Discussiones Mathematicae
Probability and Statistics 20 (2000) 25–50

AN ADAPTIVE METHOD OF ESTIMATION AND OUTLIER DETECTION IN REGRESSION APPLICABLE FOR SMALL TO MODERATE SAMPLE SIZES

Brenton R. Clarke

Mathematics and Statistics, Division of Science and Engineering
Murdoch University
Murdoch, W.A., 6150, Australia

Abstract

In small to moderate sample sizes it is important to make use of all the data when there are no outliers, for reasons of efficiency. It is equally important to guard against the possibility that there may be single or multiple outliers which can have disastrous effects on normal theory least squares estimation and inference. The purpose of this paper is to describe and illustrate the use of an adaptive regression estimation algorithm which can be used to highlight outliers, either single or multiple of varying number. The outliers can include 'bad' leverage points. Illustration is given of how 'good' leverage points are retained and 'bad' leverage points discarded. The adaptive regression estimator generalizes its high breakdown point adaptive location estimator counterpart and thus is expected to have high efficiency at the normal model. Simulations confirm this. On the other hand, examples demonstrate that the regression algorithm given highlights outliers and 'potential' outliers for closer scrutiny.

The algorithm is computer intensive for the reason that it is a global algorithm which is designed to highlight outliers automatically. This also obviates the problem of searching out "local minima" encountered by some algorithms designed as fast search methods. Instead the objective here is to assess all observations and subsets of observations with the intention of culling all outliers which can range up to as much as approximately half the data. It is assumed that the distributional form of the data less outliers is approximately normal. If this distributional assumption fails, plots can be used to indicate such failure, and, transformations may be ;required before potential outliers are deemed as outliers. A well known set of data illustrates this point.

Keywords: outlier; least median of squares regression; least trimmed squares; trimmed likelihood; adaptive estimation; leverage.

1991 Mathematics Subject Classification: 62T05.

1. Introduction

For fitting the linear model

$$Y_j = x_j^T \beta_0 + \epsilon_j, \qquad j = 1, \ldots, n,$$

the assumption on which much inference is based is that the ϵ_j are independent random variables with a normal cumulative distribution function, say $\Phi(./\sigma)$. Here Φ is the standard normal cumulative distribution and σ is a scale parameter. Also x_j and β_0 are p-dimensional vectors of covariates and regression coefficients. The assumption of normality is used for inference based on least squares estimates of β_0 and is also used in formulating many of the tests for outliers discussed in Cook and Weisberg (1982), and Barnett and Lewis (1994).

While least squares estimates are known to be efficient under the normal model they are greatly affected by the presence of outliers. The number or prevalence of outliers is not usually known to the data analyst prior to analysis and may constitute up to as much as approximately one half of the data. This motivated the introduction of the least median of squares or LMS estimation technique of Rousseeuw (1984) and Rouseeuw and Leroy (1987). But even LMS which is robust to such large contamination of the data can only be regarded as an 'exploratory tool' for identification of outliers in the view of Atkinson (1986a) and Fung (1993). Hettmansperger and Sheather (1992) in fact criticise LMS for its slow rate of convergence to a non-normal asymptotic distribution and also possible numerical instability. On the other hand Atkinson (1994) perseveres with the use of LMS in the case of regression, and the Minimum Volume Ellipsoid in the case of multivariate estimation, to identify multiple outliers using his 'stalactite' plots.

In order to obtain more efficient estimates a common suggestion has been to use a one-step M-estimator or as recently investigated by Simpson, Ruppert and Carroll (1992) a one-step GM-estimator. These estimators retain the high-breakdown point of their initial estimators, such as LMS. However their motivation for use is essentially asymptotic. This is also the motivation behind such methods as Rousseeuw and Yohai's (1984) S-estimators, see also Davies (1990), Ruppert (1992), and the MM, and

tau estimators of Yohai (1987) and Yohai and Zamar (1988). Performances in small to moderate sample sizes rely on the hope that the asymptotics will somehow still be operating for finite sample sizes n, albeit that there may be points of influence, for example caused by outlying x observations.

The alternative approach to using high breakdown estimators which is frequently employed in finite sample estimation is the use of 'hands on' techniques for the detection of multiple outliers, such as introduced by Kianifard and Swallow (1989, 1990), Swallow and Kianifard (1996) and also recently by Hadi and Simonoff (1993). The approach of Kianifard and Swallow as observed by the latter authors, involves first ordering the observations by a single case diagnostic, then using the recursive residuals of Brown, Durbin, and Evans (1975), based on the p observations with smallest diagnostic, to identify outlying observations. The fact that single-case diagnostics are susceptible to masking, where the presence of an outlier goes undetected because of another or perhaps other multiple outliers, means that a case with smallest diagnostic can be a true outlier which would then lead to incorrect identifications from the recursive residuals. On the other hand the proposed procedure of Hadi and Simonoff (1993) requires as its starting point an initial clean subset M initially of size h = integer part of $(n+p-1)/2 = [(n+p-1)/2]$. Interestingly the problem of finding approximately one half of the observations that are outlier free is in essence the motivation behind LMS, for instance, in the case of simple linear regression it corresponds to the narrowest strip covering half of the observations. Hadi and Simonoff (1993) offer two proposed methods to find an initial clean subset, which in itself should lead to approximately 50% breakdown point estimators, but as they readily admit the breakdown points of their overall methods is an open problem.

Rousseeuw (1984) also proposed an approximately 50% breakdown estimator which he called least trimmed squares, LTS, and which was also described independently by Butler (1982). This estimator minimizes

$$\sum_{i=1}^{h} r_{i:n}^2, \tag{1.1}$$

where $r_{1:n}^2 \leq r_{2:n}^2 \leq \cdots \leq r_{n:n}^2$ are the ordered squared residuals given by $r_j = Y_j - x_j\beta$. Rousseeuw and Leroy (1987, pp. 124 and 132) choose h to be [n/2]+[(p+1)/2] in order to maximize the breakdown point. Bednarski and Clarke (1993) arrived at this estimator from a more general form, treating the class of trimmed likelihood estimators and discussing a compact related

differentiability of the resulting estimators of location and scale. The proportion of trimming, $\alpha = g/n$ where $g = n - h$, is allowed to vary in Clarke (1994) and is chosen to minimize the estimated asymptotic variance of the estimator. In this way possible multiple outliers are identified and confidence intervals which have robustness of efficiency and robustness of validity even in relatively small samples are given. The estimators and consequent procedures were shown to have an approximate breakdown point of 50%. In this paper we describe the algorithm for the regression setting which identifies outliers of varying number and investigate its performance on several data sets as well as by simulation. The method has high power when the proportion of outliers is small as evidenced by Monte Carlo studies in the ilk of those of Kianifard and Swallow (1989). When dealing with larger proportions of contamination the outliers are easily identified if they appear as symmetric contamination, but if they appear as asymmetric contamination the degree to which the aberrant observations are outlying determines whether the whole subset of outliers is identified. This is particularly so for high leverage outliers.

The LTS estimator converges to a normal distribution, but has only 7.1% asymptotic efficiency at the normal model. However, allowing the trimming proportion to vary adaptively, Clarke (1994) demonstrates through simulations that the resulting adaptive trimmed likelihood estimators exhibit high efficiency, indeed for sample sizes of n=10 or n=20 the average trimming proportion over all samples in the simulations from normal data was 0.0069 or 0.0006, respectively. Moreover, the adaptive estimates are robust to departures from the normal model. The basis of the adaptive trimmed likelihood estimator relies on choosing to trim $0 \leq g \leq G^*(n)$ to minimize the estimated asymptotic variance of the location estimator. This is an idea synonomous with adaptive trimmed mean estimates of Tukey and McLaughlin (1963) and Jaeckel (1971), though those authors never suggested that their adaptive trimmed mean approach could be used for identifying outliers, perhaps because as demonstrated in Clarke (1994) this approach to outlier detection does not work for the trimmed mean, whereas it is completely successful for the trimmed likelihood mean estimator. The choice of an upper bound on the proportion of outliers possibly present, $G^*(n)$, is governed by three possibilities: the computation time required to compute the adaptive trimmed likelihood estimator; the breakdown bound for the adaptive trimmed likelihood estimator; or the users prior knowledge about a value of $G^*(n)$.

The regression version of the adaptive trimmed likelihood algorithm which we shall denote by **ATLA** supplies simultaneously the regression estimates and outliers, where the regression estimates are exactly those of the least squares estimates based on the sample with the outliers removed. This obviates successive analyses using least squares that may be required because of masking (Atkinson, 1986b). While an automatic procedure that we give via the algorithm of Section 2 is proposed it should be understood that the outliers are identified subject to the normal model. The algorithm also provides an order of importance by which observations or groups of observations may be classed to be potentially outlying. Plots of estimated variances and or regression estimates for different proportions of trimming using trimmed likelihood estimation can highlight potential non-normality and observations that are potential outliers worthy of further consideration.

If it were known that there were g outliers, typical outlier identification techniques involve computation of $N = \binom{n}{g}$ least squares estimates. **ATLA** sets $G^*(n)$ to be an upper bound on the number of outliers whence computation can be intensive. Apart from computational considerations the maximum value of $G^*(n)$ is chosen in relation to the breakdown bound. It is shown in Clarke (1994) that the location adaptive trimmed likelihood estimator inherits the maximum breakdown bound for LTS when 'trimming' $G^*(n) = [n/2]$ observations. Hence in analogy, the maximum value for $G^*(n)$ to be chosen in regard to **ATLA** in regression is

$$G^*(n) = n - [n/2] - [(p+1)/2], \tag{1.2}$$

since 'trimming' this number of observations in LTS gives the maximum breakdown bound (Rousseeuw and Leroy, 1987, pp. 124 and 132). Clearly, computing N least squares estimates when g is given by this value of $G^*(n)$ is only feasible for small n. But the situation for small n is that an efficient estimator is needed, such as provided by **ATLA**. Bigger sample sizes lead to natural computation time bounds on the upper proportion of outliers that can be detected by **ATLA**, but these bounds increase with advances in computing power.

In Section 3 we analyse the 'modified wood specific gravity' data of Rousseeuw (1984). This data was used to motivate LMS. **ATLA** yields exactly the outliers known to be the contamination. Another set of outliers constituting a larger proportion of the data is identified from the telephone data used to introduce the S-estimator of Rousseeuw and Yohai (1984). A major data set in the statistical literature is the Brownlee stack loss

data examined by many authors including Atkinson (1986a), Chambers and Heathcote (1981) and others mentioned in Hampel *et al.* (1986). Here the algorithm is useful for identifying the potential outliers but the suggestion is that the data apart from the 4 outliers do not support the assumption of normality. As an example where $G^*(n)$ must be reduced because of the computation time, the Scottish Hill Races data of Atkinson (1986b) is analysed. The outliers are identified nevertheless by **ATLA**. To illustrate the usefulness of **ATLA**, simulations of a simple linear regression model for a sample size $n = 15$ are carried out with both low leverage outliers and high leverage outliers that range in number from one up to $G^*(n)$ equal to seven. **ATLA** is successful in identifying the outliers and moreover trims few data samples when there are no outliers. These analyses are reported in Section 4. Analyses of Sections 3 and 4 follow the preliminary definitions in Chapter 2 which give the regression algorithm **ATLA**.

2. Definitions and the regression estimation algorithm

For a parametric family of densities $\{ f_\theta : \theta \in \Theta\}$ Bednarski and Clarke (1993) introduced the idea of estimating functionals $T[.]$, defined on the space of distribution functions, G, that were obtained through a trimmed likelihood principle: to trim a proportion $0 \leq \alpha < 1$ observations which are least likely to occur as indicated by the likelihood. Specifically, define a functional on the product space $G \times \Theta$:

$$S(F, \theta) = \int \log f_\theta(x) J[F\{y : \log f_\theta(y) \leq \log f_\theta(x)\}] dF(x),$$

where

$$J(t) = \begin{cases} 0 & \text{if } t \leq \alpha \\ 1 & \text{if } \alpha < t \leq 1 \end{cases}.$$

Then define the estimating functional T at F as

$$T[F] = \arg_\Theta \max_\Theta \{S(F, \theta)\}.$$

When the parametric family is normal and $\theta = (\mu, \sigma)$, where μ is a location parameter and F_n is the empirical distribution function formed from a sample $Y_1, ..., Y_n$, where $Y_i = \mu_0 + \epsilon_i$, then

$$T[F_n]$$
$$= \arg_\theta \max_\theta \left\{ \int \left[-\frac{1}{2} ln\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2} \right] J[F_n\{y : (y-\mu)^2 \geq (x-\mu)^2\}] dF_n(x) \right\}.$$

In practice, if $r_i(\mu) = Y_i - \mu$ and $r^2_{1:n} \leq r^2_{2:n} \leq ... \leq r^2_{n:n}$ are the ordered squared residuals, trimming exactly $\alpha = g/n$ observations in the likelihood is equivalent to choosing the location estimator to satisfy

$$\hat{\mu}(g) = \arg_\mu \quad \min_\mu \sum_{i=1}^{h} r^2_{i:n}, \tag{2.1}$$

where $h = n - g$. The variance estimator is

$$\overline{\sigma}^2(g) = \frac{1}{h} \sum_{i=1}^{h} r^2(\hat{\mu})_{i:n} \tag{2.2}$$

(Bednarski and Clarke, 1993, Clarke, 1994). The estimator for variance is asymptotically biased. For a fixed α a consistent estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{(1-\alpha)\overline{\sigma}^2}{1 - \alpha - \sqrt{\frac{2}{\pi}} z_{\alpha/2} \exp(-\frac{1}{2} z^2_{\alpha/2})}, \tag{2.3}$$

where $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. When $h = [\frac{n}{2}] + 1$ equation (2.1) defines the LTS estimator of Rousseeuw (1984). For fixed α the asymptotic distribution of $\sqrt{n}(\hat{\mu} - \mu_0)$ is normal with asymptotic variance

$$\mathrm{var}(\sigma^2, \alpha) = \frac{\sigma^2}{\{1 - \alpha - \sqrt{\frac{2}{\pi}} z_{\alpha/2} \exp(-\frac{1}{2} z^2_{\alpha/2})\}^2}. \tag{2.4}$$

In Clarke (1994) it is suggested to choose the proportion of trimming adaptively to minimize $var(\sigma^2, \alpha)$ using appropriate estimates of σ^2 based on trimmed samples. Note that if in fact there are g outliers , one would expect the estimator $\overline{\sigma}^2(g)$ to be the best estimate of σ^2, rather than $\hat{\sigma}^2(g)$ which would be slightly inflated. Denoting

$$V_n(g) = \mathrm{var}(\overline{\sigma}^2(g), g/n), \tag{2.5}$$

the adaptive trimmed likelihood estimator is then $\hat{\mu}_{ATL} = \hat{\mu}(\tilde{g})$, where $\tilde{g}$ satisfies

$$V_n(\tilde{g}) = \min_{0 \leq g \leq G^*(n)} V_n(g). \tag{2.6}$$

Empirically the location estimator of this adaptive approach has high efficiency at the normal model as described by option (4) of Clarke (1994).

Generalizing to regression we get

$$\hat{\beta}(g) = \arg_\beta \quad \min_\beta \quad \sum_{i=1}^{h} (r^2(\beta))_{i:n}, \tag{2.7}$$

where $r_i(\beta) = Y_i - x_i^T\beta$, and the analogous estimator for variance adjusted to be "unbiased" is then

$$\tilde{\sigma}^2(g) = \frac{1}{h-p}\sum_{i=1}^{h}(r^2(\hat{\beta}))_{i:n}.$$

An algorithm to solve for $\hat{\beta}(g)$ is given here using adaption of notation from Ruppert (1992).

Given g , the observations 'trimmed', in the sense of trimming the likelihood, are those observations corresponding to r_i's not found in the sum (2.7). These are potential outliers. The algorithm which the author calls **ATLA** follows.

Define J_l to be the l'th sample $\{j_{1l}, ..., j_{gl}\}$ from the N subsamples chosen from the first n positive integers. Consider the following algorithm:

1. Initialize $S^2_{0\alpha} = \infty, \quad l = 1$ and $\tilde{J} = \{1, ..., g\}$.

REPEAT WHILE $l \leq N$.

2. Let β_l solve $Y_i = x_i^T\beta_l$ for $i \notin J_l$, (fit by least squares).

3. Let $S_l^2 = \sum_{i=1}^{h}(r^2(\beta_l))_{i:n}$.

4. If $S_l^2 < S^2_{0\alpha}$, then do:

(i) $\tilde{\beta} \leftarrow \beta_l$,

(ii) $S^2_{0\alpha} \leftarrow S_l$,

(iii) $\tilde{J} \leftarrow J_l$.

5. $l \leftarrow l + 1$.

On running the above algorithm for fixed g then $\hat{\beta}(g) = \tilde{\beta}, \quad \tilde{\sigma}^2(g) = S^2_{0\alpha}/(h-p)$ and the set of observation numbers, $\tilde{J}$, is effectively a function identifying g potential outliers; call it $\tilde{J}(g)$. To define the ATL estimator let $V_n(g) = var(\tilde{\sigma}^2(g), g/n)$ and choose $\tilde{g}$ to satisfy (2.6). Then

$$\hat{\beta}_{ATL} = \hat{\beta}(\tilde{g}) \quad \text{and} \quad \hat{\sigma}^2_{ATL} = \tilde{\sigma}^2(\tilde{g}).$$

The choice for $G^*(n)$ is (1.2) or something smaller chosen by the statistician.

The outliers identified by **ATLA** are then given by $\tilde{J}(\tilde{g})$. Plots illustrating the stability or otherwise of the trimmed likelihood estimates can be made for $\hat{\sigma}^2(g)$ for different g. Here $\hat{\sigma}^2(g)$ is given by (2.3) with $\alpha = g/n$ and $\overline{\sigma}^2(g)$ replaced by $\tilde{\sigma}^2(g)$. Storing in tables $\tilde{J}(g)$ for different g highlights the order of importance of potential outliers. Ryan (1995) has highlighted the idea of plotting $\tilde{J}(g)$ for different g. **ATLA** goes beyond listing simply the $\tilde{J}(g)$ in actually identifying the outliers $\tilde{J}(\tilde{g})$ specifically.

3. Examples of ATLA analysis

Example 3.1. One influential data point.
The technique of using **ATLA** is illustrated in this example with one point of influence, corresponding to a high leverage point. Cook and Weisberg (1982) discuss influential points in terms of leverage and more recently Rousseeuw and Van Zomeron (1990) advocate a display to clasify data into regular observations, vertical outliers, good leverage points, and bad leverage points. Generating seven data points using a model $Y_j = 1 + x_j + \epsilon_j$; $j = 1, ..., 7$, where the ϵ_j are then obtained by drawing from a standard normal distribution, and choosing the $\{x_j\}$ to give one clear leverage point, gave values $\{(x_j, Y_j)\}$ to be $(0, 1.61), (1, 1.54), (2, 2.81), (3, 5.2), (4, 5.74), (5, 7.93)$ and $(20, 20.95)$. Clearly $x_7 = 20$ yields a large leverage point, (x_7, Y_7) being a good point of influence as ϵ_7 was generated from the normal distribution. The simple linear model $Y = \beta_0 + \beta_1 x + \epsilon$ is fitted using **ATLA** yielding Table 3.1.1.

Table 3.1.1. Example with one 'good point' of influence.

	g	$V_n(g)$	$\tilde{J}(g)$	$\hat{\beta}(g)$	
$\tilde{g} =$	0	0.844	-	1.62	0.98
	1	2.348	6	1.39	0.98
	2	3.861	1,7	-0.07	1.57
	3	9.957	1,4,7	-0.21	1.57

The good point of influence is retained with $\tilde{g} = 0$, yielding an efficient estimate under the model.

If in fact the leverage point was a 'bad' point of influence given by $(x_7, Y_7) = (20, -14)$ so that $\epsilon_7 = -35.0$, an unlikely value to be drawn from the standard normal distribution, then we obtain the analysis of Table 3.1.2.

Figure 1. Least squares and ATLA estimates for data with one "good" point of influence.

Figure 2. Least squares fitted line — and ATLA fitted line - - - for data with one bad point of influence ×.

Table 3.1.2. Example with one 'bad' point of influence.

	g	$V_n(g)$	$\tilde{J}$	$\hat{\beta}(g)$	
	0	93.94	-	5.99	-0.89
$\tilde{g} =$	1	1.97	7	0.81	1.33
	2	3.86	1,7	-0.07	1.57
	3	9.96	1,4,7	-0.21	1.57

Figures 1 and 2 illustrate the difference between least squares fitted lines when fitting the respective data sets. The least squares and **ATLA** fits agree when there is one 'good' point of influence. The **ATLA** fit on the other hand is equivalent to the least squares fit of the data with the 'bad' point of influence excluded in the second analysis. ■

Example 3.2. Modified data of wood specific gravity.
Rousseeuw (1984) motivated LMS with an example containing multidimensional real data. Table 2 of that paper gives 20 observations for which there are five independent variables and an intercept. The data were formulated by replacing 4 observations from a data set given by Draper and Smith (1967, p. 227) by outliers. Observations 4,6,8, and 19 are identified by LMS as being outlying. These observations do not appear to be obvious outliers from the least squares analysis. Table 3.2.1 demonstrates how **ATLA** picked out the 4 outliers. Here (1.2) gives $G^*(20) = 7$. When $g = 7$ there are N=77,520 least squares estimates to be calculated to evaluate $\hat{\beta}(7)$. This took the MATLAB algorithm 1,735 cpu seconds to run on a Sun Sparcstation.

Table 3.2.1. Analysis on modified data on wood specific gravity from Rousseeuw (1984, p. 875).

	g	$V_n(g) \times 10^4$	$\tilde{J}(g)$
	0	5.8	-
	1	7.1	11
	2	9.0	3,11
	3	10.7	7,11,14
$\tilde{g} =$	4	4.5	4,6,8,19
	5	4.7	4,5,6,8,19
	6	5.9	4,5,6,8,12,19
	7	5.9	1,4,5,6,7,8,19

With $\tilde{g} = 4$ exactly those four observations previously identified above are given by $\tilde{J}(\tilde{g})$. The resulting **ATLA** estimator corresponds to the least squares estimator with these observations deleted from the sample, thus giving the more efficient estimated regression equation $\hat{y} = x^T \hat{\beta}_{ATL}$ given by

$$\hat{y} = 0.2174x_1 - 0.0850x_2 - 0.5643x_3 - 0.4003x_4 + 0.6074x_5 + 0.3773,$$

where $x_1, ..., x_5$ represent the five independent variables. It is more efficient than LTS or LMS. ■

Example 3.3 . Telephone data of the belgian statistical survey. Rousseeuw and Yohai (1984) introduce S-estimators which have breakdown point one half and illustrate their estimator in an example with a large fraction of outliers. The data given in Table 1 of that paper consists of 24 observations from 1950 to 1973 with the dependent variable being the total number of international calls made, and the independent variable is the year. The data contains heavy contamination and observations 14 to 21 or years 1963 to 1970 are spurious, since in fact according to Rousseeuw and Yohai, in this period another recording system was used, which only gave the total number of minutes of these calls. Table 3.3.1 gives an analysis using **ATLA**. Since for these data it was observed that $\tilde{J}(g) \supset \tilde{J}(g-1), \quad g = 1, ..., 11$, only the observation in $\tilde{J}(g) \backslash \tilde{J}(g-1)$ is reported. Formula (1.2) gives $G^*(24)$=11.

Table 3.3.1. Analysis of telephone data.

g	$V_n(g)$	$\tilde{J}(g) \backslash \tilde{J}(g-1)$		g	$V_n(g)$	$\tilde{J}(g) \backslash \tilde{J}(g-1)$
0	31.61	-		6	2.53	15
1	44.03	20		7	0.42	21
2	52.95	19	$\tilde{g} =$	8	0.28	14
3	59.37	18		9	0.33	1
4	59.33	17		10	0.37	22
5	48.96	16		11	0.37	2

The ATL estimator corresponds to calculating the least squares estimator on the data set with observations in $\tilde{J}(8)$ deleted. Here $\tilde{J}(8)$ corresponds

exactly to observations 14 to 21 and the estimated regression line is $\hat{y} = 0.1085x - 5.16$, not much different from the S-estimate of Rousseeuw and Yohai of $\hat{y} = 0.1121x - 5.42$. ■

Example 3.4 . Stack loss data.
The **ATLA** analysis of the stack loss data of Brownlee (1965, p. 454) illustrates that the algorithm should not be applied routinely in all situations. The data are 21 observations on losses of ammonia from an oxidation plant and there are three independent variables and a constant. Several authors mentioned in the introduction arrive at the conclusion that there are four suspected outliers, corresponding to observations 1, 3, 4 and 21. Curiously in the **ATLA** investigation of the data it was observed that as for the telephone data $\tilde{J}(g) \supset \tilde{J}(g-1)$ for $g = 1, 8$. Since $G^*(21) = 9$ in formula (1.2) proved out of bounds for computing $\hat{\beta}(9)$ because of computational time needed, it was assumed that $\tilde{J}(9) \supset \tilde{J}(8)$ whence an obvious adjustment to the algorithm allowed easy computation of $\hat{\beta}(9)$ under that assumption.

Table 3.4.1. Analysis of Brownlee stack loss data.

g	$V_n(g)$	$\tilde{J}(g)\backslash\tilde{J}(g-1)$	g	$V_n(g)$	$\tilde{J}(g)\backslash\tilde{J}(g-1)$
0	10.52	-	5	12.27	13
1	12.38	21	6	15.90	20
2	12.11	4	7	18.94	2
3	14.84	3	8	15.97	14
4	11.71	1	9*	17.00	8

* Entry for $g = 9$ calculated assuming $\tilde{J}(9) \supset \tilde{J}(8)$.

Importantly, the first four observations given by the algorithm as potential outliers were observations 21,4,3 and 1. However, routine **ATLA** analysis of the data in Table 3.4.1 yields $\tilde{g} = 0$, giving zero observations trimmed, whereupon **ATLA** and least squares agree. The objective function minimized has another local minima $V_{21}(4)$ not much different to $V_{21}(0)$. $\tilde{J}(4)$ indicates the aforementioned alleged outliers. Continuing with further investigation of the variance estimates given by (2.3), there is a suggestion that the data may not be normal. Estimates $\hat{\sigma}^2(g)$ of σ^2 are obtained under the assumption of normality and one expects that, apart from small sample biases, estimates $\hat{\sigma}^2(g)$ would converge to σ^2 fairly uniformly over the range

$0 \leq g \leq 9$. But plotting $\hat{\sigma}^2(g)$ along with the simulated expected values of $\hat{\sigma}^2(g)$ when the variances are estimated from normal data of size $n = 21$ generated with variance $\sigma^2 = 10.52$ gives the plots in Figure 3. Clearly the departure from the expected variances is an indication of non-normality.

Figure 3. Plot of $\hat{\sigma}^2(g)$ for Brownlee stack loss data. Also plotted is the expected value of $\hat{\sigma}^2(g)$ for data generated from a normal sample with variance $\sigma^2 = 10.52$ as obtained through simulation. This is denoted by —-.

This evidence supports observations by Atkinson (1981, 1986) and Chambers and Heathcote (1981) that the data are not normally distributed, and should be transformed. It also highlights the observations which do not appear to agree with the bulk of the data when a transformation is not used. ■

Example 3.5 . Scottish hill races data.
Atkinson (1986b) originally reported and analysed this data set. A more recent analysis is given in Venables and Ripley (1994). The dependent variable is the record time in minutes and the independent variables are distance in miles and climb feet. There are 35 observations whence a bound on computation time of **ATLA** required $G^*(35) = 8$, whereas formula (1.2) yields $G^*(35) = 16$. Atkinson's analysis illustrates how normal plots of studentized

residuals reveal observations 7 and 18 as outliers. When these two observations are deleted, then an additional plot reveals that observation 33 is an outlier, the point being that observations 7 and 18 mask the presence of an outlier in observation 33. LMS in fact identifies observations 7,18,11,33 and 35 as having the largest absolute residuals and worthy of further attention, but Atkinson finds observations 11 and 35 agree with the bulk of the data.

Figure 4. Plot of $\hat{\sigma}^2(g)$ for Scottish hill races data.

In a single analysis using **ATLA** the observations 7,18 and 33 are identified as outlying in Table 3.5.1.

Table 3.5.1. Analysis of Scottish hill races data.

	g	$V_n(g)$	$\tilde{J}(g)\backslash\tilde{J}(g-1)$	g	$V_n(g)$	$\tilde{J}(g)\backslash\tilde{J}(g-1)$
	0	217.9	-	5	89.8	6
	1	124.3	18	6	103.1	8
	2	82.8	7	7	113.9	14
$\tilde{g} =$	3	69.8	33	8	130.0	30
	4	78.2	19			

For these data the assumption of normality appears to be supported for observations other than the identified $\tilde{g} = 3$ outliers. A plot of $\hat{\sigma}^2(g)$ against g indicates that estimates of σ^2 stabilize for $g \geq 3$ as one would hope if indeed the data were normal. The plot in Figure 4 is extended up to $G^*(n) = 16$ by assuming $\tilde{J}(g) \supset \tilde{J}(g-1)$ for $g = 9, ..., 16$. This allowed for easy calculation beyond the computed $\tilde{J}(8)$ which was arrived at by **ATLA**. ■

4. Monte carlo performance

In this section we report results of Monte Carlo experiments to examine the size and power of the method discussed in this article. All simulations were carried out using MATLAB version 4.2c on a Sun Sparc10 using an operating system SunOS 4.x . The study is limited by the need to keep the computing time for **ATLA** within bounds. For this reason the sample size is set to $n = 15$ and p is set to 2 (simple regression). As an indication of the computing time needed, in carrying out **ATLA** on 1,000 different simulated samples generated from a normal regression model, the computing time could be as much as 70 hours. In this sense the study is extensive. Here we report similar experiments to Kianifard and Swallow (1989, 1990) and Hadi and Simonoff (1993).

The data sets for both the null and alternative cases are generated from the model

$$y_j = \beta_0 + \beta_1 x_j + \epsilon_j, \qquad j = K+1, ..., 15 \quad ,$$

where $x_j \sim U(0, 15)$ for $j = K+1, ..., 15$ and where $\beta_0 = 1$ and $\beta_1 = 1$. Values of K are chosen for each experiment between 1 and $G^*(15) = 7$. Observations for $j = 1, ..., K$ are generated via

$$y_j = \beta_0 + \beta_1 x_j + c + \epsilon_j \quad ,$$

where to investigate low leverage outliers values $x_j = 8.5 - 0.25(j-1)$ for $j = 1, ..., K$ are chosen, and to investigate high leverage outliers values $x_j = 20 - 0.25(j-1)$ for $j = 1, ..., K$ are chosen respectively. For all $j = 1, ..., 15$ the ϵ_j are random standard normal errors. Values of c are chosen to either reflect the normal regression model, viz., $c = 0$, or potential outliers, where $c = -4$ or less.

Simulations for the null regression model with $c = 0$ were carried out for both low leverage outliers and for high leverage outliers. For each K 1,000 samples were generated. For each simulation the number of samples trimmed by **ATLA** was fewer than 10. This suggests the size of the ATL

outlier detection procedure is less than 0.01, much smaller than either of the procedures of Hadi and Simonoff (1993) which require as an inbuilt feature of their respective methods a size of $'\alpha = 0.05'$.

To investigate the power of **ATLA** to identify outliers, the value of c was allowed to range from -4 to -13. However, in all cases the power increased as c decreased until a plateau was reached. This was apart from minor variation due to sampling error. Rather than report power for all values of c only values of c are reported up until a plateau is reached.

Simulation results are summarized in Table 4.1 for low leverage points and Table 4.2 for high leverage points. We report the frequency with which exactly the first K observations are identified as outliers, p_1, the frequency with which at least one outlier is identified, p_2, the frequency with which there is false identification, p_3, and the frequency with which at least all the K outliers are trimmed, p_4. In addition the overall number of samples trimmed by **ATLA** is recorded as p_5. All of the simulations for values of c different from zero and each K involved 100 samples.

To investigate the nature of local and global minima of the objective function $V_n(g)$ the frequency with which these occurred at each value of $g = 0, 1, ..., 7$ was recorded. The local minima are indicated by LM and the global minima by GM.

Table 4.1. Low leverage points.

$K = 1$

c	p_1	p_2	p_3	p_4	p_5	$g =$	0	1	2	3	4	5	6	7
0	0	1	6	1	6	GM	994	2	2	0	1	0	0	1
						LM	998	2	11	26	35	55	85	135
-4	95	100	5	100	100	GM	0	95	1	0	1	1	1	1
						LM	0	98	1	2	3	7	6	10

$K = 2$

0	0	2	6	0	6	GM	994	3	0	0	0	2	1	0
						LM	997	3	12	18	36	36	78	135
-4	89	100	11	100	100	GM	0	0	89	3	2	1	2	3
						LM	100	0	97	3	4	7	8	14
-5	97	100	3	100	100	GM	0	0	97	3	0	0	0	0
						LM	100	0	97	3	1	5	5	11

$K = 3$

0	0	3	9	0	10	GM	990	5	1	2	1	1	0	0
						LM	994	6	3	20	34	48	73	153
-4	89	100	11	100	100	GM	0	0	0	89	6	1	1	3
						LM	100	0	0	92	7	2	5	13
-5	91	100	9	100	100	GM	0	0	0	91	4	1	3	1
						LM	100	0	0	96	4	1	7	10

$K = 4$

0	0	2	4	0	4	GM	996	2	0	0	1	0	1	0
						LM	998	2	13	18	33	50	66	145
-4	92	100	8	100	100	GM	0	0	0	0	92	4	2	2
						LM	100	0	0	0	94	4	8	14
-5	91	100	9	100	100	GM	0	0	0	0	91	4	3	2
						LM	100	0	1	0	95	4	5	12

$K = 5$

0	0	6	4	0	6	GM	994	2	0	0	1	2	0	1
						LM	998	2	11	22	29	44	73	141
-4	75	90	16	90	91	GM	0	0	0	0	0	84	11	5
						LM	100	0	1	1	0	89	11	10
-5	84	100	16	100	100	GM	0	0	0	0	0	84	11	5
						LM	100	0	1	1	0	89	11	10

$K = 6$

0	5	5	0	0	5	GM	993	3	0	1	1	0	1	1
						LM	997	3	12	13	39	52	88	132
-4	26	39	21	37	47	GM	53	0	1	0	2	2	26	16
						LM	100	0	3	6	22	4	72	21
-5	63	78	18	77	81	GM	19	0	1	1	0	1	63	15
						LM	100	0	7	13	12	1	82	16
-6	81	92	14	92	95	GM	5	0	0	0	2	1	81	11
						LM	100	0	7	13	12	1	82	16
-12	90	100	10	100	100	GM	0	0	0	0	0	0	90	10
						LM	100	0	2	14	12	0	90	10

$K = 7$

0	0	1	4	0	4	GM	996	1	1	0	1	0	1	0
						LM	999	1	16	23	29	48	86	139
-4	7	9	19	7	26	GM	74	1	3	3	5	3	2	9
						LM	99	1	6	11	29	26	12	66
-5	15	20	18	15	33	GM	67	0	1	4	4	3	3	18
						LM	99	1	8	16	23	23	11	78
-6	37	40	20	37	57	GM	43	1	0	4	6	3	4	39
						LM	99	1	4	18	27	29	7	90
-9	91	91	6	91	97	GM	3	1	0	1	1	3	0	91
						LM	99	1	7	15	26	31	0	99

It is a feature of the results that for low leverage outliers and $K = 1, ..., 4$ that the power of the **ATLA** in identifying exactly the K outliers, as illustrated by values of p_1, is remarkably high in the order of 90% when $c = -4$. The fact that in this case 100% of samples have a subsample that contains the K outliers which is trimmed by **ATLA** as indicated by values of p_4 demonstrates that the algorithm tends to at least cull the outliers in these situations. For $K = 5$, which indicates that a significant proportion of the data are outliers, the power remains high, with $p_1 = 75\%$ and $p_4 = 90\%$ for $c = -4$, while for $c = -5$ a value $p_4 = 100\%$ is observed. For this asymmetric contamination the value of c has to increase in magnitude for the power to be retained when $K = 6$ and $K = 7$, but nevertheless **ATLA** can identify large outlier sets in this case.

Table 4.2. High leverage points.

$K = 1$

c	p_1	p_2	p_3	p_4	p_5	$g =$	0	1	2	3	4	5	6	7
0	0	1	7	1	7	GM	993	4	1	0	1	1	0	0
						LM	996	4	8	24	41	45	75	148
-4	95	100	5	100	100	GM	0	95	1	0	1	1	1	1
						LM	0	98	1	2	3	7	6	10

$K = 2$

0	0	2	3	1	4	GM	996	3	0	0	0	0	1	0
						LM	997	3	8	23	34	40	91	127
-4	89	100	11	100	100	GM	0	0	89	3	2	1	2	3
						LM	100	0	97	3	4	7	8	14
-5	97	100	3	100	100	GM	0	0	97	3	0	0	0	0
						LM	100	0	97	3	1	5	5	11

$K = 3$

0	0	3	5	0	7	GM	993	4	1	1	0	1	0	0
						LM	996	4	8	21	31	54	71	126
-4	82	91	11	91	93	GM	7	1	0	82	5	1	1	3
						LM	99	1	0	92	6	3	6	14
-5	89	98	9	98	98	GM	2	0	0	89	4	1	3	1
						LM	99	1	0	92	6	3	6	14
-6	93	100	7	100	100	GM	0	0	0	96	4	1	7	10
						LM	100	0	0	96	4	1	7	10

$K = 4$

0	0	2	2	0	2	GM	998	0	0	1	0	1	0	0
						LM	1000	0	7	14	36	50	75	137
-4	39	46	8	46	47	GM	53	0	0	0	39	4	2	2
						LM	100	0	2	0	93	4	9	13
-5	69	78	12	78	81	GM	19	0	0	0	70	4	4	3
						LM	99	1	4	0	94	4	6	13
-6	85	98	13	98	98	GM	2	0	0	0	85	5	5	3
						LM	99	1	2	0	93	6	6	13
-7	90	99	9	99	99	GM	1	0	0	0	90	6	2	1
						LM	100	0	1	0	94	6	5	12

$K = 5$

0	0	3	1	0	3	GM	997	2	0	0	0	0	1	0
						LM	998	2	8	18	43	48	79	129
-4	5	10	9	9	14	GM	86	1	1	0	0	5	6	7
						LM	99	1	8	11	5	68	13	17
-5	25	36	17	34	42	GM	58	1	1	2	0	25	7	6
						LM	96	2	7	11	3	80	15	13
-6	45	55	17	55	62	GM	38	0	0	1	2	46	6	7
						LM	100	0	4	9	3	88	6	14
-7	62	74	18	73	80	GM	20	2	0	1	0	64	8	5
						LM	98	2	5	5	1	88	10	14
-8	68	85	24	85	92	GM	8	3	1	0	1	69	11	7
						LM	98	3	11	6	1	85	11	14
-9	80	95	17	95	97	GM	3	0	0	1	0	81	7	8
						LM	99	1	6	10	0	90	7	18

$K = 6$

0	0	3	2	0	4	GM	996	1	1	1	1	0	0	0
						LM	999	1	11	25	39	47	76	132
-4	0	3	12	1	12	GM	88	1	3	1	2	0	1	4
						LM	99	1	9	16	21	19	29	21
-6	8	14	13	9	21	GM	79	2	1	0	2	1	9	0
						LM	98	2	19	12	24	15	53	20
-8	25	39	27	33	52	GM	48	2	0	1	3	1	32	13
						LM	97	3	4	19	19	6	65	22
-10	50	66	28	66	78	GM	22	1	1	1	3	3	53	16
						LM	97	3	10	18	21	6	69	20
-12	76	85	17	85	93	GM	7	0	0	0	1	2	79	11
						LM	99	1	5	22	17	2	86	11

$K = 7$

0	0	7	4	0	7	GM	993	2	2	0	1	1	0	1
						LM	998	2	16	23	44	45	69	140
-4	0	3	24	0	24	GM	76	5	4	5	4	4	0	2
						LM	93	5	11	17	19	18	5	21
-8	5	12	45	5	50	GM	50	4	3	6	9	10	10	8
						LM	91	8	10	19	28	27	21	41
-12	24	26	41	24	65	GM	35	2	5	4	10	10	7	27
						LM	90	5	17	20	27	39	13	59

ATLA has the ability to trim either low or high leverage outliers. For high leverage outliers with $K = 1, 2, 3$ values of p_4 remain over 90% when $c = -4$. For increasing magnitude of c the power increases. However, for an outlying data set with almost half of the data outlying in an asymmetric fashion such as when $K = 7$ the power does not become large with c as much as -12. On the other hand the power is retained for large outlying data sets with high leverage if the contamination is symmetric. For instance, choosing $y_j = \beta_0 + \beta_1 x_j + (-1)^j c + \epsilon_j$ for $j = 1, ..., 6$ gave a value of $(p_1, p_2, p_3, p_4, p_5) = (69, 96, 19, 84, 96)$ for $c = -4$ and $(p_1, p_2, p_3, p_4, p_5) = (82, 100, 18, 96, 100)$ for $c = -5$.

Interestingly one can frequently expect more than one local minima of the objective function. Often there are two or three local minima for data generated by the normal regression model ($c = 0$). When data are generated with outliers ($c \neq 0$), there may still exist local minima at $g = 0$ for $V_n(g)$ and sometimes these local minima are global minima if **ATLA** is unable to detect the outliers (as for example with $K = 6, 7$). Hence if one suspects a group of outliers and a local minima is obtained for which $\tilde{J}(g)$ identifies the outliers, and where the global minima is given by $V_n(0)$, then further investigation is warranted. This is illustrated by the discussion of the Brownlee Stack Loss Data. The simulations do on the other hand, suggest that if $\tilde{g} > 0$, then one should regard the subsample $\tilde{J}(\tilde{g})$ as outliers.

5. Discussion

The idea of minimizing an asymptotic variance associated with the trimmed likelihood estimator follows from the discussion of the univariate location

estimation problem (Clarke, 1994). In the application to regression one observes a significant increase in computing time. For example, an approximate upper bound for the number of least squares estimates carried out in order to determine the ATL estimator in a sample of size n is 2^{n-1}. Naturally if $G^*(n)$ is much less than $[n/2]$, either because p is large or simply because $G^*(n)$ is chosen to be much smaller such as in Example 3.5, then this increases the possible sample size n for which **ATLA** can feasibly be applied. An important advantage of **ATLA** is that it does not necessarily require as a matter of course the interpretation of plots which may require some statistical expertise. If **ATLA** identifies several points as outlying the simulations of Section 4 and the data analysis of Section 3 would suggest that these points are definitely outliers which should be rejected from the sample. The global algorithm given is illustrated to work well on several known data sets and in simulation.

The one case where the global algorithm does not identify the outliers explicitly, but does highlight them implicitly, is for the Brownlee stack loss data. This illustrates that plots may be required if the global minima of V_n occurs at $g = 0$ and this competes with a local minima that is not too different in magnitude and where $g > 0$. For this particular data, failure of automatic outlier detection using the global algorithm is accompanied by a general departure from distributional assumptions for the bulk of the data, as is demonstrated by Figure 3.

The proposed use of **ATLA** yields a highly efficient estimator at the normal model in the ATL estimator, and adapts to highlight and cull outliers which may be multiple in number when outliers are present in the sample.

A suggestion by a referee was that one should take into account the true asymptotic variance for regression, which should be $var(\sigma^2, \alpha)(X'X)^{-1}$. One possibility is to use a generalised variance, which would be $var(\sigma^2, \alpha)$ $/det(X'X)$. However, this would not help matters, especially in the case of the Brownlee stack loss data, well known for its leverage points. For example if X_{-g} is the design matrix with g observations omitted, then it is a simple observation that $1/det(X'X) \leq 1/det(X'_{-g}X_{-g})$ and this would tend to bias the procedure further toward selecting $g = 0$ as a global minima in any generalization of (2.6), when the evidence is against this choice.

The approach introduced here for multiple outlier detection is different from the *elemental subset* method of Hawkins, Bradu and Kass (1984). Those authors introduce four variants of their procedure, and the introduction here of a single method that copes with influential data points is an advance. It is shown here the adaptive approach works in identifying

outliers. It is a single high breakdown approach, which can work automatically in identifying outliers. It does not rely on the choice of critical points for the size of the test for outliers as do methods of Marasinghe (1985) and Hadi and Simonoff (1993).

The current discussion focuses on the detection of outliers in regression using **ATLA**. It is not intended to discuss here the wider problems of statistical inference in regression or indeed the analogous multivariate estimation problem, both of which can be seen as extensions to this work. Discussion of the multivariate estimation and outlier detection problem is found in Woodruff and Rocke (1994) and Atkinson (1994). The latter article highlights several references to multiple outlier detection, for example the discussion of Davies and Gather (1993) which provides a more formal approach to outlier detection in univariate samples.

References

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

[2] A.C. Atkinson, *Masking unmasked,* Biometrika **73** (1986a), 533–41.

[3] A.C. Atkinson, *Comment : Aspects of diagnostic regression analysis*, Statistical Science **1** (1986b), 397–401.

[4] A.C. Atkinson, *Fast very robust methods for the detection of multiple outliers*, Journal of the American Statistical Association **89** (1994), 1329–1339.

[5] V. Barnett and T. Lewis, Outliers in Statistical Data, 3rd ed., New York, Wiley, 1994.

[6] T. Bednarski and B.R. Clarke, *Trimmed likelihood estimation of location and scale of the normal distribution*, Australian Journal of Statistics **35** (1993), 141–153.

[7] M.D. Brown, J. Durbin and J.M. Evans, *Techniques for testing the constancy of regression relationships over time*, Journal of the Royal Statistical Society, Series B **37** (1975), 149–192.

[8] K.A. Brownlee, Statistical Theory and Methodology in Science and Engineering, 2^{nd} ed., New York, Wiley, 1965.

[9] R.W. Butler, *Nonparametric interval and point prediction using data trimmed by a Grubbs-type outlier rule*, Annals of Statistics **10** (1982), 197–204.

[10] R.L. Chambers and C.R. Heathcote, *On the estimation of slope and the identification of outliers in linear regression*, Biometrika **68** (1981), 21–33.

[11] B.R. Clarke, *Empirical evidence for adaptive confidence intervals and identification of outliers using methods of trimming*, Australian Journal of Statistics **36** (1994), 45–58.

[12] R.D. Cook and S. Weisberg, Residuals and Influence in Regression, New York and London, Chapman and Hall 1982.

[13] P.L. Davies, *The asymptotics of S-estimators in the Linear Regression Model*, Annals of Statistics **18** (1990), 1651–1675.

[14] P.L. Davies and U. Gather, *The identification of multiple outliers (with discussion)*, Journal of the American Statistical Association **88** (1993), 782–801.

[15] N.R. Draper and H. Smith, Applied Regression Analysis, New York, Wiley, 1966.

[16] W. Fung, *Unmasking outliers and leverage points : A confirmation*, Journal of the American Statistical Association **88** (1993), 515–519.

[17] A.S. Hadi and J.S. Simonoff, *Procedures for the identification of multiple outliers in linear models*, Journal of the American Statistical Association **88** (1993), 1264–1272.

[18] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.J. Stahel, Robust Statistics, the Approach Based on Influence Functions, New York, Wiley, 1986.

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

[20] T.P. Hettmansperger and S.J. Sheather, *A cautionary note on the method of least median squares*, American Statistician **46** (1992), 79–83.

[21] L.A. Jaeckel, *Some flexible estimates of location*, Annals of Mathematical Statistics **42** (1971), 1540–1552.

[22] F. Kianifard and W.H. Swallow, *Using recursive residuals, calculated on adaptively-ordered observations, to identify outliers in linear regression*, Biometrics **45** (1989), 571–585.

[23] F. Kianifard and W.H. Swallow, *A Monte Carlo comparison of five procedures for identifying outliers in linear regression*, Communications in Statistics, Part A-Theory and Methods **19** (1990), 1913–1938.

[24] M.G Marasinghe, *A multistage procedure for detecting several outliers in linear regression*, Technometrics **27** (1985), 395–399.

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

[26] P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection, New York, Wiley, 1987.

[27] P.J. Rousseeuw and B.C. van Zomeren, *Unmasking multivariate outliers and leverage points*, Journal of the American Statistical Association **85** (1990), 633–651.

[28] P.J. Rousseeuw and V.J. Yohai, *Robust regression by means of S-estimators*, in: Robust and Nonlinear Time Series Analysis, eds., J. Franke, W. Härdle and R.D. Martin, (Lecture Notes in Statistics), New York, Springer-Verlag, (1984), 256–272.

[29] D. Ruppert, *Computing S-estimators for regression and multivariate location/dispersion*, Journal of Computational and Graphical Statistics **1** (1992), 253–270.

[30] T.P. Ryan, *Comment on Hadi and Simonoff*, Letters to the Editor, Journal of the American Statistical Association **90** (1995), 811.

[31] G. Simpson, D. Ruppert and R.J. Carroll, *On one-step GM estimates and stability of inferences in linear regression*, Journal of the American Statistical Association **87** (1992), 439–450.

[32] W.H. Swallow and F. Kianifard, *Using robust scale estimates in detecting multiple outliers in linear regression*, Biometrics **52** (1996), 545–556.

[33] J.W Tukey and D.H. McLaughlin, *Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization* 1, Sankhya **25** (A) (1963), 331–352.

[34] W.N. Venables and B.D. Ripley, Modern Applied Statistics with S-Plus, New York, Springer-Verlag, 1994.

[35] D.L. Woodruff and D.M. Rocke, *Computable robust estimation of multivariate location and shape in high dimension using compound estimators*, Journal of the American Statistical Association **89** (1994), 888–896.

[36] V.J. Yohai, *High breakdown point and high-efficiency robust estimates for regression*, Annals of Statistics **15** (1987), 642–656.

[37] V.J. Yohai and R.H. Zamar, *High breakdown-point estimates of regression by means of the minimization of an efficient scale*, Journal of the American Statistical Association **83** (1988), 406–413.

Received 15 November 1998