12-1999

Applications and Algorithms for Least Trimmed Sum of Absolute Deviations Regression

Douglas M. Hawkins
University of Minnesota - Twin Cities

David Olive
Southern Illinois University Carbondale, dolive@math.siu.edu

Published in *Computational Statistics & Data Analysis* 32, 119-134.

Applications and Algorithms for Least Trimmed Sum of Absolute Deviations Regression

Douglas M. Hawkins and David Olive*

University of Minnesota

August 3, 2003

Abstract

High breakdown estimation (HBE) addresses the problem of getting reliable parameter estimates in the face of outliers that may be numerous and badly placed. In multiple regression, the standard HBE's have been those defined by the least median of squares (LMS) and the least trimmed squares (LTS) criteria. Both criteria lead to a partitioning of the data set's n cases into two "halves" – the covered "half" of cases are accommodated by the fit, while the uncovered "half", which is intended to include any outliers, are ignored. In LMS, the criterion is the Chebyshev norm of the residuals of the covered cases, while in LTS the criterion is the sum of squared residuals of the covered cases. Neither LMS nor LTS is

*Douglas M. Hawkins is Professor and David Olive is Visiting Assistant Professor, School of Statistics, University of Minnesota, St. Paul, MN 55108, U S A. The authors are grateful to the editors and referees for a number of helpful suggestions for improvement in the article.

entirely satisfactory. LMS has a statistical efficiency of zero if the true residuals are normal, and so is unattractive, particularly for large data sets. LTS is preferable on efficiency grounds, but its exact computation turns out to involve an intolerable computational load in any but quite small data sets.

The criterion of least trimmed sum of absolute deviations (LTA) is found by minimizing the sum of absolute residuals of the covered cases. We show in this article that LTA is an attractive alternative to LMS and LTS, particularly for large data sets. It has a statistical efficiency that is not much below that of LTS for outlier-free normal data and better than LTS for more peaked error distributions. As its computational complexity is of a lower order than LMS and LTS, it can also be evaluated exactly in much larger samples than either LMS or LTS. Finally, just as its full-sample equivalent, the L1 norm, is robust against outliers on low leverage cases, LTA is able to cover larger subsets than LTS in those data sets where not all outliers are on high leverage cases.

For samples too large for exact evaluation of the LTA, we outline a "feasible solution algorithm", which provides excellent approximations to the exact LTA solution using quite modest computation.

KEY WORDS: High Breakdown; Least Median of Squares; Least Trimmed Sum of Squares; Missing Values; Outliers; Robust Estimation; L1 Norm.

1 INTRODUCTION

Consider the Gaussian regression model

$$Y = X\beta + \epsilon \tag{1.1}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, and ϵ is an $n \times 1$ vector of errors. The ith case (y_i, x_i^T) corresponds to the ith element y_i of $\boldsymbol{Y}$ and the ith row x_i^T of $\boldsymbol{X}$. We will also consider models with "clean" cases and contaminated cases.

When we have a subsample

$$J_i = \{j_1, ..., j_h\}$$

of size $h \geq p$ of the original data, by applying any convenient fitting criterion to the data (Y_{J_i}, X_{J_i}), we can obtain an estimator b_{J_i} of β. Possible criteria include ordinary least squares (OLS), the Chebyshev (minimum maximum absolute deviation) norm, and the L1 norm. To compute the criterion $Q(b_{J_i})$, we need the n residuals

$r_1(b_{J_i}), \ldots, r_n(b_{J_i})$ where

$$r_k(b_{J_i}) = y_k - x_k^T b_{J_i}, \tag{1.2}$$

and these three criteria aim to minimize in b

$LTS : \sum_{i=1}^{h} |r(b)|_{(i)}^2$

$LMS : |r(b)|_{(h)}$

$LTA : \sum_{i=1}^{h} |r(b)|_{(i)}$

where $|r(b)|_{(i)}$ is the ith smallest absolute residual from fit b. LMS and LTS were proposed by Rousseeuw (1984) and LTA by Bassett (1991), Hössjer (1991, 1994) and Tableman

(1994a,b).

It is conventional to set $h = [(n + p + 1)/2]$, a choice that maximizes the breakdown of the resulting estimator. It is frequently valuable to use larger values of h (for example in data sets where large numbers of outliers are unlikely and we want to get the benefit of statistical efficiency from covering more cases); and to explore the fits for a range of values of h.

Finding the LMS, LTS or LTA estimator leads to a two-stage problem – identifying the "best" subset of size h to cover; and then finding the Chebyshev, OLS or L1 fit to this subset. In general, there is no completely reliable method other than full enumeration to identify the "best" subset to cover (that is, the subset whose fit criterion will be the smallest among all possible subsets of size h), and so computing any of these HBE's involves a substantial combinatorial problem. We will show that, while LTA also involves a combinatorial search, it is smaller than that required for LMS and far smaller than that required for LTS. This much smaller computational requirement, in part, motivates a closer consideration of the statistical properties of LTA.

2 THE LTA ESTIMATOR

Both LTA and LTS involve the parameter h, the number of "covered" cases. The remaining $n-h$ cases, by being ignored, are "trimmed". If $h = h_n$ is a sequence of integers such that $h/n \to \gamma$, then $1 - \gamma$ is the approximate amount of trimming. The LTA(γ)

estimator $\hat{\beta}_{LTA}$ is the fit that minimizes

$$Q_{LTA}(b) = \sum_{i=1}^{h} |r(b)|_{(i)} \tag{2.1}$$

where $|r(b)|_{(i)}$ is the ith smallest absolute residual from fit b. Several authors have examined the LTA estimator in the location model (a model including an intercept, but no nontrivial predictors). For the location model, Bassett(1991) gives an algorithm, and Tableman (1994a,b) derives the influence function and asymptotics. In the regression model, LTA is a special case of the R-estimators of Hössjer (1991, 1994).

2.1 Breakdown and Bias of LTS, LMS, and LTA

The three estimators LTS(γ), LMS(γ) and LTA(γ) have breakdown value

$$\min(1-\gamma, \gamma).$$

See Hössjer (1994, p. 151). Breakdown proofs in Rousseeuw and Bassett (1991) and Niinimaa, Oja, and Tableman (1990) could also be modified to give the result. Yohai and Zamar (1993, p. 1832 for LTA) show that LTS, LMS, and LTA have finite maximum asymptotic bias when the contamination proportion is less than $1-\gamma$ where $0.5 < \gamma < 1$. Croux, Rousseeuw, and Van Beal (1996, p. 219) show that the maxbias curve of LTA is lower than that of LTS.

2.2 Asymptotic variances of LTA and LTS

Many regression estimators $\hat{\beta}$ satisfy

$$\sqrt{n}(\hat{\beta} - \beta) \to N(0, V(\hat{\beta}, F)\ W)$$

where

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

For example, Koenker and Bassett (1978) and Bassett and Koenker (1978) show that the L1 estimator has greater statistical efficiency than OLS for more "peaked" distributions in which $f(0) > 1/(2\sigma)$ where $\sigma^2 = \mathrm{Var}(\epsilon_i)$.

Let the zero median error distribution F be continuous and strictly increasing on its interval support with a symmetric, unimodal, density f. Also assume that f is differentiable on $(0, \infty)$. Then the conjectured asymptotic variance of LTS(γ) is

$$V(LTS(\gamma), F) = \frac{\int_{F^{-1}(1/2-\gamma/2)}^{F^{-1}(1/2+\gamma/2)} x^2 dF(x)}{[\gamma - 2F^{-1}(1/2+\gamma/2) f(F^{-1}(1/2+\gamma/2))]^2}. \tag{2.2}$$

See Rousseeuw and Leroy (1987, p. 180, p. 191), Tableman (1994a, p. 337), and remark 2.7 of Stromberg, Hawkins, and Hössjer (1997).

Combining Tableman(1994b, p. 392) with Hössjer (1994, p. 150) leads to the conjecture that the asymptotic variance for LTA(γ) is

$$V(LTA(\gamma), F) = \frac{\gamma}{4[f(0) - f(F^{-1}(1/2+\gamma/2))]^2}. \tag{2.3}$$

Rigorous proofs for these conjectures have only been given in the location model - see Tableman (1994b) and Butler (1982). As $\gamma \to 1$, the efficiency of LTS approaches that of OLS and the efficiency of LTA approaches that of L1. The results of Oosterhoff (1994) suggest that when $\gamma = 0.5$, LTA will be more efficient than LTS only for sharply peaked distributions such as the double exponential; we will explore this issue below.

The normal case. At the standard normal

$$V(LTS(\gamma), \Phi) = \frac{1}{\gamma - 2k\phi(k)} \tag{2.4}$$

while

$$V(LTA(\gamma), \Phi) = \frac{\gamma}{4[\phi(0) - \phi(k)]^2} \tag{2.5}$$

where ϕ is the standard normal pdf and

$$k = \Phi^{-1}(0.5 + \gamma/2).$$

The double-exponential case. For a double exponential DE(0,1) random variable,

$$V(LTS(\gamma), DE(0,1)) = \frac{2 - (2 + 2k + k^2)\exp(-k)}{[\gamma - k\ \exp(-k)]^2}$$

while

$$V(LTA(\gamma), DE(0,1)) = \frac{1}{\gamma}$$

where $k = -\log(1-\gamma)$. Note that LTA(0.5) and OLS have the same asymptotic efficiency at the double exponential distribution.

The Cauchy case. Since the Cauchy distribution has infinite variance, so does the OLS estimator, though the full-sample L1 estimator and the trimmed estimators have finite variance. Hence

$$V(LTS(\gamma), C(0,1)) = \frac{2k - \pi\gamma}{\pi[\gamma - \frac{2k}{\pi(1+k^2)}]^2}$$

and

$$V(LTA(\gamma), C(0,1)) = \frac{\gamma}{4[\frac{1}{\pi} - \frac{1}{\pi(1+k^2)}]^2}$$

where $k = \tan(\pi\gamma/2)$. The LTA sampling variance converges to a finite value as $\gamma \to 1$ while that of LTS increases without bound. LTS(0.5) is slightly more efficient than LTA(0.5), but LTA pulls ahead of LTS as the amount of trimming is reduced.

We simulated LTA and LTS for the location model (that is, an intercept but no non-trivial predictors) using the above three models. For the location model, computation of the estimators is easy and fast. Find the order statistics $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$ of the data, and evaluate the variance (for LTS) and the sum of absolute deviations from the median (for LTA) of each of the $n - h + 1$ half-samples $Y_{(i)}, \ldots, Y_{(i+h-1)}$, for $i = 1, \ldots, n - h + 1$. The minimum across these half-samples then defines the LTA and LTS estimates.

We computed the sample standard deviations (SD) of the resulting location estimates from 1000 runs of each sample size studied. Tables 1, 2, and 3 list $\sqrt{n}$ SD from the simulations. The entry $n = \infty$ lists the asymptotic standard deviation multiplied by $\sqrt{n}$. Table 4 shows the Monte Carlo OLS relative efficiencies. The finite-sample variance of LTS is known to converge to the asymptotic limit very slowly when the errors are Gaussian, and this tendency is evident in table 1. The rate of convergence for C(0,1) data and DE(0,1) data seems to be faster. LTA exhibits similar behavior.

3 COMPUTATION OF THE LTA ESTIMATOR

One of the practical attractions of the LTA estimator is the relative ease (compared with the LTS or even the LMS estimator) with which it can be computed. The LTA estimator has the property that it is an L1 fit to some subset of size h of the data, the subset being

those cases for which this L1 norm is minimized. As an L1 regression corresponds to an exact fit to some subset of size p, the LTA is similarly characterized as a two-part problem - identifying the correct subset of size h to cover with the LTA fit, and determining the subset of size p that minimizes the L1 norm of the fit to these h cases. Denote the number of subsets of size h from a sample of size n by nCh. There are nCp "elemental" subsets (subsets of size p), – a much smaller number than nCh in typical applications – and one of these must provide an LTA solution for the full data set. By reversing the order of the two-part search therefore, we can dramatically reduce its computational complexity.

Exact fits to subsets of size p have a special place in the area of high breakdown estimation - these elemental sets have long been used to generate approximations to other high breakdown fit criteria such as LMS or LTS (Rousseeuw and Leroy 1987). In the case of LTA though they yield exact solutions and not just approximations.

3.1 Exact calculation of the LTA

The characterization of the LTA as an elemental regression parallels that of the LMS fit (which is a Chebyshev fit to a suitably chosen subset of size $p+1$ - Portnoy (1987)), and leads to an LTA counterpart to Stromberg's (1993) exact algorithm for LMS:-

- Generate every elemental set.
- For each elemental set, compute the exactly fitting regression function, and get residuals on all cases in the data set.
- Find the sum of the smallest h absolute values among these residuals.

- The LTA is given by the elemental fit for which this sum is smallest.

The MVELMS code of Hawkins and Simonoff (1993) contains a provision for generating all possible elemental sets. While this code was written to implement an approximate LMS fit, all that is needed to convert it into an LTA code is a change to the calculation of the criterion value, changing this from the h^{th} smallest absolute order statistic to the sum of the h smallest absolute values.

This exhaustive algorithm requires the generation of all nCp subsets of size p. From the viewpoint of computational complexity, it is thus inherently smaller than Stromberg's exact algorithm for LMS, which requires the generation of all $nC(p+1)$ subsets of size $p+1$. The ratio of these numbers of subsets, $(p+1)/(n-p)$ is substantial for moderate p, particularly with large n.

Table 5 shows some values of nCp. To gain some feeling for these numbers, evaluating a million regressions is quite a small computation on a desktop personal computer, but a billion is excessive. So exhaustive enumeration to get the exact LTA fit is a modest computation for all the n values listed with a simple linear regression; for $n \leq 200$ with $p = 3$, $n < 100$ with $p = 4$ and $n < 50$ with $p = 5$. These maximum sample sizes are larger than the small text-book size range often discussed in writings on HBE methods.

Exact algorithms for LTA, LMS and LTS all comprise generating all subsets of cases of some appropriate size, performing a fit to these cases, and evaluating the fit on all data in the sample. The subset sizes and the type of fit are:-

Criterion	Subset size	Fit type	Number of possible subsets
LTA	p	Exact	nCp
LMS	p+1	Chebyshev	nC(p+1)
LTS	h	Least squares	nCh

Croux, Rousseeuw, and Hössjer (1994) propose an exact algorithm for the least quantile of differences (LQD) estimator using the fact that LQD is just LMS applied to the set of case differences.

For data sets of interesting size, the number of subsets required for exact evaluation of the LTA is far smaller than those required for either LMS or LTS. For example, if $n = 100$ and $p = 4$, then LTA involves some 4 million fits, and LMS 75 million. The default choice of h for LTS, $h = 52$, would lead to 9×10^{28} subsets. While LTA leads to a reasonable computation on a modest personal computer, LMS does not, and LTS is far beyond the bounds of the thinkable.

3.2 A "feasible solution" method for LTA

While exhaustive study of 500 cases and 3 predictors (an intercept and two slopes, for example) is manageable, going to 4 predictors takes the problem out of the realm of exact computation, and this shows the need for some other method suitable for approximating the LTA in large data sets. The LTA is defined as the L1 fit to a suitably chosen "half" of the cases. It has the property that the absolute residuals of all cases that it covers are less than or equal to the absolute residuals of all cases that it does not cover. This

characterization leads to the following "feasible solution algorithm" (FSA) for LTA:-

1. Generate a random elemental set, and calculate the residuals it gives rise to on each case in the data set.

2. If the current elemental set gives the L1 fit to the h cases with the smallest absolute residuals, then it is a feasible solution.

3. If not, then it can not be the solution. Refine it by replacing one of the cases in the elemental set with a "better" one.

4. Continue until you reach a feasible elemental.

5. Repeat the algorithm with a large number of random starts.

6. Use the feasible solution with the smallest sum of absolute deviations on the h covered cases.

Clearly, provided this algorithm is started with enough random starts, it must converge to the global LTA. The third step involves the replacement of one case in the current elemental set with a "better" one, as can be done effectively with a single step of Bloomfield and Steiger's (1980) algorithm for the L1 fit to a data set. This algorithm starts with an arbitrary elemental set and computes the residuals on all cases. If the current elemental set does not provide the L1 fit, then it is improved by replacing one of the cases in it with the case whose residual defines a suitably weighted median of the residuals. Bloomfield and Steiger motivated an heuristic to identify a good case to remove from the current elemental set, and claimed that the resulting algorithm was inherently faster than any other algorithm then known.

A simple adaptation of Bloomfield and Steiger's full-sample algorithm is suitable for our problem. At the stage of selecting which case to bring into the current basis, instead of searching over all n cases, restrict the search to the h cases with the smallest residuals from the current trial elemental solution. Perform a single exchange, as in the full-sample version. This leads to a new elemental set which is improved in the sense of being closer to the L1 fit to the h cases with the currently smallest absolute residuals. The residuals on all cases are then recomputed using this new elemental regression, during which the set of cases with the h smallest absolute residuals might change. If the current elemental set does not provide the L1 fit to the cases with the h smallest residuals, then a further step of the modified Bloomfield-Steiger algorithm is applied. This process continues until a feasible solution is reached.

Since the sum of the h smallest absolute residuals decreases at each step of this algorithm, it follows at once that the algorithm must converge.

FORTRAN codes implementing this feasible solution algorithm, and the exact code obtained by expanding MVELMS, are at the following website (go to the software icon). http://www.stat.umn.edu

We defined our feasible solution algorithm for LTA (FLTA) by the property that a feasible solution gives the L1 fit to the subset of cases that have the h smallest absolute residuals. We can use parallel definitions to define feasible solution approaches to LMS and LTS (FLMS and FLTS, respectively). These are not necessarily the same definitions that have been used in published definitions of these algorithms (Hawkins 1993, 1994), but are both particularly suitable for large n and better for comparability with the LTA

we have defined. Ruppert (1992) also uses subset refinement to approximate LMS and LTS.

Each of the feasible solution algorithms starts with a trial subset and refines it. In the worst case, if the initial subset contains one or more outlying cases, the algorithm may converge to a feasible solution that still includes one or more outlying cases. We can therefore get a conservative estimate of the ability of any of these estimators to locate outliers by calculating the probability that the initial elemental set consists entirely of "clean" cases.

Table 6 shows some illustrative figures for the LTA and LTS algorithms for samples with 10% contamination – a level that is not unrealistically high for many types of data. We omit LMS because its zero efficiency makes it particularly unattractive in large samples. The table shows the common log of the probability that a single starting subset will consist entirely of clean cases.

An entry less than -3 indicates that the probability of getting a clean starting set is less than 1 in 1000, which implies that the algorithm is likely to fail unless it is restarted with at least several thousand initial random subsets. This is the case for LTS for all of the subset sizes of 100 or above, indicating that the LTS algorithm (at least as we have defined it here) will not work for these larger sample sizes unless it uses a very large number of random starts. For all of the entries in the table, however, the LTA algorithm has a high proportion of clean starting subsets, and so a high probability of reaching a correct identification of sufficiently severe outliers when started with a modest number of random starting sets.

For example, at $n = 500, p = 5$, the probability that a starting subset is "clean" is $10^{-0.2249} = 0.6$, so a majority of starting subsets will be clean and there is little cause for concern about the algorithm ending in a contaminated subset.

We generated data sets containing 10% of severe outliers, all of them on high leverage cases, and investigated the ability of the FLTA and FLTS to converge to a solution in which all the outliers were uncovered. The results are summarized in table 7, which shows the execution time per starting subset and the proportion of starting subsets that converge to a valid solution. The runs used 5,000 random starts. The execution times are explained quite well by the empiric model

$$\text{FLTA} \quad time = 0.062(np)^{1.5}.$$

We tested the ability of the FSA for LTA to handle large data sets by analyzing a simulated data set with 10,000 cases (one third of them outlying) and 10 predictors. The feasible solution algorithm required under one minute per random start. All feasible solutions (of which we found 343 in 5,000 random starts) correctly identified the outliers.

Table 8 covers the illustrative case $p = 10$, and shows as a function of n and of the contamination proportion δ, the common log of the proportion of starting subsets that are entirely clean. Using again the rough guidance that a figure above -3 is acceptable (one random start in 1,000 will be clean so that a few thousand random starts will converge reliably), we see that the LTA gives acceptable performance across the board, while LTS (at least if implemented using only the concentration necessary condition) is guaranteed

to work well only for data sets of modest size and limited contamination.

4 THE IMPACT OF CASE LEVERAGE ON h

There is one final point we have not discussed in detail – the interplay between case leverage and regression outliers. OLS is affected by all regression outliers, regardless of their position in $\boldsymbol{X}$ space. The L1 norm by contrast is resistant to regression outliers occurring on low-leverage cases – this has been a strong argument for the routine use of the L1 full-sample norm. For example, Hampel, Rousseeuw, Ronchetti, and Stahel (1986, p. 328) state that L1 has 25% breakdown for uniform design and approximately 24% breakdown for Gaussian design. L1 is however not robust to regression outliers on high leverage cases, and for this reason has the same zero breakdown as does OLS. When using LTS, it is necessary to pick h sufficiently low such that all outliers can be trimmed; with LTA it is sometimes enough only to trim regression outliers on high leverage cases. This means that it is reasonable to use higher values for the coverage h when using LTA than when using LTS. Those who would carry an umbrella regardless of the weather forecast will continue to stick to the maximum breakdown choice $h = [(n + p + 1)/2]$ since it will accommodate the worst possible case of the maximum possible number of outliers, all of them on high leverage cases, but others might increase the coverage h, and thereby get back some of the statistical efficiency lost at normal data by using the L1 rather than least squares norm.

For example, in order to obtain Gaussian efficiency roughly twice that of LTS(0.5), it may be reasonable to use $h = 2n/3$ with LTA. This will still handle close to 50% outliers,

provided these split at random between high and low leverage cases.

5 EXAMPLES

Modified octane data set. The modified octane data set (Atkinson 1994) is a well-known data set of 82 cases and 4 predictors, containing 7 outliers planted on high leverage cases. It can in principle be analyzed using LTS. However in view of the data set size, it is impossible to find the exact LTS fit, all one can do is look for a good approximation.

We analyzed the data set using the exact LTA procedure using enumeration of all possible elemental sets. We found the LTA for all coverages h from 63 to 82 cases. Table 9 and table 10 show some summary values for the different h values in the range. The most striking feature is perhaps that all the summary values seem quite stable for h values up to 74, but then started to change substantially as the previously-excluded outliers are included in the covered set. It is interesting that the change starts one h value before one might have expected; this is due to case 21 which, while not particularly outlying in comparison with cases 71-77, is sufficiently different and of high enough leverage to impact the fit once it is accommodated. This computation took 15.75 hours on a HP 712/60 workstation, a substantial though still reasonable computation on that machine, but a much smaller one on a more modern desktop machine.

A "missing data" set. Missing values of the predictors are always a problem in multiple regression. Perhaps the most common approach is to either delete cases that have missing data, or (in cases where the analyst is not committed to a model using all predictors) to delete predictors that are not present on all cases. Either approach presents

a severe dilemma to the data analyst; if one knew the predictor were not needed, one would rather keep the case and lose the predictor, but if the predictor were needed, its removal would invalidate the whole analysis. Data sets with a sprinkling of missing values therefore typically involve many iterations of deleting some mix of affected cases and affected predictors to get complete data sets that can be fitted and evaluated.

High breakdown methodologies provide a possible third approach. This is to include all data, but with missing values on predictors assigned some extreme value like the traditional 9999 to make them massively influential. Faced with a case with a missing predictor, the high breakdown estimator may then choose to either cover the case, using a zero coefficient on the predictor with the missing information; or to exclude the case, as will be necessary if the predictor with the missing value really is informative about the dependent. Thus a single pass with a high breakdown analysis such as LTA can provide a starting picture of which predictors appear to be needed and which do not, despite even quite high levels of missing information.

To illustrate this possibility, along with the handling of conventional outlying values, we analyzed a physical anthropology data set from the literature. The data set (from Gladstone 1905-06) investigates the relationship between brain weight measured post mortem and a number of body dimensions measurable in vivo. This data set contained 276 cases. We used 7 predictors – cause of death (coded as either chronic or acute), cranium height, length, breadth, volume and circumference, and cephalic index. There were 77 cases missing information on one or more predictors. The data set also included five infants less than 7 months old. We carried out an LTA fit using the feasible solution

algorithm. The coverage was $h = 182$, a choice motivated by our $2n/3$ suggestion, and one that allows for all 77 cases that had any missing information to be deleted along with 17 outliers.

The results showed several useful features of this approach. First, recall that a feasible solution is a regression that is an L1 fit to the smallest h residuals it generates, and that there may be many feasible solutions in a data set. In this data set, three feasible solutions were particularly interesting; their coefficients (with all variables in standardized units) and those of the OLS fit using just complete cases were

Fit	Cause	Height	Length	Breadth	Vol	Circum	C I
OLS	-0.05	0.29	0.35	0.14	0.05	0.02	0.04
LTA 1	0.00	0.38	-3.07	3.73	0.19	0.00	-3.10
LTA 2	0.00	0.41	-3.40	4.02	0.17	0.03	-3.39
LTA 3	0.00	-0.81	-3.83	0.81	3.69	0.00	-2.20

All LTA solutions dropped the predictor "Cause", preferring to retain the many cases that were missing this predictor.

The first solution also dropped "Circum", the head circumference. In this regression, all the infants were inliers. The second solution retained "Circum" as a predictor, but trimmed another formerly inlying case that was missing "Circum". With this decision, the infants brain weights were again inliers. The third solution dropped "Circum" substituting "Vol", and making the infants outliers. The mean absolute residual of the covered cases was 0.08, and the infants' absolute residuals were all in excess of 2.5. This shows the interplay between case characteristics and the apparent importance or otherwise of predictors.

We do not know whether to prefer the models that contrasts breadth and length; largely, it would seem, to accommodate the infants, or to prefer the model that has contrasts length and volume, recognizing that the resulting model does not describe infants adequately. It is however very much to the credit of the feasible solution LTA fit that it identifies these three possibilities for the analyst's attention.

6 CONCLUSIONS

High breakdown estimation in large data sets is a challenging problem. For data sets with "normal in the middle" residuals, least trimmed squares (LTS) is attractive on the grounds of statistical efficiency, but there is no workable way of finding the exact LTS fit on any but quite small data sets. Switching the criterion to the L1 norm gives the least trimmed sum of absolute values (LTA) estimator. This sacrifices some statistical efficiency for "normal in the middle" data, but is more efficient for peaked error distributions. It is also far easier to compute, both exactly and approximately. These properties make it potentially very attractive, particularly for the analysis of large data sets.

To the extent that there is concern about the efficiency lost, LTA may be used in the conventional two-stage way, taking its coefficients as the starting point for an MM, S or τ estimate. As the estimate has $O_P(n^{-0.5})$ convergence along with its high breakdown, it will serve this purpose as well as does any other initial $O_P(n^{-0.5})$ high breakdown estimator, and will do so better than the traditional LMS estimator which has worse asymptotics and is harder to compute.

Exact computation of the LTA involves enumeration of all elemental subsets of the

data, and has been implemented through a modification of the Hawkins-Simonoff elemental set code. A feasible solution algorithm gives good approximations for data sets too large for exact enumeration. These make the LTA a potentially useful tool for practical data analysis.

Another less conventional use of the LTA is as a tool for modeling data sets with missing observations on predictors. By coding missing values as extremes in an LTA analysis, one may get in a single run indications of which predictors are important (with the implication that the cases missing those predictors must be either completed in some way or dropped) and which are not. This use may greatly streamline the initial modeling step of data with a sprinkling of missing values of predictors.

7 REFERENCES

Atkinson, A.C. (1994), "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association,* 89, 1329-1339.

Bassett, G.W. (1991), "Equivariant, Monotonic, 50% Breakdown Estimators," *The American Statistician,* 45, 135-137.

Bassett, G.W., and Koenker, R.W. (1978), "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association,* 73, 618-622.

Bloomfield, P., and Steiger, W. (1980), "Least Absolute Deviations Curve-Fitting," *SIAM Journal of Statistical Computing,* 1, 290-301.

Butler, R.W. (1982), "Nonparametric Interval and Point Prediction Using Data Trimming by a Grubbs-Type Outlier Rule," *The Annals of Statistics,* 10, 197-204.

Croux, C., Rousseeuw, P.J., and Hössjer, O. (1994), "Generalized S-Estimators," *Journal of the American Statistical Association,* 89, 1271-1281.

Croux, C., Rousseeuw, P.J., and Van Bael, A. (1996), "Positive-Breakdown Regression by Minimizing Nested Scale Estimators," *Journal of Statistical Planning and Inference,* 53, 197-235.

Gladstone, R. J. (1905-1906), "A Study of the Relations of the Brain to the Size of the Head," *Biometrika,* 4, 105-123.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics,* John Wiley and Sons, Inc., NY.

Hawkins, D.M. (1993), "The Feasible Set Algorithm for Least Median of Squares Regression," *Computational Statistics and Data Analysis,* 16, 81-101.

Hawkins, D.M. (1994), "The Feasible Solution Algorithm for Least Trimmed Squares Regression," *Computational Statistics and Data Analysis,* 17, 185-196.

Hawkins, D.M., and Simonoff, J.S. (1993), "High Breakdown Regression and Multivariate Estimation," *Applied Statistics,* 42, 423-432.

Hössjer , O. (1991), Rank-Based Estimates in the Linear Model with High Breakdown Point, Ph.D. Thesis, Report 1991:5, Department of Mathematics, Uppsala University, Uppsala, Sweden.

Hössjer , O. (1994), "Rank-Based Estimates in the Linear Model with High Breakdown Point," *Journal of the American Statistical Association,* 89, 149-158.

Koenker, R.W., and Bassett, G. (1978), "Regression Quantiles," *Econometrica,* 46, 33-50.

Niinimaa, A., Oja, H., and Tableman, M. (1990), "The Finite-Sample Breakdown Point of the Oja Bivariate Median and of the Corresponding Half-Samples Version," *Statistics and Probability Letters,* 10, 325-328.

Oosterhoff, J. (1994), " Trimmed Mean or Sample Median?" *Statistics and Probability Letters,* 20 401-409.

Portnoy, S., (1987), "Using Regression Fractiles to Identify Outliers", in Y. Dodge, *Statistical Data Analysis Based on the L_1 Norm and Related Methods*, North Holland, Amsterdam.

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

Rousseeuw, P.J., and Bassett, G.W. (1991), "Robustness of the p-Subset Algorithm for Regression with High Breakdown Point," in *Directions in Robust Statistics and Diagnostics: Part 1*, eds. W. Stahel and S. Weisberg, New York: Springer-Verlag, 185-194.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* New York: John Wiley.

Ruppert, D. (1992), "Computing S-Estimators for Regression and Multivariate Location/Dispersion," *Journal of Computational and Graphical Statistics,* 1, 253-270.

Stromberg, A.J. (1993), "Computing the Exact Least Median of Squares Estimate and Stability Diagnostics in Multiple Linear Regression," *SIAM Journal of Scientific and Statistical Computing,* 14, 1289-1299.

Stromberg, A.J., Hawkins, D.M., and Hössjer , O. (1997), "The Least Trimmed Differ-

ences Regression Estimator and Alternatives," submitted for publication.

Tableman, M. (1994a), "The Influence Functions for the Least Trimmed Squares and the Least Trimmed Absolute Deviations Estimators," *Statistics and Probability Letters,* 19, 329-337.

Tableman, M. (1994b), "The Asymptotics of the Least Trimmed Absolute Deviations (LTAD) Estimator," *Statistics and Probability Letters,* 19, 387-398.

Yohai, V.J., and Zamar, R.H. (1993), "A Minimax-Bias Property of the Least $\alpha-$Quantile Estimates," *The Annals of Statistics,* 21, 1824-1842.

Table 1: $\sqrt{n}$ SD for N(0,1) Data

n	OLS	L1	LTA(0.5)	LTS(0.5)
20	1.00	1.23	2.20	2.12
40	1.01	1.21	2.56	2.41
100	0.99	1.24	3.13	2.94
400	0.98	1.22	3.84	3.34
600	1.00	1.25	3.87	3.31
∞	1.00	1.25	4.36	3.74

Table 2: $\sqrt{n}$ SD for C(0,1) Data

n	OLS	L1	LTA(0.5)	LTS(0.5)
20	593	1.67	2.07	1.91
40	2969	1.63	2.09	1.99
100	7360	1.65	2.24	2.03
400	1394	1.62	2.18	1.98
600	524	1.54	2.18	1.98
∞	∞	1.57	2.22	2.03

Table 3: $\sqrt{n}$ SD for DE(0,1) Data

n	OLS	L1	LTA(0.5)	LTS(0.5)
20	1.40	1.12	1.72	1.58
40	1.39	1.10	1.73	1.68
100	1.41	1.05	1.74	1.70
400	1.40	1.06	1.64	1.72
600	1.44	1.05	1.57	1.71
∞	1.41	1.00	1.41	1.68

Table 4: Monte Carlo OLS Relative Efficiencies

dist	n	L1	LTA(0.5)	LTS(0.5)	LTA(0.75)
N(0,1)	20	.668	.206	.223	.377
N(0,1)	40	.692	.155	.174	.293
N(0,1)	100	.634	.100	.114	.230
N(0,1)	400	.652	.065	.085	.209
N(0,1)	600	.643	.066	.091	.209
N(0,1)	∞	.637	.053	.071	.199
DE(0,1)	20	1.560	.664	.783	1.157
DE(0,1)	40	1.596	.648	.686	1.069
DE(0,1)	100	1.788	.656	.684	1.204
DE(0,1)	400	1.745	.736	.657	1.236
DE(0,1)	600	1.856	.845	.709	1.355
DE(0,1)	∞	2.000	1.000	.71	1.500

Table 5: Number of elemental regressions as a function of n and p

	p			
n	2	3	4	5
10	45	120	210	252
20	190	1140	4845	15504
30	435	4060	27405	142506
50	1225	19600	230300	2e6
100	4950	161700	3e6	8e6
200	19900	1e6	65e6	3e9
500	124750	21e6	3e9	3e11

Table 6: Log10 of probability that a random starting set is clean

		p			
n		2	3	4	5
25	FLTA	-.0740	-.1135	-.1549	-.1984
	FLTS	-.6576	-.7368	-.7368	-.8239
50	FLTA	-.0732	-.1110	-.1496	-.1891
	FLTS	-1.3359	-1.4151	-1.4151	-1.4981
75	FLTA	-.0857	-.1295	-.1739	-.2190
	FLTS	-2.2851	-2.3761	-2.3761	-2.4700
100	FLTA	-.0824	-.1242	-.1665	-.2092
	FLTS	-2.9665	-3.0547	-3.0547	-3.1449
150	FLTA	-.0854	-.1286	-.1720	-.2158
	FLTS	-4.5941	-4.6852	-4.6852	-4.7777
200	FLTA	-.0869	-.1307	-.1748	-.2191
	FLTS	-6.2209	-6.3135	-6.3135	-6.4071
500	FLTA	-.0897	-.1347	-.1797	-.2249
	FLTS	-15.9785	-16.0736	-16.0736	-16.1692

Table 7: Execution time and success rate per random start

			n			
p			50	100	200	500
5	FLTA	time	261	718	1858	6560
		% good	50	52	53	55
10	FLTA	time	583	1697	4948	18587
		% good	25	26	31	37
15	FLTA	time	1122	3483	10395	42116
		% good	15	15	18	24

Table 8: Log10 probability that a random starting set is clean

		δ							
n		0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
50	FLTA	-.20	-.51	-.73	-1.08	-1.34	-1.75	-2.05	-2.53
	FLTS	-.81	-2.14	-3.11	-4.75	-5.98	-8.16	-9.94	-13.67
100	FLTA	-.23	-.48	-.74	-1.02	-1.32	-1.64	-1.99	-2.36
	FLTS	-1.79	-3.73	-5.87	-8.23	-10.88	-13.93	-17.54	-22.05
200	FLTA	-.23	-.47	-.72	-.99	-1.28	-1.59	-1.93	-2.29
	FLTS	-3.35	-6.97	-10.94	-15.31	-20.20	-25.77	-32.26	-40.17
500	FLTA	-.22	-.46	-.71	-.98	-1.26	-1.57	-1.89	-2.24
	FLTS	-8.03	-16.72	-26.21	-36.65	-48.29	-61.47	-76.76	-95.14

Table 9: Fitted coefficients by coverage

h	Q	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
63	14.10	96.170	-0.106	-0.074	-0.038	2.650
64	14.72	96.254	-0.104	-0.083	-0.039	2.574
65	15.34	96.254	-0.104	-0.083	-0.039	2.574
66	15.97	96.254	-0.104	-0.083	-0.039	2.574
67	16.68	96.254	-0.104	-0.083	-0.039	2.574
68	17.43	96.370	-0.105	-0.082	-0.039	2.558
69	18.17	96.149	-0.103	-0.062	-0.040	2.589
70	18.95	96.395	-0.102	-0.067	-0.042	2.488
71	19.77	99.046	-0.109	-0.108	-0.062	1.923
72	20.55	99.044	-0.110	-0.108	-0.062	1.930
73	21.47	99.044	-0.110	-0.108	-0.062	1.930
73	21.47	99.044	-0.110	-0.108	-0.062	1.930
74	22.44	98.900	-0.110	-0.112	-0.063	2.044
75	23.58	98.289	-0.108	-0.110	-0.053	2.034
76	27.54	98.156	-0.107	-0.103	-0.051	1.979
77	31.34	97.217	-0.102	-0.090	-0.041	1.995
78	35.20	96.914	-0.100	-0.102	-0.035	1.887
79	38.92	97.066	-0.090	-0.153	-0.036	1.475
80	42.60	96.877	-0.083	-0.224	-0.029	1.139
81	45.13	92.974	-0.070	-0.328	0.027	1.098
82	48.13	92.503	-0.068	-0.332	0.028	1.297

Table 10: Some summary numbers of the modified octane data set

h	r_{71}	r_{72}	r_{73}	r_{74}	r_{75}	r_{76}	r_{77}
63	-4.02	-4.12	-4.73	-4.35	-6.52	-6.69	-6.37
64	-3.99	-4.07	-4.63	-4.27	-6.28	-6.49	-6.23
65	-3.99	-4.07	-4.63	-4.27	-6.28	-6.49	-6.23
66	-3.99	-4.07	-4.63	-4.27	-6.28	-6.49	-6.23
67	-3.99	-4.07	-4.63	-4.27	-6.28	-6.49	-6.23
68	-4.00	-4.08	-4.66	-4.30	-6.37	-6.58	-6.31
69	-3.97	-4.06	-4.72	-4.33	-6.44	-6.59	-6.28
70	-3.95	-4.02	-4.67	-4.29	-6.32	-6.51	-6.25
71	-4.18	-4.13	-4.71	-4.37	-6.43	-6.84	-6.68
72	-4.19	-4.13	-4.71	-4.37	-6.44	-6.85	-6.69
73	-4.19	-4.13	-4.71	-4.37	-6.44	-6.85	-6.69
74	-4.26	-4.22	-4.76	-4.42	-6.49	-6.89	-6.69
75	-4.08	-4.05	-4.60	-4.28	-6.26	-6.65	-6.53
76	-3.98	-3.95	-4.54	-4.21	-6.19	-6.57	-6.48
77	-3.71	-3.71	-4.33	-4.01	-5.86	-6.19	-6.18
78	-3.51	-3.50	-4.07	-3.79	-5.42	-5.81	-5.94
79	-3.23	-3.16	-3.51	-3.31	-4.14	-4.76	-5.31
80	-2.91	-2.80	-2.83	-2.75	-2.77	-3.64	-4.66
81	-1.83	-1.83	-1.28	-1.45	-0.04	-1.16	-3.04
82	-1.92	-1.95	-1.33	-1.50	0.00	-1.10	-2.95