A Note on Partitioning

David J. Olive *

Southern Illinois University

July 28, 2003

Abstract

The computational complexity of algorithms for robust regression and multivariate location and dispersion often increases exponentially with the number of variables. Many algorithms use K_n trial fits. Partitioning screens out bad trial fits by evaluating the fits on a subset of the data. The best fits are kept and evaluated on the entire data set.

Assume that the data set of $n = hC$ cases contains d outliers, and partition the data set into C disjoint sets of size n/C. It will be shown that each cell contains approximately d/C outliers if d is large and C is fixed.

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

The *multiple linear regression* model is

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

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

A *multivariate location and dispersion model* is a joint distribution

$$f(\boldsymbol{z}) \equiv f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

for a $p \times 1$ random vector $\boldsymbol{x}$ that is completely specified by a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. Hence $P(\boldsymbol{x} \in A) = \int_A f(\boldsymbol{z})d\boldsymbol{z}$ for suitable sets A. The data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are n iid $p \times 1$ random vectors from $f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the ith case is $\boldsymbol{x}_i$.

Elemental sets are subsets just large enough estimate the unknown coefficients. For regression p cases are used to estimate $\boldsymbol{\beta}$ while for multivariate location and dispersion, $p+1$ cases are used to estimate $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In the elemental basic resampling algorithm, K_n elemental sets are randomly selected, producing the estimators $\boldsymbol{S}_{1,n}, ..., \boldsymbol{S}_{K_n,n}$. Then the algorithm estimator $\boldsymbol{S}_{A,n}$ is the elemental fit that minimized the criterion Q.

In a concentration algorithm the half set of cases that have the smallest absolute residuals or Mahalanobis distances from the ith trial fit $\boldsymbol{S}_{i,0,n} \equiv \boldsymbol{S}_{i,n}$ is found. Then an estimator $\boldsymbol{S}_{i,j,n}$ is computed and the process is repeated for k_i steps. Often $k_i = 10$ for all i or the iteration is performed until convergence. The estimator $\boldsymbol{S}_{i,k_i,n}$ is called the

ith attractor of the ith start $\boldsymbol{S}_{i,0,n}$. Then the algorithm estimator $\boldsymbol{S}_{A,n}$ is the attractor that minimized the criterion Q.

In a partitioning algorithm, C subsets J_i of size h cases are randomly selected. Then D elemental subsets are drawn from each subset J_i, and concentration and evaluation of the fit uses only the h cases in the subset. Of the $CD = K$ subsets, the M fits with the smallest criterion values are retained, and then these fits are used as starts on the entire data set.

Woodruff and Rocke (1994) introduced partitioning for robust algorithms, and the partitioning step is often much faster than evaluating K elemental sets on all n cases. Rousseeuw and Van Driessen (1999ab) implement the partitioning step in their concentration step. The basic idea is that sampling theory suggests that if h is large enough, then fits that have small criterion values evaluated on the h cases should also have small criterion values when evaluated on all n cases. Hence partitioning is useful for eliminating bad fits.

Suppose that the data set has n cases and that d of these cases are outliers. If the data is randomly assigned to $C = 2$ groups of equal size, then sampling theory suggests that both subgroups will be similar to the full data set; however, the group size is half the sample size, and one group will usually have a smaller proportion of outliers than the other. The following section uses results from multinomial theory to estimate the proportion of outliers in the subset that contains the fewest outliers.

2 Outliers and Partitioning

We will partition the data into C cells each of size n/C. Suppose the total number of outliers in the data set is d. Then the expected number of outliers in any cell is d/C. We will show that the cell with the smallest number of outliers still has about

$$\frac{d}{C} - k\sqrt{\frac{d}{C}} \approx \frac{d}{C}$$

outliers when d is large and C is fixed. Hence if d is large compared to C, then even the cleanest of the C partitions has a level of contamination broadly commensurate with that of the full sample.

First we give some notation. Suppose d of the n cases are contaminated. Then the proportion of contaminated cases is

$$\gamma = \frac{d}{n}.$$

If d identical balls are placed randomly into C urns, and if d_i denotes the number of balls in the ith urn, then the joint distribution of $(d_1, ..., d_C)$ is multinomial$(d, 1/C, ..., 1/C)$. Since we are constraining each cell to have n/C cases, the distribution of the C cells will not be multinomial, but a multinomial approximation may be good if

$$C < \frac{n(1-\gamma)^2}{16\gamma}$$

or

$$7C < n.$$

Johnson and Young (1960) argue that the joint distribution

$$\frac{1}{\sqrt{\frac{d}{C}\frac{C-1}{C}}}(d_1 - \frac{d}{C}, ..., d_C - \frac{d}{C})$$

$$\approx \sqrt{\frac{C}{C-1}}(Z_1 - \bar{Z}_C, ..., Z_C - \bar{Z}_C)$$

where $Z_1, ..., Z_C$ are iid standard normal. Thus the largest number of outliers in a cell

$$d_{(C)}$$

and

$$\frac{d}{C} + \sqrt{\frac{d}{C}}(Z_{(C)} - \bar{Z}_C) \overset{D}{=} \frac{d}{C} + \sqrt{\frac{d}{C}}(\bar{Z}_C - Z_{(1)})$$

have approximately the same distribution. One approximation for the upper 100α percentage point of $d_{(C)}$ from a symmetric multinomial distribution is

$$\frac{d}{C} + \frac{d}{C}\sqrt{\frac{C-1}{d}}\Phi^{-1}(1 - \frac{\alpha}{C}) \tag{2.1}$$

where Φ is the standard normal cdf. See equation 5 of Johnson and Young (1960) combined with equation 23 of Nair (1948), David (1981, p. 113), and Kozelka (1956). For the exact distribution and other approximations, see Freeman (1979). Hence the upper $100(1-\alpha)$ percentage point of $d_{(1)}$, the fewest number of outliers in a cell, is approximately

$$[\max(\frac{d}{C} - \frac{d}{C}\sqrt{\frac{C-1}{d}}\Phi^{-1}(1 - \frac{\alpha}{C}), 0)]. \tag{2.2}$$

From Johnson and Young (1960) and Kozelka (1956), the approximation should be useful for $\alpha = 0.05$ or $\alpha = 0.01$ and for

$$C \leq \min(15, \frac{n}{7}).$$

Note that if $\alpha = 0.05$, then Equation 2.2 is equal to 0 when

$$n \leq \frac{(C-1)[\Phi^{-1}(1 - \frac{0.05}{C})]^2}{\gamma}.$$

A small simulation of 1000 partitions was performed. The 0.05 percentile and the 0.01 percentile of $d_{(1)}$ were close for each value of C, n, and γ used in the simulation. Table 1 compares (2.2) with the observed 0.05 percentile of $d_{(1)}$ when 1000 partitions were generated. Although the approximation (2.2) had small error, replacing α by $\alpha/5$ in (2.2) gave better empirical results.

For algorithm design, note that if we partition the data into C cells M times where $1/M = \alpha$, we might find one cell with a contamination proportion as low as

$$\gamma - \sqrt{\gamma}\sqrt{\frac{C-1}{n}}\Phi^{-1}(1 - \frac{\alpha}{C}). \tag{2.3}$$

The above approximations are used when the number of cells C is small. When C is large, the probability that j cells are clean has an approximate Poisson(λ) distribution with

$$\lambda = C \ \exp(\frac{-d}{C}).$$

See Feller (1957, p. 92-94). Hence

$$1 - \exp(-C \ \exp(\frac{-n}{2C})) \leq 1 - \exp(-C \ \exp(\frac{-d}{C})) \approx P(d_{(1)} = 0).$$

With d outliers and C cells, we expect about

$$C(1 - \frac{1}{C})^d$$

of the cells to be clean. See Feller (1957, p. 226).

Table 1: Observed 0.05 Percentile for $d_{(1)}$ vs (10.2)

						C		
n	γ	d	4	4	6	6	12	12
			obs	(2.2)	obs	(2.2)	obs	(2.2)
24	.042	1	0	0	0	0	0	0
24	.125	3	0	0	0	0	0	0
24	.25	6	0	0	0	0	0	0
24	.5	12	1	0	0	0	0	0
48	.042	2	0	0	0	0	0	0
48	.125	6	0	0	0	0	0	0
48	.25	12	0	0	0	0	0	0
48	.5	24	3	1	1	0	0	0
96	.042	4	0	0	0	0	0	0
96	.125	12	0	0	0	0	0	0
96	.25	24	2	1	1	0	0	0
96	.5	48	7	5	4	1	1	0

Table continued

			C					
n	γ	d	4	4	6	6	12	12
			obs	(2.2)	obs	(2.2)	obs	(2.2)
240	.042	10	0	0	0	0	0	0
240	.125	30	3	2	1	0	0	0
240	.25	60	9	7	4	3	1	0
240	.5	120	22	19	13	10	5	2
480	.042	20	1	0	0	0	0	0
480	.125	60	8	7	4	3	1	0
480	.25	120	21	19	12	10	4	2
480	.5	240	49	44	30	26	12	8
960	.042	40	4	3	2	1	0	0
960	.125	120	21	19	11	10	3	2
960	.25	240	47	44	28	26	11	8
960	.5	480	105	98	66	60	28	24

3 References

David, H.A. (1981), *Order Statistics,* 2nd ed., John Wiley and Sons, Inc., NY.

Feller, W. (1957), *An Introduction to Probability Theory and Its Applications,* Vol. 1, 2nd ed., John Wiley and Sons, Inc., NY.

Freeman, P.R. (1979), "Exact Distribution of the Largest Multinomial Frequency," *Applied Statistics,* 28, 333-336.

Johnson, N.L., and Young, D.H. (1960), "Some Applications of Two Approximations to the Multinomial Distribution," *Biometrika,* 47, 463-468.

Kozelka, R.M. (1956), "Approximate Upper Percentage Points for Extreme Values in Multinomial Sampling," *The Annals of Mathematical Statistics,* 27, 507-512.

Nair, K.R. (1948), "The Distribution of the Extreme Deviate from the Sample Mean and Its Studentized Form," *Biometrika,* 35, 118-144.

Rocke, D.M., and Woodruff, D.L. (1996), "Identification of Outliers in Multivariate Data," *Journal of the American Statistical Association,* 91, 1047-1061.

Rousseeuw, P.J., and Van Driessen, K. (1999a), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Rousseeuw, P.J., and Van Driessen, K. (1999b), "Computing LTS Regression for Large Data Sets," Technical report, University of Antwerp.

Woodruff, D.L., and Rocke, D.M. (1994), "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators," *Journal of the American Statistical Association,* 89, 888-896.