Two Simple Resistant Regression Estimators

David J. Olive*

Southern Illinois University

January 13, 2005

Abstract

Two simple resistant regression estimators with $O_P(n^{-1/2})$ convergence rate are presented. Ellipsoidal trimming can be used to trim the cases corresponding to predictor variables $\boldsymbol{x}$ with large Mahalanobis distances, and the forward response plot of the residuals versus the fitted values can be used to detect outliers. The first estimator uses ten forward response plots corresponding to ten different trimming proportions, and the final estimator corresponds to the "best" forward response plot. The second estimator is similar to the elemental resampling algorithm, but sets of $O(n)$ cases are used instead of randomly selected elemental sets.

These two estimators should be regarded as new tools for outlier detection rather than as replacements for existing methods. Outliers should always be examined

[*] David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@math.siu.edu. This research was supported by NSF grant DMS 0202922. The author is grateful to the editors and referees for a number of helpful suggestions for improvement in the article.

to see if they follow a pattern, are recording errors, or if they could be explained adequately by an alternative model. Using scatterplot matrices of fitted values and residuals from several resistant estimators is a very useful method for comparing the different estimators and for checking the assumptions of the regression model.

KEY WORDS: diagnostics; outliers; robust regression.

1 INTRODUCTION

Consider the multiple linear regression (MLR) model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{1.1}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of errors. The ith case $(y_i, \boldsymbol{x}_i^T)$ corresponds to the ith element y_i of $\boldsymbol{Y}$ and the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$.

Most regression methods attempt to find an estimate $\boldsymbol{b}$ for $\boldsymbol{\beta}$ which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals where the ith residual $r_i = r_i(\boldsymbol{b}) = y_i - \boldsymbol{x}_i^T\boldsymbol{b}$. Two of the most used classical regression methods are ordinary least squares (OLS) and least absolute deviations (L_1). OLS and L_1 choose $\hat{\boldsymbol{\beta}}$ to minimize

$$Q_{OLS}(\boldsymbol{b}) = \sum_{i=1}^{n} r_i^2 \quad \text{and} \quad \mathrm{Q_{L_1}}(\boldsymbol{b}) = \sum_{\mathrm{i=1}}^{\mathrm{n}} |\mathrm{r_i}|, \tag{1.2}$$

respectively.

Some high breakdown robust regression methods can fit the bulk of the data even if certain types of outliers are present. Let $r_{(i)}^2(\boldsymbol{b})$ denote the squared residuals sorted from smallest to largest. Suppose that the integer valued parameter $c_n \approx n/2$. Then the least median of squares (LMS(c_n)) estimator (Hampel 1975) minimizes the criterion

$$Q_{LMS}(\boldsymbol{b}) = r_{(c_n)}^2(\boldsymbol{b}). \tag{1.3}$$

The least trimmed sum of squares (LTS(c_n)) estimator (Rousseeuw 1984) minimizes the criterion

$$Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c_n} r_{(i)}^2(\boldsymbol{b}), \tag{1.4}$$

and the least trimmed sum of absolute deviations (LTA(c_n)) estimator (Hawkins and Olive 1999) minimizes the criterion

$$Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^{c_n} |r|_{(i)}(\boldsymbol{b}). \tag{1.5}$$

Robust regression estimators tend to be judged by their Gaussian efficiency and breakdown value. To formally define breakdown (see Zuo 2001 for references), the following notation will be useful. Let $\boldsymbol{W}$ denote the $n \times (p+1)$ data matrix where the ith case corresponds to the ith row $(y_i, \boldsymbol{x}_i^T)$ of $\boldsymbol{W}$. Let $\boldsymbol{W}_d^n$ denote the data matrix where any d of the cases have been replaced by arbitrarily bad contaminated cases. Then the contamination fraction is $\gamma = d/n$.

If $T(\boldsymbol{W})$ is a $p \times 1$ vector of regression coefficients, then the breakdown value of T is

$$B(T, \boldsymbol{W}) = \min\{\frac{d}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \le d \le n$.

A regression estimator basically "breaks down" if d outliers can make the median absolute residual arbitrarily large. Consider a fixed data set $\boldsymbol{W}_d^n$ with ith row $(z_i, \boldsymbol{w}_i^T)$. If the regression estimator $T(\boldsymbol{W}_d^n) = \hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}}\| = M$ for some constant M, then the median absolute residual $\text{MED}(|z_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{w}_i|)$ is bounded by $\max_{i=1,...,n} \|y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i\| \le \max_{i=1,...,n}[|y_i| + \sum_{j=1}^p M|x_{i,j}|]$ if $d < n/2$.

Now suppose that $\|\hat{\boldsymbol{\beta}}\| = \infty$. Since the absolute residual is the vertical distance of the observation from the hyperplane, the absolute residual $|r_i| = 0$ if the ith case lies on the regression hyperplane, but $|r_i| = \infty$ otherwise. Hence the median absolute residual will equal ∞ if fewer than half of the cases lie on the regression hyperplane. This will occur unless the proportion of outliers $d/n > (n/2 - q)/n \to 0.5$ as $n \to \infty$ where q is the number of "good" cases that lie on a hyperplane of lower dimension than p. In the literature it is usually assumed that the original data is in general position: $q = p - 1$. For example, if $p = 2$, then $q = 1$ if all cases are distinct: a vertical line can be formed with one "good" case and with d outliers placed on a point mass.

This result implies that (due to asymptotic equivalence if the breakdown value ≤ 0.5) breakdown can be computed using the median absolute residual $\text{MED}(|r_i|(\boldsymbol{W}_d^n))$ instead of $\|T(\boldsymbol{W}_d^n)\|$. This result also implies that the breakdown value of a regression estimator is more of a y–outlier property than an $\boldsymbol{x}$–outlier property. If the y_i's are fixed, arbitrarily large $\boldsymbol{x}$–outliers tend to drive the slope estimates to zero. The result also implies that the LMS estimator is "best" in terms of breakdown since the LMS estimator minimizes the "median" squared absolute residual.

Perhaps the simplest affine equivariant high breakdown regression estimator can be found by computing OLS on the set S of approximately $n/2$ cases that have $y_i \in [\text{MED}(y_i) \pm \text{MAD}(y_i)]$ where $\text{MED}(y_i)$ is the median and $\text{MAD}(y_i) = \text{MED}(|y_i - \text{MED}(y_i)|)$ is the median absolute deviation of the response variable. To see this, suppose that n is odd and that the model has an intercept β_1. Consider the estimator

$$\hat{\boldsymbol{\beta}}_M = (\text{MED}(y_i), 0, ..., 0)^T$$

which yields the predicted values $\hat{y}_i \equiv \text{MED}(y_i)$. The squared residual

$$r_i^2(\hat{\boldsymbol{\beta}}_M) \leq (\text{MAD}(y_i))^2$$

if the ith case is in S. Hence the OLS fit $\hat{\boldsymbol{\beta}}_S$ to the cases in S has

$$\sum_{i \in S} r_i^2(\hat{\boldsymbol{\beta}}_S) \leq n(\text{MAD}(y_i))^2,$$

and

$$\text{MED}(|r_i(\hat{\boldsymbol{\beta}}_S)|) \leq \sqrt{n}\text{MAD}(y_i) < \infty$$

if $\text{MAD}(y_i) < \infty$. Hence the estimator has a high breakdown value, but it only resists large y–outliers.

There is an enormous literature on the detection of outliers and influential cases for the multiple linear regression model. The "elemental (basic) resampling" algorithm for robust regression estimators uses K_n randomly selected "elemental" subsets of p cases where p is the number of predictors. An estimator is computed from the elemental set and then a criterion function that depends on all n cases is computed. The algorithm returns the elemental fit that optimizes the criterion. The efficiency and resistance properties of the elemental resampling algorithm estimator turn out to depend strongly on the number of starts K_n used, and many of the most used algorithm estimators are inconsistent with zero breakdown – see Hawkins and Olive (2002).

Many types of outlier configurations occur in real data, and no single estimator can perform well on every outlier configuration. A resistant estimator should have good statistical properties on "clean data" and perform well for several of the most commonly occuring outlier configurations. Sections 2 and 3 describe two simple resistant estimators.

2 The Trimmed Views Estimator

Ellipsoidal trimming can be used to create resistant estimators. To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed from the predictor variables where T is a $p \times 1$ multivariate location estimator and $\boldsymbol{C}$ is a $p \times p$ symmetric positive definite dispersion estimator. Then the ith squared Mahalanobis distance is the scalar

$$D_i^2 \equiv D_i^2(T, \boldsymbol{C}) = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x}_i - T) \tag{2.1}$$

for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distance $D_{(j)}$ is unique, then j of the $\boldsymbol{x}_i$'s are in the ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \le D_{(j)}^2\}. \tag{2.2}$$

The ith case $(y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Then an estimator of $\boldsymbol{\beta}$ is computed from the untrimmed cases. For example, if $j \approx 0.9n$, then about 10% of the cases are trimmed, and OLS or L_1 could be used on the untrimmed cases. Trimming using $(T, \boldsymbol{C})$ computed from a subset of the predictors may be useful if some of the predictors are categorical.

A forward response plot is a plot of the fitted values $\hat{y}_i$ versus the response y_i. Since MLR is the study of the conditional distribution of $y_i | \boldsymbol{x}_i^T \boldsymbol{\beta}$, the forward response plot is used to visualize this conditional distribution. If the MLR model holds and the MLR estimator is good, then the plotted points will scatter about the identity line that has unit slope and zero intercept. The identity line is added to the plot as a visual aid, and the vertical deviations from the identity line are equal to the residuals since $y_i - \hat{y}_i = r_i$.

Modifying the Olive (2002) procedure (for visualizing g in models of the form $y_i = g(\boldsymbol{\beta}^T \boldsymbol{x}_i, e_i)$) results in a resistant MLR estimator similar to one proposed by Rousseeuw and van Zomeren (1992). First compute $(T, \boldsymbol{C})$ using the *Splus* function `cov.mcd` (see Rousseeuw and Van Driessen 1999). Trim the $M\%$ of the cases with the largest Mahalanobis distances, and then compute the MLR estimator $\hat{\boldsymbol{\beta}}_M$ from the untrimmed cases. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten forward response plots of the fitted values $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i$ versus y_i using all n cases. (Fewer plots are used for small data sets if $\hat{\boldsymbol{\beta}}_M$ can not be computed for large M.) These plots are called "trimmed views," and as a resistant MLR estimator, the final trimmed views (TV) estimator $\hat{\boldsymbol{\beta}}_{T,n}$ corresponds to the plot where the bulk of the plotted points follow the identity line with smallest variance function, ignoring any outliers. The following example helps illustrate the procedure.

Example 1. Buxton (1920, pp. 232–5) gives 20 measurements of 88 men. *Height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet! OLS was used on the untrimmed cases and Figure 1 shows four trimmed views corresponding to 90%, 70%, 40% and 0% trimming. The OLS TV estimator used 70% trimming since this trimmed view was best. Since the vertical distance from a plotted point to the identity line is equal to the case's residual, the outliers had massive residuals for 90%, 70% and 40% trimming. Notice that the OLS trimmed view with 0% trimming "passed through the outliers" since the cluster of outliers is scattered about the identity line.

For this data set, the relationship between the response variable and the predictors is very weak, and Hawkins and Olive (2002) suggest that the exact LMS, LTS and LTA estimators will also pass through the outliers. (If the outliers were pulled towards $-\infty$, then the high breakdown estimators would eventually give the outliers weight zero.) As will be seen in the following section, the estimators produced by the *Splus* functions `lmsreg` and `ltsreg` also pass through the outliers. When `lmsreg` replaced OLS in the TV estimator, the outliers had massive residuals except for the 0% trimming proportion.

The TV estimator $\hat{\boldsymbol{\beta}}_{T,n}$ has good statistical properties if the estimator applied to the untrimmed cases $(\boldsymbol{X}_{M,n}, \boldsymbol{Y}_{M,n})$ has good statistical properties. Candidates include OLS, L_1, Huber's M–estimator, Mallows' GM–estimator or the Wilcoxon rank estimator. See Rousseeuw and Leroy (1987, pp. 12-13, 150). The basic idea is that if an estimator with $O_P(n^{-1/2})$ convergence rate is applied to a set of $n_M \propto n$ cases, then the resulting estimator $\hat{\boldsymbol{\beta}}_{M,n}$ also has $O_P(n^{-1/2})$ rate provided that the response y was not used to select the n_M cases in the set. If $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ for $M = 0, ..., 90$ then $\|\hat{\boldsymbol{\beta}}_{T,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ by Pratt (1959).

Let $\boldsymbol{X}_n = \boldsymbol{X}_{0,n}$ denote the full design matrix. Often when proving asymptotic normality of an MLR estimator $\hat{\boldsymbol{\beta}}_{0,n}$, it is assumed that

$$\frac{\boldsymbol{X}_n^T \boldsymbol{X}_n}{n} \to \boldsymbol{W}^{-1}.$$

If $\hat{\boldsymbol{\beta}}_{0,n}$ has $O_P(n^{-1/2})$ rate and if for big enough n all of the diagonal elements of

$$\left(\frac{\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}}{n}\right)^{-1}$$

are all contained in an interval $[0, B)$ for some $B > 0$, then $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$.

The distribution of the estimator $\hat{\boldsymbol{\beta}}_{M,n}$ is especially simple when OLS is used and the errors are iid $N(0, \sigma^2)$. Then

$$\hat{\boldsymbol{\beta}}_{M,n} = (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1} \boldsymbol{X}_{M,n}^T \boldsymbol{Y}_{M,n} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n})^{-1})$$

and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}) \sim N_p(\boldsymbol{0}, \sigma^2(\boldsymbol{X}_{M,n}^T \boldsymbol{X}_{M,n}/n)^{-1})$. Notice that this result does not imply that the distribution of $\hat{\boldsymbol{\beta}}_{T,n}$ is normal.

3 The MBA Estimator

Next we describe a simple resistant algorithm estimator, called the *median ball algorithm* (MBA). The Euclidean distance of the ith vector of predictors $\boldsymbol{x}_i$ from the jth vector of predictors $\boldsymbol{x}_j$ is

$$D_i \equiv D_i(\boldsymbol{x}_j) \equiv D_i(\boldsymbol{x}_j, \boldsymbol{I}_p) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}.$$

For a fixed $\boldsymbol{x}_j$ consider the ordered distances

$$D_{(1)}(\boldsymbol{x}_j), ..., D_{(n)}(\boldsymbol{x}_j).$$

Next, let $\hat{\boldsymbol{\beta}}_j(\alpha)$ denote the OLS fit to the $\min(p+3+[\alpha n/100], n)$ cases with the smallest distances where the approximate percentage of cases used is $\alpha \in \{1, 2.5, 5, 10, 20, 33, 50\}$. (Here $[x]$ is the greatest integer function so $[7.7] = 7$. The extra $p+3$ cases are added so that OLS can be computed for small n and α.) This yields seven OLS fits corresponding to the cases with predictors closest to $\boldsymbol{x}_j$. A fixed number K of cases are selected at random without replacement to use as the $\boldsymbol{x}_j$. We use $K = 7$ as the default. A robust criterion Q, such as the median squared residual, is used to evaluate the $7K$ fits and the OLS fit to all of the data. Hence $7K + 1$ OLS fits are generated and the OLS MBA estimator is the fit that minimizes the criterion Q.

This estimator is simple to program and easy to modify. For example change the criterion Q or change K. Alternatively, replacing the $7K + 1$ OLS fits by L_1 fits results in the more resistant L_1 MBA estimator. In the discussion below, the MBA estimator is the OLS MBA estimator.

Three ideas motivate this estimator. First, $\boldsymbol{x}$–outliers, which are outliers in the predictor space, tend to be much more destructive than y–outliers which are outliers in the response variable. Suppose that the proportion of outliers is γ and that $\gamma < 0.5$. We would like the algorithm to have at least one "center" $\boldsymbol{x}_j$ that is not an outlier. The probability of drawing a center that is not an outlier is approximately $1 - \gamma^K > 0.99$ for $K \geq 7$ and this result is free of p. Secondly, by using the different percentages of coverages, for many data sets there will be a center and a coverage that contains no outliers. Thirdly, since only a fixed number $(7K + 1)$ of fits with $O_P(n^{-1/2})$ rate are computed, the MBA estimator has an $O_P(n^{-1/2})$ convergence rate (by Pratt 1959).

Example 1 continued. When comparing different estimators, it is useful to make an RR plot which is simply a scatterplot matrix of the residuals from the various estimators. Figure 2 shows the RR plot applied to the Buxton (1920) data for the *Splus* estimators `lsfit`, `l1fit`, `lmsreg` (denoted by ALMS), `ltsreg` (denoted by ALTS), and the MBA estimator. Note that only the MBA estimator gives large absolute residuals to the outliers.

Table 1 compares the TV, MBA, `lmsreg`, `ltsreg`, L_1 and OLS estimators on 7 data sets available from the author's website (http://www.math.siu.edu/olive/ol-bookp.htm). The column headers give the file name while the remaining rows of the table give the sample size n, the number of predictors p, the amount of trimming M used by the TV estimator, the correlation of the residuals from the TV estimator with the corresponding alternative estimator, and the cases that were outliers. If the correlation was greater than 0.9, then the method was effective in detecting the outliers, and the method failed, otherwise. Sometimes the trimming percentage M for the TV estimator was picked after fitting the bulk of the data in order to find the good leverage points and outliers.

Notice that the TV, MBA and OLS estimators were the same for the Gladstone data and for the *major* data which had two small y–outliers. For the Gladstone data, there is a cluster of infants that are good leverage points, and we attempt to predict *brain weight* with the head measurements *height, length, breadth, size* and *cephalic index.* Originally, the variable *length* was incorrectly entered as 109 instead of 199 for case 119, and the *glado* data contains this outlier. In 1997, `lmsreg` was not able to detect the outlier while `ltsreg` did. Due to changes in the *Splus* 2000 code, `lmsreg` now detects the outlier but `ltsreg` does not.

Both the TV and MBA estimators have resistance comparable to that of `lmsreg`. A data set in Table 1 where `lmsreg` outperforms the MBA estimator is the Douglas M. Hawkins' *nasty* data. The MBA estimator may be superior to `lmsreg` for data sets such as the Buxton data where the bulk of the data follow a very weak linear relationship and there is a single cluster of outliers. The `ltsreg` estimator should not be used since it is inconsistent and is rarely able to detect x–outliers.

The MBA estimator depends on the sample of 7 centers drawn and changes each time the function is called. After running MBA several times, sometimes there is a forward response plot or RR plot that differs greatly from the other plots. This feature is useful for data sets like the *nasty* data. On the other hand, in ten runs on the Buxton data, about nine RR plots will look like Figure 2, but in about one RR plot the MBA estimator will also pass through the outliers.

4 Conclusions and Extensions

The author's website contains a file *rpack.txt* of several *Splus* functions including the `mba` and `tv` functions. When some of the variables are categorical, the TV estimator may not work because the covariance estimator used for trimming is singular. A simple solution is to perform the trimming using only the continuous predictors. This technique is not necessary for the MBA estimator since the Euclidean distance works for categorical and continuous predictors.

In the literature there are many high breakdown estimators that are impractical to compute such as the CM, maximum depth, GS, LQD, LMS, LTS, LTA, MCD, MVE, projection, repeated median and S estimators. Two stage estimators that use an initial high breakdown estimator from the above list are even less practical to compute. These estimators include the cross–checking, MM, one step GM, one step GR, REWLS, tau and t type estimators. Implementations of the two stage estimators tend to use an inconsistent zero breakdown initial estimator, resulting in a zero breakdown final estimator that is often inconsistent. No single robust algorithm estimator seems to be very good, and for any given estimator, it is easy to find outlier configurations where the estimator fails. Hawkins and Olive (2002) discuss outlier configurations that can cause problems for robust regression algorithm estimators.

Often the assumptions needed for large sample theory are better approximated by the distribution of the untrimmed data than by the entire data set, and it is often suggested that the statistical analysis should be run on the "cleaned data set" where the outliers have been deleted. For the MLR model, the forward response plot should always be made and is a useful diagnostic for goodness of fit and for detecting outliers. The TV and MBA estimators use these facts to produce simple resistant estimators with the good $O_P(n^{-1/2})$ convergence rate. These two estimators should be regarded as new tools for outlier detection rather than as replacements for existing methods.

There are two approaches that are useful for detecting outliers in the MLR setting. The first approach is to compute several algorithm estimators as well as OLS and L_1. Then use plots to detect outliers, to check the goodness of fit of the MLR model, and to compare the different estimators. In particular, make the forward response plots and residuals plots for each estimator. Then make the RR plot and the FF plot, which is a scatterplot matrix of the response and the fitted values from the different estimators. An advantage of the FF plot is that the forward response plots of the different estimators appear in the scatterplot matrix. This technique can be modified if a parametric model is used. For example, add the maximum likelihood estimator, a Bayesian estimator or an estimator that works well in the presence of heteroscedasticity.

The second approach is to make an adaptive estimator from two or more estimators. The cross–checking estimator uses an asymptotically efficient estimator if it is close to the robust estimator but uses the robust estimator otherwise. If the robust estimator is a high breakdown consistent estimator, then the cross–checking estimator is both high breakdown and asymptotically efficient. Plots of residuals and fitted values from both estimators should still be made since the probability that the robust estimator is chosen when outliers are present is less than one. The proofs in He (1991, p. 304), He and Portnoy (1992, p. 2163) and Davies (1993, pp. 1889-1891) need the robust estimator to be consistent, and `lmsreg` and `ltsreg` are inconsistent since they use a fixed number

(3000) of elemental sets. It needs to be shown that using n elemental starts or using a consistent start in an LTS concentration algorithm (see Hawkins and Olive 2002) results in a consistent estimator. The conjectured consistency of such an algorithm is in the folklore (see Maronna and Yohai 2002), but no proofs of these conjectures are available.

Although both the TV and MBA estimators have the good $O_P(n^{-1/2})$ convergence rate, their efficiency under normality may be very low. (We could argue that the TV and OLS estimators are asymptotically equivalent on clean data if 0% trimming is always picked when all 10 plots look good.) Using the TV and MBA estimators as the initial estimator in the cross–checking estimator results in a resistant (easily computed but zero breakdown) asymptotically efficient final estimator. High breakdown estimators that have high efficiency tend to be impractical to compute.

The ideas used in this paper have the potential for making many methods resistant. First, suppose that the MLR model (1.1) holds but $\text{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{V}'$ where $\boldsymbol{V}$ is known and nonsingular. Then $\boldsymbol{V}^{-1}\boldsymbol{Y} = \boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}^{-1}\boldsymbol{e}$, and the TV and MBA estimators can be applied to $\tilde{\boldsymbol{Y}} = \boldsymbol{V}^{-1}\boldsymbol{Y}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{V}^{-1}\boldsymbol{X}$ provided that OLS is fit without an intercept. Similarly, the minimum chi squared estimators for several generalized linear models can be fit with an OLS regression (without an intercept) that uses appropriate $\tilde{Y}$ and $\tilde{\boldsymbol{X}}$. See Agresti (2002, p. 611).

Secondly, many 1D regression models where y_i is independent of $\boldsymbol{x}_i$ given the sufficient predictor $\boldsymbol{x}_i^T\boldsymbol{\beta}$ can be made resistant by making EY plots of the estimated sufficient predictor $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ versus y_i for the 10 trimming proportions. Since 1D regression is the study of the conditional distribution of y_i given $\boldsymbol{x}_i^T\boldsymbol{\beta}$, the EY plot is used to visualize this distribution and needs to be made anyway. These plots were called trimmed views by Olive (2002) where the data sets were assumed to be clean.

Thirdly, for nonlinear regression models of the form $y_i = m(\boldsymbol{x}_i, \boldsymbol{\beta}) + e_i$, the fitted values are $\hat{y}_i = m(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})$ and the residuals are $r_i = y_i - \hat{y}_i$. The points in the FY plot of the fitted values versus the response should follow the identity line. The TV estimator would make FY and residual plots for each of the trimming proportions. The MBA estimator with the median squared residual criterion can also be used for many of these models.

5 References

Agresti, A. (2002), *Categorical Data Analysis*, 2nd ed., John Wiley and Sons, Hoboken, NJ.

Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Davies, P.L. (1993), "Aspects of Robust Linear Regression," *The Annals of Statistics,* 21, 1843-1899.

Hampel, F.R. (1975), "Beyond Location Parameters: Robust Concepts and Methods," *Bulletin of the International Statistical Institute,* 46, 375-382.

Hawkins, D.M., and Olive, D. (1999), "Applications and Algorithms for Least Trimmed Sum of Absolute Deviations Regression," *Computational Statistics and Data Analysis,* 32, 119-134.

Hawkins, D.M., and Olive, D.J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm" (with discussion), *Journal of the American Statistical Association,* 97, 136-159.

He, X. (1991), "A Local Breakdown Property of Robust Tests in Linear Regression," *Journal of Multivariate Analysis,* 38, 294-305.

He, X., and Portnoy, S. (1992), "Reweighted LS Estimators Converge at the Same Rate as the Initial Estimator," *The Annals of Statistics,* 20, 2161-2167.

Maronna, R.A., and Yohai, V.J. (2002), "Comment on 'Inconsistency of Resampling Algorithms for High Breakdown Regression and a New Algorithm' by D.M. Hawkins and D.J. Olive," *Journal of the American Statistical Association,* 97, 154-155.

Olive, D.J. (2002), "Applications of Robust Distances for Regression," *Technometrics,* 44, 64-71.

Pratt, J.W. (1959), "On a General Concept of 'in Probability'," *The Annals of Mathematical Statistics,* 30, 549-558.

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

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* John Wiley and Sons, NY.

Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Rousseeuw, P.J., and van Zomeren, B.C. (1992), "A Comparison of Some Quick Algorithms for Robust Regression," *Computational Statistics and Data Analysis,* 14, 107-116.

Zuo, Y. (2001), "Some Quantitative Relationships between Two Types of Finite Sample Breakdown Point," *Statistics and Probability Letters,* 51, 369-375.

Table 1: Summaries for Seven Data Sets, cor(TV,Method) is the Correlation of the Residuals from TV(M) and the Alternative Method

summary/file	Buxton	Gladstone	glado	hbk	major	nasty	wood
cor(TV,MBA)	0.997	1.0	0.455	0.960	1.0	-0.004	0.9997
cor(TV,LMSREG)	-0.114	0.671	0.938	0.977	0.981	0.9999	0.9995
cor(TV,LTSREG)	-0.048	0.973	0.468	0.272	0.941	0.028	0.214
cor(TV,L1)	-0.016	0.983	0.459	0.316	0.979	0.007	0.178
cor(TV,OLS)	0.011	1.0	0.459	0.780	1.0	0.009	0.227
outliers	61-65	none	119	1-10	3,44	2,6,...,30	4,6,8,19
n	87	247	247	75	112	32	20
p	5	7	7	4	6	5	6
M	70	0	30	90	0	90	20

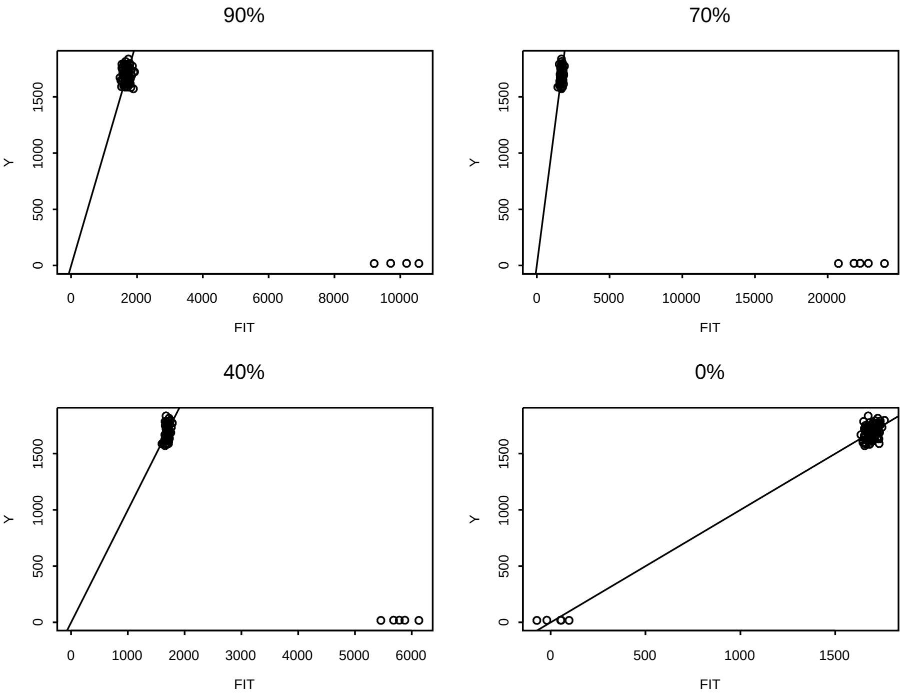

Figure 1: 4 Trimmed Views for the Buxton Data

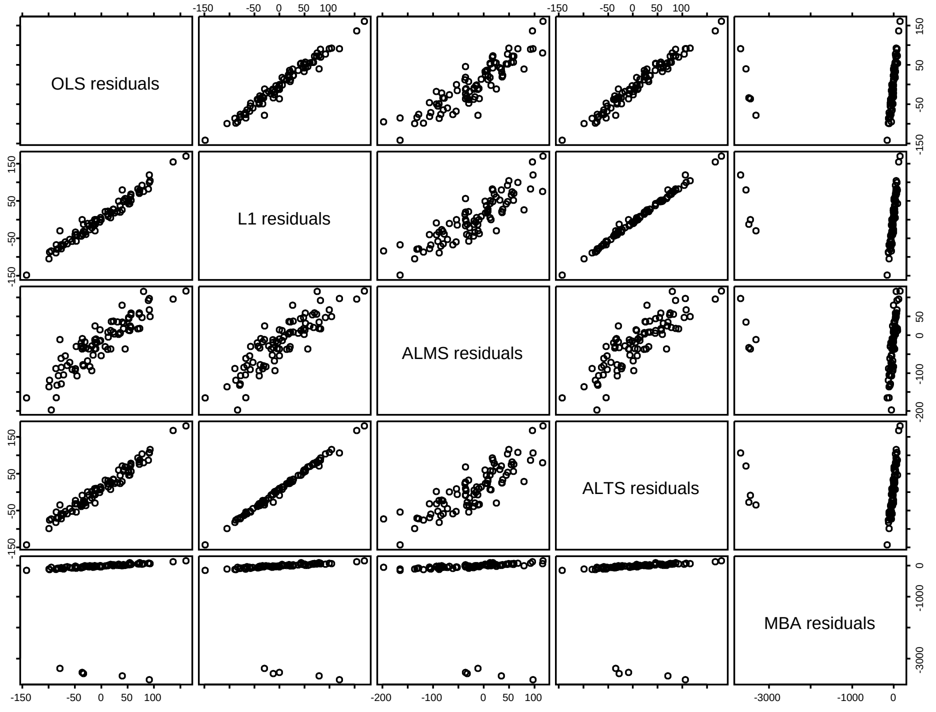

Figure 2: RR Plot for the Buxton Data