

Regression-Free and Robust Estimation of Scale for Bivariate Data

Peter J. Rousseeuw and Mia Hubert

*Department of Mathematics and Computer Science, U.I.A.,
Universiteitsplein 1, B-2610 Antwerp, Belgium*

Abstract: In this paper we present robust estimators for the dispersion of the errors in simple linear regression. Existing scale estimators are based on the residuals from an estimator of the regression itself. Instead, we propose scale estimators that do not depend on any previous estimate of the regression parameters. For this purpose we consider triangles formed by data points, and define their vertical height. Taking the repeated median of all such heights leads to a 50% breakdown point estimator. A second estimator is obtained from the 0.278-quantile of all triangle heights, and results in a breakdown point of 34.7%. When we restrict ourselves to the heights of adjacent triangles and take their 0.4-quantile, we obtain a much faster estimator with a 20% breakdown point. Simulations are carried out to study the computation time and statistical performance of these estimators.

Keywords: Breakdown point; Linear model; Outliers; Regression invariance.

1 Introduction

The central model in this paper is that of simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ for } i = 1, \dots, n, \tag{1.1}$$

where the distribution of ε_i has a scale parameter σ . Like the response y_i , also the regressor x_i can be observational, as is the case in many applications. In the model (1.1) one typically begins by estimating the parameters β_0 and β_1 by regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, and then uses the latter to estimate σ . For instance, the ordinary least squares (LS) method estimates σ by

$$\hat{\sigma}_{LS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n r_i^2}$$

where the residuals are $r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ for all $i = 1, \dots, n$.

It is well-known that the LS estimators of β_0 , β_1 and σ are extremely vulnerable to outliers, whether in the y_i or in the x_i (see, e.g., Rousseeuw and Leroy 1987). Outliers can be thought of as a minority of data points $z_i = (x_i, y_i)$ which do not obey the model assumptions. Because of this sensitivity, many methods of robust regression have been considered. In all cases we know of, the estimate of σ is always based on the residuals r_i which in turn are based on the robust estimates of β_0 and β_1 .

For univariate data a similar situation exists, in the sense that robust scale estimators are typically based on an initial estimator of location. However, recently some univariate scale estimators were proposed (Rousseeuw and Croux 1993) which are equally robust but location-free. In the same vein, our purpose is to construct robust estimators of σ in (1.1) which do not depend on the choice of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

Our original motivation for addressing this question was twofold. First of all, such a regression-free $\hat{\sigma}$ may be used as an initial estimator in the construction of certain rank-based estimators and tests for $\hat{\beta}_0$ and $\hat{\beta}_1$ which are presently being developed (J. Jurečková, personal communication). And secondly, by comparing a regression-based scale estimate with a regression-free scale estimate on the same data, it is possible to check the validity of the model (1.1) as is done in Section 6.

The structure of the paper is as follows. In Section 2 the estimators are constructed, and in Section 3 their breakdown points are given. Section 4 presents a simulation study of their performance, whereas Section 5 discusses related approaches. In Section 6 some examples are considered. Finally, Section 7 draws conclusions and looks at possible extensions. All proofs are collected in the Appendix.

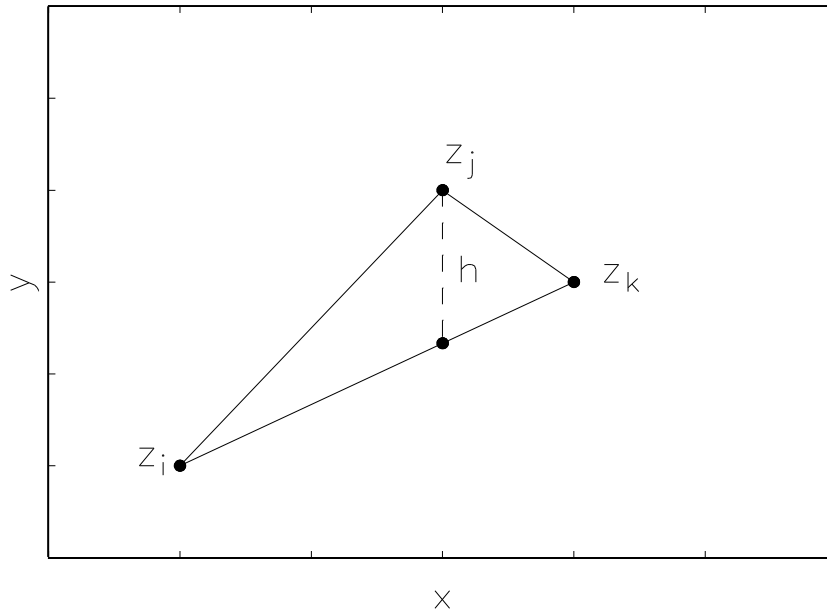


Figure 1: Vertical height of a triangle

2 Description of the estimators

Any scale estimator in the linear model needs to be *regression invariant*, which means that replacing the data points (x_i, y_i) by $(x_i, y_i + ax_i + b)$ should yield the same result. In other words, adding a perfect linear relation with arbitrary slope a and intercept b makes no difference to the error scale. This property is the natural analog to the well-known translation invariance of univariate scale estimators.

The basic idea of the estimators in this paper is to consider triplets of data points. Any three data points z_i, z_j and z_k determine a (possibly degenerate) triangle $\Delta(z_i, z_j, z_k)$. It is possible to measure the scale of these three points in a regression invariant way, by computing the vertical height of the triangle. If we assume that the points are labelled such that $x_i < x_j < x_k$ (as in Figure 1), the height h is defined as the length of the vertical line segment between z_j and $z_i z_k$.

We can easily verify that

$$h(z_i, z_j, z_k) = \left| y_j - y_i - \frac{(y_k - y_i)(x_j - x_i)}{x_k - x_i} \right| \quad \text{if } x_i < x_j < x_k \quad (2.1)$$

which becomes zero when z_j lies on $z_i z_k$, in which case the triangle degenerates to a line segment. Formula (2.1) also holds when $x_i \leq x_j < x_k$ or $x_i < x_j \leq x_k$. In the special case where $x_i = x_j = x_k$ we obtain a degenerate triangle for which we set $h(z_i, z_j, z_k) = 0$.

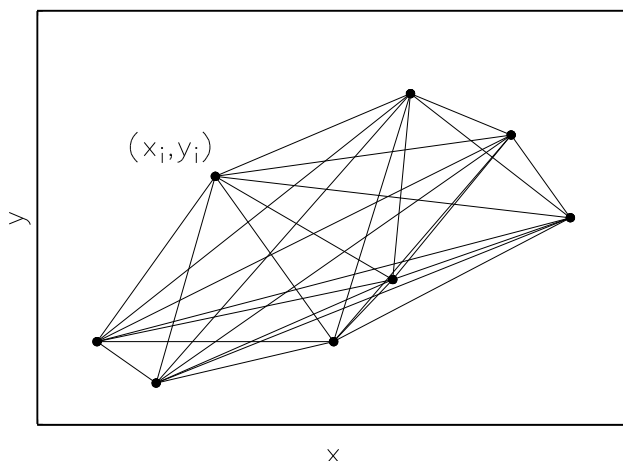


Figure 2: All triangles formed by the data points

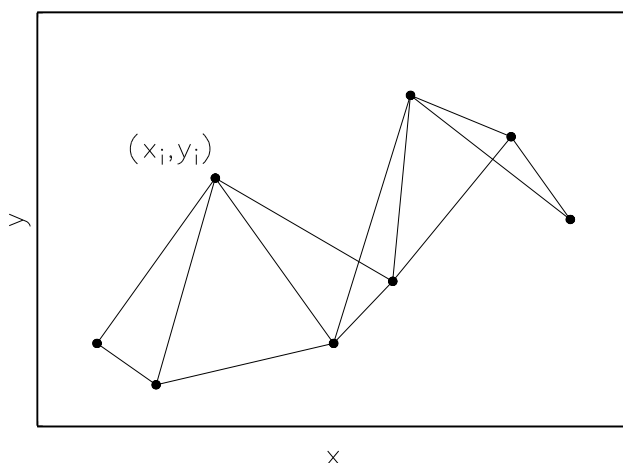


Figure 3: Adjacent triangles formed by the data points

Now we have to choose which collection of triangles we are going to consider. We investigate two cases:

Case 1: We take *all* triangles formed by the n data points, as in Figure 2. There are $\binom{n}{3}$ such triangles. (Note that the x -components do not need to be ordered in this approach.)

Case 2: We consider the *adjacent* triangles, as shown in Figure 3. Let us first label the data points to have increasing x -coordinates. Then we consider the triangles $\Delta(z_1, z_2, z_3)$, $\Delta(z_2, z_3, z_4)$, $\Delta(z_3, z_4, z_5)$, \dots which means that two consecutive triangles always have one edge in common. In all, we form $n - 2$ adjacent triangles:

$$\{\Delta(z_i, z_{i+1}, z_{i+2}); i = 1, \dots, n - 2\}.$$

Now we can construct the scale estimators. For case 1 we propose an estimator based on

the α -quantile of all heights:

$$Q_{all}^\alpha = c_1 \{h(z_i, z_j, z_k); 1 \leq i < j < k \leq n\}_{[\alpha \binom{n}{3}]} \quad (2.2)$$

where $0 < \alpha \leq 1$, and the repeated median height

$$R = c_2 \underset{i}{\text{med}} \left\{ \underset{j \neq i}{\text{med}} \left\{ \underset{k \neq i, j}{\text{med}} h(z_i, z_j, z_k) \right\} \right\}. \quad (2.3)$$

The expression for R can be interpreted as follows: first fix $i \in \{1, \dots, n\}$ and $j \neq i$. Then take $\text{med}_{k \neq i, j} h(z_i, z_j, z_k)$. Doing this for all $j \neq i$ results in $n - 1$ values, of which we take the median once again. Finally we repeat this for all i , and obtain R as the median of all partial results. Note that each height occurs six times in the course of this computation. If we want small storage requirements, we have to compute each height when it is needed. On the other hand it is faster to store all heights in memory, but for this we need $O(n^3)$ storage space. To obtain Q_{all}^α we consider all the heights just once, and take their $[\alpha \binom{n}{3}]$ -th order statistic. In the next section we will select an appropriate value of α .

In the case of adjacent triangles, we analogously define the estimator

$$Q_{adj}^\alpha = c_3 \{h(z_i, z_{i+1}, z_{i+2}); i = 1, \dots, n - 2\}_{[\alpha(n-2)]} \quad (2.4)$$

where again $0 < \alpha \leq 1$. The constants c_1, c_2 and c_3 can be chosen to make the estimators consistent for σ in (1.1) under a given type of distribution (see Section 4).

By construction, all three estimators are regression invariant. Note that the computation of R and Q_{all}^α as described above takes $O(n^3)$ time, whereas Q_{adj}^α only needs $O(n \log n)$ time if we use an efficient algorithm for sorting the x_i .

Note that Q_{adj}^α is related to a scale estimator proposed by Rice (1984) and Gasser et al. (1986) in the context of nonparametric regression with a fixed and equidistant design (see also Müller and Stadtmüller 1993). For each data point z_i they consider the residual $r_{i,LS}$ obtained by a least squares fit to z_i and the surrounding points z_{i-1} and z_{i+1} . The error variance is then estimated by the average of all $r_{i,LS}^2$. Now $|r_{i,LS}| = \frac{2}{3}h(z_{i-1}, z_i, z_{i+1})$ for an equidistant design, hence both kernels are equivalent in that case. But in our model we assume the explanatory variable to be random. Therefore Q_{adj}^α differs from Rice's estimator in two ways. First, instead of $r_{i,LS}$ we use the residual r_{i,L_1} from an L_1 -fit to (z_{i-1}, z_i, z_{i+1}) . The absolute value of r_{i,L_1} equals the vertical height (2.1) because an L_1 -fit to 3 points contains the first and the last. Secondly, instead of taking an average which is vulnerable to even a single outlier, we use a quantile. Because of this, Q_{adj}^α can withstand a substantial fraction of outliers. This robustness aspect is worked out in detail in the next section.

3 Breakdown points

We will now investigate the breakdown points of our estimators. For any sample $Z = \{z_1, \dots, z_n\}$ the *finite-sample breakdown point* (see Donoho and Huber, 1983) of a scale estimator S is defined as

$$\varepsilon_n^*(S, Z) = \min \{\varepsilon_n^+(S, Z), \varepsilon_n^-(S, Z)\}$$

where

$$\varepsilon_n^+(S, Z) = \min \left\{ \frac{m}{n}; \sup_{Z'} S(Z') = \infty \right\}$$

and

$$\varepsilon_n^-(S, Z) = \min \left\{ \frac{m}{n}; \inf_{Z'} S(Z') = 0 \right\}.$$

Here, Z' ranges over all data sets obtained by replacing any m observations of Z by arbitrary values. The ε_n^+ and ε_n^- are called the *explosion* and *implosion* breakdown points.

Note that the breakdown point of S depends on the original data set Z . For many estimators S however, the breakdown point remains the same over all data sets in *general position*. A bivariate data set Z is said to be in general position if no three data points lie on a straight line. This implies that any three observations form a nondegenerate triangle, and that at most two different replicates can occur at each x_i .

Although the breakdown point depends on the sample size n , it often does not vary much with n (as long as Z is in general position). In most cases it tends to a meaningful limit, denoted by

$$\varepsilon^*(S) = \lim_{n \rightarrow \infty} \varepsilon_n^*(S, Z),$$

which may be called the (asymptotic) breakdown point of S . Analogously, we denote

$$\varepsilon^+(S) = \lim_{n \rightarrow \infty} \varepsilon_n^+(S, Z) \quad \text{and} \quad \varepsilon^-(S) = \lim_{n \rightarrow \infty} \varepsilon_n^-(S, Z).$$

Theorem 1. *At any sample $Z = \{z_1, \dots, z_n\}$ in general position we have*

$$\varepsilon_n^+(R, Z) = \left\lfloor \frac{n-1}{2} \right\rfloor / n \quad \text{and} \quad \varepsilon_n^-(R, Z) = \left\lfloor \frac{n}{2} \right\rfloor / n$$

hence the breakdown point of R is 50%.

Theorem 2. For each $0 < \alpha \leq 1$ the estimator Q_{all}^α has explosion breakdown point

$$\varepsilon^+(Q_{all}^\alpha) = 1 - \sqrt[3]{\alpha}$$

and implosion breakdown point

$$\varepsilon^-(Q_{all}^\alpha) = \begin{cases} \frac{1}{2} - \frac{1}{2}\cos(\theta_\alpha) + \frac{\sqrt{3}}{2}\sin(\theta_\alpha) & \text{if } 0 < \alpha \leq \frac{1}{2} \\ \frac{1}{2} & \text{if } \alpha = \frac{1}{2} \\ \frac{1}{2} + \frac{1}{2}\cos(\theta_\alpha) + \frac{\sqrt{3}}{2}\sin(\theta_\alpha) & \text{if } \frac{1}{2} < \alpha \leq 1 \end{cases} \quad (3.1)$$

where

$$\theta_\alpha = \frac{1}{3} \text{Arctan}\left(\frac{\sqrt{\alpha(1-\alpha)}}{\frac{1}{2}-\alpha}\right).$$

The maximal breakdown point of Q_{all}^α is obtained by putting $\alpha = 0.278$, which results in $\varepsilon^*(Q_{all}^\alpha) = 34.7\%$.

Theorem 3. For each $0 < \alpha \leq 1$ the estimator Q_{adj}^α satisfies

$$\varepsilon_n^+(Q_{adj}^\alpha, Z) = \left\lceil \frac{(n-1) - [\alpha(n-2)]}{3} \right\rceil / n$$

and

$$\varepsilon_n^-(Q_{adj}^\alpha, Z) = \left\lceil \frac{[\alpha(n-2)]}{2} \right\rceil / n$$

at any Z in general position. As a consequence of this result, the breakdown point of Q_{adj}^α reaches its maximal value $\varepsilon^*(Q_{adj}^\alpha) = 20\%$ by taking $\alpha = 0.4$.

Theorem 2 allows us to select the value of α yielding the most robust version of Q_{all}^α , and Theorem 3 does the same for Q_{adj}^α . From now on we will simply denote

$$Q_{all} = Q_{all}^{0.278} \quad \text{and} \quad Q_{adj} = Q_{adj}^{0.4}$$

and concentrate mainly on these versions rather than those with general α .

From the proof of Theorem 2, we see that taking a higher quantile for Q_{all}^α (e.g. $\alpha = 0.5$) would yield an estimator with a lower explosion breakdown point, making it more sensitive to vertical outliers. The fact that R has the maximum (50%) breakdown point is similar to the situation for slope estimators. There the estimator of Theil (1950) and Sen (1968), defined as the overall median of the pairwise slopes, has a 29.3% breakdown point. On the other hand Siegel's (1982) repeated median slope estimator, based on the same kernel, resists up to 50% of contamination. (Note that the Theil-Sen and Siegel regression methods do not provide an estimate of scale.)

Table 1: Average computation time in seconds

n	R	Q_{all}	Q_{adj}
15	4.5	.75	.02
35	61	38	.04
100			.13
1000			1.8

4 Simulations and numerical results

In this section we present some simulation results to investigate the finite-sample behavior of the proposed estimators. Because of regression invariance, we may assume $\beta_0 = \beta_1 = 0$. To set up the experiment, we first looked at computation time. For different values of n , Table 1 gives the computation time in seconds for R and Q_{all} (averaged over 300 samples) and Q_{adj} (averaged over 1000 samples). The computations were carried out on a 486-PC using Gauss, a software package which interprets macros. If one would execute a compiled version (e.g. in Fortran), the speed would be much higher.

For each n in Table 2 we generated many samples of n data points (x_i, y_i) , where y_i always had a standard gaussian distribution. The variable x_i was generated according to :

1. the standard gaussian distribution;
2. a bimodal distribution, with half of its mass uniform on $[-2, -1]$ and the other half uniform on $[1, 2]$;
3. the negative exponential distribution on $[0, \infty)$;
4. the Cauchy distribution.

For each of these four situations we verified, by means of $Q - Q$ plots, that the estimates obtained from R , Q_{all} , and Q_{adj} are approximately normally distributed.

The first columns of Table 2 give the average value of R and Q_{all} over 300 samples. Because of computation time, these values were only obtained up to $n = 35$. For Q_{adj} we were able to run the simulation 1000 times, for n up to 1000. Note that these average values were obtained without including the constants c_1 to c_3 which occur in the definitions (2.2) to (2.4). These constants will be determined below.

Table 2: Simulation results for different distributions of x_i

distribution	n	average value			standardized variance		
		R	Q_{all}	Q_{adj}	R	Q_{all}	Q_{adj}
Gaussian	15	0.76	0.48	0.64	1.22	0.98	3.06
	35	0.76	0.46	0.66	0.86	0.72	3.10
	100			0.67			2.59
	1000			0.68			2.61
Bimodal	15	0.79	0.48	0.64	1.19	1.06	2.87
	35	0.78	0.47	0.67	0.92	0.81	2.95
	100			0.67			2.78
	1000			0.67			2.81
Exponential	15	0.79	0.49	0.63	1.17	1.04	3.34
	35	0.76	0.46	0.65	0.94	0.80	3.16
	100			0.67			2.64
	1000			0.67			2.92
Cauchy	15	0.79	0.49	0.65	1.15	0.96	2.95
	35	0.78	0.47	0.66	0.85	0.78	3.01
	100			0.66			2.70
	1000			0.67			2.86

The second part of Table 2 shows the standardized variance

$$n(\text{var}_m(S))/(\text{ave}_m(S))^2 \quad (4.1)$$

for each estimator S , where $\text{var}_m(S)$ and $\text{ave}_m(S)$ are the variance and the average value of S over $m = 300$ or $m = 1000$ samples. We use the standardized variance to obtain a natural measure of accuracy for scale estimators (Bickel and Lehmann 1976), which does not change when S is multiplied by a constant factor. Note that the standardized variance of Q_{adj} is substantially higher (hence its finite-sample efficiency is lower) than those of R and Q_{all} , due to the fact that we restrict ourselves to the adjacent triangles rather than considering all of them.

Table 3 lists the results of a numerical computation of the asymptotic value of R , Q_{all} , and Q_{adj} . This procedure was quite tedious. First we considered the functional corresponding with each estimator. For Q_{all} we have the functional

$$Q_{all}(F) = c_1 H^{-1}(0.278)$$

where H is the cumulative distribution function of $h(Z_1, Z_2, Z_3)$ when Z_1, Z_2 and Z_3 follow the distribution F , hence

$$\begin{aligned} H(u) &= P(h(Z_1, Z_2, Z_3) \leq u) \\ &= P\left(\left|Y_2 - Y_1 - \frac{(Y_3 - Y_1)(X_2 - X_1)}{X_3 - X_1}\right| \leq u \mid X_1 < X_2 < X_3\right). \end{aligned}$$

Therefore the finite-sample Q_{all} can be seen as an estimator of $Q_{all}(F)$. (This parallels the situation for univariate data $x_1, \dots, x_n \sim G$ where the sample median is a consistent estimator of the functional $T(G) = \text{med}(G)$, and where the median absolute deviation estimates the functional $S(G) = \text{med}|X - \text{med}(G)|$.) The functional $Q_{all}(F)$ has the advantage that it exists at all distributions F , since no moment conditions are needed. By generating a sample of 15000 triplets from each distribution F , we approximated H by the corresponding empirical distribution function. The functional $Q_{all}(F)$ is then computed by interpolation.

The same technique is used to obtain the asymptotic value of R . It can be written as the functional

$$R(F) = c_2 \text{med}_{Z_1} f_1(Z_1)$$

with

$$f_1(z_1) = \text{med}_{Z_2} f_2(z_1, Z_2) \quad \text{and} \quad f_2(z_1, z_2) = \text{med}_{Z_3} h(z_1, z_2, Z_3).$$

Table 3: Numerically obtained asymptotic values of R , Q_{all} , and Q_{adj} for the same distributions

distribution	R	Q_{all}	Q_{adj}
Gaussian	0.765	0.456	0.676
Bimodal	0.810	0.478	0.675
Exponential	0.800	0.459	0.675
Cauchy	0.792	0.461	0.674

This expression is equivalent with

$$R(F) = c_2 G^{-1}\left(\frac{1}{2}\right) \quad \text{where} \quad G(u) = P(f_1(Z_1) \leq u).$$

The functions f_1 and f_2 can also be written as the inverse of distribution functions in the point $\frac{1}{2}$. We approximated them by generating many observations and calculating the corresponding empirical distribution functions. A similar computation was used for $Q_{adj}(F)$.

In the simulations of Table 2 we have simply set $c_1 = c_2 = c_3 = 1$, which is also what we recommend for analyzing data when we don't know which distribution generated it. It is only possible to compute consistency factors c_1, c_2, c_3 when we assume that (a majority of) the data come from a given distribution. Then c_1 to c_3 can be obtained as the inverses of the corresponding entries in Table 3. This then makes the functionals Fisher consistent for σ in the model (1.1), meaning that

$$Q_{all}(F) = R(F) = Q_{adj}(F) = \sigma.$$

Therefore it suffices to set $c_1 = 1/H^{-1}(0.278)$, and to assign analogous values to c_2 and c_3 .

We see that the average values in Table 2, obtained by simulation, are closely approximated by the asymptotic values in Table 3. In Table 2 the average values and variances remain almost the same across widely different x -distributions, which is important when using these estimators in practice. Note that Q_{all} belongs to the class of generalized L-statistics (Serfling 1984) and hence is asymptotically normal. We expect Q_{adj} to be asymptotically normal as well, because of its close relation to the incomplete generalized L-statistics of Hössjer (1994). Moreover, some estimators based on repeated medians were recently shown to be asymptotically normal (Hössjer, Rousseeuw and Croux 1994, 1995) by techniques that may be generalizable to R .

Table 4: Simulation results of different scale estimators, with and without outliers

<i>outliers</i>	average value				standardized variance			
	R	Q_{all}	Q_{adj}	$\hat{\sigma}_{LS}$	R	Q_{all}	Q_{adj}	$\hat{\sigma}_{LS}$
no outliers	0.98	1.01	0.94	0.97	1.17	1.04	3.70	0.61
vertical outliers	1.23	1.43	1.33	3.23	1.15	1.02	2.98	0.13
bad leverage points	1.21	1.45	1.29	44.49	1.11	0.97	3.16	4.30

We also verify the robustness of our scale estimators. This is done in the setting described in (Rousseeuw and Leroy 1987, page 209) with $n = 20$. We first generated n data points (x_i, y_i) following the linear relation

$$y_i = x_i + 1 + \varepsilon_i$$

with $x_i \sim N(0, 100)$ and $\varepsilon_i \sim N(0, 1)$. Then we replaced 10% of the data by vertical outliers with $\varepsilon_i \sim N(10, 1)$. In a third situation we used 10% of bad leverage points instead, for which the x_i have mean 100 and the y_i remain as before. The scale estimates were averaged over 300 runs and listed in Table 4, together with the corresponding standardized variances. (Here the estimators have been multiplied with the constants c_1, c_2 and c_3 .) The last column shows the least squares estimate $\hat{\sigma}_{LS}$ of σ . We see that $\hat{\sigma}_{LS}$ explodes in the presence of the outliers, whereas R, Q_{all} and Q_{adj} still give reasonable values at the contaminated data sets. Also their standardized variance remains quite stable.

5 Related approaches

If we replace the medians by averages in (2.3), the repeated median becomes a repeated average, which simply equals the average of all heights $h(z_i, z_j, z_k)$ with distinct indices i, j , and k . We obtain exactly the same estimator by replacing the quantile in (2.2) by an average. A different estimator is obtained from (2.4), yielding the average height of the $n - 2$ adjacent triangles. Simulations have confirmed that these modified estimators have lower standardized variances than the original R, Q_{all} , and Q_{adj} , but of course their breakdown point becomes zero due to the nonrobustness of the average. Therefore, these estimators were not considered further.

By inverting the above reasoning, we can regard the estimators in this paper as generalizations of U -statistics with the same h . Indeed, several positive-breakdown estimators are based on replacing sums by medians (Rousseeuw 1984). In the univariate scale problem, Rousseeuw and Croux (1992) constructed new estimators by robustifying the formulas of U -statistics (as well as those of M - and L -estimators) by replacing sums by medians, quantiles, and nested operations. Such modifications yield estimators that are easy to apply but hard to analyze mathematically. In the present paper we use data triplets (z_i, z_j, z_k) rather than the pairwise differences $|y_i - y_j|$ of the univariate setting.

Another possibility would therefore be to replace the height $h(z_i, z_j, z_k)$ in R, Q_{all} , and Q_{adj} by a different kernel. However, few other regression invariant measures of scale based on three points are readily available. A convenient choice would be the least squares scale estimate on three points, given by $(r_{LS}^2(z_i) + r_{LS}^2(z_j) + r_{LS}^2(z_k))^{\frac{1}{2}}$ which is easy to compute. We have repeated the simulations in Table 2 for the versions of R, Q_{all} , and Q_{adj} with this kernel, but it turned out that the standardized variances increased substantially in each case.

Finally, one could use other configurations of triangles than those in Figures 2 and 3. Instead of adjacent triangles, each of which has two data points in common with its predecessor, we may also consider *vertex-incident triangles* which intersect in a single data point, as in Figure 4. (Analogously, the adjacent triangles might have been called "edge-incident".) Going one step further, we can also consider a configuration of *isolated triangles* based on consecutive triplets, as in Figure 5. This leads to the two estimators

$$Q_{vertex} = \underset{i}{\text{med}}\{h(z_i, z_{i+1}, z_{i+2}); i = 1, 3, 5, \dots, 2 \left\lfloor \frac{n-1}{2} \right\rfloor - 1\} \quad (5.1)$$

and

$$Q_{iso} = \underset{i}{\text{med}}\{h(z_i, z_{i+1}, z_{i+2}); i = 1, 4, 7, \dots, 3 \left\lfloor \frac{n}{3} \right\rfloor - 2\}. \quad (5.2)$$

We see that, depending on n , one or two data points may be left over: in Figures 4 and 5 some points do not belong to any triangle of the configuration. We have verified that Q_{vertex} has a breakdown point of 12.5%, while Q_{iso} attains a breakdown point of 16.6%. (Replacing the medians in (5.1) and (5.2) by other quantiles does not increase these values.) Therefore, Q_{vertex} and Q_{iso} are dominated by Q_{adj} in this respect, whereas their computational complexity is the same. Moreover, when repeating the simulations of Table 1 for these estimators we found that their standardized variances were systematically higher than that of Q_{adj} , which reflects the fact that (5.1) and (5.2) use the information in the data points very sparingly.

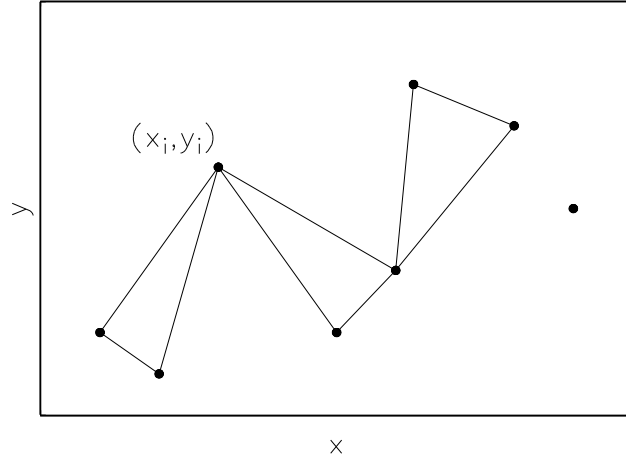


Figure 4: Vertex-incident triangles formed by the data points

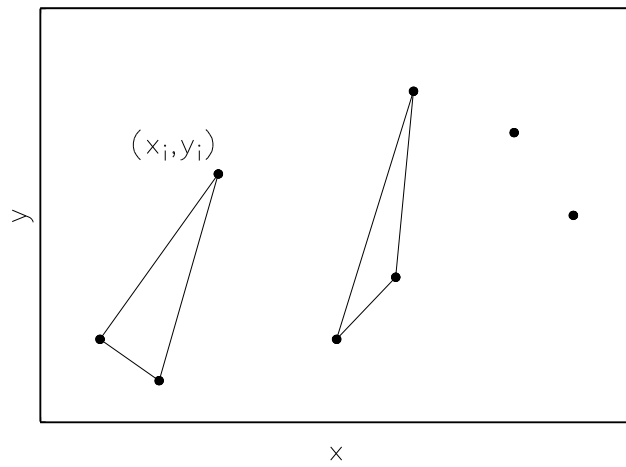


Figure 5: Isolated triangles formed by the data points

Because of these results, the estimators Q_{vertex} and Q_{iso} are not recommended in practice.

6 Application: testing linearity

A potential application of regression-free scale estimators is to testing linearity. It seems that Q_{adj} is suitable for this purpose because it only uses the local structure of the data, as can be seen in Figure 3. On the other hand, a regression-based scale estimator S starts by fitting a straight line to the data (thus using the data globally) and then measures the scale of the residuals. When there really is a linear model underlying the data (possibly with the exception of some outliers), then Q_{adj} and S will give similar results. But if there is curvature, Q_{adj} will typically be smaller than S because the latter is based on a linear relation which does not fit well. Therefore, we propose the test statistic

$$T = \frac{S}{Q_{adj}} \quad (6.1)$$

which is dimensionless and regression invariant. The null hypothesis of linearity is rejected at the level γ if (6.1) exceeds a critical value c_γ . It is possible to determine c_γ in advance for various distributions of x_i by means of numerical methods. Another approach is to consider the observed x_i as fixed, and to simulate m gaussian samples $\{y_i; i = 1, \dots, n\}$ of i.i.d. observations. For each such sample we can compute the test statistic (6.1) and denote it by $t^{(j)}$ for $j = 1, \dots, m$. Then the $[(1 - \gamma)m]$ -th order statistic of these values $t^{(j)}$ provides an approximation to c_γ . Analogously, $\#\{t^{(j)} > t\}/m$ is an approximation to the p -value $P[T > t]$.

We will illustrate this idea on two data sets. For S we use the robust scale estimator

$$\hat{\sigma}_{LQS} = c \sqrt{r_{([1-\alpha]n)}^2(\hat{\beta}_{LQS})} \quad (6.2)$$

where $\hat{\beta}_{LQS}$ is the least quantile squares (LQS) estimator (Rousseeuw and Leroy 1987, p.124) defined as

$$\hat{\beta}_{LQS} = \operatorname{argmin}_{\beta} r_{([1-\alpha]n)}^2(\beta).$$

For $\alpha = 20\%$ the estimators $\hat{\beta}_{LQS}$ and $\hat{\sigma}_{LQS}$ have a 20% breakdown point just like Q_{adj} , and we choose $c = 1.28$ to make $\hat{\sigma}_{LQS}$ consistent at a gaussian error distribution.

Our first example is taken from Campbell (1974, p. 44). A biologist investigates the response of 21 sea anemones to a standard stimulus. The explanatory variable x is the size of

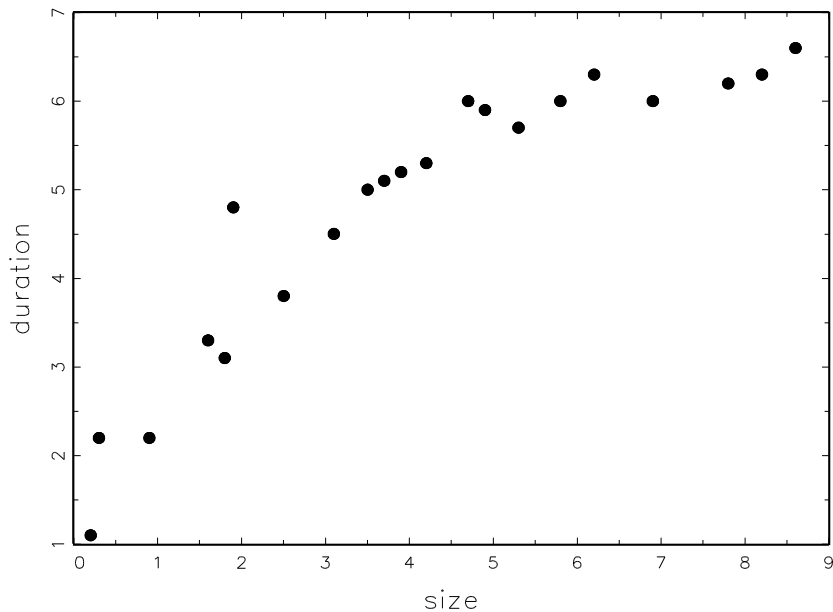


Figure 6: Duration of response versus size of sea anemones.

the anemone, whereas y measures the duration of the response. The plot (Figure 6) indicates a nonlinear relationship, whereas we also observe a vertical outlier. The test statistic becomes

$$T = \frac{0.820}{0.148} = 5.538.$$

The critical values were obtained by simulating 1000 gaussian samples $\{y_i\}$. This yielded $c_{0.01} = 3.246$, hence the linear hypothesis is rejected manifestly. Moreover, the test statistic is robust against the outlier.

Our second example are the Kootenay data, which report the water flow of the Kootenay river at Libby and Newgate in 13 successive years. Originally from Ezekiel and Fox (1959), Hampel et al. (1986) changed a good leverage point into a bad one (see Figure 7). For this data set T equals $\frac{1.916}{1.290} = 1.485$, yielding a simulated p -value of 0.36, hence the linearity hypothesis is maintained. Also here, the test is unaffected by the outlier.

7 Conclusions and outlook

In this paper we have proposed three scale estimators which are regression-free, in the sense that they do not depend on any prior regression estimator. Among these three scale estimators there is a tradeoff between robustness and efficiency on the one hand, and computation time on the other. The estimators R and Q_{all} have a high breakdown point (50% and 34.7%)

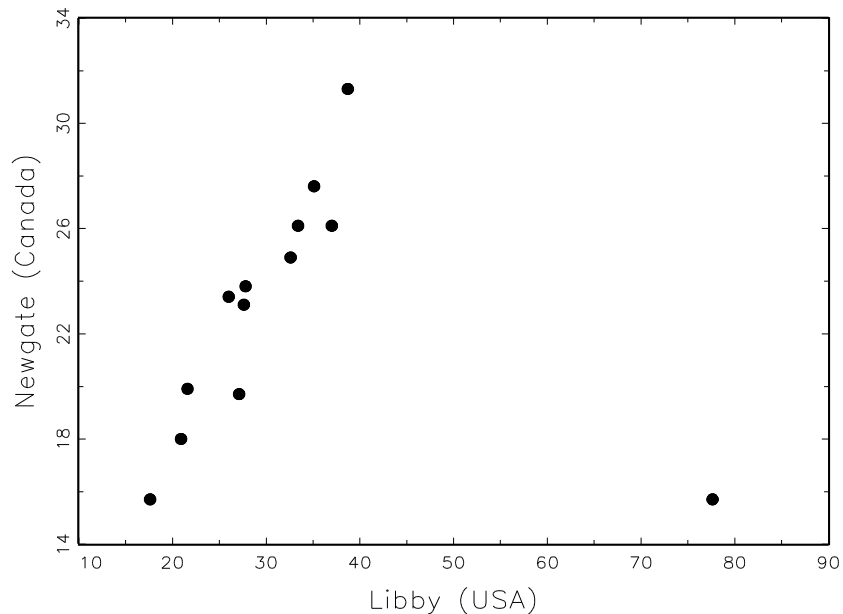


Figure 7: Water flow of the Kootenay river in Libby and Newgate.

and are relatively efficient, but their computational complexity is $O(n^3)$ which restricts their use to sample sizes up to about $n = 100$. The third estimator, Q_{adj} , has a 20% breakdown point which is lower but still reasonable, as well as a substantially higher variance, but its $O(n \log n)$ complexity makes it easy to compute even for large n .

Several extensions could be investigated. For instance, in view of the above tradeoff it might be interesting to study configurations in which the number of triangles lies between $n - 2$ and $\binom{n}{3}$, because this could yield estimators intermediate between Q_{adj} and Q_{all} in terms of breakdown point, efficiency and computation time. Also, it might be possible to construct faster algorithms for R and Q_{all} , because they are generalizations of the univariate location-free scale estimators S_n and Q_n for which faster algorithms have been constructed (Croux and Rousseeuw 1992).

A natural question is about multiple regression. In that case the data triangles become simplices (for instance, when there are two explanatory variables we obtain tetrahedrons). The height of a simplex is measured as before, along a vertical line, from the vertex with interior x -component to the hyperplane formed by the other vertices. Using a theorem of de la Vallée Poussin (see Cheney 1966, page 41) this height can be computed by a simple two-step procedure. In the first step, we apply least squares regression to the $p + 1$ points $\{z_{i_1}, \dots, z_{i_{p+1}}\}$, yielding $p + 1$ residuals $r_{LS}(z_{i_1}), \dots, r_{LS}(z_{i_{p+1}})$. Then the desired height

equals

$$h(z_{i_1}, \dots, z_{i_{p+1}}) = 2 \frac{r_{LS}^2(z_{i_1}) + \dots + r_{LS}^2(z_{i_{p+1}})}{|r_{LS}(z_{i_1})| + \dots + |r_{LS}(z_{i_{p+1}})|}. \quad (7.1)$$

The estimators R and Q_{all} can be generalized immediately by using all data simplices, but then the need for faster algorithms becomes urgent. For instance, one could adapt the code of Hawkins et al (1994). To extend Q_{adj} we first have to set up an adjacency relation, for instance by triangulating the collection $\{x_i; 1 \leq i \leq n\}$.

8 Appendix

Proof of Theorem 1. We first prove $\varepsilon_n^+(R) \leq \lfloor \frac{n-1}{2} \rfloor / n$. Denote $m = \lfloor \frac{n-1}{2} \rfloor$. We replace the points $z_i = (x_i, y_i)$ by $z'_i = (x_i, y_i + L^i)$ for $i = 1, \dots, m$ and $L > 0$. Then we can easily verify that the height of a triangle containing at least one contaminated point can be made arbitrarily large by letting L go to infinity. We now take some z_i from the original sample. If z'_j is a contaminated point, it is clear that $\text{med}_k h(z_i, z'_j, z_k) \rightarrow \infty$. If z_j is an original point and z'_k a contaminated one, then $h(z_i, z_j, z'_k) \rightarrow \infty$ as well. Because this is true for $\lfloor \frac{n-2}{2} \rfloor$ of the $n-2$ possible values of k , $\text{med}_k h(z_i, z_j, z_k) \rightarrow \infty$. Both cases lead to $\text{med}_j \text{med}_k h(z_i, z_j, z_k) \rightarrow \infty$. Now there are at least $\lfloor \frac{n}{2} \rfloor$ original points z_i , which implies that R can be made arbitrary large.

On the other hand, $\varepsilon_n^+(R) \geq \lfloor \frac{n-1}{2} \rfloor / n$. Take any sample Z' where at least $n-m+1 = \lfloor \frac{n}{2} \rfloor + 2$ of the original points are kept. For each triplet (z_i, z_j, z_k) of original points, $h(z_i, z_j, z_k) \leq \max_{i < j < k} h(z_i, z_j, z_k) = M < \infty$. We can then verify that the number of original points is big enough, so that $R \leq c_2 M < \infty$. For this we note that the median of n numbers is bounded as soon as this is the case for $\lfloor \frac{n}{2} \rfloor + 1$ of them.

Secondly, $\varepsilon_n^-(R) \leq \lfloor \frac{n}{2} \rfloor / n$. We construct a new sample Z' by setting $z_1 = z_2 = \dots = z_{m+1}$ with $m = \lfloor \frac{n}{2} \rfloor$. Fix z_i with $1 \leq i \leq m+1$. For $j \neq i$ with $1 \leq j \leq m+1$ it holds that $\text{med}_k h(z_i, z_j, z_k) = 0$. When $m+2 \leq j \leq n$ it follows that $h(z_i, z_j, z_k) = 0$ for all $1 \leq k \leq m+1$. Because the median of n positive numbers equals zero as soon as $\lfloor \frac{n}{2} \rfloor + 1$ of them are zero, we can again verify that $\text{med}_k h(z_i, z_j, z_k) = 0$. Therefore, $\text{med}_j \text{med}_k h(z_i, z_j, z_k) = 0$. This holds for $m+1$ values of i , and thus $R = 0$.

Finally we prove that R does not implode at a contaminated sample Z' with only $m = \lfloor \frac{n}{2} \rfloor - 1$ points changed. Denote $\delta = \frac{1}{2} \min_{i < j < k} h(z_i, z_j, z_k)$ where the minimum is taken over all the points of the original sample Z . Because they are in general position,

$\delta > 0$. If we take (z_i, z_j, z_k) from the original sample, it is clear that $h(z_i, z_j, z_k) > \delta$. It follows that $R > c_2\delta > 0$, by noting that the median of n numbers is larger than δ if at least $\lceil \frac{n}{2} \rceil$ of them are.

Proof of Theorem 2. To prove that $\varepsilon^+(Q_{all}^\alpha) = 1 - \sqrt[3]{\alpha}$ we note that we can always replace m points such that only the heights of the triangles consisting of original points are bounded. We can therefore use the same construction as in the proof of Theorem 1. This implies that Q_{all}^α explodes if and only if

$$\binom{n}{3} - \binom{n-m}{3} \geq \binom{n}{3} - \left\lceil \alpha \binom{n}{3} \right\rceil + 1. \quad (8.1)$$

Putting $\varepsilon = \frac{m}{n}$ and taking the limit for $n \rightarrow \infty$ leads to the cubic inequality

$$\varepsilon^3 - 3\varepsilon^2 + 3\varepsilon + (\alpha - 1) \geq 0. \quad (8.2)$$

The smallest positive solution of (8.2) then corresponds to $\varepsilon^+(Q_{all}^\alpha) = 1 - \sqrt[3]{\alpha}$.

Next, we have to prove that $\varepsilon^-(Q_{all}^\alpha)$ satisfies (3.1). If we replace m points by setting $z_1 = z_2 = \dots = z_{m+1}$ we see that $\binom{m+1}{3} + \binom{m+1}{2}(n-m-1)$ heights become zero. Therefore, we must find the smallest m such that

$$\binom{m+1}{3} + \binom{m+1}{2}(n-m-1) \geq \left\lceil \alpha \binom{n}{3} \right\rceil. \quad (8.3)$$

With $\varepsilon = \frac{m}{n}$, taking limits in (8.3) yields

$$\varepsilon^3 - \frac{3}{2}\varepsilon^2 + \frac{\alpha}{2} \geq 0. \quad (8.4)$$

Using the formulas of Cardano (see, e.g., van der Waerden 1971) we obtain (3.1) as the smallest positive solution of (8.4).

It remains to show that replacing m points, where

$$\binom{m+1}{3} + \binom{m+1}{2}(n-m-1) \leq \left\lceil \alpha \binom{n}{3} \right\rceil - 1, \quad (8.5)$$

keeps Q_{all}^α strictly positive. It can be verified that there exists a positive δ_1 such that at least $m \binom{n-m-1}{2}$ of the heights containing one contaminated point are larger than δ_1 . If $\delta_2 > 0$ denotes the smallest height in the original sample, we can conclude that at least $k = \binom{n-m}{3} + m \binom{n-m-1}{2}$ heights in the new sample are larger than $\delta = \frac{1}{2} \min\{\delta_1, \delta_2\}$. Because (8.5) implies that $k \geq \binom{n}{3} - \left\lceil \alpha \binom{n}{3} \right\rceil + 1$, we conclude that $Q_{all}^\alpha > c_1\delta > 0$.

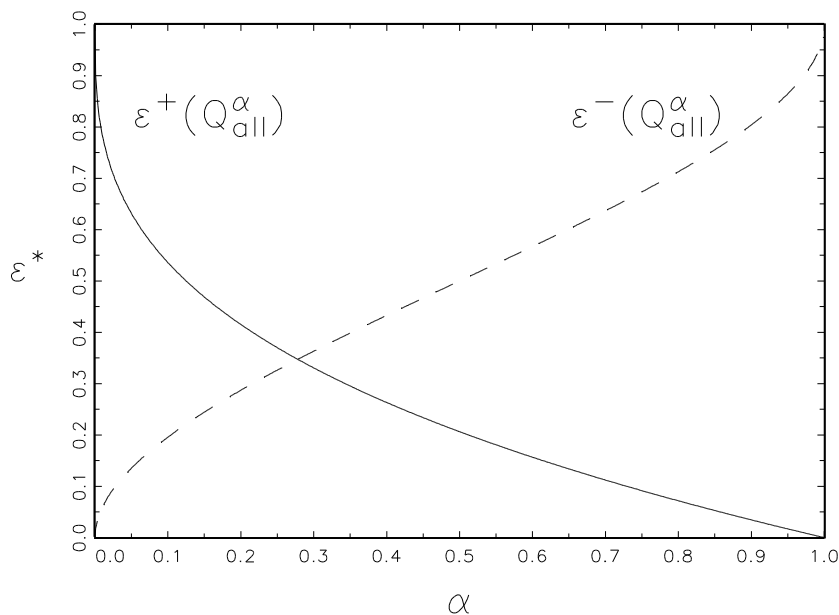


Figure 8: Implosion and explosion breakdown point of Q_{all}^α as a function of α

Figure 8 shows $\varepsilon^+(Q_{all}^\alpha)$ and $\varepsilon^-(Q_{all}^\alpha)$ as a function of α . From a numerical calculation it follows that $\varepsilon^*(Q_{all}^\alpha) = \min\{\varepsilon^+(Q_{all}^\alpha), \varepsilon^-(Q_{all}^\alpha)\}$ reaches its maximal value $\varepsilon^*(Q_{all}^\alpha) = 34.7\%$ for $\alpha = 0.278$.

Proof of Theorem 3. We may assume that the original data points are ordered in the x -direction, hence $x_i \leq x_j$ if $i < j$. First we show that $\varepsilon_n^+(Q_{adj}^\alpha) \leq \left\lceil \frac{(n-1) - [\alpha(n-2)]}{3} \right\rceil / n$. Suppose we replace the points z_{3i} for all $i = 1, \dots, m = \left\lceil \frac{(n-1) - [\alpha(n-2)]}{3} \right\rceil$ by $z'_{3i} = (x_{3i}, y_{3i} + L)$. Then the height of any triangle containing one of these new points becomes arbitrarily large when L grows. If $m < \left\lceil \frac{n}{3} \right\rceil$ then z'_{3i} always belongs to 3 triangles. Because $3m \geq (n-1) - [\alpha(n-2)]$ it follows that Q_{adj}^α will explode. If $m = \left\lceil \frac{n}{3} \right\rceil$, each triangle has exactly one contaminated vertex, so that Q_{adj}^α of course becomes unbounded as well.

We can also verify that $\varepsilon_n^+(Q_{adj}^\alpha) \geq m/n$. Indeed, when we replace at most $m-1$ points, they can at most change $3(m-1)$ heights. Now $3(m-1) \leq (n-2) - [\alpha(n-2)]$, so that at least $[\alpha(n-2)]$ heights remain bounded. We may thus conclude that $\varepsilon_n^+(Q_{adj}^\alpha) = m/n$.

For the proof of the implosion breakdown point, we construct a new sample Z' by setting $z'_{i+2} = z_{i+1}$ for $i = 1, \dots, m = \left\lceil \frac{[\alpha(n-2)]}{2} \right\rceil$. Then $h(z_i, z_{i+1}, z'_{i+2}) = 0 = h(z_{i+1}, z'_{i+2}, z_{i+3})$ for all $i = 1, \dots, m$. So at least $[\alpha(n-2)]$ heights become zero, hence $Q_{adj}^\alpha = 0$. This implies that $\varepsilon_n^-(Q_{adj}^\alpha) \leq m/n$.

Finally, we need to show that $\varepsilon_n^-(Q_{adj}^\alpha) \geq \left\lceil \frac{[\alpha(n-2)]}{2} \right\rceil / n$. This is based on the fact that by replacing m points at most $2m$ heights can become arbitrarily small. To prove this we consider two situations :

1. a replaced point belongs to 3 triangles, for the remainder formed by original points.
2. two replaced points together belong to 5 triangles of otherwise original points.

In both situations we can prove that at least one height of these new triangles has to be larger than a positive constant δ_1 which only depends on the original sample. This result is due to the fact that no three original points are collinear. As in the proof of Theorem 2 we will leave out the details. In all other situations k new points belong to at most $2k$ triangles, in accordance with the claim above. Putting $\delta_2 = \min_{i < j < k} h(z_i, z_j, z_k)$ and $\delta = \frac{1}{2} \min\{\delta_1, \delta_2\}$ we find that when replacing at most $m = \left\lceil \frac{[\alpha(n-2)]}{2} \right\rceil - 1$ points it holds that $Q_{adj}^\alpha > c_3 \delta > 0$.

Because $\varepsilon^+(Q_{adj}^\alpha) = \frac{1-\alpha}{3}$ and $\varepsilon^-(Q_{adj}^\alpha) = \frac{\alpha}{2}$ we can easily see that $\varepsilon^*(Q_{adj}^\alpha) = \min\{\varepsilon^+(Q_{adj}^\alpha), \varepsilon^-(Q_{adj}^\alpha)\}$ is at its highest when $\alpha = 0.4$, in which case $\varepsilon^*(Q_{adj}^\alpha) = 20\%$.

Acknowledgement

We would like to thank Luc Meuleners for assistance with the programming, and the referees for their constructive remarks.

References

- Bickel, P.J., and Lehmann, E.L. (1976), Descriptive statistics III: dispersion, *The Annals of Statistics*, 4, 1139-1158.
- Campbell, R.C. (1974), *Statistics for Biologists*, Cambridge: University Press.
- Cheney, E.W. (1966), *Introduction to Approximation Theory*, New York: McGraw-Hill.
- Croux, C., and Rousseeuw, P.J. (1992), Time-efficient algorithms for two highly robust estimators of scale, in *Computational Statistics, Volume 1*, eds. Y. Dodge and J. Whittaker, Heidelberg: Physika-Verlag, 411-428.

- Donoho, D.L., and Huber, P.J. (1983), The notion of breakdown point, in *A Festschrift for Erich Lehmann*, eds. P. Bickel, K. Doksum, and J.L. Hodges, Jr., California: Wadsworth.
- Ezekiel, M., and Fox, K.A. (1959), *Methods of Correlation and Regression Analysis*, New York: John Wiley.
- Gasser, T., Stroka, L. and Jenner, C. (1986), Residual variance and residual pattern in nonlinear regression, *Biometrika*, 73, 625-633.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986), *Robust Statistics: the Approach based on Influence Functions*, New York: John Wiley.
- Hawkins, D.M., Simonoff, J.S., and Stromberg, A.J. (1994), Distributing a computationally intensive estimator: the case of exact LMS regression, *Computational Statistics*, 9, 83-95.
- Hössjer, O. (1994), Incomplete generalized L-statistics, *Technical Report, Lund Institute of Technology, Sweden*.
- Hössjer, O., Rousseeuw, P.J., and Croux, C. (1994), Asymptotics of the repeated median slope estimator, *The Annals of Statistics*, 22, 1478-1501.
- Hössjer, O., Rousseeuw, P.J., and Croux, C. (1995), Asymptotics of an estimator of a robust spread functional, *Technical Report, University of Antwerp*.
- Müller, H.G., and Stadtmüller, U. (1993), On variance function estimation with quadratic forms, *Journal of Statistical Planning and Inference*, 35, 213-231.
- Rice, J. (1984), Bandwidth choice for nonparametric regression, *The Annals of Statistics*, 12, 1215-1230.
- Rousseeuw, P.J. (1984), Least median of squares regression, *Journal of the American Statistical Association*, 79, 871-880.
- Rousseeuw, P.J., and Croux, C. (1992), Explicit scale estimators with high breakdown point, in *L₁-Statistical Analysis and Related Methods*, ed. Y. Dodge, Amsterdam: North-Holland, 77-92.
- Rousseeuw, P.J., and Croux, C. (1993), Alternatives to the median absolute deviation, *Journal of the American Statistical Association*, 88, 1273-1283.

- Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley.
- Sen, P.K. (1968), Estimates of the regression coefficient based on Kendall's tau, *Journal of the American Statistical Association*, 63, 1379-1389.
- Serfling, R.J. (1984), Generalized L-, M-, and R-Statistics, *The Annals of Statistics*, 12, 76-86.
- Siegel, A.F. (1982), Robust regression using repeated medians, *Biometrika*, 69, 242-244.
- Theil, H. (1950), A rank-invariant method of linear and polynomial regression analysis, I, II and III, *Koninklijke Nederlandse Akademie van Wetenschappen, Proceedings* 53, 386-392, 521-525, 1397-1412.
- van der Waerden, B.L. (1971), *Algebra 1*, Berlin: Heidelberger Taschenbücher.