'

An excellent review of different robust regression methods with their relative advantages and shortcomings are available in Maronna, Martin and Yohai (2006). Among these methods the L–estimator (Maronna and Yohai, 2000), the M–estimator (Huber, 1973), the least median of squares (LMS) (Rousseeuw, 1984), the least trimmed squares (LTS) (Rousseeuw and Leroy, 1987), the bounded influence or generalized M (GM) estimator (Mallows, 1975), the S–estimator (Rousseeuw and Yohai, 1984), the MM–estimator (Yohai, 1987), the blocked adaptive computationally-efficient outlier nominators (BACON) (Billor, Hadi and Vellman, 2000) and the fast iterative estimators (Atkinson, 1994; Hawkins, 1994; Rousseeuw and van Driessen, 2000; Salibian-Barrera and Yohai, 2005) have become popular with the statisticians. The robust approach of outlier detection in regression is to fit the regression model first by any suitable robust regression method. Then identify observations as outliers which fail to match with the robust fit. For example, Rousseeuw and Leroy (1987) suggested an outlier detection method where the regression line is fitted by the LMS (or LTS) method. The $i$-th LMS residual $e_i$ and the scale estimate $\hat{\sigma}_{LMS}$ are obtained from this robust fit. Observations for which

$$\frac{|e_i|}{\hat{\sigma}_{LMS}} > \quad 2.5, \quad i = 1, \ldots, n \tag{3.1}$$

are identified as outliers. They extended this idea of robust outlier detection to get another robust fit, known as the reweighted least squares (RLS), where outliers thus identified are omitted before fitting the model by the OLS and the responses (and also the residuals) corresponding to the deleted points are re-estimated.

Robust Mahalanobis squared distance (RMSD) based on the minimum volume ellipsoid (MVE) (Rousseeuw, 1984) or the minimum covariance determinants (MCD) (Rousseeuw, 1984) and clustering techniques (Pena and Prieto, 2001; Maronna and Zamar, 2002) are well-known robust methods for identifying $X-$ outliers.

It is now evident that diagnostic methods of outlier detection are subject to masking and/or swamping in the study of the identification of multiple outliers. Most of the robust techniques are based on the most compact data set (nearly half of the observations). Those techniques indeed may be very robust and may possess very high breakdown, but they have a tendency to identify innocent observations as outliers. Robust estimators of scale are often seriously underestimated. This gives a spurious precision to subsequent analysis so that the detection procedures may identify observations as outliers which are actually not. Cook and Hawkins (1990) noted that even considering a sample of 20 five-dimensional N(0,1) data the robust techniques identify 6 cases as outliers. They described this situation as a case of 'outliers everywhere.' This type of unfortunate consequences in the identification of $X-$ outliers and influential observations are reported by Imon (2002, 2005) and Midi, Ramli and Imon (2008).

# 4 Diagnostic-Robust Approach of Outlier Identification

The diagnostic–robust approach of outlier detection is a combination of diagnostic and robust approaches. This approach consists of three steps:

Step 1: Find all suspect outliers by robust techniques.

Step 2: Apply diagnostic techniques to verify whether the suspected cases are genuine outliers are not.

Step 3: 'Innocent' observations are put back into the data set if they are wrongly identified as outliers by robust techniques.

At first we present group deletion residuals and leverages that are basis of diagnostic–robust approach of outlier detection. Denote a set of cases 'remaining' in the analysis by $R$ and a set of cases 'deleted' by $D$. Also suppose that $R$ contains $(n-d)$ cases after $d < (n-k)$ cases in $D$ are deleted. Without loss of generality, assume that these observations are the last $d$ rows of $X$ and $Y$ so that they can be partitioned as

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix} \quad , \quad Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix}$$

Then the weight matrix $W = X(X^T X)^{-1} X^T$ becomes

$$W = \begin{bmatrix} U_R & V \\ V^T & U_D \end{bmatrix}$$

where $U_R = X_R(X^T X)^{-1} X_R^T$ and $U_D = X_D(X^T X)^{-1} X_D^T$ are symmetric matrices of order $(n-d)$ and $d$ respectively, and $V = X_R(X^T X)^{-1} X_D^T$ is an $(n-d) \times d$ matrix. Hence using the result of Henderson and Searle (1981), $(X_R^T X_R)^{-1}$ can be expressed as

$$\begin{aligned}(X_R^T X_R)^{-1} &= (X^T X - X_D^T X_D)^{-1} \\ &= (X^T X)^{-1} + (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} X_D (X^T X)^{-1} \qquad (4.1)\end{aligned}$$

where $I_D$ is an identity matrix of order $d$. Then the vector of estimated parameters after the deletion of $d$ observations, denoted by $\hat{\beta}^{(-D)}$, is obtained using (4.1) as

$$\hat{\beta}^{(-D)} = (X_R^T X_R)^{-1} X_R^T Y_R = \hat{\beta} - (X^T X)^{-1} X_D^T (I_D - U_D)^{-1} \hat{\in}_D \qquad (4.2)$$

where $\hat{\in}_D = Y_D - X_D \hat{\beta}$. Thus an $n \times 1$ vector of deletion residuals can be defined as

$$\hat{\in}^{(-D)} = Y - X \hat{\beta}^{(-D)} \qquad (4.3)$$

From this the $i$-th deletion residual is defined by

$$\hat{\in}_i^{(-D)} = y_i - x_i^T \hat{\beta}^{(-D)}, \quad i = 1, \ldots, n \qquad (4.4)$$

To develop the idea of group-deleted leverages, Imon (2002) extended the definition of a single case deleted potential to group deletion. He named the resulting values as generalized potentials (GP) and studied their usefulness when a group of high leverage cases were masked. Define

$$w_{ii}^{(-D)} = x_i^T (X_R^T X_R)^{-1} x_i, \quad i = 1, \ldots, n \tag{4.5}$$

so that $w_{ii}^{(-D)}$ is the $i$-th diagonal element of $X(X_R^T X_R)^{-1} X^T$ matrix. When $D = i$, we observe from (2.4) and (4.5) that

$$w_{ii}^{(-i)} = x_i^T \left( X_{(i)}^T X_{(i)} \right)^{-1} x_i = p_{ii}. \tag{4.6}$$

Suppose now that a further point $i$ is removed from the remaining subset $R$ and joins the deleted subset $D$. For any such $i$, it is easy to show that

$$w_{ii}^{-(D+i)} = \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} \tag{4.7}$$

This tells us that the potential value of any case $i$, generated externally should be equivalent to the quantity $\frac{w_{ii}^{(-D)}}{1-w_{ii}^{(-D)}}$, when $w_{ii}^{(-D)}$ is generated internally on a reduced sample space $R$.

## 4.1 Identification of Outliers

Here we outline a diagnostic-robust method for the identification of multiple outliers in linear regression. At first we employ any robust technique to fit the model. We prefer using the BACON technique proposed by Billor, Hadi and Vellman (2000) to find the 'clean subset' because it is equally as reliable as most of the robust techniques but computationally less extensive. After choosing a clean subset indexed by $R$, we compute scaled residuals for the entire data set. The expressions (4.4) and (4.7) can be combined together to define generalized Studentized (GS) residuals (see Imon, 2005) for the entire data set as

$$\begin{aligned} t_i^* &= \frac{\hat{\epsilon}_i^{(-D)}}{\hat{\sigma}_{R-i}\sqrt{1 - w_{ii}^{(-D)}}} \qquad \text{for} \quad i \in R \\ &= \frac{\hat{\epsilon}_i^{(-D)}}{\hat{\sigma}_{R}\sqrt{1 + w_{ii}^{(-D)}}} \qquad \text{for} \quad i \in D \end{aligned} \tag{4.8}$$

Generalized Studentized residuals defined in (4.8) are analogous to residuals suggested by Hadi and Simonoff (1993) and Atkinson (1994). The scaled residuals for the $R$ set are in fact the Studentized residuals on $R$. On the other hand for the $D$ set they are the externally Studentized residuals on $R$. For this reason Imon (2005) renamed these residuals as generalized Studentized residuals. We call observations outliers if their corresponding GS residuals exceed 3 in absolute terms. Since generalized Studentized residuals are measured in a similar scale, it should not matter too much if a point possessing low residuals is included in the deletion set. But we prefer to put them back into the estimation subset $R$ sequentially (observation with the least absolute GS value will be replaced first) for the improved estimation of standard errors and thus we get correct inference. We continue this process until all members of the deletion set individually possesses absolute GS residuals bigger than 3. The points thus identified will be finally declared as outliers. Another diagnostic-robust approach for multiple outlier detection is the best omitted from the ordinary least squares (BOFOLS) technique proposed by Davies, Imon and Ali (2004).

We revisit the New York rivers data once again. When we apply the diagnostic-robust technique to fit the data we observe (see Figure 5(a)) that outliers cannot affect the fitting of the model any more and two outliers are clearly separated from the rest of the data. The index plot of the generalized Studentized residuals (see Figure 5(b)) also shows that two outliers have much bigger value than 3 in absolute term and are clearly identified as outliers.

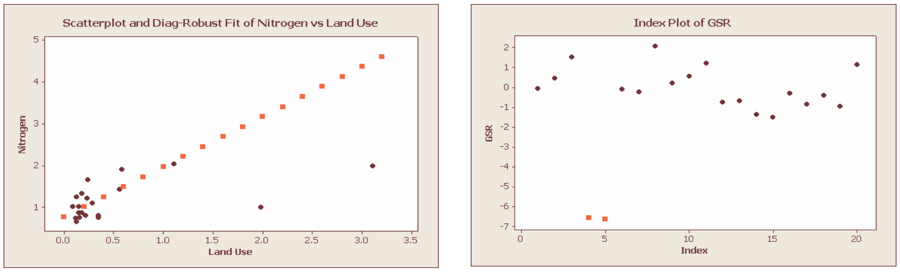


Figure 5: (a) Diagnostic-robust fit (b) Index plot of GSR for the New York rivers data

## 4.2 Identification of Multiple High Leverage Points

Here we introduce a stepwise procedure (see Midi, Ramli and Imon, 2008) for the identification of multiple high leverage points in linear regression. We first employ any suitable robust technique (we prefer MCD) to identify suspect high leverage points. We form the deletion set $D$ that contains all suspect cases. After the formation of the $D$ set, we use (4.5) and (4.7) to compute the generalized potentials proposed by Imon (2002) defined as

$$\begin{aligned} p_{ii}^* &= \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} \qquad \text{for} \quad i \in R \\ &= w_{ii}^{(-D)} \qquad \text{for} \quad i \in D, \end{aligned} \tag{4.9}$$

where $D$ is any arbitrary deleted set of points. There exists no finite upper bound for $p_{ii}^*$'s and it may not be easy to derive a theoretical distribution of them. But this does not make any problem to obtain a suitable confidence bound type cut-off point for them. We consider $p_{ii}^*$ to be large if

$$p_{ii}^* > Median\ (p_{ii}^*) + 3MAD\ (p_{ii}^*) \tag{4.10}$$

where $MAD\ (p^*_{ii}) = Median\ \{|p^*_{ii} - Median\ (p^*_{ii})|\}$. The final step of this process is the checking for swamping. We put back observations sequentially to the estimation subset if they do not exceed the above cut-off point and continue this process until all genuine high leverage points are identified.

## 5 Identification of Multiple Influential Observations

Here we outline a method for the identification of multiple influential observations in regression. We would like to consider outliers and high leverage points separately as suspects and observations that are considered either as outliers or high leverage points become members of the deletion set. For the identification of suspect outliers we could use any robust regression technique like LMS, LTS or BACON. For the identification of suspect high leverage points one could consider robust methods like generalized potentials based on MCD, MVE or RMD. Observations thus identified either as outliers or high leverage points will form the deletion set $D$. We compute the generalized difference in fits ($GDFFITS$) proposed by Imon (2005) for the entire data set defined as

$$\begin{aligned} GDFFITS_i &= \frac{\hat{y}_{i(R)} - \hat{y}_{i(R-i)}}{\hat{\sigma}_{R-i}\sqrt{w_{ii(R)}}} \qquad \text{for} \quad i \in R \\ &= \frac{\hat{y}_{i(R+i)} - \hat{y}_{i(R)}}{\hat{\sigma}_{R}\sqrt{w_{ii(R+i)}}} \qquad \text{for} \quad i \notin R. \end{aligned} \tag{5.1}$$

The detection rule suggested for single case deletion DFFITS may apply to the $GDFFITS_i$'s. We suggest considering observations as influential if

$$|GDFFITS_i| \ \geq \ 3\sqrt{k/\,(n-d)}.$$

We anticipate that sometimes the rules for the selection of initial deletion set may be very sensitive and that is why, some of the good observations are swamped in as outliers either in the $X$-space or in the $Y$-space or both. So it may be necessary for checking in swamping before the declaration of any of the observations to be influential. Sometimes we may observe that one or more of the members of the initial deletion set do not satisfy rule (4.12). So these are not potential influential cases. At this stage, we can put back the observations into the estimation subset sequentially; observations possessing the lowest $GDFFITS_i$ values will come back first into the estimation subset. We will continue this process until all of the potential influential cases individually satisfy rule (4.12) and observations thus identified are finally declared as influential observations.

Here we consider the Hawkins-Bradu-Kass (1984) data to identify high leverage points and influential observations. We know that cases 1-14 are high leverage points and the diagnostic-robust generalized potentials can identify all 14 cases (see Figure 6(a)) successfully. As we have already mentioned that the Hawkins-Bradu-Kass data contains 10 high leverage outliers (observations 1-10) which should be the most influential cases followed

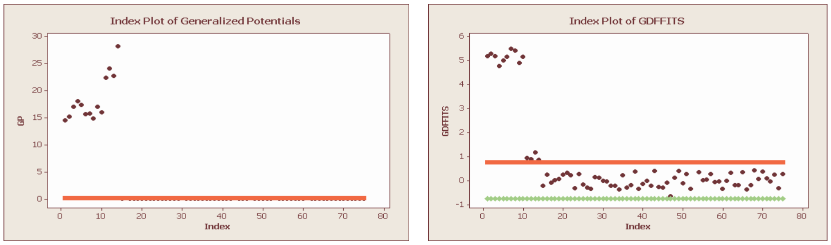


Figure 6: Index plot of (a) generalized potentials (b) $GDFFITS$ for Hawkins et al. (1984) data

by a relatively less influential group (cases 11-14) which are high leverage points only, the diagnostic-robust $GDFFITS$ give a clear picture about influences of all observations. We observe from the index plot of $GDFFITS$ as given in Figure 6(b) that cases 1-10 appear as the most influential cases followed by cases 11-14 and they are clearly separated from the rest of the data.

## 6 Conclusions

Identification of outliers in regression is really important because their presence can often cause huge problems in inference. Regression diagnostics are often used to detect outliers and they are very popular for their simple nature. But diagnostic procedures suffer huge setback when data set contains multiple outliers. There exist robust methods which are very successful in the identification of outliers but they have tendencies to identify good observations as outliers. Diagnostic-robust approach of outlier detection is proposed as a compromise between these two approaches; robust techniques are employed first to identify suspect outliers and then diagnostics are used to confirm their status. In this paper we present several examples which show that the diagnostic-robust approach avoids some drawbacks of diagnostic or robust approach in the identification of outliers.

## Acknowledgements

The author is thankful to the Referees and Managing Editor for their valuable comments and suggestions.

## References

[1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.F. (1972). *Robust Estimates of Location*, Princeton University Press, New Jersey.

[2] Atkinson, A.C. (1981). Two graphical displays for outlying and influential observations in regression, Biometrika, 68, 13-20.

[3] Atkinson, A.C. (1994). Fast very robust methods for the detection of multiple outliers, *J. Amer. Stat. Assoc.*, **89**, 1329-1339.

[4] Barnett, V. and Lewis, T.B. (1994). *Outliers in Statistical Data*, 3rd ed., Wiley, New York.

[5] Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.

[6] Billor, N., Hadi, A.S., and Velleman, F. (2000). BACON: Blocked adaptive computationally-efficient outlier nominator, *Comput. Stat. Data Anal.*, **34,** 279-298.

[7] Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression*, Wiley, New York.

[8] Chatterjee, S. and Hadi, A.S. (2006). *Regression Analysis by Examples*, $4^{th}$ ed., Wiley, New York.

[9] Cook, R.D. and Hawkins, D.M. (1990). Comment on 'Unmasking multivariate outliers and leverage points' by Rousseeuw, P.J. and van Zomeren, B.C*., J. Amer. Stat. Assoc.*, **85**, 640-644.

[10] Davies, P., Imon, A. H. M. R., and Ali, M. M. (2004). A conditional expectation method for improved residual estimation and outlier identification in linear regression, *Int. Jour. Statist. Sci.* (Special issue in honour of Professor M.S. Haq), 191-208.

[11] Hadi, A.S. (1992). A new measure of overall potential influence in linear regression, *Comput. Stat. Data. Anal.*, **14**, 1-27.

[12] Hadi, A.S. and Simonoff, J.S. (1993). Procedure for the identification of outliers in linear models, *J. Amer. Stat. Assoc.*, **88**, 1264-1272.

[13] Hadi, A.S., Imon, A.H.M.R, and Werner, M. (2008a). *Outliers Identification* in '*Wiley Interdisciplinary Reviews: Computational Statistics*', Wiley, New York (Submitted for publication).

[14] Hadi, A.S., Imon, A.H.M.R, and Werner, M. (2008b). *Identification of Outliers in Large Statistical Data*, Wiley, New York (Work in Progress).

[15] Haith, D.A. (1976). Land use and water quality in New York rivers, J. Env. Eng. Div., ASCE 102 (No. EEI. Proc. Paper 11902), 1-15.

[16] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Function*, Wiley, New York.

[17] Hawkins, D.M. (1994). The feasible solution algorithm for least trimmed squares regression, *Comput. Stat. Data Anal.*, **17**, 185-196.

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

[19] Henderson, H.V. and Searle, S.R. (1981). On deriving the inverse of a sum of matrices, *SIAM Rev.* **22**, 53-60.

[20] Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and ANOVA, *Am. Stat.*, **32**, 17-22.

[21] Huber, P.J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo, *Ann. Stat.*, **1**, 799-821.

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

[23] Imon, A.H.M.R. (2002). Identifying multiple high leverage points in linear regression, *J. Statist. Stud.* (Special Volume in Honour of Professor Mir Masoom Ali), 207-218.

[24] Imon, A.H.M.R. (2005). Identifying multiple influential observations in linear regression, *J. Appl. Stat.*, **32**, 929-946.

[25] Imon, A.H.M.R. (2008). Deletion residuals in the detection of heterogeneity of variances in linear regression, (Accepted for publication) *J. Appl. Stat.*

[26] Imon, A.H.M.R. and Alam, M.N. (2008). The effect of outliers in regression variable selection (Submitted for publication).

[27] Imon, A.H.M.R. and Khan, M.A.I. (2003). A solution to the problem of multicollinearity caused by the presence of multiple high leverage points, *Int. Jour. Stat. Sci.*, **2**, 37-50.

[28] Mallows, C.L. (1975). On some topics in robustness, Unpublished memorandum, Bell Telephone Laboratories, New Jersey.

[29] Maronna, R.A. and Yohai, V.J. (2000). Robust regression with both continuous and categorical predictors, *J. Stat. Plan. Inf.*, **89**, 197-214.

[30] Maronna, R.A. and Zamar, R.H. (2002). Robust estimates of location and dispersion for high-dimensional data sets, *Technometrics*, **44**,307-313.

[31] Maronna, R.A., Martin, R.D., and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*, Wiley, New York.

[32] Midi, H., Ramli, N., and Imon, A.H.M.R (2008). The performance of diagnostic–robust generalized potential approach for the identification of multiple high leverage points in linear regression, (Accepted for publication) *J. Appl. Stat.*

[33] Peña, D. and Prieto, F.J. (2001). Multivariate outlier detection and robust covariance estimation, *Technometrics*, **43**, 286-310.

[34] Rousseeuw, P.J. (1984). Least median of squares regression, *J. Amer. Stat. Assoc.*, **79**, 871-880.

[35] Rousseeuw, P.J. and Leroy, A. (1987). *Robust Regression and Outlier Detection*, Wiley, New York.

[36] Rousseeuw, P.J. and van Driessen, K. (2000). An algorithm for positive-breakdown regression based on concentration steps, In '*Data Analysis: Modeling and Practical Applications*,' Gaul, W., Opitz, O. and Schader, M. (eds.), 335-346, Springer, New York.

[37] Rousseeuw, P.J. and Yohai, V. (1984). Robust regression by means of S-estimators, In '*Robust and Nonlinear Time Series*', Franke, J., Hardle, W. and Martin, R.D. (eds.), Lecture Notes in Statistics **26**, 256-272, Springer, New York.

[38] Salibian-Barrera, M. and Yohai, V. (2005). A fast algorithm for S – regression estimates, *J. Comput. Graph. Stat.*, **15**, 1-14.

[39] Stigler, S.M. (1973). Simon Newcombe, Percy Daniell, and the history of robust estimation 1885-1920, *J. Amer. Stat. Assoc.*, **68**, 872-879.

[40] Vellman, P.F. and Welsch, R.E. (1981). Efficient computing of regression diagnostics, *Am. Stat.*, **35**, 234-242.

[41] Yohai, V. (1987). High breakdown-point and high efficiency estimates for regression, *Ann. Stat.*, **15**, 642-656.