The Breakdown of Breakdown

David J. Olive and Douglas M. Hawkins *

Southern Illinois University and University of Minnesota

June 12, 2008

Abstract

High breakdown multivariate location and dispersion, high breakdown regression, and outlier diagnostics have generated an enormous literature. High breakdown estimators are often computed with sub-sampling algorithms, and changes to these algorithm estimators are needed to produce estimators that are useful for both outlier detection and inference. For regression, the response and residual plots are useful for outlier detection and for visualizing the model.

KEY WORDS: least trimmed sum of squares estimator, minimum covariance determinant estimator, outliers, robust regression.

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. Douglas M. Hawkins is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455-0493, USA. Their work was supported by the National Science Foundation under grants DMS 0600933, DMS 0306304, DMS 9803622 and ACI 9619020.

1 Introduction

The *multiple linear regression (MLR) model* is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients 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 $\boldsymbol{Y}$.

A *multivariate location and dispersion (MLD) model* is a joint distribution 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 $\boldsymbol{\Sigma}$. The multivariate normal distribution is an important MLD model.

High breakdown (HB) estimators of MLR and of MLD are longstanding objects of statistical research. These publications tend to concentrate on the theoretical properties of the estimators they propose, and to downplay their computational aspects. The two are, however, inextricably connected. If an estimator can not be computed in a tolerable amount of time, then most of its theoretical properties are of only academic interest. What is of interest is the properties of the estimator as it will actually be computed, and these properties are something entirely different.

To illustrate the problem, consider an estimator whose computational complexity is $O(n^p)$, where n is the number of cases and p is the dimensionality. This complexity (which applies for example to the least median of squares LMS criterion) is very modest in the field of HB estimation; many criteria have complexity $O(n^{n/2})$, but is enough to highlight the problem. Suppose n and p are 200 and 10 respectively so that the data set is small by modern standards. Then $n^p \approx 10^{23}$. A computer that could analyze one

candidate solution per microsecond would take 5 billion years to evaluate the theoretical estimator. This means HB estimators such as LMS, least trimmed sum of squares (LTS), minimum covariance determinant (MCD), minimum volume ellipsoid (MVE), repeated median, S-estimators, Stahel-Donoho estimators and many depth estimators are far too slow to be practical as defined theoretically. Two stage estimators that use an initial estimator from the above list are also impractical, including the cross checking, MM, one step GM, one step GR, τ and t-type estimators. See Maronna, Martin and Yohai (2006) for details on most of these estimators.

Section 2 reviews the theory for the sub-sampling algorithm estimators actually used. Section 3 shows that computing HB MLR estimators is simple, and Section 4 gives a graphical method for detecting MLR outliers.

2 Review of Theory for HB Algorithm Estimators

The actual estimators do not implement the theoretical criterion, but use some approximation. The workhorse of most estimators is the "basic resampling" or "elemental set" method where random subsets of size p (for MLR but $p+1$ for MLD) are drawn to get trial estimators. In the "pure" basic resampling scheme, the final estimator used is the best of K such trial solutions. Since each trial estimator is inconsistent, so is the final estimator if K is fixed.

More recent methods tend to use these initial elemental estimators as a starting point for refinement. Some perform a fixed number of refinement steps; others refine to convergence. The final estimator is then the best of the "attractors" uncovered from

these K starts.

None of these approaches is guaranteed to give a final solution that is close to the theoretical solution whose good properties are shown, and theoretical analyses of the procedure as actually implemented, rather than as conceptually defined, are rare.

The inadequacy of the original PROGRESS algorithm with $K \leq 3000$ and no refinement has long been familiar. Concentration estimators like FAST-LTS use k refinement steps of an initial elemental start to identify "attractors" that are fits to $\approx n/2$ of the cases. By direct application of results in He and Portnoy (1992) and Lopuhaä (1999), Hawkins and Olive (2002) showed that the best attractor has no better convergence than its starting point, and so is not consistent if K and k are fixed and free of n. Hence no matter how the attractor is chosen, the resulting estimator is not consistent. This is equally true of the MLR or MLD problems.

More specifically, He and Portnoy (1992) and Lopuhaä (1999) show that if, for MLR and MLD respectively a start $\boldsymbol{b}$ or $(T, \boldsymbol{C})$ is a consistent estimator of $\boldsymbol{\beta}$ or $(\boldsymbol{\mu}, s\boldsymbol{\Sigma})$, then the attractor is a consistent estimator of $\boldsymbol{\beta}$ or $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where $a, s > 0$ are some constants. The start and the attractor have the same convergence rate; if the start is inconsistent, then so is the attractor. The classical estimator applied to a randomly drawn elemental set is an inconsistent estimator, so the K starts and the K attractors are inconsistent. The final estimator is an attractor and thus inconsistent.

Iterating some starts to convergence so that k is not fixed, as suggested by Hubert, Rousseeuw and Van Aelst (2008), produces inconsistent estimators if the attractor of a randomly drawn elemental start is inconsistent.

Another popular algorithm uses K randomly chosen directions to define weights for

a weighted MLD estimator. If each of the K sets of weights results in an inconsistent estimator, then the final estimator is inconsistent. Such approximations are completely different estimators than the impractical estimators that have theory, such as the Stahel-Donoho estimator.

The central thesis of Hawkins and Olive (2002) was that, given the disconnect between the theoretically defined estimator and what can actually be computed, the theoretical properties of the former do not necessarily give useful guidance on the properties of the latter. Nearly all of the literature appears to overlook this disconnect, including Agulló, Croux and Van Aelst (2008), Berrendero, Mendes and Tyler (2007), Hubert, Rousseeuw and Van Aelst (2008), Maronna, Martin and Yohai (2006) and Zuo, Cui and He (2004).

3 Computing HB MLR Estimators

Turning to a somewhat different topic, consider breakdown in the MLR problem (similar remarks apply to MLD). If d of the cases have been replaced by arbitrarily bad contaminated cases, then the contamination fraction is $\gamma = d/n$. Then the breakdown value of the MLR $\hat{\boldsymbol{\beta}}$ is the smallest value of γ needed to make $\|\hat{\boldsymbol{\beta}}\|$ arbitrarily large. Olive (2005) showed that an MLR estimator is high breakdown if the median absolute or squared residual stays bounded under high contamination: if $\|\hat{\boldsymbol{\beta}}\| = \infty$, then $med(|r_i|) = \infty$, and if $\|\hat{\boldsymbol{\beta}}\| = M$ then $med(|r_i|)$ is bounded if fewer than half of the cases are outliers.

The folklore says that HB estimators are good: they make outliers have large absolute residuals, and that an estimator should be judged by its breakdown value and Gaussian efficiency; however, the property of being a high breakdown estimator is weaker than the

property of being an asymptotically unbiased estimator.

To see that the folklore is false, consider the LMS estimator. In a plot of $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{LMS}$ vs Y_i, the LMS(c_n) estimator is determined by the "narrowest strip" covering $c_n \approx n/2$ cases. Picture a "planet and moon" data set in which the majority of the data constitute a spheroid, indicating that in the majority of the cases there is no relationship between $\boldsymbol{x}$ and Y, while a minority of the cases form a compact "moon", located some distance from the planet. Then as long as the moon remains within some cone, the narrowest strip is likely to consist of all the outliers plus enough planetary points to make up the half-sample. Suppose for example that a large data set comprises a single predictor x and response Y, both being independent N(0,1). Now replace 40% of the cases with outlier values clustered tightly about $(x, Y) = (k, bk)$. It can be shown that the LMS line will accommodate the outliers provided $|b|$ is no larger than about 5. Only for larger values of $|b|$ will the outliers be excluded from coverage, so cases located at say $(100, 500)$ could lie within the covered half-sample and have near-zero residuals. As for the fitted line, the maximum absolute value of the slope of the LMS line is no more than about 5. As $5 < \infty$, the breakdown property is indeed respected, but this may be small comfort to a user who assumed that the fitted LMS slope would not be very far from the true value of 0 that describes the inliers.

Breakdown involves the estimate of $\boldsymbol{\beta}$ tending to infinity or remaining bounded as data values are replaced by arbitrary values. A very simple procedure gives high breakdown, along with $\sqrt{n}$ asymptotics at clean data. Suppose the MLR model has an intercept β_1. Let $\boldsymbol{b} = (med(Y_i), 0, ..., 0)^T$. Define the estimator $\hat{\boldsymbol{\beta}}_{HB}$ to equal the least squares estimator $\hat{\boldsymbol{\beta}}_{OLS}$ if $med(r_i^2(\hat{\boldsymbol{\beta}}_{OLS})) \leq med(r_i^2(\boldsymbol{b}))$ and equal $\boldsymbol{b}$ otherwise. Then $\hat{\boldsymbol{\beta}}_{HB}$ is HB

since its median squared residual is less than $(\text{MAD}(Y_i))^2$, the squared median absolute deviation. Consistent estimators such as OLS and LMS have a median squared residual that converges to $med(e^2)$ where the random variable e has the same distribution as the iid errors e_i. Hence the criterion for the OLS estimator gets arbitrarily close to the criterion for the LMS estimator, and if at least one of the slopes β_i is nonzero, then the probability that $med(r_i^2(\hat{\boldsymbol{\beta}}_{OLS})) \leq med(r_i^2(\boldsymbol{b}))$ goes to one. Hence $\hat{\boldsymbol{\beta}}_{HB}$ is high breakdown with 100% asymptotic Gaussian efficiency since it is asymptotically equivalent to OLS. This estimator is greatly superior to LMS because it is practical to compute and $\sqrt{n}$ consistent, but its outlier resistance is similar to that of OLS.

4 Graphical Detection of MLR Outliers

One application of the MLR algorithm estimators is outlier detection. Outlier diagnostics such as Cook's distances CD_i from Cook (1977) and the weighted Cook's distances WCD_i from Peña (2005) are also sometimes useful. For detection of outliers and influential cases, it is crucial to make the residual plot of $\hat{Y}$ vs r and the response plot of $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ vs Y with the identity line with zero intercept and unit slope added as a visual aid. Vertical deviations from the identity line are the residuals $r_i = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$.

Olive and Hawkins (2005) also showed that the two plots are crucial for visualizing the MLR model and for examining lack of fit. If $n > 10p$ and if the plotted points scatter about the identity line and the $r = 0$ line in an evenly populated band, then the MLR model with iid e_i where $\text{VAR}(e_i) = \sigma^2$ may be reasonable. Deviations from the evenly populated band suggest that something is wrong with the MLR model, and

are often easily detected even if OLS is used. In the following examples, cases in the plots with $CD_i > \min(0.5, 2p/n)$ are highlighted with squares, and cases with $|WCD_i - med(WCD_i)| > 4.5MAD(WCD_i)$ are highlighted with an open triangle in `R` and a cross in `Splus`.

Example 1. Buxton (1920, p. 232-5) gives 20 measurements of 88 men. Consider predicting *stature* using an intercept, *head length, nasal height, bigonal breadth*, and *cephalic index*. One case was deleted since it had missing values. Five individuals, numbers 61-65, were reported to be about 0.75 inches tall with head lengths well over five feet! Figure 1 shows the OLS response plot and residual plot (made in `R`) for the Buxton data. Notice that the OLS fit passes through the outliers, but the response plot is resistant to Y–outliers since Y is on the vertical axis. Also notice that although the outlying cluster is far from $\overline{Y}$ only two of the outliers had large Cook's distance and only case 62 had a large WCD_i. Influence diagnostics are the most effective when there is a single cluster about the identity line.

Example 2. Wood (1973) provides octane data where the octane number is predicted from 3 feed compositions and the log of a combination of process conditions. Figure 1 shows the OLS response plot and residual plot (made in `Splus`). Although none of the cases had large influence diagnostics, the two plots suggest that the iid error MLR model with constant variance is not appropriate for this data. There seems to be three groups of data and the largest group has a left opening megaphone shape.

Tremendous profit can be gained by raising the octane number by one point. To use the response plot to visualize the MLR model, mentally examine the response plot for a

narrow vertical strip about fit $= w$. The cases in the strip have octane numbers near w on average. The two cases with the largest fitted values were of the greatest interest.

Using the response and residual plots, 26 benchmark data sets were examined for large CD_i and WCD_i. Cook's distances were useful for 8 of the data sets while Peña's distances were useful for 4 data sets; however, the numerical diagnostics did not provide much information that could not be seen in the response and residual plots. The plots often showed two or more groups of data, and for several data sets the outliers caused an obvious tilt in the residual plot.

5 Conclusions

Many papers, for example the discussion of asymptotics of LMS and MCD in Kim and Pollard (1990) and Butler, Davies and Jhun (1993), prove impressive large sample theory. The practical usefulness of these theoretical properties however is clouded by the fact that these estimators are unimplementable in moderate to large samples. The fact that they are consistent and HB was used by Olive in 2004 to create MLR and MLD estimators that are HB, $\sqrt{n}$ consistent, easy to compute, but with much more outlier resistance than $\hat{\boldsymbol{\beta}}_{HB}$.

There is an important but unfilled need for fuller theoretical analysis of actual implementable HBEs to supplement the current results that may be mathematically elegant, but do not address the questions of practical performance. For MLR, the response and residual plots should be made to check whether the MLR model is reasonable.

6 References

Agulló, J., Croux, C., Van Aelst, S. (2008), "The Multivariate Least-Trimmed Squares Estimator," *Journal of Multivariate Analysis*, 99, 311-338.

Berrendero, J. R., Mendes, B. V. M., and Tyler, D. E. (2007), "On the Maximum Bias Functions of MM-estimates and Constrained M-estimates of Regression," *The Annals of Statistics,* 34, 13-40.

Butler, R. W., Davies, P. L., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics,* 21, 1385-1400.

Buxton, L. H. D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Cook, R. D. (1977), "Deletion of Influential Observations in Linear Regression," *Technometrics,* 19, 15-18.

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.

He, X., and Portnoy, S. (1992), "Reweighted LS Estimators Converge at the Same Rate as the Initial Estimator," *The Annals of Statistics,* 20, 2161-2167.

Hubert, M., Rousseeuw, P. J., and Van Aelst, S. (2008), "High Breakdown Multivariate Methods," *Statistical Science*, to appear.

Kim, J., and Pollard, D. (1990), "Cube Root Asymptotics," *The Annals of Statistics,* 18, 191-219.

Lopuhaä, H. P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location

and Scatter," *The Annals of Statistics,* 27, 1638-1665.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, John Wiley and Sons, Hoboken, NJ.

Olive, D. J. (2005), "Two Simple Resistant Regression Estimators," *Computational Statistics and Data Analysis*, 49, 809-819.

Olive, D. J., and Hawkins, D. M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Peña, D. (2005), "A New Statistic for Influence in Regression," *Technometrics*, 47, 1-12.

Wood, F. S. (1973), "The Use of Individual Effects and Residuals in Fitting Equations to Data," *Technometrics*, 15, 677-695.

Zuo, Y., Cui, H., and He, X. (2004), "On the Stahel-Donoho Estimator and Depth-Weighted Means of Multivariate Data," *The Annals of Statistics,* 32, 167-188.

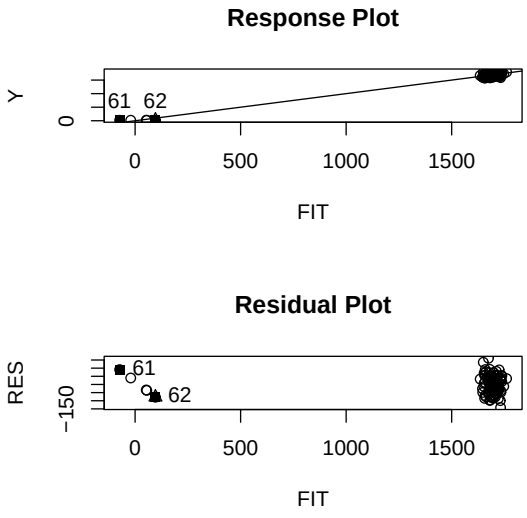

Figure 1: Buxton Data

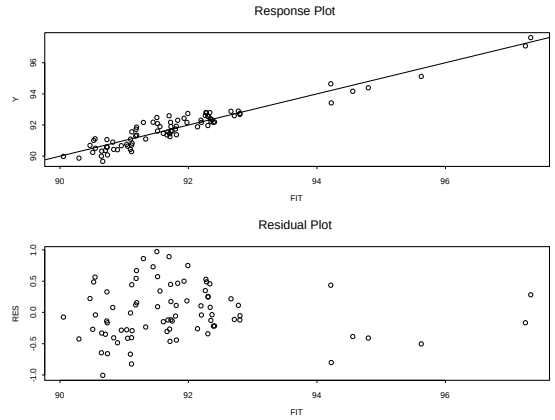

Figure 2: Octane Data