# A Note on High Breakdown Regression Estimators

by

R.W. Oldford

Technical Report No. 39 August 1983

Massachusetts Institute of Technology
Center for Computational Research in Economics and Management Science
Alfred P. Sloan School of Management

## Abstract

High breakdown, without other measures of estimator resistance, is an inadequate goal for regression estimators. This is shown by constructing an easily computed regression estimator with 50% breakdown. The estimator is essentially least squares.



## Acknowledgements

Special thanks go to David Belsley, David Donoho, Peter Rousseeuw and Alexander Samarov for their helpful comments on an earlier draft and to Karen Martel for typing this paper.

## 1. Introduction

Suppose we have the regression model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \qquad i = 1, ..., n$$

where $\boldsymbol{\beta} \in \Re^p$ is the parameter vector and $(\boldsymbol{x}_i^T, y_i) \in \Re^{p+1}$ is the $i^{\text{th}}$ observation vector. Let $D_n = \{(\boldsymbol{x}_i^T, y_i) \ : \ i = 1, ..., n\}$ denote the sample of $n$ observations and $\boldsymbol{b}(D_n)$ be any estimator of $\boldsymbol{\beta}$ based on $D_n$. Further let $D_{n,m} \subset \Re^{p+1}$ be any set of cardinality $n$ such that $D_{n,m} \cap D_n$ has cardinality $m$ and let $\mathcal{D}_{n,m}$ be the set containing all such sets for fixed $m$ and $n$.

The finite sample replacement breakdown (Donoho and Huber (1983)) is defined to be $\epsilon^* = m/n$ where $m$ is the smallest integer such that

$$\sup_{D_{n,m} \in \mathcal{D}_{n,m}} \| \boldsymbol{b}(D_{n,m}) - \boldsymbol{b}(D_n) \| = \infty \tag{1.1}$$

and $\| \cdot \|$ denotes the Euclidean norm.

Recently [Siegel (1982), Donoho and Huber (1983), Rousseeuw (1982)] there has been increasing interest in the construction of regression estimators which have high breakdown. It is generally felt that the resulting estimates will provide good starting values for more efficient robust estimation procedures [e.g., Andrews (1974)]. Alternatively, such estimates may be used for exploratory data analytic purposes.

The purpose of the present note is to demonstrate the inadequacy of the definition (1.1) of breakdown for regression estimators. We do this by proposing a new estimator which has breakdown of 1/2, but which is clearly unsuitable for data analysis. In addition, this estimator is equivariant to non-singular transformations of the explanatory variables, highly efficient at the standard Gaussian model and trivially computed. Other high breakdown estimators, repeated medians (RM) [Siegel (1979, 1982)], and least median square (LMS) [Rousseeuw (1982)], do not possess all of those attributes.

The estimator exploits a peculiarity inherent in the definition of breakdown (1.1), namely, that the supremum must be infinite. Since the parameters being estimated are slope and/or intercept parameters, most estimators will give infinite estimates only if some of the $|y_i|$'s $\rightarrow \infty$. The proposed estimator prevents this occurrence.

## 2. The Estimator

Let $y'$ be the median of the $y_i$'s and MAD be their median absolute deviation, median $(|y_i - y'|)$. An estimator $b^*(D_n)$ may be constructed as a weighted least squares estimate with weights $w_i$ given by

$$w_i = \begin{cases} 1 & \text{if } \; |y_i - y'| \leq c \cdot \text{MAD} \\ 0 & \text{otherwise,} \end{cases} \tag{2.1}$$

where $c$ is some constant satisfying $1 \leq c < \infty$. The weights ignore the $x$-data entirely. Even so, we may prove that it has breakdown $1/2$ (as $n \to \infty$).

We make the following assumptions on the elements of $D_n$:

A1. $y_i \in \Re, \quad |y_i| < \infty \quad i = 1, ..., n$

A2. $\boldsymbol{x}_i \in \Re^p, \quad \| \boldsymbol{x}_i \| < \infty \quad i = 1, ..., n$

A3. $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are in general position, that is no $(p+1)$ points lie on a $(p-1)$ dimensional linear manifold.

And we introduce the following notation: let

(i) $\| A \|$ denote the spectral norm of a matrix $A$,

(ii) $J_m \subset \{1, ..., n\}$ of cardinality $m$ and $\overline{J}_m$ its complement in this set be defined implicitly by

$$D_{n,m} = \{\boldsymbol{z}_i \; : \; \boldsymbol{z}_i^T = (\boldsymbol{u}_i^T, v_i) \;\; \forall\, i \in J_m \;\text{ and }\; \boldsymbol{z}_i^T = (\boldsymbol{x}_i^T, y_i) \;\; \forall\, i \in \overline{J}_m\} \subset \Re^{p+1},$$

(iii) $W(D) = \text{diag}(w_1(D), ..., w_n(D))$ where $w_i(D)$'s are weights as in (2.1) based on a data set $D$,

(iv) $X(D)$ and $\boldsymbol{y}(D)$ be the usual $X$ matrix and $y$-vector constructed from a data set $D$,

(v) $\lambda_j(A)$ the $j^{\text{th}}$ largest eigenvalue of $A$,

(vi) $I(D) = \{i \; : \; w_i(D) = 1\} \subset \{1, ..., n\}$,

(vii) $[a]$ denote the largest integer less than or equal to $a$.

We may now prove the following result.

**Theorem**: Given A1-A3, the weighted least squares estimator $\boldsymbol{b}^*(D_n)$ with weights given by (2.1) has finite sample replacement breakdown of

$$\epsilon^* \geq \frac{n - [n/2] - p}{n}. \tag{2.2}$$

**Proof**: By A1-A3, $\| \boldsymbol{b}^*(D_n) \| < \infty$ and we need only examine $\| \boldsymbol{b}^*(D_{n,m}) \|$ in (1.1). By a property of the spectral norm [Wilkinson (1965)]

$$\begin{aligned} \| \boldsymbol{b}^*(D_{n,m}) \| \leq & \| (X^T(D_{n,m})W(D_{n,m})X(D_{n,m}))^{-1}X(D_{n,m})^T W(D_{n,m}) \| \\ & \cdot \| W(D_{n,m})\boldsymbol{y}(D_{n,m}) \| \end{aligned} \tag{2.3}$$

and

$$\begin{aligned} & \| (X^T(D_{n,m})W(D_{n,m})X(D_{n,m}))^{-1}X(D_{n,m})^T W(D_{n,m}) \|^2 \\ & \quad = \lambda_p^{-1} \left( \sum_{i \in J_m} w_i(D_{n,m})\boldsymbol{u}_i\boldsymbol{u}_i^T + \sum_{i \in \overline{J}_m} w_i(D_{n,m})\boldsymbol{x}_i\boldsymbol{x}_i^T \right) \\ & \quad \leq \lambda_p^{-1} \left( \sum_{i \in \overline{J}_m} w_i(D_{n,m})\boldsymbol{x}_i\boldsymbol{x}_i^T \right) \\ & \quad = \lambda_p^{-1} \left( \sum_{i \in \overline{J}_m \cap I(D_{n,m})} \boldsymbol{x}_i\boldsymbol{x}_i^T \right). \end{aligned} \tag{2.4}$$

The inequality may be found for example in Lawson and Hanson (1974). By A3, (2.4) is finite provided the cardinality of $\overline{J}_m \cap I(D_{n,m})$ is greater than $p$. The cardinality of $\overline{J}_m$ is $n - m$. Let that of $I(D_{n,m})$ be $k$. The cardinality of their intersection is guaranteed to be greater than $p$ if $(n-m)+k-n > p$ or $m < k - p$. Since $k \geq n - [n/2]$ the last quantity is certainly finite if $m < n - [n/2] - p$. Also,

$$\| W(D_{n,m})\boldsymbol{y}(D_{n,m}) \|^2 = \sum_{i \in J_m} w_i(D_{n,m})v_i^2 + \sum_{i \in \overline{J}_m} w_i(D_{n,m})y_i^2$$

which is always finite provided $m < \left[\frac{n}{2}\right]$. Therefore

$$\| \boldsymbol{b}^*(D_{n,m}) \| < \infty \qquad \forall D_{n,m}$$

for all $m < n - [n/2] - p$ and the theorem is proved.

Since $\epsilon^*$ is bounded from above by $\frac{1}{2}$, as $n \to \infty$, $\epsilon^* \to \frac{1}{2}$. Note also that the theorem could be proven with the inclusion of an intercept term.

## 3. Discussion

In addition to having a high breakdown value, the estimator is equivariant to non-singular transformations and easily computed. For $c$ sufficiently large in (2.1), the estimator is essentially least squares and therefore highly efficient at the usual Gaussian Model. Further, $c$ can be chosen large enough so that, for practical purposes, the estimator is equivariant to location transformations of the regression parameters.

However, it is clearly not a resistant regression estimator. While sup $\| \boldsymbol{b}^*(D_n) - \boldsymbol{b}(D_{n,m}) \|$ is bounded for $m < n - [n/2] - p$, it may be quite large for $m = 1$ (finite-sample-influence or sensitivity function). Figure 1 demonstrates that this can also be the case for the LMS estimator. Here a single point essentially determines the line. Rousseeuw (1982) gives an example with $p = 2$, where RM is substantially affected by 40% contamination but LMS is not. Clearly these estimators differ on other resistance properties.

Breakdown describes a worst-possible-case scenario. Because of this it is often more easily assessed than an estimator's sensitivity curve (e.g., LMS). However, without assessment of other resistance properties, it may be misleading. High breakdown implies bounded sensitivity but the bounds may be high enough to be ineffectual.

At least two alternative definitions of breakdown are possible. The first would replace $\infty$ in (1.1) with some constant $k$. Breakdown would now be a function $\epsilon^*(k)$ and more difficult to assess. Moreover, breakdown would no longer be invariant to a linear transformation of $\boldsymbol{\beta}$.

The second possibility is closer in spirit to the original finite sample version of Andrews et al. (1971). It consists of fixing the sets $D_n$ and $\mathcal{D}_{n,m}$ to be investigated. In such cases $\infty$ is usually replaced by $k < \infty$.

Donoho and Rousseeuw (1983; personal communication) have suggested that the *exact fit property* (EFP) of a regression estimator be investigated. Basically the set $D_n$ is chosen to be such that all $(\boldsymbol{x}_i^T, y_i)$ lie exactly on a plane. $\mathcal{D}_{n,m}$ is as before and $\boldsymbol{b}(D_n)$ is said to have the exact fit property $\epsilon = m/n$ if $m$ is the largest integer such that $\sup_{D_{n,m} \in \mathcal{D}_{n,m}} \| \boldsymbol{b}(D_n) - \boldsymbol{b}(D_{n,m}) \| = 0$. Siegel (1979) and Rousseeuw (1983) proved that the RM and LMS respectively have EFP of $\epsilon = 1/2$.

Other possibilities for $D_n$ and $\mathcal{D}_{n,m}$ should be investigated. It may be the case that estimators which are reasonable on other grounds "break down" only for pairs $(D_n, D_{n,m})$ which rarely occur in practice. It would be helpful to have breakdown type assessments for commonly occurring $(D_n, D_{n,m})$ and exploratory techniques for recognizing those infrequent pairs $(D_n, D_{n,m})$ which are dangerous to the estimator being used.

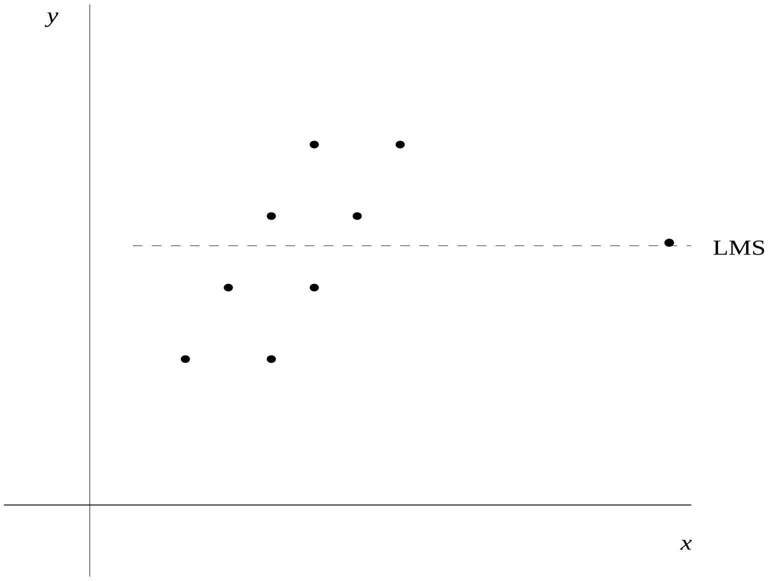


Figure 1. The LMS estimate for a particular $D_{n,m}$.

**References**


Andrews, D.F., P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers, and J.W. Tukey (1972), *Robust Estimates of Location: Survey and Advances*, Princeton University Press: Princeton, New Jersey.

Andrews, D.F. (1974), "A Robust Method for Multiple Linear Regression", *Technometrics 16*, pp. 523-531.

Donoho, D.L. and P.J. Huber (1983), "The Notion of Breakdown Point", *E.L. Lehmann Festschriftt*, Bickel, P.J., K. Doksum and J.L. Hodges, Jr. (eds.), Wadsworth Press.

Lawson, C.R. and R.J. Hanson (1974), *Solving Least Squares Problems*, Prentice-Hall: Englewood Cliffs, New Jersey.

Rousseeuw, P.J. (1982), "Least Median of Squares Regression", Technical Report, Centrum voor Statistiek en Operationeel Onderzoek, Vrije Universiteit Brussel.

Siegel, A.F. (1979), "The Repeated Median Algorithm", unpublished working paper, University of Wisconsin.

Siegel, A.F. (1982), "Robust Regression using Repeated Medians", *Biometrika, 69*, pp. 242-244.

Wilkinson, J.H. (1965), *The Algebraic Eigenvalue Problem*, Oxford University Press, London.