THE RESISTANT LINE AND RELATED REGRESSION METHODS

by

Iain Johnstone
Stanford University
and
Mathematical Sciences Research Institute

and

Paul Velleman
Cornell University

ORION 018
JUNE 1983

PROJECT ORION

Department of Statistics
Stanford University
Stanford, California
The Resistant Line and Related Regression Methods

by

Iain Johnstone *, Stanford University, and MSRI
Paul Velleman, Cornell University

Abstract

In a bivariate \((x,y)\) scatter-plot the three-group resistant line is that line for which the median residual in each outer third of the data (ordered on \(x\)) is zero. It was proposed by Tukey as an exploratory method resistant to outliers in \(x\) and \(y\) and suited to hand calculation. A dual plot representation of the procedure yields a fast, convergent algorithm, non-parametric confidence intervals for the slope, consistency, influence function and asymptotic normality results. Monte-Carlo results show that small sample efficiencies exceed their asymptotic values in important cases and are adequate for exploratory work. Replacing medians by other M-estimators of location increases efficiency substantially without affecting breakdown or computational complexity. The method thus combines bounded influence and acceptable efficiency with conceptual and computational simplicity.

Key Words: Resistant line, bounded influence, robust regression, dual plot, variance-reduction swindle, Brown-Mood regression, Berry-Esseen theorem, M-estimator, confidence interval, repeated median regression.

*Work by Professor Johnstone was supported in part at Cornell by an Australian National University Postbachelor Travelling Scholarship, at Stanford by ONR contract # N00014-81-K-0340, and at the Math Sciences Research Institute by NSF. Work by Professor Velleman was supported in part by ARO(Durham) #DAAG29-82-C-0178 awarded to Princeton University. The authors thank David Hoaglin for helpful comments on an earlier draft of this paper and Cornell University for the computer funds.
1. INTRODUCTION AND SUMMARY

Tukey (1971) proposed the (three group) resistant line as an exploratory data analytic tool for quickly fitting a straight line to bivariate \( (x, y) \) data. The data points are divided into three groups according to smallest, middle or largest \( x \) values and the line with an equal number of points above and below it in each of the outer groups is fitted. The resulting parameter estimates are resistant to the effects of data points extreme in \( y \), or \( x \) or both. (For detailed discussions including examples see Velleman and Hoaglin (1981) and Emerson and Hoaglin (1983a).) Designed as a "pencil and paper" method, the resistant line is also useful as a computationally quick nonparametric regression in computer-assisted data analyses. Although it compares favorably to other well-known nonparametric regression methods and is widely available in statistics packages (for example, Minitab (Ryan et al. 1981), and P-Stat (Buhler and Buhler, 1981), the method has not been analyzed theoretically.

This article presents an inexpensive computational algorithm (§2), nonparametric confidence intervals (§3), and asymptotic results on consistency, influence function, and asymptotic normality (§4) for the resistant line. A class of "\( \psi \)-resistant" lines constructed from \( M \)-estimators of location is then introduced (§5). While retaining the computational and resistance properties of the resistant line, these estimators can have substantially higher efficiency. Finally, we use the preceding theory and the results of a Monte Carlo simulation experiment of small sample properties (§6) to compare the resistant lines with related line-fitting procedures. The experiment uses two new variance reduction
"swindles" and provides small-sample results not previously available for most of the estimators studied.

In considering the merits of various alternatives to least squares regression methods in exploratory work our primary criteria are:

1. resistance; including protection from data values extreme in $x$ as well as those extreme in $y$, ("bounded influence"),

2. cost of computation, and

3. asymptotic and small-sample efficiency for both Gaussian and heavy-tailed error distributions.

The resistant line performs well for (1) and (2), and in many cases has an efficiency of roughly 60%—exceeding the 50% usually considered adequate for exploratory data analyses. The modifications proposed in Section 5 can increase the practical efficiency to roughly 75% under a wide range of error distribution and increase the asymptotic relative efficiency to 80% under classical regression assumptions.

Precedent Literature

Tukey's method combines features of two traditional approaches to linear regression. Wald (1940) studied the errors-in-variables problem of fitting a straight line $y = a + bx$ when both $y$ and $x$ are subject to error. He partitions the observed data points $(x_i, y_i)$ into two sets $L = \{ (x_i, y_i) : x_i \leq \text{med } x \}$ and $R = \{ (x_i, y_i) : x_i > \text{med } x \}$ and uses an estimator based on group means:

$$b_W = (\bar{y}_R - \bar{y}_L) / (\bar{x}_R - \bar{x}_L).$$

Independently of Wald, Nair and Shrivastava (1942) applied a similar estimator to fixed, equally spaced $x$-values divided into three groups. They showed for Gaussian $y$ that an optimal (minimum variance) estimator of the form (1.1) is obtained by using only $n/3$ points in each
of the extreme groups, increasing its efficiency relative to least squares from $3/4$ to $8/9$. Bartlett (1949) made a similar observation. Mosteller (1946) noted the advantage of three groups in simple partition-based estimation of the correlation coefficient of a bivariate Gaussian $(x,y)$ population and found that the optimal division places about 27% of the data in each of the outer groups. The same optimality calculation (Section 4) applies to the regression problem and in a different context leads to the so called "27% rule" in psychometrics (c.f. Kelley (1939), and McCabe (1980) for some references). Further results on optimal allocation for other $x$-distributions and small sample approximations are given by Gibson and Jowett (1957a) and Barton and Casley (1958).

The second approach (Mood, 1950; Brown and Mood, 1951) is non-parametric: divide the data into two groups with the split at the median $x$ (as in Wald's method), and find the slope $b_{BM}$ making the median residual in the left group equal to the median of the residuals in the right group. The intercept $a_{BM}$ is then chosen to make the median of all the residuals zero. Let $\bar{x}_L = \text{med}(x_i: x_i \in L)$, $\bar{y}_L = \text{med}(y_i: x_i \in L)$, etc. Mood's (1950) algorithm for finding the slope (repeated in Tukey, 1971) begins with $b^{(1)} = (\bar{y}_R - \bar{y}_L)/(\bar{x}_R - \bar{x}_L)$, computes residuals $e_i^{(1)}$, and iterates:

\[(1,2) \quad b^{(k+1)} = b^{(k)} + (e_R^{(k)} - e_L^{(k)})/(\bar{x}_R - \bar{x}_L).\]

Here $e_R^{(k)}$ and $e_L^{(k)}$ are the medians of the $k^{th}$ step residuals in the right and left groups, respectively. The iteration is to continue until the two group medians are equal or the correction is as small as required. This appears to be the only published algorithm for Brown-Mood regression.
However, experience shows that it fails to converge often enough to make it inappropriate for practical data analysis. Section 2 offers an explanation and an algorithm whose rapid convergence is guaranteed.

Brown and Mood, and Daniels (1954) provided tests for the hypothesis \((a, b) = (a_0, b_0)\) assuming fixed \(x\)-values. Bhaskar (1961) gave some consistency results and Hill (1962) proved asymptotic normality for \((a_{BM}, b_{BM})\). Both authors proposed tests for linearity based on Brown-Mood regression. Hogg (1975) fit regression lines based on percentiles other than the median of the residuals in each half of the data. Kildea (1981) obtained asymptotic normality of a class of Brown-Mood estimators using weighted medians and proposed an optimally weighted, but non-robust estimator.

The resistant line thus combines the resistance to outliers of the Brown-Mood median regression with the improved efficiency obtained from three partitions.

In addition to the regression estimators discussed above we include three others of current interest in the comparisons. Two are more expensive to compute than those discussed thus far, but reward the extra effort with higher breakdown values and/or greater efficiencies. These are the median-of-pairwise-slopes procedure studied by Theil (1950) and Sen (1968); \(b_{TS} = \text{med}\{(y_j - y_i)/(x_j - x_i), x_j > x_i\}\), which has generally high efficiency, and the related repeated median estimator \(b_{RM} = \text{med}\{\text{med}_{i \neq j}(y_j - y_i)/(x_j - x_i)\}\) introduced by Siegel (1982), which has optimal 50% breakdown. The third, least absolute residual \((L_1)\) regression (e.g. Laplace 1818, Bassett and Koenker 1978), has been suggested as an improvement on least squares for heavy-tailed error
distributions. It has an influence function that is unbounded in \( x \), and thus can be sensitive to extreme \( x \) values. With this caveat, \( L_1 \) regression performs comparably to the \( \psi \)-resistant lines with respect to the criteria (1) - (3) above. We have not included a bounded influence robust regression of the Krasker-Welsch (1982) type because there is no agreement yet on how to apply the definitions to simple regression.

Some of the results in this paper were first announced in Johnstone and Velleman (1982).

\[ \text{§2. PROPERTIES OF THE RESISTANT LINE} \]

Let data points \((x_i, y_i), \ldots, (x_n, y_n)\) be given with \( x_1 \leq x_2 \leq \ldots \leq x_n \). Partition the \( x \)-values into three groups \( L, M, \) and \( R \) containing \( n_L, n_M, \) and \( n_R \) data points respectively \((n_L + n_M + n_R = n)\); where \( x_L = \max(L) \leq x_R = \min(R) \) and \( M \subset (x_L, x_R) \) might be empty. (See Velleman and Hoaglin (1981) for a detailed algorithm.) We shall occasionally abuse notation by writing \( L \) for \( \{i : x_i \in L\} \), etc. The resistant line slope estimate \( b_{RL} \) is then defined as the solution to

\[
(2.1) \quad \text{med}_{i \in L} (y_i - bx_i) = \text{med}_{i \in R} (y_i - bx_i).
\]

The intercept estimate \( a_{RL} \) may, for example, be chosen to make the median residuals of both groups zero. Brown-Mood regression is the special case in which \( n_M = 0, n_L = n_R = n/2 \). The resistant line is a special case of Kildea's (1981) weighted median estimators that uses only 0-1 weights.
2.1 A Dual Plot  The dual plot interchanges the roles of points and lines. Ignore the intercept, \( a \), and plot each residual, \( e_i(b) = y_i - x_i b \), against \( b \). This yields a line with slope \(-x_i\) and intercept \( y_i\) to correspond to the original data point \((x_i, y_i)\). Two dual lines \( e_i(b) \) and \( e_j(b) \) will intersect at a point having \( b\)-value

\[ b_{ij} = \frac{(y_i - y_j)}{(x_i - x_j)}, \]

namely the slope of the line in \((x,y)\) space joining \((x_i, y_i)\) and \((x_j, y_j)\) (interpreted as \(\infty\) if \(x_i = x_j\)). For an illustrative example, see Emerson and Hoaglin (1983a). Dolby (1960) and Daniels (1954) have used the dual plot in simple linear regression contexts.

The median trace of the left partition is the piecewise-linear function

\[ T_L(b) = \text{med}_{i \in L} (y_i - x_i b). \]

If \( n_L \) is odd, \( T_L \) consists of line segments from the dual lines in \( L \). If \( n_L \) is even, the line segments in \( T_L \) average adjacent pairs of dual lines. Since \( x_L < x_R \) the right median trace

\[ T_R(b) = \text{med}_{i \in R} (y_i - bx_i) \]

intersects \( T_L \) exactly once. The \( b\)-value of this intersection is the slope of the resistant line. (See figure A.) This graphical construction shows that \( b_{RL} \) always exists and is unique.

The dual plot reveals the difficulties encountered by the Mood-Tukey algorithm.

Let \( \tilde{x}_R = \text{med}_{i \in R} x_i \), \( \tilde{y}_R = \text{med}_{i \in R} y_i \), etc. and \( \Delta x = \tilde{x}_R - \tilde{x}_L \). Then the first estimated slope is \( b^{(1)} = (\tilde{y}_R - \tilde{y}_L) / \Delta x \).

On the dual plot the distance between the median residuals in the left and right groups. Thus, from (1,2), \( b^{(k+1)} = b^{(k)} + [T_R(b^{(k)}) - T_L(b^{(k)})] / \Delta x \). Now, if
Figure A. The Dual Plot displays for each data point \((x, y)\) the line 
\[ e = y - xb. \]
Here lines are shown only for points in the Left or the Right partitions of the data set. The median traces, \(T_L\) and \(T_R\), follow the median residual, \(e\), at each slope, \(b\). The abcissa of their intersection is the resistant line slope, \(b_{RL}\).
(2.2) \[ |[T_R(b^{(k)}) - T_L(b^{(k)})]/\Delta x| > 2|b^{(k)} - b_{RL}| \]

then \[ |b^{(k+1)} - b_{RL}| > |b^{(k)} - b_{RL}| \], and this step of the iteration has made the approximation worse. Inequality (2.2) can be written

(2.3) \[ \left| \frac{T_R(b^{(k)}) - T_R(b_{RL})}{b^{(k)} - b_{RL}} - \frac{T_L(b^{(k)}) - T_L(b_{RL})}{b^{(k)} - b_{RL}} \right| > 2\Delta x \]

which bounds the difference in the slopes of the median traces near \( b_{RL} \).

A median trace can be steep only if an x-value in its partition is large in magnitude. Thus the iteration can be stymied by high-leverage points with extraordinary x-values. See Emerson and Hoaglin (1983) for a specific example constructed by A. Siegel.

2.2 An Algorithm for Resistant Lines. The dual plot suggests an algorithm for the resistant line that will always converge. The zero of the piecewise-linear monotone-decreasing function \( E(b) = T_L(b) - T_R(b) \) yields the resistant line slope. Among algorithms for finding zeros of functions the one known as ZEROIN (Wilkinson (1967), Dekker (1969), Brent (1973)) is well-suited to this problem. The algorithm requires choice of a tolerance, TOL, which is the amount of error in \( b \) that can be tolerated, and an interval \( (b_{\min}, b_{\max}) \) to search. Brent (1973) shows that this algorithm will always converge in fewer than \( \left[ \log_2\left(\frac{b_{\max} - b_{\min}}{TOL} \right) \right]^2 \) steps, where \( TOL = 0.5TOL + 2.0 \times \epsilon \times |b_{\max}| \) and \( \epsilon \) is the machine epsilon. Brent claims that the algorithm can usually be expected to converge much more quickly, and for this application our experience confirms this.

Each step requires the median of the residuals in each of the outer groups. Asymptotically, the median of n numbers can be found with \( O(n) \).
comparisons. (See, for example, Aho, Hopcroft, and Ullman (1974) pp. 97-99.) However, for moderate sample size, \( n/3 \) is too small to make the large overhead of specialized methods worthwhile. The convergence rate of ZEROIN is independent of \( n \), so this algorithm requires \( O(n) \) operations.

Because \( E(b) \) is monotone, any \( b_{\text{max}} \) and \( b_{\text{min}} \) that satisfy \( \text{sign}(E(b_{\text{max}})) \neq \text{sign}(E(b_{\text{min}})) \) are suitable; the first two estimates from Mood's original iteration often suffice. Velleman and Hoaglin (1981) provide portable programs for finding resistant lines using this algorithm.

2.3 Unbiasedness. Consider the bivariate linear model
\[
y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n
\]
in which the \( \{\epsilon_i\} \) are i.i.d. and
(I) the \( x_i \) are fixed, or (II) the \( x_i \) are i.i.d. and independent of \( \epsilon_i \). The results are given for model (I) and follow for model (II) by conditioning on \( x \).

If \( n_L \) and \( n_R \) are both odd, then \( b_{RL} \) is median unbiased for \( \beta \) without any assumption of symmetry on the \( \epsilon_i \):
\[
P(b_{RL} > \beta) = P(\text{med}_{i \in R} \epsilon_i > \text{med}_{i \in L} \epsilon_i) \geq 1/2 ,
\]
and similarly \( P(b_{RL} < \beta) \geq 1/2 \). Suppose now also that the errors \( \epsilon_i \) are symmetric about 0: in this case \( b_{RL} \) (and \( a_{RL} \)) are symmetrically distributed about \( \beta \) (and \( \alpha \)), and thus are median unbiased and unbiased (if \( E|\epsilon_i| < \infty \)) for all values of \( n_L \) and \( n_R \). To see this note that
\[
\{b_{RL} - \beta > c\} = \{\text{med}_{i \in R} (\epsilon_i - cx_i) > \text{med}_{i \in L} (\epsilon_i - cx_i)\}
\]
and that \( \{b_{RL} - \beta < -c\} \) is equivalent to the same event with every \( \epsilon_i \) replaced by \( -\epsilon_i \).
Breakdown Bounds. The (gross error) breakdown point of an estimator measures its ability to resist the effects of outliers (Hampel 1971). Loosely, the breakdown point is the largest fraction of the data that can be changed arbitrarily without changing the estimator beyond all bounds. The resistant line has a breakdown value of
\[ \frac{1}{n} \min\{ \frac{(n_L - 1)/2}{}, \frac{(n_R - 1)/2}{}, \} \]
because we need only alter half of the data points in one of the outer partitions to take complete control of the slope estimate. If \( n_L = n_M = n_R \), this is approximately equal to 1/6.

A Probabilistic Breakdown Value. While a breakdown value of 1/6 is certainly resistant (least squares and least-absolute-value regression both have breakdowns of 0%), it is poorer than comparable values for Brown-Mood (25%), Theil-Sen regression (29%), or repeated median regression (50%). However, breakdown concentrates on worst-case behavior. In reality, not all subsets of size 1/6 are equally dangerous for the resistant line; for the line to break down, all bad points must lie in the same outer third. If we restrict attention to gross errors in \( y \) (so that the \( x \)-values are held fixed), and assume that extreme \( y \)-values occur independently with probability \( \alpha \), we can calculate a "probabilistic breakdown value" (c.f. Donoho and Huber, 1982, Section 5.1). This gives a plausible rather than worst-case bound on the protection available. For the resistant line (including Brown-Mood), breakdown occurs if there are more than \( \left\lfloor \frac{(n_L - 1)/2}{}, (n_R - 1)/2 \right\rfloor \) extreme \( y \) values in the left (right) groups. For convenience take \( n_L = n_R = 2k + 1 \); then the breakdown probability becomes
\[ 1 - \Pr(\text{Bin}(2k+1, \alpha) \leq k)^2 = 1 - I_{1-\alpha} (k+1, k+1)^2, \]
where \( I_x(p,q) \) is Pearson's incomplete Beta function.

Table 1 gives some values for the breakdown probability under the above "random gross error model". For the resistant line, \( n_L = n_R = n/3 \) is used, while for Brown-Mood regression, we take \( n_L = n_R = \left\lceil n/2 \right\rceil \), which yields a larger value of \( k \) and hence lower breakdown probabilities.
For comparison, results for repeated median and Theil-Sen regression are given. In these cases, the possibility of breakdown is affected only by the size of the subset drawn, so that the breakdown probability is given by a single binomial probability.

Table 1 shows for example that for sample size 27, in 90% of cases the resistant line can tolerate about 25% contamination with outliers (and this still assumes that all outliers reinforce each other). For practical applications, a value near 25% seems appropriate as a guide in determining the applicability of the resistant line method. Note that the Theil-Sen method is much more sensitive to α-values close to and exceeding its nominal breakdown point.

1. **Illustrative Breakdown Probabilities**

\[ α = \text{Probability of Gross Error} \]

<table>
<thead>
<tr>
<th>method</th>
<th>.20</th>
<th>.25</th>
<th>.33</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resistant line</td>
<td>.04</td>
<td>.10</td>
<td>.26</td>
</tr>
<tr>
<td>Brown-Mood</td>
<td>.014</td>
<td>.05</td>
<td>.19</td>
</tr>
<tr>
<td>Repeated Median</td>
<td>.0002</td>
<td>.003</td>
<td>.04</td>
</tr>
<tr>
<td>Theil-Sen</td>
<td>.07</td>
<td>.25</td>
<td>.56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>method</th>
<th>.20</th>
<th>.25</th>
<th>.33</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resistant line</td>
<td>.002</td>
<td>.013</td>
<td>.10</td>
</tr>
<tr>
<td>Brown-Mood</td>
<td>.0002</td>
<td>.003</td>
<td>.05</td>
</tr>
<tr>
<td>Repeated Median</td>
<td>0</td>
<td>0</td>
<td>.002</td>
</tr>
<tr>
<td>Theil-Sen</td>
<td>.02</td>
<td>.14</td>
<td>.37</td>
</tr>
</tbody>
</table>
2.5 Resistant Line and Repeated Median Regression

The repeated median regression estimate of Siegel (1982),

\[ b_{RM} = \text{med}_{i} \text{med}_{j\neq i} b_{ij} \]

pays for its optimal gross error breakdown of nearly 50% with a computational complexity of \( O(n^2) \) operations. Interestingly, modifying the repeated median by computing \( b_{ij} \) only for \( x_i \in L \) and \( x_j \in R \) is equivalent to the \( O(n) \) resistant line estimator.

**Proposition 2.1.** If \( n_L \) and \( n_R \) are both odd, then \( b_{RL} = \text{med}_{r \in R} \text{med}_{\ell \in L} b_{\ell r} \).

**Proof.** In the dual plot, let the intersection of \( e_r(b) = y_r - x_r b \) \((x_r \in R)\) with \( T_L(b) \) occur at \( b_r \). From the definition \( T_L(b_r) = \text{med}_{\ell \in L} e_{\ell}(b_r) \) and the relation

\[ e_{\ell}(b_r) \begin{cases} > 0 \\ < 0 \end{cases} \Rightarrow T_L(b_r) \text{ respectively as } b_{\ell r} \begin{cases} < 0 \\ > 0 \end{cases} \]

it follows (since \( n_L \) is odd) that \( b_r = \text{med}_{\ell \in L} b_{\ell r} \).

\( 2^0 \). Since \( b_{RL} \) satisfies \( \text{med}_{r \in R} e_r(b_{RL}) = T_L(b_{RL}) \) by definition, it is enough to note that

\[ e_r(b_{RL}) \begin{cases} > 0 \\ < 0 \end{cases} \Rightarrow T_L(b_{RL}) \text{ respectively as } b_r \begin{cases} < 0 \\ > 0 \end{cases} b_{RL} \]

**Remark.** In an ordered set \( x_1 \leq x_2 \leq \cdots \leq x_{2n} \) of even cardinality, we denote \( x_n \) by lomed \( x_i \) and \( x_{n+1} \) by himed \( x_i \). If \( n_L \) is even, the conclusion of \( 1^0 \) is weakened to \( b_r \leq \text{himed} b_{\ell r} \), and if both \( n_L \) and \( n_R \) are even (for example) the conclusion of the proposition is just that

\[ \text{lomed lomed } b_{\ell r} \leq b_{RL} \leq \text{himed himed} b_{\ell r}. \]
53. CONFIDENCE INTERVALS FOR $\beta$

Non-parametric confidence intervals for the slope $\beta$ in a regression model may be derived by an analogue of the method that gives interval estimates for the population median. Consider Model I of §2.3 and assume that the $\epsilon_i$ have a continuous distribution; the results extend to the random carriers model as before.

Write $e_i(\beta) = y_i - x_i\beta$ for the residual of $(x_i, y_i)$ at $\beta$ (ignoring $\alpha$) in the dual plot. Let $e_{(r)}^L(\beta)$ denote the $r$th smallest residual in the left group, $\{y_i - x_i\beta, i \in L\}$, and define $e_{(s)}^R(\beta)$ similarly in the right group. Assume that $n_L = n_R = m$ (say) for convenience: there is no difficulty in extending the theory or computations to unequal $n_L$ and $n_R$. As a convention, set $s = m + 1 - r$. Let $\hat{\beta}_r$ be the unique solution of

\begin{equation}
(3.1) \quad e_{(r)}^L(\beta) = e_{(s)}^R(\beta).
\end{equation}

Existence and uniqueness of $\hat{\beta}_r$ follow from the dual plot as before.

Of course, if $m$ is odd then $\hat{\beta}[(m+1)/2] = \hat{\beta}_{m/2}$, and, with an appropriate choice of intercept $\alpha$, finding $\hat{\beta}_r$ amounts to locating the line in $(x, y)$ space such that $r$ data points in the left group lie on or below the line and $r$ points in the right group lie on or above the line.

Clearly $\hat{\beta}_1 \geq \hat{\beta}_2 \geq \ldots \geq \hat{\beta}_m$, and we now show that pairs $(\hat{\beta}_s, \hat{\beta}_r)$ ($r < s$) yield nonparametric confidence intervals for $\beta$. The function $g(\beta) = e_{(s)}^R(\beta) - e_{(r)}^L(\beta)$ is strictly decreasing, and hence if the model $y_i = \alpha + \beta x_i + \epsilon_i$ obtains,
\[ P_\beta(\beta > \beta_r) = P_\beta(e^L_r(\beta) > e^R_s(\beta)) = P\{\varepsilon^L_r > \varepsilon^R_s\}, \]

where \( \varepsilon^L_1 \leq \ldots \leq \varepsilon^L_m \) and \( \varepsilon^R_1 \leq \ldots \leq \varepsilon^R_m \) are the ordered \( \varepsilon_i \) in the outer groups. The quantity \( U^m_r = \text{number of values in right group which exceed} \ varepsilon^L_r \), is an exceedance statistic and has probability mass function

\[ (3.2) \quad P(U^m_r = x) = \frac{(m-x+r-1)!}{(r-1)!x!(2m)!}, \quad x = 0,1,\ldots,m. \]

as may be seen from a simple combinatorial argument. Now

\[ P_\beta(\beta > \beta_r) = P(U^m_r < r), \quad \text{and by a symmetrical argument (since} \ s = m+1-r), \]

\[ P_\beta(\beta_s > \beta_r) = P(U^m_r < r) \text{ also. Thus the interval} \ [\beta_s, \beta_r] \text{ covers} \ \beta \text{ with probability} \ 1 - 2 P(U^m_r < r). \]

The latter expression has been tabulated by Epstein (1954) for \( m = 2(1)15 \ (5)20, r \leq m \), and more extensively by Bechtel (1982).

Remarks:

1) Finiteness of the set \( \{\beta_1, \ldots, \beta_m\} \) limits the number of possible confidence levels. This can be a non-trivial restriction if \( m \) is small.

2) The confidence interval above can be inverted in the usual way to give a test of \( H_0: \beta = \beta_0 \) against general alternatives. The test is based on the number of data points in the left group that lie on or above the line with slope \( \beta_0 \) and intercept chosen so that \( m \) points in \( L \cup R \) lie on or above it. This test is discussed in the case \( n_M = 0 \), using normal approximations, by Hogg (1975), along with extensions to percentile regression.
3) If instead of Models I or II, it is assumed that given \( x, \alpha + \beta x \) is the \((100 \, r/m)\)th percentile of the \( Y \) distribution, then an estimate of the kind considered by Hogg (1975) is obtained (with \( n_M \) not restricted to be zero) by solving for \( \beta \) in anologue of (3.1):

\[
\frac{e^L}{e^R}(\beta) = \frac{e^R}{e^L}(\beta).
\]

**Asymptotic approximation for** \( P(W^m_r < r) \) **The expression (3.2) is a negative hypergeometric probability and as such may be thought of in terms of fair coin-tossing as** \( P[X = x | X + Y = m] \) **where** \( X \) **and** \( Y \) **are independent negative binomial variables recording the number of tails before the \((m - r + 1)^{st}\) and \(r^{th}\) heads respectively. Switching to i.i.d.**

\( X_i = 1 \) if heads, \(-1\) if tails, with \( S^m_m = \sum_1^m X_i \), we have \( \{X + Y = m\} = \{S^m_{2m} = 0, S^m_{2m+1} = 1\} \) and \( \{X < r\} = \{(S^m_m + m)/2 \geq m - r + 1\} \), so that if one sets \( r = (m + 2 - x \sqrt{m})/2 \), then

\[
P(W^m_r < r) = P(S^m_m / \sqrt{2m} \geq x / \sqrt{2} \mid S^m_{2m} / \sqrt{2m} = 0)
\]

(3.3)

\[
\rightarrow P(W^0(\frac{1}{2}) \geq x / \sqrt{2}) = 1 - \Phi(x / \sqrt{2}),
\]

where \( W^0(t) \) is a standard Brownian Bridge on \([0,1]\) (Billingsley, 1968).

A less revealing demonstration can be based directly on the negative hypergeometric probabilities using the (local) normal approximation given in Feller (1968, p. 194). The approximate confidence intervals that result have coverage probabilities close to the nominal level even for moderate \( m \).
The standard continuity correction (replace $x$ by $x-1/\sqrt{m}$ in equation 3.3) yields the formula

\[(3.4) \quad r_c = \frac{1}{2}(m + 1 - \sqrt{m/2} \, z_{1-\alpha})\]

for a two-sided approximate $100(1 - 2\alpha)\%$ confidence interval. For example, if $m = 20$ and $\alpha = .05$, then $r = 7,9$, which rounds to 8. The interval $(\beta_{13}, \beta_{8})$ has exact coverage probability $1 - 2P\left(U_{8}^{20} \leq 7\right) = .89$ (Epstein 1954). Bechtel (1982) has shown that values of $r_c$ from equation 3.4 round to the exact values even at the smallest applicable sample sizes.

When $r_c$ is not an integer, we can interpolate for the edges of the interval. Interpolation of $\log P(U_r^n < r - 1)$ vs $r/n$ appears (from data analysis) to be satisfactory. This can overcome the restricted choice of $\alpha$ values imposed by integer $r$. (See remark 1 of the previous section.) Of course, the coverage probability is no longer exactly distribution-free.
§4. ASYMPTOTIC PROPERTIES OF THE RESISTANT LINE

To study asymptotic properties, we extend the definition of the resistant line slope estimate from bivariate scatter plots to arbitrary bivariate distributions $H$ of random variables $(X, Y)$ with values in $\mathbb{R}^2$.

Let $x_L \leq x_R$ be lower $p_L$th and upper $p_R$th percentiles of the marginal distribution of $X$ such that $p_L = P(X \leq x_L)$ and $p_R = P(X \geq x_R)$ and assume that either $\text{med}(X|X \leq x_L) < x_L$ or $\text{med}(X|X \geq x_R) > x_R$ or both. Assume that $0 < p_L \leq 1 - p_R < 1$, and that the partitioning is symmetric: $p_L = p_R$, although this is done for notational convenience only.

The lower and upper medians of a distribution function $G$ are defined by $\text{lomed} G = \sup(x: G(x) < 1/2)$ and $\text{himed} G = \inf(x: G(x) > 1/2)$, and we take $\text{med} G = (\text{lomed} G + \text{himed} G)/2$. Define the resistant line slope functional $\beta_{RL}(H)$ as the crossing point of zero of the function

$$h(\beta) = \text{med}(Y - \beta X|X \geq x_R) - \text{med}(Y - \beta X|X \leq x_L).$$

Existence and uniqueness are clear because the slope of $h(\beta)$ is always less than $-(x_R - x_L)$ and $h(\beta)/\beta \rightarrow +\left[\text{med}(X|X \leq x_L) - \text{med}(X|X \geq x_R)\right]$ as $\beta \rightarrow +\infty$.

4.1 Consistency

Fisher consistency of $\beta_{RL}(H)$ holds for linear regression models of the form $Y = \alpha + \beta X + \varepsilon$, where $\text{med}(\varepsilon|X) = \text{med} \varepsilon$ almost surely.

Indeed, $\text{med}(Y - \beta X|X \geq x_R) = \text{med}(\alpha + \varepsilon|X \geq x_R) = \text{med}(\alpha + \varepsilon) = \text{med}(\alpha + \varepsilon|X \leq x_L) = \text{med}(Y - \beta X|X \leq x_L)$ so that $\beta_{RL}(H) = \beta$.

In location problems the median of i.i.d. samples from $G$ is consistent iff $\text{lomed} G = \text{himed} G$. To describe the analogous strong consistency
properties of $\beta_{RL}(H)$, we introduce

$$h^*(\beta) = \text{himed}(Y - \beta X \mid X \geq x_R) - \text{lomed}(Y - \beta X \mid X \leq x_L),$$

and

$$h_*(\beta) = \text{lomed}(Y - \beta X \mid X \geq x_R) - \text{himed}(Y - \beta X \mid X \leq x_L).$$

Clearly $h^*(\beta) \geq h(\beta) \geq h_*(\beta)$, and the unique zeros $\beta^*, \beta_{RL}, \beta_*$ of the respective functions satisfy $\beta^* \geq \beta_{RL} \geq \beta_*$. 

**Example 4.1** Suppose that $H$ places equal mass on the four points $(-1,-1), (-1,0), (1,0), (1,1)$. Then $h^*(\beta) = 2 - 2\beta$, $h(\beta) = 1 - 2\beta$, $h_*(\beta) = -2\beta$, and $\beta^* = 1$, $\beta_{RL} = 1/2$, $\beta_* = 0$. For samples $H_n$ of size $n$ from $H$ it is clearly possible to have $\beta_{RL}(H_n) = 0$ or $1$.

Let $H_n$ be a sequence of bivariate distributions with corresponding partition points $x_{L,n} \prec x_{R,n}$, probabilities $p_{L,n}$, $p_{R,n}$, and resistant line estimators $\beta_{RL}^{(n)}$. The notations $H_n \overset{\psi}{\Rightarrow} H$ and $\mathcal{L}(X_n) \rightarrow \mathcal{L}(X)$ denote convergence in distribution of distribution functions and random variables respectively.

**Proposition 4.2** Suppose that $H_n \overset{\psi}{\Rightarrow} H$, and that $p_{L,n} \rightarrow p_L$, $p_{R,n} \rightarrow p_R$. Then

$$\beta_* \leq \lim \beta_{RL}^{(n)} \leq \lim \beta_{RL}^{(n)} \leq \beta^*.$$ 

**Proof.** We prove that $\beta_* \leq \lim \beta_{RL}^{(n)} \leq \lim \beta_{RL}^{(n)} \leq \beta^*$. It is enough to check, for example, that $\lim h_*(b) > h_*(b) > 0$ for any $b < \beta_*$, for this implies that ultimately $\beta_{RL}^{(n)} < b$. 
Appendix 1 shows that $\mathcal{L}(Y - \hat{b}X \mid X \geq x_{R,n}, H_n) \rightarrow \mathcal{L}(Y - bX \mid X \geq x_R, H)$ together with the corresponding version for $x_{L,n}$. The result is thus a consequence of the lower (upper) semi-continuity of the function $\text{lomed}(\cdot)$ (resp. $\text{uim}(\cdot)$) for weak convergence of distribution.

Remark 1. Consequently, if $\beta_{RL}(H)$ is uniquely defined and $H_n \overset{\text{d}}{\rightarrow} H$, (which occurs with probability one if $H_n$ is the empirical measure of i.i.d. observations from $H$)

then $\beta_{RL}(H_n) \rightarrow \beta_{RL}(H)$. In particular, if $H$ lies in the family of linear regression models $H_{\alpha} = \{H: Y = \alpha_0 + \beta_0 X + \varepsilon; \varepsilon, X \text{ independent}\}$ and either $0$ is the unique median of $\varepsilon$, or $X$ has an everywhere positive density, then $\beta_{RL}$ is strongly consistent for $\beta_0$. Note that the conditions on $x_{L,n}$ and $x_{R,n}$ permit choice of the partition points based on the sample itself.

Remark 2. As with Wald's estimator, the resistant line may be fitted to any bivariate scatter plot, including those in which the data are believed to follow an errors-in-variables model. Issues of identifiability and consistency arise just as for Wald's estimator. We mention without proof only one such result in the spirit of Neyman and Scott (1951, Theorem 2). Let $(X_i, Y_i)$, $i = 1, \ldots, n$ be a sample from the structural model $X = U + \eta$, $Y = \beta U + \varepsilon$ where $U, \eta$, and $\varepsilon$ are independent, and $\eta$ and $\varepsilon$ have smooth positive densities. Let the convex hull of the support of $\eta$ be the interval $[\mu, \nu]$, and $L = \{X_i; X_i \leq x_L\}$, $R = \{X_i; X_i \geq x_R\}$. Then $\beta_{RL}$ is a consistent estimator of $\beta$ if and only if $P(x_L - \nu < U < x_L - \mu) = P(x_R - \nu < U < x_R - \mu) = 0$. 

4.2 Influence Function. Several authors (Mallows 1973, Huber 1973, Velleman and Ypelaar 1980, Krasker and Welsch 1982) have noted the importance of bounding the influence function of a regression estimator in both \( y \) and \( x \). The resistant line influence function is so bounded. Therefore, in contrast to, say \( L_1 \) regression, \( \beta_{RL} \) is not unduly affected by a data point with an extreme \( x \)-value. Let \( H_\delta = (1 - \delta)H + \delta \varepsilon(x,y) \), where \( \varepsilon(x,y) \) is a point probability mass at \((x,y)\). The influence function of \( \beta_{RL} \) is defined by

\[
\text{IC}(\beta_{RL}, H, (x,y)) = \lim_{\delta \to 0} \frac{\beta_{RL}(H_\delta) - \beta_{RL}(H)}{\delta}.
\]

Suppose that \( H \in \mathcal{H}_r \)

\[
\mathcal{H}_r = \{ H : Y = \alpha + \beta X + \varepsilon, \text{ for some } \alpha, \beta \in \mathbb{R}, \varepsilon \text{ and } X \text{ independent} \},
\]

and assume that \( \varepsilon \) has a density \( g \) positive at its median, 0, and that \( E|X| < \infty \). Let \( \mu_L = E(X|X \leq x_L) \) and \( \mu_R = E(X|X \geq x_R) \). Then Appendix 2 shows that under mild smoothness conditions on the distribution of \( X \),

\[
\text{IC}(\beta_{RL}, H, (x,y)) = \begin{cases} 
+ c_H \psi_0(y - \beta_0 x - \alpha_0) & \text{for } x \geq x_R, \\
(-) & \text{for } x \leq x_L, \\
0 & \text{for } x_L < x < x_R,
\end{cases}
\]

where \( \psi_0(x) = I\{x > 0\} - I\{x < 0\} \) is the influence function of the median and \( c_H = [2g(0) \rho_L(\mu_R - \mu_L)]^{-1} \).

Qualitatively, (4.1) accords with intuition from dual plots. The position of the point \((\beta_0, y - \beta_0 x)\) with respect to the appropriate (left if \( x \leq x_L \), right if \( x \geq x_R \)) median trace evaluated at \( \beta_0 \) determines the direction of change in slope. The analogy between this influence function and that of the median in location problems is pursued further in Section 5.

For an arbitrary bivariate distribution \( H \), the form of the influence function remains the same except that \( c_H \) takes on two values (given in Appendix 2) according as \( x \geq x_R \) or \( x \leq x_L \).
4.3 Asymptotic Normality. A standard heuristic argument (e.g. Huber, 1981) asserts that if \( H_n \) denotes the empirical distribution function of a sample \((X_i, Y_i)\) of size \( n \) from \( H \), then the statistic \( \beta_{RL}(H_n) \) is asymptotically normal with mean \( \beta_{RL}(H) \) and asymptotic variance given by \( E_H[IC(\beta_{RL}, H, (X,Y))]^2 \). If \( H \in \mathcal{H} \), then we use (4.1) to obtain

\[
\text{Var} \sqrt{n} [\beta_{RL}(H_n) - \beta_{RL}(H)] = \frac{1}{2p_L 2^2 (0)(\mu_R - \mu_L)^2}
\]

We state a version of the asymptotic normality theorem for the resistant line slope estimator, proved in Appendix 3 for the more general setting of Section 5.

Theorem 4.3. Suppose that \((X_i, Y_i), i = 1, \ldots, n\) is an i.i.d. sample from \( H \in \mathcal{H} \). Suppose also that \( E|X| < \infty \), \( \epsilon \) has a continuous positive density at its median, 0, \( p = P_{n_L/n + p_L, n_R/n + p_R} \) and (for convenience only) \( p_L = p_R \). Then

\[
\sqrt{n} (\beta_{RL} - \beta) \overset{d}{\to} \text{Gau}(0, \{2p_L(\mu_R - \mu_L)^2 2^2 (0)\}^{-1})
\]

Remarks. 1) If \( p_L \neq P(X < x_L) \) for some \( x_L \), then \( \mu_L \) is interpreted as \( p_L \int_0^{x_L} F^{-1}(p)dp \), where \( F^{-1}(p) \) denotes the left continuous inverse of \( F \), the marginal distribution of \( X \). Further details are given in Appendix 3.

2) A similar result is obtained by Kildea (1981) in the context of weighted medians when the \( X_i \) are considered as fixed rather than random.

3) If a heteroscedastic model of the form \( Y = \alpha + \beta X + \epsilon/\sigma(X) \) obtains, where \( X \) and \( \epsilon \) are still assumed independent, then the argument of Appendix 3 extends (with appropriate modifications of the regularity conditions) to
show that \( \sqrt{n} \left( \beta_{RL} - \beta \right) \) is asymptotically centered Gaussian with variance
\[
\{2p_L(\tilde{m}_R - \tilde{m}_L)^2 + \frac{2}{\gamma^2(0)}\}^{-1} \{\tau_L^{-2} + \tau_R^{-2}\}/2, \quad \text{where } \tau_L = E(\sigma(X)|X \leq x_L),
\]
\( \tilde{m}_L = E(\sigma(X)|X \leq x_L)/\tau_L \) and \( \tau_R, \tilde{m}_R \) are defined correspondingly.

4) Inspection of the proof reveals that of all choices of sets \( L \) and \( R \) (with \( L < R \)) such that \( n_L/n + p_L \) and \( n_R/n + p_R \), the sets composed of the smallest \( n_L \) and largest \( n_R \) \( X_i \)'s are optimal (in the sense of minimizing asymptotic variance).

In the special case that \( L = \{X_i: X_i \leq x_L\} \) and \( R = \{X_i: X_i \geq x_R\} \), a direct proof is easy. The event \( \{\sqrt{n}(\beta_{RL} - \beta_0) > c\} \) is equivalent to \( \{Z_{n_R} = \text{med}(\varepsilon_i - cn^{-1/2}X_i) > \text{med}(\varepsilon_i - cn^{-1/2}X_i) = W_{n_L}\} \).

The special form of \( L \) allows \( W_{n_L} \) to be approximated to order \( o(n^{-1/2}) \) by the median \( W_m \) of \( m = [np_L] \) values of \( \varepsilon_i - cn^{-1/2}X_i \), these being regarded as i.i.d copies of \( \varepsilon - cn^{-1/2}X \), where \( \xi(X^L) = \xi(X|X \leq x_L) \). Similarly one obtains \( Z_m \), independent of \( W_m \), corresponding to the right group. Standard asymptotic theory for medians of i.i.d. variables extends to show that \( \sqrt{m}(W_m, Z_m) \sim N(-c(\mu_L, \mu_R), [\gamma^2(0)]^{-1}I_2) \). Thus \( P(\sqrt{n}(\beta_{RL} - \beta_0) > c) = P(Z > W) + o(1) \) and this leads to the claimed conclusion.

Generally, however, the groups \( L \) and \( R \) are not chosen with reference to fixed values \( x_L \) and \( x_R \), but rather using sample quantiles. To cover these cases it is convenient to use the Berry-Esseen method of Appendix 3.
4.4 Asymptotic Efficiency

Given the marginal distribution $F$ of $X$, the partition point $x_L$ (and from symmetry $x_R$ also) can be determined to minimize asymptotic variance by maximizing $p_L(\mu_R - \mu_L)^2$. If $F$ is symmetric about 0 and $p_L = p_R$, then $\mu_R = -\mu_L$ and it suffices to maximize

$$h(x_L) = \left(\int_{-\infty}^{x_L} s \, dF(s)\right)^2 / F(x_L).$$

If $E X^2 < \infty$, then the Cauchy-Schwartz inequality shows that $h(x_L) \to 0$ as $x_L \to \infty$. However, if the tails of $\mathcal{F}$ vary as $|x|^{-\alpha}$ for $\alpha \leq 2$, then $h(x)$ fails to converge to zero as $x \to -\infty$. (c.f. the remarks following (4.3).) When $E X^2 < \infty$ and $F$ has a density $f$, any critical point $x_0 \in (-\infty, 0)$ of (4.2) satisfies $\mu_L(x_0) = 2x_0$, and a sufficient condition for $x_0$ to maximize (4.2) is that $f$ be unimodal and $F(x_0) \geq 1/6$.

Specific cases were considered by Nair and Shrivastava, Bartlett, Barton and Casley, and Gibson and Jowett, who obtained $2\sigma^2 F(x_L)(\mu_R - \mu_L)^2$ as the variance of the Wald estimator (1.1) under normal theory assumptions. We list a couple of these cases and remark that Section 5 shows that asymptotically, the optimal choice of division points depends only on the marginal distribution of $X$ and not on the particular $M$-estimator applied to the residuals. If $X$ is uniformly distributed on some interval or places equal mass at equally spaced points, then (4.2) is maximized by $F(x_L) = 1/3$, which lends support to the choice of "thirds" as the partition points. If $F(x) = \phi(x)$, then (4.2) becomes $\phi^2(x_L)/\phi(x_L)$ (where $\phi(x) = \phi'(x)$), which is maximized by $x_L = -.6121$, corresponding to $\phi(x_L) = .2702$, and yields the "27%-rule" described in the introduction.
We now consider the asymptotic relative efficiency of the resistant line estimator for models in $\mathcal{H}_r$. Let $\{T_n\}$ be any sequence of regression invariant estimates of $\beta$ based on samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ of size $n$; i.e., $T_n(X_1, \alpha + \beta X_1 + \varepsilon_1) = T_n(X_1, \varepsilon_1) + \beta$.

The local asymptotic minimax results of Hajek (1972, Theorem 4.2) imply that for appropriately smooth models in $\mathcal{H}_r$

$$\liminf_{n \to \infty} n E_{\alpha, \beta} \left( T_n - \beta \right)^2 \geq \left\{ \Var X E[g'(\varepsilon)/g(\varepsilon)]^2 \right\}^{-1} \sigma_{X, \varepsilon}^2,$$

and that equality can occur only if $\sqrt{n}(T_n - \beta)$ is asymptotically $N(0, \sigma_{X, \varepsilon}^2)$. It is therefore natural to consider the ratio (c.f. Section 6)

$$(4.3) \quad \sigma_{X, \varepsilon}^2 / [\text{AsVar}_{RL}^\beta] = \{ p_L(\mu_R - \mu_L)^2/2 \Var X \} \{ 4g^2(0)/I_g \},$$

where $I_g = E[g'(\varepsilon)/g(\varepsilon)]^2$ is the Fisher information of the density $g$.

Clearly (4.3) can be made arbitrarily small by appropriate choice of $g$.

This situation can, in principle, be remedied by replacing the median by an (adaptive, efficient) estimate of location (c.f. Section 5). However the ratio depending on the $X$ distribution can also be made arbitrarily small by taking densities for $X$ with tails of order $x^{-(3+\varepsilon)}$ for $\varepsilon > 0$. For example, if

$$f_X(x) = \frac{1}{2}(\alpha - 1)x^{-\alpha} I\{|x| > 1\},$$

then $\sup_{xL < 0} p_L(\mu_R - \mu_L)^2/(2 \Var X) = (\alpha - 3)(\alpha - 1)(\alpha - 2)^{-2}/4$. Thus there can be a loss of efficiency when the carrier has a distribution with very heavy tails.

In the familiar case in which $\{\varepsilon_i\}$ are i.i.d. $N(0,1)$, the least squares estimate $\hat{\beta}_{LS} = \Sigma(x_1 - \bar{x})y_1/\Sigma(x_1 - \bar{x})^2$ is asymptotically efficient, with $\text{AsVar}_{LS}^\beta / \text{AsVar}_{RL}^\beta = p_L(\mu_R - \mu_1)^2/(\pi \Var X)$. If $X$ is uniform, or equispaced, then the Gaussian efficiency of the resistant line (with $p_L = 1/3$) is 56.6% while if $X$ is $N(0,1)$, and $p_L = 1/3$, the efficiency becomes 50.5% (rising to 51.6% if $p_L = .27$). We note
that efficiencies of greater than 50% are considered adequate for exploratory data analytic methods. The corresponding efficiencies for Brown-Mood regression \( (p_L = 1/2) \) are 47.8% and 40.5% respectively, so that neglecting an appropriate middle group of observations increases efficiency by a factor of about one-fifth.

§5. THE CLASS OF \( \psi \)-RESISTANT LINES

A general class of regression estimates that includes the resistant line, Wald-Nair-Shrivastava-Bartlett, and Brown-Mood methods can be defined by replacing the medians applied to the residuals in the definition of the resistant line with an M-estimator of location. This class is introduced because of the substantial improvements in efficiency that can be achieved without significantly affecting breakdown or (in some cases) computational expense. A similar treatment can in principle be based on \( L \)- or \( R \)-estimators.

Let \( T_n(Z_1) = T_n(Z_1, \ldots, Z_n) \) be an M-estimator of location with monotone influence function \( \psi \). (We make some remarks below concerning redescending \( \psi \)). As before, suppose that bivariate data \( (x_i, y_i) \), \( i = 1, \ldots, n \) are available. Let \( T_L(Z_i) \) and \( T_R(Z_i) \) denote \( T \) applied to those \( Z_i \) with indices \( i \) corresponding to the \( n_L \) smallest and \( n_R \) largest \( X_i \) respectively. Define \( \beta_\psi = \beta_\psi(T_n) \) as the zero of the function

\[
(5.1) \quad h(b) = T_R(y_i - bx_i) - T_L(y_i - bx_i),
\]

The monotonicity of \( \psi \) and translation invariance of \( T \) ensure that \( h \) is monotone decreasing with slope at most \( \frac{x(n_L) - x(n-n_R+1)}{n(n-1)} \).
which assures the existence of a unique solution to (5.1). Locating the zero of such a function is possible using (say) the ZERION algorithm as described in Section 2 and thus requires the same order of computation effort as computation of $\mathbf{T}$. The gross error breakdown value of $\beta_\psi$ is one third of that of the corresponding $M$-estimator if $n_L = n_R = n/3$.

Remark In practice, one would have to estimate the scale of the residuals also, and (5.1) might for example then be replaced by

$$
(5.2) \quad T_L \left( \frac{y_i - bx_i}{\widehat{\sigma}} \right) = T_R \left( \frac{y_i - bx_i}{\widehat{\sigma}} \right),
$$

where $\widehat{\sigma}$ is a robust estimate of scale which, in homoscedastic situations can be based on all the data. In the presence of heteroscedastic errors $\widehat{\sigma}$ could be replaced by $\widehat{\sigma}_L$ and $\widehat{\sigma}_R$ estimated separately in the two groups.

To define the $\psi$-resistant line as a functional on theoretical distributions $H$ of the pair $(X,Y)$, we proceed as follows. Let $x_L < x_R$ be lower $p_L^{th}$ and upper $p_R^{th}$ percentiles of the distribution of $X$, and write $F_{L,b} = F_{L,b}(H)$ and $F_{R,b} = F_{R,b}(H)$ for the respective distribution functions of $Y - b X \mid X \leq x_L$ and $Y - b X \mid X \geq x_R$. Let $t(F)$ be the $M$-estimator of location obtained from a monotone increasing function $\psi(x)$ by solving $\lambda(F,t) = \int \psi(x - t) \, dF(x) = 0$. The $\psi$-resistant line estimator $b_\psi(H)$ is defined as the (unique, since $\psi$ is monotone) solution to
\[ h(b) = T_R(Y-bX) - T_L(Y-bX) = 0 \]

where \( T_R(Y-bX) \) is a mnemonic for \( t(F_R,b) \), etc.

The consistency of \( \beta_\psi \) can be discussed in a fashion exactly analogous to that of \( \beta_{RL} \). To replace \( \text{hmmed} F \) and \( \text{iomed} F \), we use

\[ T^*(F) = \inf\{t: \lambda(F,t) < 0\} > T^*_R(F) = \sup\{t: \lambda(F,t) > 0\} \]

The statement and proof of Proposition 4.2 remain valid (with the obvious changes) under the extra technical restriction that \( \psi \) be bounded. This facilitates the proof that \( T^*_R(F) \) and \( T^*(F) \) are semi continuous. (See Appendix 1.)

In particular, if \( Y = \alpha_0 + \beta_0 X + \varepsilon \) with \( X \) and \( \varepsilon \) independent and \( T(\varepsilon) \) uniquely defined, then \( \beta_\psi \) is strongly consistent for \( \beta_0 \).

If \( \psi \) is bounded, the \( \psi \)-resistant line retains the property of bounded influence in both \( x \) and \( y \) possessed by the ordinary resistant line. Let \( H_0 = (1-\delta)H + \delta \varepsilon_{[x,y]} \). Under appropriate regularity conditions, the argument in Appendix 2 shows that if \( H \in H_R \) and \( \varepsilon \) has density \( g \) with \( t(\varepsilon) = 0 \), and \( E|X| < \infty \), then

\[
(5.3) \quad IC(\beta_\psi,H,(x,y)) = \begin{cases} 
+ \frac{\psi(y - \beta_0 x - \alpha_0)}{p_L(\mu_R - \mu_L) \int_{-\infty}^{\infty} g \, d\psi} & \text{as } x \geq x_R \\
(-) & \text{as } x \leq x_L \\
0 & \text{if } x_L < x < x_R.
\end{cases}
\]

Thus the influence function is (for \( x \) values in the outer thirds), a constant multiple of the influence function of the location estimate \( T \).
If the observations \((X_i, Y_i)\) come from a model \(H \in \mathcal{H}_L\) then asymptotic normality of \(\beta_{\psi}\) can be established under fairly general conditions on \(\psi, X, \varepsilon\) and the partitioning rule. Let \(V_T\) denote the asymptotic variance of \(\{T_n\}\). Then

\[
(5.4) \quad \mathcal{L}\{\sqrt{n} \beta_{\psi} - \beta\} \overset{d}{\rightarrow} \text{Gau}(0, 2p_L^{-1}(\mu_R - \mu_L)^2, V_T).
\]

The regularity conditions and proof are detailed in Appendix 3. Essentially we require that \(\psi\) be monotone, \(\psi(\varepsilon - t)\) have third moments, \(\lambda(t) = \mathbb{E} \psi(\varepsilon - t)\) be differentiable at \(t = 0\) and either \(\mathbb{E} \lambda^2 < \infty\) or \(\lambda(t) = O(\vert t \vert)\) as \(t \to \infty\). The hypotheses cover the extremes in which \(T_n\) is the median or the mean (at least for most \(\varepsilon\)). The number of points \(n_L\) (resp. \(n_R\)) in the left and right groups need only to be chosen so that \(\frac{n_L}{n} \sim p_L(p_R)\). The proof conditions on the observed \(X\)'s and uses a Berry-Esseen theorem to obtain a uniform normal approximation.

It is apparent from (5.4) that the asymptotic relative efficiency of two \(\psi\)-resistant lines is the relative efficiency of the two corresponding location estimators, so that the choice of \(\psi\) could be tailored to the distribution of \(\varepsilon\), if known (approximately). For example, use of \(\psi(x) = x(\vert x \vert < k) + k \text{ sign } x(\vert x \vert > k)\) (which is minimax over contamination neighborhoods of the Gaussian distribution (Huber 1964)) with \(k = 1.5\) raises the A.R.E. with respect to least squares (for uniform \(X\) and Gaussian \(Y\)) from 57% to 85%, while retaining good breakdown behavior.
Reduced The use of non-monotone ("descending") $\psi$-functions has been advocated to afford greater protection from outliers. However, solutions are no longer guaranteed to exist or be unique and care is required in estimating the scale, $\sigma$ (c.f. 5.2). Without unimodality assumptions on the error distribution, the resulting estimators need not even be consistent (Diaconis and Freedman, 1982). Nevertheless, such estimators perform well in practice. We have included a $\psi$-resistant line using a one-step biweight (see, e.g. Mosteller and Tukey, 1977) in the experiment discussed in Section 6. A heuristic argument (outlined in Appendix 4) shows, under regularity conditions ensuring asymptotic uniqueness of the solution, that the asymptotic normality result (5.4) is preserved for a wide class of non-monotone $\psi$-functions. Further, the Gaussian efficiency of the biweight is close to that of the optimal "hyperbolic tangent" descending estimators (Hampel, Rousseeuw, Rochetti; 1981).
§6. SMALL SAMPLE EFFICIENCIES

We studied small-sample properties of ten simple regression methods under the standard linear model assumptions, \( \mathcal{H}_r \) (section 4.2), using a Monte Carlo experiment. The experiment design was a complete crossing of three sample sizes \((n = 10, 22, \text{ and } 40)\) by two x configurations (fixed equispaced and random Gaussian \((0,1)\) by three error distributions (Gaussian \((0,1)\), a mixture of 90% Gau\((0,1)\) and 10% Gau\((0,9)\), and "slash"). The Gaussian mixture is one for which the mean and median have approximately equal asymptotic variances. The "slash" distribution, generated as a unit Gaussian variate divided by a Uniform \((0,1)\) variate, has infinite variance but is not as peaked as the Cauchy density.

The methods studied fall into three categories. (I) non-resistant (breakdown = 0) methods with computing effort \(O(n)\): Least Squares (LS), Least Absolute Residual (LAR) (algorithm from IMSL, 1980), and the partitioned line method of Nair, Shrivastava, and Bartlett (NSB).

(II) Resistant methods with computing effort \(O(n^2)\): Repeated Median (RM) and Sen’s version of the global median pairwise slope (SEN). (III) Resistant methods with computing effort \(O(n)\):

Brown-Mood regression (BM), Resistant line (RL), \(\psi\)-Resistant line using one biweight step \((c=6)\) to improve each median and the global median absolute deviation from the median (MAD) as a scale estimate, the \(\psi\)-Resistant line using one biweight step but a local MAD computed only within each partition (RLBW3), and the \(\psi\)-Resistant line using a fully iterated Huber \(\psi\)-function and local scale (RLHU3). All of the partition-based lines (NSB, BM, RL, RLBW, RLBW3, RLHU3) were computed using a modified version of the code published by Velleman and Hoaglin (1981) for the resistant line. Technical details of the experiment are reported in Appendix 6.
The experiment employed a "Gaussian over independent" variance reduction"swindle". (See, for example, Simon (1976), or Goodfellow and Martin (1976).) We also developed a new swindle for this study to improve performance on non-Gaussian error distributions. The swindle is defined in Appendix 5 and discussed more extensively in Johnstone and Velleman (1983). The swindles increase the effective number of replications, typically by factors of 3 to 8. (That is, the standard errors of our efficiency estimates are as small as those of a naive simulation experiment with between 3 and 8 times the number of replications.) Occasionally (especially for the more efficient estimators) swindle gain factors reached 160.

Table 2 presents the small-sample efficiencies for fixed equi-spaced x. The efficiencies have been computed relative to the Cramer-Rao lower bound for each situation, so 100% efficiency will only be attainable in the Gaussian case here. In this sense, the efficiencies reported for the mixed Gaussian and slash error distributions are conservative. Few small-sample results have been published for any of these methods. Siegel (1982) reports efficiencies for the repeated median (RM) at sample size 10 and 20 (for fixed x: .69 and .73; for Gaussian X, .53 and .61, respectively) that agree with ours.

Except for slash distributed errors at sample size 10 where all estimators predictably did badly), the resistant line performs well, staying above the nominal 50% value for exploratory methods and usually exceeding 60% efficiency. The increase in efficiency over asymptotic values it exhibits for Gaussian and mixed Gaussian errors at smaller sample sizes is a pleasant benefit in a technique often computed by hand for small data sets. It appears to derive from the similar behavior of the...
### Fixed Equispaced x

#### 2. Efficiency Relative to Cramer-Rao Lower Bound

(Standard error)

<table>
<thead>
<tr>
<th></th>
<th>Gaussian</th>
<th></th>
<th>Mixed Gaussian</th>
<th></th>
<th>Slash</th>
<th></th>
<th>Triefficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>22</td>
<td>40</td>
<td>80</td>
<td>10</td>
<td>22</td>
<td>40</td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LS</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LAR</td>
<td>61.9(.4)</td>
<td>63.3(.5)</td>
<td>63.6(.8)</td>
<td>63.7</td>
<td>36.2(1.3)</td>
<td>60.9(1.7)</td>
<td>68.0(1.1)</td>
</tr>
<tr>
<td>NSB</td>
<td>89.2(.1)</td>
<td>89.1(.7)</td>
<td>89.2(.3)</td>
<td>89</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RM</td>
<td>69.7(.4)</td>
<td>73.0(.4)</td>
<td>72.3(.7)</td>
<td>72.3</td>
<td>37.0(.9)</td>
<td>59.4(.7)</td>
<td>65.9(1.1)</td>
</tr>
<tr>
<td>SEN</td>
<td>87.9(.2)</td>
<td>91.4(.2)</td>
<td>93.0(.3)</td>
<td>95.5</td>
<td>35.1(1.1)</td>
<td>56.0(.7)</td>
<td>62.5(1.1)</td>
</tr>
<tr>
<td>III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RM</td>
<td>50.0(.4)</td>
<td>49.0(.6)</td>
<td>49.2(.8)</td>
<td>48</td>
<td>51.2(.5)</td>
<td>51.5(.6)</td>
<td>53.1(.8)</td>
</tr>
<tr>
<td>RL</td>
<td>65.3(.4)</td>
<td>60.5(.5)</td>
<td>58.8(.8)</td>
<td>57</td>
<td>64.0(.6)</td>
<td>63.6(.5)</td>
<td>63.0(.8)</td>
</tr>
<tr>
<td>RLBW3</td>
<td>73.5(.3)</td>
<td>68.6(.5)</td>
<td>66.9(.7)</td>
<td>82.3</td>
<td>65.6(.6)</td>
<td>69.4(.5)</td>
<td>70.8(.7)</td>
</tr>
<tr>
<td>RLBW</td>
<td>76.2(.4)</td>
<td>77.8(.4)</td>
<td>78.2(.6)</td>
<td>80</td>
<td>61.8(.6)</td>
<td>73.2(.5)</td>
<td>77.9(.6)</td>
</tr>
<tr>
<td>CRLB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>0.10</td>
<td>9.9454</td>
<td>0.025</td>
<td>0.025</td>
<td>0.1256</td>
<td>0.0571</td>
<td>0.0314</td>
</tr>
</tbody>
</table>

Key:
- **I**: Breakdown = 0%, Cost = O(n)
- **II**: Breakdown = 20%(Sen), 50%(RM), Cost = O(n^2)
- **III**: Breakdown = 25%(RM), 16%(others), Cost = O(n^2)

Results for n = 10 Based on 10,000 replications
Results for n = 22 Based on 5,000 replications
Results for n = 40 Based on 2,000 replications

- **LAR**: Least Absolute Residual
- **NSB**: Nair-Shrivastava (1942); Bartlett (1949)
- **RM**: Repeated Median; Siegel (1982)
- **SEN**: Sen (1968); Theil (1950)
- **BM**: Brown-Mood (1950)
- **RL**: Resistant line; Tukey (1971)
- **RLBW3**: ψ-Resistant line, One-step biweight, local MAD, c = 6
- **RLBW**: ψ-Resistant line, Huber ψ (e.g. Huber 1981), local MAD, c = 1.5
- **RLBW**: ψ-Resistant line, one-step biweight, global MAD, c = 6
median, which is 100% efficient (relative to the mean) in Gaussian samples of one or two and declines in efficiency to an ARE of $2/\pi$. The reverse trend for slash errors appears to be due to breakdown of the median for the samples of $n/3$ found in the outer partitions of small samples.

The resistant line is thus appropriate for hand calculation on small samples with the caveat that data sets should be a bit larger (at least 20, preferrably 40) if very heavy-tailed error distributions are suspected.

Among the estimators ordinarily found only with the aid of a computer, the biweight $\psi$-resistant line (RLBW) performs quite well. It has the highest maximim efficiency ("tri efficiency") at sample sizes 22 and 40 among methods studied here. It clearly dominates at moderate sample sizes among estimators with bounded influence and inexpensive (0(n)) computation order. Because the biweight is a single step improvement of the medians in the resistant line, this estimator can be computed by hand if a programmable calculator is available. For its favorable combination of properties, we prefer this estimator among those studied here.

When the scale estimate is computed only from the local partition (RLBW3, RLHU3) we can relax the assumption of homoscedastic errors, but we will pay about 5% in efficiency if that assumption is in fact true.

The least absolute residual (LAR) line performs similarly to the RLBW. Its primary weakness is that it is not resistant to the effects of large fluctuations at extreme x values.
The method of Theil and Sen, and Siegel's repeated median reward the $O(n^2)$ computing expense with higher Gaussian efficiencies and (for the repeated median) higher breakdown. However, their efficiency advantage is eroded at the more severe error distributions, and their computations are too complex for hand calculation and too expensive to be applied to large data sets or generalized to more than 3 or 4 dimensions.

The efficiencies reported in Table 3 for random Gaussian X follow the same general pattern, but are somewhat lower than the corresponding fixed-X efficiencies. For the resistant line family estimators, some of the loss is due to the suboptimal placement of $x_L$ and $x_R$ at the 33% rather than the 27% points in X.

§7. MULTIPLE REGRESSION

Partition-based lines can be extended to the multiple regression model in a variety of ways: no definitive answers have yet emerged. Gibson and Jowett (1957b) generalize Bartlett's (1949) method to two carriers, but their method is not practical for more than two carriers. Andrews (1974) proposes building multiple regressions by "sweeping" with a simple regression related to the resistant line. Beaton and Tukey (1974) sweep in a different order to build a multiple regression from a simple resistant regression. Emerson and Hoaglin (1983b) follow Andrew's sweeping order to build a multiple regression from the resistant line. Substituting more efficient \psi-resistant lines in these sweep methods will improve their efficiency with almost no increase in computing effort.
<table>
<thead>
<tr>
<th>Random Gaussian X</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 10 )</td>
<td>114.0 (1.3)</td>
<td>105.0 (1.3)</td>
<td>102.6 (1.3)</td>
</tr>
<tr>
<td>( I )</td>
<td>63.0 (1.6)</td>
<td>62.6 (1.6)</td>
<td>66.2 (1.6)</td>
</tr>
<tr>
<td>( II )</td>
<td>79.7 (1.6)</td>
<td>80.1 (1.6)</td>
<td>78.7 (1.6)</td>
</tr>
<tr>
<td>( III )</td>
<td>56.1 (1.6)</td>
<td>61.7 (1.6)</td>
<td>64.2 (1.6)</td>
</tr>
</tbody>
</table>

### Residual Distribution

| Residual Distribution | Gaussian
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 20 )</td>
<td>22.0</td>
</tr>
<tr>
<td>( n = 40 )</td>
<td>22.0</td>
</tr>
</tbody>
</table>

### Mixed Gaussian

<table>
<thead>
<tr>
<th>Mixed Gaussian</th>
<th>40.0</th>
</tr>
</thead>
</table>

### Interference

<table>
<thead>
<tr>
<th>Interference</th>
<th>22.0</th>
</tr>
</thead>
</table>
One extrapolation of equation (2.1) requires the residuals from a multiple regression fit to have zero \( \psi \)-resistant line slope against each carrier: For \( p \) carriers (including the constant carrier) and coefficient vector, \( \beta \), define \( e = y - X\beta \). We seek to satisfy the \( p \) simultaneous conditions

\[
\text{T}(e_i^*) = \text{T}(e_i), \quad j = 2, \ldots, p
\]

\[
t_{i \in L_j} \quad i \in R_j
\]

\[
T(e_i) = 0
\]

for \( M \)-estimator, \( T \), and partition sets \( (L_j, R_j) \) defined by the rank ordering of carrier \( x_j \). Other possibilities include first orthogonalizing the carriers in some (possibly resistant) fashion.

The experience of Emerson and Hoaglin indicates that (7.1) can be approximated inexpensively. No \( \psi \)-resistant multiple regression will compete in efficiency with, for example, the bounded influence regressions of Krasker and Welsch (1982). However, they provide a less expensive exploratory-quality bounded influence robust multiple regression estimate, which will be sufficient for many applications. When greater efficiency is desired, the \( \psi \)-resistant multiple regression provides a better starting value for the Krasker-Welsch iteration than least squares or least absolute residual regression.
A.1 ADDENDA TO CONSISTENCY RESULTS

We show first that under the condition of Proposition 4.2,
\( \mathcal{L}(Y - bX | X \leq x_L, n, \mathbb{H}_n) \rightarrow^d \mathcal{L}(Y - bX | X \leq x_L, \mathbb{H}) \). Let \( z \) be a continuity point of the latter, and let \( A = \{Y - bX \leq z\} \). By hypothesis \( \mathbb{H}_n(X \leq x_L, n) = p_{L,n} + p_L = H(X \leq x_L) \), so that it remains to check that \( \mathbb{H}_n(A, X \leq x_L, n) \rightarrow H(A, X \leq x_L) \). For this purpose write \( x_n \) and \( x \) for \( x_{L,n} \) and \( x_L \),

\[ H_n(A, x_n) \text{ for } H_n(A \cap \{(-\infty, x_n] \times \mathbb{R}\}) \text{ and simply } H_n(x) \text{ when } H_n(A) = 1. \]

Using right continuity of \( H(x) \), and given \( \varepsilon > 0 \), choose \( y > x, y \)
a continuity point of \( H \), so that \( H(x) \leq H(y) < H(x) + \varepsilon \). From this we obtain

\[ |H_n(A, x_n) - H(A, x)| \leq |H_n(A, y) - H_n(A, x_n)| + |H_n(A, y) - H(A, y)| + |H(A, y) - H(A, x)|. \]

The first term on the right side is bounded by \( H_n([-\infty, y] \times (-\infty, x_n] \times \mathbb{R}) \)
where \( [-\infty, y] \times \mathbb{R} \) denotes \( \{\infty, y\} \times \mathbb{R} \). The latter equals \( H_n(x_n) - H_n(y) \)
and converges to \( |H(x) - H(y)| < \varepsilon \) as \( n \to \infty \). The second term vanishes asymptotically since \( A \cap \{\infty, y\} \) is a continuity set of \( H \) by assumption. Finally, the third term is bounded by \( |H(y) - H(x)| < \varepsilon \). Since \( \varepsilon \) is arbitrary it follows that \( H_n(A, x_n) \rightarrow H(A, x) \), as was required.

Secondly, we prove that \( T_*(F) = \sup \{t : \lambda(F, t) > 0\} \) is lower semi-
continuous if \( \psi \) is bounded and monotone. Suppose that \( F_n \rightarrow^d F \):
then \( \lambda(F, t) \rightarrow \lambda(F, t) \) for \( t \in \mathbb{R} \setminus N_F \), where \( N_F \) is a countable set possibly depending on \( F \). To see this, note that the monotone function

\[ x \mapsto \psi(x-t) \] is bounded and continuous on \( \mathbb{R} \setminus N_t \), where \( N_t \) is countable. If \( F(N_t) = 0 \), then \( t \notin N_F \). Now \( N_t = N_0 + t \) since \( \psi \) is translation invariant; and this implies that \( N_F \subset \{t : F(N_t) - F(N_0 + t) > 0\} \) is at most countable. It is now easy to check that \( \lim T_*(F_n) \geq T_*(F) = t_0 \), say.

Indeed, choose \( \varepsilon > 0 \) such that \( \lambda(F, t_0 - \varepsilon) \rightarrow \lambda(F, t_0 - \varepsilon) > 0 \); this implies that for all sufficiently large \( n \), \( T_*(F_n) \geq t_0 - \varepsilon \), and this suffices.
A.2 DERIVATION OF INFLUENCE FUNCTION

For brevity in what follows we shall assume that \( F_L, b(z), F_R, b(z) \) and \( \psi(z) \) are sufficiently smooth in \( b \) and \( z \) that the necessary differentiations and interchanges are valid. In the case of the \( \psi \) function corresponding to \( T(F) = \text{med}(F) \), the results will hold except on an easily identified null set.

The estimate \( b(\delta) = b(\psi(H_\delta)) \) is defined implicitly as the solution to \( h(\delta, b) = t_R(\delta, b) - t_L(\delta, b) = 0 \), where \( t_R(\delta, b) = t(F_R, b, \delta) \) and \( F_R, b, \delta = F_R, b(H_\delta) \) (and similarly for \( t_L(\delta, b) \)). In turn, \( t_R(\delta, b) \) is defined implicitly by

\[
g_R(\delta, b, t) = \int \psi(z-t) \, dF_R, b, \delta(z) = 0
\]

It follows by implicit differentiation that

\[
(A.1) \quad \text{IC}(b_\psi, (x, y)) = b'(0) = [-D_1 h/D_2 h](0, b(0)) = \left[ \frac{D_1 g_{L}}{D_3 g_{L}} - \frac{D_1 g_{R}}{D_3 g_{R}} \right] \left[ \frac{D_2 g_{R}}{D_3 g_{R}} - \frac{D_2 g_{L}}{D_3 g_{L}} \right]^{-1} (0, b(0), t_0),
\]

where \( t_0 = t_L(0, b(0)) = t_R(0, b(0)) \).

To evaluate \( (A.1) \), note that \( H_\delta = (1-\delta) \, H + \delta \varepsilon_{\{x, y\}} \) implies that

\[
F_R, \delta, b(t) = H_\delta(Y - bX \leq t | X > x_\delta) = (1 - \alpha_\delta^R) \, F_R, b(t) + \alpha_\delta^R \{ y - bx \leq t \},
\]

where we write \{A\} instead of \( I(\{A\}) \), and
\[ \alpha_\delta^R = \frac{\delta\{x \geq x_R\}}{(1-\delta)p_R + \delta\{x > x_R\}} \]

is the probability that a wild point is drawn given that it lies in R.

A similar expression holds for \( F_L, \delta, b \) and hence \( g_L(\delta, b, t) \).

Assume for now that \( H \in H_1 \) in order to simplify (A.1). In this case \( F_R, 0, \beta = F_L, 0, \beta = \xi(\varepsilon + \alpha) \), so it is easy to check that \( b_\psi(0) = \beta \) and \( t = t_R(0, b(0)) = t_L(0, b(0)) = t(\xi(\varepsilon + \alpha)) = \alpha \).

This implies that \( g_L(0, b(0), t) = g_R(0, b(0), t) \) and hence \( D_3g_R = D_3g_L \) at \((0, \beta, t_0)\). A calculation shows that \( D_3^R \alpha_\delta^R \big|_{\delta=0} = \{x \geq x_R\}/p_R \) and hence that \( D_3g_R(0, \beta, t_0) = \psi(y-\beta, x-\alpha) \{x \geq x_R\}/p_R \). Analogous results hold for \( D_2g_L \).

It remains to find \( D_2g_L \) and \( D_2g_R \). Assume now that \( \varepsilon \) has a density \( g \) and that \( E[X] < \infty \). Write \( F_R(x) = P(X \leq x | X \geq x_R) \) and \( \mu_R \) for \( E(X | X \geq x_R) \). If \( \psi(v)F_R, b(v + t) \to 0 \) as \( v \to -\infty \) and \( \psi(v)[1 - F_R, b(v + t)] \to 0 \) as \( v \to +\infty \), then

\[
\begin{align*}
g_R(0, b, t) &= \int \psi(v) dF_R, b(v + t) \\
&= \psi(0) - \int_{-\infty}^0 F_R, b(v + t) \psi(v) \, dv + \int_0^\infty [1 - F_R, b(v + t)] \psi(v) \, dv .
\end{align*}
\]

Thus \( D_2g_R(0, b, t) = -\int_{-\infty}^\infty d\psi(v) \int_{x_R}^{\infty} dF_R(z) z \psi((b-\beta)z + v + t) \), and

\[
D_2g_R(0, \beta, t_0) = -\mu_R \int_{-\infty}^\infty g(v) \psi(v) ,
\]

and a similar result holds for \( D_2g_L \). Inserting these expressions in (A.1) leads to the influence functions (4.1) and (5.3).
The influence function takes a more complicated form if $H \notin \mathcal{H}_L$, and we give only the result for the ordinary resistant line. In this case, a straightforward calculation shows that

$$IC(\beta_{RL}, H, (x, y)) = \frac{\psi(y-b_0x-t_0)}{2p_L} \left[ \frac{\{x \geq x_R\}}{f_{RL, b_0}(t_0)} - \frac{\{x \leq x_L\}}{f_{RL, b_0}(t_0)} \right] \left[ \frac{D_bF_{RL, b_0}(t_0)}{f_{RL, b_0}(t_0)} - \frac{D_bF_{RL, b_0}(t_0)}{f_{RL, b_0}(t_0)} \right]^{-1}$$

where $b_0 = b_{RL}(0)$, and $D_b$ denotes partial differentiation with respect to $b$. 
A.3 ASYMPTOTIC NORMALITY

We use freely the notation of Sections 4 and 5. Suppose that

\((X_i, Y_i)\) is an i.i.d. sample from \(H \in \mathcal{H}_r\) and impose the following

conditions on \(X, \varepsilon, \) and the \(\psi\) function corresponding to \(T_n^*\).

(A) \(\psi\) is monotone, with \(\psi(x) \geq 0\) for \(x > 0\) and \(\psi(x) \leq 0\) for \(x < 0\).

If \(T_n = T_n(Z_1, \ldots, Z_n)\) is any point in \([T_n^*, T_n^{**}]\), where

\(T_n^* = \sup\{t : \Sigma_1^n \psi(Z_i - t) > 0\}\) and \(T_n^{**} = \inf\{t : \Sigma_1^n \psi(Z_i - t) < 0\}\), then

\[(A.2) \quad \{\Sigma_1^n \psi(Z_i - w) < 0\} \subset \{T_n \leq w\} \subset \Sigma_1^n \psi(Z_i - w) \leq 0\} .

(B) (i) \(\lambda(t) = E\psi(\varepsilon - t)\) is continuously differentiable at \(t = 0\), with

\(\lambda(0) = 0\).

(ii) The function \(t \rightarrow E\psi^2(\varepsilon - t)\) is continuous at \(0\).

(C) Either \(EX^2 < \infty\) or \(\lambda(t) = O(|t|)\) as \(t \rightarrow \infty\).

(D) \(E|\psi(\varepsilon + w + cx)|^3 < \infty\) for all \(w, c \in \mathbb{R}\).

(E) \(n_L/n \overset{p}{\rightarrow} p_L, \quad n_R/n \overset{p}{\rightarrow} p_R\).

Under these assumptions, we show for the \(\psi\)-resistant line estimator

\(\hat{\beta} = \hat{\beta}_\psi\) that \(\sqrt{n}(\hat{\beta} - \beta_0)\) is asymptotically centered Gaussian with variance

\([p_L^{-1} + p_R^{-1}]\mu_R^{-1}\mu_L^{-1} \varepsilon^2/[\lambda'(0)]^2\). Here \(\mu_L = \int_0^{p_L} F^{-L}(p)dp\) \(p_L^{-1}\) and \(\mu_R = \int_{1-p_R}^{1} F^{-L}(p)dp\), where \(F^{-L}(p)\) is the left continuous inverse of

the distribution of \(X\). When \(p_L = P(X \leq x_L)\) for some \(x_L\), it may be shown that \(\mu_L = E(X \mid X \leq x_L)\), and similarly \(\mu_R = E(X \mid X \geq x_R)\) when

\(p_R = P(X \geq x_R)\).
Proof. From the monotonicity of \( h \) (see (5.1)) and the model
\[ Y = \alpha_0 + \beta_0 X + \varepsilon \]
we have
\[ \{ \sqrt{n}(\hat{\beta} - \beta_0) > c \} = \{ T_L(\varepsilon_1 \sim c n^{-1/2} X_1) < T_R(\varepsilon_1 \sim c n^{-1/2} X_1) \} . \]

We show \((T_L, T_R) = (T_L(\varepsilon_1 \sim c n^{-1/2} X_1), T_R(\varepsilon_1 \sim c n^{-1/2} X_1))\) converges in probability to a jointly normal distribution with mean \(-c(\mu_L, \mu_R)\) and covariance matrix \([\lambda'(0)]^{-2} \mathcal{E} \psi^2 . \text{diag}(p_{L}^{-1}, p_{R}^{-1})\), from which the result follows easily.

Conditional on \( X \), \( T_L \) and \( T_R \) are independent, so that
\[ P(\sqrt{n} T_L \leq w, \sqrt{n} T_R \leq z) \]
(A.3)
\[ = E[P(T_L(\varepsilon_1 \sim c n^{-1/2} X_1) \leq w/\sqrt{n}| X) P(T_R(\varepsilon_1 \sim c n^{-1/2} X_1) \leq z/\sqrt{n}| X)] . \]

Consider for now the first term, which in view of (A.2) may be sandwiched between

\[ P_L(X) = P[ \sum_{i \in L} \psi(\varepsilon_1 \sim c n^{-1/2}(w+cX_1)) < 0 | X] \]
(A.4)

and the corresponding expression with non-strict inequality. It will be seen that it suffices to consider (A.4). Apply the Berry Esseen theorem for independent r.v.'s (e.g. Feller (1971), p. 544) to approximate (A.4).

Let \( \bar{E} \) denote conditional expectation given \( X \), and set
\[ Z_{n,i} = \psi(\varepsilon_1 \sim n^{-1/2}(w+cX_1)) \], \( \mu_{n,i} = \bar{E}Z_{n,i} = \lambda[n^{-1/2}(w+cX_1)] \), \( \sigma_{n,i}^2 = \text{Var}(Z_{n,i}) \)
and \( \rho_{n,i} = \bar{E}|Z_{n,i} - \mu_{n,i}|^3 \). Now write \( \mu_n = \frac{1}{i \in L} \mu_{n,i} \), \( s_n^2 = \frac{1}{i \in L} \sigma_{n,i}^2 \) and
\[ r_n = \sum_{i \in L} \rho_{n,i}, \]
where as always \( L = \{ i: X(i) \leq X(n_L) \} \). Note that all these
quantities are functions of $X_n$, From Berry-Esseen follows

$$|P_L(X_n) - \Phi(-\mu_n/s_n)| \leq 6r_n/s_n^3.$$  

Now the second factor in (A.3) can be handled analogously, so an appeal to the Bounded Convergence Theorem reduces the proof to showing that

(A.5) $$r_n/s_n^3 \xrightarrow{P} 0,$$  

and

(A.6) $$\mu_n/s_n \xrightarrow{P} -(\omega+c\mu_L)(P_L(\lambda(0)))^2(\psi')^{-1/2}.$$  

We begin by showing that $s^2_n/n \xrightarrow{P} P_L\psi^2$. Write

(A.7) $$s^2_n/n = (nL/n)\psi^2 + n^{-1} \sum_{i \in L} \bar{E}[[\psi^2(e_i + n^{-1/2}\eta_i) - \psi^2(e_i)] - n^{-1} \sum_{i \in L} \bar{E}\psi(e_i + n^{-1/2}\eta_i)^2,$$

where $\eta_i = -(\omega+cX_i)$. The Markov inequality shows that the second term in (A.7) tends to 0 in probability if $W_n = \bar{E}\psi^2(e + n^{-1/2} \eta) \xrightarrow{L^1} \bar{E}\psi^2(e) = W$.

This follows if a) $W_n \xrightarrow{P} W$ and b) $W_n$ is uniformly integrable. The continuity property (B)(ii) implies a); while b) follows from

monotonicity of $\psi$ and condition (D). Indeed, $\psi^2(e + n^{-1/2} \eta)\mathbb{I}[\psi^2(e + n^{-1/2} \eta) > a] \leq [\psi^2(e) + \psi^2(c + \eta)]\mathbb{I}[\psi^2(e) \lor \psi^2(c + \eta) > a]$, and the latter is integrable.

The third term of (A.7) is $o_P(1)$ for similar reasons, except that one appeals to the continuity of $\lambda(t)$ at 0 (and $\lambda(0) = 0$) instead of (B)(ii).

Consequently (A.5) can be established by showing that $r_n$ is $o_P(n^{3/2})$. Indeed $r_n/n \leq n^{-1} \sum 1 \rho_{n,i}$, which is $o_P(1)$ because
\[ \mathbb{E}_n \psi(c + \eta^{-1/2} - \lambda(n^{-1/2} + \eta))^{\frac{3}{2}} \]

which is uniformly bounded in \( n \) by virtue of assumption (D) and monotonicity of \( \psi \).

To establish (A.6) we write

\[
\frac{n^{-1/2}(w+cX)}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i \in I} \int_{\mathbb{R}} \left[ \lambda'(s) - \lambda'(0) \right] ds + \frac{\lambda'(0)}{n} \sum_{i \in I} (w+cX_i).
\]

That the second term converges in probability to \( p_L \lambda'(0)(w+c\mu_L) = -p_L(w+c\mu_L)[\lambda'(0)]^2 \) may be seen by writing \( n^{-1} \sum_{i \in I} X_i \) as

\[
\frac{n}{n} \int_0^1 F_n^{-1}(p) dp, \text{ where } F_n \text{ is the empirical d.f. of } X_1, \ldots, X_n. \]

The first term is \( o_p(1) \): this is a consequence of the next lemma applied to

\[ Y_i = w + cX_i \] and \( g(y) = \frac{1}{y} \int_0^y \left[ \lambda'(s) - \lambda'(0) \right] ds. \]

**Lemma.** Let \( g(y) \) be a function such that \( g(y) \to 0 \) as \( y \to 0 \); and \( Y_1, Y_2, \ldots \) be i.i.d. r.v.'s. Suppose that either (i) \( EY^2 < \infty \) or (ii) \( E|Y| < \infty \) and \( g(y) \) is bounded. Let \( L_n \) be a (possibly random) subset of \( \{1, \ldots, n\} \). Then

\[
\frac{1}{n} \sum_{i \in L_n} Y_i g(n^{-1/2} Y_i) \overset{P}{\to} 0.
\]

We omit the proof, remarking only that in case (i) \( n^{-1/2} \max_{1 \leq i \leq n} |Y_i| \overset{P}{\to} 0 \) while in case (ii) \( E|Yg(n^{-1/2} Y)| \to 0 \). Assumption B and C ensure that the lemma applies to the present situation.
A.4 HEURISTIC ARGUMENT FOR REDESCENDING $\psi$

In the following, we condition on the observed sequence of carriers $X = (x_1, x_2, \ldots)$ and set $\delta = 1$ in (5.2). Define $T(y_1, \ldots, y_m)$ as a value of $t$ minimizing $|\sum_1^m \psi(y_i - t)|$ in a prespecified interval $[-M, M]$. Suppose for the sequence $(x_1, x_2, \ldots)$ that $n_L / n \rightarrow p_L$, $n_L^{-1} \sum_1^n x_i \rightarrow \mu_L$ and $n_L^{-1} \sum_1^n (x_i - \mu_L)^2 \epsilon \delta(1)$, and that $\lambda(t) = E\psi(\epsilon - t)$ has a unique zero at $t = 0$ with $\lambda'(0) \neq 0$. Then under further smoothness and integrability conditions one can show using Huber (1967) that

$$\sqrt{n} T_L_n (\epsilon_i - \gamma n^{-1/2} (x_i - \mu_L)) \xrightarrow{d} N(0, p_L^{-1} V_L)$$

with an analogous result valid for $T_R_n$.

These results hold for each value of $\gamma$. Under appropriate conditions the random process

$$X_n(\gamma) = \sqrt{n} T_L_n (\epsilon_i - \gamma n^{-1/2} x_i) - \sqrt{n} T_R_n (\epsilon_i - \gamma n^{-1/2} x_i), \gamma \in \mathbb{R}$$

with paths in $D(-\infty, \infty)$ converges weakly to the degenerate Gaussian process $X(\gamma) = (2p_L^{-1} V_L)^{1/2} - (\mu_R - \mu_L) \gamma$, where $Z \sim N(0, 1)$.

The mapping $h: D(-\infty, \infty) \rightarrow \mathbb{R}$ that locates the minimum in $[-M, M]$ of $\gamma \rightarrow |X(\gamma)|$ (with conventions to handle non-uniqueness and jump discontinuities) is continuous on the support of $\{X(\gamma)\}$, and so from the continuous mapping theorem

$$\sqrt{n}(\beta - \beta) = h(X_n(\gamma)) \xrightarrow{d} h(X(\gamma)) = (2p_L^{-1} V_L)^{1/2} / (\mu_R - \mu_L).$$
A.5 VARIANCE REDUCTION METHODS

Gaussian over Independent Swindles. This is an extension also discussed by Goodfellow and Martin (1976) to the regression situation of a common method for location problems (e.g. Andrews et al (1972), Simon (1976)). Assume that a linear model $y = X \beta + \epsilon$ obtains, where $Y$ is an $n \times 1$ column vector, $X$ a fixed $n \times p$ matrix of carriers, $\beta$ a $p \times 1$ parameter vector, and $\epsilon$ is an $n \times 1$ vector of i.i.d. variables $\epsilon_i = z_i / \omega_i$. Here $z_i \sim N(0, \sigma^2)$ and $\omega_i$ are i.i.d., positive and independent of $z_i$.

Conditional on the vector $w = (\omega_1, \ldots, \omega_n)$, standard normal theory yields the (complete) sufficient statistics $\hat{\beta}_w$, $\hat{\sigma}_w^2$ as the weighted least squares estimates of $\beta$, $\sigma^2$ and the (ancillary) standardized residuals $\hat{e} = (y - X \hat{\beta}_w) / \hat{\sigma}_w$. Let $b(y)$ be an arbitrary regression invariant estimator of $\beta$, i.e.

satisfying $b(cY + Xd) = c b(Y) + d$ for $c \in \mathbb{R}$, $d \in \mathbb{R}_p$. Conditional on $w$, the sufficiency ancillarity theorem (e.g. Simon (1976)) ensures independence of $\hat{\beta}_w$, $\hat{\sigma}_w^2$ and $\hat{\sigma}$. From this, it follows the decomposition (under the assumption that $\beta = 0$, $\sigma^2 = 1$, and that $b(y)$ is unbiased)

$$
\text{Var } b(y) = E(X^T \Lambda^2 X)^{-1} + \text{Var } b(\hat{\epsilon}).
$$

where $\Lambda = \text{diag}(\omega_i)$ and $\text{Var}$ denotes the variance-covariance matrix. In many cases the first term can be evaluated analytically, or once-and-for-all by Monte Carlo, while $\text{Var } b(\hat{\epsilon})$, being smaller than $\text{Var } b(y)$, can usually be estimated more accurately. Based on $I = 1, \ldots, N$ replications,

$$
\text{Var } b(\hat{\epsilon}) = E b(\hat{\epsilon}) b(\hat{\epsilon})^T
$$

is estimated via the method of moments as

$$
\frac{1}{N} \Sigma b(e(I)) b(e(I))^T,
$$

with the variability of the latter estimated in a standard way using sums of fourth moments. Finally we note that the decomposition (1) remains valid if the rows of $x$ are i.i.d. random vectors, this following from the unbiasedness of $b(y)$. 

Score function swindle. This is a general method, not limited to the regression context, or to distributions from the Gaussian-over-independent family. For more details see Johnstone and Velleman (1983). Consider for example a simple linear regression model \( Y_i = \alpha + \beta (X_i - \mu_X) + \epsilon_i, \ i = 1, \ldots, n \); where \( \alpha \) and \( \beta \) are unknown scalars, \( \epsilon_i \) are i.i.d. with density \( g \), and the \( X_i \) may be either fixed (in which case \( \mu_X = \bar{X} \)) or i.i.d. variables (with \( \mu_X = E X \)). The efficient score function for estimation of \( \beta \) is given by

\[
S_n = -\sum_{i=1}^{n} (X_i - \mu_X) (\log g)' (Y_i - \alpha - \beta (X_i - \mu_X)),
\]

and has (in regular cases) expectation zero and variance \( n \sigma^2_{X} I(g) \), where \( I(g) = \int (g')^2 / g \) is the Fisher information (for location) of the density \( g \).

The Cramer-Rao inequality says that \( L = [n \sigma^2_{X} I(g)]^{-1} \) is a lower bound to the variance of any unbiased estimator \( b(y) \) of \( \beta \), and that \( b(y) - LS_n \) and \( LS_n \) are uncorrelated (when \( \beta = 0 \)). This gives the decomposition (for \( \beta = 0 \), and a regression invariant estimator \( b(y) \))

\[
\text{Var} b(y) = L + \text{Var} b(y - LS_n).
\]

Again, in many cases \( L \) can be evaluated explicitly (perhaps with the aid of a numerical integration), while \( \text{Var}(b(y) - LS_n) \) can usually be estimated from simulation more accurately than \( \text{Var}(b(y)) \). The effectiveness of the swindle will depend on how close the (usually unattainable) bound \( L \) is to the variance of the Pitman estimate, and on the amount of correlation between \( b(y) \) and \( LS_n \).
A.6 MONTE CARLO EXPERIMENT DETAILS

Without loss of generality we set the true $\beta = 0$. Pseudo-random errors were generated using the IMSL (1980) linear congruential random number generator GGBUS (seed = 714235803) and Box-Muller Gaussian transformation in GGNOR. Gaussian variates were taken directly from GGNOR. Mixed Gaussian were generated by multiplying unit Gaussian values by 3.0 with probability 0.1. Slash variates were generated as the quotient of a random Gaussian from GGNOR and an independent random uniform from GGBUS.

For sample size 10 we performed 10,000 replications; for sample size 22, 5,000; for sample size 40, 2,000. In each run, the same samples were presented to all estimators in parallel, so comparisons of efficiencies among estimators in the same run are somewhat more reliable than the reported standard errors would indicate.

All programs were written in Fortran 66, compiled either by IBM's Fortran H compiler or by IBM's FORTVS compiler at optimization level 3. Trials were run on Cornell University's IBM 370/168 under a modified OS/MVT operating system using CMS.
REFERENCES


Bartlett, M. S. (1949), "Fitting a Straight Line When Both Variables are Subject to Error," Biometrics, 5, 207-212.


de'Laplace, Pierre Simon (1818), Deuxième Supplement à la Théorie Analytique des Probabilités, Paris: Courcier.


