# Chapter 6
# Robust and Resistant Regression

The brand name high breakdown regression estimators discussed in the last chapter take too long to compute, but the LMS, LTA, and LTS criteria are used in practical regression algorithms to screen attractors. The practical algorithms in the literature tend to be zero breakdown and inconsistent. Chapter 5 showed that the response plot is useful for detecting MLR outliers, defined MLR breakdown, and the MLR concentration algorithm. This chapter gives several practical outlier resistant MLR estimators that are $\sqrt{n}$ consistent.

## 6.1 Resistant Multiple Linear Regression

The first outlier resistant regression method was given by Application 3.3. Call the estimator the *MLD set MLR estimator*. Let the $i$th case $\boldsymbol{w}_i = (Y_i, \boldsymbol{x}_i^T)^T$ where the continuous predictors from $\boldsymbol{x}_i$ are denoted by $\boldsymbol{u}_i$ for $i = 1, ..., n$. Now let $D$ be the RMVN set $U$, the RFCH set $V$, or the `covmb2` set $B$. Find $D$ by applying the MLD estimator to the $\boldsymbol{u}_i$, and then run the MLR method on the $m$ cases $\boldsymbol{w}_i$ corresponding to the set $D$ indices $i_1, ..., i_m$, where $m \geq n/2$. The set $B$ can be used even if $p > n$. The theory of the MLR method applies to the cleaned data set since $Y$ was not used to pick the subset of the data. Efficiency can be much lower since $m$ cases are used where $n/2 \leq m \leq n$, and the trimmed cases tend to be the "farthest" from the center of $\boldsymbol{u}$. The *rpack* function `getu` gets the RMVN set $U$. See the following *R* code for the Buxton (1920) data where we could use the `covmb2` set $B$ instead of the RMVN set $U$ by replacing the command *getu(x)* by `getB(x)`.

```
Y <- buxy
x <- buxx
indx <- getu(x)$indx #u = x for this example
```

```
Yc <- Y[indx]
Xc <- x[indx,]
length(Y) - length(Yc) #the RMVN set (= cleaned data)
#omitted 4 inliers and 5 outliers
MLRplot(Xc,Yc) #right click Stop two times,
#response plot for cleaned data
out<-lsfit(Xc,Yc)
ESP <- x%*%out$coef[-1] + out$coef[1]
plot(ESP,Y)
abline(0,1)  #response plot using the resistant
#MLR estimator and all of the data
```

A good resistant estimator is the Olive (2005a) *median ball algorithm* (MBA or `mbareg`). The Euclidean distance of the $i$th vector of predictors $\boldsymbol{x}_i$ from the $j$th vector of predictors $\boldsymbol{x}_j$ is

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

For a fixed $\boldsymbol{x}_j$ consider the ordered distances $D_{(1)}(\boldsymbol{x}_j), ..., D_{(n)}(\boldsymbol{x}_j)$. Next, let $\hat{\boldsymbol{\beta}}_j(\alpha)$ denote the OLS fit to the $\min(p + 3 + \lfloor \alpha n/100 \rfloor, n)$ 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 $\alpha$.) This yields seven OLS fits corresponding to the cases with predictors closest to $\boldsymbol{x}_j$. A fixed number of $K$ 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 (attractors) are generated and the MBA estimator is the fit that minimizes the criterion. The median squared residual is a good choice for $Q$.

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 $\gamma$ 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. Third, the MBA estimator is a $\sqrt{n}$ consistent estimator of the same parameter vector $\boldsymbol{\beta}$ estimated by OLS under mild conditions on the zero mean error distribution. This result occurs since each of the $7K + 1$ attractors is $\sqrt{n}$ consistent when there are no outliers. See Remark 6.1 and Theorem 6.1.

Ellipsoidal trimming can be used to create resistant multiple linear regression (MLR) estimators. To perform ellipsoidal trimming, an estimator $(T, \boldsymbol{C})$ is computed and used to create the squared Mahalanobis distances $D_i^2$ for

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

$$\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x} - T) \leq D^2_{(j)}\}. \tag{6.1}$$

The $i$th case $(Y_i, \boldsymbol{x}_i^T)^T$ is trimmed if $D_i > D_{(j)}$. Then an estimator of $\boldsymbol{\beta}$ is computed from the remaining cases. For example, if $j \approx 0.9n$, then about 10% of the cases are trimmed, and OLS or $L_1$ could be used on the cases that remain. Ellipsoidal trimming differs from the *MLD set MLR estimator* that uses the MLD set on the $\boldsymbol{x}_i$, since the MLD set uses a random amount of trimming. (The ellipsoidal trimming technique can also be used for other regression models, and the theory of the regression method tends to apply to the method applied to the cleaned data that was not trimmed since the response variables were not used to select the cases. See Chapter 9.)

Use ellipsoidal trimming on the RFCH, RMVN, or `covmb2` set applied to the continuous predictors to get a fit $\hat{\boldsymbol{\beta}}_C$. Then make a response and residual plot using all of the data, not just the cleaned data that was not trimmed.

The Olive (2005a) resistant trimmed views estimator combines ellipsoidal trimming and the response plot. First compute $(T, \boldsymbol{C})$ on the $\boldsymbol{x}_i$, perhaps using the RMVN estimator. Trim the $M$% of the cases with the largest Mahalanobis distances, and then compute the MLR estimator $\hat{\boldsymbol{\beta}}_M$ from the remaining cases. Use $M = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten response plots of the fitted values $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i$ versus $Y_i$ using all $n$ cases. (Fewer plots are used for small data sets if $\hat{\boldsymbol{\beta}}_M$ can not be computed for large $M$.) These plots are called "trimmed views." The TV estimator will also be called the `tvreg` estimator. Since each of the 10 attractors $\hat{\boldsymbol{\beta}}_M$ is $\sqrt{n}$ consistent, so is the TV estimator. See Theorem 6.1.

**Definition 6.1.** The trimmed views (TV) estimator $\hat{\boldsymbol{\beta}}_{T,n}$ corresponds to the trimmed view where the bulk of the plotted points follow the identity line with smallest variance function, ignoring any outliers.

**Example 6.1.** For the Buxton (1920) data, *height* was the response variable while an intercept, *head length, nasal height, bigonal breadth,* and *cephalic index* were used as predictors in the multiple linear regression model. Observation 9 was deleted since it had missing values. Five individuals, cases 61–65, were reported to be about 0.75 inches tall with head lengths well over five feet! OLS was used on the cases remaining after trimming, and Figure 6.1 shows four trimmed views corresponding to 90%, 70%, 40%, and 0% trimming. The OLS TV estimator used 70% trimming since this trimmed view was best. Since the vertical distance from a plotted point to the identity line is equal to the case's residual, the outliers had massive residuals for 90%, 70%, and 40% trimming. Notice that the OLS trimmed view with 0% trim-

ming "passed through the outliers" since the cluster of outliers is scattered about the identity line.

The TV estimator $\hat{\boldsymbol{\beta}}_{T,n}$ has good statistical properties if an estimator with good statistical properties is applied to the cases $(\boldsymbol{X}_{M,n}, \boldsymbol{Y}_{M,n})$ that remain after trimming. Candidates include OLS, $L_1$, Huber's M–estimator, Mallows' GM–estimator, or the Wilcoxon rank estimator. See Rousseeuw and Leroy (1987, pp. 12-13, 150). The basic idea is that if an estimator with $O_P(n^{-1/2})$ convergence rate is applied to a set of $n_M \propto n$ cases, then the resulting estimator $\hat{\boldsymbol{\beta}}_{M,n}$ also has $O_P(n^{-1/2})$ rate provided that the response $Y$ was not used to select the $n_M$ cases in the set. If $\|\hat{\boldsymbol{\beta}}_{M,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ for $M = 0, ..., 90$ then $\|\hat{\boldsymbol{\beta}}_{T,n} - \boldsymbol{\beta}\| = O_P(n^{-1/2})$ by Pratt (1959). See Theorems 6.1 and 11.17.

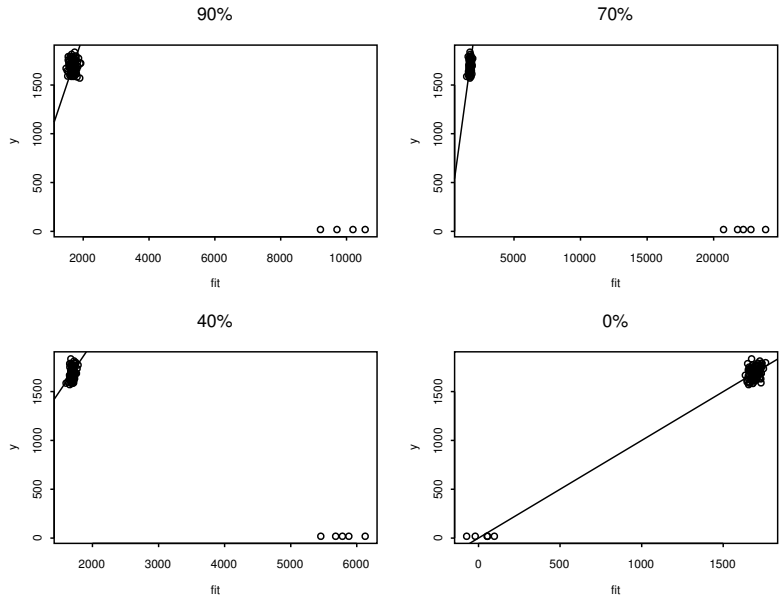


**Fig. 6.1** 4 Trimmed Views for the Buxton Data

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

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

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

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

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

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

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

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

**Remark 6.1.** When $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e$, MLR estimators tend to estimate the same slopes $\beta_2, ..., \beta_p$, but the constant $\beta_1$ tends to depend on the estimator unless the errors are symmetric. The MBA and trimmed views estimators do estimate the same $\boldsymbol{\beta}$ as OLS asymptotically, but samples may need to be huge before the MBA and trimmed views estimates of the constant are close to the OLS estimate of the constant. If the trimmed views estimator is modified so that the LTS, LTA, or LMS criterion is used to select the final estimator, then a conjecture is that the limiting distribution is similar to that of the variable selection estimator: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MTV} - \boldsymbol{\beta}) \xrightarrow{D} \sum_{i=1}^{k} \pi_i \boldsymbol{w}_i$ where $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{k} \pi_i = 1$. The index $i$ corresponds to the fits considered by the modified trimmed views estimator with $k = 10$. For the MBA estimator and the modified trimmed views estimator, the prediction region method, described in Section 7.5, may be useful for testing hypotheses. Large sample sizes may be needed if the error distribution is not symmetric since the constant $\hat{\boldsymbol{\beta}}_1$ needs large samples. See Olive (2017b, p. 444) for an explanation for why large sample sizes may be needed to estimate the constant.

### *6.1.1* The rmreg2 Estimator

The Olive (2017b) robust multiple linear regression estimator `rmreg2` is the classical multiple linear regression estimator applied to the RMVN set when RMVN is computed from the vectors $\boldsymbol{u}_i = (x_{i2}, ..., x_{ip}, Y_i)^T$ for $i = 1, ..., n$. Hence $\boldsymbol{u}_i$ is the $i$th case with $x_{i1} = 1$ deleted. This estimator is one of the most outlier resistant practical robust MLR estimators. The `rmreg2` estimator has been shown to be consistent if the $\boldsymbol{u}_i$ are iid from a large class of elliptically contoured distributions, which is a much stronger assumption than having iid errors $e_i$.

First we will review some results for multiple linear regression. Let $\boldsymbol{x} = (1, \boldsymbol{w}^T)^T$ and let

$$\mathrm{Cov}(\boldsymbol{w}) = \mathrm{E}[(\boldsymbol{w} - \mathrm{E}(\boldsymbol{w}))(\boldsymbol{w} - \mathrm{E}(\boldsymbol{w}))^{\mathrm{T}}] = \boldsymbol{\Sigma_w}$$

and $\mathrm{Cov}(\boldsymbol{w}, Y) = E[(\boldsymbol{w} - E(\boldsymbol{w}))(Y - E(Y))] = \boldsymbol{\Sigma_{wY}}$. Let $\boldsymbol{\beta} = (\alpha, \boldsymbol{\eta}^T)^T$ be the population OLS coefficients from the regression of $Y$ on $\boldsymbol{x}$ ($\boldsymbol{w}$ and a constant), where $\alpha$ is the constant and $\boldsymbol{\eta}$ is the vector of slopes. Let the OLS estimator be $\hat{\boldsymbol{\beta}} = (\hat{\alpha}, \hat{\boldsymbol{\eta}}^T)^T$. Then the population coefficients from an OLS regression of $Y$ on $\boldsymbol{x}$ are

$$\alpha = E(Y) - \boldsymbol{\eta}^T E(\boldsymbol{w}) \text{ and } \boldsymbol{\eta} = \boldsymbol{\Sigma_w^{-1}} \boldsymbol{\Sigma_{w\mathrm{Y}}}. \tag{6.2}$$

Then the OLS estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. The sample covariance matrix of $\boldsymbol{w}$ is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T \text{ where the sample mean } \overline{\boldsymbol{w}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}} \boldsymbol{w}_{\mathrm{i}}.$$

Similarly, define the sample covariance vector of $\boldsymbol{w}$ and $Y$ to be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{w}_i - \overline{\boldsymbol{w}})(Y_i - \overline{Y}).$$

Suppose that $(Y_i, \boldsymbol{w}_i^T)^T$ are iid random vectors such that $\boldsymbol{\Sigma_w^{-1}}$ and $\boldsymbol{\Sigma_{wY}}$ exist. Then a second way to compute the OLS estimator is

$$\hat{\alpha} = \overline{Y} - \hat{\boldsymbol{\eta}}^T\overline{\boldsymbol{w}} \xrightarrow{P} \alpha$$

and

$$\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} \xrightarrow{P} \boldsymbol{\eta} \text{ as } \mathrm{n} \to \infty.$$

A common technique to try to get a robust MLR estimator is to plug a robust MLD estimator $(T, \boldsymbol{C})$ for the above quantities. These techniques were not very good because the robust MLD estimators were poor before the FCH, RFCH, and RMVN estimators. The `rmreg2` estimator is the OLS estimator computed from the cases in the RMVN set and the plug in estimator where $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the RMVN set when RMVN is applied to vectors $\boldsymbol{u}_i$ for $i = 1, ..., n$ (could use $(T, \boldsymbol{C})$ = RMVN estimator since the scaling does not matter for this application). Then $(T, \boldsymbol{C})$ is a $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu_u}, c\ \boldsymbol{\Sigma_u})$ if the $\boldsymbol{u}_i$ are iid from a large class of $EC_p(\boldsymbol{\mu_u}, \boldsymbol{\Sigma_u}, g)$ distributions. Thus `rmreg2` estimator is a $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$ if the $\boldsymbol{u}_i$ are iid from a large class of elliptically contoured distributions. This assumption is quite strong, but the robust estimator is useful for detecting outliers. When there are categorical predictors or the joint distribution of $\boldsymbol{u}$ is not elliptically contoured,

it is possible that the robust estimator is bad and very different from the good classical least squares estimator.

The *rpack* function `rmreg2` computes the `rmreg2` estimator and produces the response and residual plots. The function `rmreg3` computes the estimator without the plots. See the following *R* code.

```
rmreg2(buxx,buxy) #right click Stop 2 times
rmreg3(buxx,buxy)
```

The conditions under which the `rmreg2` estimator has been shown to be $\sqrt{n}$ consistent are quite strong, but it seems likely that the estimator is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under mild conditions where the parameter vector $\boldsymbol{\beta}$ is not, in general, the parameter vector estimated by OLS. For MLR, the *rpack* function `rmregboot` bootstraps the `rmreg2` estimator, and the function `rmregbootsim` can be used to simulate `rmreg2`. Both functions use the residual bootstrap where the residuals come from OLS. See the *R* code below.

```
out<-rmregboot(belx,bely)
plot(out$betas)
ddplot4(out$betas) #right click Stop

out<-rmregboot(cbrainx,cbrainy)
ddplot4(out$betas) #right click Stop
```

## 6.2 A Practical High Breakdown Consistent Estimator

Olive and Hawkins (2011) showed that the practical `hbreg` estimator is a high breakdown $\sqrt{n}$ consistent robust estimator that is asymptotically equivalent to the least squares estimator for many error distributions. This section follows Olive (2017b, pp. 420-423).

The outlier resistance of the `hbreg` estimator is not very good, but roughly comparable to the best of the practical "robust regression" estimators available in *R* packages as of 2020. The estimator is of some interest since it proved that practical high breakdown consistent estimators are possible. Other practical regression estimators that claim to be high breakdown and consistent appear to be zero breakdown because they use the zero breakdown elemental concentration algorithm. See Theorem 5.13.

The following theorem is powerful because it does not depend on the criterion used to choose the attractor, and proves that the `mbareg` and `tvreg` estimators are $\sqrt{n}$ consistent. Suppose there are $K$ consistent estimators $\hat{\boldsymbol{\beta}}_j$ of $\boldsymbol{\beta}$, each with the same rate $n^\delta$. If $\hat{\boldsymbol{\beta}}_A$ is an estimator obtained by choosing one of the $K$ estimators, then $\hat{\boldsymbol{\beta}}_A$ is a consistent estimator of $\boldsymbol{\beta}$ with rate $n^\delta$ by Pratt (1959). See Theorem 11.17.

**Theorem 6.1.** Suppose the algorithm estimator chooses an attractor as the final estimator where there are $K$ attractors and $K$ is fixed.

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

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

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

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

The consistency of the algorithm estimator changes dramatically if $K$ is fixed but the start size $h = h_n = g(n)$ where $g(n) \to \infty$. In particular, if $K$ starts with rate $n^{1/2}$ are used, the final estimator also has rate $n^{1/2}$. The drawback to these algorithms is that they may not have enough outlier resistance. Notice that the basic resampling result below is free of the criterion.

**Theorem 6.2.** Suppose $K_n \equiv K$ starts are used and that all starts have subset size $h_n = g(n) \uparrow \infty$ as $n \to \infty$. Assume that the estimator applied to the subset has rate $n^\delta$.

i) For the $h_n$-set basic resampling algorithm, the algorithm estimator has rate $[g(n)]^\delta$.

ii) Under regularity conditions (e.g. given by He and Portnoy 1992), the k–step CLTS estimator has rate $[g(n)]^\delta$.

**Proof.** i) The $h_n = g(n)$ cases are randomly sampled without replacement. Hence the classical estimator applied to these $g(n)$ cases has rate $[g(n)]^\delta$. Thus all $K$ starts have rate $[g(n)]^\delta$, and the result follows by Pratt (1959). ii) By He and Portnoy (1992), all $K$ attractors have $[g(n)]^\delta$ rate, and the result follows by Pratt (1959). □

**Remark 6.2.** Theorem 5.11 shows that $\hat{\boldsymbol{\beta}}$ is HB if the median absolute or squared residual (or $|r(\hat{\boldsymbol{\beta}})|_{(c_n)}$ or $r^2_{(c_n)}$ where $c_n \approx n/2$) stays bounded under high contamination. Let $Q_L(\hat{\boldsymbol{\beta}}_H)$ denote the LMS, LTS, or LTA criterion for an estimator $\hat{\boldsymbol{\beta}}_H$; therefore, the estimator $\hat{\boldsymbol{\beta}}_H$ is high breakdown if and only if $Q_L(\hat{\boldsymbol{\beta}}_H)$ is bounded for $d_n$ near $n/2$ where $d_n < n/2$ is the number of outliers. The concentration operator refines an initial estimator by successively reducing the LTS criterion. If $\hat{\boldsymbol{\beta}}_F$ refers to the final estimator (attractor) obtained by applying concentration to some starting estimator $\hat{\boldsymbol{\beta}}_H$ that is high breakdown, then since $Q_{LTS}(\hat{\boldsymbol{\beta}}_F) \leq Q_{LTS}(\hat{\boldsymbol{\beta}}_H)$, applying concentration to

a high breakdown start results in a high breakdown attractor. See Theorem 5.15.

High breakdown estimators are, however, not necessarily useful for detecting outliers. Suppose $\gamma_n < 0.5$. On the one hand, if the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down parallel to the $Y$ axis, then for high breakdown estimators, $\hat{\boldsymbol{\beta}}$ and MED($|r_i|$) will be bounded. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large, suggesting that the high breakdown estimator is useful for outlier detection. On the other hand, if the $Y_i$'s are fixed at any values and the $\boldsymbol{x}$ values perturbed, sufficiently large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not $\infty$. For many estimators, including LTS, LMS, and LTA, a cluster of $Y$ outliers can be moved arbitrarily far from the bulk of the data but still, by perturbing their $\boldsymbol{x}$ values, have arbitrarily small residuals. See Example 6.2.

Our practical high breakdown procedure is made up of three components.
1) A practical estimator $\hat{\boldsymbol{\beta}}_C$ that is consistent for clean data. Suitable choices would include the full-sample OLS and $L_1$ estimators.
2) A practical estimator $\hat{\boldsymbol{\beta}}_A$ that is effective for outlier identification. Suitable choices include the `mbareg`, `rmreg2`, `lmsreg`, or FLTS estimators.
3) A practical high-breakdown estimator such as $\hat{\boldsymbol{\beta}}_B$ from Definition 5.42 with $k = 10$.

By selecting one of these three estimators according to the features each of them uncovers in the data, we may inherit some of the good properties of each of them.

**Definition 6.2.** The `hbreg` estimator $\hat{\boldsymbol{\beta}}_H$ is defined as follows. Pick a constant $a > 1$ and set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_C$. If $aQ_L(\hat{\boldsymbol{\beta}}_A) < Q_L(\hat{\boldsymbol{\beta}}_C)$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_A$. If $aQ_L(\hat{\boldsymbol{\beta}}_B) < \min[Q_L(\hat{\boldsymbol{\beta}}_C), aQ_L(\hat{\boldsymbol{\beta}}_A)]$, set $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_B$.

That is, find the smallest of the three scaled criterion values $Q_L(\hat{\boldsymbol{\beta}}_C)$, $aQ_L(\hat{\boldsymbol{\beta}}_A)$, $aQ_L(\hat{\boldsymbol{\beta}}_B)$. According to which of the three estimators attains this minimum, set $\hat{\boldsymbol{\beta}}_H$ to $\hat{\boldsymbol{\beta}}_C, \hat{\boldsymbol{\beta}}_A$, or $\hat{\boldsymbol{\beta}}_B$ respectively.

Large sample theory for `hbreg` is simple and given in the following theorem. Let $\hat{\boldsymbol{\beta}}_L$ be the LMS, LTS, or LTA estimator that minimizes the criterion $Q_L$. Note that the impractical estimator $\hat{\boldsymbol{\beta}}_L$ is never computed. The following theorem shows that $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_C$ on a large class of zero mean finite variance symmetric error distributions. Thus if $\hat{\boldsymbol{\beta}}_C$ is $\sqrt{n}$ consistent or asymptotically efficient, so is $\hat{\boldsymbol{\beta}}_H$. Notice that $\hat{\boldsymbol{\beta}}_A$ does not need to be consistent. This point is crucial since `lmsreg` is not consistent and it is not known whether FLTS is consistent. The clean data are in *general position* if any $p$ clean cases give a unique estimate of $\hat{\boldsymbol{\beta}}$.

**Theorem 6.3.** Assume the clean data are in general position, and suppose that both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$ where the regression

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

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

**Remark 6.3.** i) Let $\hat{\boldsymbol{\beta}}_C = \hat{\boldsymbol{\beta}}_{OLS}$. Then hbreg is asymptotically equivalent to OLS when the errors $e_i$ are iid from a large class of zero mean finite variance symmetric distributions, including the $N(0, \sigma^2)$ distribution, since the probability that hbreg uses OLS instead of $\hat{\boldsymbol{\beta}}_A$ or $\hat{\boldsymbol{\beta}}_B$ goes to one as $n \to \infty$.

ii) The above theorem proves that practical high breakdown estimators with 100% asymptotic Gaussian efficiency exist; however, such estimators are not necessarily good.

iii) The theorem holds when both $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_C$ are consistent estimators of $\boldsymbol{\beta}$, for example, when the iid errors come from a large class or zero mean finite variance symmetric distributions. For asymmetric distributions, $\hat{\boldsymbol{\beta}}_C$ estimates $\boldsymbol{\beta}_C$ and $\hat{\boldsymbol{\beta}}_L$ estimates $\boldsymbol{\beta}_L$ where the constants usually differ. The theorem holds for some distributions that are not symmetric because of the penalty $a$. As $a \to \infty$, the class of asymmetric distributions where the theorem holds greatly increases, but the outlier resistance decreases rapidly as $a$ increases for $a > 1.4$.

iv) The default hbreg estimator used OLS, mbareg, and $\hat{\boldsymbol{\beta}}_B$ with $a = 1.4$ and the LTA criterion. For the simulated data with symmetric error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give biased estimates of the slopes. However, for the simulated data with right skewed error distributions, $\hat{\boldsymbol{\beta}}_B$ appeared to give good estimates of the slopes but not the constant estimated by OLS, and the probability that the hbreg estimator selected $\hat{\boldsymbol{\beta}}_B$ appeared to go to one.

v) Both MBA and OLS are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$, even for a large class of skewed distributions. Using $\hat{\boldsymbol{\beta}}_A = \hat{\boldsymbol{\beta}}_{MBA}$ and removing $\hat{\boldsymbol{\beta}}_B$ from the hbreg estimator results in a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ when $\hat{\boldsymbol{\beta}}_C =$ OLS is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$, but massive sample sizes were still needed to get good estimates of the constant for skewed error distributions. For skewed distributions, if OLS needed $n = 1000$ to estimate the constant well, mbareg might need $n >$ one million to estimate the constant well.

The situation is worse for multivariate linear regression when hbreg is used instead of OLS, since there are $m$ constants to be estimated. If the distribution of the iid error vectors $\boldsymbol{e}_i$ is not elliptically contoured, getting

all $m$ `mbareg` estimators to estimate all $m$ constants well needs even larger sample sizes.

vi) The outlier resistance of `hbreg` is not especially good.

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

We suggest using the LTA criterion since in simulations, `hbreg` behaved like $\hat{\boldsymbol{\beta}}_C$ for smaller sample sizes than those needed by the LTS and LMS criteria. We want $a$ near 1 so that `hbreg` has outlier resistance similar to $\hat{\boldsymbol{\beta}}_A$, but we want $a$ large enough so that `hbreg` performs like $\hat{\boldsymbol{\beta}}_C$ for moderate $n$ on clean data. Simulations suggest that $a = 1.4$ is a reasonable choice. The default `hbreg` program from *rpack* uses the $\sqrt{n}$ consistent outlier resistant estimator `mbareg` as $\hat{\boldsymbol{\beta}}_A$.

There are at least three reasons for using $\hat{\boldsymbol{\beta}}_B$ as the high breakdown estimator. First, $\hat{\boldsymbol{\beta}}_B$ is high breakdown and simple to compute. Second, the fitted values roughly track the bulk of the data. Lastly, although $\hat{\boldsymbol{\beta}}_B$ has rather poor outlier resistance, $\hat{\boldsymbol{\beta}}_B$ does perform well on several outlier configurations where some common alternatives fail.

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

Simulations were run in *R* with the $x_{ij}$ (for $j > 1$) and $e_i$ iid $N(0, \sigma^2)$ and $\boldsymbol{\beta} = \mathbf{1}$, the $p \times 1$ vector of ones. Then $\hat{\boldsymbol{\beta}}$ was recorded for 100 runs. The mean and standard deviation of the $\hat{\beta}_j$ were recorded for $j = 1, ..., p$. For $n \geq 10p$ and OLS, the vector of means should be close to $\mathbf{1}$ and the vector of standard deviations should be close to $\mathbf{1}/\sqrt{n}$. The $\sqrt{n}$ consistent high breakdown `hbreg` estimator performed like OLS if $n \approx 35p$ and $2 \leq p \leq 6$, if $n \approx 20p$ and $7 \leq p \leq 14$, or if $n \approx 15p$ and $15 \leq p \leq 40$. See Table 7.7 for $p = 5$ and 100 runs. ALTS denotes `ltsreg`, HB denotes `hbreg`, and BB denotes $\hat{\boldsymbol{\beta}}_B$. In the simulations, `hbreg` estimated the slopes well for the highly skewed lognormal data, but not the OLS constant. Use the *rpack* function `hbregsim`.

**Table 6.1** MEAN $\hat{\beta}_i$ and SD($\hat{\beta}_i$)

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

As implemented in *rpack*, the `hbreg` estimator is a practical $\sqrt{n}$ consistent high breakdown estimator that appears to perform like OLS for moderate $n$ if the errors are unimodal and symmetric, and to have outlier resistance comparable to competing practical "outlier resistant" estimators.

The `hbreg`, `lmsreg`, `ltsreg`, OLS, and $\hat{\boldsymbol{\beta}}_B$ estimators were compared on the same 25 benchmark data sets. Also see Park et al. (2012). The HB estimator $\hat{\boldsymbol{\beta}}_B$ was surprisingly good in that the response plots showed that it was the best estimator for 2 data sets and that it usually tracked the data, but it performed poorly in 7 of the 25 data sets. The `hbreg` estimator performed well, but for a few data sets `hbreg` did not pick the attractor with the best response plot, as illustrated in the following example.

**Example 6.2.** The LMS, LTA, and LTS estimators are determined by a "narrowest band" covering half of the cases. Hawkins and Olive (2002) suggested that the fit will pass through outliers if the band through the outliers is narrower than the band through the clean cases. This behavior tends to occur if the regression relationship is weak, and if there is a tight cluster of outliers where $|Y|$ is not too large. Also see Wang and Suter (2003). As an illustration, Buxton (1920, pp. 232-5) gave 20 measurements of 88 men. Consider predicting *stature* using an intercept, *head length, nasal height, bigonal breadth*, and *cephalic index*. One case was deleted since it had missing values. Five individuals, numbers 61-65, were reported to be about 0.75 inches tall with head lengths well over five feet! Figure 6.2 shows the response plots for `hbreg`, OLS, `ltsreg`, and $\hat{\boldsymbol{\beta}}_B$. Notice that only the fit from $\hat{\boldsymbol{\beta}}_B$ (BBFIT) did not pass through the outliers, but `hbreg` selected the OLS attractor. There are always outlier configurations where an estimator will fail, and `hbreg` should fail on configurations where LTA, LTS, and LMS would fail.

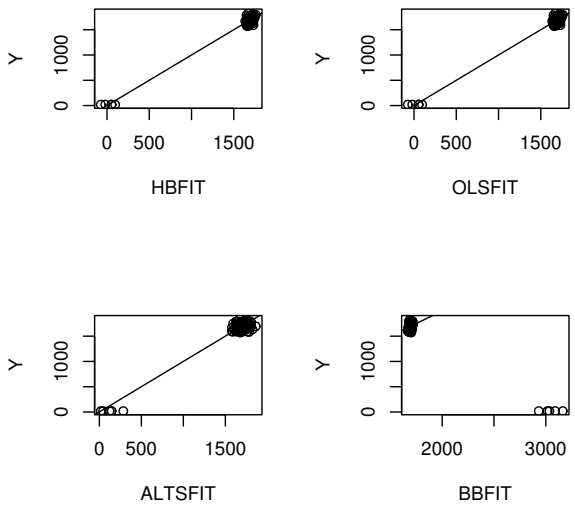


**Fig. 6.2** Response Plots Comparing Robust Regression Estimators

The *rpack* functions `ffplot2` and `rrplot2` make FF and RR plots using OLS, ALMS from `lmsreg`, ALTS from `ltsreg`, `mbareg`, an outlier detector `mbalata`, `BB`, and `rmreg2`. The `mbalata` estimator is described in Olive (2017b, § 12.6.2). OLS, `BB`, and `mbareg` are the three trial fits used by the default version of the $\sqrt{n}$ consistent high breakdown `hbreg` estimator. The top row of `ffplot2` shows the response plots. The *R* code below is useful and shows how to get some of the text's data sets into *R*.

```
library(MASS)
rrplot2(buxx,buxy)
ffplot2(buxx,buxy)
#The following three data sets can be obtained with
#the source("G:/robdata.txt") command
#if the data file is on flash drive G.
rmreg2(buxx,buxy)       #right click Stop twice
rmreg2(cbrainx,cbrainy)
rmreg2(gladox,gladoy)

hbk <- matrix(scan(),nrow=75,ncol=5,byrow=T)
hbk <- hbk[,-1]
rmreg2(hbk[,1:3],hbk[,4]) #Outliers are clear
#but fit avoids good leverage points.
```

```
nasty <- matrix(scan(),nrow=32,ncol=6,byrow=T)
nasty <- nasty[,-1]
rmreg2(nasty[,1:4],nasty[,5])

wood <- matrix(scan(),nrow=20,ncol=7,byrow=T)
wood <- wood[,-1]
rmreg2(wood[,1:5],wood[,6]) #failed to find
#the outliers

major <- matrix(scan(),nrow=112,ncol=7,byrow=T)
major <- major[,-1]
rmreg2(major[,1:5],major[,6])
```

**Example 6.1, continued.** The FF and RR plots for the Buxton (1920) data are shown in Figures 6.3 and 6.4. Note that only the last four estimators gives large absolute residuals to the outliers. The top row of Figure 6.3 gives the response plots for the estimators. If there are two clusters, one in the upper right and one in the lower left of the response plot, then the identity line goes through both clusters. Hence the fit passes through 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 Figures 6.3 and 6.4, but in about three plots the MBA estimator will also pass through the outliers.

**Table 6.2** Summaries for Seven Data Sets, the Correlations of the Residuals from TV(M) and the Alternative Method are Given in the 1st 5 Rows

| Method | Buxton | Gladstone | glado | hbk | major | nasty | wood |
|---|---|---|---|---|---|---|---|
| MBA | 0.997 | 1.0 | 0.455 | 0.960 | 1.0 | -0.004 | 0.9997 |
| LMSREG | -0.114 | 0.671 | 0.938 | 0.977 | 0.981 | 0.9999 | 0.9995 |
| LTSREG | -0.048 | 0.973 | 0.468 | 0.272 | 0.941 | 0.028 | 0.214 |
| L1 | -0.016 | 0.983 | 0.459 | 0.316 | 0.979 | 0.007 | 0.178 |
| OLS | 0.011 | 1.0 | 0.459 | 0.780 | 1.0 | 0.009 | 0.227 |
| outliers | 61-65 | none | 115 | 1-10 | 3,44 | 2,6,...,30 | 4,6,8,19 |
| n | 87 | 267 | 267 | 75 | 112 | 32 | 20 |
| p | 5 | 7 | 7 | 4 | 6 | 5 | 6 |
| M | 70 | 0 | 30 | 90 | 0 | 90 | 20 |

Table 6.2 compares the TV, MBA (for MLR), `lmsreg`, `ltsreg`, $L_1$, and OLS estimators on 7 data sets available from the text's website. The column headers give the file name while the remaining rows of the table give the sample size $n$, the number of predictors $p$, the amount of trimming $M$ used by the TV estimator, the correlation of the residuals from the TV estimator with the corresponding alternative estimator, and the cases that were outliers. If the correlation was greater than 0.9, then the method was effective in detecting the outliers, and the method failed, otherwise. Sometimes the

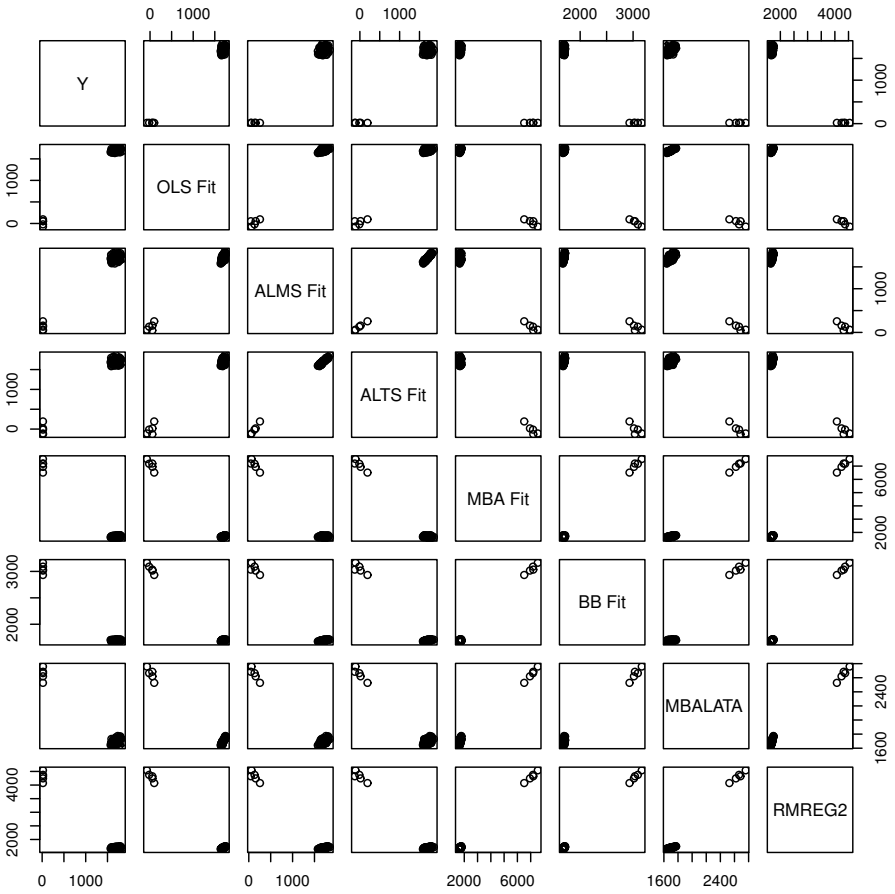


**Fig. 6.3** FF Plots for Buxton Data

trimming percentage $M$ for the TV estimator was picked after fitting the bulk of the data in order to find the good leverage points and outliers. Each model included a constant.

Notice that the TV, MBA, and OLS estimators were the same for the Gladstone (1905) data and for the Tremearne (1911) *major* data which had two small $Y$–outliers. For the Gladstone data, there is a cluster of infants that are good leverage points, and we attempt to predict *brain weight* with the head measurements *height, length, breadth, size,* and *cephalic index*. Originally, the variable *length* was incorrectly entered as 109 instead of 199 for case 115, and the *glado* data contains this outlier. In 1997, `lmsreg` was not able to detect the outlier while `ltsreg` did. Due to changes in the *Splus* 2000

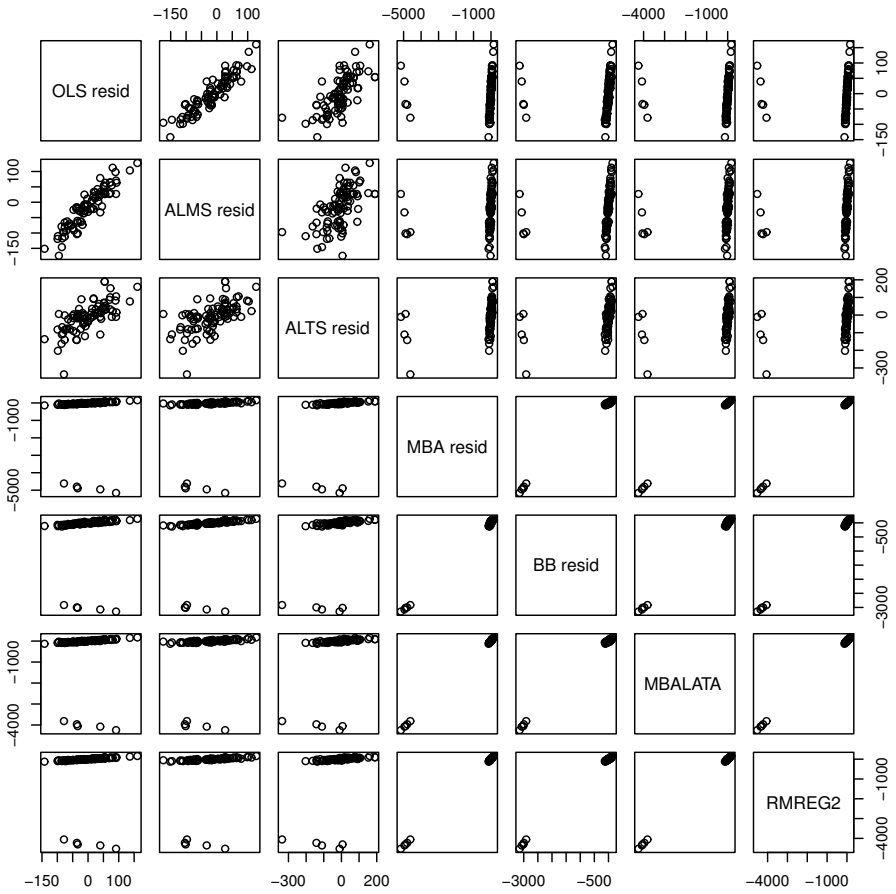


**Fig. 6.4** RR Plots for Buxton Data

code, `lmsreg` detected the outlier but `ltsreg` did not. These two functions change often, not always for the better.

## 6.3 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 and that the clean data are in general position. Following Section 5.8, 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; i.e., the breakdown value of the regression estimator is close to 0.5.

Perhaps the first HB MLR estimator proposed was the least median of squares (LMS) estimator. Let $|r(\boldsymbol{b})|_{(i)}$ denote the $i$th ordered absolute residual from the estimate $\boldsymbol{b}$ sorted from smallest to largest, and let $r^2_{(i)}(\boldsymbol{b})$ denote the $i$th ordered squared residual. Next, three of the most important robust criteria are defined, but the robust estimators take too long to compute. In the literature, LMS($c_n$) is used more that LQS($c_n$), but the term "LMS" makes the most sense when $c_n/n \to 0.5$ as $n \to \infty$.

**Definition 6.3.** The *least quantile of squares* (LQS($c_n$)) estimator minimizes the criterion

$$Q_{LQS}(\boldsymbol{b}) \equiv Q_{LMS}(\boldsymbol{b}) = r^2_{(c_n)}(\boldsymbol{b}). \tag{6.3}$$

The LQS($c_n$) estimator is also known as the *least median of squares* LMS($c_n$) estimator (Hampel 1975, p. 380).

**Definition 6.4.** The *least trimmed sum of squares* (LTS($c_n$)) estimator (Rousseeuw 1984) minimizes the criterion

$$Q_{LTS}(\boldsymbol{b}) = \sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}). \tag{6.4}$$

**Definition 6.5.** The *least trimmed sum of absolute deviations* (LTA($c_n$)) estimator (Hössjer 1991) minimizes the criterion

$$Q_{LTA}(\boldsymbol{b}) = \sum_{i=1}^{c_n} |r(\boldsymbol{b})|_{(i)}. \tag{6.5}$$

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. Let $\lfloor x \rfloor$ be the greatest integer less than or equal to $x$. For example, $\lfloor 7.7 \rfloor = 7$.

**Definition 6.6.** 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,

$$c_n = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor \tag{6.6}$$

is often used as the default.

**Remark 6.4. 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 moderate data sets. The second category consists of estimators that have rigorous asymp-

totic 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 zero breakdown inconsistent estimators. 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. Section 6.2 shows how to fix the practical algorithms so that the resulting estimator is $\sqrt{n}$ consistent and high breakdown.

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 $\tau$ tends to 1, but the breakdown value $1 - \tau$ decreases to 0, although asymptotic normality of LTA has not yet been proven. We will use the unifying notation LTx($\tau$) 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 consistency and resistance properties of the ALTx estimator depend strongly on the number of starts $K_n$ used.

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 $\alpha$ decreases, the Gaussian efficiency of the $\alpha$ trimmed mean increases to 1, but the breakdown value decreases to 0.

### *6.3.1* Theoretical Properties

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(0, V(\hat{\boldsymbol{\beta}}, F)\ \boldsymbol{W}) \tag{6.7}$$

when $\dfrac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{W}^{-1}$, 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 Bassett and Koenker (1978). Broffitt (1974) compares OLS, $L_1$, and $L_\infty$ in the location model and shows that the rate of convergence of the Chebyshev estimator is often very poor.

**Remark 6.5.** Obtaining asymptotic theory for LTA and LTS is a very challenging problem. Mašiček (2004), Čížek (2006) and Víšek (2006) claim to have shown asymptotic normality of LTS under general conditions. Čížek (2008) shows that LTA is $\sqrt{n}$ consistent. 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($\tau$) 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. All of the "brand name" high breakdown regression estimators take far too long to compute, and 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 breakdown results for the LTx estimators are well known. See Hössjer (1994, p. 151). See Section 5.8 for the definition of breakdown.

**Theorem 6.4: Breakdown of LTx Estimators.** Assume the clean data are in general position. Then LMS($\tau$), LTS($\tau$), and LTA($\tau$) have breakdown value

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

**Theorem 6.5.** Under regularity conditions similar to those in Conjecture 6.1 below, a) the LMS($\tau$) converges at a cubed root rate to a non-Gaussian limit. b) The estimator $\hat{\boldsymbol{\beta}}_{LTS}$ satisfies Equation (6.7) and

$$V(LTS(\tau), F) = \frac{\int_{F^{-1}(1/2-\tau/2)}^{F^{-1}(1/2+\tau/2)} w^2 dF(w)}{[\tau - 2F^{-1}(1/2+\tau/2) f(F^{-1}(1/2+\tau/2))]^2}. \tag{6.8}$$

The proof of Theorem 6.5a 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 6.1.** 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 (6.7) and

$$V(LTA(\tau), F) = \frac{\tau}{4[f(0) - f(F^{-1}(1/2 + \tau/2))]^2}. \tag{6.9}$$

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

Čížek (2008a) shows that LTA is $\sqrt{n}$ consistent, but does not prove that LTA is asymptotically normal. *Assume Conjecture 6.1 is true for the following LTA remarks in this section.* Then as $\tau \to 1$, the efficiency of LTS approaches that of OLS and the efficiency of LTA approaches that of $L_1$. Hence for $\tau$ 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 2.27).

**Theorem 6.6.** Under the symmetry conditions given in Conjecture 6.1,

$$V(LTS(\tau), F) = \frac{\tau \sigma^2_{TF}(-k, k)}{[\tau - 2kf(k)]^2} \tag{6.10}$$

and

$$V(LTA(\tau), F) = \frac{\tau}{4[f(0) - f(k)]^2} \tag{6.11}$$

where

$$k = F^{-1}(0.5 + \tau/2). \tag{6.12}$$

**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 $W_T$ is $V(W_T) = \sigma^2_{TF}(-k, k) = \dfrac{\int_{-k}^{k} w^2 dF(w)}{F(k) - F(-k)}$ by Definition 2.27. 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 11. The following examples illustrate the result. See Hawkins and Olive (1999b).

**Example 6.3: N(0,1) Errors.** If $Y_T$ is a $N(0, \sigma^2)$ truncated at $a = -k\sigma$ and $b = k\sigma$, $V(Y_T) = \sigma^2[1 - \frac{2k\phi(k)}{2\Phi(k) - 1}]$. At the standard normal

$$V(LTS(\tau), \Phi) = \frac{1}{\tau - 2k\phi(k)} \tag{6.13}$$

$$\text{while } \mathrm{V(LTA}(\tau), \Phi) = \frac{\tau}{4[\phi(0) - \phi(\mathrm{k})]^2} = \frac{2\pi\tau}{4[1 - \exp(-\mathrm{k}^2/2)]^2} \tag{6.14}$$

where $\phi$ is the standard normal pdf and $k = \Phi^{-1}(0.5+\tau/2)$. Thus for $\tau \geq 1/2$, LTS($\tau$) has breakdown value of $1 - \tau$ and Gaussian efficiency

$$\frac{1}{V(LTS(\tau), \Phi)} = \tau - 2k\phi(k). \tag{6.15}$$

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 6.4: 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,

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

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

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 6.5: 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 $V(X_T) = \frac{k - \tan^{-1}(k)}{\tan^{-1}(k)}$. Hence

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

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

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.

### *6.3.2* Computation and Simulations

**Theorem 6.7.** 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,

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

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,

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

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. □

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

**Table 6.3** 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) | $\infty$ | .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) | $\infty$ | 2.000 | 1.000 | .71 | 1.500 |

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), Bertsimas and Mazumder (2014), Hofmann et al. (2010), and Klouda (2015).

These three estimators have $O(n^p)$ complexity or higher, and estimators with $O(n^4)$ or higher complexity take too long to compute and will rarely be used. The literature on estimators with $O(n^p)$ complexity typically claims that the estimator can be computed for up to a few hundred cases if $p \leq 4$, while simulations use $p \leq 2$. Since estimators need to be widely used before they are trustworthy, the brand name HB robust regression estimators are untrustworthy for $p > 2$.

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. See Section 2.12.

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 6.1. 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.

## 6.4 Complements

Olive (2008, ch. 7-9) covers robust and resistant regression. Also see Hawkins and Olive (1999b), Olive and Hawkins (2003) and Olive (2005a, 2017b). The outlier resistance of elemental algorithms decreases rapidly as $p$ increases. However, for $p < 10$, such elemental algorithms are often useful for outlier detection. They can perform better than MBA, trimmed views, and `rmreg2` if $p$ is small and the outliers are close to the bulk of the data or if $p$ is small and there is a mixture distribution: the bulk of the data follows one MLR model, but "outliers" and some of the clean data are fit well by another MLR model.

A promising resistant regression estimator is given by Park et al. (2012). Bassett (1991) suggested the LTA estimator for location and Hössjer (1991) suggested the LTA regression estimator. Oldford (1983) proves that $\hat{\boldsymbol{\beta}}_B$ is high breakdown.

The LMS, LTA, and LTS estimators 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 criteria are used in the MBA estimator and in the easily computed $\sqrt{n}$ consistent high breakdown `hbreg` estimator.

In addition to the LMS, LTA, and LTS estimators, 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 et al. 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.

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. See Bernholt (2005, 2006).

A large number of impractical "brand name" high breakdown regression estimators have been proposed, including LTS, LMS, LTA, S, LQD, $\tau$, constrained M, repeated median, cross checking, one step GM, one step GR, t-type, and regression depth estimators. See Rousseeuw and Leroy (1987) and Maronna et al. (2019). The practical algorithms used in the software use a brand name criterion to evaluate a fixed number of trial fits and should be

denoted as an F-brand name estimator such as FLTS. Two stage estimators, such as the MM estimator, that need an initial consistent high breakdown estimator often have the same breakdown value and consistency rate as the initial estimator.

These impractical "brand name" estimators have at least $O(n^p)$ complexity, while the practical estimators used in the software have not been shown to be both high breakdown and consistent. See Hawkins and Olive (2002), Hubert et al. (2002), and Maronna and Yohai (2002). Huber and Ronchetti (2009, pp. xiii, 8-9, 152-154, 196-197) suggested that high breakdown regression estimators do not provide an adequate remedy for the ill effects of outliers, that their statistical and computational properties are not adequately understood, that high breakdown estimators "break down for all except the smallest regression problems by failing to provide a timely answer!" and that "there are no known high breakdown point estimators of regression that are demonstrably stable."

A massive problem with "robust high breakdown regression" research is the claim that a brand name impractical estimator is being used since the software nearly always actually replaces the brand name estimator by a practical F-brand name estimator that is not backed by theory, such as FLTS. In particular, the claim that "LTS can be computed with Fast-LTS" is false. See Theorem 5.13. An estimator implemented with a zero breakdown inconsistent initial estimator tends to be 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`. Recent examples are Bondell and Stefanski (2013) and Jiang et al. (2019).

Maronna and Yohai (2015) used OLS and 500 elemental sets as the 501 trial fits to produce an FS estimator used as the initial estimator for an FMM estimator. Since the 501 trial fits are zero breakdown, so is the FS estimator. Since the FMM estimator has the same breakdown as the initial estimator, the FMM estimator is zero breakdown. For regression, they show that the FS estimator is consistent on a large class of zero mean finite variance symmetric distributions. Consistency follows since the elemental fits and OLS are unbiased estimators of $\boldsymbol{\beta}_{OLS}$ but an elemental fit is an OLS fit to $p$ cases. Hence the elemental fits are very variable, and the probability that the OLS fit has a smaller S-estimator criterion than a randomly chosen elemental fit (or $K$ randomly chosen elemental fits) goes to one as $n \to \infty$. (OLS and the S-estimator are both $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$, so the ratio of their criterion values goes to one, and the S-estimator minimizes the criterion value.) Hence the FMM estimator is asymptotically equivalent to the MM estimator that has the smallest criterion value for a large class of iid zero mean finite variance symmetric error distributions. This FMM estimator is asymptotically equivalent to the FMM estimator that uses OLS as the initial estimator. When the error distribution is skewed the S-estimator and OLS population constant are not the same, and the probability that an elemental

fit is selected is close to one for a skewed error distribution as $n \to \infty$. (The OLS estimator $\hat{\boldsymbol{\beta}}$ gets very close to $\boldsymbol{\beta}_{OLS}$ while the elemental fits are highly variable unbiased estimators of $\boldsymbol{\beta}_{OLS}$, so one of the elemental fits is likely to have a constant that is closer to the S-estimator constant while still having good slope estimators.) Hence the FS estimator is inconsistent, and the FMM estimator is likely inconsistent for skewed distributions. No practical method is known for computing a $\sqrt{n}$ consistent FS or FMM estimator that has the same breakdown and maximum bias function as the S or MM estimator that has the smallest S or MM criterion value.

## 6.5 Problems

**R Problems** Some *R* code for homework problems is at (http://parker.ad.siu.edu/Olive/robRhw.txt).

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

The "asymptotic variance" for LTA in Problems 8.1, 8.2 and 8.3 is actually the conjectured asymptotic variance for LTA if the multiple linear regression model is used instead of the location model.

**6.1.** a) Download the *R* 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?

**6.2.** a) Download the *R* 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?

**6.3.** a) Download the *R* 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?

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

b) Enter the command *mbamv(belx,bely)* in *R*. 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*. Click on the rightmost mouse button (and in *R*, click on *Stop*). You need to do this 14 times before the program ends. There are four predictors $x_1, ..., x_4$ 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*.

**6.5*.** 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("G:/rpack.txt")* and *source("G:/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*.

**6.6.** a) In addition to the *source("G:/rpack.txt")* command, also use the *source("G:/robdata.txt")* command (and in *R*, type the *library(MASS)* command).

b) Type the command *tvreg(buxx,buxy,ii=1)*. Click the rightmost mouse button (and in *R*, highlight *Stop*). The response plot should appear. Repeat 10 times and remember which plot percentage $M$ (say M = 0) had the best response plot. Then type the command *tvreg2(buxx,buxy, M = 0)* (except use your value of M, not 0). Again, click the rightmost mouse button (and in *R*, highlight *Stop*). The response plot should appear. Hold down the *Ctrl* and *c* keys to make a copy of the plot. Then paste the plot in *Word*.

c) The estimated coefficients $\hat{\boldsymbol{\beta}}_{TV}$ from the best plot should have appeared on the screen. Copy and paste these coefficients into *Word*.