Elemental Fits are Dense

David J. Olive *

Southern Illinois University

July 26, 2003

Abstract

Elemental sets are subsets of the data which are just large enough to produce an estimate $\boldsymbol{b}$ of the coefficients $\boldsymbol{\beta}$. In the elemental 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. Suppose that $K_n \propto n$ elemental sets are randomly selected. Let $\boldsymbol{b}_{o,n}$ be the "best" elemental fit examined by the algorithm. Then $\|\boldsymbol{b}_{o,n} - \boldsymbol{\beta}\| = O_P(n^{-1/p})$, and elemental fits are "dense" since $\boldsymbol{\beta}$ can be replaced by any vector $\boldsymbol{c}$.

KEY WORDS: Combinatorics; Elemental Sets; Outliers; Robust Estimation.

*David J. Olive is Assistant Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA.

1 Introduction

Consider the Gaussian regression model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e} \tag{1.1}$$

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 $(y_i, \boldsymbol{x}_i^T)$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith row of $\boldsymbol{Y}$.

Elemental sets are subsets of p cases and are just large enough to produce an estimate $\boldsymbol{b}$ of the coefficients $\boldsymbol{\beta}$. In the elemental set or 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

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

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

2 Behavior of the Best Elemental Fit

The main result of this paper is an analytic proof that the best elemental subset has a $n^{-1/p}$ convergence rate if the errors are Gaussian and $K_n = [n/p]$ nonoverlapping elemental sets from the n cases are used. Let

$$J_i = \{j_1, ..., j_p\}$$

be the ith of these. Let $b_{J_1,m}, \ldots, b_{J_K,m}$ be the K_n coefficients for the mth predictor variable among the K fits obtained from these disjoint elemental sets. Let

$$v_{ki} = 1/\sqrt{\boldsymbol{A}_{i,kk}}$$

be the inverse of the square root of the kth diagonal element of $\boldsymbol{A}_i = (\boldsymbol{X}_{J_i}^T \boldsymbol{X}_{J_i})^{-1}$.

We make the following two assumptions on the Gaussian regression model.

H1) Assume that $\boldsymbol{A}_i$ is nonsingular for $i = 1, ..., K_n$.

2) Let $q \geq p$. Assume that $[n/q]$ of the v_{ki} satisfy

$$0 < a \leq v_{ki} \leq b.$$

These assumptions are slightly different than those of Hawkins (1993). The proof of the following lemma follows from the proof of Theorem 2.3.

Lemma 2.1 (Hawkins 1993). Under H1) and 2), for any real number c_m,

$$d_m \equiv \min_{i=1,\ldots,K} |b_{J_i,m} - c_m| = O_P(n^{-1}).$$

If all p components of $\boldsymbol{b}_{J_i}$ satisfied the above equation, and if the components were independent, then

$$d_o \equiv \min_{i=1,\ldots,K} \|\boldsymbol{b}_{J_i} - \boldsymbol{c}\| = O_P(n^{-1/p}) \tag{2.1}$$

where the mth component of the $p \times 1$ vector $\boldsymbol{c}$ is c_m. In particular, if $\boldsymbol{c} = \boldsymbol{\beta}$, then the best fit obtained from the disjoint elemental sets may have a very poor rate. Hence the rate for the fit selected by the algorithm would be even worse.

Theorem 2.3 below will show that Equation 2.1 holds even if the vector components are not independent provided that the sizes h_i of the disjoint subsets are bounded. We will choose at least $[n/r]$ nonoverlapping sets of size h_i, $p \leq h_i \leq r$, from the n cases, and we will let

$$J_{i,n} = J_i = \{j_1, ..., j_{h_i}\}$$

be the ith of these. Let

$$\boldsymbol{A}_{i,n} = \boldsymbol{A}_i = (\boldsymbol{X}_{J_i}^T \boldsymbol{X}_{J_i})^{-1},$$

and let

$$\boldsymbol{B}_{i,n} = \boldsymbol{B}_i = \boldsymbol{X}_{J_i}^T \boldsymbol{X}_{J_i}. \tag{2.2}$$

Note that $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are $p \times p$ matrices and that the jth diagonal element $\boldsymbol{B}_{i,jj}$ is bounded if the jth predictor is bounded. If we bound the determinant $det(\boldsymbol{B}_i)$ from below and the largest diagonal element of $\boldsymbol{B}_i$ from above, we will be able to bound $f_{\boldsymbol{b}_{J_i}}(\boldsymbol{x}^T)$ from below when $\boldsymbol{x}$ falls in a bounded closed set.

We add one assumption to the Gaussian regression model.

A1) Let $K_n = [n/q]$ where $q \geq r$. Assume that there is an N such that for $n \geq N$, at least K_n of the $\boldsymbol{X}_{J_i}$ are disjoint and satisfy $0 < a \leq \sqrt{det(\boldsymbol{B}_i)}$, $\max_{k,j} |\boldsymbol{X}_{J_i,kj}| \leq L$, and $p \leq h_i \leq r$.

This assumption says that if $n > N$, then some percentage of the disjoint sets J_i have a determinant $det(\boldsymbol{B}_i)$ that is bounded below by some positive number a^2. So for elemental sets, the condition becomes $0 < a < det(\boldsymbol{X}_{J_i})$. The main purpose of assumption A1) is to bound the density corresponding to the fit $\boldsymbol{b}_{J_i}$ in some neighborhood of a fixed p-vector $\boldsymbol{c}$. If a is a number between 0 and the smallest positive computer number, then the first part of A1) must hold or the estimator can not be computed. In other words, if $det(\boldsymbol{B}_i)$ is too close to zero, then the fit $\boldsymbol{b}_{J_i}$ can not be computed numerically. The second part of A1) implies that some fraction of the cases have predictors that are bounded from above. Since $\boldsymbol{B}_i$ is a symmetric positive definite matrix if $det(\boldsymbol{B}_i) > 0$, the element of $\boldsymbol{B}_i$ with the largest magnitude lies on the diagonal. Moreover, the jth diagonal element of $\boldsymbol{B}_i$ is the sum of h_i squared observations from the jth predictor. Hence the magnitudes of these elements are bounded above by $D = rL^2$ if $\boldsymbol{X}_{J_i}$ satisfies A1).

Lemma 2.2. Suppose $\boldsymbol{X}_{J_i}$ satisfies condition A1). Let $\boldsymbol{c}$ be a $p \times 1$ vector, and let $0 < \delta$. If the $p \times 1$ vector $\boldsymbol{x}$ is contained in a cube centered at $\boldsymbol{c}$ with edge length 2δ, that is, if $x_i \in [c_i - \delta, c_i + \delta]$ for $i = 1, ..., p$, then

$$f_{b_{J_i}}(\boldsymbol{x}^T) \geq \frac{a}{\sigma^p (2\pi)^{p/2}} \exp[-h_\delta D]$$

where $D = rL^2$ and

$$h_\delta \to \frac{p^2}{2\sigma^2} \max_i (c_i - \beta_i)^2$$

as $\delta \to 0$.

Proof. As noted by Hawkins (1993),

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n),$$

and

$$\boldsymbol{Y}_{J_i} \sim N_{h_i}(\boldsymbol{X}_{J_i}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_{h_i}).$$

Hence

$$\boldsymbol{b}_{J_i} = (\boldsymbol{X}_{J_i}^T \boldsymbol{X}_{J_i})^{-1} \boldsymbol{X}_{J_i}^T \boldsymbol{Y}_{J_i} \sim N_p(\boldsymbol{\beta}, \sigma^2 \boldsymbol{A}_i).$$

Thus

$$f_{b_{Ji}}(\boldsymbol{x}^T) = \frac{\sqrt{det(\boldsymbol{B}_i)}}{\sigma^p (2\pi)^{p/2}} \exp[-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\beta})^T \boldsymbol{B}_i (\boldsymbol{x}-\boldsymbol{\beta})]$$

$$= \frac{\sqrt{det(\boldsymbol{B}_i)}}{\sigma^p (2\pi)^{p/2}} \exp[-\frac{1}{2\sigma^2} \sum_{k=1}^{p} \sum_{j=1}^{p} (x_k - \beta_k)(x_j - \beta_j) \boldsymbol{B}_{i,kj}].$$

Since $\boldsymbol{B}_i$ is positive definite and symmetric,

$$|\boldsymbol{B}_{i,kj}| \leq \max(\boldsymbol{B}_{i,kk}, \boldsymbol{B}_{i,jj}) \leq \max_j \boldsymbol{B}_{i,jj}.$$

See Datta (1995, p. 23).

Since $x_k \in [c_k \pm \delta]$,

$$|\frac{1}{2\sigma^2} \sum_{k=1}^{p} \sum_{j=1}^{p} (x_k - \beta_k)(x_j - \beta_j) \boldsymbol{B}_{i,kj}| \leq$$

$$\frac{1}{2\sigma^2} \sum_{k=1}^{p} \sum_{j=1}^{p} \max_{k, x_k \in [c_k \pm \delta]} |x_k - \beta_k| \max_{j, x_j \in [c_j \pm \delta]} |x_j - \beta_j| \max_j \boldsymbol{B}_{i,jj} \leq$$

$$\frac{p^2}{2\sigma^2} [\max_{k, x_k \in [c_k \pm \delta]} |x_k - \beta_k|]^2 D = h_\delta D$$

where $D = rL^2$. Hence

$$\exp[-\frac{1}{2\sigma^2} \sum_{k=1}^{p} \sum_{j=1}^{p} (x_k - \beta_k)(x_j - \beta_j) \boldsymbol{B}_{i,kj}] \geq \exp[-h_\delta D]$$

for $x_k \in [c_k - \delta, c_k + \delta]$ where

$$h_\delta \rightarrow \frac{p^2}{2\sigma^2} \max_k (c_k - \beta_k)^2$$

as $\delta \rightarrow 0$, and

$$f_{b_{Ji}}(\boldsymbol{x}^T) \geq \frac{a}{\sigma^p (2\pi)^{p/2}} \exp[-h_\delta D].$$

QED

Theorem 2.3. Suppose the regression model with iid Gaussian errors holds. If A1) holds and $\boldsymbol{c}$ is a p-dimensional vector, then

$$d_o = \min_{i=1,\dots,K_n} \|\boldsymbol{b}_{J_i} - \boldsymbol{c}\| = O_P(n^{-\frac{1}{p}}). \tag{2.3}$$

Proof. Relabel the $\boldsymbol{X}_{J_i}$ such that the first K_n $\boldsymbol{b}_{J_i}$ satisfy condition A1). If the vector $\boldsymbol{x}$ is contained in a sphere of radius δ centered at $\boldsymbol{c}$, then $\boldsymbol{x}$ is contained in the cube of Lemma 2.2 and

$$f_{b_{Ji}}(\boldsymbol{x}^T) \geq \frac{a}{\sigma^p (2\pi)^{p/2}} \exp[-h_\delta D].$$

The independence of the $\boldsymbol{b}_{J_i}$ implies that

$$P(n^{1/p} d_o > \gamma) = \prod_{i=1}^{K} P(\|\boldsymbol{b}_{J_i} - \boldsymbol{c}\| > \gamma/n^{1/p})$$

$$= \prod_{i=1}^{K} [1 - P(\|\boldsymbol{b}_{J_i} - \boldsymbol{c}\| \leq \gamma/n^{1/p})]$$

$$\leq \prod_{i=1}^{K} [1 - \int_{c_1 - \frac{\gamma}{\sqrt{2} n^{1/p}}}^{c_1 + \frac{\gamma}{\sqrt{2} n^{1/p}}} \cdots \int_{c_p - \frac{\gamma}{\sqrt{2} n^{1/p}}}^{c_p + \frac{\gamma}{\sqrt{2} n^{1/p}}} f_{b_{J_i}}(w_1, \dots, w_p) dw_1 \dots dw_p]$$

since if $\boldsymbol{b}_{J_i}$ is in a sphere centered at $\boldsymbol{c}$ with radius $\gamma/n^{1/p}$, then $\boldsymbol{b}_{J_i}$ is in a cube centered at $\boldsymbol{c}$ with edge length $\sqrt{2}\gamma/n^{1/p}$. For large enough n, Lemma 2.2 can be applied and hence

$$P(n^{1/p} d_o > \gamma) \leq \prod_{i=1}^{K} [1 - \frac{a e^{-h_\delta D}}{\sigma^p (2\pi)^{p/2}} (\frac{\sqrt{2}\gamma}{n^{1/p}})^p]$$

$$= [1 - \frac{\frac{a e^{-h_\delta D}}{\sigma^p (2\pi)^{p/2}} (\sqrt{2}\gamma)^p}{n}]^K = [1 - \frac{\frac{K}{n} \frac{a e^{-h_\delta D}}{\sigma^p (2\pi)^{p/2}} (\sqrt{2}\gamma)^p}{K}]^K$$

$$\to \exp[-\frac{a e^{-h_\delta D}}{q \sigma^p (2\pi)^{p/2}} (\sqrt{2}\gamma)^p]$$

which can be made arbitrarily small by making γ large. QED

The proofs in Hawkins and Olive (2002) are much simpler, but it is useful to have multiple proofs of results and this proof corrects some errors in Hawkins (1993).

3 References

Datta, B.N. (1995), *Numerical Linear Algebra and Applications,* Pacific Grove: Brooks/Cole Publishing Company.

Hawkins, D.M. (1993), "The Accuracy of Elemental Set Approximations for Regression," *Journal of the American Statistical Association,* 88, 580-589.

Hawkins, D.M., and Olive, D.J. (2002), "Inconsistency of Resampling Algorithms for High Breakdown Regression Estimators and a New Algorithm," With Discussion. *Journal of the American Statistical Association,* 97, 136-159.