Chapter 9

Resistance and Equivariance

9.1 Resistance of Algorithm Estimators

In spite of the inconsistency of resampling algorithms that use a fixed number K of elemental starts, these algorithms appear throughout the robustness literature and in *R/Splus* software. Proposition 8.7 on p. 267 suggests that the algorithms can be useful for small data sets.

The previous chapter used the *asymptotic paradigm* to show that the algorithm estimators are inconsistent. In this paradigm, it is assumed that the data set size n is increasing to ∞ and we want to know whether an estimator $\hat{\boldsymbol{\beta}}_n$ converges to $\boldsymbol{\beta}$ or not.

Definition 9.1. Suppose that a subset of h cases is selected from the n cases making up the data set. Then the subset is *clean* if none of the h cases are outliers.

In this chapter we will consider the *perfect classification paradigm* where the goal is to analyze a single *fixed data set* of n cases of which $0 \leq d < n/2$ are outliers. The remaining $n - d$ cases are "clean." The main assumption of the perfect classification paradigm is that the algorithm can perfectly classify the clean and outlying cases; ie, the outlier configuration is such that if K subsets of size $h \geq p$ are selected, then the subset J_o corresponding to the fit that minimizes the criterion Q will be clean, and the (eg OLS or L_1) fit $\boldsymbol{b}_{J_o}$ computed from the cases in J_o will perfectly classify the $n - d$ clean cases and d outliers. Then a separate analysis is run on each of the two groups. Although this is a very big assumption that is almost impossible to verify,

the paradigm gives a useful initial model for the data. The assumption is very widely used in the literature for diagnostics and robust statistics.

Remark 9.1. Suppose that the data set contains n cases with d outliers and $n - d$ clean cases. Suppose that $h \geq p$ cases are selected at random without replacement. Let W count the number of the h cases that were outliers. Then W is a hypergeometric$(d, n - d, h)$ random variable and

$$P(W = j) = \frac{\binom{d}{j}\binom{n-d}{h-j}}{\binom{n}{h}} \approx \binom{h}{j}\gamma^j(1-\gamma)^{h-j}$$

where the *contamination proportion* $\gamma = d/n$ and the binomial$(h, \rho \equiv \gamma = d/n)$ approximation to the hypergeometric$(d, n - d, h)$ distribution is used. In particular, the probability that the subset of h cases is clean $= P(W = 0) \approx (1 - \gamma)^h$ which is maximized by $h = p$. Hence using elemental sets maximizes the probability of getting a clean subset. Moreover, computing the elemental fit is faster than computing the fit from $h > p$ cases.

Remark 9.2. Now suppose that K elemental sets are chosen with replacement. If W_i is the number of outliers in the ith elemental set, then the W_i are iid hypergeometric$(d, n - d, p)$ random variables. Suppose that it is desired to find K such that the probability P(that at least one of the elemental sets is clean) $\equiv P_1 \approx 1 - \alpha$ where $\alpha = 0.05$ is a common choice. Then $P_1 = 1-$ P(none of the K elemental sets is clean)

$$\approx 1 - [1 - (1 - \gamma)^p]^K$$

by independence. Hence

$$\alpha \approx [1 - (1 - \gamma)^p]^K$$

or

$$K \approx \frac{\log(\alpha)}{\log([1 - (1 - \gamma)^p])} \approx \frac{\log(\alpha)}{-(1 - \gamma)^p} \tag{9.1}$$

using the approximation $\log(1 - x) \approx -x$ for small x. Since $\log(.05) \approx -3$, if $\alpha = 0.05$, then

$$K \approx \frac{3}{(1 - \gamma)^p}.$$

Frequently a clean subset is wanted even if the contamination proportion $\gamma \approx 0.5$. Then for a 95% chance of obtaining at least one clean elemental set, $K \approx 3\ (2^p)$ elemental sets need to be drawn.

Table 9.1: Largest p for a 95% Chance of a Clean Subsample.

					K			
γ	500	3000	10000	10^5	10^6	10^7	10^8	10^9
0.01	509	687	807	1036	1265	1494	1723	1952
0.05	99	134	158	203	247	292	337	382
0.10	48	65	76	98	120	142	164	186
0.15	31	42	49	64	78	92	106	120
0.20	22	30	36	46	56	67	77	87
0.25	17	24	28	36	44	52	60	68
0.30	14	19	22	29	35	42	48	55
0.35	11	16	18	24	29	34	40	45
0.40	10	13	15	20	24	29	33	38
0.45	8	11	13	17	21	25	28	32
0.50	7	9	11	15	18	21	24	28

Notice that number of subsets K needed to obtain a clean elemental set with high probability is an exponential function of the number of predictors p but is free of n. Hence if this choice of K is used in an elemental or concentration algorithm (that uses k concentration steps), then the algorithm is inconsistent and has (asymptotically) zero breakdown. Nevertheless, many practitioners use a value of K that is free of both n and p (eg $K = 500$ or $K = 3000$).

This practice suggests fixing both K and the contamination proportion γ and then finding the largest number of predictors p that can be in the model such that the probability of finding at least one clean elemental set is high. Given K and γ, P(at least one of K subsamples is clean) $= 0.95 \approx$ $1 - [1 - (1 - \gamma)^p]^K$. Thus the largest value of p satisfies

$$\frac{3}{(1-\gamma)^p} \approx K,$$

or

$$p \approx \left\lfloor \frac{\log(3/K)}{\log(1-\gamma)} \right\rfloor \tag{9.2}$$

if the sample size n is very large. Again $\lfloor x \rfloor$ is the greatest integer function: $\lfloor 7.7 \rfloor = 7$.

Table 9.1 shows the largest value of p such that there is a 95% chance that at least one of K subsamples is clean using the approximation given by Equation (9.2). Hence if $p = 28$, even with one billion subsamples, there is a 5% chance that none of the subsamples will be clean if the contamination proportion $\gamma = 0.5$. Since clean elemental fits have great variability, an algorithm needs to produce many clean fits in order for the best fit to be good. When contamination is present, all K elemental sets could contain outliers. Hence basic resampling and concentration algorithms that only use K elemental starts are doomed to fail if γ and p are large.

Remark 9.3: Breakdown. The breakdown value of concentration algorithms that use K elemental starts is bounded above by K/n. (See Section 9.4 for more information about breakdown.) For example if 500 starts are used and $n = 50000$, then the breakdown value is at most 1%. To cause a regression algorithm to break down, simply contaminate one observation in each starting elemental set so as to displace the fitted coefficient vector by a large amount. Since K elemental starts are used, at most K points need to be contaminated.

This is a worst-case model, but sobering results on the outlier resistance of such algorithms for a fixed data set with d gross outliers can also be derived. Assume that the LTA(c), LTS(c), or LMS(c) algorithm is applied to a fixed data set of size n where $n - d$ of the cases follow a well behaved model and $d < n/2$ of the cases are gross outliers. If $d > n - c$, then every criterion evaluation will use outliers, and every attractor will produce a bad fit even if some of the starts are good. If $d < n - c$ and if the outliers are far enough from the remaining cases, then clean starts of size $h \geq p$ may result in clean attractors that could detect certain types of outliers (that may need to be hugely discrepant on the response scale).

Proposition 9.1. Let γ_o be the highest percentage of massive outliers that a resampling algorithm can detect reliably. Then

$$\gamma_o \approx \min(\frac{n-c}{n}, 1 - [1 - (0.2)^{1/K}]^{1/h})100\% \tag{9.3}$$

if n is large.

Proof. In Remark 9.2, change p to h to show that if the contamination proportion γ is fixed, then the probability of obtaining at least one clean subset of size h with high probability (say $1 - \alpha = 0.8$) is given by $0.8 =$

$1 - [1 - (1 - \gamma)^h]^K$. Fix the number of starts K and solve this equation for γ. QED

The value of γ_o depends on $c > n/2$ and h. To maximize γ_o, take $c \approx n/2$ and $h = p$. For example, with $K = 500$ starts, $n > 100$, and $h = p \leq 20$ the resampling algorithm should be able to detect up to 24% outliers provided every clean start is able to at least partially separate inliers from outliers. However if $h = p = 50$, this proportion drops to 11%.

Remark 9.4: Hybrid Algorithms. More sophisticated algorithms use both concentration and partitioning. Partitioning evaluates the start on a subset of the data, and poor starts are discarded. This technique speeds up the algorithm, but the consistency and outlier resistance still depends on the number of starts. For example, Equation (9.3) agrees very well with the Rousseeuw and Van Driessen (1999) simulation performed on a hybrid MCD algorithm. (See Section 10.6.)

9.2 Advice for the Practitioner

Results from the previous section and chapter suggest several guidelines for the practitioner. Also see Section 6.3.

1) Make a response plot of $\widehat{Y}$ versus Y and a residual plot of $\widehat{Y}$ versus r. These plots are the most important diagnostics for multiple linear regression (MLR), and the list of real MLR "benchmark" data sets with outlier configurations that confound both plots is currently rather small. In general, do not overlook classical (OLS and L1) procedures and diagnostics. They often suffice where the errors e_i and their propensity to be outlying are independent of the predictors $\boldsymbol{x}_i$.

2) Theorem 8.8 shows how to modify elemental basic resampling and concentration algorithms so that the easily computed modified estimator is a $\sqrt{n}$ consistent HB estimator. The basic idea is simple: in addition to using K attractors from randomly selected elemental starts, also use two carefully chosen attractors. One should be an easily computed but biased HB attractor and the other attractor should be a $\sqrt{n}$ consistent estimator such as $\hat{\boldsymbol{\beta}}_{OLS}$. (Recall that the attractor = the start for the basic resampling algorithm.)

3) For 3 or fewer variables, use graphical methods such as scatterplots

and 3D plots to detect outliers and other model violations.

4) Make a scatterplot matrix of the predictors and the response if p is small. Often isolated outliers can be detected in the plot. Also, removing strong nonlinearities in the predictors with power transformations can be very useful.

5) Use several estimators – both classical and robust. (We recommend using OLS, L_1, the CLTS estimator from Theorem 8.8, `lmsreg`, the `tvreg` estimator from Section 11.3, `mbareg` and the MBA estimator using the LATA criterion (see Problem 7.5).) Then make a scatterplot matrix of i) the residuals and ii) the fitted values and response from the different fits. Also make a scatterplot matrix of the Mahalanobis distances of $\boldsymbol{x}_i$ using several of the distances discussed in Chapter 10. If the multiple linear regression model holds, then the subplots will be strongly linear if consistent estimators are used. Thus these plots can be used to detect a wide variety of violations of model assumptions.

6) Use subset refinement – concentration. Concentration may not improve the theoretical convergence rates, but concentration gives dramatic practical improvement in many data sets.

7) Compute the median absolute deviation of the response variable $\text{mad}(y_i)$ and the median absolute residual $\text{med}(|r|_i(\hat{\boldsymbol{\beta}}))$ from the estimator $\hat{\boldsymbol{\beta}}$. If $\text{mad}(y_i)$ is smaller, then the constant $\text{med}(y_i)$ fits the data better than $\hat{\boldsymbol{\beta}}$ according to the median squared residual criterion.

Other techniques, such as using *partitioning* to screen out poor starts, are also important. See Remark 9.4 and Woodruff and Rocke (1994). The *line search* may also be a good technique. Let $\boldsymbol{b}_b$ be the fit which currently minimizes the criterion. Ruppert (1992) suggests evaluating the criterion Q on

$$\lambda \boldsymbol{b}_b + (1 - \lambda)\boldsymbol{b}$$

where $\boldsymbol{b}$ is the fit from the current subsample and λ is between 0 and 1. Using $\lambda \approx 0.9$ may make sense. If the algorithm produces a good fit at some stage, then many good fits will be examined with this technique.

9.3 Desirable Properties of a Regression Estimator

There are many desirable properties for regression estimators including (perhaps in decreasing order of importance)
a) conditions under which $\hat{\boldsymbol{\beta}}_n$ is a consistent estimator,
b) computability (eg in seconds, or hours, or days),
c) the limiting distribution of $n^{\delta}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$,
d) rate and tightness results (see Definition 8.7): $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \asymp_P n^{-\delta}$ or $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = O_P(n^{-\delta})$,
e) conditions under which the slopes $(\hat{\beta}_{2,n}, ..., \hat{\beta}_{p,n})$ are consistent estimators of the population slopes $(\beta_2, ..., \beta_p)$ when the errors are asymmetric,
f) conditions under which $\hat{\boldsymbol{\beta}}_n$ is a consistent estimator of $\boldsymbol{\beta}$ when heteroscedasticity is present,
g) resistance of $\hat{\boldsymbol{\beta}}_n$ for a fixed data set of n cases of which $d < n/2$ are outliers,
h) equivariance properties of $\hat{\boldsymbol{\beta}}_n$, and
i) the breakdown value of $\hat{\boldsymbol{\beta}}_n$.

To some extent Chapter 8 and Remark 9.3 gave negative results: for the typical computable HB algorithms that used a fixed number of K elemental starts, the algorithm estimator $\boldsymbol{b}_{A,n}$ is inconsistent with an asymptotic breakdown value of zero. Section 9.1 discussed the resistance of such algorithm estimators for a fixed data set containing d outliers. Theorem 8.8 showed how to modify some of these algorithms, resulting in easily computed $\sqrt{n}$ consistent HB estimators, but the outlier resistance of the Theorem 8.8 estimators decreases rapidly as p increases.

Breakdown and equivariance properties have received considerable attention in the literature. Several of these properties involve transformations of the data. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the original data, then the vector of the coefficient estimates is

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}), \tag{9.4}$$

the vector of predicted values is

$$\widehat{\boldsymbol{Y}} = \widehat{\boldsymbol{Y}}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}), \tag{9.5}$$

and the vector of residuals is

$$\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}. \tag{9.6}$$

If the design $\boldsymbol{X}$ is transformed into $\boldsymbol{W}$ and the dependent variable $\boldsymbol{Y}$ is transformed into $\boldsymbol{Z}$, then $(\boldsymbol{W}, \boldsymbol{Z})$ is the new data set. Several of these important properties are discussed below, and we follow Rousseeuw and Leroy (1987, p. 116-125) closely.

Definition 9.2. Regression Equivariance: Let $\boldsymbol{u}$ be any $p \times 1$ vector. Then $\widehat{\boldsymbol{\beta}}$ is regression equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}) = T(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u} = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{u}. \quad (9.7)$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$, and $\boldsymbol{Z} = \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{u}$, then

$$\widehat{\boldsymbol{Z}} = \widehat{\boldsymbol{Y}} + \boldsymbol{X}\boldsymbol{u},$$

and

$$\boldsymbol{r}(\boldsymbol{W}, \boldsymbol{Z}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}).$$

Note that the residuals are invariant under this type of transformation, and note that if

$$\boldsymbol{u} = -\widehat{\boldsymbol{\beta}},$$

then regression equivariance implies that we should not find any linear structure if we regress the residuals on $\boldsymbol{X}$. Also see Problem 9.3.

Definition 9.3. Scale Equivariance: Let c be any scalar. Then $\widehat{\boldsymbol{\beta}}$ is scale equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, c\boldsymbol{Y}) = T(\boldsymbol{X}, c\boldsymbol{Y}) = cT(\boldsymbol{X}, \boldsymbol{Y}) = c\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \quad (9.8)$$

Hence if $\boldsymbol{W} = \boldsymbol{X}$, and $\boldsymbol{Z} = c\boldsymbol{Y}$ then

$$\widehat{\boldsymbol{Z}} = c\widehat{\boldsymbol{Y}},$$

and

$$\boldsymbol{r}(\boldsymbol{X}, c\boldsymbol{Y}) = c\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}).$$

Scale equivariance implies that if the Y_i's are stretched, then the fits and the residuals should be stretched by the same factor.

Definition 9.4. Affine Equivariance: Let $\boldsymbol{A}$ be any $p \times p$ nonsingular matrix. Then $\widehat{\boldsymbol{\beta}}$ is affine equivariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = T(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}T(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \quad (9.9)$$

Hence if $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{A}$ and $\boldsymbol{Z} = \boldsymbol{Y}$, then

$$\widehat{\boldsymbol{Z}} = \boldsymbol{W}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{A}\boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{Y}},$$

and

$$\boldsymbol{r}(\boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y}) = \boldsymbol{Z} - \widehat{\boldsymbol{Z}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}).$$

Note that both the predicted values and the residuals are invariant under an affine transformation of the independent variables.

Definition 9.5. Permutation Invariance: Let $\boldsymbol{P}$ be an $n \times n$ permutation matrix. Then $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is an $n \times n$ identity matrix and the superscript T denotes the transpose of a matrix. Then $\widehat{\boldsymbol{\beta}}$ is permutation invariant if

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = T(\boldsymbol{X}, \boldsymbol{Y}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{9.10}$$

Hence if $\boldsymbol{W} = \boldsymbol{P}\boldsymbol{X}$, and $\boldsymbol{Z} = \boldsymbol{P}\boldsymbol{Y}$, then

$$\widehat{\boldsymbol{Z}} = \boldsymbol{P}\widehat{\boldsymbol{Y}},$$

and

$$\boldsymbol{r}(\boldsymbol{P}\boldsymbol{X}, \boldsymbol{P}\boldsymbol{Y}) = \boldsymbol{P}\ \boldsymbol{r}(\boldsymbol{X}, \boldsymbol{Y}).$$

If an estimator is not permutation invariant, then swapping rows of the $n \times (p+1)$ augmented matrix $(\boldsymbol{X}, \boldsymbol{Y})$ will change the estimator. Hence the case number is important. If the estimator is permutation invariant, then the position of the case in the data cloud is of primary importance. Resampling algorithms are not permutation invariant because permuting the data causes different subsamples to be drawn.

9.4 The Breakdown of Breakdown

This section gives a standard definition of breakdown and then shows that if the median absolute residual is bounded in the presence of high contamination, then the regression estimator has a high breakdown value. The following notation will be useful. Let $\boldsymbol{W}$ denote the data matrix where the ith row corresponds to the ith case. For regression, $\boldsymbol{W}$ is the $n \times (p+1)$ matrix with ith row $(\boldsymbol{x}_i^T, Y_i)$. Let $\boldsymbol{W}_d^n$ denote the data matrix where any d of the cases have been replaced by arbitrarily bad contaminated cases. Then the contamination fraction is $\gamma = d/n$.

Definition 9.6. If $T(\boldsymbol{W})$ is a $p \times 1$ vector of regression coefficients, then the breakdown value of T is

$$B(T, \boldsymbol{W}) = \min\{\frac{d}{n} : \sup_{\boldsymbol{W}_d^n} \|T(\boldsymbol{W}_d^n)\| = \infty\}$$

where the supremum is over all possible corrupted samples $\boldsymbol{W}_d^n$ and $1 \leq d \leq n$.

The following result greatly simplifies some breakdown proofs and shows that a regression estimator basically breaks down if the median absolute residual MED($|r_i|$) can be made arbitrarily large. The result implies that if the breakdown value ≤ 0.5, breakdown can be computed using the median absolute residual MED($|r_i|(\boldsymbol{W}_d^n)$) instead of $\|T(\boldsymbol{W}_d^n)\|$.

Suppose that the proportion of outliers is less that 0.5. If the $\boldsymbol{x}_i$ are fixed, and the outliers are moved up and down the Y axis, then for high breakdown (HB) estimators, $\hat{\boldsymbol{\beta}}$ and MED($|r_i|$) will eventually be bounded. Thus if the $|Y_i|$ values of the outliers are large enough, the $|r_i|$ values of the outliers will be large.

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers tend to drive the slope estimates to 0, not ∞. If both $\boldsymbol{x}$ and Y can be varied, then a cluster of outliers can be moved arbitrarily far from the bulk of the data but still have small residuals. For example, move the outliers along the regression plane formed by the clean cases.

Proposition 9.2. If the breakdown value ≤ 0.5, computing the breakdown value using the median absolute residual MED($|r_i|(\boldsymbol{W}_d^n)$) instead of $\|T(\boldsymbol{W}_d^n)\|$ is asymptotically equivalent to using Definition 9.6.

Proof. Consider a fixed data set $\boldsymbol{W}_d^n$ with ith row $(\boldsymbol{w}_i^T, Z_i)^T$. If the regression estimator $T(\boldsymbol{W}_d^n) = \hat{\boldsymbol{\beta}}$ satisfies $\|\hat{\boldsymbol{\beta}}\| = M$ for some constant M, then the median absolute residual MED($|Z_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{w}_i|$) is bounded by $\max_{i=1,\ldots,n} |Y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i| \leq \max_{i=1,\ldots,n}[|Y_i| + \sum_{j=1}^p M|x_{i,j}|]$ if $d < n/2$.

Now suppose that $\|\hat{\boldsymbol{\beta}}\| = \infty$. Since the absolute residual is the vertical distance of the observation from the hyperplane, the absolute residual $|r_i| = 0$ if the ith case lies on the regression hyperplane, but $|r_i| = \infty$ otherwise. Hence MED($|r_i|$) $= \infty$ if fewer than half of the cases lie on the regression hyperplane. This will occur unless the proportion of outliers $d/n > (n/2 - q)/n \to 0.5$ as $n \to \infty$ where q is the number of "good" cases that lie on a

hyperplane of lower dimension than p. In the literature it is usually assumed that the original data is in general position: $q = p - 1$. QED

If the $(\boldsymbol{x}_i^T, Y_i)$ are in general position, then the contamination could be such that $\hat{\boldsymbol{\beta}}$ passes exactly through $p - 1$ "clean" cases and d "contaminated" cases. Hence $d + p - 1$ cases could have absolute residuals equal to zero with $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large (but finite). Nevertheless, if T possesses reasonable equivariant properties and $\|T(\boldsymbol{W}_d^n)\|$ is replaced by the median absolute residual in the definition of breakdown, then the two breakdown values are asymptotically equivalent. (If $T(\boldsymbol{W}) \equiv \boldsymbol{0}$, then T is neither regression nor affine equivariant. The breakdown value of T is one, but the median absolute residual can be made arbitrarily large if the contamination proportion is greater than $n/2$.)

If the Y_i's are fixed, arbitrarily large $\boldsymbol{x}$-outliers will rarely drive $\|\hat{\boldsymbol{\beta}}\|$ to ∞. The $\boldsymbol{x}$-outliers can drive $\|\hat{\boldsymbol{\beta}}\|$ to ∞ if they can be constructed so that the estimator is no longer defined, eg so that $\boldsymbol{X}^T\boldsymbol{X}$ is nearly singular. The following examples may help illustrate these points.

Example 9.1. Suppose the response values Y are near 0. Consider the fit from an elemental set:

$$\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$$

and examine Equations (8.2), (8.3), and (8.4) on p. 254. Now

$$\|\boldsymbol{b}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \, \|\boldsymbol{Y}_J\|,$$

and *since x-outliers make $\|\boldsymbol{X}_J\|$ large, x-outliers tend to drive $\|\boldsymbol{X}_J^{-1}\|$ and $\|\boldsymbol{b}_J\|$ towards zero not towards ∞.* The x-outliers may make $\|\boldsymbol{b}_J\|$ large if they can make the trial design $\|\boldsymbol{X}_J\|$ nearly singular. Notice that Euclidean norm $\|\boldsymbol{b}_J\|$ can easily be made large if one or more of the elemental response variables is driven far away from zero.

Example 9.2. Without loss of generality, assume that the clean Y's are contained in an interval $[a, f]$ for some a and f. Assume that the regression model contains an intercept β_1. Then there exists an estimator $\boldsymbol{b}_o$ of $\boldsymbol{\beta}$ such that $\|\boldsymbol{b}_o\| \leq \max(|a|, |f|)$ if $d < n/2$.

Proof. Let $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ and $\text{MAD}(n) = \text{MAD}(Y_1, ..., Y_n)$. Take $\boldsymbol{b}_o = (\text{MED}(n), 0, ..., 0)^T$. Then $\|\boldsymbol{b}_o\| = |\text{MED}(n)| \leq \max(|a|, |f|)$. Note that the median absolute residual for the fit $\boldsymbol{b}_o$ is equal to the median absolute

deviation $\text{MAD}(n) = \text{MED}(|Y_i - \text{MAD}(n)|, i = 1, ..., n) \leq f - a$ if $d < \lfloor (n+1)/2 \rfloor$. QED

A high breakdown regression estimator is an estimator which has a bounded median absolute residual even when close to half of the observations are arbitrary. Rousseeuw and Leroy (1987, p. 29, 206) conjecture that high breakdown regression estimators can not be computed cheaply, and they conjecture that if the algorithm is also affine equivariant, then the complexity of the algorithm must be at least $O(n^p)$. The following counterexample shows that these two conjectures are false.

Example 9.3. If the model has an intercept, then a scale and affine equivariant high breakdown estimator $\hat{\boldsymbol{\beta}}_{WLS}(k)$ can be found by computing OLS to the set of cases that have $Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm k \text{ MAD}(Y_1, ..., Y_n)]$ where $k \geq 1$ (so at least half of the cases are used). When $k = 1$, this estimator uses the "half set" of cases closest to $\text{MED}(Y_1, ..., Y_n)$.

Proof. This estimator has a median absolute residual bounded by $\sqrt{n}\ k\ \text{MAD}(Y_1, ..., Y_n)$. To see this, consider the weighted least squares fit $\hat{\boldsymbol{\beta}}_{WLS}(k)$ obtained by running OLS on the set J consisting of the n_j observations which have

$$Y_i \in [\text{MED}(Y_1, ..., Y_n) \pm k\text{MAD}(Y_1, ..., Y_n)] \equiv [\text{MED}(n) \pm k\text{MAD}(n)]$$

where $k \geq 1$ (to guarantee that $n_j \geq n/2$). Consider the estimator

$$\hat{\boldsymbol{\beta}}_M = (\text{MED}(n), 0, ..., 0)^T$$

which yields the predicted values $\hat{Y}_i \equiv \text{MED}(n)$. The squared residual

$$r_i^2(\hat{\boldsymbol{\beta}}_M) \leq (k \text{ MAD}(n))^2$$

if the ith case is in J. Hence the weighted LS fit has

$$\sum_{i \in J} r_i^2(\hat{\boldsymbol{\beta}}_{WLS}) \leq n_j (k \text{ MAD}(n))^2.$$

Thus

$$\text{MED}(|r_1(\hat{\boldsymbol{\beta}}_{WLS})|, ..., |r_n(\hat{\boldsymbol{\beta}}_{WLS})|) \leq \sqrt{n_j}\ k \text{ MAD}(n) < \infty.$$

Hence $\hat{\boldsymbol{\beta}}_{WLS}$ is high breakdown, and it is affine equivariant since the design is not used to choose the observations. It is scale equivariant since for $c = 0$, $\hat{\boldsymbol{\beta}}_{WLS} = \mathbf{0}$, and for $c \neq 0$ the set of cases used remains the same under scale transformations and OLS is scale equivariant. If k is huge and MAD$(n) \neq 0$, then this estimator and $\hat{\boldsymbol{\beta}}_{OLS}$ will be the same for most data sets. Thus high breakdown estimators can be very nonrobust.

Proposition 9.3. If a high breakdown start is added to a LTA, LTS or LMS concentration algorithm, then the resulting estimator is HB.

Proof. Concentration reduces the HB criterion that is based on $c_n \geq n/2$ absolute residuals, so the median absolute residual of the resulting estimator is bounded as long as the criterion applied to the HB estimator is bounded. QED

For example, consider the LTS(c_n) criterion. Suppose the ordered squared residuals from the mth start $\boldsymbol{b}_{0m} = \hat{\boldsymbol{\beta}}_{WLS}(1)$ are obtained. Then $\boldsymbol{b}_{1m}$ is simply the OLS fit to the cases corresponding to the c_n smallest squared residuals. Denote these cases by $i_1, ..., i_{c_n}$. Then

$$\sum_{i=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{1m}) \leq \sum_{j=1}^{c_n} r^2_{i_j}(\boldsymbol{b}_{0m}) = \sum_{j=1}^{c_n} r^2_{(i)}(\boldsymbol{b}_{0m})$$

where the second inequality follows from the definition of the OLS estimator. Hence concentration steps reduce the LTS criterion. If $c_n = (n+1)/2$ for n odd and $c_n = 1 + n/2$ for n even, then the criterion is bounded iff the median squared residual is bounded.

Example 9.4. Consider the smallest computer number A greater than zero and the largest computer number B. Choose k such that $kA > B$. Define the estimator $\hat{\boldsymbol{\beta}}$ as above if MAD$(Y_i, i = 1, ..., n)$ is greater than A, otherwise define the estimator to be $\hat{\boldsymbol{\beta}}_{OLS}$. Then we can just run OLS on the data without computing MAD$(Y_i, i = 1, ..., n)$.

Notice that if $\boldsymbol{b}_{0m} = \hat{\boldsymbol{\beta}}_{WLS}(1)$ is the $m = (K+1)$th start, then the attractor $\boldsymbol{b}_{km}$ found after k LTS concentration steps is also a HB regression estimator. Let $\hat{\boldsymbol{\beta}}_{k,B} = 0.99\boldsymbol{b}_{km}$. Then $\hat{\boldsymbol{\beta}}_{k,B}$ is a HB biased estimator of $\boldsymbol{\beta}$ (biased if $\boldsymbol{\beta} \neq \mathbf{0}$), and $\hat{\boldsymbol{\beta}}_{k,B}$ could be used as the biased HB estimator needed in Theorem 8.8. The following result shows that it is easy to make a HB estimator that is asymptotically equivalent to any consistent estimator, although the outlier resistance of the HB estimator is poor.

Proposition 9.4. Let $\hat{\boldsymbol{\beta}}$ be any consistent estimator of $\boldsymbol{\beta}$ and let $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}$ if $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}})) \leq \text{MED}(r_i^2(\hat{\boldsymbol{\beta}}_{k,B}))$. Let $\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}}_{k,B}$, otherwise. Then $\hat{\boldsymbol{\beta}}_H$ is a HB estimator that is asymptotically equivalent to $\hat{\boldsymbol{\beta}}$.

Proof. Since $\hat{\boldsymbol{\beta}}$ is consistent, $\text{MED}(r_i^2(\hat{\boldsymbol{\beta}})) \rightarrow \text{MED}(e^2)$ in probability where $\text{MED}(e^2)$ is the population median of the squared error e^2. Since the LMS estimator is consistent, the probability that $\hat{\boldsymbol{\beta}}$ has a smaller median squared residual than the biased estimator $\hat{\boldsymbol{\beta}}_{k,B}$ goes to 1 as $n \rightarrow \infty$. Hence $\hat{\boldsymbol{\beta}}_H$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}$. QED

The affine equivariance property can be achieved for a wide variety of algorithms. The following lemma shows that if $T_1, \ldots, T_K$ are K equivariant regression estimators and if T_Q is the T_j which minimizes the criterion Q, then T_Q is equivariant, too. A similar result is in Rousseeuw and Leroy (1987, p. 117). Also see Rousseeuw and Bassett (1991).

Lemma 9.5. Let $T_1, \ldots, T_K$ be K regression estimators which are regression, scale, and affine equivariant. Then if T_Q is the estimator whose residuals minimize a criterion which is a function Q of the absolute residuals such that

$$Q(|cr_1|, \ldots, |cr_n|) = |c|^d Q(|r_1|, \ldots, |r_n|)$$

for some $d > 0$, then T_Q is regression, scale, and affine equivariant.

Proof. By the induction principle, we can assume that $K = 2$. Since the T_j are regression, scale, and affine equivariant, the residuals do not change under the transformations of the data that define regression and affine equivariance. Hence T_Q is regression and affine equivariant. Let $r_{i,j}$ be the residual for the ith case from fit T_j. Now without loss of generality, assume that T_1 is the method which minimizes Q. Hence

$$Q(|r_{1,1}|, \ldots, |r_{n,1}|) < Q(|r_{1,2}|, \ldots, |r_{n,2}|).$$

Thus

$$Q(|cr_{1,1}|, \ldots, |cr_{n,1}|) = |c|^d Q(|r_{1,1}|, \ldots .|r_{n,1}|) <$$
$$|c|^d Q(|r_{1,2}|, \ldots, |r_{n,2}|) = Q(|cr_{1,2}|, \ldots, |cr_{n,2}|),$$

and T_Q is scale equivariant. QED

Since least squares is regression, scale, and affine equivariant, the fit from an elemental or subset refinement algorithm that uses OLS also has these

properties provided that the criterion Q satisfies the condition in Lemma 9.2. If

$$Q = \text{MED}(r_i^2),$$

then $d = 2$. If

$$Q = \sum_{i=1}^{h} (|r|_{(i)})^{\tau}$$

or

$$Q = \sum_{i=1}^{n} w_i |r_i|^{\tau}$$

where τ is a positive integer and $w_i = 1$ if

$$|r_i|^{\tau} < k \text{ MED}(|r_i|^{\tau}),$$

then $d = \tau$.

Remark 9.5. Similar breakdown results hold for multivariate location and dispersion estimators. See Section 10.5.

Remark 9.6. There are several important points about breakdown that do not seem to be well known. First, a breakdown result is weaker than even a result such as an estimator being asymptotically unbiased for some population quantity such as $\boldsymbol{\beta}$. This latter property is useful since if the asymptotic distribution of the estimator is a good approximation to the sampling distribution of the estimator, and if many independent samples can be drawn from the population, then the estimator can be computed for each sample and the average of all the different estimators should be close to the population quantity. The breakdown value merely gives a yes or no answer to the question of whether the median absolute residual can be made arbitrarily large when the contamination proportion is equal to γ, and having a bounded median absolute residual does not imply that the high breakdown estimator is asymptotically unbiased or in any way useful.

Secondly, the literature implies that the breakdown value is a measure of the global reliability of the estimator and is a lower bound on the amount of contamination needed to destroy an estimator. These interpretations are not correct since the complement of complete and total failure is *not* global reliability. The breakdown value d_T/n is actually an upper bound on the amount of contamination that the estimator can tolerate since the estimator can be made arbitrarily bad with d_T maliciously placed cases.

In particular, the breakdown value of an estimator tells nothing about more important properties such as consistency or asymptotic normality. Certainly we are reluctant to call an estimator robust if a small proportion of outliers can drive the median absolute residual to ∞, but this type of estimator failure is very simple to prevent. Notice that Example 9.3 suggests that many classical regression estimators can be approximated arbitrarily closely by a high breakdown estimator: simply make k huge and apply the classical procedure to the cases that have $Y_i \in [\text{MED}(n) \pm k \text{ MAD}(n)]$. Of course these high breakdown approximations may perform very poorly even in the presence of a single outlier.

Remark 9.7. The breakdown values of the LTx, RLTx, and LATx estimators was given by Proposition 7.5 on p. 236.

Since breakdown is a very poor measure of resistance, alternative measures are needed. The following description of resistance expands on remarks in Rousseeuw and Leroy (1987, p. 24, 70). Suppose that the data set consists of a cluster of clean cases and a cluster of outliers. Set $\boldsymbol{\beta} = \mathbf{0}$ and let the dispersion matrix of the "clean" cases $(\boldsymbol{x}_i^T, y_i)^T$ be the identity matrix $\boldsymbol{I}_{p+1}$. Assume that the dispersion matrix of the outliers is $c\boldsymbol{I}_{p+1}$ where $0 \leq c \leq 1$ and that γ is the proportion of outliers. Then the mean vectors of the clusters can be chosen to make the outliers bad leverage points. (This type of data set is frequently used in simulations where the affine and regression equivariance of the estimators is used to justify these choices.) It is well known that the LMS(c_n), LTA(c_n) and LTS(c_n) are defined by the "narrowest strip" covering c_n of the cases where the width of the strip is measured in the vertical direction with the L_∞, L_1, and L_2 criterion, respectively. We will assume that $c_n \approx n/2$ and focus on the LMS estimator since the narrowness of the strip is simply the vertical width of the strip.

Figure 9.1 will be useful for examining the resistance of the LMS estimator. The data set consists of 300 $N_2(\mathbf{0}, \boldsymbol{I}_2)$ clean cases and 200

$$N_2((9,9)^T, 0.25\boldsymbol{I}_2)$$

cases. Then the narrowest strip that covers only clean cases covers $1/[2(1-\gamma)]$ of the clean cases. For the artificial data, $\gamma = 0.4$, and 5/6 of the clean cases are covered and the width of the strip is approximately 2.76. The strip shown in Figure 9.1 consists of two parallel lines with y-intercepts of ± 1.38 and covers approximately 250 cases. As this strip is rotated counterclockwise

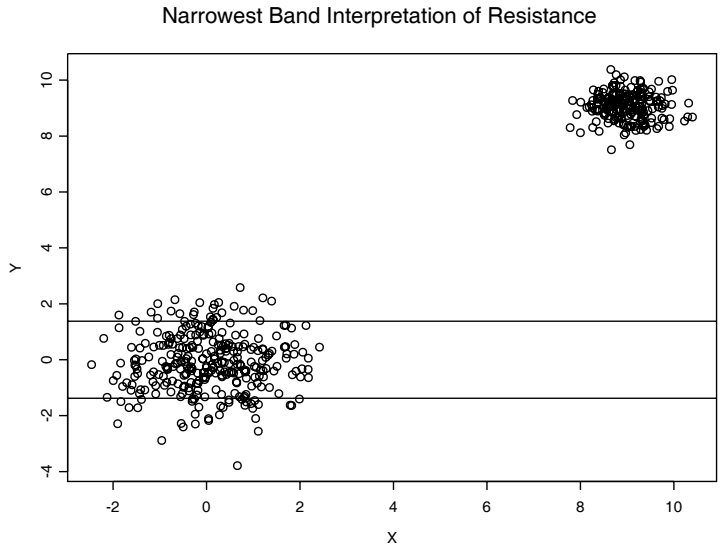

Figure 9.1: 300 N$(\mathbf{0}, \boldsymbol{I}_2)$ cases and 200 N$((9,9)^T, 0.25\boldsymbol{I}_2)$ cases

about the origin until it is parallel to the y-axis, the vertical width of the strip increases to ∞. Hence LMS will correctly produce a slope near zero if no outliers are present. Next, stop the rotation when the center of the strip passes through the center of both clusters, covering nearly 450 cases. The vertical width of the strip can be decreased to a value less than 2.76 while still covering 250 cases. Hence the LMS fit will accommodate the outliers, and with 40% contamination, an outlying cluster can tilt the LMS fit considerably. As $c \to 0$, the cluster of outliers tends to a point mass and even greater tilting is possible.

Also notice that once the narrowest band that determines the LMS estimator is established, the cluster of outliers can be moved along the band in such a way that the LMS estimator does not change. Hence masking will occur for the cluster **even if the cluster of outliers is arbitrarily far from the bulk of the data.** Notice that the response plot and the residual plot of fitted values versus residuals can be used to detect outliers with distant Y's. Since LMS is a HB estimator, if the $\boldsymbol{x}$'s of the outliers are fixed and the Y's go to ∞, then LMS will eventually give the outliers 0 weight, even if the outliers form a 40% point mass.

Next suppose that the 300 distinct clean cases lie exactly on the line

through the origin with zero slope. Then an "exact fit" to at least half of the data occurs and any rotation from this line can cover at most 1 of the clean cases. Hence a point mass will not be able to rotate LMS unless it consists of at least 299 cases (creating 300 additional exact fits). Similar results hold for the LTA and LTS estimators.

These remarks suggest that the narrowest band interpretation of the LTx estimators gives a much fuller description of their resistance than their breakdown value. Also, setting $\boldsymbol{\beta} = \mathbf{0}$ may lead to misleading simulation studies.

The band interpretation can also be used to describe the resistance of the LATx estimators. For example, the LATS(4) estimator uses an adaptive amount of coverage, but must cover at least half of the cases. Let $\boldsymbol{b}$ be the center of a band covering c_n cases. Then the LATS criterion inflates the band to cover $C_n(\boldsymbol{b})$ cases. If $\boldsymbol{b}$ passes through the center of both clusters in Figure 9.1, then nearly 100% of the cases will be covered. Consider the band with the x-axis as its center. The LATS criterion inflates the band to cover all of the clean cases but none of the outliers. Since only 60% of the cases are covered, the LATS(4) criterion is reduced and the outliers have large residuals. Although a large point mass can tilt the LATx estimators if the point mass can make the median squared residual small, the LATx estimators have a very strong tendency to give outlying clusters zero weight. In fact, the LATx estimator may tilt slightly to avoid a cluster of "good leverage" points if the cluster is far enough from the bulk of the data.

Problem 7.5 helps illustrate this phenomenon with the MBA estimators that use the MED(r_i^2) and LATA criteria. We suggest that the residuals and fitted values from these estimators (and from OLS and L_1) should be compared graphically with the RR and FF plots of Sections 6.3 and 7.6.

9.5 Complements

Feller (1957) is a great source for examining subsampling behavior when the data set is fixed. Hampel, Ronchetti, Rousseeuw and Stahel (1986, p. 96-98) and Donoho and Huber (1983) provide some history for breakdown. Maguluri and Singh (1997) have interesting examples on breakdown. Morgenthaler (1989) and Stefanski (1991) conjectured that high breakdown estimators with high efficiency are not possible. These conjectures have been shown to be false.

9.6 Problems

9.1 a) Enter or download the following *R/Splus* function

```
pifclean <- function(k, gam){
        p <- floor(log(3/k)/log(1 - gam))
list(p = p) }
```

b) Include the output from the commands below that are used to produce the second column of Table 9.1.

```
> zgam <- c(.01,.05,.1,.15,.2,.25,.3,.35,.4,.45,.5)
> pifclean(3000,zgam)
```

9.2. a) To get an idea for the amount of contamination a basic resampling or concentration algorithm can tolerate, enter or download the `gamper` function (with the *source("A:/rpack.txt")* command) that evaluates Equation (9.3) at different values of $h = p$.

b) Next enter the following commands and include the output in *Word*.

```
> zh <- c(10,20,30,40,50,60,70,80,90,100)
> for(i in 1:10) gamper(zh[i])
```

9.3. Assume that the model has a constant β_1 so that the first column of $\boldsymbol{X}$ is $\mathbf{1}$. Show that if the regression estimator is regression equivariant, then adding $\mathbf{1}$ to $\boldsymbol{Y}$ changes $\hat{\beta}_1$ but does not change the slopes $\hat{\beta}_2, ..., \hat{\beta}_p$.