Chapter 7

Robust and Resistant Regression

7.1 High Breakdown Estimators

Assume that the multiple linear regression model

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

is appropriate for all or for the bulk of the data. For a high breakdown (HB) regression estimator $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the median absolute residual

$$\text{MED}(|r|_i) \equiv \text{MED}(|r(\boldsymbol{b})|_1, ..., |r(\boldsymbol{b})|_n)$$

stays bounded even if close to half of the data set cases are replaced by arbitrarily bad outlying cases; ie, the breakdown value of the regression estimator is close to 0.5. The concept of breakdown will be made more precise in Section 9.4.

Perhaps the first HB regression estimator proposed was the least median of squares (LMS) estimator. Let $|r(\boldsymbol{b})|_{(i)}$ denote the ith ordered absolute residual from the estimate $\boldsymbol{b}$ sorted from smallest to largest, and let $r^2_{(i)}(\boldsymbol{b})$ denote the ith ordered squared residual. Three of the most important robust estimators are defined below.

Definition 7.1. The *least quantile of squares* (LQS(c_n)) estimator minimizes the criterion

$$Q_{LQS}(\boldsymbol{b}) = r^2_{(c_n)}(\boldsymbol{b}). \tag{7.1}$$

When $c_n/n \to 1/2$, the LQS(c_n) estimator is also known as the *least median of squares* estimator (Hampel 1975, p. 380).

Definition 7.2. The *least trimmed sum of squares* (LTS(c_n)) estimator (Rousseeuw 1984) minimizes the criterion

$$Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}). \tag{7.2}$$

Definition 7.3. The *least trimmed sum of absolute deviations* (LTA(c_n)) estimator (Hössjer 1991) minimizes the criterion

$$Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^{c_n} |r(\boldsymbol{b})|_{(i)}. \tag{7.3}$$

These three estimators all find a set of fixed size $c_n = c_n(p) \geq n/2$ cases to cover, and then fit a classical estimator to the covered cases. LQS uses the Chebyshev fit, LTA uses L_1, and LTS uses OLS.

Definition 7.4. The integer valued parameter c_n is the *coverage* of the estimator. The remaining $n - c_n$ cases are given weight zero. In the literature and software,

$$c_n = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor \tag{7.4}$$

is often used as the default. Here $\lfloor x \rfloor$ is the greatest integer less than or equal to x. For example, $\lfloor 7.7 \rfloor = 7$.

Remark 7.1. Warning: In the literature, HB regression estimators seem to come in two categories. The first category consists of estimators that have no rigorous asymptotic theory but can be computed for very small data sets. The second category consists of estimators that have rigorous asymptotic theory but are impractical to compute. Due to the high computational complexity of these estimators, they are rarely used; however, the criterion are widely used for fast approximate algorithm estimators that can detect certain configurations of outliers. These approximations are typically inconsistent estimators with low breakdown. One of the most disappointing aspects of robust literature is that frequently no distinction is made between the impractical HB estimators and the inconsistent algorithm estimators used to detect outliers. Chapter 8 shows how to fix some of these algorithms so that the resulting estimator is $\sqrt{n}$ consistent and high breakdown.

7.2 Two Stage Estimators

The LTA and LTS estimators are very similar to trimmed means. If the coverage 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. There is a tradeoff in that the Gaussian efficiency of LTA and LTS seems to rapidly increase to that of the L_1 and OLS estimators, respectively, as τ tends to 1, but the breakdown value $1 - \tau$ decreases to 0. 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 $ALTx(\tau)$ estimator.

Many algorithms use K_n randomly selected "elemental" subsets of p cases called a "start," from which the residuals are computed for all n cases. The efficiency and resistance properties of the ALTx estimator depend strongly on the number of starts K_n used. Chapter 8 describes such approximations in much greater detail.

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, in 2000 *Splus* 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.

Similar issues occur in the location model: as the trimming proportion α decreases, the Gaussian efficiency of the α trimmed mean increases to 1, but the breakdown value decreases to 0. Chapters 2 and 4 described the following procedure for obtaining a robust two stage trimmed mean. The metrically trimmed mean M_n computes the sample mean of the cases in the interval

$$[\text{MED}(n) - k\text{MAD}(n), \text{MED}(n) + k\text{MAD}(n)]$$

where MED(n) is the sample median and MAD(n) is the sample median absolute deviation. A convenient value for the trimming constant is $k = 6$. Next, find the percentage of cases trimmed to the left and to the right by M_n, and round both percentages up to the nearest integer and compute the corresponding trimmed mean. Let $T_{A,n}$ denote the resulting estimator. For example, if M_n trimmed the 7.3% smallest cases and the 9.76% largest cases, then the final estimator $T_{A,n}$ is the (8%, 10%) trimmed mean. $T_{A,n}$ is asymptotically equivalent to a sequence of trimmed means with an asymptotic

variance that is easy to estimate.

To obtain a regression generalization of the two stage trimmed mean, compute the ALTx(c_n) estimator where $c_n \equiv c_{n,1}$ is given by Equation (7.4). Consider a finite number L of coverages $c_{n,1}$ and $c_{n,j} = \lfloor \tau_j \ n \rfloor$ where $j = 2, ..., L$ and $\tau_j \in G$. We suggest using $L = 5$ and $G = \{0.5, 0.75, 0.90, 0.99, 1.0\}$. The exact coverages c are defined by $c_{n,1} \equiv c_n$, $c_{n,2} = \lfloor .75 \ n \rfloor$, $c_{n,3} = \lfloor .90 \ n \rfloor$, $c_{n,4} = \lfloor .99 \ n \rfloor$, and $c_{n,5} = n$. (This choice of L and G is illustrative. Other choices, such as $G = \{0.5, 0.6, 0.7, 0.75, 0.9, 0.99, 1.0\}$ and $L = 7$, could be made.)

Definition 7.5. The RLTx(k) estimator is the ALTx(τ_R) estimator where τ_R is the largest $\tau_j \in G$ such that $\lfloor \tau_j \ n \rfloor \leq C_n(\hat{\boldsymbol{\beta}}_{ALTx(c_n)})$ where

$$C_n(\boldsymbol{b}) = \sum_{i=1}^{n} I[|r|_{(i)}(\boldsymbol{b}) \leq k \ |r|_{(c_n)}(\boldsymbol{b})] = \sum_{i=1}^{n} I[r^2_{(i)}(\boldsymbol{b}) \leq k^2 \ r^2_{(c_n)}(\boldsymbol{b})]. \quad (7.5)$$

The two stage trimmed mean inherits the breakdown value of the median and the stability of a trimmed mean with a low trimming proportion. The RLTx estimator can be regarded as an extension of the two stage mean to regression. The RLTx estimator inherits the high breakdown value of the ALTx(0.5) estimator, and the stability of the ALTx(τ_R) where τ_R is typically close to one.

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 generally has higher coverage and therefore higher statistical efficiency than ALTx(0.5). Notice that although L estimators ALTx($c_{n,j}$) were defined, only two are needed: ALTx(0.5) to get a resistant scale and define the coverage needed, and the final estimator ALTx(τ_R). The computational load is typically less than twice that of computing the ALTx(0.5) estimator since the computational complexity of the ALTx(τ) estimators decreases as τ increases from 0.5 to 1.

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 $\tau_R \in G$ and compute the RLTx estimator. (The RLTx estimator uses C_n in a manner analogous to the way that the two stage mean uses $k\text{MAD}(n)$.) 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. On the other hand, 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 the ALTx(0.5) estimator. Theoretical results and a simulation study, based on Olive and Hawkins (2003) and presented in Sections 7.4 and 7.5, suggest that the RLTx estimator is simultaneously more stable and more resistant than the ALTx(0.75 n) estimator for x = A and S. Increasing the coverage for the LQS criterion is not suggested since the Chebyshev fit tends to have less efficiency than the LMS fit.

7.3 Estimators with Adaptive Coverage

Estimators with adaptive coverage (EAC estimators) are also motivated by the idea of varying the coverage to better model the data set, but differ from the RLTx estimators in that they move the determination of the covered cases "inside the loop". Let c_n and C_n be given by (7.4) and (7.5). Hence

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

Definition 7.6. The *least adaptive quantile of squares* (LATQ(k)) estimator is the L_∞ fit that minimizes

$$Q_{LATQ}(\boldsymbol{b}) = r^2_{(C_n(\boldsymbol{b}))}(\boldsymbol{b}).$$

The *least adaptively trimmed sum of squares* (LATS(k)) estimator is the OLS fit that minimizes

$$Q_{LATS}(\boldsymbol{b}) = \sum_{i=1}^{C_n(\boldsymbol{b})} r^2_{(i)}(\boldsymbol{b}).$$

The *least adaptively trimmed sum of absolute deviations* (LATA(k)) estimator is the L_1 fit that minimizes

$$Q_{LATA}(\boldsymbol{b}) = \sum_{i=1}^{C_n(\boldsymbol{b})} |r|_{(i)}(\boldsymbol{b}).$$

Note that the adaptive estimators reduce to the highest breakdown versions of the fixed coverage estimators if $k = 1$ and (provided there is no exact fit to at least c_n of the cases) to the classical estimators if $k = \infty$.

These three adaptive coverage estimators simultaneously achieve a high breakdown value with high coverage, as do the RLTx estimators, but there are important outlier configurations where the resistance of the two estimators differs. The notation LATx will sometimes be used.

7.4 Theoretical Properties

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(0, V(\hat{\boldsymbol{\beta}}, F)\ \boldsymbol{W}) \tag{7.6}$$

when

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \rightarrow \boldsymbol{W}^{-1},$$

and when the errors e_i are iid with a cdf F and a unimodal pdf f that is symmetric with a unique maximum at 0. When the variance $V(e_i)$ exists,

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

See Koenker and Bassett (1978) and 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.

Remark 7.2. Obtaining asymptotic theory for LTA and LTS is a very challenging problem. Mašïček (2004), Čížek (2006) and Víšek (2006) claim to have shown asymptotic normality of LTS under general conditions. For the location model, Yohai and Maronna (1976) and Butler (1982) derived asymptotic theory for LTS while Tableman (1994ab) derived asymptotic theory for LTA. Shorack (1974) and Shorack and Wellner (1986, section 19.3) derived the asymptotic theory for a large class of location estimators that use random coverage (as do many others). 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, and Davies 1993, p. 1897), and it is known that scale estimators based on regression residuals behave well (see Welsh 1986).

Negative results are easily obtained. 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.

The asymptotic theory for RLTx depends on that for ALTx. **Most ALTx implementations have terrible statistical properties,** but an exception is the easily computed $\sqrt{n}$ consistent HB CLTS estimator given in Theorem 8.8 (and Olive and Hawkins 2007b, 2008). The following lemma can be used to estimate the coverage of the RLTx estimator given the error distribution F.

Lemma 7.1. 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

$$\frac{C_n(\hat{\boldsymbol{\beta}}_n)}{n} \xrightarrow{P} \tau_F \equiv \tau_F(k) = F(k\ F^{-1}(0.75)) - F(-k\ F^{-1}(0.75)). \qquad (7.7)$$

Proof. 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 the proportion of such cases can be made arbitrarily close to one as n increases to ∞ by increasing M. QED

Under the same conditions of Lemma 7.1,

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

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 7.2 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. Theorem 7.2 shows that the RLTQ estimator has an $O_P(n^{-1/3})$ rate if the exact LTQ estimator is used, but this estimator would be impractical to compute. ALTS could be the CLTS estimator of Theorem 8.8, but the resulting RLTS estimator is inferior to the CLTS estimator.

Theorem 7.2. 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 7.3 shows that the RLTx estimator is asymptotically equivalent to an ALTx estimator that typically has high coverage.

Theorem 7.3. 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.

The next theorem gives a case where RLTx can have an $O_P(n^{-1/2})$ convergence rate even though the ALTx(0.5) rate is poor. The result needs to be modified slightly for uniform data since then the ALTx constant is not consistent even if the slopes are.

Theorem 7.4. Assume that the conditions of Lemma 7.1 hold, that the predictors are bounded, and that the errors e_i have support on $[-d, d]$. If the ALTx(0.5) estimators are consistent and if $k > d/F^{-1}(0.75)$, then the RLTx estimators are asymptotically equivalent to the L_1, L_∞, and OLS estimators for x = A, Q, and S respectively.

Proof. The proof of Lemma 7.1 shows that $k|r|_{(c_n)}(\boldsymbol{b})$ converges to $kF^{-1}(0.75) > d$ where $\boldsymbol{b}$ is the ALTx(0.5) estimator and that the residuals r_i converge to the errors e_i. Hence the coverage proportion converges to one in probability. QED

Choosing a suitable k for a target distribution F is simple. Assume Equation (7.7) 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

$$P(-4\sigma < e_i < 4\sigma, i = 1, ..., 1000) \approx (0.9999)^{1000} > 0.90.$$

On the other hand, $|r|_{(c_n)}$ is converging to $\Phi^{-1}(0.75)\sigma \approx 0.67\sigma$. Hence if $k \geq 6.0$ and $n < 1000$, the RLTS estimator will cover all cases with high probability if the errors are Gaussian. 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 $n < 200$.

Proposition 7.5: Breakdown of LTx, RLTx, and LATx Estimators. LMS(τ), LTS(τ), and LTA(τ) have breakdown value

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

The breakdown value for the LATx estimators is 0.5, and the breakdown value for the RLTx estimators is equal to the breakdown value of the ALTx(c_n) estimator.

The breakdown results for the LTx estimators are well known. 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. See Section 9.4 for the definition of breakdown.

Theorem 7.6. Under regularity conditions similar to those in Conjecture 7.1 below,

a) the LMS(τ) converges at a cubed root rate to a non-Gaussian limit.
b) The estimator $\hat{\boldsymbol{\beta}}_{LTS}$ satisfies Equation (7.6) and

$$V(LTS(\tau), F) = \frac{\int_{F^{-1}(1/2-\tau/2)}^{F^{-1}(1/2+\tau/2)} w^2 dF(w)}{[\tau - 2F^{-1}(1/2+\tau/2) f(F^{-1}(1/2+\tau/2))]^2}. \tag{7.8}$$

The proof of Theorem 7.6a is given in Davies (1990) and Kim and Pollard (1990). Also see Davies (1993, p. 1897). The proof of b) is given in Mašiček (2004), Čížek (2006) and Víšek (2006).

Conjecture 7.1. Let the iid errors e_i have a cdf F that is continuous and strictly increasing on its interval support with a symmetric, unimodal, differentiable density f that strictly decreases as $|x|$ increases on the support.

Then the estimator $\hat{\boldsymbol{\beta}}_{LTA}$ satisfies Equation (7.6) and

$$V(LTA(\tau), F) = \frac{\tau}{4[f(0) - f(F^{-1}(1/2+\tau/2))]^2}. \tag{7.9}$$

See Tableman (1994b, p. 392) and Hössjer (1994).

As $\tau \to 1$, the efficiency of LTS approaches that of OLS and the efficiency of LTA approaches that of L_1. Hence for τ close to 1, LTA will be more efficient than LTS when the errors come from a distribution for which the

sample median is more efficient than the sample mean (Koenker and Bassett, 1978). The results of Oosterhoff (1994) suggest that when $\tau = 0.5$, LTA will be more efficient than LTS only for sharply peaked distributions such as the double exponential. To simplify computations for the asymptotic variance of LTS, we will use truncated random variables (see Definition 2.17).

Lemma 7.7. Under the symmetry conditions given in Conjecture 7.1,

$$V(LTS(\tau), F) = \frac{\tau \sigma^2_{TF}(-k, k)}{[\tau - 2kf(k)]^2} \tag{7.10}$$

and

$$V(LTA(\tau), F) = \frac{\tau}{4[f(0) - f(k)]^2} \tag{7.11}$$

where

$$k = F^{-1}(0.5 + \tau/2). \tag{7.12}$$

Proof. Let W have cdf F and pdf f. Suppose that W is symmetric about zero, and by symmetry, $k = F^{-1}(0.5 + \tau/2) = -F^{-1}(0.5 - \tau/2)$. If W has been truncated at $a = -k$ and $b = k$, then the variance of the truncated random variable W_T by

$$\text{VAR}(W_T) = \sigma^2_{TF}(-k, k) = \frac{\int_{-k}^{k} w^2 dF(w)}{F(k) - F(-k)}$$

by Definition 2.17. Hence

$$\int_{F^{-1}(1/2-\tau/2)}^{F^{-1}(1/2+\tau/2)} w^2 dF(w) = \tau \sigma^2_{TF}(-k, k)$$

and the result follows from the definition of k.

This result is useful since formulas for the truncated variance have been given in Chapter 4. The following examples illustrate the result. See Hawkins and Olive (1999b).

Example 7.1: N(0,1) Errors. If Y_T is a $N(0, \sigma^2)$ truncated at $a = -k\sigma$ and $b = k\sigma$, $\text{VAR}(Y_T) =$

$$\sigma^2[1 - \frac{2k\phi(k)}{2\Phi(k) - 1}].$$

At the standard normal

$$V(LTS(\tau), \Phi) = \frac{1}{\tau - 2k\phi(k)} \tag{7.13}$$

while

$$V(LTA(\tau), \Phi) = \frac{\tau}{4[\phi(0) - \phi(k)]^2} = \frac{2\pi\tau}{4[1 - \exp(-k^2/2)]^2} \tag{7.14}$$

where ϕ is the standard normal pdf and

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

Thus for $\tau \geq 1/2$, LTS(τ) has breakdown value of $1 - \tau$ and Gaussian efficiency

$$\frac{1}{V(LTS(\tau), \Phi)} = \tau - 2k\phi(k). \tag{7.15}$$

The 50% breakdown estimator LTS(0.5) has a Gaussian efficiency of 7.1%. If it is appropriate to reduce the amount of trimming, we can use the 25% breakdown estimator LTS(0.75) which has a much higher Gaussian efficiency of 27.6% as reported in Ruppert (1992, p. 255). Also see the column labeled "Normal" in table 1 of Hössjer (1994).

Example 7.2: Double Exponential Errors. The double exponential (Laplace) distribution is interesting since the L_1 estimator corresponds to maximum likelihood and so L_1 beats OLS, reversing the comparison of the normal case. For a double exponential $DE(0, 1)$ random variable,

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

while

$$V(LTA(\tau), DE(0, 1)) = \frac{\tau}{4[0.5 - 0.5\exp(-k)]^2} = \frac{1}{\tau}$$

where $k = -\log(1 - \tau)$. Note that LTA(0.5) and OLS have the same asymptotic efficiency at the double exponential distribution. Also see Tableman (1994ab).

Example 7.3: Cauchy Errors. Although the L_1 estimator and the trimmed estimators have finite variance when the errors are Cauchy, the

OLS estimator has infinite variance (because the Cauchy distribution has infinite variance). If X_T is a Cauchy $C(0,1)$ random variable symmetrically truncated at $-k$ and k, then

$$\mathrm{VAR}(X_T) = \frac{k - \tan^{-1}(k)}{\tan^{-1}(k)}.$$

Hence

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

and

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

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

7.5 Computation and Simulations

In addition to the LMS estimator, there are at least two other regression estimators, the *least quantile of differences* (LQD) and the *regression depth* estimator, that have rather high breakdown and rigorous asymptotic theory. The LQD estimator is the LMS estimator computed on the $(n-1)n/2$ pairs of case difference (Croux, Rousseeuw and Hössjer 1994). The regression depth estimator (Rousseeuw and Hubert 1999) is interesting because its criterion does not use residuals. The large sample theory for the depth estimator is given by Bai and He (1999). The LMS, LTS, LTA, LQD and depth estimators can be computed exactly only if the data set is tiny.

Proposition 7.8. a) There is an LTS(c) estimator $\hat{\boldsymbol{\beta}}_{LTS}$ that is the OLS fit to the cases corresponding to the c smallest LTS squared residuals.
b) There is an LTA(c) estimator $\hat{\boldsymbol{\beta}}_{LTA}$ that is the L_1 fit to the cases corresponding to the c smallest LTA absolute residuals.
c) There is an LQS(c) estimator $\hat{\boldsymbol{\beta}}_{LQS}$ that is the Chebyshev fit to the cases corresponding to the c smallest LQS absolute residuals.

Proof. a) By the definition of the LTS(c) estimator,

$$\sum_{i=1}^{c} r^2_{(i)}(\hat{\boldsymbol{\beta}}_{LTS}) \leq \sum_{i=1}^{c} r^2_{(i)}(\boldsymbol{b})$$

where $\boldsymbol{b}$ is any $p \times 1$ vector. Without loss of generality, assume that the cases have been reordered so that the first c cases correspond to the cases with the c smallest residuals. Let $\hat{\boldsymbol{\beta}}_{OLS}(c)$ denote the OLS fit to these c cases. By the definition of the OLS estimator,

$$\sum_{i=1}^{c} r^2_i(\hat{\boldsymbol{\beta}}_{OLS}(c)) \leq \sum_{i=1}^{c} r^2_i(\boldsymbol{b})$$

where $\boldsymbol{b}$ is any $p \times 1$ vector. Hence $\hat{\boldsymbol{\beta}}_{OLS}(c)$ also minimizes the LTS criterion and thus $\hat{\boldsymbol{\beta}}_{OLS}(c)$ is an LTS estimator. The proofs of b) and c) are similar. QED

Definition 7.7. In regression, an *elemental set* is a set of p cases.

One way to compute these estimators exactly is to generate all $C(n, c)$ subsets of size c, compute the classical estimator $\boldsymbol{b}$ on each subset, and find the criterion $Q(\boldsymbol{b})$. The robust estimator is equal to the $\boldsymbol{b}_o$ that minimizes the criterion. Since $c \approx n/2$, this algorithm is impractical for all but the smallest data sets. Since the L_1 fit is an elemental fit, the LTA estimator can be found by evaluating all $C(n, p)$ elemental sets. See Hawkins and Olive (1999b). Since any Chebyshev fit is also a Chebyshev fit to a set of $p + 1$ cases, the LQS estimator can be found by evaluating all $C(n, p+1)$ cases. See Stromberg (1993ab) and Appa and Land (1993). The LMS, LTA, and LTS estimators can also be evaluated exactly using branch and bound algorithms if the data set size is small enough. See Agulló (1997, 2001).

Typically HB algorithm estimators should not be used unless the criterion complexity is $O(n)$. The complexity of the estimator depends on how many fits are computed and on the complexity of the criterion evaluation. For example the LMS and LTA criteria have $O(n)$ complexity while the depth criterion complexity is $O(n^{p-1} \log n)$. The LTA and depth estimators evaluates $O(n^p)$ *elemental sets* while LMS evaluates the $O(n^{p+1})$ subsets of size $p+1$. The LQD criterion complexity is $O(n^2)$ and evaluates $O(n^{2(p+1)})$ subsets of case distances.

Consider the algorithm that takes a subsample of n^δ cases and then computes the exact algorithm to this subsample. Then the complexities

of the LTA, LMS, depth and LQD algorithms are $O(n^{\delta(p+1)})$, $O(n^{\delta(p+2)})$, $O(n^{\delta(2p-1)} \log n^{\delta})$ and $O(n^{\delta(2p+4)})$, respectively. The convergence rates of the estimators are $n^{\delta/3}$ for LMS and $n^{\delta/2}$ for the remaining three estimators (if the LTA estimator does indeed have the conjectured $\sqrt{n}$ convergence rate). These algorithms rapidly become impractical as n and p increase. For example, if $n = 100$ and $\delta = 0.5$, use $p < 7, 6, 4, 2$ for these LTA, LMS, depth, and LQD algorithms respectively. If $n = 10000$, this LTA algorithm may not be practical even for $p = 3$. These results suggest that the LTA and LMS approximations will be more interesting than depth or LQD approximations unless a computational breakthrough is made for the latter two estimators.

We simulated LTA and LTS for the location model using normal, double exponential, and Cauchy error models. For the location model, these estimators can be computed exactly: find the order statistics

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$$

of the data. For LTS compute the sample mean and for LTA compute the sample median (or the low or high median) and evaluate the LTS and LTA criteria of each of the $n-c+1$ "c-samples" $Y_{(i)}, \ldots, Y_{(i+c-1)}$, for $i = 1, \ldots, n-c+1$. The minimum across these samples then defines the LTA and LTS estimates.

We computed the sample standard deviations of the resulting location estimate from 1000 runs of each sample size studied. The results are shown in Table 7.1. For Gaussian errors, the observed standard deviations are smaller than the asymptotic standard deviations but for the double exponential errors, the sample size needs to be quite large before the observed standard deviations agree with the asymptotic theory.

Table 7.2 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), double exponential, 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 mean of 5.5 was intended as a case where the ALTS(0.5) estimator should outperform the RLTS estimator, as these contaminants are just small enough that many pass the $k = 6$ screen, and the mean of 12 to test how the estimators handled catastrophic contamination.

Table 7.1: Monte Carlo Efficiencies Relative to OLS.

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

The simulation used $n = 100$ and $p = 6$ (5 slopes and an intercept) over 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 – a traditional elemental algorithm with 3000 starts and a concentration algorithm (see Chapter 8). 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.

All of the estimators in this small study are inconsistent zero breakdown estimators, but some are useful for detecting outliers. (A better choice than the inconsistent estimators is to use the easily computed $\sqrt{n}$ consistent HB CLTS estimator given in Theorem 8.8.) The concentration algorithm 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

Table 7.2: $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2/p$, 1000 runs

pop.-alg.	ALTS (.5)	ALTS (.75)	ALTS (.9)	RLTS (6)	% of runs that τ_R = .5,.75,.9,.99 or 1
N(0,1)-conc	0.0648	0.0350	0.0187	0.0113	0,0,6,18,76
DE(0,1)-conc	0.1771	0.0994	0.0775	0.0756	0,0,62,23,15
U$(-1,1)$-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
N(0,1)-elem	0.1391	0.1163	0.1051	0.0975	0,0,1,6,93
DE(0,1)-elem	0.9268	0.8051	0.7694	0.7522	0,0,20,28,52
U$(-1,1)$-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

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 suggests that the RLTx method with concentration is a better approach for improving the resistance and performance of the inconsistent *Splus* `ltsreg` estimator than increasing the coverage from 50% to 90%. The simulation also suggests that even the inconsistent version of RLTx used

in the study is useful for detecting outliers. The concentration RLTx estimator would be improved if $\max(n, 500)$ starts were used instead of 50 starts. Although the easily computed $\sqrt{n}$ consistent HB CLTS estimator of Theorem 8.8 can be used to make a $\sqrt{n}$ consistent HB RLTS estimator (as soon as the CLTS estimator is available from the software), the CLTS estimator may be superior to the resulting RLTS estimator.

7.6 Resistant Estimators

Definition 7.8. A regression estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is a *resistant estimator* if $\hat{\boldsymbol{\beta}}$ is known to be useful for detecting certain types of outliers. (Often we also require $\hat{\boldsymbol{\beta}}$ to be a consistent estimator of $\boldsymbol{\beta}$.)

Typically resistant estimators are useful when the errors are iid from a heavy tailed distribution. Some examples include the L_1 estimator, which can be useful for detecting Y-outliers, and some M, R, GM, and GR estimators. M-estimators tend to obtain a tradeoff between the resistance of the L_1 estimator and the Gaussian efficiency of the OLS estimator. This tradeoff is especially apparent with the *Huber* M-estimator. Street, Carroll, and Ruppert (1988) discuss the computation of standard errors for M-estimators. R-estimators have elegant theory similar to that of OLS, and the Wilcoxon rank estimator is especially attractive. See Hettmansperger and McKean (1998, ch. 3). GM-estimators are another large class of estimators. Carroll and Welsh (1988) claim that **only the Mallows class of GM-estimators are consistent for slopes if the errors are asymmetric**. Also see Simpson, Ruppert, and Carroll (1992, p. 443). The Mallows estimator may have a breakdown value as high as $1/(p+1)$. A discussion of GR-estimators is in Hettmansperger and McKean (1998, ch. 5). The resistant trimmed views estimator (`tvreg`) is presented in Section 11.3.

For illustration, we will construct a simple resistant algorithm estimator, called the *median ball algorithm* (MBA or `mbareg`). The Euclidean distance of the ith vector of predictors $\boldsymbol{x}_i$ from the jth vector of predictors $\boldsymbol{x}_j$ is

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

For a fixed $\boldsymbol{x}_j$ consider the ordered distances $D_{(1)}(\boldsymbol{x}_j), ..., D_{(n)}(\boldsymbol{x}_j)$. Next, let $\hat{\boldsymbol{\beta}}_j(\alpha)$ denote the OLS fit to the $\min(p + 3 + \lfloor \alpha n/100 \rfloor, n)$ cases with

the smallest distances where the approximate percentage of cases used is $\alpha \in \{1, 2.5, 5, 10, 20, 33, 50\}$. (Here $\lfloor x \rfloor$ is the greatest integer function so $\lfloor 7.7 \rfloor = 7$. The extra $p+3$ cases are added so that OLS can be computed for small n and α.) This yields seven OLS fits corresponding to the cases with predictors closest to $\boldsymbol{x}_j$. A fixed number K of cases are selected at random without replacement to use as the $\boldsymbol{x}_j$. Hence $7K$ OLS fits are generated. We use $K = 7$ as the default. A robust criterion Q is used to evaluate the $7K$ fits and the OLS fit to all of the data. Hence $7K + 1$ OLS fits are generated and the MBA estimator is the fit that minimizes the criterion. The median squared residual, the LTA criterion, and the LATA criterion are good choices for Q. Replacing the $7K + 1$ OLS fits by L_1 fits increases the resistance of the MBA estimator.

Three ideas motivate this estimator. First, $\boldsymbol{x}$-outliers, which are outliers in the predictor space, tend to be much more destructive than Y-outliers which are outliers in the response variable. Suppose that the proportion of outliers is γ and that $\gamma < 0.5$. We would like the algorithm to have at least one "center" $\boldsymbol{x}_j$ that is not an outlier. The probability of drawing a center that is not an outlier is approximately $1 - \gamma^K > 0.99$ for $K \geq 7$ and this result is free of p. Secondly, by using the different percentages of coverages, for many data sets there will be a center and a coverage that contains no outliers.

Thirdly, the MBA estimator is a $\sqrt{n}$ consistent estimator. To see this, assume that n is increasing to ∞. For each center $\boldsymbol{x}_{j,n}$ there are 7 spheres centered at $\boldsymbol{x}_{j,n}$. Let $r_{j,h,n}$ be the radius of the hth sphere with center $\boldsymbol{x}_{j,n}$. Fix an extremely large N such that for $n \geq N$ these $7K$ regions in the predictor space are fixed. Hence for $n \geq N$ the centers are $\boldsymbol{x}_{j,N}$ and the radii are $r_{j,h,N}$ for $j = 1, ..., K$ and $h = 1, ..., 7$. Since only a fixed number $(7K + 1)$ of $\sqrt{n}$ consistent fits are computed, the final estimator is also a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$, regardless of how the final estimator is chosen (by Pratt 1959).

Section 11.3 will compare the MBA estimator with other resistant estimators including the *R/Splus* estimator `lmsreg` and the *trimmed views* estimator. *Splus* also contains other regression estimators (such as `ltsreg`, `lmRobMM` and `rreg`), but the current (as of 2000) implementations of `ltsreg` and `rreg` are not very useful for detecting outliers. Section 6.3 suggested using resistant estimators in RR and FF plots to detect outliers. Chapter 8 discusses some of the more conventional algorithms that have appeared in the literature.

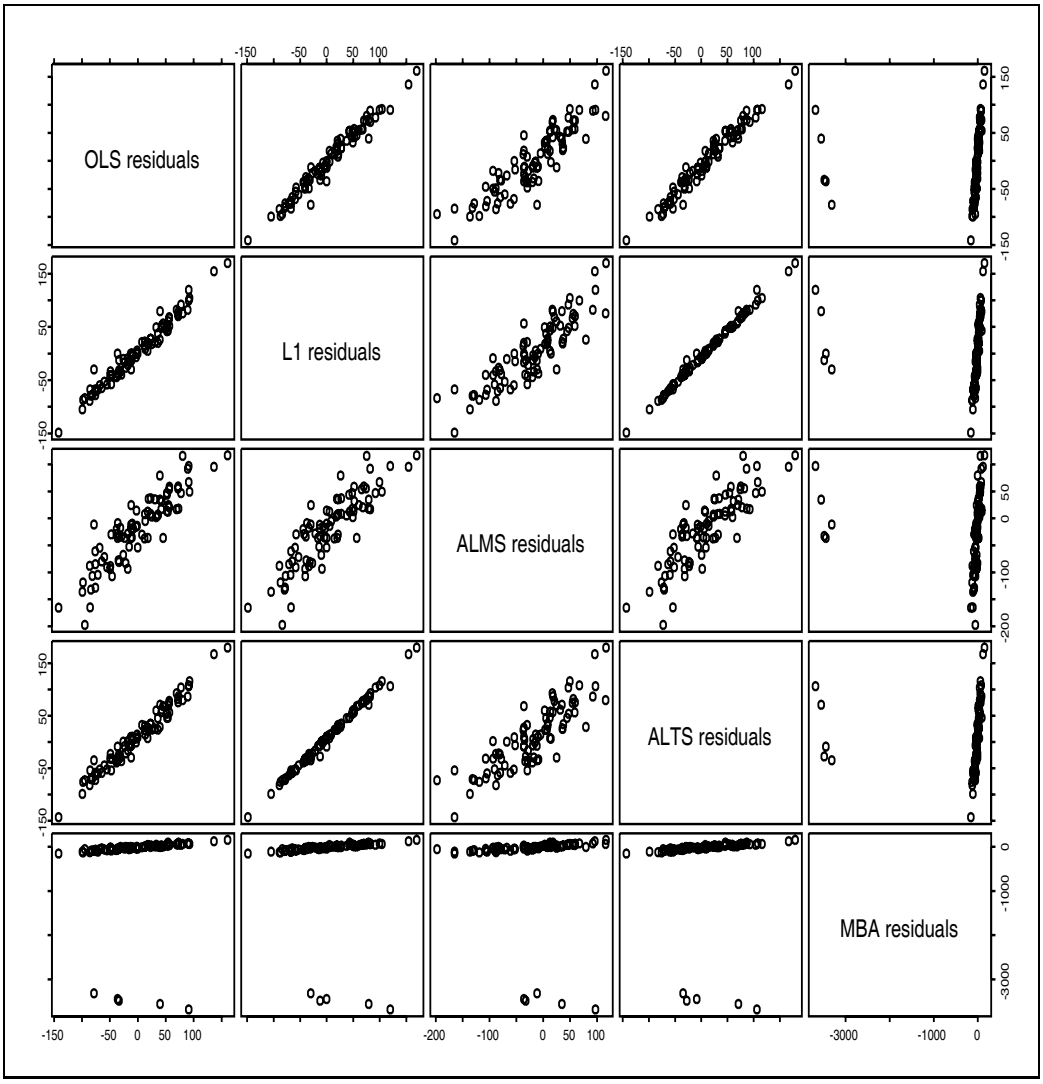

Figure 7.1: RR plot for the Buxton Data

Example 7.4. Buxton (1920, p. 232-5) gives 20 measurements of 88 men. *Height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, numbers 62–66, were reported to be about 0.75 inches tall with head lengths well over five feet! Figure 7.1 shows the RR plot for the *Splus 2000* estimators `lsfit`, `l1fit`, `lmsreg`, `ltsreg` and the MBA estimator. Note that only the MBA estimator gives large absolute residuals to the outliers. One feature of the MBA estimator is that it depends on the sample of 7 centers drawn and changes each time the function is called. In ten runs, about seven plots will look like Figure 7.1, but in about three plots the MBA estimator will also pass through the outliers.

7.7 Complements

The LTx and LATx estimators discussed in this chapter are not useful for applications because they are impractical to compute; however, the criterion are useful for making resistant or robust algorithm estimators. In particular the robust criterion are used in the MBA algorithm (see Problem 7.5) and in the easily computed $\sqrt{n}$ consistent HB CLTS estimator described in Theorem 8.8 and in Olive and Hawkins (2007b, 2008).

Section 7.3 is based on Olive and Hawkins (1999) while Sections 7.2, 7.4, 7.5 and 7.6 follow Hawkins and Olive (1999b), Olive and Hawkins (2003) and Olive (2005).

Several HB regression estimators are well known, and perhaps the first proposed was the least median of squares (LMS) estimator. See Hampel (1975, p. 380). For the location model, Yohai and Maronna (1976) and Butler (1982) derived asymptotic theory for LTS. Rousseeuw (1984) generalized the location LTS estimator to the LTS regression estimator and the minimum covariance determinant estimator for multivariate location and dispersion (see Chapter 10). Bassett (1991) suggested the LTA estimator for location and Hössjer (1991) suggested the LTA regression estimator.

Two stage regression estimators compute a high breakdown regression (or multivariate location and dispersion) estimator in the first stage. The initial estimator is used to weight cases or as the initial estimator in a one

step Newton's method procedure. The goal is for the two stage estimator to inherit the outlier resistance properties of the initial estimator while having high asymptotic efficiency when the errors follow a zero mean Gaussian distribution. The theory for many of these estimators is often rigorous, but the estimators are even less practical to compute than the initial estimators. There are dozens of references including Jureckova and Portnoy (1987), Simpson, Ruppert and Carroll (1992), Coakley and Hettmansperger (1993), Chang, McKean, Naranjo and Sheather (1999), and He, Simpson and Wang (2000). The "cross checking estimator," see He and Portnoy (1992, p. 2163) and Davies (1993, p. 1981), computes a high breakdown estimator and OLS and uses OLS if the two estimators are sufficiently close.

The easily computed HB CLTS estimator from Theorem 8.8 makes two stage estimators such as the cross checking estimator practical for the first time. However, CLTS is asymptotically equivalent to OLS, so the cross checking step is not needed.

The theory of the RLTx estimator is very simple, but it can be used to understand other results. For example, Theorem 7.3 will hold as long as the initial estimator $\boldsymbol{b}$ used to compute C_n is consistent. Suppose that the easily computed $\sqrt{n}$ consistent HB CLTS estimator $\boldsymbol{b}$ (from Theorem 8.8) is used. The CLTS(0.99) estimator is asymptotically equivalent to OLS, so the RLTS estimator that uses $\boldsymbol{b}$ as the initial estimator will have high Gaussian efficiency. Similar results have appeared in the literature, but their proofs are very technical, often requiring the theory of empirical processes.

The major drawback of high breakdown estimators that have nice theoretical results such as high efficiency is that they tend to be impractical to compute. If an inconsistent zero breakdown initial estimator is used, as in most of the literature and in the simulation study in Section 7.5, then the final estimator (including even the simplest two stage estimators such as the cross checking and RLTx estimators) also has zero breakdown and is often inconsistent. Hence $\sqrt{n}$ consistent resistant estimators such as the MBA estimator often have higher outlier resistance than zero breakdown implementations of HB estimators such as `ltsreg`.

Another drawback of high breakdown estimators that have high efficiency is that they tend to have considerably more bias than estimators such as LTS(0.5) for many outlier configurations. For example the fifth row of Table 7.2 shows that the RLTS estimator can perform much worse than the ALTS(0.5) estimator if the outliers are within the $k = 6$ screen.

7.8 Problems

R/Splus Problems

Warning: Use the command *source("A:/rpack.txt")* to download the programs. See Preface or Section 14.2. Typing the name of the `rpack` function, eg *mbamv*, will display the code for the function. Use the `args` command, eg *args(mbamv)*, to display the needed arguments for the function.

7.1. a) Download the *R/Splus* function `nltv` that computes the asymptotic variance of the LTS and LTA estimators if the errors are N(0,1).
b) Enter the commands *nltv(0.5), nltv(0.75), nltv(0.9)* and *nltv(0.9999).* Write a table to compare the asymptotic variance of LTS and LTA at these coverages. Does one estimator always have a smaller asymptotic variance?

7.2. a) Download the *R/Splus* function `deltv` that computes the asymptotic variance of the LTS and LTA estimators if the errors are double exponential DE(0,1).
b) Enter the commands *deltv(0.5), deltv(0.75), deltv(0.9)* and *deltv(0.9999).* Write a table to compare the asymptotic variance of LTS and LTA at these coverages. Does one estimator always have a smaller asymptotic variance?

7.3. a) Download the *R/Splus* function `cltv` that computes the asymptotic variance of the LTS and LTA estimators if the errors are Cauchy C(0,1).

b) Enter the commands *cltv(0.5), cltv(0.75), cltv(0.9)* and *cltv(0.9999).* Write a table to compare the asymptotic variance of LTS and LTA at these coverages. Does one estimator always have a smaller asymptotic variance?

7.4*. a) If necessary, use the commands *source("A:/rpack.txt")* and *source("A:/robdata.txt").*

b) Enter the command *mbamv(belx,bely)* in *R/Splus.* Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this 7 times before the program ends. There is one predictor x and one response Y. The function makes a scatterplot of x and y and cases that get weight one are shown as highlighted squares. Each MBA sphere covers half of the data. When you find a good fit to the bulk of the data, hold down the *Ctrl* and c keys to make a copy of the plot. Then paste the plot in *Word.*

c) Enter the command *mbamv2(buxx,buxy)* in *R/Splus.* Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this 14 times before the program ends. There is one predictor x and one response Y. The function makes the response and residual plots based on the OLS fit to the

highlighted cases. Each MBA sphere covers half of the data. When you find a good fit to the bulk of the data, hold down the *Ctrl* and *c* keys to make a copy of the two plots. Then paste the plots in *Word*.

7.5*. This problem compares the MBA estimator that uses the median squared residual MED(r_i^2) criterion with the MBA estimator that uses the LATA criterion. On clean data, both estimators are $\sqrt{n}$ consistent since both use 50 $\sqrt{n}$ consistent OLS estimators. The MED(r_i^2) criterion has trouble with data sets where the multiple linear regression relationship is weak and there is a cluster of outliers. The LATA criterion tries to give all x–outliers, including good leverage points, zero weight.

a) If necessary, use the commands *source("A:/rpack.txt")* and *source("A:/robdata.txt")*. The `mlrplot2` function is used to compute both MBA estimators. Use the rightmost mouse button to advance the plot (and in *R*, highlight stop).

b) Use the command *mlrplot2(belx,bely)* and include the resulting plot in *Word*. Is one estimator better than the other, or are they about the same?

c) Use the command *mlrplot2(cbrainx,cbrainy)* and include the resulting plot in *Word*. Is one estimator better than the other, or are they about the same?

d) Use the command *mlrplot2(museum[,3:11],museum[,2])* and include the resulting plot in *Word*. For this data set, most of the cases are based on humans but a few are based on apes. The MBA LATA estimator will often give the cases corresponding to apes larger absolute residuals than the MBA estimator based on MED(r_i^2).

e) Use the command *mlrplot2(buxx,buxy)* until the outliers are clustered about the identity line in one of the two response plots. (This will usually happen within 10 or fewer runs. Pressing the "up arrow" will bring the previous command to the screen and save typing.) Then include the resulting plot in *Word*. Which estimator went through the outliers and which one gave zero weight to the outliers?

f) Use the command *mlrplot2(hx,hy)* several times. Usually both MBA estimators fail to find the outliers for this artificial Hawkins data set that is also analyzed by Atkinson and Riani (2000, section 3.1). The *lmsreg* estimator can be used to find the outliers. In *Splus*, use the command *ffplot(hx,hy)* and in *R* use the commands *library(MASS)* and *ffplot2(hx,hy)*. Include the resulting plot in *Word*.