Robust Regression with High Coverage

David J. Olive and Douglas M. Hawkins*

Southern Illinois University and University of Minnesota

July 21, 2003

Abstract

An important parameter for several high breakdown regression algorithm estimators is the number of cases given weight one, called the coverage of the estimator. Increasing the coverage is believed to result in a more stable estimator, but the price paid for this stability is greatly decreased resistance to outliers. A simple modification of the algorithm can greatly increase the coverage and hence its statistical performance while maintaining high outlier resistance.

KEY WORDS: Elemental Sets; LMS; LTA; LTS; Outliers.

*David J. Olive is Assistant Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@math.siu.edu. Douglas M. Hawkins is Professor, School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church St. SE, Minneapolis, MN 55455-0493, USA. E-mail address: doug@stat.umn.edu. This research was supported by NSF grants DMS 9806584 and DMS 0202922. The authors are grateful to the editor and a referee for a number of helpful suggestions for improvement in the article.

1 INTRODUCTION

Consider the regression 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 row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element y_i of $\boldsymbol{Y}$.

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

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

respectively. The less frequently used Chebyshev (L_∞) method minimizes the maximum absolute residual.

Some high breakdown (HB) robust regression methods can fit the bulk of the data even if certain types of outliers are present. Let $|r|_{(i)}(\boldsymbol{b})$ denote the absolute residuals sorted from smallest to largest. The least quantile of squares (LQS(c_n)) estimator minimizes the criterion

$$Q_{LQS}(\boldsymbol{b}) = r^2_{(c_n)}(\boldsymbol{b}). \tag{1.3}$$

When $c_n/n \to 1/2$, the LQS(c_n) estimator is also known as the least median of squares (LMS) estimator (Hampel 1975). The least trimmed sum of squares (LTS(c_n)) estimator

(Rousseeuw 1984) minimizes the criterion

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

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

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

These three estimators all "cover" a set of fixed size $c_n = c_n(p) \geq n/2$ cases, fitting a classical estimator to the covered cases. LQS uses the Chebyshev fit, LTA uses L_1, and LTS uses OLS. If c_n is a sequence of integers such that $c_n/n \to \tau \geq 0.5$, then $1 - \tau$ is the approximate amount of trimming. For the LTA and LTS estimators there is a tradeoff in that the Gaussian efficiency increases as τ tends to 1, but the breakdown value $1 - \tau$ decreases to zero. Let $[x]$ denote the greatest integer function. Hence $[7.7] = 7$. The integer valued parameter c_n is called the *coverage* of the estimator and the choice

$$c_n = [n/2] + [(p+1)/2] \tag{1.6}$$

corresponding to $\tau = \frac{1}{2}$ maximizes the breakdown of the estimator.

We will use the unifying notation LTx(τ) for the LTx(c_n) estimator where x is A, Q, or S for LTA, LQS, and LTS, respectively. Since the exact algorithms for the LTx criteria have very high computational complexity, approximations based on iterative algorithms are generally used. We will call the algorithm estimator $\hat{\boldsymbol{\beta}}_A$ the ALTx(τ) estimator.

Historically, the workhorse of high breakdown algorithms has been the "basic resampling", or "elemental set" algorithm. This uses K_n "starts" – randomly selected "elemental" subsets of p from which the residuals are computed for all n cases. The algorithm returns the elemental fit that optimizes the criterion. The efficiency and resistance properties of the ALTx estimator turn out to depend strongly on the number of starts K_n used – see Hawkins and Olive (2002). For a fixed choice of K_n, increasing the coverage c_n in the LTx criterion seems to result in a more stable ALTA or ALTS estimator. For this reason, *Splus* has increased the default coverage of the `ltsreg` function to $0.9n$ while Rousseeuw and Hubert (1999) recommend $0.75n$. The price paid for this stability is greatly decreased resistance to outliers.

Rather than using a fixed coverage such as 0.75, we suggest using a highly resistant initial estimator to determine the variable trimming proportion. Section 2 defines the estimator and Section 3 provides some theory. Earlier work on combining efficiency and high breakdown includes Jureckova and Portnoy (1987) and He (1991).

2 Obtaining Stability and Resistance

Combining the two concepts of variable coverage and a two–stage process of identification followed by estimation leads to a class of regression estimators. Define a set of $L = 5$ estimators ALTx($c_{n,j}$) corresponding to coverages $\tau_j \in G = \{0.5, 0.75, 0.9, 0.99, 1.0\}$. The exact coverages c are defined by $c_{n,1} \equiv c_n$ as given by equation (1.6); and $c_{n,2} = [.75\ n]$, $c_{n,3} = [.90\ n]$, $c_{n,4} = [.99\ n]$, and $c_{n,5} = n$. (This choice of L and G is for illustration.)

Then the RLTx(k) estimator is the ALTx(τ_R) estimator where τ_R is the largest $\tau_j \in G$

such that $[\tau_j \; n] \leq C_n(\hat{\boldsymbol{\beta}}_{ALTx(c_n)})$ where

$$C_n(\boldsymbol{b}) = \sum_{i=1}^{n} I[|r|_{(i)}(\boldsymbol{b}) \leq k \; |r|_{(c_n)}(\boldsymbol{b})] = \sum_{i=1}^{n} I[r^2_{(i)}(\boldsymbol{b}) \leq k^2 \; r^2_{(c_n)}(\boldsymbol{b})]. \qquad (2.1)$$

Notice that although L estimators ALTx($c_{n,j}$) were defined, there is no need to compute all of them; only two are needed – ALTx(0.5) to get a resistant scale and define the coverage needed, and the final estimator ALTX(τ_R).

Section 3 will show that the RLTx estimator has a high degree of resistance along with high stability. The tuning parameter $k \geq 1$ controls the amount of trimming. The inequality $k \geq 1$ implies that $C_n \geq c_n$, so the RLTx(k) estimator has coverage at least as high as the LTx(0.5), and in "clean" data will commonly have 100% coverage.

The behavior of the RLTx estimator is easy to understand. Compute the most resistant ALTx estimator $\hat{\boldsymbol{\beta}}_{ALTx(c_n)}$ and obtain the corresponding residuals. Count the number C_n of absolute residuals that are no larger than $k \; |r|_{(c_n)} \approx k\text{MED}(|r|_i)$. Then find the corresponding $\tau_R \in G$ and compute the RLTx estimator. If $k = 6$, and the regression model holds, the RLTx estimator will be the classical estimator or the ALTx estimator with 99% coverage for a wide variety of data sets. The method has the "exact fit" property – if $\hat{\boldsymbol{\beta}}_{ALTx(c_n)}$ fits c_n cases exactly, then $|r|_{(c_n)} = 0$ and RLTx = ALTx(c_n).

The RLTx estimator has the same breakdown point as its starting ALTx(0.5) estimator. Hence the RLTx estimator for $x =$ A and S is simultaneously more stable and more resistant than the fixed–coverage ALTx estimators with $\tau = 0.75$ or $\tau = 0.9$, but takes about twice as long to compute. Increasing the coverage for the LQS criterion is possible but inadvisable since the Chebyshev fit tends to have less efficiency than the LMS fit.

3 Theoretical properties

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N(0, V(\hat{\boldsymbol{\beta}}, F)\ W) \text{ where } \frac{X^T X}{n} \to W^{-1},$$

and the errors e_i are iid with zero median and have a distribution F with symmetric unimodal density f. When the variance $V(e_i)$ exists,

$$V(OLS, F) = V(e_i) = \sigma^2 \text{ while } V(L_1, F) = \frac{1}{4f^2(0)}.$$

See Bassett and Koenker (1978). Broffitt (1974) compares OLS, L_1, and L_∞ in the location model and shows that the rate of convergence of the Chebyshev estimator is often very poor.

Obtaining asymptotic theory for LTA and LTS is a very challenging problem and there are currently no results outside of the location model – see Hawkins and Olive (2002) for further discussion. For the location model, Butler (1982) derived asymptotic theory for LTS while Tableman (1994ab) derived asymptotic theory for LTA. In the regression setting, it is known that LQS(τ) converges at a cube root rate to a non-Gaussian limit (Davies 1990, Kim and Pollard 1990).

Some negative results are immediate; if the "shortest half" is not unique, then LQS, LTA, and LTS are inconsistent. For example, the shortest half is not unique for the uniform distribution.

Surprisingly, some useful asymptotic theory for RLTx is easily derived. The following lemma will be useful for estimating the coverage of the RLTx estimator given the error distribution F.

Lemma 3.1. Assume that the errors are iid with a density f that is symmetric about 0 and positive and continuous in neighborhoods of $F^{-1}(0.75)$ and $kF^{-1}(0.75)$. If the predictors $\boldsymbol{x}$ are bounded in probability and $\hat{\boldsymbol{\beta}}_n$ is consistent for $\boldsymbol{\beta}$, then

$$\frac{C_n(\hat{\boldsymbol{\beta}}_n)}{n} \xrightarrow{P} \tau_F = \tau_F(k) = F(k\ F^{-1}(0.75)) - F(-k\ F^{-1}(0.75)). \tag{3.1}$$

Proof. See appendix.

Under the same conditions of Lemma 3.1,

$$|r|_{(c_n)}(\hat{\boldsymbol{\beta}}_n) \xrightarrow{P} F^{-1}(0.75).$$

This result can be used as a diagnostic – compute several regression estimators including OLS and L_1 and compare the corresponding median absolute residuals.

A competitor to RLTx is to compute ALTx, give zero weight to cases with large residuals, and fit OLS to the remaining cases. He and Portnoy (1992) prove that this two–stage estimator has the same rate as the initial estimator. Theorem 3.2 gives a similar result for the RLTx estimator, but the RLTx estimator could be an OLS, L_1 or L_∞ fit to a subset of the data. In particular, if the exact LTx estimators are used, Theorem 3.2 shows that the RLTQ estimator has an $O_P(n^{-1/3})$ rate but suggests that the RLTA and RLTS estimators converge at an $O_P(n^{-1/2})$ rate.

Theorem 3.2. If $\|\hat{\boldsymbol{\beta}}_{ALTx(\tau_j)} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for all $\tau_j \in G$, then $\|\hat{\boldsymbol{\beta}}_{RLTx} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$.

Proof. Since G is finite, this result follows from Pratt (1959). QED

Theorem 3.3 shows that the RLTx estimator is asymptotically equivalent to an LTx estimator that typically has high coverage.

Theorem 3.3. Assume that $\tau_j, \tau_{j+1} \in G$. If $P[C_n(\hat{\boldsymbol{\beta}}_{ALTx(0.5)})/n \in (\tau_j, \tau_{j+1})] \xrightarrow{P} 1$, then the RLTx estimator is asymptotically equivalent to the ALTx(τ_j) estimator.

Choosing a suitable k for a target distribution F is simple. Assume equation (3.1) holds where τ_F is not an element of G. If n is large, then with high probability τ_R will equal the largest $\tau_i \in G$ such that $\tau_i < \tau_F$. Small sample behavior can also be predicted. For example, if the errors follow a $N(0, \sigma^2)$ distribution and $n = 1000$, then $P(-4\sigma < e_i < 4\sigma, i = 1, ..., 1000) \approx (0.9999)^{1000} > 0.90$, while $|r|_{(c_n)}$ is converging to $\Phi^{-1}(0.75)\sigma \approx 0.67\sigma$. Hence if $k \geq 6.0$, $n < 1000$, and the errors are Gaussian, the RLTS estimator will cover all cases with high probability. To include heavier tailed distributions, increase k. For example, similar statements hold for distributions with lighter tails than the double exponential distribution if $k \geq 10.0$ and $n < 200$.

Table 1 presents the results of a small simulation study. We compared ALTS(τ) for $\tau = 0.5, 0.75$, and 0.9 with RLTS(6) for 6 different error distributions – the normal(0,1), Laplace, uniform($-1, 1$) and three 60% N(0,1) 40 % contaminated normals. The three contamination scenarios were: N(0,100) for a "scale" contaminated setting; and two "location" contaminations – N(5.5,1) and N(12,1). The shift of 5.5 is perhaps a worst case for the RLTS estimator, as these contaminants are just small enough that many pass the $k = 6$ screen. The shift of 12 tests the estimators under catastrophic contamination.

The simulation used $n = 100$ and $p = 6$ (5 slopes and an intercept) over at least 1000 runs and computed $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2/6$ for each run. Note that for the three CN scenarios the number of contaminants is a binomial random variable which, with probability 6% will exceed the 47 that the maximum breakdown setting can accommodate.

The means from the 1000 values are displayed. Their standard errors are at most 5% of the mean. The last column shows the percentage of times that τ_R was equal to .5, .75, .9, .99 and 1.0. Two fitting algorithms were used. One was a traditional elemental algorithm with 3000 starts. As discussed in Hawkins and Olive (2002) this choice, chosen to match much standard practice, is far fewer than we would recommend with a raw elemental algorithm.

The other was a concentration algorithm. This used 300 starts for the location contamination distributions, and 50 starts for all others, preliminary experimentation having indicated that this many starts were sufficient. Comparing the 'conc' mean squared errors with the corresponding 'elem' confirms the recommendations in Hawkins and Olive (2002) that far more than 3000 elemental starts are necessary to achieve good results. The 'elem' runs also verify that second-stage refinement, as supplied by the RLTS approach, is not sufficient to overcome the deficiencies in the poor initial estimates provided by the raw elemental approach.

The RLTS estimator was, with one exception, either the best of the 4 estimators or barely distinguishable from the best. The single exception was the concentration algorithm with the contaminated normal distribution $F(x) = 0.6\Phi(x) + 0.4\Phi(x - 5.5)$, where most of the time it covered all cases. We already noted that location contamination with this mean and this choice of k is about the worst possible for the RLTS estimator, so that this worst-case performance is still about what is given by the more recent recommendations for ALTx coverage – 75% or 90% – is positive. This is reinforced by RLTS' excellent performance with 12σ location outliers.

The simulation therefore supports the use of the RLTx method as an approach that can provide the resistance of a traditional 50% high breakdown estimator with the greater stability and statistical efficiency associated with higher coverage.

Appendix

Proof of Lemma 3.1. First assume that the predictors are bounded. Hence $\|\boldsymbol{x}\| \leq M$ for some constant M. Let $0 < \gamma < 1$, and let $0 < \epsilon < 1$. Since $\hat{\boldsymbol{\beta}}_n$ is consistent, there exists an N such that

$$P(A) = P(\hat{\beta}_{j,n} \in [\beta_j - \frac{\epsilon}{4pM}, \beta_j + \frac{\epsilon}{4pM}], j = 1, ..., p) \geq 1 - \gamma$$

for all $n \geq N$. If $n \geq N$, then on set A,

$$\sup_{i=1,...,n} |r_i - e_i| = \sup_{i=1,...,n} |\sum_{i=1}^{p} x_{i,j}(\beta_j - \hat{\beta}_{j,n})| \leq \frac{\epsilon}{2}.$$

Since ϵ and γ are arbitrary,

$$r_i - e_i \xrightarrow{P} 0.$$

This result also follows from Rousseeuw and Leroy (1987, p. 128). In particular,

$$|r|_{(c_n)} \xrightarrow{P} \text{MED}(|e_1|) = F^{-1}(0.75).$$

Now there exists N_1 such that

$$P(B) \equiv P(|r_i - e_i| < \frac{\epsilon}{2}, i = 1, ..., n \ \& \ |\ |r|_{(c_n)} - \text{MED}(|e_1|)| < \frac{\epsilon}{2k}) \geq 1 - \gamma$$

for all $n \geq N_1$. Thus on set B,

$$\frac{1}{n}\sum_{i=1}^{n} I[-k\text{MED}(|e_1|) + \epsilon \leq e_i \leq k\text{MED}(|e_1|) - \epsilon] \leq \frac{C_n(\hat{\boldsymbol{\beta}}_n)}{n}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} I[-k\text{MED}(|e_1|) - \epsilon \leq e_i \leq k\text{MED}(|e_1|) + \epsilon],$$

and the result follows since γ and ϵ are arbitrary and the three terms above converge to τ_F almost surely as ϵ goes to zero.

When $\boldsymbol{x}$ is bounded in probability, fix M and suppose M_n of the cases have predictors $\boldsymbol{x}_i$ such that $\|\boldsymbol{x}_i\| \leq M$. By the argument above, the proportion of absolute residuals of these cases that are below $|r|_{(c_{M_n})}$ converges in probability to τ_F. But by increasing M, the proportion of such cases can be made arbitrarily close to one as n increases. QED

4 References

Bassett, G.W., and Koenker, R.W. (1978), Asymptotic theory of least absolute error regression, *J. Amer. Statis. Assoc.* **73,** 618-622.

Broffitt, J.D. (1974), An example of the large sample behavior of the midrange, *Amer. Statis.* **28,** 69-70.

Butler, R.W. (1982), Nonparametric interval and point prediction using data trimming by a Grubbs-type outlier rule, *Ann. Statis.* **10,** 197-204.

Davies, P.L. (1990), The asymptotics of S-estimators in the linear regression model, *Ann. Statis.* **18,** 1651-1675.

Hampel, F.R. (1975), Beyond location parameters: robust concepts and methods, *Bull. Internat. Statis. Instit.* **46,** 375-382.

Hawkins, D.M., and Olive, D. (1999), Applications and algorithms for least trimmed sum of absolute deviations regression, *Comput. Statist. and Data Anal.* **32,** 119-134.

Hawkins, D.M., and Olive, D.J. (2002), Inconsistency of resampling algorithms for high breakdown regression estimators and a new algorithm, *J. Amer. Statis. Assoc.* **96,** 136-148.

He, X. (1991), A local breakdown property of robust tests in linear regression, *J. Mult. Anal.* **38,** 294-305.

He, X., and Portnoy, S. (1992), Reweighted LS estimators converge at the same rate as the initial estimator, *Ann. Statis.* **20,** 2161-2167.

Jureckova, J., and Portnoy, S. (1987), Asymptotics for one-step M-estimators in regression with application to combining efficiency and high breakdown point, *Comm. Statis. Theory Methods* **16,** 2187-2199.

Kim, J., and Pollard, D. (1990), Cube root asymptotics, *Ann. Statis.* **18,** 191-219.

Pratt, J.W. (1959), On a general concept of 'in probability', *Ann. Math. Statis.* **30,** 549-558.

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

Rousseeuw, P.J., and Hubert, M. (1999), Regression depth, *J. Amer. Statis. Assoc.* **94,** 388-433.

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

Tableman, M. (1994a), The influence functions for the least trimmed squares and the least trimmed absolute deviations estimators, *Statis. and Probab. Lett.* **19,** 329-337.

Tableman, M. (1994b), The asymptotics of the least trimmed absolute deviations (LTAD) estimator, *Statis. and Probab. Lett.* **19,** 387-398.

Table 1: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/p$, 1000 runs

pop.	alg.	ALTS(.5)	ALTS(.75)	ALTS(.9)	RLTS(6)	% of runs that τ_R = .5,.75,.9,.99 or 1
Normal	conc	0.0648	0.0350	0.0187	0.0113	0,0,6,18,76
Laplace	conc	0.1771	0.0994	0.0775	0.0756	0,0,62,23,15
uniform	conc	0.0417	0.0264	0.0129	0.0039	0,0,2,6,93
scale CN	conc	0.0560	0.0622	0.2253	0.0626	2,96,2,0,0
5.5 loc CN	conc	0.0342	0.7852	0.8445	0.8417	0,4,19,9,68
12 loc CN	conc	0.0355	3.5371	3.9997	0.0405	85,3,2,0,9
normal	elem	0.1391	0.1163	0.1051	0.0975	0,0,1,6,93
Laplace	elem	0.9268	0.8051	0.7694	0.7522	0,0,20,28,52
uniform	elem	0.0542	0.0439	0.0356	0.0317	0,0,0,1,98
scale CN	elem	4.4050	3.9540	3.9584	3.9439	0,14,40,18,28
5.5 loc CN	elem	1.8912	1.6932	1.6113	1.5966	0,0,1,3,96
12 loc CN	elem	8.3330	7.4945	7.3078	7.1701	4,0,1,2,92