**Report Title**: Robust Regression Using Repeated Medians.

**Authors**: Andrew F. Siegel

**Performing Organization Name and Address**: Princeton University, Princeton, NJ 08540

**Controlling Office Name and Address**: U.S. Army Research Office, Post Office Box 12211, Research Triangle Park, NC 27709

**Abstract**: 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.
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

ABSTRACT

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

\[
\hat{B} = \text{Median} \left\{ \text{Median}_{i \neq j} B(i,j) \right\} \quad (2.1)
\]

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:

\[
\hat{A}^{(1)} = \text{Median}_{i} (Y_i - \hat{B}X_i) \quad (2.2)
\]

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

\[
\hat{A}^{(2)} = \text{Median}_{i} \left\{ \text{Median}_{j \neq i} A(i,j) \right\} \quad (2.3)
\]
where $A(i,j) = (X_j Y_i - X_i Y_j)/(X_j - X_i)$ 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 $\theta$ whenever there is a positive integer $k$ such that every subset of $k$ data points determines a value of $\theta$; 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.

$$
\left(\frac{n}{k}\right)^{-1} \sum_{1 \leq i_1 < \ldots < i_k \leq n} \theta(i_1, \ldots, i_k) \quad (2.4)
$$

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

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

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 $k-1$. 
\[
\theta(i_1, \ldots, i_{k-1}, \cdot) = \text{Median}_{i_k \notin \{i_1, \ldots, i_{k-1}\}} \theta(i_1, \ldots, i_k)
\] (2.6)

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

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

For example, in the multiple regression model

\[
Y = A + B_1X_1 + B_2X_2
\] (2.8)

\(B_1\) would be estimated using a triple median

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

where \(B_1(i, j, k)\) is the \(B_1\) coefficient of the plane (2.8) determined by points \(i, j,\) and \(k\). Collinearity 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

$$1 + \sum_{i=1}^{k-1} \frac{i}{\Pi_{j=1}^{n-j}} = O(n^{k-1})$$

(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).
FIGURE 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 \( 1 - 2^{-2}(1/k) \). 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:

**Theorem.** 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.

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

$$\theta_\alpha(i_1, \ldots, i_{k-1}, *) = \text{Median}_{i_k \in \{i_1, \ldots, i_{k-1}\}} \theta_\alpha(i_1, \ldots, i_k)$$

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 $\alpha$, the $k$-fold repeated median of $\theta_\alpha(i_1, \ldots, i_k)$ is identical to the $(k-1)$-fold repeated median of $\theta_\alpha(i_1, \ldots, i_{k-1}, *)$. These are seen to be bounded by using the induction hypothesis. \[\square\]
Proof. 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

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

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 $\alpha$, 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. \qed
4. UNBIASEDNESS, FISHER CONSISTENCY, AND EFFICIENCY

The repeated median estimates are unbiased in the bivariate linear model

\[ Y_i = A + BX_i + \varepsilon_i, \quad i=1, \ldots, n \] (4.1)

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

\[
\hat{B} - B = \text{Median}_i \left[ \frac{e_j - e_i}{x_j - x_i} \right] \\
= \left\{ \begin{array}{l}
\text{Median}_i \left[ \frac{-e_j - \varepsilon_i}{x_j - x_i} \right] \\
\text{Median}_j \left[ \frac{-e_j - \varepsilon_i}{x_j - x_i} \right]
\end{array} \right.
\]

\[
= \frac{\text{Median}_i \left[ \frac{e_j - e_i}{x_j - x_i} \right] - \text{Median}_j \left[ \frac{e_j - e_i}{x_j - x_i} \right]}{\text{Median}_i \left[ \frac{e_j - e_i}{x_j - x_i} \right] - \text{Median}_j \left[ \frac{e_j - e_i}{x_j - x_i} \right]}
\]

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 \((X, Y - A - BX) = (X, -(Y-A-BX))\). 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

\[
\hat{B} = \text{Median}_{(X,Y)\sim F} \left[ \text{Median}_{(X',Y')\sim F} \frac{Y' - Y}{X' - X} \right]
\]

(4.3)

This is algebraically equivalent to

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

\[
= \left( \text{Median}_{(X,Y)\sim F} \left[ \text{Median}_{(X',Y')\sim F} \frac{-(Y' - A - B X') - (Y - A - B X)}{X' - X} \right] \right)
\]

\[
= -(\hat{B} - B)
\]

(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 asymptotically \( 2/\pi \approx 0.64 \) (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 \( \Phi \) 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

<table>
<thead>
<tr>
<th>X design</th>
<th>Gaussian</th>
<th>evenly spaced</th>
<th>even percentiles</th>
<th>random</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>.69</td>
<td>.64</td>
<td>.53</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>.73</td>
<td>.65</td>
<td>.61</td>
</tr>
</tbody>
</table>
REFERENCES


