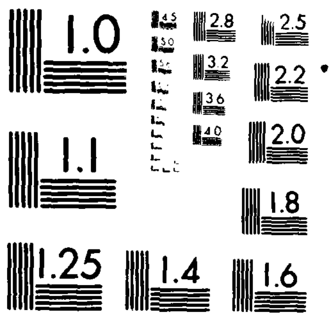
AD-A092 660 PRINCETON UNIV NJ DEPT OF STATISTICS F/G 12/1
ROBUST REGRESSION USING REPEATED MEDIANS.(U)
SEP 80 A F SIEGEL DAAG29-79-C-0205
UNCLASSIFIED TR-172-SER-2 ARO-16669.2-M NL

END
DATE FILMED
1 81
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

LEVEL II

12

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 19 16669.2-M 18
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): 6 Robust Regression Using Repeated Medians.
5. TYPE OF REPORT & PERIOD COVERED: 9 Technical
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): 10 Andrew F. Siegel
8. CONTRACT OR GRANT NUMBER(s): 15 DAAG29-79-C-0205
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Princeton University, Princeton, NJ 08540
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: 11 U. S. Army Research Office, Post Office Box 12211, Research Triangle Park, NC 27709
12. REPORT DATE: Sep 80 12 18
13. NUMBER OF PAGES: 15
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

DTIC ELECTE DEC 0 8 1980 S D E

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

breakdown value
U-statistic
resistance
regression
algorithms
estimation
iteration

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The repeated median algorithm is a robustified U-statistic in which nested medians replace the single mean. Unlike many generalizations of the univariate median, repeated median estimates maintain the high 50% breakdown value and can resist the effects of outliers even when they comprise nearly half of the data. Because they are calculated directly, not iteratively, repeated median procedures can be used as starting values for iterative robust estimation methods. For bivariate linear regression with symmetric errors, repeated median estimates are unbiased and Fisher consistent, and their efficiency under Gaussian sampling can be comparable to the efficiency of the univariate median.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD A092660

DDC FILE COPY

80 12 01 127

ROBUST REGRESSION USING REPEATED MEDIANS

by

Andrew F. Siegel*
Princeton University

Technical Report No. 172, Series 2
Department of Statistics
Princeton University

September 1980

*Andrew F. Siegel is Assistant Professor, Department of Statistics, Princeton University, Princeton, New Jersey, 08544. Research was supported in part by U.S. Army Research Office Contract DAAG29-79-C-0205 awarded to the Statistics Department, Princeton University.

ROBUST REGRESSION USING REPEATED MEDIANS

by

Andrew F. Siegel
Princeton University

A B S T R A C T

The repeated median algorithm is a robustified U-statistic in which nested medians replace the single mean. Unlike many generalizations of the univariate median, repeated median estimates maintain the high 50% breakdown value and can resist the effects of outliers even when they comprise nearly half of the data. Because they are calculated directly, not iteratively, repeated median procedures can be used as starting values for iterative robust estimation methods. For bivariate linear regression with symmetric errors, repeated median estimates are unbiased and Fisher consistent, and their efficiency under Gaussian sampling can be comparable to the efficiency of the univariate median.

Key Words: Breakdown Value, U-Statistic, Resistance.

1. INTRODUCTION

Robust regression procedures based on medians have been considered by Thiel(1950), Mood(1950, p.406), Brown and Mood(1951), Sen(1968), Maritz(1979), and others. Such high-breakdown procedures are of interest for several reasons. First, some applied problems, including the editing of data, require maximal protection against the presence of outliers. Siegel and Benson (1980) provide an example of this need in the comparison of shapes. Secondly, many of the more efficient robust procedures, including M-estimates (Huber, 1973) are iterative and require directly computable resistant starting values (Andrews, 1974) to guard against convergence to a non-robust local optimum near the least-squares solution. Finally, the extreme case of high-breakdown estimates should be well understood.

The repeated median algorithm is defined in Section 2 as a modified U-statistic in which nested medians are used instead of a single mean, and their computational complexity is found. The breakdown value is shown in Section 3 to be 50%, the best possible for unbounded invariant estimators and an improvement upon previously considered median procedures. Under suitable conditions, repeated median estimates are unbiased and Fisher consistent, as shown in Section 4, and their efficiency under Gaussian sampling can be comparable to the efficiency of the univariate median.

2. THE REPEATED MEDIAN ALGORITHM

We first consider the bivariate linear case of fitting a robust line $Y=A+BX$ to the data (X_i, Y_i) , $i=1,\ldots,n$ with distinct X_i . Define the pairwise slope $B(i,j)=(Y_j-Y_i)/(X_j-X_i)$ of the line from point i to point j . These $n(n-1)/2$ slope estimates will be condensed into a single number using two stages of medians. The repeated median estimate of slope is

$$\hat{B} = \underset{i}{\text{Median}} \{\underset{j\neq i}{\text{Median}} \; B(i,j)\} \qquad (2.1)$$

The inner median is the median slope of the lines that pass through point i . We can visualize (2.1) as the median of the column medians (or row medians, by symmetry) of the $B(i,j)$ matrix, ignoring entries along the main diagonal. This is not an iterative method; if we calculate (2.1) using the residuals $R_i=Y_i-\hat{B}X_i$ in place of Y_i , we obtain zero by additive invariance of the median.

The y-intercept A can be estimated in two ways. If we use the value $\hat{B}$ just estimated, a single median will suffice for this hierarchical approach:

$$\hat{A}^{(1)} = \underset{i}{\text{Median}} \; (Y_i-\hat{B}\; X_i) \qquad (2.2)$$

Otherwise, A can be estimated directly using a double median as in (2.1) to obtain

$$\hat{A}^{(2)} = \underset{i}{\text{Median}} \{\underset{j\neq i}{\text{Median}} \; A(i,j)\} \qquad (2.3)$$

where $A(i,j) = (X_jY_i - X_iY_j)/(X_j - X_i)$ is the y-intercept of the line connecting points i and j. Less time is required for computing the hierarchical estimate (2.2), but direct estimation (as in 2.3) is invariant to the ordering of the parameters A and B.

The general repeated median algorithm is like a U-statistic (Hoeffding, 1948), except that nested medians replace the overall mean. We therefore obtain a general procedure for estimating a real parameter θ whenever there is a positive integer k such that every subset of k data points determines a value of θ; say points numbered $i_1,\ldots,i_k$ determine $\theta(i_1,\ldots,i_k)$. The mean of these estimates, if we have n data points in all, is the U-statistic.

$$\binom{n}{k}^{-1} \sum_{(1\leq i_1<\ldots<i_k\leq n)} \theta(i_1,\ldots,i_k) \tag{2.4}$$

Using a median in place of the mean, we can robustify this somewhat to

$$\underset{(1\leq i_1<\ldots<i_k\leq n)}{\text{Median}} \{\theta(i_1,\ldots,i_k)\} \tag{2.5}$$

which includes the case of regression estimates considered by Thiel(1950) and Sen(1968).

Repeated median estimates use a succession of k partial medians. Begin by reducing the number of indices from k to $k-1$.

$$\theta(i_1,\ldots,i_{k-1},\cdot) = \underset{i_k \notin \{i_1,\ldots,i_{k-1}\}}{\text{Median}} \theta(i_1,\ldots,i_k) \tag{2.6}$$

This process can be repeated, and with each median an index is deleted. Finally, the repeated median estimate is

$$\hat{\theta} = \underset{i_1}{\text{Median}} \left\{ \underset{i_2 \notin \{i_1\}}{\text{Median}} \cdots \left[\underset{i_k \notin \{i_1,\ldots,i_{k-1}\}}{\text{Median}} \theta(i_1,\ldots,i_k) \right] \cdots \right\} \tag{2.7}$$

For example, in the multiple regression model

$$Y = A + B_1X_1 + B_2X_2 \tag{2.8}$$

B_1 would be estimated using a triple median

$$\hat{B}_1 = \underset{i}{\text{Median}} \left\{ \underset{j \neq i}{\text{Median}} \left[\underset{k \neq i,j}{\text{Median}} B_1(i,j,k) \right] \right\} \tag{2.9}$$

where $B_1(i,j,k)$ is the B_1 coefficient of the plane (2.8) determined by points i,j, and k. Colinearity problems can be handled by considering only those triples that actually determine a value for B_1.

When more than one parameter is to be estimated, they can be estimated hierarchically using information on previously estimated parameters at each stage or directly using (2.7) for each parameter. These two approaches were illustrated in (2.2) and (2.3), and the same considerations apply in general.

The computational complexity of (2.7) is $O(n^k)$ because the total number of medians of $n-1$ or fewer numbers that must be performed is

$$1 + \sum_{i=1}^{k-1} \left[\prod_{j=1}^{i} (n-j) \right] = O(n^{k-1}) \qquad (2.10)$$

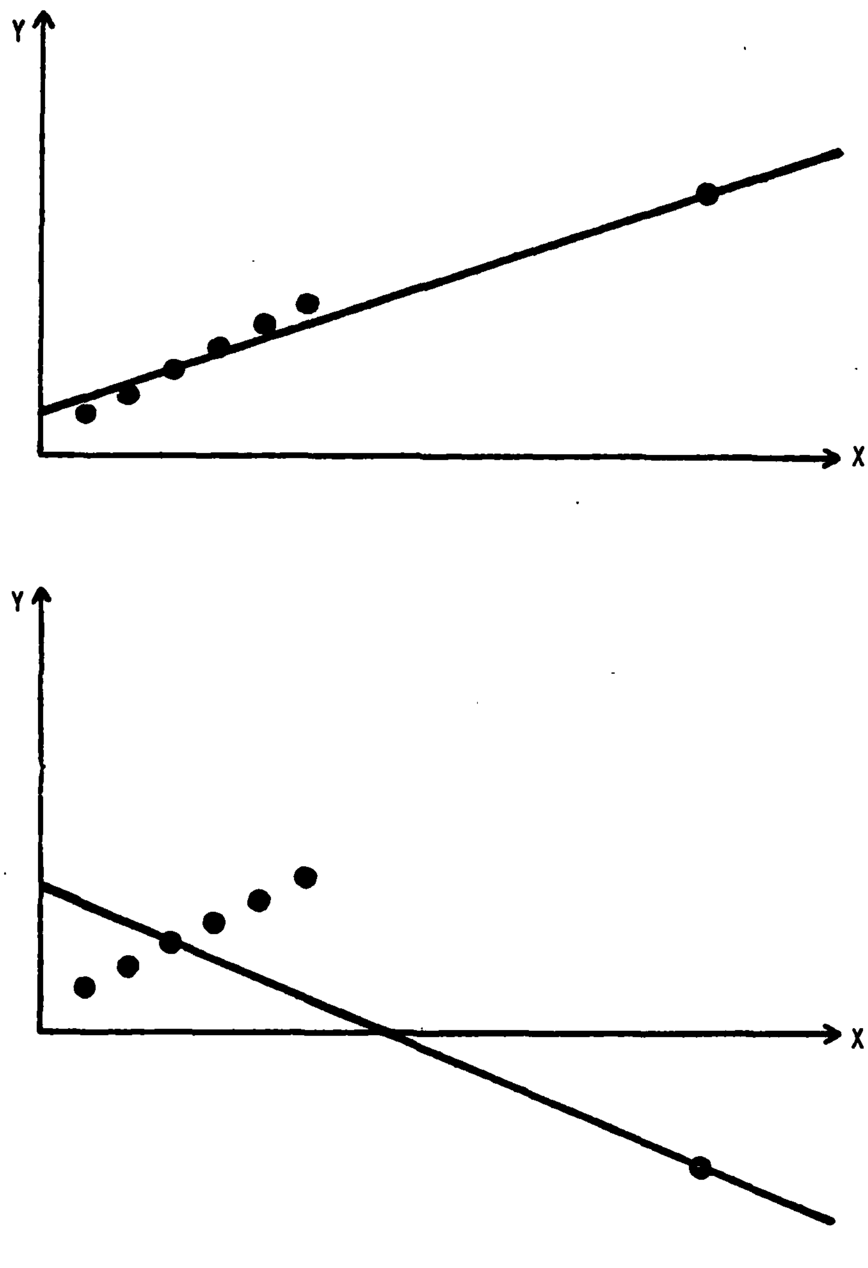
and an $O(n)$ algorithm is available for calculating the median (Knuth, Vol. III, 1973, Section 5.3.3, p. 216).

3. BREAKDOWN VALUE

Breakdown value is a measure of the ability of an estimator to resist the effects of outliers (Hodges, 1967, and Hampel, 1971). It is, roughly speaking, the largest fraction of the data that can be arbitrarily changed while the estimator is guaranteed to remain bounded. The arithmetic mean has a breakdown value of 0%, while the univariate median achieves nearly 50% because $[(n-1)/2]$ out of n points can be changed while the median remains bounded (brackets indicate the greatest integer function). This value, 50%, is the highest possible for invariant unbounded estimators.

Median-based regression methods do not necessarily preserve the highest possible 50% breakdown value of the univariate median. For example, least absolute error regression (Bassett and Koenker, 1978) has a breakdown value of zero (0%); the figure shows an example in which the least absolute error regression line can be controlled by changing only the height of a single point.

The Mood-Brown procedure for bivariate linear regression (Mood, 1950; and Brown and Mood, 1951) requires that the median residual be zero for both halves (low X and high X) of the data. Because half of the data in either group can control the estimated line, the breakdown value is 25%. The breakdown value of Andrews' median-based regression method is also at most 25% (Andrews, 1974, Section 5).



<u>FIGURE</u> 1. The height of a single influential point can control the least absolute error regression line.

The overall median procedure (2.5), studied by Thiel(1950) and Sen(1968) for bivariate linear regression, has a breakdown value of 29%. In higher dimensions, with subsets of k points at a time, the breakdown value is $1-2^{-(1/k)}$. This is found by setting the ratio of the number of unchanged to total estimates $\theta(i_1,\ldots,i_k)$ equal to 1/2, the breakdown value for the median. When the primitive estimates $\theta(i_1,\ldots,i_k)$ are themselves robust, the resulting breakdown value can be higher.

The repeated median procedure has an asymptotic breakdown value of 50% (as $n\to\infty$ with k fixed) because each nested median in (2.7) involves n or fewer terms (the overall, nonrepeated median (2.5) involves n^k terms at once in a single median). This is shown in the following theorem which finds the exact breakdown value in small samples:

<u>Theorem</u>. The repeated median estimate (2.7) will remain bounded whenever more than $(n+k-1)/2$ points are held fixed while the remaining points are arbitrarily moved, provided each subset of k of the fixed points determines a value $\theta(i_1,\ldots,i_k)$.

This theorem is a consequence of a more general lemma.

<u>Lemma</u>. Consider a class of functions $\theta_\alpha(i_1,\ldots,i_k)$ where $1\le i_j\le n$ are integers and different values of α can be thought of as different data configurations. Suppose $A\subset\{1,\ldots,n\}$ has more than $(n+k-1)/2$ elements and $\theta_\alpha(i_1,\ldots,i_k)$ are bounded (as α varies) whenever $i_1,\ldots,i_k\in A$. Then the repeated median values $\hat\theta_\alpha$ calculated from (2.7) are also bounded.

<u>Proof</u>. We proceed by induction on k . When $k=1$, this reduces to the breakdown bound of the univariate median. Now assume the hypotheses of the lemma. Performing the innermost median (2.6) in (2.7) we see that

$$\theta_\alpha(i_1,\ldots,i_{k-1},\cdot) = \underset{i_k \notin \{i_1,\ldots,i_{k-1}\}}{\text{Median}} \theta_\alpha(i_1,\ldots,i_k)$$

are bounded whenever $i_1,\ldots,i_{k-1} \in A$ because the median has $n-k$ terms, of which more than half are bounded. Note that for each α , the k-fold repeated median of $\theta_\alpha(i_1,\ldots,i_k)$ is identical to the $(k-1)$-fold repeated median of $\theta_\alpha(i_1,\ldots,i_{k-1},\cdot)$. These are seen to be bounded by using the induction hypothesis. □

<u>Proof</u>. We proceed by induction on k . When $k=1$, this reduces to the breakdown bound of the univariate median. Now assume the hypotheses of the lemma. Performing the innermost median (2.6) in (2.7) we see that

$$\theta_\alpha(i_1,\ldots,i_{k-1},\cdot) = \underset{i_k \notin \{i_1,\ldots,i_{k-1}\}}{\text{Median}} \theta_\alpha(i_1,\ldots,i_k)$$

are bounded whenever $i_1,\ldots,i_{k-1} \in A$ because the median has $n-k$ terms, of which more than half are bounded. Note that for each α , the k-fold repeated median of $\theta_\alpha(i_1,\ldots,i_k)$ is identical to the $(k-1)$-fold repeated median of $\theta_\alpha(i_1,\ldots,i_{k-1},\cdot)$. These are seen to be bounded by using the induction hypothesis. □

4. UNBIASEDNESS, FISHER CONSISTENCY, AND EFFICIENCY

The repeated median estimates are unbiased in the bivariate linear model

$$Y_i = A + BX_i + \varepsilon_i \quad , \quad i=1,\ldots, n \tag{4.1}$$

with fixed X_i and symmetric errors for which $(\varepsilon_1,\ldots, \varepsilon_n) \overset{D}{=} (-\varepsilon_1,\ldots, -\varepsilon_n)$. The slope estimate $\hat{B}$ from (2.1) is symmetrically distributed about the true slope B because

$$\begin{aligned}
\hat{B}-B &= \underset{i}{\text{Median}} \left[\underset{j \neq i}{\text{Median}} \left(\frac{\varepsilon_j - \varepsilon_i}{X_j - X_i} \right) \right] \\
&= -\left\{ \underset{i}{\text{Median}} \left[\underset{j \neq i}{\text{Median}} \left(\frac{(-\varepsilon_j)-(-\varepsilon_i)}{X_j - X_i} \right) \right] \right\} \\
&\overset{D}{=} -\left\{ \underset{i}{\text{Median}} \left[\underset{j \neq i}{\text{Median}} \left(\frac{\varepsilon_j - \varepsilon_i}{X_j - X_i} \right) \right] \right\} \\
&= -(\hat{B}-B)
\end{aligned} \tag{4.2}$$

Thus $E(\hat{B})=B$ whenever the expectation exists. We find similarly that $\hat{A}$ is symmetrically distributed about A for both the single median (2.2) and the double median (2.3) calculation.

Repeated median estimates are Fisher consistent for bivariate distributions in which Y given X is symmetrically distributed about a center that is linear in X, so that $(X, Y - A - BX) \overset{D}{=} (X, -(Y-A-BX))$. Fisher consistency requires that when we evaluate the estimator at the actual population distribution (not at a sample), we obtain the population parameter (Cox and Hinkley, 1974, p. 287). The repeated median procedure (2.7) extends to allow us to estimate the slope B given a distribution $(X, Y) \sim F$. Assume the X marginal is continuous and define

$$\hat{B} = \underset{(X,Y)\sim F}{\text{Median}} \left[\underset{(X',Y')\sim F}{\text{Median}} \quad \frac{Y'-Y}{X'-X} \right] \tag{4.3}$$

This is algebraically equivalent to

$$\hat{B}-B = \underset{(X,Y)\sim F}{\text{Median}} \left[\underset{(X',Y')\sim F}{\text{Median}} \quad \frac{(Y'-A-BX')-(Y-A-BX)}{X'-X} \right]$$

$$= -\left\{ \underset{(X,Y)\sim F}{\text{Median}} \left[\underset{(X',Y')\sim F}{\text{Median}} \quad \frac{[-(Y'-A-BX')]-[-(Y-A-BX)]}{X'-X} \right] \right.$$

$$= -(\hat{B}-B) \tag{4.4}$$

where the last equality follows by symmetry. Because these are fixed, not random, variables, (4.4) must be zero and we have $\hat{B}=B$. Similarly, it can be shown that $\hat{A}=A$ regardless of whether $\hat{A}$ is found using a single or double median.

The efficiency of repeated median regression, in the presence of Gaussian errors, is not far from the efficiency of the univariate median, as shown in the table for evenly spaced and for Gaussian X values. Efficiency here is the ratio of the variances of the least squares and median-based estimates. For the univariate median, this ratio is assymptotically $2/\pi \cong .64$ (Cramer, 1946, p. 369).

Efficiencies for repeated median regression were estimated using Monte Carlo computer simulation techniques. For each table entry, 10,000 replications were performed in order to achieve an estimated standard error of the efficiency smaller than .01. Simulations were done on Princeton University's IBM 3033 Computer using the IMSL subroutine ggnpm for pseudorandom Gaussian deviates. Three X designs were chosen: evenly spaced, even Gaussian percentiles ($\Phi^{-1}((i-\frac{1}{2})/n)$, $i=1, \ldots, n$ where Φ denotes the standard Gaussian cumulative distribution function) and random Gaussian deviates chosen independently for each replication.

TABLE 1.

Efficiency of repeated median regression
bivariate slope estimation
with
independent Gaussian errors,
by Monte Carlo simulation

	X design		
		Gaussian	
n	evenly spaced	even percentiles	random
10	.69	.64	.53
20	.73	.65	.61

REFERENCES

ANDREWS, D.F. (1974). A Robust Method for Multiple Linear Regression. Technometrics 16, 523-531.

BASSETT, G., JR., and KOENKER, R. (1978). Asymptotic Theory of Least Absolute Error Regression. Journal of the American Statistical Association 73, 618-622.

BROWN, G.W., and MOOD, A.M. (1951). On Median Tests for Linear Hypotheses. Proceedings of the Second Berkeley Symposium in Mathematical Statistics and Probability, 159-166, University of California Press.

COX, D.R., and HINKLEY, D.V. (1974). Theoretical Statistics. London: Chapman and Hall.

CRAMER, H. (1946). Mathematical Methods of Statistics. Princeton: Princeton University Press.

HAMPEL, F.R. (1971). A General Qualitative Definition of Robustness. Annals of Mathematical Statistics 42, 1887-1896.

HODGES, J.L., JR. (1967). Efficiency in Normal Samples and Tolerance of Extreme Values for Some Estimates of Location. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1, 163-186. Berkeley: University of California Press.

HOEFFDING, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. Annals of Mathematical Statistics 19, 293-325.

HUBER, P.J. (1973). Robust Regression: Asymptotics, Conjectures, and Monte Carlo. <u>Annals of Statistics</u> <u>1</u>, 799-821.

KNUTH, D.E. (1973). <u>The Art of Computer Programming</u>, Volume III, <u>Sorting and Searching</u>. Reading, MA.: Addison-Wesley.

MARITZ, J.S. (1979). On Thiel's Method in Distribution-Free Regression. <u>Australian Journal of Statistics</u> <u>21</u>, 30-35.

MOOD, A.M. (1950). <u>Introduction to the Theory of Statistics</u>. New York: McGraw-Hill.

SEN, P.K. (1968). Estimates of the Regression Coefficient Based on Kendall's Tau. <u>Journal of the American Statistical Association</u> <u>63</u>, 1379-1389.

SIEGEL, A.F., and BENSON, R.H. (1980). Estimating Allometric Change in Animal Morphology. Submitted to <u>Biometrics</u>.

THIEL, H. (1950). A Rank-Invariant Method of Linear and Polynomial Regression Analysis, I, II, and III. <u>Nederlandse Akademie van Wetenschappen Proceedings</u> <u>53</u>, 386-392, 521-525, and 1397-1412.