3-2007

Behavior of Elemental Sets in Regression

David J. Olive
Southern Illinois University Carbondale, dolive@math.siu.edu

Douglas M. Hawkins
University of Minnesota - Twin Cities

Published in *Statistics & Probability Letters*, 77, 621-624. doi: 10.1016/j.spl.2006.09.009

Behavior of elemental sets in regression

David J. Olive and Douglas M. Hawkins *

Southern Illinois University and University of Minnesota

June 12, 2005

Abstract

Elemental sets are used to produce trial estimates $\boldsymbol{b}$ of the regression coefficients $\boldsymbol{\beta}$. If $\boldsymbol{b}_o$ minimizes $\|\boldsymbol{b}-\boldsymbol{\beta}\|$ among all elemental fits $\boldsymbol{b}$, then $\|\boldsymbol{b}_o-\boldsymbol{\beta}\| = O_P(n^{-1})$, regardless of the criterion used. For any estimator $\boldsymbol{b}_A$, $\|\boldsymbol{b}_A-\boldsymbol{\beta}\|$ is at best $O_P(n^{-1/2})$. Hence restricting fits to elemental introduces asymptotically negligible error.

KEY WORDS: Breakdown; Depth; LMS; LTA; Outliers; Robust Regression.

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@math.siu.edu. Douglas M. Hawkins is Professor, School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church St. SE, Minneapolis, MN 55455-0493, USA. E-mail address: doug@stat.umn.edu. This research was supported by NSF grants DMS 9806584 and DMS 0202922.

1 Introduction

Consider the regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, and $\boldsymbol{e}$ is an $n \times 1$ vector of errors. The ith case $(\boldsymbol{x}_i^T, y_i)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$.

High breakdown (HB) estimators are used to produce "fits" that resist outliers. Important examples include the least median of squares (LMS) estimator (Hampel 1975), the least trimmed squares (LTS) estimator (Rousseeuw 1984), the least trimmed absolute deviations (LTA) estimator (Hössjer 1994) and the regression depth (RD) estimator (Rousseeuw and Hubert 1999). The computational complexities of the LTA, LMS and RD exact algorithms are $O(n^{p+1})$, $O(n^{p+2})$ and $O(n^{2p-1} \log n)$, respectively. Since these exact algorithms are impractical, approximate algorithms are generally used.

Many algorithms use subsets of p cases called "elemental sets." The oldest such method is the "basic resampling" or "elemental set" algorithm (Siegel 1982; Rousseeuw 1984; Hawkins, Bradu, and Kass 1984), and some estimators can be found by searching all $C(n,p) = \binom{n}{p}$ elemental sets. Examples include least absolute deviations (L_1), regression depth, the repeated median (Siegel 1982) and LTA.

Following Lehmann (1999, pp. 53-54), recall that the sequence of random variables W_n is *tight* or *bounded in probability,* $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that $P(|W_n| \leq D_\epsilon) \geq 1 - \epsilon$ for all $n \geq N_\epsilon$. Also $W_n = O_P(X_n)$ if $|W_n/X_n| = O_P(1)$. W_n has the same order as X_n in probability, written $W_n \asymp_P X_n$, if

$W_n = O_P(X_n)$ and $X_n = O_P(W_n)$.

If $W_n = \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\| \asymp_P n^{-\delta}$ for some $\delta > 0$, then we say that both W_n and $\hat{\boldsymbol{\beta}}_n$ *have rate* n^{δ}. Notice that if $W_n = O_P(n^{-\delta})$, then n^{δ} is a lower bound on the rate of W_n. As an example, if LMS, least squares (OLS) or L_1 is used for $\hat{\boldsymbol{\beta}}$, then $W_n = O_P(n^{-1/3})$, but $W_n \asymp_P n^{-1/3}$ for LMS while $W_n \asymp_P n^{-1/2}$ for OLS and L_1.

In the basic resampling algorithm, K_n elemental sets are randomly selected. An exact fit of the regression is performed for each subset, producing the estimators $\boldsymbol{b}_{1,n}, ..., \boldsymbol{b}_{K_n,n}$. Then the algorithm estimator $\boldsymbol{b}_{A,n}$ is the elemental fit that minimized the regression criterion Q. Let $\hat{\boldsymbol{\beta}}_{Q,n}$ denote the estimator that the algorithm is approximating, e.g., $\hat{\boldsymbol{\beta}}_{LTS,n}$. Let $\boldsymbol{b}_{o,n}$ be the "best" elemental fit examined by the algorithm in that

$$\boldsymbol{b}_{o,n} = \text{argmin}_{h=1,...,K_n} \|\boldsymbol{b}_{h,n} - \boldsymbol{\beta}\| \tag{2}$$

where K_n is the number of random starts and the Euclidean norm is used. Since the algorithm estimator is an elemental fit, $\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| \geq \|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\|$, and an upper bound on the rate of $\boldsymbol{b}_{o,n}$ is an upper bound on the rate of $\boldsymbol{b}_{A,n}$.

Hawkins and Olive (2002) proved that under weak conditions $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| \leq O_P(K_n^{-1/p})$. Since the rate of $\boldsymbol{b}_{A,n}$ is bounded above by the rate of $\boldsymbol{b}_{o,n}$ regardless of the criterion Q, this result is one of the most powerful tools for examining the behavior of robust estimators actually used in practice. For example, an estimator $\boldsymbol{b}_{A,n}$ that uses n randomly drawn elemental sets satisfies $\|\boldsymbol{b}_{A,n} - \boldsymbol{\beta}\| \leq O_P(n^{-1/p})$. When all elemental sets are searched, the rate of $\boldsymbol{b}_{o,n} \in [n^{1/2}, n]$ since the L_1 estimator is elemental and provides the lower bound. Section 2 establishes that $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(K_n^{-1/p})$ and that the number of elemental sets $\boldsymbol{b}_{i,n}$ that satisfy $\|\boldsymbol{b}_{i,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \leq 1$ is proportional to $n^{p(1-\delta)}$.

2 The anatomy of elemental sets

The following observations are useful for examining elemental sets. Let $J = J_h = \{h_1, ..., h_p\}$ be a randomly selected elemental set. Then $\boldsymbol{Y}_{J_h} = \boldsymbol{X}_{J_h}\boldsymbol{\beta} + \boldsymbol{e}_{J_h}$ where $\boldsymbol{Y}_{J_h}$ and $\boldsymbol{e}_{J_h}$ are $p \times 1$ vectors and $\boldsymbol{X}_{J_h}$ is a $p \times p$ matrix. Denote the ith entry of $\boldsymbol{Y}_{J_h}$ by y_{hi}, the ith entry of $\boldsymbol{e}_{J_h}$ by e_{hi}, and the ij entry of $\boldsymbol{X}_{J_h}$ by $x_{hi,j}$. Denote the ith elemental case by $(\boldsymbol{x}_{hi}^T, y_{hi})$. The subscript h will often be suppressed. Then the elemental data $(\boldsymbol{Y}_J, \boldsymbol{X}_J)$ produce an estimator $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1} Y_J$ of $\boldsymbol{\beta}$, and $\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{X}_J^{-1}\boldsymbol{e}_J\| \leq \|\boldsymbol{X}_J^{-1}\| \|\boldsymbol{e}_J\|$. Let $0 \leq \sigma_p \leq \sigma_{p-1} \leq \cdots \leq \sigma_1$ denote the singular values of $\boldsymbol{X}_J$. Then the following results (Golub and Van Loan 1989, pp. 57, 80) on the Euclidean norm are useful:

$$\|\boldsymbol{X}_J^{-1}\| = \frac{\sigma_1}{\sigma_p \|\boldsymbol{X}_J\|}, \tag{3}$$

$$\max_{i,j} |x_{hi,j}| \leq \|\boldsymbol{X}_J\| \leq p \max_{i,j} |x_{hi,j}|, \text{ and} \tag{4}$$

$$\frac{1}{p \max_{i,j} |x_{hi,j}|} \leq \frac{1}{\|\boldsymbol{X}_J\|} \leq \|\boldsymbol{X}_J^{-1}\|. \tag{5}$$

Two assumptions are used, but results that do not use (A2) are given later.

(A1) The errors are iid, independent of the predictors, and have a density f that is positive and continuous in a neighborhood of zero.

(A2) Let τ be proportion of elemental sets J that satisfy $\|\boldsymbol{X}_J^{-1}\| \leq B$ for some constant $B > 0$. Assume $\tau > 0$.

These assumptions are reasonable. If the errors can be arbitrarily placed, then they could cause the estimator to oscillate about $\boldsymbol{\beta}$. Hence no estimator would be consistent for $\boldsymbol{\beta}$. Note that if $\epsilon > 0$ is small enough, then $P(|e_i| \leq \epsilon) \approx 2\epsilon f(0)$. Equations (3) and (4) suggest that (A2) will hold unless the data is very badly behaved.

Theorem 1. *Assume that all $C(n,p)$ elemental subsets are searched and that (A1) and (A2) hold. Then $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(n^{-1})$.*

Proof. Let the random variable $W_{n,\epsilon}$ count the number of errors e_i that satisfy $|e_i| \leq M_\epsilon/n$ for $i = 1, ..., n$. For fixed n, $W_{n,\epsilon}$ is a binomial random variable with parameters n and P_n where $nP_n \to 2f(0)M_\epsilon$ as $n \to \infty$. Hence $W_{n,\epsilon}$ converges in distribution to a Poisson$(2f(0)M_\epsilon)$ random variable, and for any fixed integer $k > p$, $P(W_{n,\epsilon} > k) \to 1$ as $M_\epsilon \to \infty$ and $n \to \infty$. Hence if n is large enough, then with arbitrarily high probability there exists an M_ϵ such that at least $C(k,p)$ elemental sets J_{h_n} have all $|e_{h_n i}| \leq M_\epsilon/n$ where the subscript h_n indicates that the sets depend on n. By condition (A2), the proportion of these $C(k,p)$ fits that satisfy $\|\boldsymbol{b}_{J_{h_n}} - \boldsymbol{\beta}\| \leq B\sqrt{p}M_\epsilon/n$ is greater than τ. If k is chosen sufficiently large, and if n is sufficiently large, then with arbitrarily high probability, $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| \leq B\sqrt{p}M_\epsilon/n$ and the result follows. QED

Corollary 2. *Assume that $H_n \leq n$ but $H_n \to \infty$ as $n \to \infty$. If (A1) and (A2) hold, and if $K_n = H_n^p$ randomly chosen elemental sets are used, then $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(H_n^{-1}) = O_P(K_n^{-1/p})$.*

Proof. Suppose H_n cases are drawn without replacement and all $C(H_n,p) \propto H_n^p$ elemental sets are examined. Then by Theorem 1, the best elemental set selected by this procedure has rate H_n^{-1}. Hence if $K_n = H_n^p$ randomly chosen elemental sets are used and if n is sufficiently large, then the probability of drawing an elemental set J_{h_n} such that $\|\boldsymbol{b}_{J_{h_n}} - \boldsymbol{\beta}\| \leq M_\epsilon H_n^{-1}$ goes to one as $M_\epsilon \to \infty$ and the result follows. QED

Suppose that an elemental set J is "good" if $\|\boldsymbol{b}_J - \boldsymbol{\beta}\| \leq M_\epsilon H_n^{-1}$ for some constant $M_\epsilon > 0$. If $H_n = n^\delta$ where $0 < \delta \leq 1$, then the number of "good" sets is proportional to

$n^{p(1-\delta)}$.

The following argument shows that similar results hold if the predictors are iid with a multivariate density that is everywhere positive. Assume that the regression model contains a constant: $\boldsymbol{x} = (1, x_2, ..., x_p)^T$. Construct a (hyper) pyramid and place the "corners" of the pyramid into a $p \times p$ matrix $\boldsymbol{W}$. The pyramid defines p "corner regions" $R_1, ..., R_p$. The p points that form $\boldsymbol{W}$ are not actual observations, but the fit $\boldsymbol{b}_J$ can be evaluated on $\boldsymbol{W}$. Define the $p \times 1$ vector $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}$. Then $\boldsymbol{\beta} = \boldsymbol{W}^{-1}\boldsymbol{z}$, and $\hat{\boldsymbol{z}} = \boldsymbol{W}\boldsymbol{b}_J$ is the fitted hyperplane evaluated at the corners of the pyramid. If an elemental set has one observation in each corner region and if all p absolute errors are less than ϵ, then the absolute deviation $|\delta_i| = |z_i - \hat{z}_i| < \epsilon$, $i = 1, ..., p$.

Examining these pyramids in low dimensions may help clarify the idea. If $p = 2$, then the 1-dimensional pyramid is simply a line segment $[w_1, w_2]$, region $R_1 = \{x_2 : x_2 \leq w_1\}$ and let region $R_2 = \{x_2 : x_2 \geq w_2\}$. Now assume that $p = 3$ and the two nontrivial predictors are scattered about the origin. Then the three points $(a, -a/2)^T$, $(-a, -a/2)^T$, and $(0, a/2)^T$ determine a triangle where $a > 0$. Use this triangle as the pyramid and let

$$\boldsymbol{W} = \begin{bmatrix} 1 & a & -a/2 \\ 1 & -a & -a/2 \\ 1 & 0 & a/2 \end{bmatrix}.$$

The corner regions are formed by extending the three lines that form the triangle and using points that fall opposite of a corner of the triangle.

For general $p \geq 2$, form a $(p-1)$-dimensional pyramid and let $\boldsymbol{W}$ be the matrix formed from the p pyramid corners. Then each of the p corner regions is formed by extending the $p-1$ surfaces of the pyramid that form the corner. The notation $\boldsymbol{x} \in R_i$

will be used to indicate that $(x_2, ..., x_p)^T \in R_i$.

Lemma 3. *Fix the pyramid that determines* $(\boldsymbol{z}, \boldsymbol{W})$ *and consider any elemental set* $(\boldsymbol{X}_J, \boldsymbol{Y}_J)$ *with each point* $(\boldsymbol{x}_{hi}^T, y_{hi})$ *such that* $\boldsymbol{x}_{hi} \in$ *a corner region* R_i *and each absolute error* $|y_{hi} - \boldsymbol{x}_{hi}^T\boldsymbol{\beta}| \leq \epsilon$. *Then the elemental set produces a fit* $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$ *such that*

$$\|\boldsymbol{b}_J - \boldsymbol{\beta}\| \leq \|\boldsymbol{W}^{-1}\| \sqrt{p}\, \epsilon. \tag{6}$$

Proof. Let the $p \times 1$ vector $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{\beta}$, and consider any subset $J = \{h_1, h_2, ..., h_p\}$ with $\boldsymbol{x}_{hi}$ in R_i and $|e_{hi}| < \epsilon$ for $i = 1, 2, ..., p$. The fit from this subset is determined by $\boldsymbol{b}_J = \boldsymbol{X}_J^{-1}\boldsymbol{Y}_J$ so $\hat{\boldsymbol{z}} = \boldsymbol{W}\boldsymbol{b}_J$. Let the $p \times 1$ deviation vector $\boldsymbol{\delta} = (\delta_1, ..., \delta_p)^T$ where $\delta_i = z_i - \hat{z}_i$. Then $\boldsymbol{b}_J = \boldsymbol{W}^{-1}(\boldsymbol{z} - \boldsymbol{\delta})$ and $|\delta_i| \leq \epsilon$ by construction. Thus $\|\boldsymbol{b}_J - \boldsymbol{\beta}\| = \|\boldsymbol{W}^{-1}\boldsymbol{z} - \boldsymbol{W}^{-1}\boldsymbol{\delta} - \boldsymbol{W}^{-1}\boldsymbol{z}\| \leq \|\boldsymbol{W}^{-1}\|\|\boldsymbol{\delta}\| \leq \|\boldsymbol{W}^{-1}\|\sqrt{p}\, \epsilon$. QED

Next we will consider all $C(n, p)$ elemental sets and again show that best elemental fit $\boldsymbol{b}_{o,n}$ satisfies $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(n^{-1})$. To get a bound, we need to assume that the number of observations in each of the p corner regions is proportional to n. This assumption is satisfied if the nontrivial predictors are iid from a distribution with a joint density that is positive on the entire $(p-1)-$dimensional Euclidean space. We replace (A2) by

(A3): Assume that the probability that a randomly selected $\boldsymbol{x} \in R_i$ is bounded below by $\alpha_i > 0$ for large enough n and $i = 1, ..., p$.

If U_i counts the number of cases $(\boldsymbol{x}_j^T, y_j)$ that have $\boldsymbol{x}_j \in R_i$ and $|e_i| < M_\epsilon/H_n$, then U_i is a binomial random variable with success probability proportional to M_ϵ/H_n, and the number G_n of elemental fits $\boldsymbol{b}_J$ satisfying equation (6) with ϵ replaced by M_ϵ/H_n satisfies

$$G_n \geq \prod_{i=1}^{p} U_i \propto n^p (\frac{M_\epsilon}{H_n})^p.$$

Hence the probability that a randomly selected elemental set $\boldsymbol{b}_J$ that satisfies $\|\boldsymbol{b}_J - \boldsymbol{\beta}\| \leq \|\boldsymbol{W}^{-1}\| \sqrt{p}\, M_\epsilon/H_n$ is bounded below by a probability that is proportional to $(M_\epsilon/H_n)^p$. If the number of randomly selected elemental sets $K_n = H_n^p$, then

$$P(\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| \leq \|\boldsymbol{W}^{-1}\| \sqrt{p}\, \frac{M_\epsilon}{H_n}) \to 1$$

as $M_\epsilon \to \infty$. These remarks prove the following corollary.

Corollary 4. *Assume that (A1) and (A3) hold. Let $H_n \leq n$ and assume that $H_n \to \infty$ as $n \to \infty$. If $K_n = H_n^p$ elemental sets are randomly chosen then*

$$\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(H_n^{-1}) = O_P(K_n^{-1/p}).$$

In particular, if all $C(n,p)$ elemental sets are examined, then $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(n^{-1})$. The following result shows that elemental fits can be used to approximate any $p \times 1$ vector $\boldsymbol{c}$, and are thus useful for projection pursuit. Of course this result is asymptotic, and some vectors will not be well approximated for reasonable sample sizes.

Theorem 5. *Assume that (A1) and (A3) hold and that the error density f is positive and continuous everywhere. Then the closest elemental fit $\boldsymbol{b}_{c,n}$ to any $p \times 1$ vector $\boldsymbol{c}$ satisfies $\|\boldsymbol{b}_{c,n} - \boldsymbol{c}\| = O_P(n^{-1})$.*

Proof sketch. The proof is essentially the same. Sandwich the plane determined by $\boldsymbol{c}$ by only considering points such that $|g_i| = |y_i - \boldsymbol{x}_i^T \boldsymbol{c}| < \alpha$. Since the e_i's have positive density, $P(|g_i| < \alpha) \propto 1/\alpha)$ (at least for $\boldsymbol{x}_i$ in some ball of possibly huge radius R about the origin). Also the pyramid needs to lie on the $\boldsymbol{c}$-plane and the corner regions will have smaller probabilities. By placing the pyramid so that $\boldsymbol{W}$ is in the "center" of the $\boldsymbol{X}$ space, we may assume that these probabilities are bounded away from zero, and make M_ϵ so large that the probability of a "good" elemental set is larger than $1 - \epsilon$. QED

3 References

Golub, G.H., and Van Loan, C.F. (1989), *Matrix Computations* (John Hopkins University Press, Baltimore, MD., 2nd ed.).

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

Hawkins, D.M., Bradu, D., and Kass, G.V. (1984), Location of several outliers in multiple regression data using elemental sets, *Technom.* **26**, 197-208.

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

Hössjer, O. (1994), Rank-based estimates in the linear model with high breakdown point, *J. Amer. Statist. Assoc.* **89**, 149-158.

Lehmann, E.L. (1999), *Elements of Large–Sample Theory*, (Wiley, New York).

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

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

Siegel, A.F. (1982), Robust regression using repeated medians, *Biometrika.* **69**, 242-244.