

Practical High Breakdown Regression

David J. Olive and Douglas M. Hawkins*

Southern Illinois University and University of Minnesota

February 8, 2011

Abstract

This paper shows that practical high breakdown \sqrt{n} consistent regression estimators exist. The response plot of the fitted values versus the response and the residual plot of the fitted values versus the residuals are shown to be useful for detecting outliers and groups of high leverage cases.

KEY WORDS: LMS; LTS; Outliers; Response Plot.

*David J. Olive is Associate Professor (E-mail: dolive@math.siu.edu), Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408, USA. Douglas M. Hawkins is Professor (E-mail: dhawkins@umn.edu), School of Statistics, University of Minnesota, Minneapolis, MN 55455-0493, USA. Their work was supported by the National Science Foundation under grants DMS 0600933, DMS 0306304, DMS 9803622 and ACI 9619020.

1. INTRODUCTION

A long standing question in Statistics is whether high breakdown regression is a viable field of study: are there high breakdown consistent regression estimators that are practical to compute? Huber and Ronchetti (2009, pp. xiii, 8-9, 152-154, 196-197) suggest that high breakdown regression estimators do not provide an adequate remedy for the ill effects of outliers, that their statistical and computational properties are not adequately understood, that high breakdown estimators “break down for all except the smallest regression problems by failing to provide a timely answer!” and that “there are no known high breakdown point estimators of regression that are demonstrably stable.” This paper provides practical high breakdown \sqrt{n} consistent estimators, providing a partial remedy for these concerns.

The *multiple linear regression model* is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where \mathbf{Y} is an $n \times 1$ vector of response variables, \mathbf{X} is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients and \mathbf{e} is an $n \times 1$ vector of errors. The i th case (\mathbf{x}_i^T, Y_i) corresponds to the i th row \mathbf{x}_i^T of \mathbf{X} and the i th element of \mathbf{Y} , and $Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + e_i$ for $i = 1, \dots, n$.

If d_n of the cases have been replaced by arbitrarily bad contaminated cases, then the contamination fraction is $\gamma_n = d_n/n$. Then the breakdown value of $\hat{\boldsymbol{\beta}}$ is the smallest value of γ_n needed to make $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large. High breakdown regression estimators have $\gamma_n \rightarrow 0.5$ as $n \rightarrow \infty$ if the clean (uncontaminated) data are in general position: any p clean cases give a unique estimate of $\boldsymbol{\beta}$. *For the remainder of this paper, assume that the clean data are in general position.* Estimators are zero breakdown if $\gamma_n \rightarrow 0$ and positive breakdown if $\gamma_n \rightarrow \gamma > 0$ as $n \rightarrow \infty$.

Since Hampel (1975) proposed the least median of squares (LMS) estimator, a large number of high breakdown regression estimators have been proposed as alternatives to least squares (OLS). The computational complexity of the “brand name” high breakdown estimators is too high: the fastest regression estimators that have been shown to be high breakdown and consistent are LMS and the least trimmed sum of absolute deviations (LTA) estimator with $O(n^p)$ complexity. The least trimmed sum of squares (LTS), least quantile of differences, repeated median and regression depth complexities are far higher, and there may be no known method for computing S, τ , projection based, constrained M and MM estimators. See Maronna, Martin and Yohai (2006, ch. 2) for references. Theory for LTA, LTS and LMS is given by Čížek (2006, 2008) and Kim and Pollard (1990) while computation is discussed by Bernholt (2005) and Hawkins and Olive (1999).

Since the above estimators take too long to compute, they are replaced by practical estimators that have not been shown to be both consistent and high breakdown. Often practical “robust estimators” generate a sequence of K trial fits called *attractors*: $\mathbf{b}_1, \dots, \mathbf{b}_K$. Then some criterion is evaluated and the attractor \mathbf{b}_A that minimizes the criterion is used as the final estimator. One way to obtain attractors is to generate trial fits called *starts*, and then use the *concentration* technique. Let $\mathbf{b}_{0,j}$ be the j th start and compute all n residuals $r_i(\mathbf{b}_{0,j}) = Y_i - \mathbf{x}_i^T \mathbf{b}_{0,j}$. Let $[n/2] \leq c_n \leq [n/2] + [(p+1)/2]$. At the next iteration, the OLS estimator $\mathbf{b}_{1,j}$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest squared residuals $r_i^2(\mathbf{b}_{0,j})$. This iteration can be continued for k steps resulting in the sequence of estimators $\mathbf{b}_{0,j}, \mathbf{b}_{1,j}, \dots, \mathbf{b}_{k,j}$. Then $\mathbf{b}_{k,j}$ is the j th attractor for $j = 1, \dots, K$. Using $k = 10$ concentration steps often works well, and the basic resampling algorithm is a special case with $k = 0$, i.e., the attractors are the starts. Elemental starts

are the fits from randomly selected “elemental sets” of p cases.

Many criteria for screening the attractors have been suggested. The $LMS(c_n)$ criterion is $Q_{LMS}(\mathbf{b}) = r_{(c_n)}^2(\mathbf{b})$ where $r_{(1)}^2 \leq \dots \leq r_{(n)}^2$ are the ordered squared residuals, and the $LTS(c_n)$ criterion is $Q_{LTS}(\mathbf{b}) = \sum_{i=1}^{c_n} r_{(i)}^2(\mathbf{b})$. The $LTA(c_n)$ criterion is $Q_{LTA}(\mathbf{b}) = \sum_{i=1}^{c_n} |r(\mathbf{b})|_{(i)}$ where $|r(\mathbf{b})|_{(i)}$ is the i th ordered absolute residual.

Although dozens of papers note that the high breakdown consistent regression estimators that have appeared in the literature are hard to compute, hundreds of papers claim to use practical, consistent, “brand name high breakdown” estimators such as LMS or LTS. In the published literature, to our knowledge, regression estimators proven to be both consistent and high breakdown have at least $O(n^p)$ complexity, conversely, if the estimator is practical to compute, then it has not been proven to be both high breakdown and consistent.

A common mistake in the literature is to claim that a brand name high breakdown estimator can be practically computed by using a few hundred randomly selected attractors and choosing the attractor that minimizes the brand name criterion. For example, Rousseeuw and Leroy (1987) use the elemental basic resampling algorithm estimators while Hubert, Rousseeuw and Van Aelst (2008), Rousseeuw, Van Aelst and Hubert (1999, p. 425) and Rousseeuw and Van Driessen (2006) claim that the LTS estimator can be computed with the FAST-LTS elemental concentration algorithm. These claims are false since Hawkins and Olive (2002) proved that elemental concentration algorithms are zero breakdown and that elemental basic resampling estimators are zero breakdown and inconsistent.

Hubert, Rousseeuw and Van Aelst (2002) reported that they appreciate this work,

but ignore it in their later papers. Maronna and Yohai (2002) correctly note that the algorithm estimators are inconsistent if the number of concentration steps k is finite, but consistency is not known if the concentration is iterated to convergence. So it is not known whether FAST-LTS is consistent.

Next consider a two stage estimator that has theory if an initial high breakdown consistent estimator such as LMS is used. Since estimators proven to be both high breakdown and consistent have $O(n^p)$ or higher complexity, this two stage estimator also has $O(n^p)$ or higher complexity. A common mistake is to claim, without proof, that the two stage estimator that is backed by theory can be computed with a practical initial estimator that has not been shown to be high breakdown. For example Maronna and Yohai (2002) and Maronna, Martin and Yohai (2006, p. 124) claim that a high breakdown consistent estimator can be iterated with a smooth objective function such as the S, MM or τ estimators. If the two stage estimator is backed by theory when the impractical LMS estimator is the initial estimator, then the practical implementation of the two stage estimator that uses the zero breakdown inconsistent `lmsreg` initial estimator is not backed by theory.

Some of the practical estimators, such as FAST-LTS, are useful for outlier detection. Also, a few practical robust estimators may eventually be proven to be high breakdown and consistent. For example, the Salibián-Barrera, Willems and Zamar (2008) FAST- τ estimator uses a \sqrt{n} consistent estimator and a high breakdown estimator as “starts” for iteration, but proving that the FAST- τ estimator is consistent will require impressive large sample theory.

Section 3 will develop a large class of practical \sqrt{n} consistent high breakdown `hblog`

estimators by using 3 attractors. One attractor will be a practical outlier resistant estimator that may not be backed by theory, such as FAST-LTS or FAST- τ . Another attractor will be a practical \sqrt{n} consistent estimator such as OLS while the third attractor will be a practical but inconsistent high breakdown estimator. The resulting **hbreg** estimator will be asymptotically equivalent to the consistent attractor, e.g. OLS.

Section 2 provides theory showing that many practical algorithm estimators for “high breakdown estimators” are inconsistent and zero breakdown. Section 4 demonstrates that the response and residual plots are useful for detecting outliers and leverage groups for linear models including multiple linear regression and many experimental design models.

2. THEORY FOR SOME PRACTICAL ESTIMATORS

The main point of this section is that the theory of the algorithm estimator depends on the theory of the attractors, not on the estimator corresponding to the criterion. The following theorem is due to Hawkins and Olive (2002) who show that if K randomly selected elemental starts are used with concentration to produce the attractors, then the resulting estimator is inconsistent and zero breakdown if K and k are fixed and free of n . Hence no matter how the attractor is chosen, the resulting estimator is not consistent. The proof uses results from He and Portnoy (1992) who show that if a start \mathbf{b} is a consistent estimator of $\boldsymbol{\beta}$, then the attractor is a consistent estimator of $\boldsymbol{\beta}$ with the same rate as the start. If the start is inconsistent, then so is the attractor.

If concentration is iterated to convergence so that k is not fixed, then it is not known whether the attractor is inconsistent. If k is fixed, Olive and Hawkins (2007) show that the “best attractor” that minimizes $\|\mathbf{b}_{k,j} - \boldsymbol{\beta}\|$ has rate $K_n^{1/p}$ if $K_n \rightarrow \infty$ as $n \rightarrow \infty$. The

“best attractor” is not an estimator since β is unknown.

Theorem 1. Suppose K is fixed. Then the elemental concentration algorithm and the elemental basic resampling algorithm estimators are zero breakdown. The estimators are inconsistent if k is fixed.

Note that each elemental start can be made to breakdown by changing one case. Hence the breakdown value of the final regression estimator is bounded by $K/n \rightarrow 0$ as $n \rightarrow \infty$. The classical estimator applied to a randomly drawn elemental set is an inconsistent estimator, so the K starts and the K attractors are inconsistent. Since the breakdown value of the FAST-LTS estimator is bounded by $500/n \rightarrow 0$, the claim that FAST-LTS can be used to compute LTS is false.

Suppose that there are K estimators $\hat{\beta}_j$ where K is fixed, and that $\hat{\beta}_A$ is an estimator obtained by choosing one of the K estimators as the final estimator. If each estimator is consistent with the same rate n^δ , then $\hat{\beta}_A$ is a consistent estimator of β with rate n^δ by Pratt (1959). On the other hand, if $P(\text{randomly selected attractor } \hat{\beta}_i \text{ gets arbitrarily close to } \beta) \rightarrow 0$ as $n \rightarrow \infty$, then $\hat{\beta}_A$ is inconsistent since none of the attractors gets arbitrarily close to β . This condition is slightly stronger than each attractor being inconsistent, but algorithms where all K of the attractors are inconsistent are untrustworthy. If the algorithm needs to use many attractors to achieve outlier resistance, then the individual attractors have little outlier resistance. Such estimators include elemental concentration algorithms, heuristic and genetic algorithms and projection algorithms.

The following theorem can be used to prove that the Olive (2005) MBA estimator is \sqrt{n} consistent, and the theorem is powerful because it does not depend on the criterion used to choose the attractor.

Theorem 2. Suppose the algorithm estimator chooses an attractor as the final estimator where there are K attractors and K is fixed.

i) If all of the attractors are consistent, then the algorithm estimator is consistent.

ii) If all of the attractors are consistent with the same rate, e.g., n^δ where $0 < \delta \leq 0.5$, then the algorithm estimator is consistent with the same rate as the attractors.

iii) If all of the attractors are high breakdown, then the algorithm estimator is high breakdown.

Proof. i) Choosing from K consistent estimators results in a consistent estimator. ii) follows from Pratt (1959). iii) Let $\gamma_{n,i}$ be the breakdown value of the i th attractor if the clean data are in general position. The breakdown value γ_n of the algorithm estimator can be no lower than that of the worst attractor: $\gamma_n \geq \min(\gamma_{n,1}, \dots, \gamma_{n,K}) \rightarrow 0.5$ as $n \rightarrow \infty$. \square

3. PRACTICAL HIGH BREAKDOWN REGRESSION ESTIMATORS

It is not hard to find an estimator that has high breakdown, though, as we shall see, the high breakdown property does not, in and of itself, mean that the estimator has much practical value. To explore this, Olive (2005) showed that $\hat{\beta}$ is high breakdown if the c_n th largest absolute residual $|r(\hat{\beta})|_{(c_n)}$ stays bounded under high contamination. This follows since, assuming general position and except in the degenerate case of exact fit, if $\|\hat{\beta}\| = \infty$, then $\text{median}(|r_i|) = \infty$, and conversely, if $\|\hat{\beta}\|$ is bounded then $\text{median}(|r_i|)$ is bounded if fewer than half of the cases are outliers.

Abusing notation slightly for ease of expression, we will refer to this c_n th largest order statistic of the residuals as the ‘median’. Let $Q_L(\hat{\beta}_H)$ denote the LMS, LTS or

LTA criterion for an estimator $\hat{\beta}_H$; therefore, the estimator $\hat{\beta}_H$ is high breakdown if and only if $Q_L(\hat{\beta}_H)$ is bounded for d_n near $n/2$.

The concentration operator refines an initial estimator by successively reducing the LTS criterion. If $\hat{\beta}_F$ refers to the final estimator obtained by applying concentration to some starting estimator $\hat{\beta}_H$ that is high breakdown, then since $Q_{LTS}(\hat{\beta}_F) \leq Q_{LTS}(\hat{\beta}_H)$, applying concentration to a high breakdown start results in a high breakdown attractor.

High breakdown estimators are, however, not necessarily useful for detecting outliers. Suppose $\gamma_n < 0.5$. On the one hand, if the \mathbf{x}_i are fixed, and the outliers are moved up and down parallel to the Y axis, then for high breakdown estimators, $\hat{\beta}$ and $\text{MED}(|r_i|)$ will be bounded. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large, suggesting that the high breakdown estimator is useful for outlier detection. On the other hand, if the Y_i 's are fixed at any values and the \mathbf{x} values perturbed, sufficiently large \mathbf{x} -outliers tend to drive the slope estimates to 0, not ∞ . For many estimators, including LTS, LMS and LTA, a cluster of Y outliers can be moved arbitrarily far from the bulk of the data but still, by perturbing their \mathbf{x} values, have arbitrarily small residuals.

A high breakdown estimator is easy to find. Recall that the median absolute deviation $\text{MAD}(w_i) = \text{median}(|w_i - \text{median}(w_i)|)$, and assume that the multiple linear regression model contains an intercept and that $\text{MAD}(Y_i)$ is finite. Make an OLS fit to the c_n cases whose Y values are closest to the median Y , and use this fit as the start for concentration. Write $\hat{\beta}_B$ for the final attractor attained in this way. Since the initial estimator has an LTS criterion value bounded by $n[\text{MAD}(Y_i)]^2$, it is high breakdown. And the subsequent concentration will further reduce this criterion, implying that $\hat{\beta}_B$

will also be high breakdown.

This start is reminiscent of literature proposals to make an initial OLS fit, find the c_n cases with the smallest residuals, and then use the OLS fit to these cases as the starting point for concentration. Note, however, that this proposal, unlike $\hat{\beta}_B$ is not guaranteed to have high breakdown.

With these preliminaries, we now define our high breakdown procedure. This is made up of three components:

- A practical estimator $\hat{\beta}_C$ that is consistent for clean data. Suitable choices would include the full-sample OLS and L1 estimators.
- A practical estimator $\hat{\beta}_A$ that is effective for outlier identification. Suitable choices would include the conventionally-implemented `lmsreg` or FAST-LTS estimators.
- A practical high-breakdown estimator such as $\hat{\beta}_B$.

By selecting one of these three estimators according to the features each of them uncovers in the data, we may inherit the good properties of each of them. Specifically, our proposed `hbreg` estimator $\hat{\beta}_H$ is defined as follows. Pick a constant $a > 1$ and find the smallest of the three scaled criterion values $Q_L(\hat{\beta}_C)$, $aQ_L(\hat{\beta}_A)$, $aQ_L(\hat{\beta}_B)$. According to which of the three estimators attains this minimum, set $\hat{\beta}_H$ to $\hat{\beta}_C$, $\hat{\beta}_A$ or $\hat{\beta}_B$ respectively.

Large sample theory for `hbreg` is simple and given in the following theorem. Let $\hat{\beta}_L$ be the LMS, LTS or LTA estimator that minimizes the criterion Q_L . Note that the impractical estimator $\hat{\beta}_L$ is never computed.

Theorem 3. Suppose that both $\hat{\beta}_L$ and $\hat{\beta}_C$ are consistent estimators of β where the

regression model contains a constant. Then the **hbreg** estimator $\hat{\beta}_H$ is high breakdown and asymptotically equivalent to $\hat{\beta}_C$.

Proof. Since the clean data are in general position and $Q_L(\hat{\beta}_H) \leq aQ_L(\hat{\beta}_B)$ is bounded for γ_n near 0.5, the **hbreg** estimator is high breakdown. Let $Q_L^* = Q_L$ for LMS and $Q_L^* = Q_L/n$ for LTS and LTA. As $n \rightarrow \infty$, consistent estimators $\hat{\beta}$ satisfy $Q_L^*(\hat{\beta}) - Q_L^*(\beta) \rightarrow 0$ in probability. Since LMS, LTS and LTA are consistent and the minimum value is $Q_L^*(\hat{\beta}_L)$, it follows that $Q_L^*(\hat{\beta}_C) - Q_L^*(\hat{\beta}_L) \rightarrow 0$ in probability, while $Q_L^*(\hat{\beta}_L) < aQ_L^*(\hat{\beta})$ for any estimator $\hat{\beta}$. Thus with probability tending to one as $n \rightarrow \infty$, $Q_L(\hat{\beta}_C) < a \min(Q_L(\hat{\beta}_A), Q_L(\hat{\beta}_B))$. Hence $\hat{\beta}_H$ is asymptotically equivalent to $\hat{\beta}_C$. \square

The family of **hbreg** estimators is enormous and depends on the practical high breakdown estimator, $\hat{\beta}_C$, $\hat{\beta}_A$, a and on the criterion Q_L . Note that the theory needs the error distribution to be such that both $\hat{\beta}_C$ and $\hat{\beta}_L$ are consistent. Sufficient conditions for LMS, LTS and LTA to be consistent are rather strong. To have reasonable sufficient conditions for the **hbreg** estimator to be consistent, $\hat{\beta}_C$ should be consistent under weak conditions. Hence OLS is a good choice that results in 100% asymptotic Gaussian efficiency.

We suggest using the LTA criterion since in simulations, **hbreg** behaved like $\hat{\beta}_C$ for smaller sample sizes than those needed by the LTS and LMS criteria. Want a near 1 so that **hbreg** has outlier resistance similar to $\hat{\beta}_A$, but want a large enough so that **hbreg** performs like $\hat{\beta}_C$ for moderate n on clean data. Simulations suggest that $a = 1.4$ is a reasonable choice.

There are at least three reasons for using $\hat{\beta}_B$ as the high breakdown estimator. First, $\hat{\beta}_B$ is high breakdown and simple to compute. Second, the fitted values roughly track

Table 1: MEAN $\hat{\beta}_i$ and SD($\hat{\beta}_i$)

| n | method | mn or sd | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ | $\hat{\beta}_5$ | |
|------|--------|----------|-----------------|-----------------|-----------------|-----------------|-----------------|--------|
| 25 | HB | mn | 0.9921 | 0.9825 | 0.9989 | 0.9680 | 1.0231 | |
| | | sd | 0.4821 | 0.5142 | 0.5590 | 0.4537 | 0.5461 | |
| | OLS | mn | 1.0113 | 1.0116 | 0.9564 | 0.9867 | 1.0019 | |
| | | sd | 0.2308 | 0.2378 | 0.2126 | 0.2071 | 0.2441 | |
| | ALTS | mn | 1.0028 | 1.0065 | 1.0198 | 1.0092 | 1.0374 | |
| | | sd | 0.5028 | 0.5319 | 0.5467 | 0.4828 | 0.5614 | |
| | BB | mn | 1.0278 | 0.5314 | 0.5182 | 0.5134 | 0.5752 | |
| | | sd | 0.4960 | 0.3960 | 0.3612 | 0.4250 | 0.3940 | |
| | 400 | HB | mn | 1.0023 | 0.9943 | 1.0028 | 1.0103 | 1.0076 |
| | | | sd | 0.0529 | 0.0496 | 0.0514 | 0.0459 | 0.0527 |
| OLS | | mn | 1.0023 | 0.9943 | 1.0028 | 1.0103 | 1.0076 | |
| | | sd | 0.0529 | 0.0496 | 0.0514 | 0.0459 | 0.0527 | |
| ALTS | | mn | 1.0077 | 0.9823 | 1.0068 | 1.0069 | 1.0214 | |
| | | sd | 0.1655 | 0.1542 | 0.1609 | 0.1629 | 0.1679 | |
| BB | | mn | 1.0184 | 0.8744 | 0.8764 | 0.8679 | 0.8794 | |
| | | sd | 0.1273 | 0.1084 | 0.1215 | 0.1206 | 0.1269 | |

the bulk of the data. Lastly, although $\hat{\beta}_B$ has rather poor outlier resistance, $\hat{\beta}_B$ does perform well on several outlier configurations where some common alternatives fail. See the first three examples in Section 4.

Next we will show that the **hbreg** estimator implemented with $a = 1.4$ using Q_{LTA} , $\hat{\beta}_C = \text{OLS}$ and $\hat{\beta}_B$ can greatly improve the estimator $\hat{\beta}_A$. We will use $\hat{\beta}_A = \text{lbsreg}$ in *R* and *Splus 2000*. Depending on the implementation, the **lbsreg** estimators use the elemental resampling algorithm, the elemental concentration algorithm or a genetic algorithm. Coverage is 50%, 75% or 90%. The *Splus 2000* implementation is an unusually poor genetic algorithm with 90% coverage. The *R* implementation appears to be the zero breakdown inconsistent elemental basic resampling algorithm that uses 50% coverage.

Simulations were run in *R* with the x_{ij} (for $i > 1$) and e_i iid $N(0, \sigma^2)$ and $\beta = \mathbf{1}$, the $p \times 1$ vector of ones. Then $\hat{\beta}$ was recorded for 100 runs. The mean and standard deviation of the $\hat{\beta}_j$ were recorded for $j = 1, \dots, p$. For $n \geq 10p$ and OLS, the vector of means should be close to $\mathbf{1}$ and the vector of standard deviations should be close to $\mathbf{1}/\sqrt{n}$. The \sqrt{n} consistent high breakdown **hbreg** estimator performed like OLS if $n \approx 35p$ and $2 \leq p \leq 6$, if $n \approx 20p$ and $7 \leq p \leq 14$, or if $n \approx 15p$ and $15 \leq p \leq 40$. See Table 1 for $p = 5$ and 100 runs. ALTS denotes **lbsreg**, HB denotes **hbreg** and BB denotes $\hat{\beta}_B$.

4. PLOTS FOR OUTLIER DETECTION

Consider the linear model $Y_i = \mathbf{x}_i^T \beta + e_i$ where $i = 1, \dots, n$. Multiple linear regression and many experimental design models are linear models. Let the *iid error model* be the linear model where the zero mean constant variance errors are iid from a unimodal distribution that is not highly skewed. Then the zero mean assumption $E(e_i) \equiv 0$ holds

without loss of generality if the linear model contains a constant.

Huber and Ronchetti (2009, p. 154) note that efficient identification of outliers and leverage groups “is an open, perhaps unsolvable, diagnostic problem.” Such groups are often difficult to detect with residuals and regression diagnostics, but often have outlying fitted values and responses. The OLS fit often passes through a cluster of outliers, causing a large gap between a cluster corresponding to the bulk of the data and the cluster of outliers. When such a gap appears, it is possible that the smaller cluster corresponds to good leverage points: the cases follow the same model as the bulk of the data. Fit the model to the bulk of the data. If the fit passes through the cluster, then the cases may be good leverage points, otherwise they may be outliers.

Olive (2005) suggests using residual, response, RR and FF plots to detect outliers. The residual plot is a plot of fitted values versus the residuals while the response plot uses fitted values versus the response. An RR plot is a scatterplot matrix of the residuals from several estimators. An FF plot replaces the residuals by the fitted values and includes the response on the top or bottom row of the scatterplot matrix, giving the response plots of the different estimators. The four plots are best for $n > 5p$.

Under the iid error model, if the fitted values take on many values, then the plotted points should scatter about the identity line with unit slope and zero intercept or about the $r = 0$ line in a roughly evenly populated band in the response and residual plots, respectively. Deviations from the evenly populated band suggest that something is wrong with the iid error model. Response and residual plots are very effective for suggesting that something is wrong with the iid error model. The plots often show two or more groups of data, and outliers often cause an obvious tilt in the residual plot.

In the following three examples, it is shown that $\hat{\beta}_B$ is sometimes useful for detecting outliers where some competing methods fail. The *Splus 2000* `ltsreg` estimator is used as $\hat{\beta}_A$ in `hblog`. Influence diagnostics such as Cook’s distances CD_i from Cook (1977) and the weighted Cook’s distances WCD_i from Peña (2005) are also sometimes useful. In the following example, cases in the plots with $CD_i > \min(0.5, 2p/n)$ are highlighted with open squares, and cases with $|WCD_i - \text{median}(WCD_i)| > 4.5\text{MAD}(WCD_i)$ are highlighted with crosses.

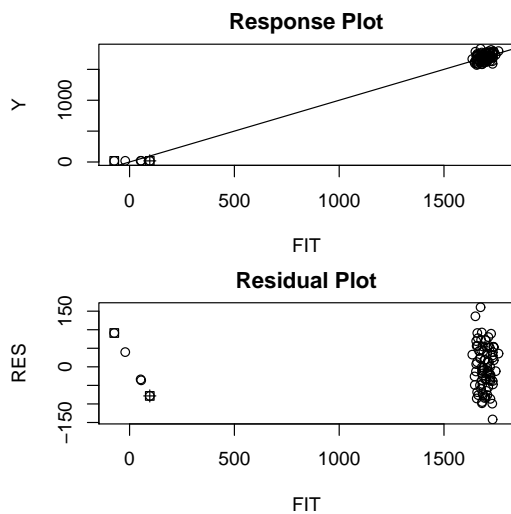


Figure 1: Buxton Data

Example 1. The LMS, LTA and LTS estimators are determined by a “narrowest band” covering half of the cases. Hawkins and Olive (2002) suggest that the fit will pass through outliers if the band through the outliers is narrower than the band through the clean cases. This behavior tends to occur if regression relationship is weak, and if there is a tight cluster of outliers where $|Y|$ is not too large. As an illustration, Buxton (1920, p. 232-5) gives 20 measurements of 88 men. Consider predicting *stature* using an intercept, *head length*, *nasal height*, *bigonal breadth*, and *cephalic index*. One case was deleted since

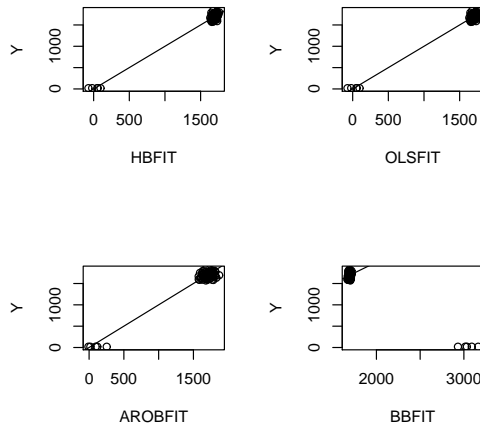


Figure 2: Response Plots for Buxton Data

it had missing values. Five individuals, numbers 61-65, were reported to be about 0.75 inches tall with head lengths well over five feet! The OLS fit to the clean data makes the absolute residuals of the outliers large.

In Figure 1, notice that the OLS fit to all of the data passes through the outliers, but the response plot is resistant to Y -outliers since Y is on the vertical axis. Also notice that only two of the outliers had large Cook's distance and only one case had a large WCD_i . Figure 2 shows the response plots for OLS, $ltsreg$, $hbreg$ and $\hat{\beta}_B$. Notice that only the fit from $\hat{\beta}_B$ (BBFIT) did not pass through the outliers. There are always outlier configurations where an estimator will fail, and $hbreg$ should fail on configurations where LTA, LTS and LMS would fail.

The CD_i and WCD_i are the most effective when there is a single cluster about the identity line. If there is a second cluster of outliers or good leverage points or if there is nonconstant variance, then these numerical diagnostics tend to fail.

Example 2. Portnoy (1987) gives an artificial data set with nine predictors and $n = 50$.

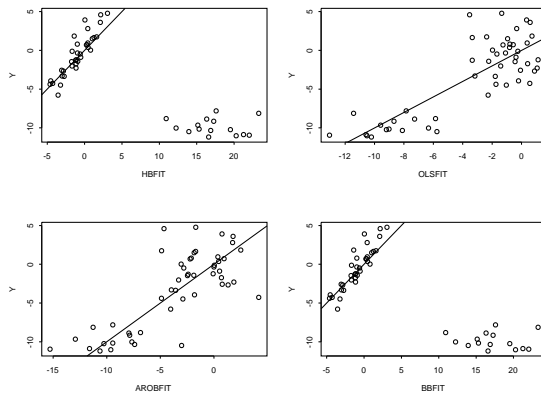


Figure 3: Response Plots for Portnoy Data

The fifteen outliers are cases 36-50. Figure 3 shows the response plots for OLS, `ltsreg`, `hbregr` and $\hat{\beta}_B$. Note that the OLS response plot shows two groups of data better than the `ltsreg` response plot. For this data set, `hbregr` is better than `ltsreg` since the outliers have massive absolute residuals in the `hbregr` response plot.

Example 3. There are data sets where the OLS response and residual plots fail. Rousseeuw and Leroy (1987, pp. 242-245) give a modified wood data set with 4 nontrivial predictors and 4 planted outliers. Figure 4 shows an FF plot for the data, using OLS, least absolute deviations (L_1), `lmsreg` (ALMS), `ltsreg` (ALTS), MBA and $\hat{\beta}_B$ (BB). The response plots are on the top row of the FF plot. The four planted outliers have the smallest values of the response, and can not be detected by the OLS response and residual plots. They can be detected by the ALMS, MBA and $\hat{\beta}_B$ response plots. For this data set, `hbregr` fails if $\hat{\beta}_A = \text{ltsreg}$, but `hbregr` works if $\hat{\beta}_A = \text{MBA}$.

As illustrated in the last three examples, there are several common outlier configurations where $\hat{\beta}_B$ is useful for outlier detection, but there are many outlier configurations where $\hat{\beta}_B$ fails. A good `hbregr` estimator should use a good outlier resistant estimator for

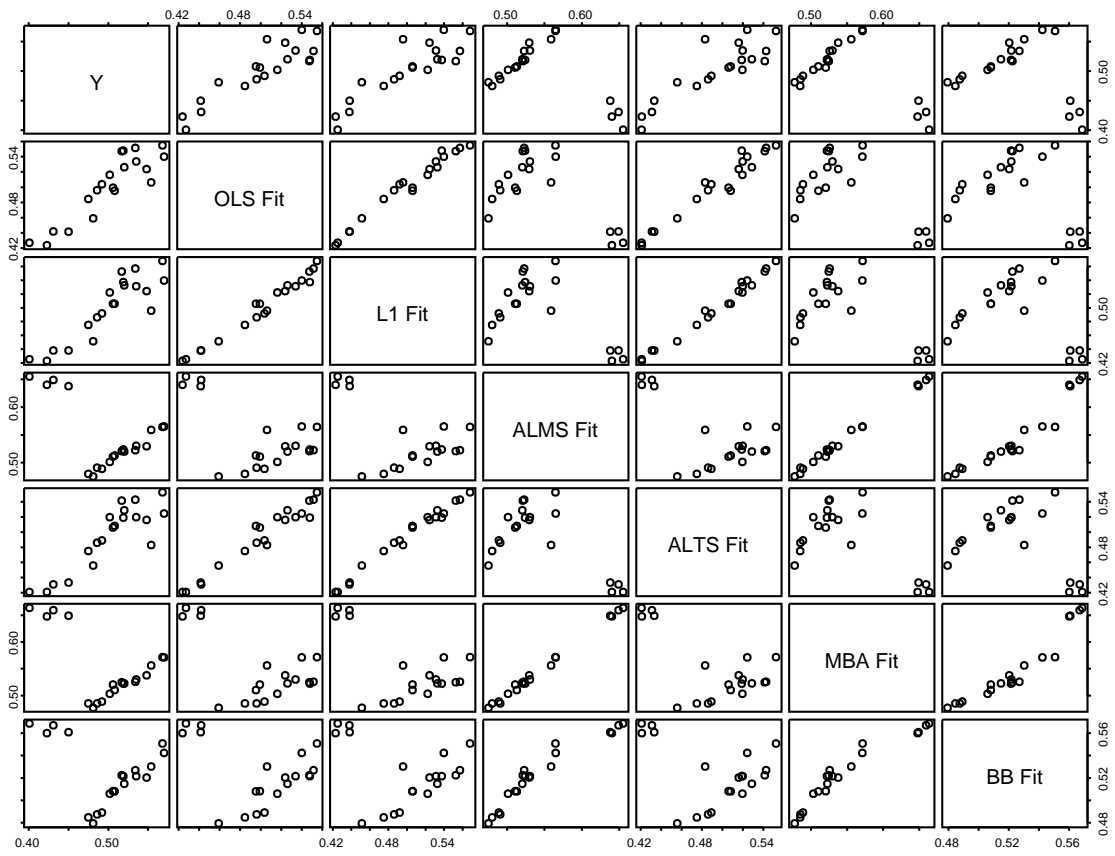


Figure 4: FF Plot for Wood Data

$\hat{\beta}_A$ such as MBA, FAST- τ , or FAST-LTS with 50% coverage. On the other hand, `hbg` greatly improves estimators like `ltsreg`.

Note that Figures 1–4 all display the regression data with response plots. The next two examples show that response and residual plots are also useful for outlier detection for experimental design models.

Example 4. Dunn and Clark (1974, p. 129) study the effects of four fertilizers on wheat yield using a Latin square design. The row blocks were 4 types of wheat, and the column blocks were 4 plots of land. Each plot was divided into 4 subplots. Case 14 had a yield of 64.5 while the next highest yield was 35.5. For the response plot in Figure 5, note that both Y and \hat{Y} are large for the high yield. Also note that \hat{Y} underestimates Y by about 10 for this case.

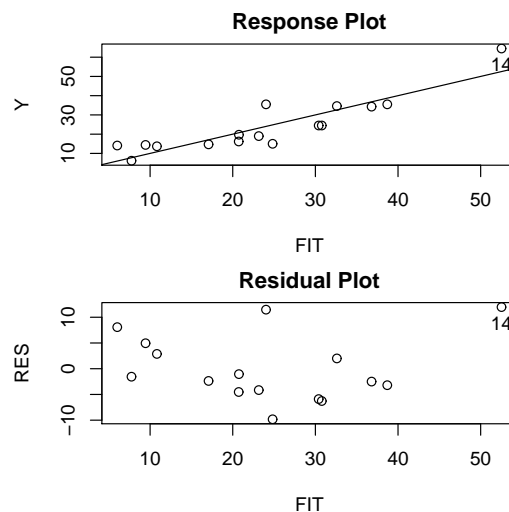


Figure 5: Latin Square Data

Example 5. Snedecor and Cochran (1967, p. 300) give a data set with 5 types of soybean seed. The response frate = number of seeds out of 100 that failed to germinate. Five blocks were used. The response and residual plots in Figure 6 suggest that case 5 is

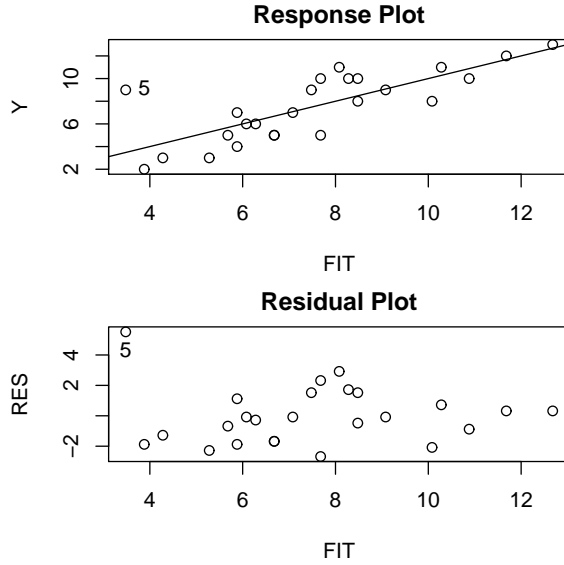


Figure 6: One Way Block Design Does Not Fit All of the Data

not fit well by the model. On further examination of the data, there seems to be a block treatment interaction, which is not allowed by the completely randomized block design.

5. CONCLUDING REMARKS

This paper discusses two important problems: i) finding outliers and high leverage cases for linear models and ii) deriving practical \sqrt{n} consistent high breakdown regression estimators.

The OLS response and residual plots are useful for detecting outliers and high leverage cases. Using the two plots speeds up the process of finding a linear model that is a useful approximation of the data, and both plots should be made before performing inference. It is important to provide researchers tools that they are willing to use, and researchers who use the residual plot to visualize $e|\beta^T \mathbf{x}$ should be willing to use the response plot to visualize $Y|\beta^T \mathbf{x}$. Cases with large “leave one out” diagnostics, such as Cook’s distances, can be highlighted in the plots.

For multiple linear regression, the RR and FF plots can be useful for detecting outliers and high leverage cases when the two OLS plots fail. The MBA estimator that uses the Olive and Hawkins (1999) LATA criterion is also useful. The fit from this estimator often tilts away from both “good” and “bad” high leverage cases. Then the clean data and one or two groups of high leverage cases can be seen in the response plot.

Note that `hbreg` technique simultaneously increases the outlier resistance of the classical estimator $\hat{\beta}_C$ while modifying the outlier resistant estimator $\hat{\beta}_A$ so that the resulting robust estimator is backed by theory. Hence the `hbreg` technique robustifies both a classical estimator and a practical outlier resistant estimator that has no large sample theory. The outlier resistance of any practical outlier resistant estimator is likely to decrease rapidly as the number of predictors p increases.

Plots and simulations were done in *R* and *Splus*. See R Development Core Team (2008). Programs are in the collection of functions *rpack.txt* available from (www.math.siu.edu/olive/rpack.txt). From *rpack*, the function `hbreg` computes the `hbreg` estimator, the function `hbregsim` can be used to reproduce the simulation, and the function `hbplot` can be used to make four response plots as in Figure 2. For these three functions, $\hat{\beta}_A$ can be MBA, `ltsreg` or MBA using the LATA criterion. *Splus* functions `ffplot` and `rrplot` and *R* functions `ffplot2` and `rrplot2` can be used to make RR and FF plots. In *R*, the function `mlrplot4` can be used as in Figure 1 to make the response and residual plots where cases with large Cook’s distances and large WCD_i are highlighted. MBA is computed with `mbareg`, and MBA using the LATA criterion is computed with `mbalata`.

REFERENCES

- Bernholt, T. (2005), “Computing the Least Median of Squares Estimator in Time $O(n^d)$,” *Proceedings of ICCSA 2005*, LNCS, 3480, 697-706.
- Buxton, L. H. D. (1920), “The Anthropology of Cyprus,” *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, 50, 183-235.
- Čížek, P. (2006), “Least Trimmed Squares Under Dependence,” *Journal of Statistical Planning and Inference*, 136, 3967-3988.
- Čížek, P., (2008), “General Trimmed Estimation: Robust Approach to Nonlinear and Limited Dependent Variable Models,” *Econometric Theory*, 24, 1500-1529.
- Cook, R. D. (1977), “Deletion of Influential Observations in Linear Regression,” *Technometrics*, 19, 15-18.
- Dunn, O. J., and Clark, V. A. (1974), *Applied Statistics: Analysis of Variance and Regression*, New York, NY: Wiley.
- 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., and Portnoy, S. (1992), “Reweighted LS Estimators Converge at the Same Rate

- as the Initial Estimator,” *The Annals of Statistics*, 20, 2161-2167.
- Huber, P. J., and Ronchetti, E. M. (2009), *Robust Statistics*, 2nd ed., Hoboken, NJ: Wiley.
- Hubert, M., Rousseeuw, P.J., and Van Aelst, S. (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, 151-153.
- Hubert, M., Rousseeuw, P. J., and Van Aelst, S. (2008), “High Breakdown Multivariate Methods,” *Statistical Science*, 23, 92-119.
- Kim, J., and Pollard, D. (1990), “Cube Root Asymptotics,” *The Annals of Statistics*, 18, 191-219.
- Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, Hoboken, NJ: Wiley.
- 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. (2005), “Two Simple Resistant Regression Estimators,” *Computational Statistics and Data Analysis*, 49, 809-819.
- Olive, D., and Hawkins, D.M. (1999), “Comment on ‘Regression Depth’ by P.J. Rousseeuw and M. Hubert,” *Journal of the American Statistical Association*, 94, 416-417.
- Olive, D. J., and Hawkins, D. M. (2007), “Behavior of Elemental Sets in Regression,” *Statistics & Probability Letters*, 77, 621-624.
- Peña, D. (2005), “A New Statistic for Influence in Regression,” *Technometrics*, 47, 1-12.

- Portnoy, S. (1987), "Using Regression Quantiles to Identify Outliers," in *Statistical Data Analysis Based on the L1 Norm and Related Methods*, ed. Dodge, Y., North Holland, Amsterdam, 345-356.
- Pratt, J. W. (1959), "On a General Concept of 'in Probability'," *The Annals of Mathematical Statistics*, 30, 549-558.
- R Development Core Team (2008), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria (www.R-project.org).
- Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York, NY: Wiley.
- Rousseeuw, P. J., Van Aelst, S., and Hubert, M. (1999), "Rejoinder to Discussion of 'Regression Depth'," *Journal of the American Statistical Association*, 94, 419-433.
- Rousseeuw, P. J., and Van Driessen, K. (2006), "Computing LTS Regression for Large Data Sets," *Data Mining and Knowledge Discovery*, 12, 29-45.
- Salibian-Barrera, M., Willems, G., and Zamar, R. H. (2008), "The Fast τ -Estimator of Regression," *Journal of Computational and Graphical Statistics*, 17, 659-682.
- Snedecor, G. W., and Cochran, W. G. (1967), *Statistical Methods*, 6th ed., Ames, IA: Iowa State College Press.