Computing the exact value of the least median of squares estimate in multiple linear regression

by Arnold J. Stromberg

University of Minnesota
Technical Report #561
May, 1991

COMPUTING THE EXACT VALUE OF THE LEAST MEDIAN OF SQUARES ESTIMATE IN MULTIPLE LINEAR REGRESSION

Arnold J. Stromberg[1]

ABSTRACT

A method for computing the least median of squares estimator ($\hat{\theta}_{\text{LMS}}$) in multiple linear regression that requires considering $\binom{n}{p+1}$ possible θ values is presented. It is based on the fact that $\hat{\theta}_{\text{LMS}}$ is the Chebyshev (or minimax) fit to half of the data. This yields a surprising easy algorithm for computing the exact LMS estimate. Several examples show how approximate algorithms can yield very different conclusions from the exact least median of squares solution.

Keywords: Least median of squares estimator, Multiple linear regression, Chebyshev, Minimax.

---

[1]Arnold J. Stromberg is Visiting Assistant Professor, Department of Applied Statistics, University of Minnesota, St. Paul, MN 55108. The author thanks Douglas Hawkins, Jeff Simonoff and David Ruppert for many helpful comments.

1. INTRODUCTION

Linear regression treats the problem of estimating θ_o where:

$$y_i = x_i\theta_o + \epsilon_i \quad i = 1, 2, \ldots, n$$

where $(x_i, y_i) \in (R^p, R)$ are data points and θ_o is an unknown p-dimensional parameter vector and the ϵ_i are unknown errors. We will denote estimators of θ_o by $\hat{\theta}$. The residuals, $y_i - x_i\theta$, $i = 1, 2, \ldots, n$, are denoted $r_i(\theta)$. The best known estimator of θ_o is the least squares estimator $\hat{\theta}_{LS}$ which is:

$$\underset{\theta}{\text{Argmin}} \sum_{i=1}^{n} r_i^2(\theta)$$

The least squares estimator, although optimal in many situations, has the drawback that it is heavily influenced by outliers. It also suffers from the problem of masking, that is, it is possible that multiple outliers may be present in the data set, yet they are not detected by common least squares diagnostic procedures.

The breakdown point of an estimator (Donoho and Huber, 1983) has been shown to be a useful measure of the robustness of an estimator. It can be thought of as the least amount of arbitrary contamination that can drive the estimate to infinity. It is clear that the breakdown point of the least squares estimate in linear regression is 1/n. Recent research (Atkinson 1986; Rousseeuw and von Zomeren 1990) has shown the usefulness of estimators with breakdown point approximately equal to 1/2. These estimators seem to be able to detect masking when least squares diagnostic procedures do not. The most studied high breakdown estimator is Rousseeuw's (1984) least median of squares (LMS) estimator. It is denoted $\hat{\theta}_{LMS}$ and defined as:

$$\underset{\theta}{\text{Argmin}} \underset{1 \le i \le n}{\text{Median}} r_i^2(\theta)$$

In order to obtain the highest possible breakdown point for $\hat{\theta}_{LMS}$ when the data are in general position, meaning that any p points give a unique determination of θ, the median is defined as the qth order statistic where $q = [n/2] + [(p+1)/2]$ and $[\cdot]$ indicates the greatest integer function.

One of the drawbacks of the least median of squares estimate is that it is quite difficult to compute. The objective function is continuous, but not differentiable and it has many local minima. Rousseeuw and Leroy's (1987) PROGRESS algorithm is the most widely used algorithm for estimating $\hat{\theta}_{LMS}$ in linear regression. For a given data set and regression function, the PROGRESS algorithm computes the exact fit, $\hat{\theta}_{ef}$, to many randomly chosen p point elemental subsets of the data set.

Denote the $\hat{\theta}_{ef}$ with the smallest median squared residual $\hat{\theta}$. If the regression function has no intercept, $\hat{\theta}$ is the PROGRESS estimate of $\hat{\theta}_{LMS}$. If an intercept is used in the model, the intercept of $\hat{\theta}$ is adjusted to yield the smallest possible median residual. This adjusted $\hat{\theta}$ is then the PROGRESS estimate of $\hat{\theta}_{LMS}$. A flow chart for the algorithm is presented in Figure 1.1. Rousseeuw and Leroy (1987) note that at the expense of additional computation time, the intercept adjustment can be done for each elemental set. Unfortunately this algorithm, which Steele and Steiger (1986) show will find the exact value of $\hat{\theta}_{LMS}$ when p=2, does not yield the exact LMS estimate in multiple linear regression where p>2.

The MVELMS algorithm of Hawkins and Simonoff (1991), which is also based on the selection of p point elemental sets but uses an intercept adjustment for all elemental sets, has been proposed as an alternative to the PROGRESS algorithm. In general, it produces estimates of θ_o with a smaller median squared residual than the PROGRESS algorithm.

Using a geometric argument, Tichavsky (1991) has argued that the exact LMS estimate in multiple linear regression can be found by considering all p+1 point elemental sets and for each one finding the values of θ where the magnitudes but not the signs of all p+1 residuals are equal. This method leads to $(2^p - 1)\binom{n}{p+1}$ values of θ that must be considered in order to compute the exact value of $\hat{\theta}_{LMS}$ in multiple linear regression. Given the complexity of the problem for moderately large n and p, Tichavsky suggests approximating $\hat{\theta}_{LMS}$ by selecting p point elemental sets and checking the median squared residual for the values of θ generated by the selected elemental sets.

A method for computing $\hat{\theta}_{LMS}$ in multiple linear regression that requires considering $\binom{n}{p+1}$ possible θ values is presented in this paper. Since $\hat{\theta}_{LMS}$ minimizes the q^{th} largest squared residual for a given data set, it must minimize the maximum squared residual for some q element subset of the data. Thus, $\hat{\theta}_{LMS}$ is the Chebyshev (or minimax) fit to that q element subset. Section 2 presents two theorems that can be used to find the Chebyshev fit for a given data set. The first implies that the LMS fit must be the Chebyshev fit to some p+1 element subset of the data, and the second provides a surprising easy method for computing the Chebyshev fit to p+1 points. Thus, the theorems can be used develop an algorithm for computing the exact value of $\hat{\theta}_{LMS}$ in multiple linear regression. Section 3 presents two examples showing how approximate algorithms can yield very different conclusions from the exact least median of squares solution.

2. THE CHEBYSHEV FIT

In this section we adapt theorems found in Cheney (1962) to provide a method for computing the Chebyshev fit, and thus the LMS fit, in linear regression. The first relevant theorem can be restated in the context of regression as follows:

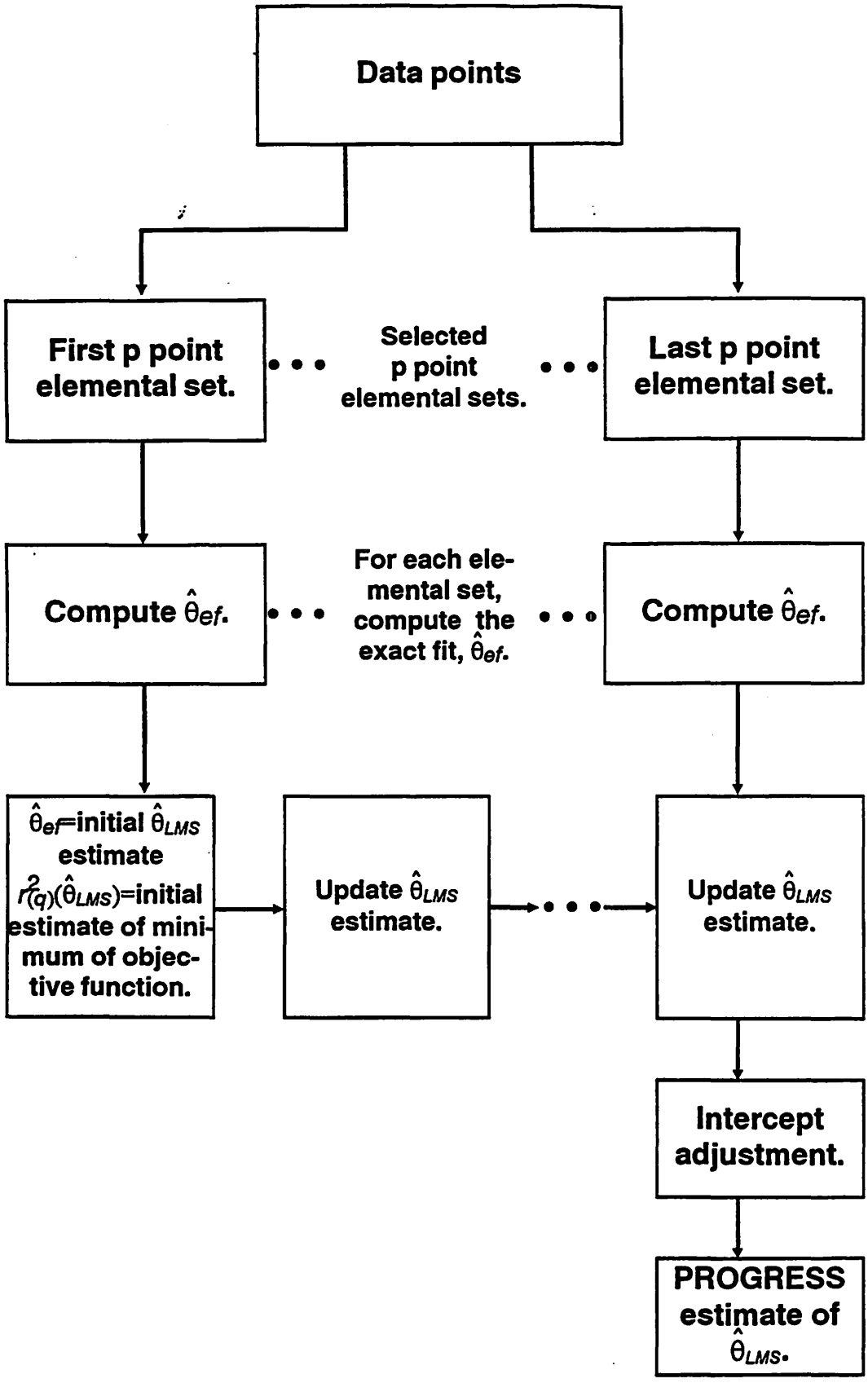

Figure 1.1 PROGRESS Algorithm for Estimating $\hat{\theta}_{LMS}$.

Theorem 1 (Cheney, 1962, p. 36)

In linear regression, the Chebyshev fit, $\hat{\theta}_c$, will be the Chebyshev fit to some p+1 element subset of the data. □

By Theorem 1, if we can find $\hat{\theta}_c$ for all p+1 element subsets of the data, then we can find $\hat{\theta}_c$ for the entire data set. Theorem 2 provides a method for finding $\hat{\theta}_c$ when the sample size is p+1. The Haar condition says that there is one and only one exact fit to any p points. Let $Y = (y_1, y_2, \ldots, y_n)^T$ and X be the $n \times p$ matrix given by $(x_1, x_2, \ldots, x_n)^T$.

Theorem 2 (Cheney, 1962, p. 41)

Consider the linear regression setting described in Section 1 with sample size p+1. Assume that the Haar condition is satisfied. Then

$$\hat{\theta}_{LS} = MY, \text{ where } M = (X^T X)^{-1} X^T$$

is the least squares fit to the data. Let

$$\epsilon = \frac{\sum_{i=1}^{p+1} r_i^2(\hat{\theta}_{LS})}{\sum_{i=1}^{p+1} \left| r_i(\hat{\theta}_{LS}) \right|}$$

and S be the p+1 dimensional vector where $s_i = \text{sgn}(r_i(\hat{\theta}_{LS}))$, i = 1, 2, ..., p+1. Then

$$\hat{\theta}_c = M(Y - \epsilon S). \qquad \square$$

Remarks: Since the LMS fit is the Chebyshev fit for some sample of p+1 points, the following algorithm (Figure 2.1) can be used for computing the exact value of $\hat{\theta}_{LMS}$ in multiple linear regression: For each p+1 point elemental set, use Theorem 2 to compute the Chebyshev fit, denoted $\hat{\theta}_c$. The $\hat{\theta}_c$ with the smallest median squared residual will be the exact LMS estimate. As with the algorithms of Section 1, implementations should take advantage of the fact that for many $\hat{\theta}_c$, computing all the squared residuals and/or the sort to find the median residual can be avoided. Suppose $\hat{\theta}$ is the current best estimate of $\hat{\theta}_{LMS}$ and $\hat{\theta}_c$ is the Chebyshev fit to the p+1 point elemental set being considered. The squared residuals at $\hat{\theta}_c$ need only be computed until $n - q$ are more than $r^2_{(q)}(\hat{\theta})$ because then it must be that $r^2_{(q)}(\hat{\theta_c}) > r^2_{(q)}(\hat{\theta})$. Should this not be the case, $\hat{\theta}_c$ becomes the new estimate of $\hat{\theta}_{LMS}$ and the squared residuals are sorted to find the qth largest squared residual at the new estimate of $\hat{\theta}_{LMS}$.

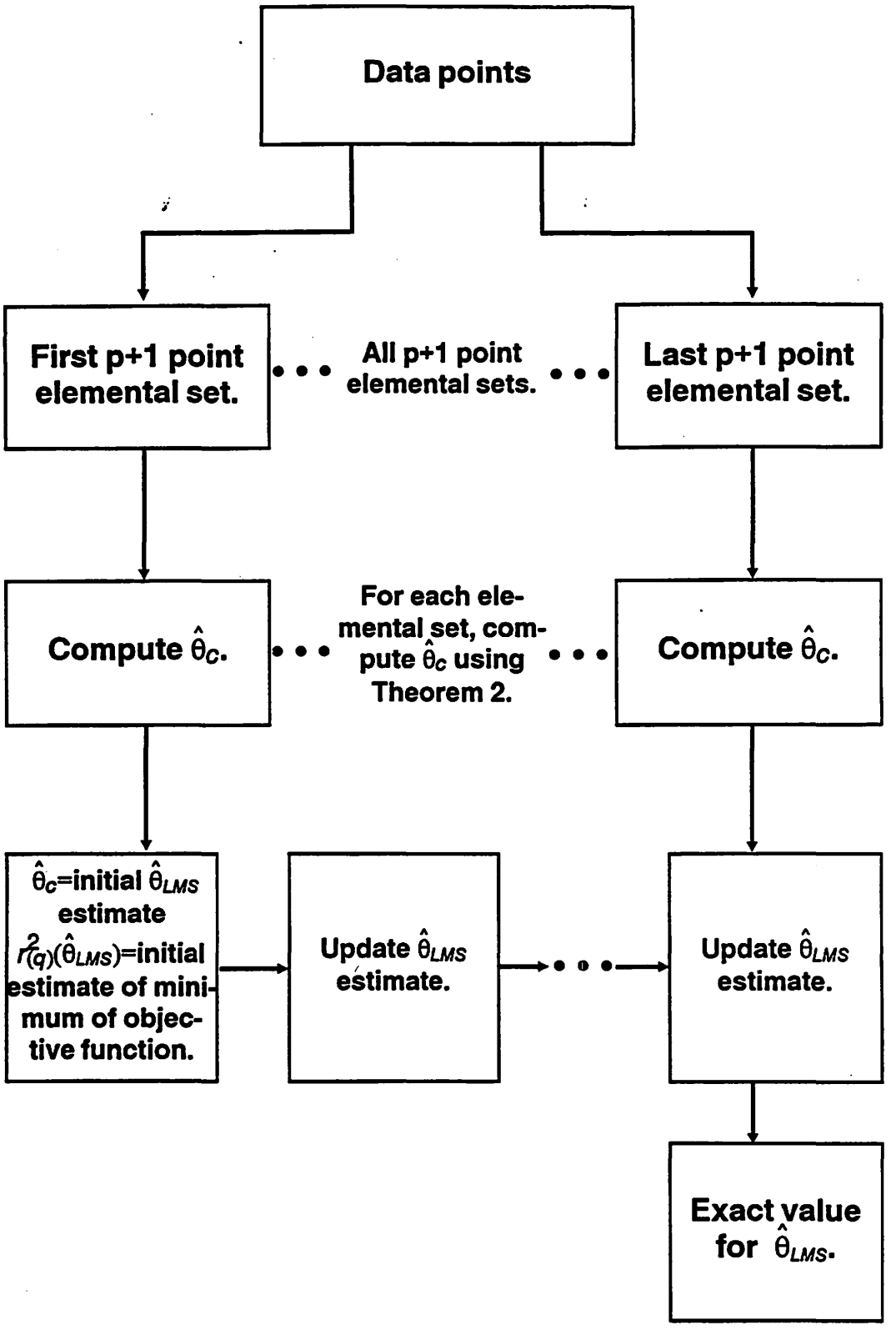

Figure 2.1 Exact Algorithm for Computing $\hat{\theta}_{LMS}$.

The fact that $\hat{\theta}_{LMS}$ is the Chebyshev fit to some p+1 point elemental set seems intuitive, but it is quite surprising that the computation of $\hat{\theta}_c$ provided by Theorem 2 is only moderately more computationally difficult than computing the exact fit, $\hat{\theta}_{ef}$ to p points as is done in the approximate algorithms of Section 1. This suggest that algorithms based on computing $\hat{\theta}_{ef}$ could be improved by computing $\hat{\theta}_c$ instead of $\hat{\theta}_{ef}$.

If the Chebyshev fits are unique, then the LMS fit will have p+1 points with squared residuals equal to the median squared residual, $q-p-1$ points with squared residuals less than the median squared residual, and $n-q$ points with squared residuals more than the median squared residual.

The exact algorithm can easily be modified to compute what can be called the percentile estimates. The $(k/n)\times 100\%$ percentile estimate for k = p+1, p+2, ..., n minimizes the k^{th} largest residual for the data set. We will denote the estimate $\hat{\theta}_{(k)}$. Note that the LMS estimate is a percentile estimate with k = q. Cook and Hawkins (1990) argue the usefulness of percentile estimates. Since each of the percentile estimates is the Chebyshev fit to some p+1 element subsample of the data, the following modification algorithm can be used to compute all the percentile estimates in one pass through the $\hat{\theta}_c$. Use the $i^{th} = k-p^{th}$ row of a $(n-p)\times p$ matrix K to hold the current best estimate of $\hat{\theta}_{(k)}$ for k = p+1, p+2, ..., n. At each $\hat{\theta}_c$, compute and sort the squared residuals. Then update $\hat{\theta}_{(k)}$ if $r^2_{(k)}(\hat{\theta}_c)$ is less than the previous smallest value for $r^2_{(k)}(\theta)$. After considering all $\hat{\theta}_c$, the matrix K will contain the exact values of $\hat{\theta}_{(k)}$ for k = p+1, p+2, ..., n.

The algorithm can also be modified to compute $\hat{\theta}_{(i)}$, the LMS estimate for the data set with the ith data point deleted. This can be done at the same time as the computation of the LMS estimate for the full data set. In general, use the ith row of a $n\times p$ matrix to hold the current best estimate of $\hat{\theta}_{(i)}$. For each $\hat{\theta}_c$, check for improvement in each of the $\hat{\theta}_{(i)}$. Of course, those $\hat{\theta}_c$ based on elemental sets that contain point i must be excluded from the possible $\hat{\theta}_{(i)}$ values. For any other $\hat{\theta}_c$, the median squared residual for the data set with the ith point deleted will be the (q + j)th largest residual for the entire data set where:

$$j = \begin{cases} 0 \text{ if } r^2_{(i)}(\hat{\theta}_c) > r^2_{(q)}(\hat{\theta}_c) \\ 1 \text{ if } r^2_{(i)}(\hat{\theta}_c) \le r^2_{(q)}(\hat{\theta}_c) \end{cases}$$

Thus, the squared residuals need only be computed once at each $\hat{\theta}_c$ to find $\hat{\theta}_{LMS}$ and $\hat{\theta}_{(i)}$, i = 1,2,...,n.

The $\hat{\theta}_{(i)}$ can be used as a diagnostic tool. If the plot of LMS residuals versus LMS fit values using $\hat{\theta}_{(i)}$ is quite different than the same residual plot using the entire data set, then point i can be considered influential.

The $\hat{\theta}_{(i)}$ can also be used to produce jackknifed standard error estimates for the the LMS estimator. Jackknifed covariance estimates for $\hat{\theta}_{LMS}$ can be computed using the method described by Efron (1982, p.18-19). He suggests the following covariance matrix:

$$\text{JACKCOV} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(.)})(\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^T \quad \text{where } \hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}.$$

The consequence is that a jackknifed estimate of the covariance matrix can be computed with little more computational effort than the computation of $\hat{\theta}_{LMS}$. Unfortunately, jackknifed standard errors do not agree with bootstrapped or Monte Carlo standard error estimates, thus it is the authors conclusion that the jackknife isn't a reasonable method for computing standard errors for $\hat{\theta}_{LMS}$. It seems possible, but unlikely because of the nonlinearity of the LMS procedure, that a modified jackknife (Wu 1986, Simonoff and Tsai 1986, 1988) may eventually yield a reasonable method for computing standard errors for the LMS estimate.

3. EXAMPLES

The most notable difference between the approximate algorithms and the exact algorithm is that the elemental sets consist of p+1 points for the the exact algorithm but only p points for the approximate algorithms. As an example of how this can effect the $\hat{\theta}_{LMS}$ fit, consider the data in Table 3.1, fit by a simple linear regression through the origin model. Both the PROGRESS and MVELMS algorithms use one point elemental sets, while the exact algorithm uses two point elemental sets. The PROGRESS and MVELMS algorithms find the the line that passes through point 8 which has slope .657225, and median squared residual .107. According to this fit, points zero through four should be considered outliers. The exact LMS fit is the Chebyshev fit to points 4 and 5, which has slope .38485 and median squared residual .075. According to the exact fit, points six through nine should be considered outliers. The regression lines are depicted in Figure 3.1.

Table 3.1

Data fit by Simple Linear Regression Through the Origin

Point #:	0	1	2	3	4	5	6	7	8	9
x:	1	2	3	4	5	1	2	3	4	5
y:	0.3302	0.6590	0.9888	1.3194	1.6495	0.6596	1.3192	1.9815	2.6289	3.3011

The $\hat{\theta}_{(i)}$ are useful in understanding the LMS fit to this data. If any of the first five points are removed, the LMS fit shifts to fit the upper five points while removing any one of the upper five points has little impact on the LMS fit. It seems that considering any of the points to be outliers when the

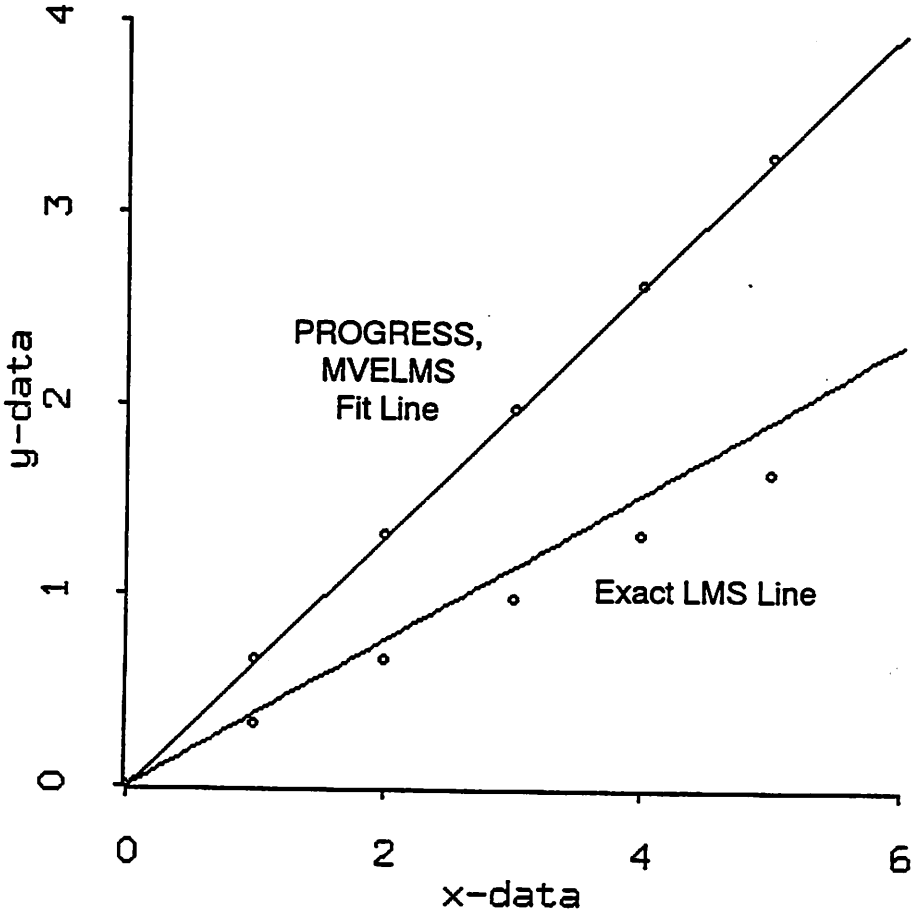

Figure 3.1. PROGRESS, MVELMS, and exact LMS fit of a Simple Linear Regression through the Origin model for the data in Table 3.1.

LMS fit to the $\hat{\theta}_{(i)}$ is so variable is questionable.

The data set in Table 3.2 (from Cook and Weisberg 1982, p. 4) summarizes the results of a cloud seeding experiment in Florida in 1975. On each of 24 days suitable for seeding, the following six explanatory variables were recorded:

A: "Action" was set to zero if no seeding took place and to one if seeding occurred.

T: "Time" was the number of days since the beginning of the experiment.

S: "Suitability" was a measure of the days suitability for seeding.

C: "Echo coverage" was the percent cloud coverage in the experimental area.

P: "Prewetness" was the total rainfall in the target area in the hour before seeding.

E: "Echo motion" was set to 1 for a moving radar echo and 2 for a stationary radar echo.

The data was fit using a multiple linear regression model with the preceding six explanatory variables and an intercept. The response variable was ln(rainfall) in a target area for a six hour period. The PROGRESS approximation to $\hat{\theta}_{LMS}$ (1.43, .695, -.016, -.455, -.039, .941, 1.08) is based on the exact fit to points 0,1,9,16,17,19, and 20. Its median squared residual was .0601. The MVELMS approximation to $\hat{\theta}_{LMS}$ (.740, 1.13, -.0047, -.567, -.056, 3.60, .990) is based on the exact fit to points 2,3,5,8,11,21, and 23. Its median squared residual was .0278. Neither of the these approximations find seven of the eight points (2,3,4,8,9,11,16,23) that determine the exact LMS fit (.715, 1.13, -.0052, -.551, -.056, 3.61, .962) which has median squared residual .0241. The MVELMS basis does have 5 points in common with the exact basis, which explains why it is close to the exact LMS fit.

The plot of the least median of squares residuals versus fit values has been suggested (Rousseeuw and Leroy, 1987) for assisting in detecting outliers in multiple linear regression. The PROGRESS, MVELMS, and exact LMS residual plots for the cloud seeding data are given in Figure 3.2. Note that with the exception of points 6, 15, the PROGRESS algorithm identifies different outliers than the other two methods. The MVELMS plot is very close to the exact LMS plot, but there would be no way to know this without computing the exact LMS fit. For this data set, none of the residual plots based on $\hat{\theta}_{(i)}$, i=1,2,..,n vary much from the residual plot for the full data set (Figure 3.2(1a)), thus none of the data points are flagged as particularly influential.

The ability to compute the exact LMS estimate allows us to study the stability of $\hat{\theta}_{LMS}$ under shifts in the observed values. Let the modified cloud seeding data be the cloud seeding data with the response at point 4 shifted from 0.8961 to 1.1061. The residual plots for the three methods are almost identical to those given if Figure 3.2. If we shift the response for point 4 from 1.1061 to 1.1161, the PROGRESS and MVELMS fits are virtually unchanged from those in Figure 3.2, but the exact LMS residual plot is now similar to the PROGRESS plot of Figure 3.2. It is interesting that although the PROGRESS and exact LMS residual plots are similar in this case, the MVELMS fit has a smaller median squared residual than the PROGRESS fit. The instabily of the LMS residual plot is shown in

Figure 3.3 where the shift of point 4 from 1.1061 to 1.1161 causes a different set of outliers to be identified. In the modified data set with point four at 1.1161, the influence of point four on the fit is evident from $\hat{\theta}_{(4)}$ which yields a residual plot quite different from the residual plot for the entire data set (Figure 3.3(b)). As expected, since only point four has been modified, the $\hat{\theta}_{(4)}$ residual plot is similar to Figure 3.3(a).

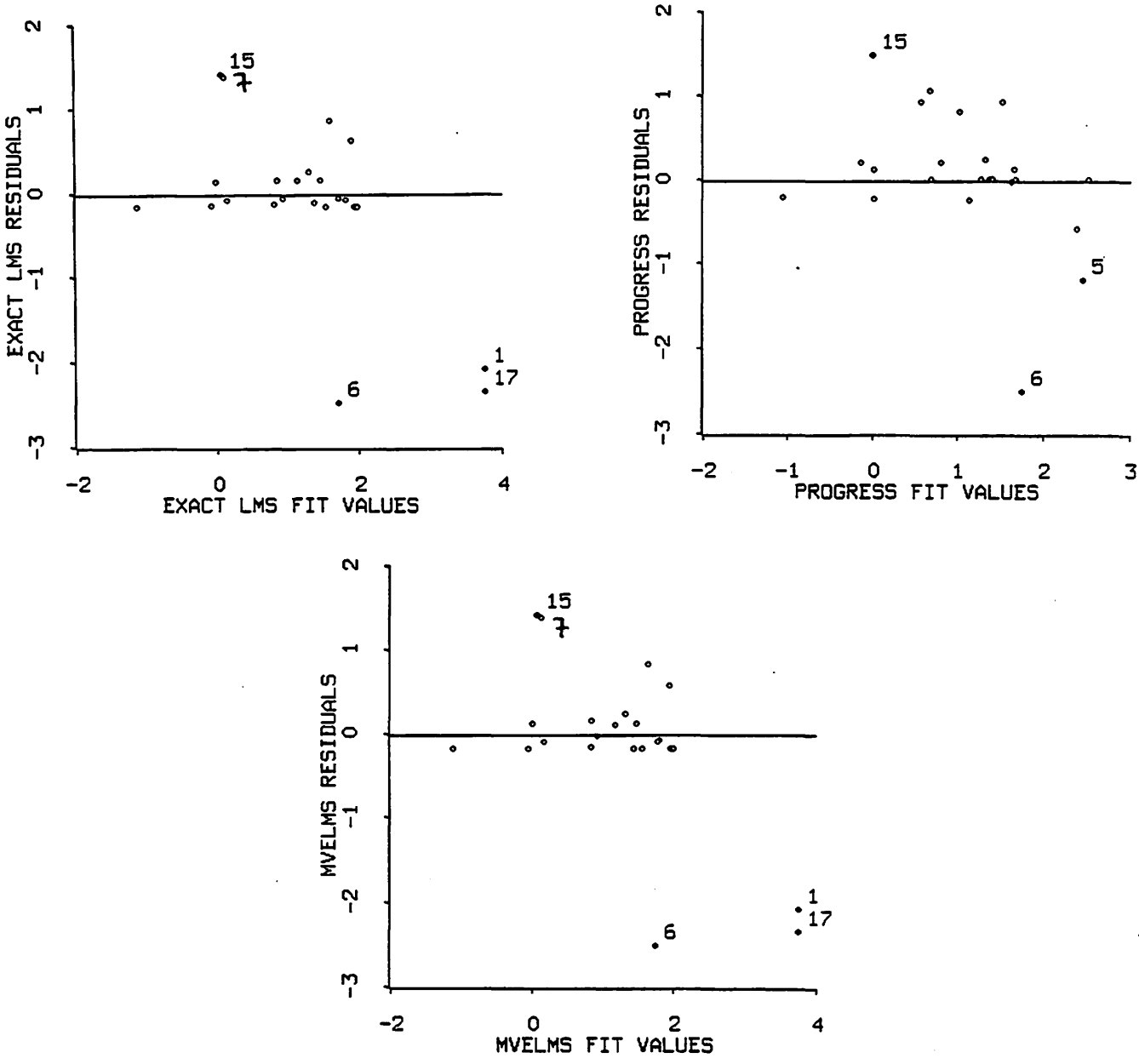

Figure 3.2. LMS residuals versus fit values for the cloud seeding data using (a) Exact LMS, (b) PROGRESS and (c)MVELMS.

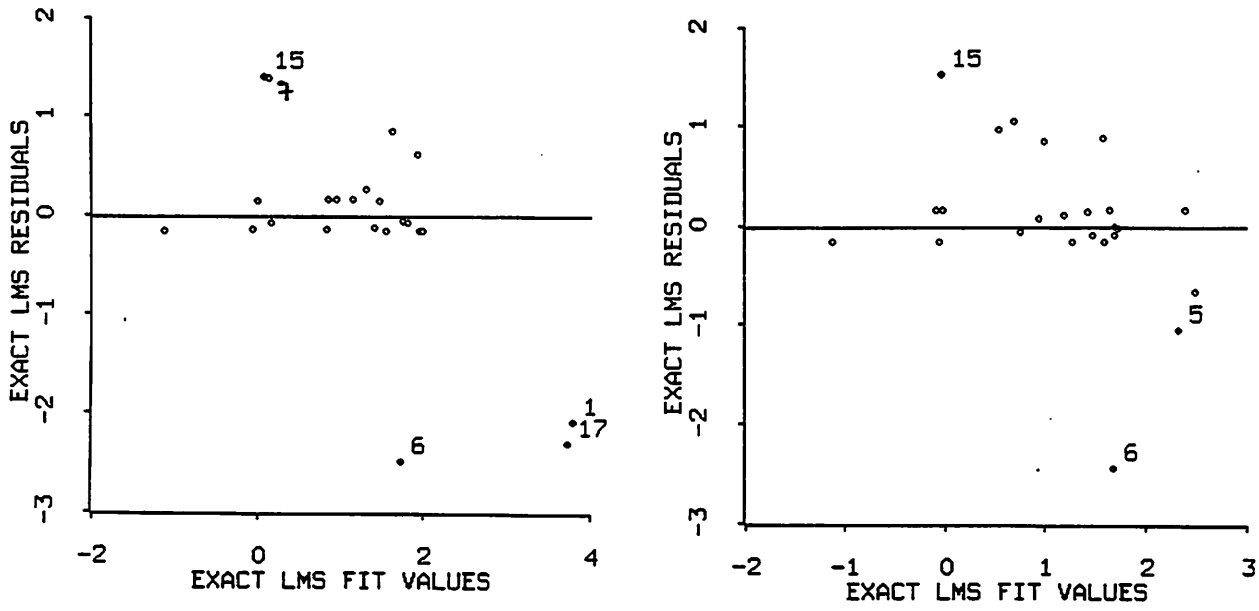

Figure 3.3 Exact LMS residuals versus fit values for the cloud seeding data with point 4 at (a) 1.1061 (b) 1.1161.

Table 3.2. Cloud Seeding Data

Index	Action	Time	Suitability	Echo Coverage	Prewetness	Echo motion	ln(Rainfall)
0	0	0	1.75	13.4	.274	2	2.5533
1	1	1	2.70	37.9	1.267	1	1.7084
2	1	3	4.1	3.9	.198	2	1.8390
3	0	4	2.35	5.3	.526	1	1.8099
4	1	6	4.25	7.1	.25	1	0.8961
5	0	9	1.6	6.9	.018	2	1.2837
6	0	18	1.3	4.6	.307	1	-0.755
7	0	25	3.35	4.9	.194	1	1.5173
8	0	27	2.85	12.1	.751	1	1.8485
9	1	28	2.2	5.2	.084	1	1.6214
10	1	29	4.4	4.1	.236	1	1.0152
11	1	32	3.1	2.8	.214	1	1.3987
12	0	33	3.95	6.8	.796	1	1.7475
13	1	35	2.9	3.0	.124	1	1.5769
14	1	38	2.05	7.0	.144	1	2.4732
15	0	39	4.0	11.3	.398	1	1.4929
16	0	53	3.35	4.2	.237	2	1.2975
17	1	55	3.7	3.3	.960	1	1.4398
18	0	56	3.8	2.2	.230	1	0.1484
19	1	59	3.4	6.5	.142	2	1.6956
20	1	65	3.15	3.1	.073	1	0.7031
21	0	68	3.15	2.6	.136	1	-.1985
22	1	82	4.01	8.3	.123	1	.00862
23	0	83	4.65	7.4	.168	1	-1.273

REFERENCES

Atkinson, A. C. (1986). "Masking Unmasked," *Biometrika*, 73: 533-542.

Cheney, E. W. (1962). *Introduction to Approximation Theory.* McGraw-Hill.

Cook, R.D and Hawkins, D. M.(1990). Comment on "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85: 640-644.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* London: Chapman-Hall.

Donoho, D. L. and Huber, P. J., (1983), "The Notion of Breakdown Point," In *A Festschrift for Erich L. Lehmann,* P. J. Bickel, K. A. Doksum, J. L. Hodges Jr. (eds.), Belmont, California: Wadsworth, 157–184.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Hawkins, D. and Simonoff, J. S. (1991), "High Breakdown Regression and Multivariate Estimation," NYU Stat/OR Dept. Working Paper #91-5.

Rousseeuw, P.J. (1984). "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79: 871-880.

Rousseeuw, P.J. and A.M. Leroy (1987). *Robust Regression and Outlier Detection.* New York: John Wiley and Sons.

Rousseeuw, P.J., and B.C. van Zomeren (1990). "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85: 633-639.

Simonoff, J. S. and Tsai, C. -L. (1986), "Jackknife-based Estimators and Confidence Regions in Nonlinear Regression," *Technometrics* 28, 103-112.

Simonoff, J. S. and Tsai, C. -L. (1988), "Jackknifing and Bootstrapping Quasi-Likelihood Estimators," *Journal of Statistical Computation and Simulation* 30, 213-232.

Steele, J.M. and W.L. Steiger (1986). "Algorithms and Complexity for Least Median of Squares Regression," *Discrete Applied Mathematics 14:93-100.*

Wu, C. F. J. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis (with discussion)," *Annals of Statistics* 14, 1261-1350.