

The Catline for Deep Regression

Mia Hubert and Peter J. Rousseeuw

July 10, 1997

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

<http://win-www.uia.ac.be/u/statis>

Abstract

Motivated by the notion of *regression depth* (Rousseeuw and Hubert 1996) we introduce the *catline*, a new method for simple linear regression. At any bivariate data set $Z_n = \{(x_i, y_i); i = 1, \dots, n\}$ its regression depth is at least $n/3$. This lower bound is attained for data lying on a convex or concave curve, whereas for perfectly linear data the catline attains a depth of n . We construct an $O(n \log n)$ algorithm for the catline, so it can be computed fast in practice. The catline is Fisher-consistent at any linear model $y = \beta x + \alpha + e$ in which the error distribution satisfies $\text{med}(e|x) = 0$, which encompasses skewed and/or heteroscedastic errors. The breakdown value of the catline is $1/3$, and its influence function is bounded. At the bivariate gaussian distribution its asymptotic relative efficiency compared to the L^1 line is 79.3% for the slope, and 88.9% for the intercept. The finite-sample relative efficiencies are in close agreement with these values. This combination of properties makes the catline an attractive fitting method.

Keywords and phrases: Algorithm; Breakdown value; Heteroscedasticity; Influence function; Regression depth; Robust regression.

AMS classification numbers: 62F35, 62J05

1 Introduction

Consider any bivariate data set $Z_n = \{(x_i, y_i); i = 1, \dots, n\}$ and any straight line of the form $y = bx + a$. Recently, Rousseeuw and Hubert (1996) introduced the *regression depth* of such a line relative to Z_n . The regression depth is an integer between 0 and n , and can be seen as a kind of ‘rank’ of the line. This allows one to compare different lines, from the viewpoint that a deeper line provides a better fit to the data.

In this paper we construct the *catline*, a new regression method which is motivated by regression depth. At any data set Z_n , the regression depth of the catline is at least $n/3$. This lower bound is attained when the data points lie on a strictly convex (or strictly concave) curve. If, on the other hand, the data points lie exactly on a straight line, then the catline’s depth attains the upper bound of n .

Section 2 outlines the notion of regression depth, whereas Section 3 defines the catline and gives its depth and equivariance properties. In Section 4 we construct an $O(n \log n)$ algorithm for the catline, so that it can easily be computed in practice. Section 5 shows that the catline is Fisher-consistent at any linear model $y = \beta x + \alpha + e$ for which $\text{med}(e|x) = 0$ at all x , which allows for asymmetric and/or heteroscedastic errors. In Section 6 the robustness of the catline is investigated. Its breakdown value is $1/3$, meaning that up to one third of the data points (x_i, y_i) may be replaced by outliers (in both the x -direction and the y -direction) without destroying the fit. Moreover, the influence functions of the slope and the intercept are bounded. Finally, Section 7 derives the efficiency properties. At a bivariate gaussian distribution the asymptotic relative efficiency of the slope compared to the L^1 line is 79.3%, and the intercept attains 88.9%. A simulation study confirms these efficiencies also for finite samples. This combination of properties makes the catline an attractive fitting method.

2 Regression depth

We start from a data set $Z_n = \{(x_i, y_i); i = 1, \dots, n\} \subset \mathbb{R}^2$. Each line of the form $y = bx + a$ will be considered as a ‘candidate fit’ to Z_n and denoted as $\boldsymbol{\theta} = (b, a)$ so the first component is the slope and the second is the intercept. The residuals of Z_n relative to $\boldsymbol{\theta}$ will be denoted as $r_i = r_i(\boldsymbol{\theta}) = y_i - bx_i - a$. In order to introduce the depth of a fit, we will first define a *nonfit*.

Definition 1. A candidate fit $\boldsymbol{\theta} = (b, a)$ to Z_n is called a **nonfit** iff there exists a real number $v_\theta = v$ which does not coincide with any x_i , and such that

$$r_i(\boldsymbol{\theta}) < 0 \quad \text{for all } x_i < v \quad \text{and} \quad r_i(\boldsymbol{\theta}) > 0 \quad \text{for all } x_i > v$$

(2.1)

or

$$r_i(\boldsymbol{\theta}) > 0 \quad \text{for all } x_i < v \quad \text{and} \quad r_i(\boldsymbol{\theta}) < 0 \quad \text{for all } x_i > v.$$

Figure 1a shows a data set with 7 observations and two nonfits $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Also the corresponding values v_θ and v_η are indicated. From this plot it is clear that the existence of v corresponds to the presence of a tilting point (marked by a cross) around which we can rotate the line until it is vertical, while not passing any observation.

Note that Definition 1 also allows for data sets Z_n with ties among the x_i . (This is because v_θ may not coincide with any x_i).

Definition 2. The **regression depth** $rdepth(\boldsymbol{\theta}, Z_n)$ of a fit $\boldsymbol{\theta} = (b, a)$ to a data set $Z_n \subset \mathbb{R}^2$ is the smallest number of observations that need to be removed to make $\boldsymbol{\theta}$ a nonfit.

Definitions 1 and 2 are due to Rousseeuw and Hubert (1996), who also provide equivalent definitions in dual space. Moreover, they show several analogies with the notion of location depth (Tukey 1975, Donoho and Gasko 1992, He and Wang 1997). However, these aspects will not be pursued in the present paper.

To illustrate Definition 2, consider the lines $\boldsymbol{\tau}$ and $\boldsymbol{\xi}$ in Figure 1b. We can make $\boldsymbol{\tau}$ a nonfit by removing observations 2 and 6 (since one can then rotate $\boldsymbol{\tau}$ about v_τ without touching any remaining observations). Since $\boldsymbol{\tau}$ cannot be made a nonfit by removing fewer observations, $rdepth(\boldsymbol{\tau}, Z_n) = 2$. The line $\boldsymbol{\xi}$ has $rdepth$ 3, since we need to remove three observations (1, 4 and 6) before it becomes a nonfit. Note that a nonfit never passes through an observation (since all residuals in (2.1) are strictly positive or strictly negative). Therefore, a line through k observations has a regression depth of at least k .

It can easily be verified that regression depth is scale invariant, regression invariant and affine invariant, according to the definitions in Rousseeuw and Leroy (1987, page 116).

In order to compute $rdepth(\boldsymbol{\theta}, Z_n)$ we first sort the observations by their x_i coordinates in $O(n \log n)$ time. Next, we denote all the *distinct* x -values by $\tilde{x}_1 < \dots < \tilde{x}_j < \dots < \tilde{x}_{\tilde{n}}$ with $\tilde{n} \leq n$. (If there are no ties, $\tilde{n} = n$.) Then we put $v_1 = \tilde{x}_1 - 1$ and $v_j = (\tilde{x}_{j-1} + \tilde{x}_j)/2$

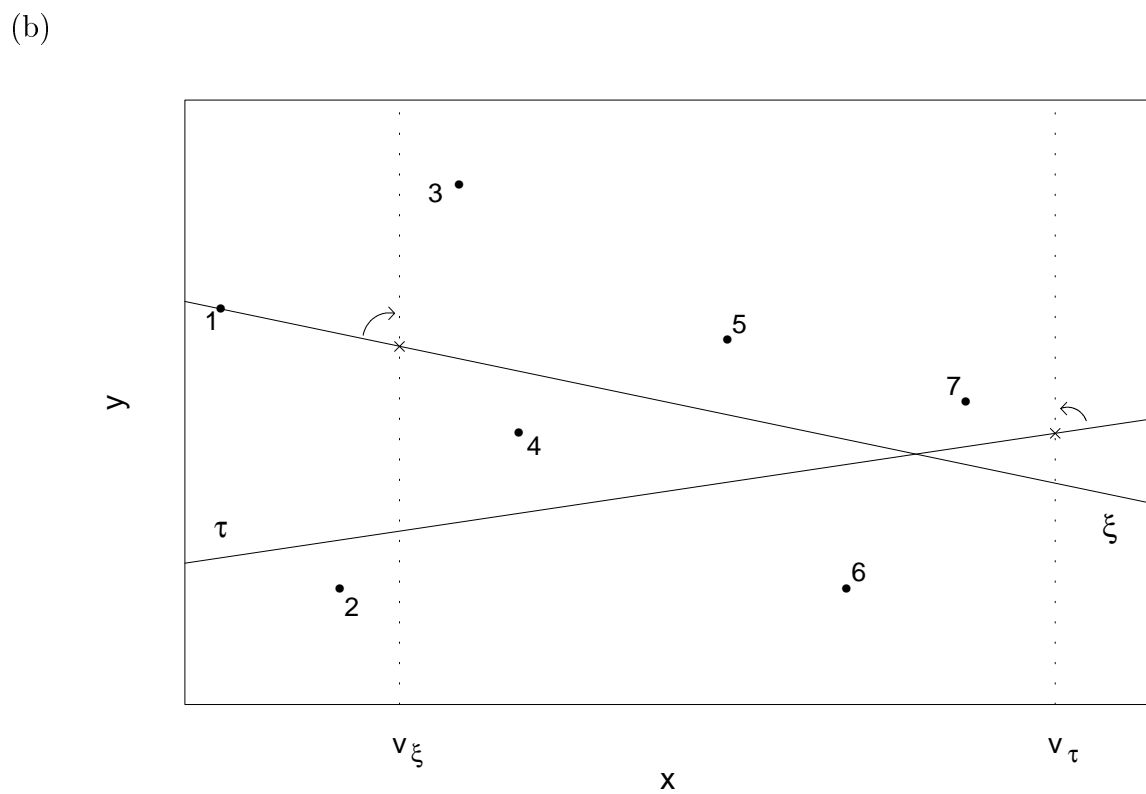
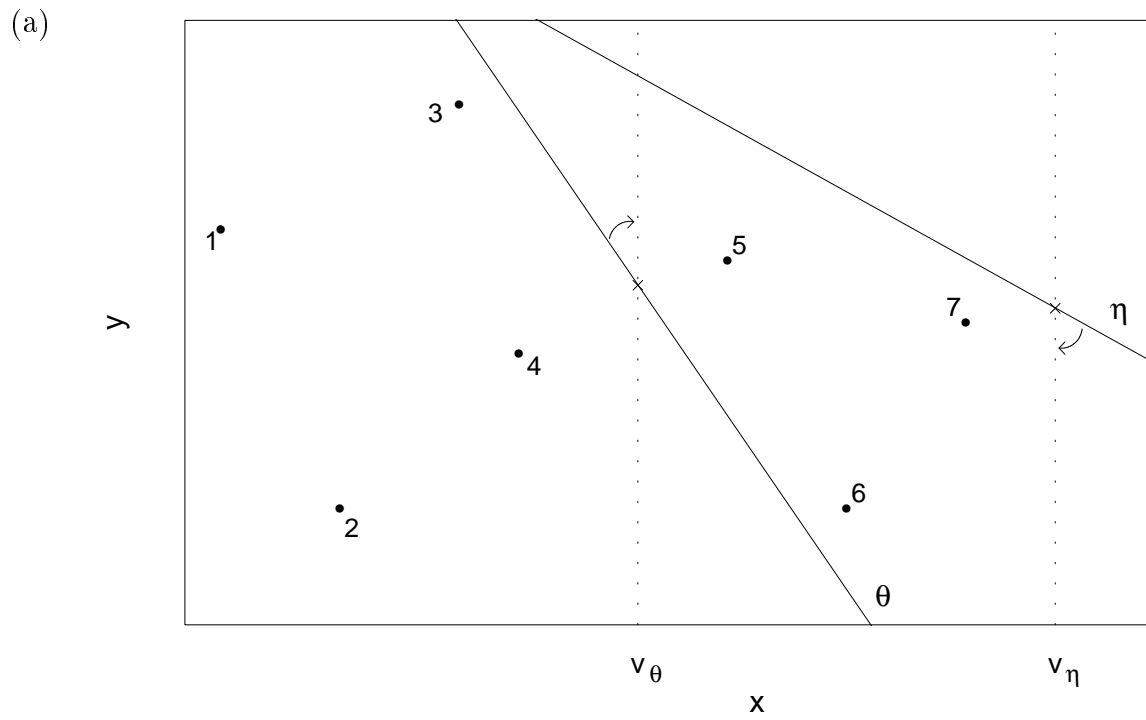


Figure 1: (a) Data set with 7 observations and two nonfits θ and η ; (b) same data set with a fit τ with regression depth 2 and a fit ξ with regression depth 3.

for all $2 \leq j \leq \tilde{n}$. The regression depth of $\boldsymbol{\theta}$ relative to Z_n is then computed in $O(n)$ time from the expression

$$rdepth(\boldsymbol{\theta}, Z_n) = \min_{1 \leq j \leq \tilde{n}} (\min\{S^+(v_j) + G^-(v_j), S^-(v_j) + G^+(v_j)\})$$

where

$$S^+(t) = \#\{i; x_i < t \text{ and } r_i \geq 0\}, \quad G^-(t) = \#\{i; x_i > t \text{ and } r_i \leq 0\},$$

and $S^-(t)$ and $G^+(t)$ are defined accordingly. It therefore suffices to update $S^+(v_j)$, $S^-(v_j)$, $G^-(v_j)$ and $G^+(v_j)$ at each $j = 1, \dots, \tilde{n}$.

We can also consider the regression depth of a fit $\boldsymbol{\theta} = (b, a)$ in the *population* case. We then assume that the random variables (X, Y) have a joint distribution H on \mathbb{R}^2 .

Definition 3. A candidate fit $\boldsymbol{\theta} = (b, a)$ to $(X, Y) \sim H$ is called a *nonfit* iff there exists a real number $v_\theta = v$ with $P(X = v) = 0$ such that

$$P(Y - bX - a > 0 \mid X < v) = 1 \quad \text{and} \quad P(Y - bX - a < 0 \mid X > v) = 1$$

or

$$P(Y - bX - a < 0 \mid X < v) = 1 \quad \text{and} \quad P(Y - bX - a > 0 \mid X > v) = 1.$$

The regression depth $rdepth(\boldsymbol{\theta}, H)$ is defined as the smallest probability mass that has to be removed to make $\boldsymbol{\theta}$ a nonfit.

3 The catline

In this section we will construct the catline and give its depth properties.

First we sort the data set $Z_n = \{(x_i, y_i); i = 1, \dots, n\}$ according to its x -values, so we may assume that $x_1 \leq x_2 \leq \dots \leq x_n$. (If ties in x_i occur, we sort the observations with identical x -values according to their y -value.) Then we divide the data set in three groups denoted by L, M and R . If n is a multiple of 3, the left group L is formed by the first $m = n/3$ data points $\{(x_1, y_1), \dots, (x_m, y_m)\}$, the group M by the middle third, and R by the rightmost third. For $n = 3m + 1$ we take $\#M = m + 1$, whereas for $n = 3m + 2$ we take $\#L = \#R = m + 1$.

Definition 4. The **catline** $\boldsymbol{\theta}_{CAT} = (b_{CAT}, a_{CAT})$ is the line $y = b_{CAT}x + a_{CAT}$ that simultaneously bisects $L \cup M$ and $M \cup R$.

(We say that a line bisects a set of N points if neither of the two open halfplanes defined by that line contains more than $\lfloor N/2 \rfloor$ points. If N is odd, the line thus must pass through at least one point). The existence of a simultaneous bisector of two finite sets in \mathbb{R}^2 follows from the Borsuk-Ulam theorem (see Edelsbrunner 1987, page 69). In the population case the catline partitions the probability mass according to (3.1), where the horizontal line indicates the catline and the vertical lines indicate the sets L , M and R :

$$\frac{q}{\frac{1}{3} - q} \left| \frac{\frac{1}{3} - q}{q} \right| \frac{q}{\frac{1}{3} - q} \quad (3.1)$$

For each distribution H on \mathbb{R}^2 , there is a unique value of $0 \leq q \leq \frac{1}{3}$ satisfying (3.1).

Roughly speaking, the catline has the property that the number of positive residuals in L equals the number of negative residuals in M and the number of positive residuals in R . We call it the catline since it Cuts All Thirds (that is, L , M , and R).

Figure 2 shows a data set with 12 observations and its catline. We have also indicated the three groups L , M and R . Here we have three positive residuals in L and in R , and three negative residuals in M .

Let us denote L^+ (resp. L^-) as the number of strictly positive (resp. strictly negative) residuals in L from a fit (b, a) , and define M^+ , M^- , R^+ , and R^- analogously.

Theorem 1. *For any data set, a necessary and sufficient condition for (b, a) to be the catline is that*

$$L^+ + M^+ \leq \lfloor n/3 \rfloor, \quad L^- + M^- \leq \lfloor n/3 \rfloor, \quad M^+ + R^+ \leq \lfloor n/3 \rfloor, \quad M^- + R^- \leq \lfloor n/3 \rfloor.$$

Note that if $n \neq 3m$ both $L \cup M$ and $M \cup R$ contain an odd number of observations, hence at least one residual is zero.

Theorem 2. *At any data set $Z_n \subset \mathbb{R}^2$,*

$$\left\lceil \frac{n}{3} \right\rceil \leq rdepth(\boldsymbol{\theta}_{CAT}, Z_n) \leq n.$$

Moreover, for any (X, Y) -distribution H on \mathbb{R}^2 it holds that

$$\frac{1}{3} \leq rdepth(\boldsymbol{\theta}_{CAT}, H) \leq 1.$$

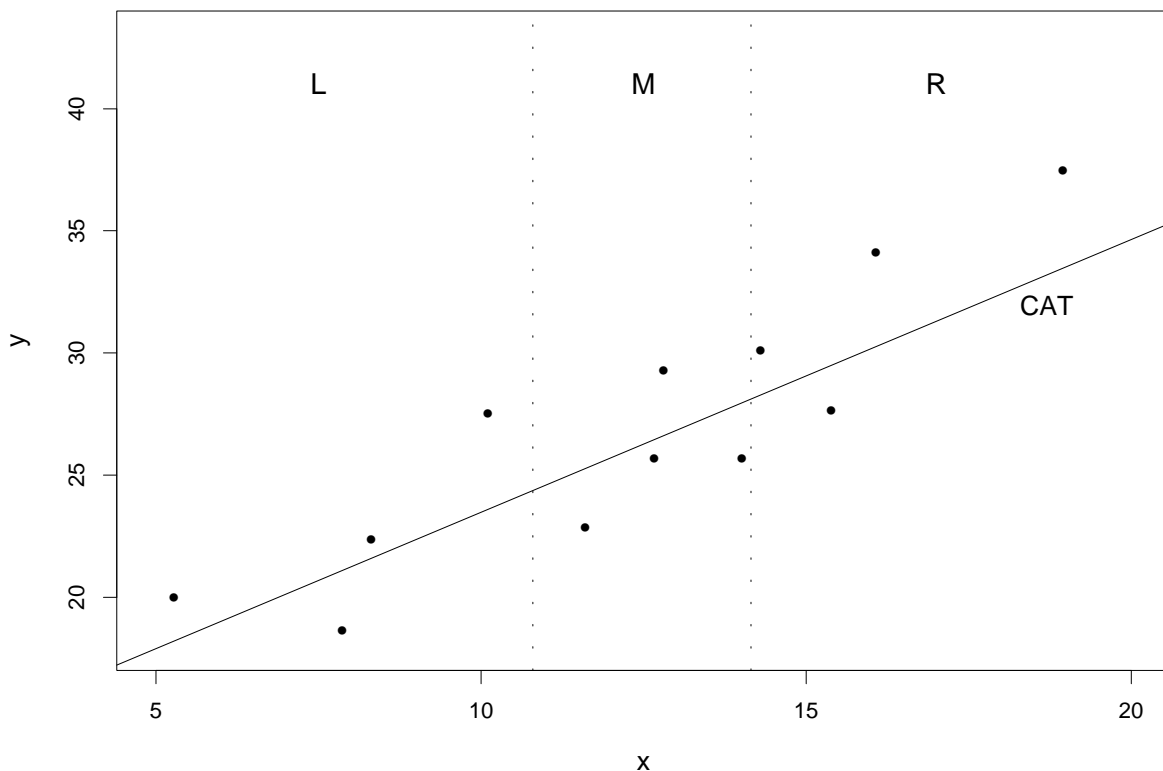


Figure 2: Example of a catline, for a data set of 12 observations. The number of positive residuals in L and in R equals the number of negative residuals in M (namely 3). The regression depth of $\boldsymbol{\theta}_{CAT}$ is 5.

(All proofs can be found in the Appendix.) Let us look again at Figure 2. The regression depth of $\boldsymbol{\theta}_{CAT}$ is 5. (For instance, we can remove the five negative residuals and use a v_θ to the left of the x -values.) In this example, $rdepth(\boldsymbol{\theta}_{CAT}, Z_n)$ is strictly larger than $\lceil n/3 \rceil = 4$.

From Definition 2 and Theorem 2 it follows that

$$\frac{1}{3} \leq \max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, H) \leq 1. \quad (3.2)$$

The following theorem shows that these bounds are sharp.

Theorem 3. (a) *If the observations of the data set Z_n have different x -values and lie on a strictly convex (or strictly concave) curve, then*

$$rdepth(\boldsymbol{\theta}_{CAT}, Z_n) = \max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_n) = \left\lceil \frac{n+2}{3} \right\rceil.$$

(b) If the probability mass of H is concentrated on a strictly convex (or concave) curve,

$$rdepth(\boldsymbol{\theta}_{CAT}, H) = \max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, H) = \frac{1}{3}.$$

(c) If all observations of Z_n lie on a straight line $y = \beta x + \alpha$, then $\boldsymbol{\theta}_{CAT} = (\beta, \alpha)$ and

$$rdepth(\boldsymbol{\theta}_{CAT}, Z_n) = \max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, Z_n) = n.$$

(d) If H is concentrated on a straight line $y = \beta x + \alpha$, then $\boldsymbol{\theta}_{CAT} = (\beta, \alpha)$ and

$$rdepth(\boldsymbol{\theta}_{CAT}, H) = \max_{\boldsymbol{\theta}} rdepth(\boldsymbol{\theta}, H) = 1.$$

To illustrate (a), Figure 3 shows a data set of 11 observations that lie on the convex curve $y = e^x$. Here the catline attains the maximal rdepth of 5.

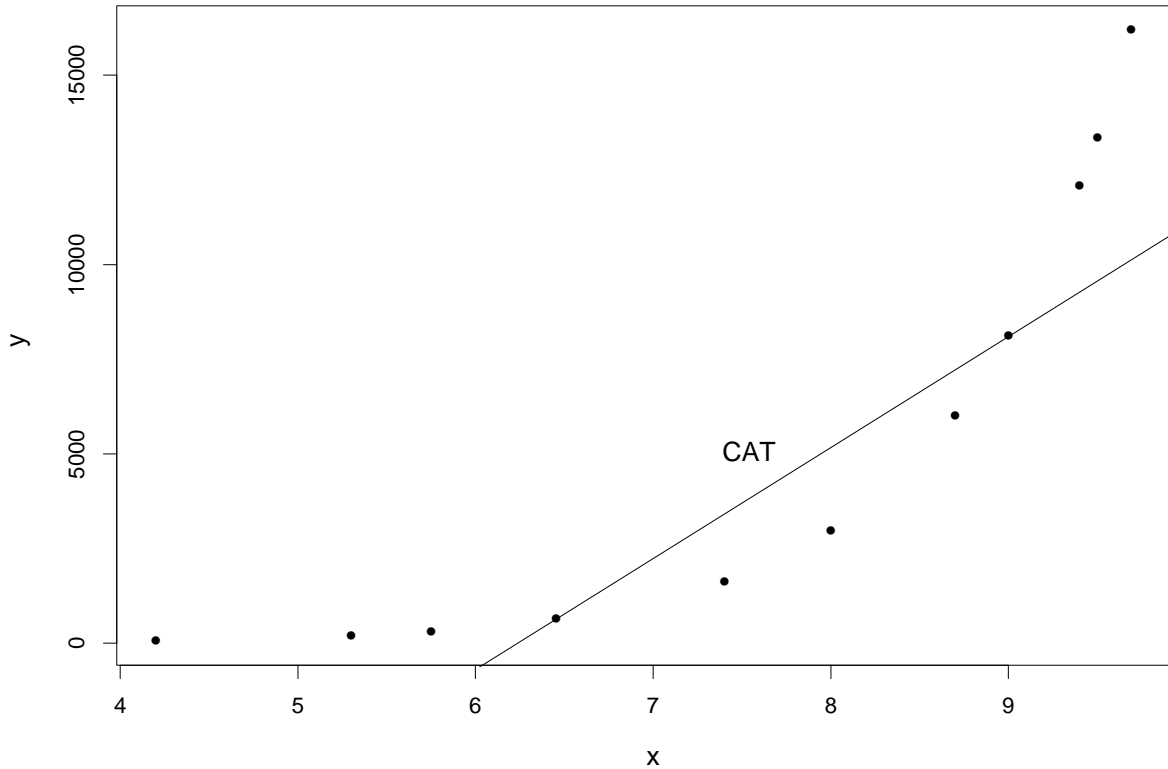


Figure 3: Plot of a data set with 11 observations that lie on the convex curve $y = e^x$. The catline has depth 5, which is the maximal rdepth for this data set.

Finally we give some equivariance and invariance properties of the catline. For convenience we denote $Z_n = (\mathbf{x}, \mathbf{y})$ where $\mathbf{x} = (x_1, \dots, x_n)^t$ and $\mathbf{y} = (y_1, \dots, y_n)^t$ are column vectors.

Theorem 4. (1) *The catline is regression equivariant, i.e. $\boldsymbol{\theta}_{CAT}(\mathbf{x}, \mathbf{y} + c\mathbf{x} + d) = \boldsymbol{\theta}_{CAT}(\mathbf{x}, \mathbf{y}) + (c, d)$ for any constants c and d .*

(2) *The catline is scale equivariant, i.e. $\boldsymbol{\theta}_{CAT}(\mathbf{x}, c\mathbf{y}) = c\boldsymbol{\theta}_{CAT}(\mathbf{x}, \mathbf{y})$ for any constant c .*

(3) *The catline is affine equivariant: put $(\hat{b}, \hat{a}) := \boldsymbol{\theta}_{CAT}(\mathbf{x}, \mathbf{y})$ and take any constants c and d with $c \neq 0$, then $\boldsymbol{\theta}_{CAT}((\mathbf{x} - d)/c, \mathbf{y}) = (c\hat{b}, \hat{a} + d\hat{b})$.*

(4) *The catline is invariant to changing the absolute magnitudes of its residuals as long as their signs remain the same.*

The fourth property reflects the fact that the catline is only defined through the signs of the corresponding residuals. We may thus enlarge (or shrink) their magnitude without affecting the estimate. This suggests that the catline will be resistant to vertical outliers in the data set, i.e. towards outlying y -values. Other estimators with this property include the line of Brown and Mood (1951), the least absolute deviations (L^1) method (see Bassett and Koenker 1978), and the resistant line (Tukey 1977, Johnstone and Velleman 1985). For a more detailed investigation of the robustness of the catline, we refer to Section 6.

Example: the pronghorn data. This data set of size $n = 29$ compares a habitat suitability index with pronghorn prevalence in 29 winter ranges in the western United States. In a previous analysis Cade and Richards (1996) used a shifted power transformation on the response variable to obtain a linear relationship (see also Box and Cox 1964, Carroll and Ruppert 1988). The transformed data set $(x_i, g(y_i))$ with $g(y) = \log_{10}(y + 1)$ yields the catline $\boldsymbol{\theta}_{CAT} = (1.75, -0.24)$. Note that in this example the catline lies close to the L^1 line $\boldsymbol{\theta}_{L^1} = (1.79, -0.27)$ obtained by Cade and Richards (1996). In Figure 4 we have plotted the original data (x_i, y_i) with the backtransformed curve $y = g^{-1}(1.75x - 0.24)$ which we will call the *catfit*. Due to the monotonicity of g , the number of positive and negative residuals in the three groups L , M and R stays the same after this transformation.

Example: the stars data. In general the catline need not be close to the L^1 estimator, which is sensitive to outliers in the x_i (i.e., leverage points). This is illustrated in Figure 5, which contains the Hertzsprung-Russell diagram of a star cluster in the direction of Cygnus (see Rousseeuw and Leroy 1987, page 27). The logarithm of the star's light intensity is

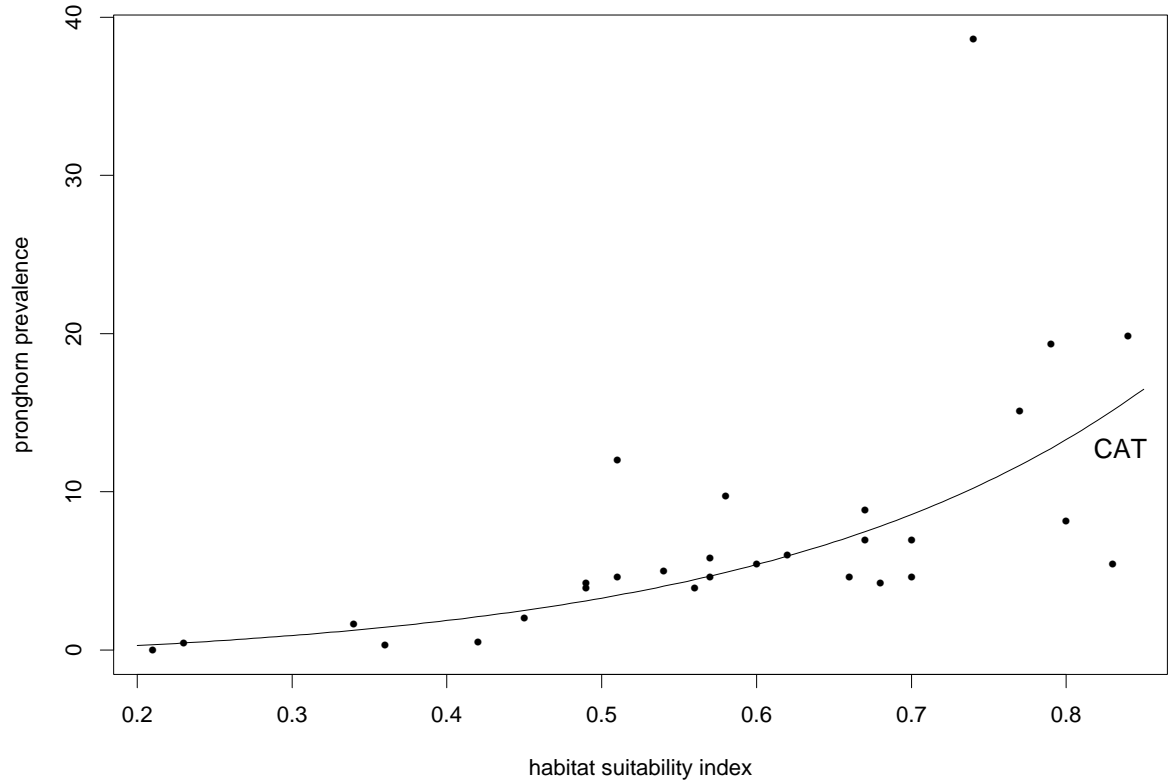


Figure 4: Plot of the pronghorn data with the catfit, obtained by backtransforming the catline of the transformed data $\{(x_i, \log_{10}(y_i + 1)); i = 1, \dots, n\}$.

plotted versus the logarithm of its surface temperature. In this plot we see the catline which fits the main sequence stars, and the L^1 line which is strongly attracted by the four giant stars in the upper right corner.

Remark. The notion of regression depth is related to the halfspace location depth, see (Rousseeuw and Hubert 1996, Section 5). In that paper also a simplicial regression depth $rdepth^{(S)}$ is constructed as a counterpart to the simplicial location depth of Liu (1990, 1995). In simple regression the general definition reduces to

$$rdepth^{(S)}(\boldsymbol{\theta}, Z_n) = \binom{n}{3}^{-1} \sum_{i < j < k} A(r_i(\boldsymbol{\theta}), r_j(\boldsymbol{\theta}), r_k(\boldsymbol{\theta}))$$

where $A(r_i, r_j, r_k)$ is 1 if the residuals r_i, r_j and r_k have alternating signs, and 0 otherwise. It is then easy to verify that as in Theorem 3(b) the catline has maximal $rdepth^{(S)}$ at convex

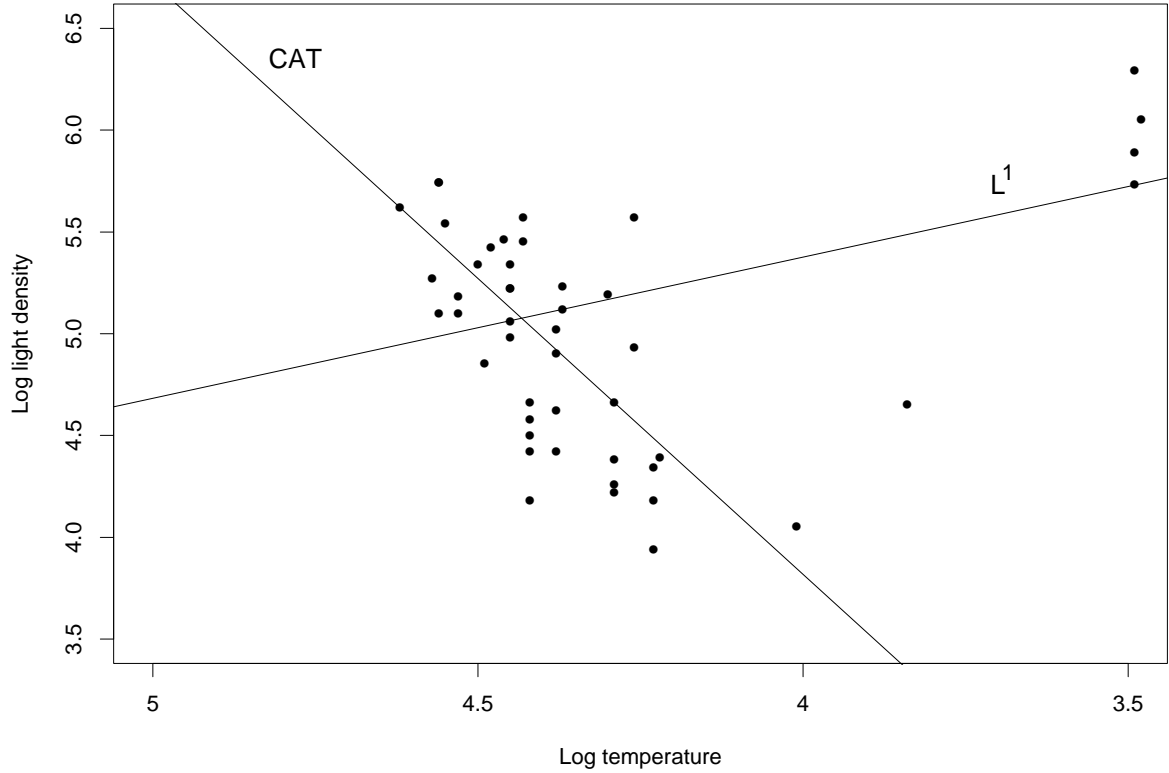


Figure 5: Hertzsprung-Russell diagram of a star cluster in the direction of Cygnus, with the catline and the least absolute deviations fit (L^1) which is attracted by the giant stars.

(concave) curves.

Theorem 5. *If the probability mass of H is concentrated on a strictly convex (or concave) curve, then*

$$rdepth^{(S)}(\boldsymbol{\theta}_{CAT}, H) = \max_{\boldsymbol{\theta}} rdepth^{(S)}(\boldsymbol{\theta}, H) = \left(\frac{1}{3}\right)^3.$$

4 Algorithm

In this section we construct a fast algorithm for computing the catline. It is essentially based on Theorem 6 below, which provides necessary and sufficient conditions for (b, a) to be the catline.

Let us denote by $r_{LM}(b, a)$ and $r_{MR}(b, a)$ the set of residuals in $L \cup M$ and $M \cup R$ relative

to a fit (b, a) . Formally, $r_{LM}(b, a) := \{y_i - bx_i - a; i \in L \cup M\}$. The median of $2k$ values $t_1 \leq \dots \leq t_{2k}$ is taken to be the interval $[t_k, t_{k+1}]$. Then Definition 4 immediately leads to:

Theorem 6. *For any data set $Z_n \subset \mathbb{R}^2$ it holds that*

(b, a) is the catline

\Downarrow

$$\text{med } r_{LM}(b, a) \cap \text{med } r_{MR}(b, a) \neq \emptyset \quad (4.1)$$

\Downarrow

$$a \in \text{med } r_{LM}(b, 0) \cap \text{med } r_{MR}(b, 0). \quad (4.2)$$

When n is not a multiple of 3 it follows that $\#(L \cup M)$ and $\#(M \cup R)$ are odd, hence both medians in (4.2) reduce to a singleton. In that case, the catline is completely characterized by

$$\text{med } r_{LM}(b_{CAT}, 0) = \text{med } r_{MR}(b_{CAT}, 0) \quad (4.3)$$

and

$$a_{CAT} := \text{med } r_{LM}(b_{CAT}, 0). \quad (4.4)$$

To obtain the slope of the catline, we therefore need to solve $f(b) = 0$ for

$$f(b) := \text{med } r_{LM}(b, 0) - \text{med } r_{MR}(b, 0)$$

and verify condition (4.1) if n is a multiple of 3. Afterwards, the intercept can be determined by (4.2) or (4.4).

To solve $f(b) = 0$ we iteratively adjust the slope estimate by means of the ‘zeroin’ algorithm of (Wilkinson 1967; Dekker 1969). This algorithm alternates interpolation and bisection in such a way that the convergence is always ensured, and that the convergence rate is independent of the sample size n (Brent 1973). Our simulations have confirmed this: on average only about four iteration steps were needed to obtain a precision of 6 digits, both for small and large data sets.

Before starting the iteration process, we first need to find two estimates b^0 and b^1 in which f takes on a different sign. For the initial value b^0 we set $l = \#L$, and take in the groups L and R the observation with $\lceil \frac{l}{2} \rceil$ -th smallest y -coordinate. Then b^0 is defined as the slope of the line through these two data points. We compute $a^0 = \text{med } r_{LM}(b^0, 0)$, and

define \hat{q} as the number of positive residuals in L from the fit (b^0, a^0) . We then compute the line through the observations in L and R with $(l - \hat{q})$ -th smallest residual, yielding b^1 . If $f(b^1)f(b^0) > 0$, we take

$$\tilde{b}^0 = \begin{cases} \max(b^0, b^1) & \text{if } f(b^0) < 0 \\ \min(b^0, b^1) & \text{if } f(b^0) > 0 \end{cases}$$

and $\Delta = -|b^1 - b^0| \text{sign}(f(b^0))$. Then we compute $\tilde{b}^1 = \tilde{b}^0 + \Delta$, $\tilde{b}^2 = \tilde{b}^1 + 2\Delta, \dots, \tilde{b}^i = \tilde{b}^{i-1} + 2^{i-1}\Delta, \dots$ until $f(\tilde{b}^j)f(\tilde{b}^0) < 0$ for some j . This will certainly happen after a finite number of steps, since f is continuous, $\lim_{b \rightarrow -\infty} f(b) = -\infty$, and $\lim_{b \rightarrow \infty} f(b) = +\infty$. (Note that f is not always monotone.)

We performed many simulations to investigate the efficiency of this algorithm. We generated data sets of size $n = 3m, n = 3m + 1$ and $n = 3m + 2$ for $m = 16, 160$ and 1600 . The regression slopes were taken to be $0, 5$, and -20 . The errors were normally distributed, but some data sets also contained 10% of outliers. On average, less than one correction step was required to get initial estimates \tilde{b}^0 and \tilde{b}^j with different signs of $f(\tilde{b}^0)$ and $f(\tilde{b}^j)$. The average number of iteration steps in the ‘zeroin’ procedure varied between approximately 3 and 8. A slight increase of this number was observed for large slopes and for large data sets. On the other hand, outliers did not affect the computation time. In conclusion, a small and fixed number of iteration steps was always sufficient. To stay on the safe side, we fixed the maximal number of iterations at 50 in our current implementation. Since we start by ordering the x -coordinates and only perform linear-time operations afterwards, the overall time complexity of this algorithm becomes $O(n \log n)$. This corresponds with the theoretical results of Cole (1984) and Edelsbrunner and Waupotitsch (1986). They derive general algorithms for computing a common bisector of two sets in two dimensions in $O((n_b + n_w) \log(n_b + n_w))$, resp. $O((n_b + n_w) \log(\min(n_b, n_w) + 1))$ time, where n_b and n_w are the number of points in the sets to be bisected.

The S-PLUS code of our algorithm for the catline can be obtained from the website <http://win-www.uia.ac.be/u/statis>.

5 Fisher-consistency at asymmetric and heteroscedastic errors

The population version (functional version) of the catline is straightforward from (4.3) and (4.4). Suppose that (X, Y) has a continuous joint distribution function H . (Throughout, we will use the same notation for a probability distribution and its cdf.) Denote by G the marginal distribution of X . Let $I_1 =]-\infty, G^{-1}(2/3)]$, $I_2 = [G^{-1}(1/3), +\infty[$, and set

$$f(b) = \text{med}(Y - bX \mid X \in I_1) - \text{med}(Y - bX \mid X \in I_2).$$

The functional $T_{CAT}(H) = (b_{CAT}, a_{CAT})(H)$ is then given by

$$f(b_{CAT}) = 0 \tag{5.1}$$

and

$$a_{CAT} = \text{med}(Y - b_{CAT}X \mid X \in I_1). \tag{5.2}$$

Theorem 7. Fisher-consistency. *If $(X, Y) \sim H$ satisfies the linear model*

$$Y = \beta X + \alpha + e \quad \text{with} \quad \text{med}(e|x) = 0 \quad \text{for all } x, \tag{5.3}$$

then $(b_{CAT}, a_{CAT})(H) = (\beta, \alpha)$.

Note that the model (5.3) does not require the errors to be symmetric or identically distributed, and hence allows for skewness and heteroscedasticity. This general form of Fisher-consistency is an important advantage of the catline over regression estimators that are only Fisher-consistent for symmetric and/or homoscedastic errors, like least squares (LS), the L^1 estimator and M-estimators (see also Carroll and Welsh 1988).

Example: skewness and heteroscedasticity. We have generated 60 observations according to the linear model $y = 2x + 1 + e$, where e follows a shifted lognormal distribution given by $e = u - 1$ where $\ln(u) \sim N(0, \sigma_x^2)$, hence $\text{med}(e|x) = 0$ for all x . The heteroscedasticity was given by $\sigma_x = (x + 2)/1.7$. Figure 6 displays the data, together with the catline and the LS fit. The least squares slope is clearly biased, whereas $b_{CAT} \approx 2$ and $a_{CAT} \approx 1$.

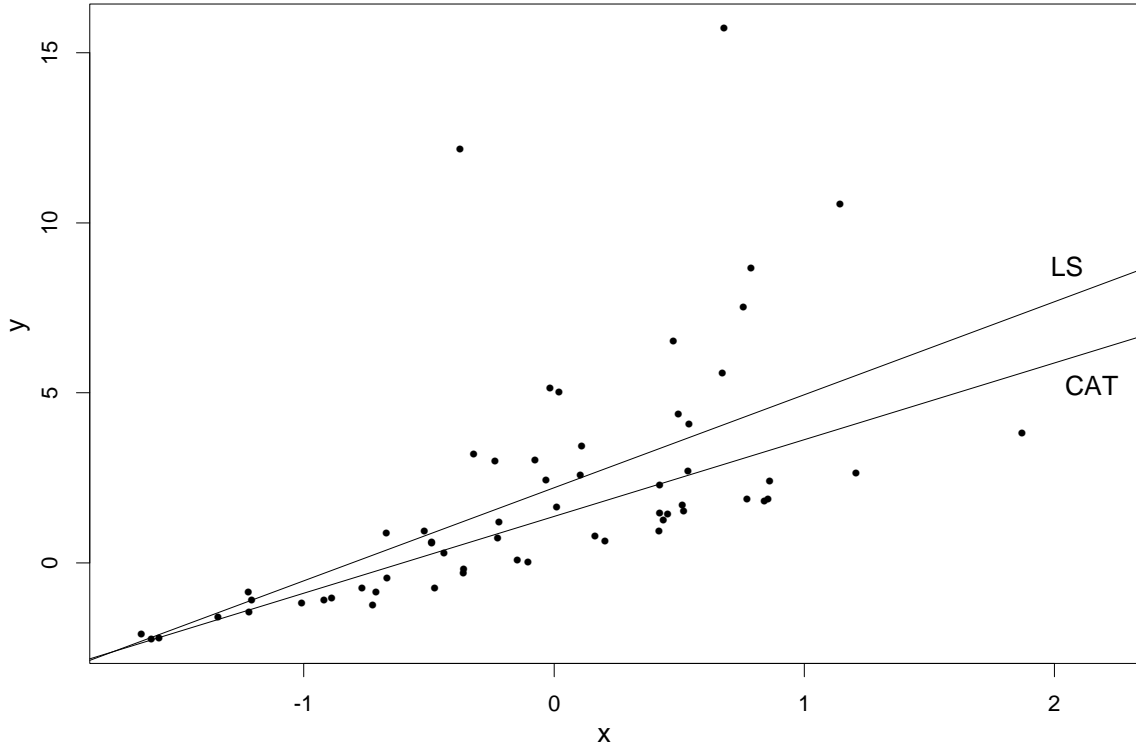


Figure 6: A generated data set ($n = 60$) following a linear model with asymmetric and heteroscedastic errors. In contrast with the catline, the least squares line (LS) is not Fisher-consistent at this model.

6 Robustness properties

6.1 Breakdown value

A well-known measure of an estimator's resistance against outliers is the *breakdown value* (Hampel et al. 1986). The finite-sample breakdown value of any estimator T_n is defined by

$$\varepsilon_n^*(T_n, Z_n) = \min \left\{ \frac{k}{n}; \sup_{Z'_n} \|T_n(Z'_n) - T_n(Z_n)\| = \infty \right\}.$$

Here Z'_n ranges over all data sets obtained by replacing any k observations of Z_n by arbitrary values. The breakdown value is thus the smallest proportion of contaminated observations that can carry the estimator beyond all bounds. Note that this contamination is not restricted to outliers in y_i , but that Z'_n may also contain outliers in x_i .

Theorem 8. *At any data set $Z_n \subset \mathbb{R}^2$ with distinct x_i we have*

$$\varepsilon_n^*(b_{CAT}, Z_n) = \varepsilon_n^*(a_{CAT}, Z_n) = \left(n - 2 \left\lceil \frac{n}{3} \right\rceil - 1 \right) / n.$$

Corollary 1. Exact Fit Property. *When at least $2\lceil \frac{n}{3} \rceil + 2$ points of Z_n lie on a straight line and have distinct x_i values, the catline coincides with that line.*

The asymptotic breakdown value of the catline is therefore $\varepsilon^* = \lim_{n \rightarrow \infty} \varepsilon_n^* = 1/3$. This is an important improvement compared to the resistant line whose breakdown value is $1/6$. This is because the resistant line bisects L and R but ignores M . In fact, the resistant line partitions the data according to

$$\begin{array}{c|c|c} \frac{1}{6} & \text{ignore} & \frac{1}{6} \\ \hline \frac{1}{6} & \text{ignore} & \frac{1}{6} \end{array}$$

using the notation of (3.1). On the other hand, the Brown-Mood estimator divides the x_i in only two sets L and R , and partitions the (x_i, y_i) according to

$$\begin{array}{c|c} \frac{1}{4} & \frac{1}{4} \\ \hline \frac{1}{4} & \frac{1}{4} \end{array}$$

Therefore its breakdown value is $1/4$, which is also below that of the catline.

Example: the Height and Weight data. Figure 7 shows the height (in cm) and weight (in kg) of 30 eleven-year-old girls attending Heaton Middle School in Bradford (Hand et al. 1994). Superimposed are the catline, the least squares line (with zero breakdown value), and the resistant line. We see that the catline is the least attracted by the outliers with large y_i .

6.2 Influence function

The influence function (see Hampel et al. 1986) of an estimator T at a distribution H measures the effect on T of adding an observation at $\mathbf{z} = (x, y)$. If we denote the point mass at \mathbf{z} by Δ_z then we can write

$$\begin{aligned} IF(\mathbf{z}, T, H) &= \lim_{\varepsilon \downarrow 0} \frac{T((1 - \varepsilon)H + \varepsilon\Delta_z) - T(H)}{\varepsilon} \\ &= \lim_{\varepsilon \downarrow 0} \frac{T(H_\varepsilon) - T(H)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(H_\varepsilon) \Big|_{\varepsilon=0} \end{aligned}$$

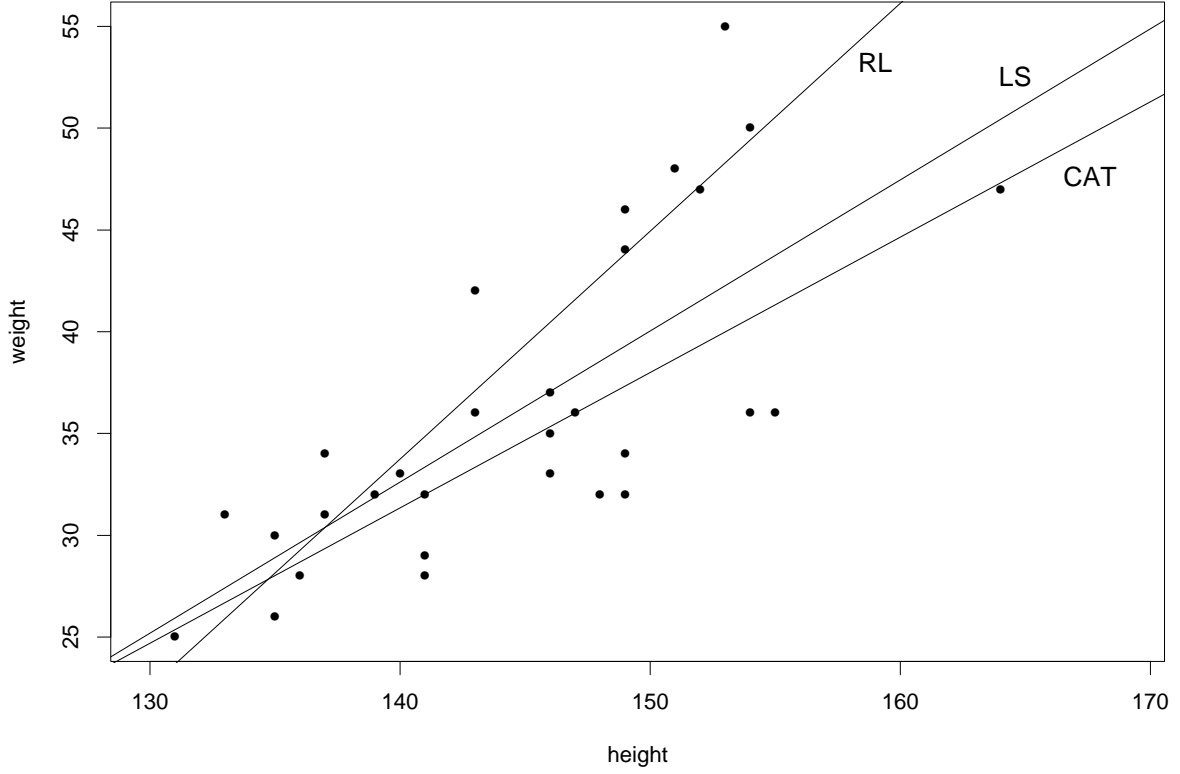


Figure 7: Plot of weight versus height for 30 schoolgirls, and three regression fits: the catline, the least squares fit (LS) and the resistant line (RL).

Theorem 9. Assume that the random variables $(X, Y) \sim H(x, y) = G(x)F(y)$ satisfy model (5.3) with $\alpha = \beta = 0$. Further assume that $E_G[|X|] < \infty$ and that F has a density f which is strictly positive at zero. Define $L =]-\infty, G^{-1}(1/3)]$, $M =]G^{-1}(1/3), G^{-1}(2/3)[$ and $R = [G^{-1}(2/3), +\infty[$ and let $\bar{x}_L := E_G[X|X \in L]$, $\bar{x}_M := E_G[X|X \in M]$ and $\bar{x}_R := E_G[X|X \in R]$. Then

$$\begin{aligned}
 IF((x, y), b_{CAT}, H) &= \frac{-3sgn(y)}{2f(0)(\bar{x}_R - \bar{x}_L)} && \text{if } x \in L \\
 &= 0 && \text{if } x \in M \\
 &= \frac{3sgn(y)}{2f(0)(\bar{x}_R - \bar{x}_L)} && \text{if } x \in R
 \end{aligned} \tag{6.1}$$

and

$$\begin{aligned}
IF((x, y), a_{CAT}, H) &= \frac{-3}{4f(0)} \operatorname{sgn}(y) \frac{(\bar{x}_L + \bar{x}_M)}{(\bar{x}_R - \bar{x}_L)} && \text{if } x \in L \\
&= \frac{3}{4f(0)} \operatorname{sgn}(y) && \text{if } x \in M \\
&= \frac{3}{4f(0)} \operatorname{sgn}(y) \frac{(\bar{x}_M + \bar{x}_R)}{(\bar{x}_R - \bar{x}_L)} && \text{if } x \in R.
\end{aligned} \tag{6.2}$$

Corollary 2. (a) If G is symmetric,

$$\begin{aligned}
IF((x, y), b_{CAT}, H) &= \frac{3}{4f(0)\bar{x}_R} \operatorname{sgn}(y) (I(x \in R) - I(x \in L)) \\
IF((x, y), a_{CAT}, H) &= \frac{3}{4f(0)} \operatorname{sgn}(y) \left(I(x \in M) + \frac{1}{2} I(x \in L \cup R) \right).
\end{aligned}$$

(b) In particular, at the bivariate gaussian distribution $N_2(\mathbf{0}, I)$ we have

$$\begin{aligned}
IF((x, y), b_{CAT}, N_2(\mathbf{0}, I)) &= \frac{\sqrt{2\pi}}{4\phi(\Phi^{-1}(\frac{2}{3}))} \operatorname{sgn}(x) \operatorname{sgn}(y) I(|x| \geq \Phi^{-1}(\frac{2}{3})) \\
IF((x, y), a_{CAT}, N_2(\mathbf{0}, I)) &= \frac{3\sqrt{2\pi}}{4} \operatorname{sgn}(y) \left(I(|x| < \Phi^{-1}(\frac{2}{3})) + \frac{1}{2} I(|x| \geq \Phi^{-1}(\frac{2}{3})) \right).
\end{aligned}$$

Since the influence functions of the slope and the intercept each take on at most 3 different absolute values, they are bounded. Figure 8a shows the influence function of the catline slope at the bivariate gaussian distribution $H = N_2(\mathbf{0}, I)$, where it coincides with that of the resistant line slope. The influence function of the catline intercept at $N_2(\mathbf{0}, I)$ is plotted in Figure 8b.

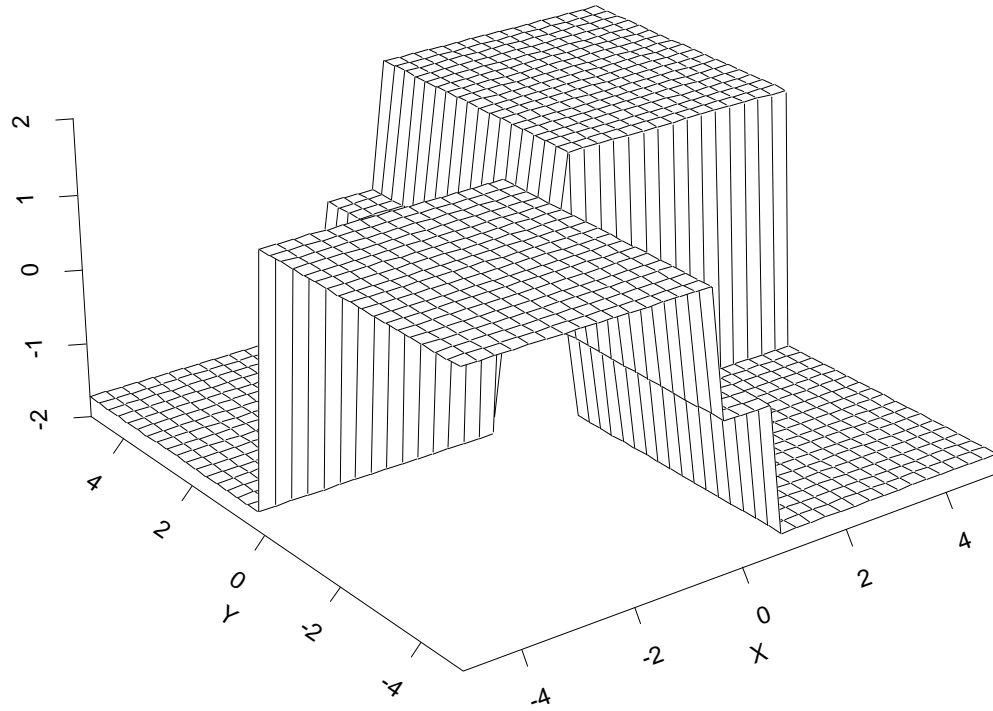
Whereas the influence function is an asymptotic concept, we also want to have a finite-sample version. For this we use the *averaged permutation-stylized sensitivity function* defined by (Rousseeuw et al. 1995). For any estimator T the sensitivity function measures the (standardized) effect of adding an observation \mathbf{z} at the sample $Z_n = \{\mathbf{z}_i; i = 1, \dots, n\}$, i.e.

$$SF_n(\mathbf{z}, T, Z_n) = n(T_{n+1}(\mathbf{z}_1, \dots, \mathbf{z}_n, \mathbf{z}) - T_n(\mathbf{z}_1, \dots, \mathbf{z}_n)). \tag{6.3}$$

The resulting sensitivity function strongly depends on the sample Z_n , but we can alleviate this effect by using a permutation-stylized data set $Z(\pi) = \{(x_i^s, x_{\pi(i)}^s); i = 1, \dots, n\}$ where $x_i^s = \Phi^{-1}(\frac{i}{n+1})$ and where π is a random permutation on $\{1, \dots, n\}$. Finally, the effect of the particular permutation π is tempered by averaging the sensitivity function over a collection of random permutations, leading to

$$APSF_n(\mathbf{z}) = \operatorname{average}_{\pi} SF_n(\mathbf{z}, T, Z(\pi)). \tag{6.4}$$

(a)



(b)

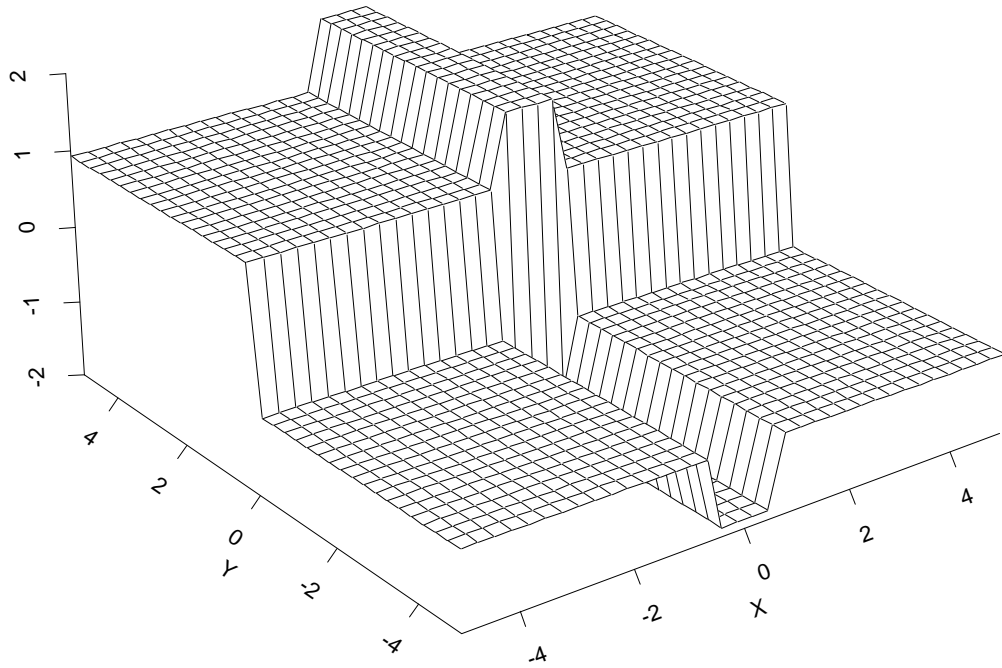


Figure 8: (a) Influence function of the catline slope at the bivariate standard gaussian distribution; (b) influence function of the catline intercept.

Figure 9a shows the sensitivity surface of the catline slope and Figure 9b that of the catline intercept, both for $n = 20$, obtained by generating $m = 350$ random permutations. We see that these smoothed surfaces approximate the asymptotic influence functions quite well.

7 Efficiency

When a functional T and its influence function are sufficiently regular, $\sqrt{n}(T(Z_n) - T(H))$ is asymptotical normal with zero mean and asymptotic variance

$$V(T, H) = \int IF(\mathbf{z}, T, H)^2 dH(\mathbf{z})$$

(see Hampel et al. 1986). Under the assumptions of Theorem 9, we find that

$$\sqrt{n}(b_{CAT}(Z_n) - b_{CAT}(H)) \longrightarrow N\left(0, \frac{3}{2f^2(0)(\bar{x}_R - \bar{x}_L)^2}\right)$$

and

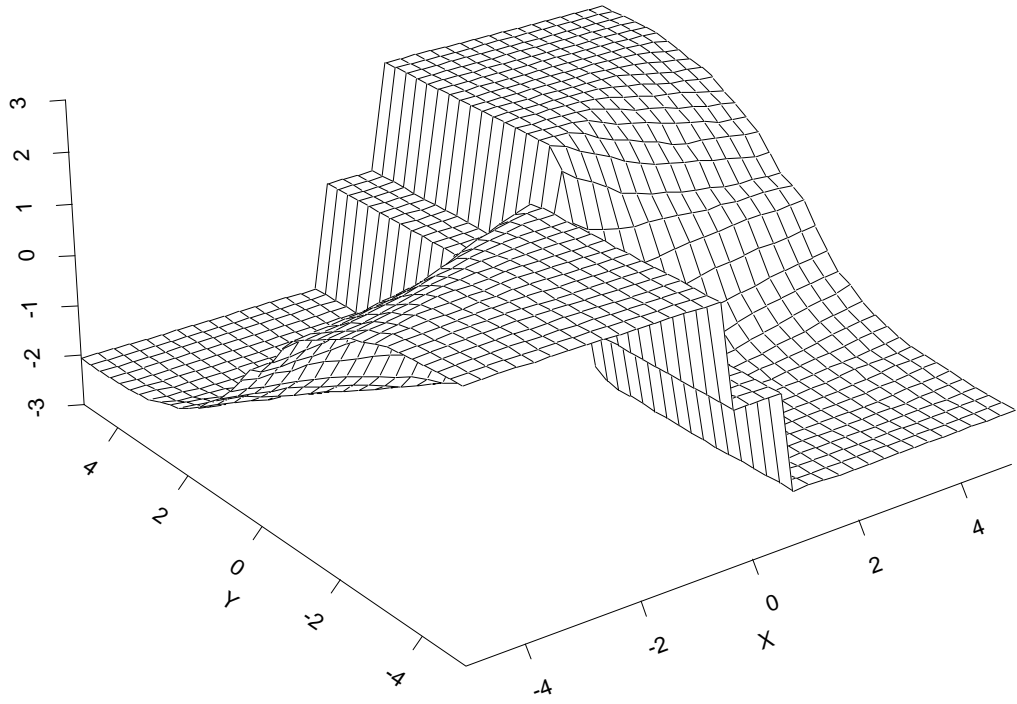
$$\sqrt{n}(a_{CAT}(Z_n) - a_{CAT}(H)) \longrightarrow N\left(0, \frac{3}{16f^2(0)}\left(1 + \frac{(\bar{x}_L + \bar{x}_M)^2 + (\bar{x}_M + \bar{x}_R)^2}{(\bar{x}_R - \bar{x}_L)^2}\right)\right).$$

At the bivariate standard gaussian distribution $H = N_2(\mathbf{0}, I)$ this yields the asymptotic variances $V(b_{CAT}, H) = 1.980$ and $V(a_{CAT}, H) = 1.767$.

Let us compare the performance of the catline with the L^1 line. The asymptotic variance of the L^1 estimator is derived in Bassett and Koenker (1978). For $H = N_2(\mathbf{0}, I)$ we have $V(b_{L^1}, H) = V(a_{L^1}, H) = 1/4\phi(0)^2 = 1.571$. The asymptotic relative efficiency of the catline compared to the L^1 line then becomes $ARE(b_{CAT}, b_{L^1}, H) = V(b_{L^1}, H)/V(b_{CAT}, H) = 79.3\%$ for the slope and $ARE(a_{CAT}, a_{L^1}, H) = V(a_{L^1}, H)/V(a_{CAT}, H) = 88.9\%$ for the intercept. We thus observe a small loss of efficiency at the bivariate gaussian distribution, but the catline achieves a much better resistance to leverage points.

Finally we have investigated whether the corresponding finite-sample relative efficiency is well approximated by the asymptotic relative efficiency. To this end we have generated $m = 10,000$ samples of various sample sizes n (see Table 1) from $N_2(\mathbf{0}, I)$, each time computing their catline $(b_{CAT}^{(k)}, a_{CAT}^{(k)})$ and their L^1 line $(b_{L^1}^{(k)}, a_{L^1}^{(k)})$ for $k = 1, \dots, m$. The catlines were computed with the $O(n \log n)$ algorithm constructed in Section 4. The L^1 regression was performed using the `l1fit` procedure in S-PLUS. For each n , Table 1 lists the average of

(a)



(b)

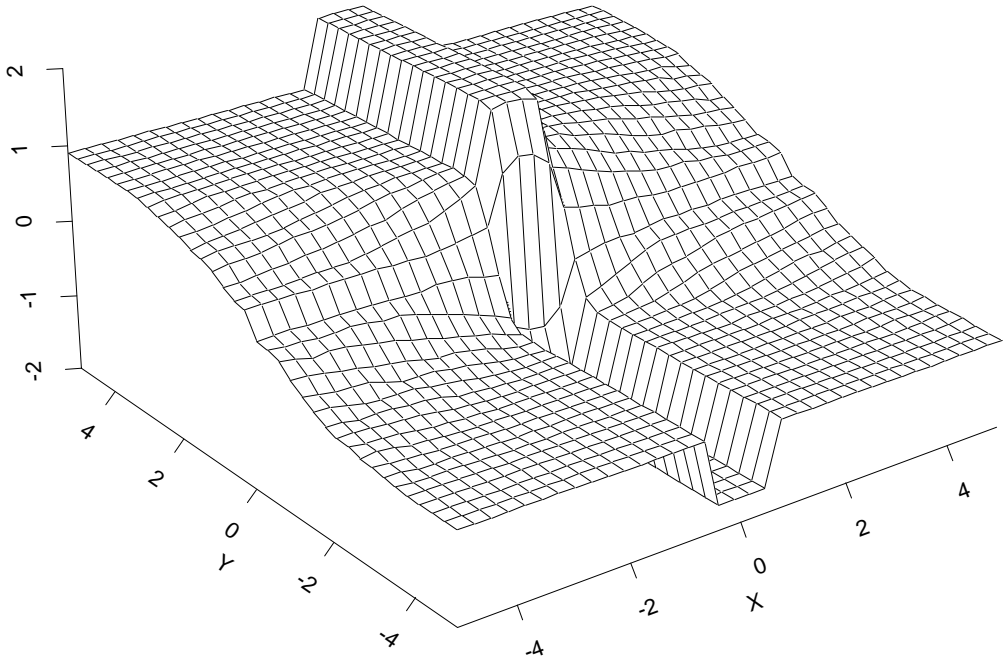


Figure 9: (a) Averaged permutation-stylized sensitivity function $APSF_n$ of the catline slope for $n = 20$; (b) $APSF_n$ of the catline intercept for $n = 20$.

Table 1: Bias of the catline and its finite-sample relative efficiency compared to the L^1 line at $H = N_2(\mathbf{0}, I)$. The simulation results are based on 10,000 samples.

| n | slope b_{CAT} | | intercept a_{CAT} | |
|----------|-----------------|-----------------|---------------------|-----------------|
| | bias | RE(CAT, L^1) | bias | RE(CAT, L^1) |
| 10 | 0.01166 | 82.7% | 0.00267 | 84.5% |
| 20 | 0.00448 | 79.7% | 0.00602 | 88.2% |
| 40 | -0.00137 | 80.5% | 0.00135 | 88.6% |
| 60 | 0.00123 | 79.5% | -0.00243 | 88.8% |
| 100 | 0.00185 | 80.5% | 0.00117 | 86.5% |
| 500 | 0.00119 | 80.3% | 0.00084 | 89.5% |
| 1,000 | -0.00031 | 80.5% | -0.00055 | 89.7% |
| 5,000 | 0.00023 | 78.6% | -0.00020 | 88.9% |
| 10,000 | 0.00001 | 78.2% | -0.00001 | 92.5% |
| 40,000 | -0.00006 | 80.0% | -0.00005 | 87.8% |
| ∞ | 0.00000 | 79.3% | 0.00000 | 88.9% |

the computed catline slopes, as well as the finite-sample relative efficiency

$$RE(b_{CAT}, b_{L^1}, H) = \frac{\text{variance}_{k=1, \dots, m} b_{L^1}^{(k)}}{\text{variance}_{k=1, \dots, m} b_{CAT}^{(k)}}.$$

It also presents the analogous results for the catline intercept. We see that the finite-sample relative efficiency is close to the asymptotic relative efficiency, which makes the asymptotics valid for both small and large data sets.

8 Appendix: Proofs

Proof of Theorem 1: immediate from Definition 4.

Proof of Theorem 2: (a) By Definition 2, the depth of any line is at most n . To prove the lower bound we consider a tilt point v to the left of $I_2 := M \cup R$. Denote by I_2^+ (resp. I_2^-, I_2^0) the number of strictly positive (resp. negative, zero) residuals in I_2 . From the

definition of the catline it follows that $I_2^+ \leq \lfloor n/3 \rfloor, I_2^- \leq \lfloor n/3 \rfloor, I_2^+ + I_2^- + I_2^0 = 2\lfloor n/3 \rfloor$ if $n = 3m$, and $I_2^+ + I_2^- + I_2^0 = 2\lfloor n/3 \rfloor + 1$ if $n \neq 3m$. Let us now tilt the line upward (downward) about v until it becomes vertical. Doing so it passes all points of I_2 with nonpositive (non-negative) residuals, hence at least $\min(I_2^+ + I_2^0, I_2^- + I_2^0) \geq \lfloor n/3 \rfloor$ points. For a tilt point v in M or R the reasoning is analogous, hence $rdepth(\boldsymbol{\theta}_{CAT}, Z_n) \geq \lfloor n/3 \rfloor$.

(b) In the population case the proof is similar, using (3.1).

Proof of Theorem 3: (a) The line with maximal depth has to pass through two observations (otherwise it could be made deeper by slightly tilting it until it does fit 2 points), thereby dividing the observations in three groups with alternating residual signs. The $rdepth$ of this line is then 2 + the size of the smallest group, which is bounded by $2 + \lfloor (n-2)/3 \rfloor = \lfloor (n+2)/3 \rfloor$. Assume the observations to be ordered by their x -coordinates, set $m = \lfloor n/3 \rfloor$ and put $\mathbf{z}_i = (x_i, y_i)$. Then the line through the observations \mathbf{z}_{m+1} and \mathbf{z}_{2m+1} (if $n \neq 3m+2$) and the line through \mathbf{z}_{m+1} and \mathbf{z}_{2m+2} (if $n = 3m+2$) has maximal depth and is also the catline. The proof of part (b) is analogous. Parts (c) and (d) are trivial.

Proof of Theorem 4: Denote $(\hat{b}, \hat{a}) := \boldsymbol{\theta}_{CAT}(\mathbf{x}, \mathbf{y})$. From Theorem 1, the catline is characterized by $L^- + M^- = \#\{i; (x_i, y_i) \in L \cup M \text{ and } r_i = y_i - \hat{b}x_i - \hat{a} < 0\} \leq \lfloor n/3 \rfloor$, and three similar relations. Because of symmetry, we will only consider the first relation. We will denote the groups in a transformed sample by \tilde{L}, \tilde{M} and \tilde{R} . The theorem now follows from the following identities:

1. Denote $(\tilde{b}, \tilde{a}) := \boldsymbol{\theta}_{CAT}(\mathbf{x}, \mathbf{y} + c\mathbf{x} + d)$. If $\#\{i; (x_i, y_i + cx_i + d) \in \tilde{L} \cup \tilde{M} \text{ and } y_i + cx_i + d - \tilde{b}x_i - \tilde{a} < 0\} \leq \lfloor n/3 \rfloor$ then $\#\{i; (x_i, y_i) \in L \cup M \text{ and } y_i - (\tilde{b}-c)x_i - (\tilde{a}-d) < 0\} \leq \lfloor n/3 \rfloor$. This also holds for the three other relations, and thus $\tilde{b} = \hat{b} + c$ and $\tilde{a} = \hat{a} + d$.
2. Denote $(\tilde{b}, \tilde{a}) := \boldsymbol{\theta}_{CAT}(\mathbf{x}, c\mathbf{y})$. First put $c > 0$. If $\#\{i; (x_i, cy_i) \in \tilde{L} \cup \tilde{M} \text{ and } cy_i - \tilde{b}x_i - \tilde{a} < 0\} \leq \lfloor n/3 \rfloor$ then $\#\{i; (x_i, y_i) \in L \cup M, y_i - \frac{\tilde{b}}{c}x_i - \frac{\tilde{a}}{c} < 0\} \leq \lfloor n/3 \rfloor$, and thus $\tilde{b} = c\hat{b}$ and $\tilde{a} = c\hat{a}$. For $c < 0$ it follows that the number of *positive* residuals is bounded by $\lfloor n/3 \rfloor$. All together we also obtain the four required inequalities, although in a different order.
3. Analogous to 1 and 2.
4. We have to show that $(\tilde{b}, \tilde{a}) := \boldsymbol{\theta}_{CAT}(x_i, \hat{y}_i + d_i r_i)$ equals (\hat{b}, \hat{a}) for all $d_i \geq 0, \hat{y}_i =$

$\hat{b}x_i + \hat{a}$, and $r_i = r_i(\boldsymbol{\theta}_{CAT})$. This is true iff $\#\{i; (x_i, \hat{y}_i + d_i r_i) \in \tilde{L} \cup \tilde{M} \text{ and } \hat{y}_i + d_i r_i - \hat{b}x_i - \hat{a} < 0\} \leq [n/3]$. The latter inequality follows from the definition of \hat{y}_i, r_i, d_i and the fact that $(x_i, y_i) \in L \cup M$.

Proof of Theorem 6: immediate from Definition 4.

Proof of Theorem 7: The Fisher-consistency of b_{CAT} follows from $\text{med}(Y - \beta X | X \in I_1) = \text{med}(\alpha + e | X \in I_1) = \alpha + \text{med}(e | X \in I_1) = \alpha = \alpha + \text{med}(e | X \in I_2) = \text{med}(\alpha + e | X \in I_2) = \text{med}(Y - \beta X | X \in I_2)$ thus $b_{CAT} = \beta$. Then $\text{med}(Y - b_{CAT}X | X \in I_1) = \alpha + \text{med}(e | X \in I_1) = \alpha$, thus also $a_{CAT} = \alpha$.

Proof of Theorem 8: First we will prove that we can make the slope of the catline arbitrarily steep by replacing $n\varepsilon_n^*$ observations of the original data set Z_n . Assume the data points are ordered by their x -coordinates. By Theorem 4(c) we may assume that all $x_i > 0$ w.l.o.g. We consider all lines through the observation with x -coordinate $x_{2m+nm \bmod 3}$ that separates the groups M and R . This means that all points in M resp. R have a positive, resp. negative residual. (For $n = 3m$ we require these lines to yield at least one positive residual in L .) Denote by Θ the set of the slopes and intercepts of these lines, and consider (b_1, a_1) its element with smallest positive slope. It is clear that $b_1 < \infty$ and $\sup_{\theta \in \Theta} b = \infty$. Moreover, $|a_1| < \infty$ and $\inf_{\theta \in \Theta} a = -\infty$. Now consider any $v > 0$ and $w < 0$. Then take a line (b, a) in Θ such that $b > \max\{b_1, v\}$ and $a < \min\{a_1, w\}$, and put all observations of L below this line (except if $n = 3m$, then we don't touch the point with positive residual). The resulting data set now has a catline (b, a) with arbitrarily large slope and arbitrarily small intercept.

Next, consider a dataset Z'_n obtained by replacing $n\varepsilon_n^* - 1$ observations from the original dataset Z_n . We will then show that $(b'_{CAT}, a'_{CAT}) = (b_{CAT}, a_{CAT})(Z'_n)$ remains bounded. Denote by L', M' and R' the three subsets of Z'_n . Denote by $\tilde{Z} = Z_n \cap Z'_n$ the set of original data points in Z'_n . Further set $L = \tilde{Z} \cap L', M = \tilde{Z} \cap M'$ and $R = \tilde{Z} \cap R'$. Assume $n = 3m$. Now consider \mathcal{Z}''_n , the collection of all data sets Z''_n obtained by adding $m - 2$ points to \tilde{Z} such that $L \subset L'', M \subset M''$ and $R \subset R''$. Set Θ the collection of all catlines of \mathcal{Z}''_n . Since $Z'_n \in \mathcal{Z}''_n$ it is clear that $(b'_{CAT}, a'_{CAT}) \in \Theta$, and thus b'_{CAT} will be bounded if the first coordinates of Θ are bounded. Take any line (b, a) in Θ , then denote by L^+ (resp. L^-, L^0) the number of strictly positive (resp. negative, zero) residuals in L (and use the analogous notations for M and R).

If this line partitions the points in \tilde{Z} such that L^+, L^-, R^+ , and R^- are all strictly positive, then it is clear that b is bounded. An unbounded slope can only be attained if (w.l.o.g.) $L^+ = L^0 = R^- = R^0 = 0$ and $M^0 \leq 1$ or if $M^- = M^0 = R^- = R^0 = 0$ and $L^0 \leq 1$. (The other situations follow by symmetry). In the first case, $L^- + M^+ + M^- + M^0 + R^+ = 2m + 2$, but $L^- + M^- \leq m$ and $M^+ + R^+ \leq m$ (since (b, a) bisects $L'' \cup M''$ and $M'' \cup R''$). In the second case, $L^+ + L^- + L^0 + M^+ + R^+ = 2m + 2$, whereas $L^+ + L^- + L^0 \leq m$ and $M^+ + R^+ \leq m$, again a contradiction. If $n \neq 3m$ we can write analogous relations, keeping in mind that a catline will then pass through at least one observation. Finally note that the second coordinates of Θ are also bounded. By definition, the intercept is bounded if $m + 1 = [n/3] + 1$ of the residuals $y_i - bx_i$ in $L'' \cup M''$ and in $M'' \cup R''$ are bounded. This is always satisfied since b is bounded and $\#L + \#M \geq m + 1$ and $\#M + \#R \geq m + 1$ (otherwise $\#R \geq m + 2$ or $\#L \geq m + 2$).

Proof of Corollary 1: From Theorem 4 we know that the catline is regression equivariant as well as scale equivariant. For any such estimator Rousseeuw and Leroy (1987, pages 123-124) showed that the breakdown property implies the exact fit property.

Proof of Theorem 9: To derive the influence function of the catline, we will use a different (but equivalent) functional form for T_{CAT} . For any (b, a) , for any bivariate distribution H and for any interval $J \in \{L, M, R\}$ we denote

$$P_{(b,a),\varepsilon}^J(t) = P(Y - bX - a \leq t \mid (X, Y) \sim H_\varepsilon \text{ and } X \in J),$$

and

$$\tilde{P}_{(b,a),\varepsilon}^J(t) = P(-(Y - bX - a) \leq t \mid (X, Y) \sim H_\varepsilon \text{ and } X \in J).$$

Let

$$G_1(t, b, a, \varepsilon) = \frac{1}{2}P_{(b,a),\varepsilon}^L(t) + \frac{1}{2}\tilde{P}_{(b,a),\varepsilon}^R(t) \quad (8.1)$$

and

$$G_2(t, b, a, \varepsilon) = \frac{1}{4}P_{(b,a),\varepsilon}^L(t) + \frac{1}{2}P_{(b,a),\varepsilon}^M(t) + \frac{1}{4}P_{(b,a),\varepsilon}^R(t), \quad (8.2)$$

then the functional $T_{CAT}(H)$ satisfies

$$\text{med}_t G_1(t, b_{CAT}, a_{CAT}, 0) = 0 \quad (8.3)$$

and

$$\text{med}_t G_2(t, b_{CAT}, a_{CAT}, 0) = 0. \quad (8.4)$$

These relations can easily be derived from (3.1). Equation (8.3) says that the union of the residuals in L and minus the residuals in R has zero median, and according to (8.4) the same is true for the (formal) ‘union’ of the residuals in $L \cup M \cup M \cup R$.

First we derive the influence function of the slope b_{CAT} . Denote $b(\varepsilon) := b_{CAT}(H_\varepsilon)$ and $a(\varepsilon) := a_{CAT}(H_\varepsilon)$. Because of Fisher-consistency at the model (5.3), $b(0) = \beta = 0$ and $a(0) = \alpha = 0$. Since $G_1(0, b(\varepsilon), a(\varepsilon), \varepsilon) = \frac{1}{2}$ we have $\frac{\partial G_1}{\partial \varepsilon}(0, b(\varepsilon), a(\varepsilon), \varepsilon)|_{\varepsilon=0} = 0$, hence

$$\begin{aligned} \frac{\partial G_1}{\partial b}(0, b(\varepsilon), a(\varepsilon), \varepsilon)|_{\varepsilon=0} \frac{db}{d\varepsilon}(\varepsilon)|_{\varepsilon=0} + \frac{\partial G_1}{\partial a}(0, b(\varepsilon), a(\varepsilon), \varepsilon)|_{\varepsilon=0} \frac{da}{d\varepsilon}(\varepsilon)|_{\varepsilon=0} \\ + \frac{\partial G_1}{\partial \varepsilon}(0, b(\varepsilon), a(\varepsilon), \varepsilon)|_{\varepsilon=0} = 0. \end{aligned}$$

Therefore

$$IF(\mathbf{z}, b_{CAT}, H) = \frac{db}{d\varepsilon}(\varepsilon)|_{\varepsilon=0} = \frac{-\frac{\partial G_1}{\partial \varepsilon} - \frac{\partial G_1}{\partial a} \frac{da}{d\varepsilon}}{\frac{\partial G_1}{\partial b}}|_{\varepsilon=0}. \quad (8.5)$$

Now

$$\frac{\partial G_1}{\partial \varepsilon}|_{t=0, \varepsilon=0} = \frac{1}{2} \frac{\partial P_{(b,a),\varepsilon}^L(t)}{\partial \varepsilon}|_{\substack{t=0, \varepsilon=0 \\ (b,a)=(0,0)}} + \frac{1}{2} \frac{\partial \tilde{P}_{(b,a),\varepsilon}^R(t)}{\partial \varepsilon}|_{\substack{t=0, \varepsilon=0 \\ (b,a)=(0,0)}}.$$

Since

$$P_{(b,a),\varepsilon}^L(t) = (1 - \delta_{\varepsilon,L})P_{(b,a),0}^L(t) + \delta_{\varepsilon,L}I(y - bx \leq t)$$

with $\delta_{\varepsilon,L} = \frac{3\varepsilon I(x \in L)}{(1-\varepsilon)+3\varepsilon I(x \in L)}$ and I the indicator function (see also Johnstone and Velleman 1985, page 1052), and since $P_{(b,a),0}^L(t)|_{\substack{t=0 \\ (b,a)=(0,0)}} = P(\varepsilon \leq 0 \mid (X, Y) \sim H \text{ and } X \in L) = \frac{1}{2} = \tilde{P}_{(b,a),0}^R(t)|_{\substack{t=0 \\ (b,a)=(0,0)}}$, we can verify that

$$\frac{\partial G_1}{\partial \varepsilon}|_{t=0, \varepsilon=0} = \frac{1}{2} \left(-\frac{3}{2}I(x \in L) + 3I(x \in L)I(y \leq 0) - \frac{3}{2}I(x \in R) + 3I(x \in R)I(y \geq 0) \right). \quad (8.6)$$

Next, we evaluate $\frac{\partial G_1}{\partial a}|_{t=0, \varepsilon=0}$. As $P_{(b,a),0}^L(t)|_{b=0} = P(Y - a \leq t \mid X \in L) = F(a + t)$, we have $\frac{\partial P_{(b,a),0}^L(t)}{\partial a}|_{\substack{t=0 \\ (b,a)=(0,0)}} = f(0)$, and analogously $\frac{\partial \tilde{P}_{(b,a),0}^R(t)}{\partial a}|_{\substack{t=0 \\ (b,a)=(0,0)}} = -f(0)$. Consequently,

$$\frac{\partial G_1}{\partial a}|_{t=0, \varepsilon=0} = 0. \quad (8.7)$$

Finally, we compute the denominator of (8.5). Now we find

$$\begin{aligned}
\frac{\partial P_{(b,a),0}^L(t)}{\partial b} \Big|_{(b,a)=(0,0)}^{t=0} &= 3 \frac{\partial}{\partial b} \int_{x \in L} \int_{-\infty}^{\infty} 1_{(y-bx \leq t)} dH(x, y) \Big|_{t=0, b=0} \\
&= 3 \int_{x \in L} \int_{-\infty}^{\infty} x 1_{(y-bx=t)} dH(x, y) \Big|_{t=0, b=0} \\
&= 3 \int_{x \in L} x \int_{-\infty}^{\infty} 1_{(e=0)} dF(e) dG(x) \\
&= 3 \int_{x \in L} x f(0) dG(x) = \bar{x}_L f(0),
\end{aligned}$$

and analogously $\frac{\partial \tilde{P}_{(b,a),0}^R(t)}{\partial b} \Big|_{(b,a)=(0,0)}^{t=0} = \bar{x}_R f(0)$, hence

$$\frac{\partial G_1}{\partial b} \Big|_{t=0, \varepsilon=0} = -\frac{1}{2} f(0) (\bar{x}_R - \bar{x}_L). \tag{8.8}$$

Combining (8.6), (8.7), (8.8) and (8.5) then yields (6.1).

The influence function of the intercept a_{CAT} is derived from (8.4), yielding

$$IF(\mathbf{z}, a_{CAT}, H) = \frac{-\frac{\partial G_2}{\partial \varepsilon} - \frac{\partial G_2}{\partial b} \frac{db}{d\varepsilon}}{\frac{\partial G_2}{\partial a}} \Big|_{\varepsilon=0}. \tag{8.9}$$

Now for each $J \in \{L, M, R\}$ it holds that

$$\begin{aligned}
\frac{\partial P_{(b,a),\varepsilon}^J(t)}{\partial \varepsilon} \Big|_{(b,a)=(0,0)}^{t=0, \varepsilon=0} &= -\frac{3}{2} I(x \in J) + 3I(x \in J)I(y \leq 0), \\
\frac{\partial P_{(b,a),\varepsilon}^J(t)}{\partial a} \Big|_{(b,a)=(0,0)}^{t=0, \varepsilon=0} &= f(0) \quad \text{and} \quad \frac{\partial P_{(b,a),\varepsilon}^J(t)}{\partial b} \Big|_{(b,a)=(0,0)}^{t=0, \varepsilon=0} = f(0) \bar{x}_J.
\end{aligned}$$

Inserting these three relations, (8.2) and (6.1) in (8.9) then yields (6.2).

References

- [1] Bassett, G., and Koenker, R. (1978). Asymptotic theory of least absolute error regression. *J. Amer. Statist. Assoc.* **73** 618-621.
- [2] Box, G.E.P., and Cox, D.R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B.* **26** 211-246.
- [3] Brent, R.P. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs.

- [4] Brown, G.W., and Mood, A.M. (1951). On median tests for linear hypotheses. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 159-166. University of California Press, Berkeley.
- [5] Cade, B.S., and Richards, J.D. (1996). Permutation tests for least absolute deviation regression. *Biometrics* **52** 886-902.
- [6] Carroll, R.J., and Ruppert, D. (1988). *Transformation and Weighting in Regression*, Chapman and Hall, New York.
- [7] Carroll, R.J., and Welsh, A.H. (1988). A note on asymmetry and robustness in linear regression. *The American Statistician* **42** 285-287.
- [8] Cole, R. (1984). Slowing down sorting networks to obtain faster sorting algorithms. In *Proceedings of the 25th Annual IEEE Symp. Found. Comp. Sci.*, pp. 225-260.
- [9] Dekker, T.J. (1969). Finding a zero by means of successive linear interpolation. In *Constructive Aspects of the Fundamental Theorem of Algebra* (B. Dejon and P. Henrici, Ed.) John Wiley, New York.
- [10] Donoho, D.L., and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20** 1803-1827.
- [11] Edelsbrunner, H. (1987). *Algorithms in Combinatorial Geometry*. Springer-Verlag, Berlin.
- [12] Edelsbrunner, H., and Waupotitsch, R. (1986). Computing a ham-sandwich cut in two dimensions. *J. Symb. Comput.* **2** 171-178.
- [13] Johnstone, I.M., and Velleman, P.F. (1985). The resistant line and related regression methods. *J. Amer. Statist. Assoc.* **80** 1041-1054.
- [14] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: the Approach based on Influence Functions*. John Wiley, New York.
- [15] Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J., and Ostrowski, E. (1994). *A Handbook of Small Data Sets*. Chapman and Hall, New York.

- [16] He, X., and Wang, G. (1997). Convergence of depth contours for multivariate datasets. *Ann. Statist.* **25** 495-504.
- [17] Liu, R.Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405-414.
- [18] Liu, R.Y. (1995). Control charts for multivariate processes. *J. Amer. Statist. Assoc.* **90** 1380-1387.
- [19] Rousseeuw, P.J., Croux, C., and Hössjer, O. (1995). Sensitivity functions and numerical analysis of the repeated median slope. *Computat. Statist.* **10** 71-90.
- [20] Rousseeuw, P.J., and Hubert, M. (1996). Regression depth. Submitted.
- [21] Rousseeuw, P.J., and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. John Wiley, New York.
- [22] Tukey, J.W. (1975). Mathematics and the picturing of data. *Proc. Intern. Congr. Math.* **2** 523-531. Vancouver.
- [23] Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading, Mass.
- [24] Wilkinson, J.H. (1967). Two algorithms based on successive linear interpolation. Technical Report STAN-CS-67-60, Computer Science Dept., Stanford University.